MBE Advance Access published December 24, 2014

Coalescent Methods Are Robust to the Simultaneous Effects of Long Branches and Incomplete Lineage Sorting

Liang Liu,*[†,1,2] Zhenxiang Xi,[†,3] and Charles C. Davis[3]

[1]Department of Statistics, University of Georgia

[2]Institute of Bioinformatics, University of Georgia

[3]Department of Organismic and Evolutionary Biology, Harvard University

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: lliu@uga.edu.

Associate editor: Hideki Innan

Abstract

It is well known that species with elevated substitution rates can give rise to disproportionately long branches in the species tree. This combination of long and short branches can contribute to long-branch artifacts (LBA). Despite efforts to remedy LBA via increased taxon sampling and methodological improvements in gene tree estimation, it remains unclear how long and short branches affect species tree estimation in the presence of incomplete lineage sorting (ILS). Here, we examine the combined influence of long external and short internal branches on concatenation and coalescent methods using both simulated and empirical data. Our results demonstrate that the presence of long and short branches alone does not obviously confound the consistency of concatenation and coalescent methods. However, when long external and short internal branches occur simultaneously with high ILS, concatenation methods can be misled, especially when two of these long branches are sister lineages. In contrast, coalescent methods are more robust under these circumstances. This is particularly relevant because this topological pattern also characterizes numerous ancient rapid radiations across the tree of life. Because short internal branches can increase the potential for ILS and gene tree discordance, our results collectively suggest that coalescent methods are more likely to infer the correct species tree in cases of ancient rapid radiations where long external and short internal branches are in close phylogenetic proximity.

Key words: **ancient rapid radiation, coalescent methods, concatenation methods, incomplete lineage sorting, long-branch artifacts.**

Introduction

Long-branch artifacts (LBA) are well known in phylogenetic inference (Felsenstein 1978). The term LBA usually refers to conditions under which statistical inconsistency in the inference method arises due to a combination of long and short branches (Huelsenbeck and Hillis 1993; Hillis et al. 1994; Sanderson et al. 2000; Bergsten 2005). The effect of LBA, in which unrelated species are incorrectly placed together due to misinterpretation of homoplastic characters, has been extensively studied in the context of gene tree estimation using both simulated and empirical data (Huelsenbeck and Hillis 1993; Hillis et al. 1994; Huelsenbeck 1995; Lyons-Weiler and Hoelzer 1997; Siddall and Whiting 1999; Sanderson et al. 2000; Anderson and Swofford 2004; Kolaczkowski and Thornton 2009; Kück et al. 2012). These studies have shown that LBA may occur when the model used in gene tree estimation is misspecified, and is especially pervasive under circumstances where substitution rates are elevated or taxon sampling is sparse (Bergsten 2005; Wiens 2005).

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 that span tens of millions of years has relied mostly on concatenation methods, in which phylogenies are inferred from a single matrix assembled by concatenating hundreds, or even thousands, of genes. This so-called "combined-analysis" or "total-evidence" approach (Kluge 1989; William and Ballard 1996; de Queiroz and Gatesy 2007) has been applied in numerous recent phylogenomic studies of animals (Dunn et al. 2008; Hejnol et al. 2009; Regier et al. 2010; Philippe, Brinkmann, Copley, et al. 2011; Smith et al. 2011), bacteria (Wu et al. 2009), plants (Finet et al. 2010; Lee et al. 2011; Wodniok et al. 2011; Timme et al. 2012), and yeasts (Hess and Goldman 2011; Salichos and Rokas 2013). These analyses implicitly assume that all genes have the same, or very similar, evolutionary histories. Empirical studies, however, have shown that incomplete lineage sorting (ILS), a major source of gene tree discordance, is perhaps common across the tree of life (Pollard et al. 2006; Degnan and Rosenberg 2009; Song et al. 2012; Wall et al. 2013; Kutschera et al. 2014). In addition, theoretical and simulation studies have shown that concatenation methods can yield misleading results, especially if the species tree is in an "anomaly zone" (Kubatko and Degnan 2007; Liu and Edwards 2009). This zone is a region of branch length space in the species tree characterized

Article

by very short internal branches in which the most frequently produced gene tree differs from the species tree topology (Degnan and Rosenberg 2006). In other words, it is a species tree characterized by a rapid radiation and/or large effective population sizes where ILS is high (Pamilo and Nei 1988; Degnan and Rosenberg 2009).

Recently developed coalescent-based methods permit gene trees to have different evolutionary histories (Liu, Yu, Kubatko, et al. 2009). Some of these coalescent methods, such as *BEAST (Heled and Drummond 2010) and BEST (Liu 2008), simultaneously estimate the gene trees and species tree. These coestimation 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, Bayzid, et al. 2014). Other gene-tree-based coalescent methods as implemented in MP-EST (Liu et al. 2010), STELLS (Wu 2012), and STEM (Kubatko et al. 2009) infer the species tree from a set of estimated gene trees. In addition, some of the recently developed consensus methods, such as ASTRAL (Mirarab, Reaz, et al. 2014), NJ_{st} (Liu and Yu 2011), STAR (Liu, Yu, Pearl, et al. 2009), and STEAC (Liu, Yu, Pearl, et al. 2009), estimate the species tree using summary statistics from the estimated gene trees. Although these 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 (Mirarab, Bayzid, et al. 2014). Thus, for simplicity, we refer to these consensus methods as gene-tree-based coalescent methods. Both simulation and empirical studies have demonstrated that coalescent methods better accommodate gene tree discordance due to ILS (Liu, Yu, Kubatko, et al. 2009; Liu, Yu, Pearl, et al. 2009; Liu et al. 2010; Song et al. 2012; Zhong et al. 2013; Mirarab, Bayzid, et al. 2014). Moreover, two recent phylogenomic studies have demonstrated that coalescent methods are more robust to elevated nucleotide substitution rates than concatenation methods using maximum likelihood (ML) (Xi et al. 2013, 2014). Despite their potential promise to the field of phylogenomics (Edwards 2009; Liu, Yu, Kubatko, et al. 2009), however, coalescent methods have not fully emerged as a major method of phylogenomic analyses, especially for phylogenetic questions spanning deep evolutionary time.

In most simulation studies, the performance of concatenation versus coalescent methods has been investigated under the assumption that the true species tree is ultrametric, and thus all branches have the same mutation rate (Kubatko and Degnan 2007; Liu and Edwards 2009; Leaché and Rannala 2011; Bayzid and Warnow 2013). However, this is seldom the case for empirical data. Rates of molecular evolution often vary widely between species (Sanderson 2002; Smith and Donoghue 2008; Lanfear et al. 2010), and species with elevated substitution rates can lead to disproportionately long branches in a species tree (all branch lengths used for this study are in mutation units). Thus, it is critical to understand the combined influence of long and short branches in species tree estimation. ML is less commonly affected by LBA in gene tree estimation if the model assumptions are not violated (Felsenstein 1978; Huelsenbeck 1995; Pol and Siddall 2001; Bergsten 2005). When there is only a negligible amount of conflict in gene tree topologies, that is, when ILS is low, the combination of long and short branches should not overly impact ML analyses of concatenated gene sequences (Pol and Siddall 2001; Kück et al. 2012). However, because short internal branches in a species tree can increase the potential for ILS and gene tree discordance (Rannala and Yang 2003; Degnan and Rosenberg 2006, 2009), it is important to understand how the combination of long external and short internal branches will affect the performance of concatenation versus coalescent methods in the presence of high ILS. Importantly, these conditions also characterize numerous ancient rapid radiations across the tree of life, that is, an initial burst in diversification followed by long descendant branches of extant lineages (Whitfield and Lockhart 2007; Whitfield and Kjer 2008). The effect of LBA under these circumstances could be further exacerbated because phylogenomic analyses frequently sample a single, or only a small number of, species for large clades to resolve such ancient rapid radiations (e.g., to minimize cost and optimize analysis efficiency). This sampling bias may artificially create long external branches that might otherwise be remedied with increased taxon sampling (Stefanović et al. 2004; Brinkmann et al. 2005; Lartillot et al. 2007; Pick et al. 2010; Xi et al. 2012).

Here, we explored the impact of long external and short internal branches on species tree estimation under varying degrees of ILS. Our simulation analyses demonstrate that the presence of long and short branches alone does not obviously confound the consistency of concatenation and coalescent methods. With low ILS, concatenation methods are more likely to recover the true species tree when the number of genes is small or the external branches are extremely long. However, when long external and short internal branches occur simultaneously with high ILS, concatenation methods can be misled especially when two of these long branches are sister lineages. In contrast, coalescent methods are more robust under these circumstances. To further investigate this phenomenon, we additionally analyzed an empirical data set of Scrotifera mammals, which represents an ancient rapid radiation. Our species tree estimation using this phylogenomic data corroborates our simulation results, and indicates that in the presence of high ILS, the combination of long external and short internal branches can lead to the failure of concatenation methods, but have less adverse effects on coalescent methods.

Results and Discussion

Simulated DNA Sequence Analyses

Simulation analyses of five-taxon species trees Q1–Q6 (fig. 1) demonstrate that when θ is low (i.e., 0.0001 for species trees Q1 and Q2), all simulated gene trees (when rooted with species *E*) are congruent with the species tree topology. When θ increases (i.e., 0.001 for species trees Q3 and Q4), on average 83% of the simulated gene trees are congruent with the species tree topology. When θ is high (i.e., 0.01 for species trees Q5 and Q6), the topologies of simulated gene trees are highly variable. Under these circumstances, on average only 20% of

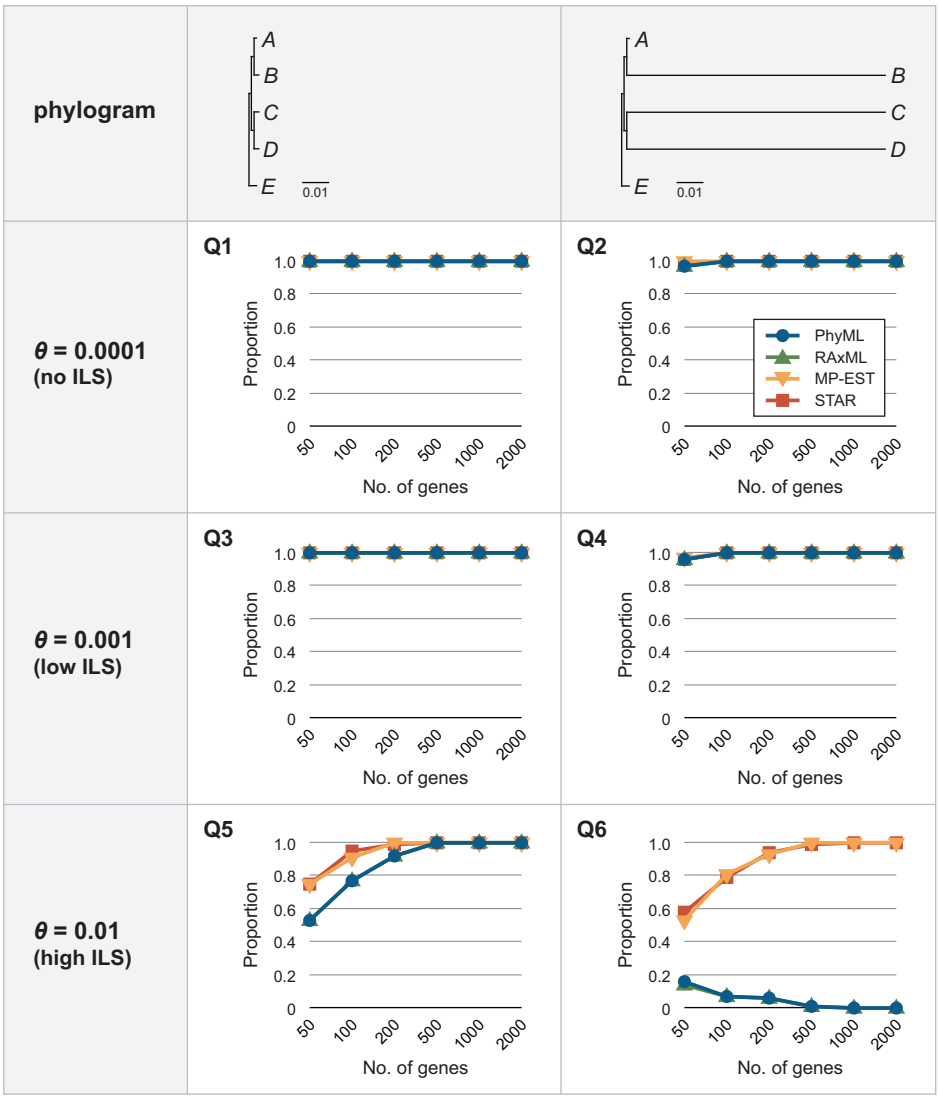

Fig. 1. DNA simulations using five-taxon species trees to investigate the impact of long external and short internal branches on concatenation (PhyML and RAxML) and coalescent (MP-EST and STAR) methods in the presence of ILS. DNA sequences were simulated on species trees Q1–Q6 under the multispecies coalescent model (Rannala and Yang 2003). The lengths of the internal branches are set to 0.001 for species trees Q1–Q6 (branch lengths are in mutation units). The lengths of the external branches leading to species *A* and *E* are set to 0.001 and 0.003, respectively, for species trees Q1–Q6. The lengths of three external branches leading to species *B*–*D* are set to 0.002 for species trees Q1, Q3, and Q5 and 0.101 for species trees Q2, Q4, and Q6. In addition, the population size parameter θ is set to 0.0001 for species trees Q1 and Q2, 0.001 for species trees Q3 and Q4, and 0.01 for species trees Q5 and Q6. Thus, these five-taxon species trees target specific cases where 1) only long external and short internal branches are present (the species tree Q2), 2) only high ILS is present (the species tree Q5), and 3) long external and short internal branches occur simultaneously with high ILS (the species tree Q6). Results shown here represent the proportion of simulations in which concatenation and coalescent methods recover the true species tree.

the simulated gene trees are congruent with the species tree topology. Importantly, despite the highly discordant topologies among gene trees, the most probable gene tree still matches the species tree topology. Thus, species trees Q5 and Q6, which are central to our subsequent discussion, are not in the anomaly zone.

When there are no long branches in the species trees (i.e., species trees Q1, Q3, and Q5; fig. 1), both concatenation (PhyML and RAxML) and coalescent (MP-EST and STAR) methods can accurately estimate the true species tree as the number of genes increases. In addition, when ILS is low (i.e., $\theta = 0.001$ for the species tree Q3), the proportion of simulations in which both methods recover the true species tree is 1.0 regardless of gene number. When ILS is high (i.e., $\theta = 0.01$ for the species tree Q5), however, more than 500 genes are required for both concatenation and coalescent methods to recover the true species tree with a proportion of 1.0. These results indicate that when there is a high degree of gene tree discordance, both methods require more genes to accurately estimate the species tree. For species trees with three long external branches but low ILS (i.e., $\theta = 0.001$ for the species tree Q4; fig. 1), both concatenation and coalescent methods recover the true species tree with a proportion of $\geq$0.96. These results suggest that in the presence of low ILS, neither method is adversely affected by the combination of long external and short internal branches in the species tree.

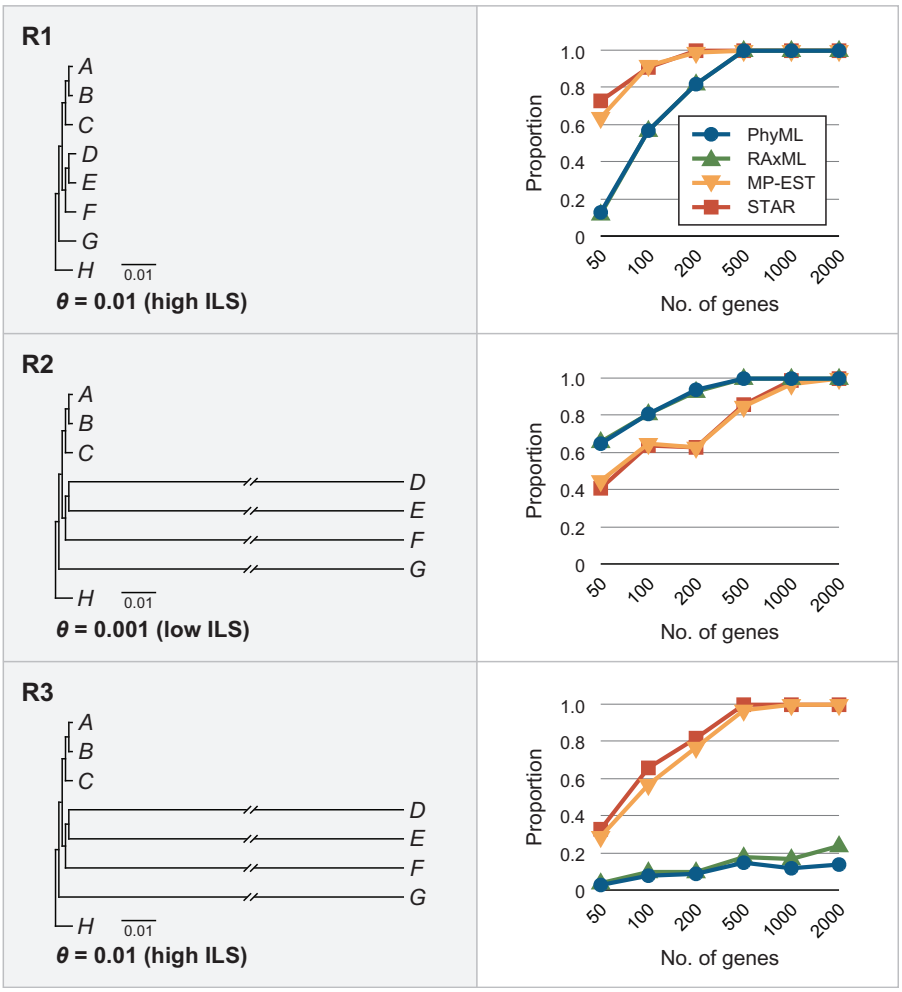

Fig. 2. DNA simulations using eight-taxon species trees to investigate the impact of long external and short internal branches on concatenation (PhyML and RAxML) and coalescent (MP-EST and STAR) methods in the presence of ILS. DNA sequences were simulated on species trees R1–R3 under the multispecies coalescent model (Rannala and Yang 2003). The lengths of the internal branches are set to 0.001 for species trees R1–R3 (branch lengths are in mutation units). The lengths of the external branches leading to species *A*, *B*, *C*, and *H* are held constant in species trees R1–R3 (0.001, 0.001, 0.002, and 0.005, respectively). For the species tree R1, the four external branches leading to species *D*–*G* are short (0.002, 0.002, 0.003, and 0.005, respectively), whereas for species trees R2 and R3, these four external branches are long (0.201, 0.201, 0.202, and 0.204, respectively). In addition, the population size parameter θ is set to 0.001 for species tree R2 and 0.01 for species trees R1 and R3. Thus, these eight-taxon species trees target specific cases where 1) only high ILS is present (the species tree R1), 2) long external and short internal branches occur simultaneously with low ILS (the species tree R2), and 3) long external and short internal branches occur simultaneously with high ILS (the species tree R3). Results shown here represent the proportion of simulations in which concatenation and coalescent methods recover the true species tree.

However, concatenation and coalescent methods differ sharply when long external and short internal branches occur simultaneously with high ILS (i.e., the species tree Q6; fig. 1). Under these circumstances, coalescent methods recover the true species tree with a proportion of $\geq$0.99 as the number of genes increases to 500. In contrast, the proportion of simulations in which concatenation methods recover the true species tree is very low ($\leq$0.16), and declines to 0 as the number of genes increases to 1,000. In these cases, even though the topology of the true species tree is symmetrical, concatenation methods consistently infer two pectinate species trees as the number of genes increases (fig. 5). These results indicate that in the presence of high ILS, the combination of long external and short internal branches can lead to the failure of concatenation methods, even when the true species tree is not in the anomaly zone.

To explore if the choice of species number in our first simulation analyses affects the performance of concatenation and coalescent methods, we also simulated DNA sequences on three 8-taxon species trees (fig. 2). When only high ILS is present (i.e., the species tree R1) or long external and short internal branches occur simultaneously with low ILS (i.e., the species tree R2), both concatenation and coalescent methods can accurately estimate the species tree as the number of genes increases. In these cases, coalescent methods recover the true species tree with a higher proportion when ILS is high and the number of genes is less than 500. In contrast, concatenation methods recover the true species tree with a

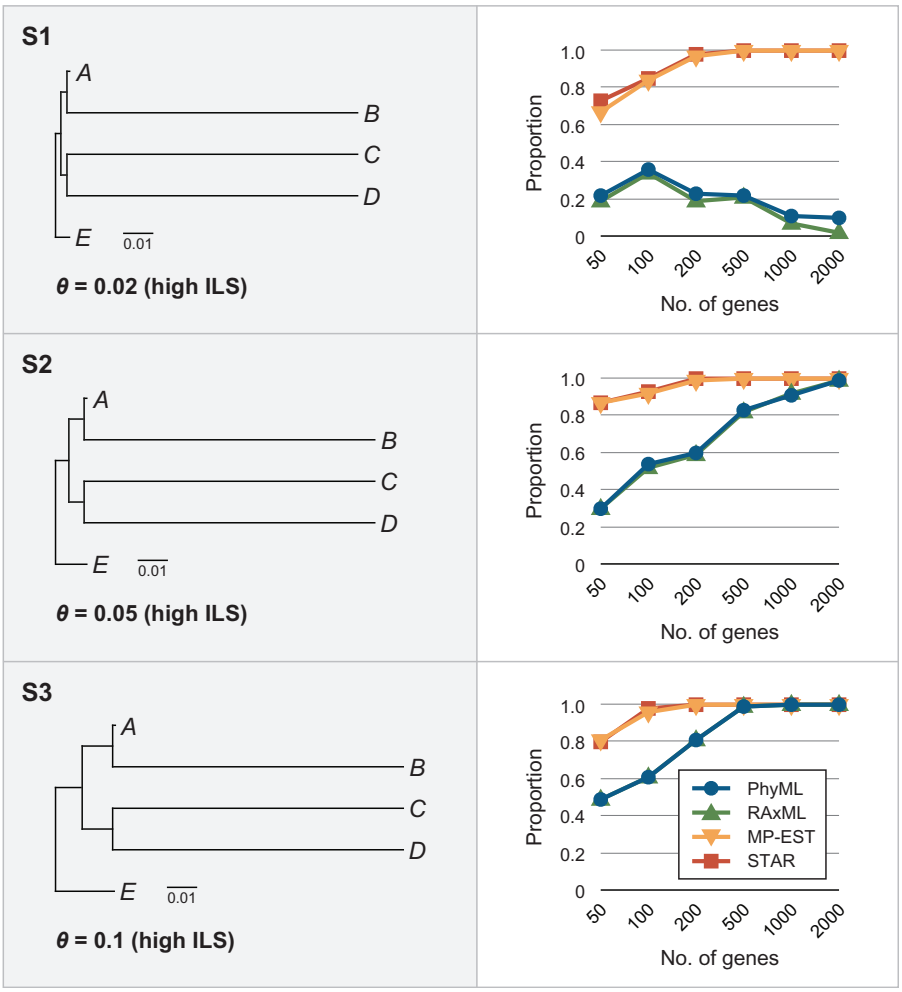

Fig. 3. DNA simulations to investigate how varying the length of the internal branches affects concatenation (PhyML and RAxML) and coalescent (MP-EST and STAR) methods in the presence of long external branches and ILS. DNA sequences were simulated on five-taxon species trees S1–S3 under the multispecies coalescent model (Rannala and Yang 2003). The lengths of the internal branches are variously set to 0.002, 0.005, and 0.01 for species trees S1–S3, respectively (branch lengths are in mutation units). The lengths of the external branches leading to ingroup species *A*–*D* are set to 0.001, 0.101, 0.101, and 0.101, respectively, for species trees S1–S3. The length of the external branch leading to outgroup species *E* is set to 0.005, 0.011, and 0.021 for species trees S1–S3, respectively. In addition, to generate the same degree of ILS as in the species tree Q6 (fig. 1), the population size parameter θ is variously set to 0.02, 0.05, and 0.1 for species trees S1–S3, respectively. Results shown here represent the proportion of simulations in which concatenation and coalescent methods recover the true species tree.

higher proportion when ILS is low and the number of genes is less than 1,000 (fig. 2). When long external and short internal branches occur simultaneously with high ILS (i.e., the species tree R3), coalescent methods still recover the true species tree with a proportion of $\geq$0.97 as the number of genes increases to 500. However, the proportion of simulations in which concatenation methods recover the true species tree under these circumstances is low ($\leq$0.24), even when the number of genes increases to 2,000. Thus, our simulation analyses using five- and eight-taxon species trees indicate that in the presence of high ILS, the combination of long external and short internal branches can lead to the failure of concatenation methods. In contrast, coalescent methods are more robust under these circumstances.

To explore how varying lengths of the internal branches affects species tree estimation in the presence of long external branches and high ILS, we additionally simulated DNA sequences on species trees S1–S3 (fig. 3). These three species trees possess the same length for external branches leading to ingroup species *A*–*D*, but various lengths for internal branches. Similar to the species tree Q6, gene trees simulated on these three species trees are highly variable, and on average only 20% of them are congruent with the species tree topology. Moreover, because the most probable gene tree matches the species tree topology, as above, all three species trees are not in the anomaly zone. For species trees S1–S3, the proportion of simulations in which coalescent methods recover the true species tree increases to 1.0 as the number of genes increases to 500 (fig. 3). These results, together with our first simulation analyses, suggest that in the presence of long external branches and high ILS, coalescent methods are robust to various internal branch lengths (i.e., 0.001 in the species tree Q6 and 0.002, 0.005, and 0.01 in species trees S1–S3, respectively). In contrast, concatenation

methods are sensitive to the internal branch length under these circumstances. Here, when internal branch lengths increase from 0.001 (i.e., the species tree Q6) to 0.002 (i.e., the species tree S1), the proportion of simulations in which concatenation methods recover the true species tree is still low (i.e., $\leq$0.10 when the number of genes increases to 2,000; fig. 3). However, if we further increase the internal branch lengths to 0.005 (i.e., the species tree S2), the proportion of simulations in which concatenation methods recover the true species tree increases to 0.99 as the number of genes increases to 2,000 (fig. 3). These results indicate that by increasing internal branch lengths, thereby decreasing the ratio of the external/internal branch lengths, there is a vast improvement in the efficiency of concatenation methods. We attribute this behavior to a reduction in the severity of LBA effects.

To more thoroughly address how varying the number and placement of long external branches affects species tree estimation in the presence of short internal branches and ILS, we further simulated DNA sequences on five-taxon species trees T1–T16 (fig. 4). Because these 16 species trees have the same internal branch lengths (i.e., $x = 0.001$), the degree of ILS depends only on the value of θ. Similar to our first simulation analyses, when $\theta = 0.0001$, all simulated gene trees are congruent with the species tree topology. When $\theta = 0.01$, the topologies of simulated gene trees are highly variable: On average 20% of the simulated gene trees matched the species tree when the topology is symmetrical (i.e., species trees T1–T5) and 16% when the topology is pectinate (i.e., species trees T6–T16). Moreover, since the most probable gene tree matches the species tree topology, as above, all 16 species trees are not in the anomaly zone.

When there is only one long external branch (i.e., species trees T1, T6, T7, and T8; fig. 4), both concatenation and coalescent methods can estimate the true species tree with a proportion of 1.0 in most cases. The proportion of simulations in which concatenation methods recover the true species tree drops below 0.90 only when high ILS (i.e., $\theta = 0.01$) is combined with an extremely long branch (i.e., $y = 0.5$ for species trees T1 and T6; fig. 4). One exception is the species tree T8, in which the external branch leading to species *D* is long. Here, coalescent methods still recover the true species tree with a proportion of 1.0. However, the proportion of simulations in which concatenation methods recover the true species tree drops when $\theta = 0.01$ and $y \geq 0.05$, and declines to 0 when $y = 0.5$ (fig. 4). In these cases, concatenation methods consistently infer an incorrect species tree (i.e., topology *b* in fig. 5). Because concatenation methods can accurately estimate the true species tree T8 when $\theta = 0.0001$, these results indicate that in the presence of high ILS, the combination of long external and short internal branches may lead to failure of concatenation methods even when there is only a single long external branch.

When there are two or three long external branches in our species trees, as long as there are no long branches sister to each other (i.e., species trees T3, T10, T11, T12, and T15; fig. 4), both concatenation and coalescent methods can estimate the true species tree with a proportion of $\geq$0.99, even when these external branches are extremely long (i.e., $y = 0.5$).

When two of these long external branches are sister lineages (i.e., species trees T2, T4, T9, T13, and T14; fig. 4), as long as there is no ILS (i.e., $\theta = 0.0001$), both concatenation and coalescent methods can still estimate the true species tree with a proportion of $\geq$0.99. The proportion of simulations in which coalescent methods recover the true species tree drops below 0.20 only when these external branches are extremely long (i.e., $y = 0.5$ for species trees T2, T9, T13, and T14; fig. 4). In contrast, concatenation methods still recover the true species tree with a proportion of $\geq$0.67 under these circumstances. Previous simulations have demonstrated that when two long external branches are sister lineages in a four-taxon gene tree (i.e., the Farris zone [Siddall 1998] or the inverse-Felsenstein zone [Swofford et al. 2001]), the accuracy of individual gene tree estimation using ML is low when genes are short (i.e., the proportion of correctly estimated gene trees is approximately 0.40 when there are 1,000 sites), but greatly improves as gene length increases (Swofford et al. 2001). Thus, with low ILS, the accuracy of ML analyses is much lower for individual gene sequences than for concatenated gene sequences. Under these circumstances, inaccurate gene tree estimation will lead to incorrect species tree estimation using gene-tree-based coalescent methods. With high ILS (i.e., $\theta = 0.01$), however, the proportion of simulations in which concatenation methods recover the true species tree declines below 0.20 when $y \geq 0.05$ for species trees T2, T4, T9, and T14 (fig. 4). In these cases, coalescent methods still accurately estimate the true species tree, and only fail to do so when these external branches are extremely long (i.e., $y = 0.5$). These results indicate that in the presence of short internal branches and high ILS, the negative impact of long external branches is more pronounced for concatenation methods when two of these long branches are sister lineages. In addition, our analyses demonstrate that in the presence of high ILS, the deleterious impact of long and short branches is more severe for concatenated gene sequences than for individual gene sequences. For example, even though the most frequently observed gene tree (on average 13% of the inferred gene trees) matches the species tree T4 when $\theta = 0.01$ and $y = 0.2$ (fig. 4), the concatenation method fails to recover the true tree. Because coalescent methods estimate species trees from individual gene trees, they are more consistent under conditions where long external and short internal branches are in close proximity. However, when these external branches are extremely long, even coalescent methods can be misled, for example, $\theta = 0.01$ and $y = 0.5$ in the species tree T4 (fig. 4). Here, neither of the two most frequently observed gene trees (on average 10.3% and 10.2% of the inferred gene trees, respectively) matches the species tree topology due to LBA, which further leads to incorrect species tree estimation using coalescent methods. Thus, even gene-tree-based coalescent methods may become inconsistent under more extreme circumstances where the estimation of individual gene trees is significantly biased due to the combination of long and short branches.

Finally, if all four external branches leading to the ingroup species *A*–*D* are long and the species tree topology is pectinate (i.e., the species tree T16; fig. 4), both concatenation and

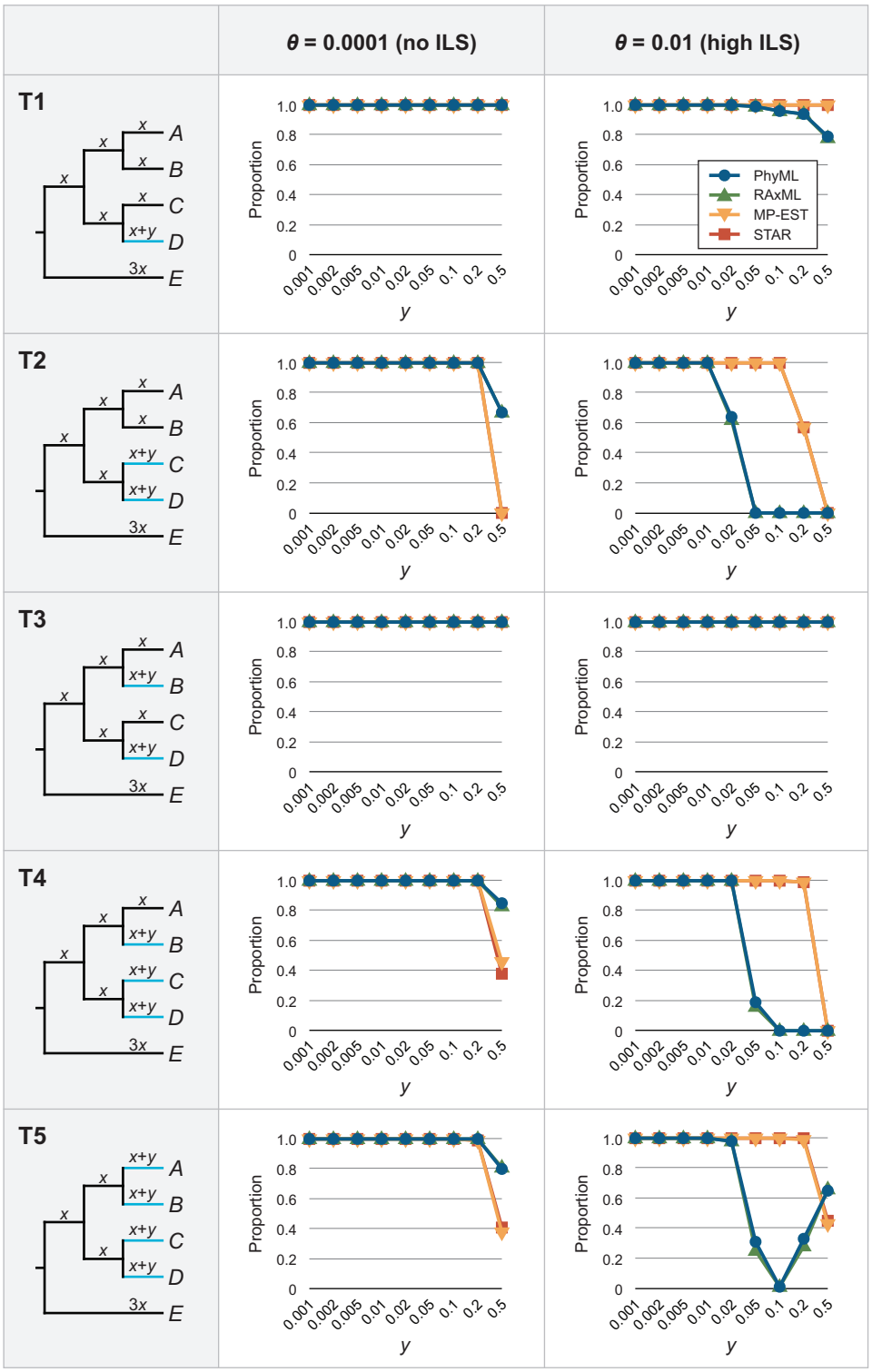

Fig. 4. DNA simulations to investigate how varying the number and placement of long external branches affects concatenation (PhyML and RAxML) and coalescent (MP-EST and STAR) methods in the presence of short internal branches and ILS. DNA sequences were simulated on five-taxon species trees T1–T16 under the multispecies coalescent model (Rannala and Yang 2003). In each of species trees T1–T16, the lengths of the internal branches are set to $x = 0.001$ (branch lengths are in mutation units). Two values of the population size parameter θ (i.e., 0.0001 and 0.01) are applied to simulate varying degrees of ILS. In addition, the external branches assigned as long are highlighted in blue, and the lengths of these long external branches vary as $y = (0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5)$. Results shown here represent the proportion of simulations in which concatenation and coalescent methods recover the true species tree.

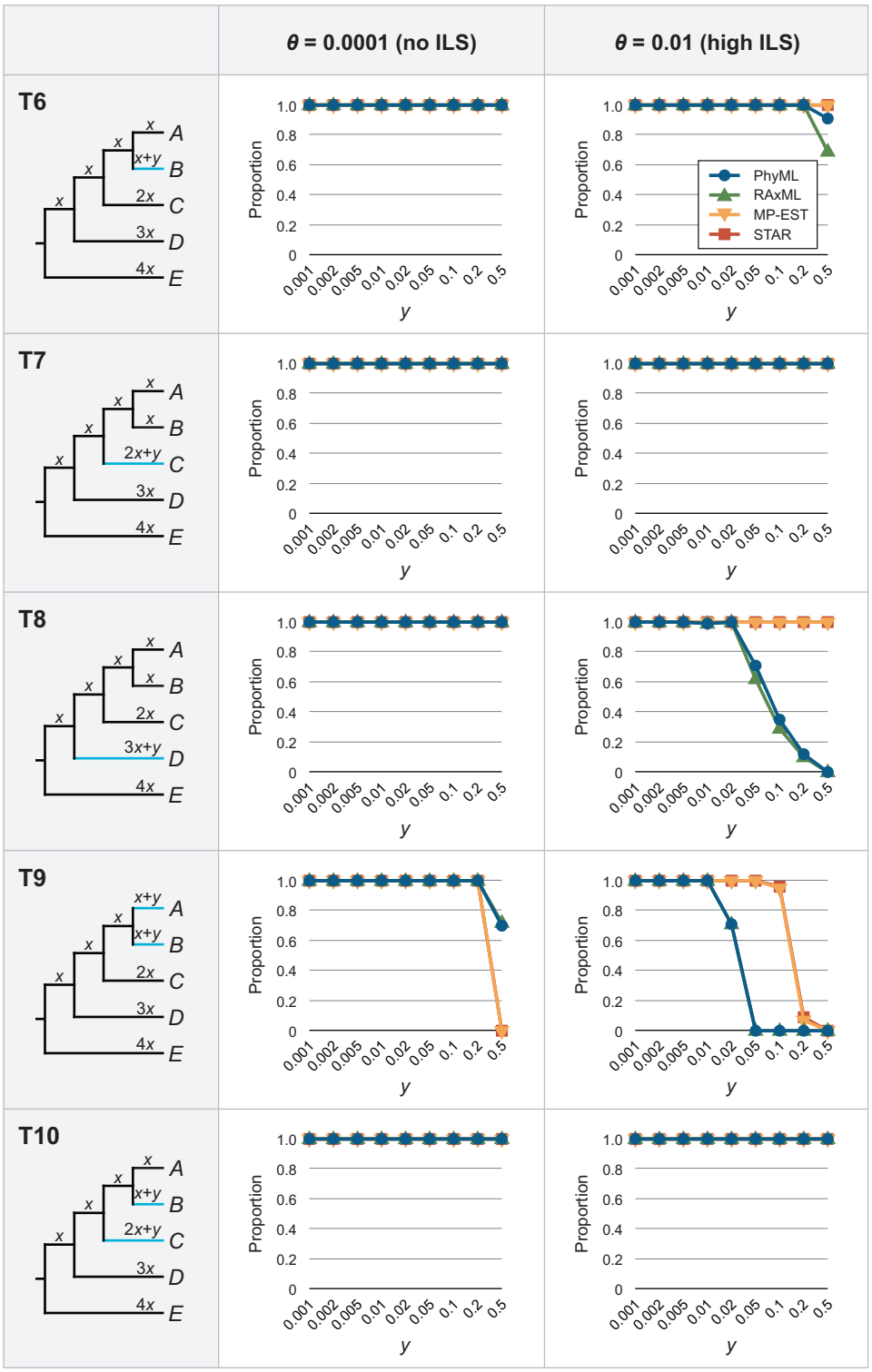

Fig. 4. Continued.

coalescent methods can estimate the true species tree with a proportion of 1.0. The lone exception is when these four branches are extremely long (i.e., $y = 0.5$). In this case, the proportion of simulations in which both methods recover the true species tree drops to as low as 0.60. If the species tree topology is symmetrical (i.e., the species tree T5; fig. 4), coalescent methods can similarly recover the true species tree with a proportion of ≥ 0.99. The proportion of simulations in which coalescent methods recover the true species tree drops to approximately 0.40 only if these four external branches are extremely long (i.e., $y = 0.5$). In contrast, when ILS is high (i.e., $\theta = 0.01$), the proportion of simulations in which concatenation methods recover the true species tree declines as the lengths of theses four external branches increase (i.e., 0.01

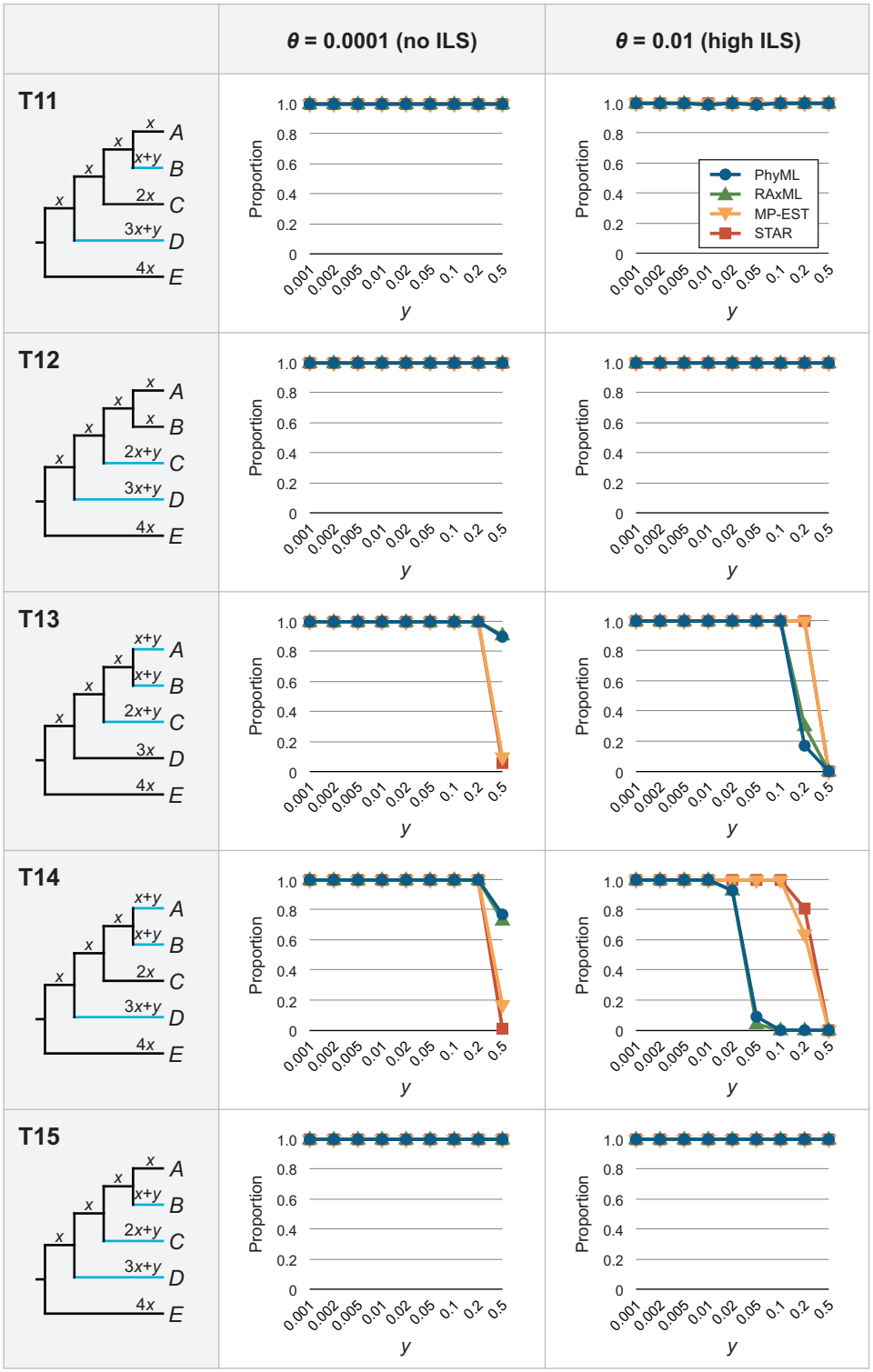

Fig. 4. Continued.

when $y = 0.1$; fig. 4). For reasons that we do not yet understand, when these four external branches are extremely long (i.e., $y = 0.5$), the proportion of simulations in which concatenation methods recovers the true species tree actually increases to 0.65 (fig. 4).

Scrotifera Mammalian Data Analyses

The two empirical data sets we analyzed (i.e., the seven- and five-taxon Scrotifera data sets) included 1,394 genes, and the average number of nucleotide sites for each gene was 1,078. Our concatenation and coalescent analyses of the

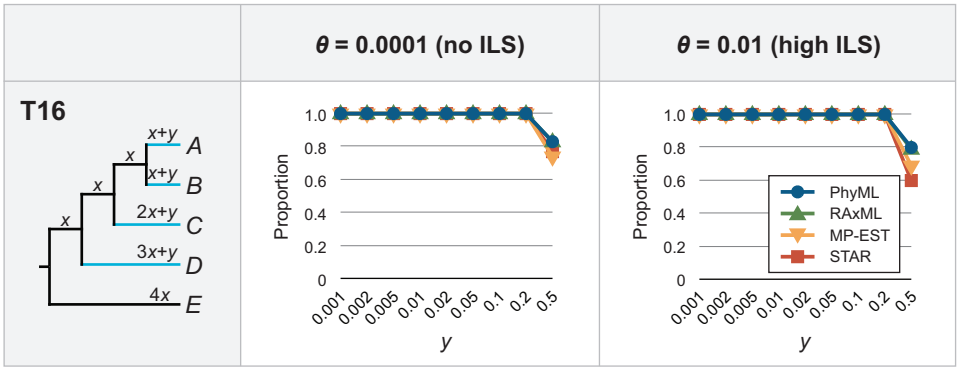

Fig. 4. Continued.

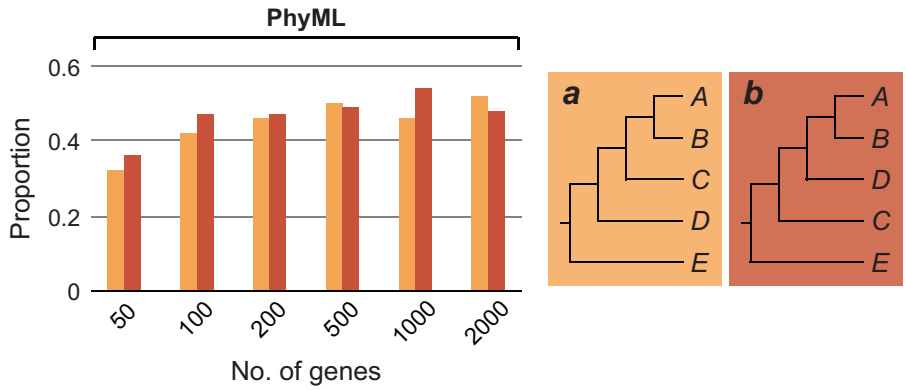

Fig. 5. Proportion of the two incorrect species trees inferred using the concatenation method (PhyML) from DNA sequences simulated on the species tree Q6 (fig. 1).

seven-taxon Scrotifera data set produced a well-supported (≥83 bootstrap percentage [BP]) species tree (fig. 6*a*), which is congruent with the species tree inferred by Tsagkogeorga et al. (2013). As expected, this noted ancient rapid radiation produces short internal branches separating the four orders in our species tree. Moreover, these data exhibit a high degree of discordance among individual gene trees (fig. 6*b*). Despite conflicting topologies among gene trees, a clade consisting of Cetartiodactyla plus Perissodactyla is supported by both concatenation and coalescent methods with 98 and 83 BP, respectively (fig. 6*a*). This sister relationship is also identified in the most frequently observed gene tree (30% of all 1,394 gene trees; fig. 6*b*).

When reducing our seven-taxon Scrotifera data set to five species (i.e., *Bos taurus*, *Canis familiaris*, *Eidolon helvum*, *Equus caballus*, and *Felis catus*) to artificially create longer external branches in this rapid radiation, we observed a similarly high degree of gene tree discordance (fig. 6*b*). Here, the clade of Cetartiodactyla plus Perissodactyla remains moderately supported in our coalescent analyses (71 and 61 BP from MP-EST and STAR, respectively; fig. 6*c*). In contrast, concatenation methods instead place Perissodactyla as sister to Carnivora with 63 BP (fig. 6*d*), despite the fact that the most frequently observed gene tree (33% of all 1,394 gene trees; fig. 6*b*) still supports Perissodactyla as sister to Cetartiodactyla. Thus, our analyses of this five-taxon Scrotifera data set corroborate our simulation results above, and suggest that in the presence of a high degree of gene tree discordance, the combination of long external and short internal branches can lead to the failure of concatenation methods. In contrast, coalescent methods are more robust under these circumstances.

Conclusions

LBA was first demonstrated by Felsenstein (1978) on a four-taxon gene tree. He demonstrated that two long external branches, separated by a short internal branch, misled parsimony or compatibility methods by consistently placing the two nonsister, long branches together. The fundamental cause of LBA in this circumstance is model misspecification, in which parsimony or compatibility methods cannot correctly handle the increasing number of identical nucleotides acquired by chance between the two nonsister lineages. Similarly, model misspecification can occur in species tree estimation. It has previously been shown that the model of species tree estimation using concatenation methods can be violated in the presence of extremely high ILS (i.e., the

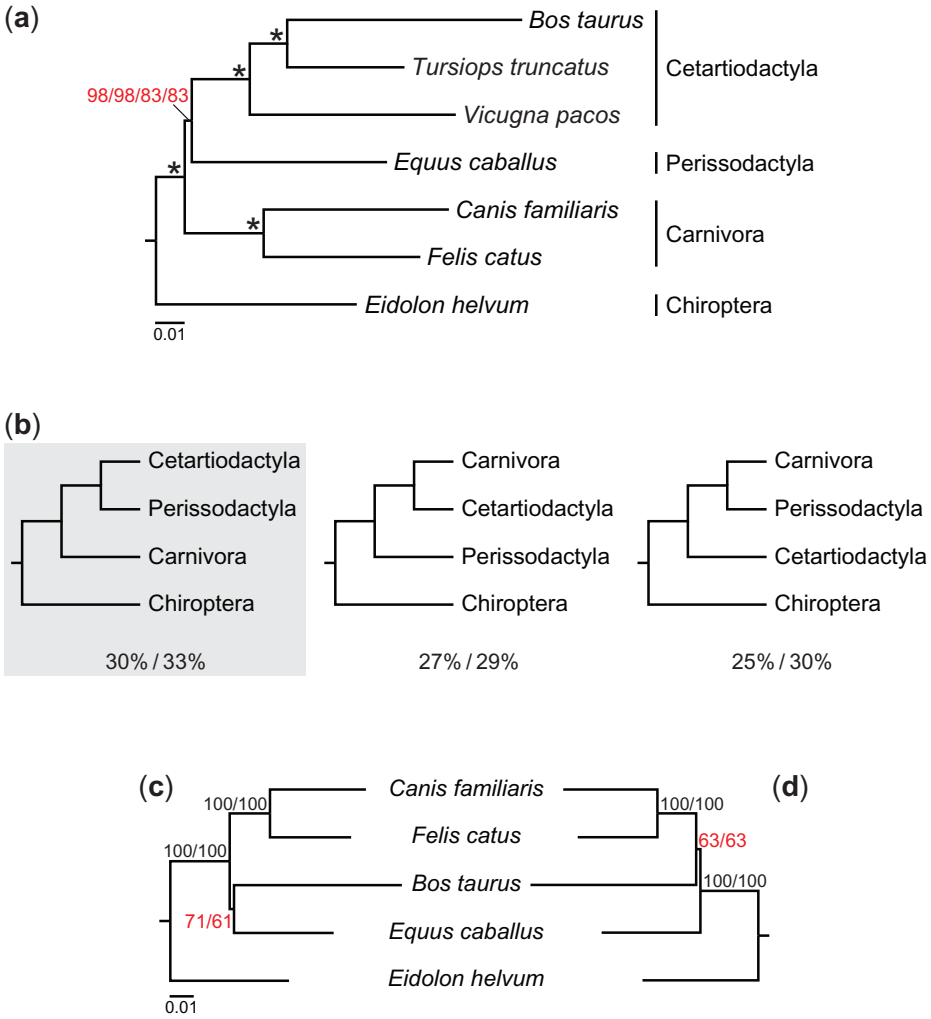

Fig. 6. Performance of concatenation (PhyML and RAxML) and coalescent (MP-EST and STAR) methods on the Scrotifera mammalian data sets, which represent an ancient rapid radiation (Zhou et al. 2012). (*a*) The species tree inferred from the seven-taxon Scrotifera data set using concatenation and coalescent methods. BPs from PhyML/RAxML/MP-EST/STAR are indicated for each branch, and an asterisk indicates that the branch is supported by 100 BPs from PhyML, RAxML, MP-EST, and STAR. Branch lengths shown here (in mutation units) were estimated from the concatenated matrix using RAxML. (*b*) Three possible relationships of Carnivora, Cetartiodactyla, Chiroptera, and Perissodactyla inferred from 1,394 genes. Percentages of each topology found in the seven-/five-taxon Scrotifera data set are indicated below. Here, the most frequently observed gene tree corresponds with the inferred species tree shown in figure 6*a*. (*c*) The species tree inferred from the five-taxon Scrotifera data set using coalescent methods (MP-EST and STAR). This five-taxon Scrotifera data set is designed to artificially create longer external branches in the species tree by removing two Cetartiodactyla species, *Tursiops truncatus* and *Vicugna pacos*, from the seven-taxon Scrotifera data set. BPs from MP-EST/STAR are indicated for each branch, and branch lengths shown here were estimated from the concatenated matrix using RAxML. (*d*) The species tree inferred from the five-taxon Scrotifera data set using concatenation methods (PhyML and RAxML). BPs from PhyML/RAxML are indicated for each branch, and branch lengths shown here were estimated from the concatenated matrix using RAxML.

anomaly zone). However, until now the combined effects of LBA and ILS have not been fully explored. Here, our simulation analyses show that the combination of long external and short internal branches alone does not obviously confound the consistency of concatenation and coalescent methods. With low ILS, concatenation methods are more likely to recover the true species tree when the number of genes is small or the external branches are extremely long. In contrast, when long external and short internal branches occur simultaneously with high ILS, concatenation methods can be misled especially when two of these long branches are sister lineages, while coalescent methods are more robust under these circumstances. Importantly, this topological pattern characterizes numerous ancient rapid radiations across the tree of life. Because short internal branches in a species tree can increase the potential for ILS and gene tree discordance, our results further suggest that coalescent methods are more likely to infer the correct species tree in cases of ancient rapid radiations where long external and short internal branches are in close phylogenetic proximity.

Previous empirical studies have demonstrated that coalescent methods better accommodate gene tree discordance (Song et al. 2012; Zhong et al. 2013) and elevated substitution rates (Xi et al. 2013, 2014). Our empirical analyses using Scrotifera mammals additionally show that when gene tree discordance is high, coalescent methods can better handle

species tree estimation in the presence of long and short branches. These results further demonstrate that reducing the effects of LBA remains critical, especially in the phylogenomics era (Heath et al. 2008; Hejnol et al. 2009; Pick et al. 2010). Thus, it remains essential to improve taxon sampling to reduce long branches when possible. However, in cases where long branches are inevitable, for example due to a high rate of extinction, it becomes essential to choose methods that are more robust to artifacts in species tree estimation. Furthermore, our results emphasize the growing consensus that we not only need additional data to resolve difficult phylogenetic problems, but also sophisticated methods that reduce systematic errors in large-scale phylogenomic analyses (Philippe, Brinkmann, Lavrov, et al. 2011; Philippe and Roure 2011).

Materials and Methods

Using Simulated Data to Examine the Impact of Long External and Short Internal Branches on Species Tree Estimation in the Presence of ILS

To investigate the impact of long external and short internal branches on species tree estimation in the presence of ILS, we first simulated DNA sequences on six 5-taxon species trees under the multispecies coalescent model (Rannala and Yang 2003). These hypothetical species trees Q1–Q6 (fig. 1), are topologically identical except with respect to branch lengths and the degree of ILS. In each of these six 5-taxon species trees, species *E* is designated as the outgroup, and one lineage was sampled from each of the species *A*–*E*. The lengths of the three short internal branches were held constant in all species trees (i.e., 0.001; branch lengths are in mutation units, that is, the number of nucleotide substitutions per site). The lengths of the external branches leading to species *A* and *E* were also held constant in all species trees (i.e., 0.001 and 0.003, respectively). In order to simulate various evolutionary rates along the same external branches across the species trees, we varied branch lengths rather than mutation rates because a single lineage was sampled from each species (fig. 1). For species trees Q1, Q3, and Q5, the three external branches leading to species *B*–*D* are short (i.e., 0.002), whereas for species trees Q2, Q4, and Q6, these three external branches are long (i.e., 0.101). In addition, we assumed that each gene lineage simulated from a branch in the species tree was subject to the same substitution rate specified for that branch. Thus, all gene trees simulated on species trees Q2, Q4, and Q6 possess longer external branches leading to species *B*–*D*, compared with gene trees simulated on species trees Q1, Q3, and Q5. Our choice of long and short branches were guided in part by our phylogenomic investigations of the angiosperm clade Malpighiales (Xi et al. 2012), which contains approximately 16,000 species and constitutes up to 40% of the understory tree diversity in tropical rain forests (Davis et al. 2005). Molecular dating using multiple fossil calibration points revealed that this corresponds to an explosive, ancient radiation during the mid-Cretaceous (Davis et al. 2005; Wurdack and Davis 2009; Xi et al. 2012).

In addition, we applied three different values of the population size parameter θ to simulate varying degrees of ILS (i.e., 0.0001 for species trees Q1 and Q2, 0.001 for species trees Q3 and Q4, and 0.01 for species trees Q5 and Q6; fig. 1). For each of species trees Q1–Q6, we assumed that population size was constant across all populations. 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. To determine if our values of θ are comparable with empirical studies, we first converted our branch lengths to coalescent units. In order to accomplish this, the branch lengths in mutation units must be divided by θ. Here, we determine that the lengths of the internal branches in species trees Q1–Q6 (i.e., 0.1, 1, and 10 coalescent units) are within the range of two well-studied examples: The branches in *Passerina* buntings (i.e., as short as 0.05 coalescent units) (Carling and Brumfield 2008; Degnan and Rosenberg 2009) and the two internal branches in the human–chimpanzee–gorilla–orangutan species tree (i.e., ~1.2 and ~4.2 coalescent units) (Rannala and Yang 2003; Degnan and Rosenberg 2006, 2009). Because species trees Q1–Q6 have the same length for all internal branches, the degree of ILS depends only on the value of θ. According to coalescent theory, the degree of ILS is positively correlated with the value of θ. Thus, a large value of θ produces gene trees with highly discordant topologies despite a common species tree. In this regard, these five-taxon species trees allowed us to target specific cases where 1) only long external and short internal branches are present (i.e., species tree Q2), 2) only high ILS is present (i.e., species tree Q5), and 3) long external and short internal branches occur simultaneously with high ILS (i.e., species tree Q6).

Next, we simulated 50, 100, 200, 500, 1,000, and 2,000 gene trees on each of species trees Q1–Q6 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 DNA sequences of 1,000 bp using Seq-Gen v1.3.3 (Rambaut and Grassly 1997) with the JC69 model (Jukes and Cantor 1969). For concatenation analyses, the ML trees were estimated from concatenated gene sequences using a single GTR+Γ model. Analyses were performed with a random starting tree using PhyML v20120412 ("-a e -b 0 -c 4 -f m -m GTR -o tlr -s SPR -v 0 –rand_start –n_rand_starts 2") (Guindon et al. 2010) and RAxML v8.1.3 ("-d -f o -m GTRGAMMAX –no-bfgs –no-seq-check") (Stamatakis 2014). For coalescent analyses, individual gene trees were first inferred using PhyML with the GTR+Γ model and rooted with species *E*. These estimated gene trees were then used to construct the species trees with MP-EST v1.4 and the STAR method as implemented in Phybase. Each simulation was repeated 100 times.

To explore if the choice of species number in our first simulation affects the performance of concatenation and coalescent methods, we also simulated DNA sequences on three 8-taxon species trees (fig. 2) under the multispecies coalescent model (Rannala and Yang 2003). In each of the species trees R1–R3, species *H* is designated as the outgroup, and one lineage was sampled from each of the species *A*–*H*. The lengths of all internal branches were held constant in species

trees R1–R3 (i.e., 0.001; fig. 2). The lengths of the external branches leading to species *A*, *B*, *C*, and *H* were also held constant in species trees R1–R3 (i.e., 0.001, 0.001, 0.002, and 0.005, respectively). For the species tree R1, the four external branches leading to species *D*–*G* are short (i.e., 0.002, 0.002, 0.003, and 0.005, respectively), whereas for species trees R2 and R3, these four external branches are long (i.e., 0.201, 0.201, 0.202, and 0.204, respectively). In addition, we applied two values of θ to simulate varying degrees of ILS (i.e., 0.001 for species tree R2, and 0.01 for species trees R1 and R3; fig. 2). Similarly, these eight-taxon species trees allowed us to target specific cases where 1) only high ILS is present (i.e., species tree R1), 2) long external and short internal branches occur simultaneously with low ILS (i.e., species tree R2), and 3) long external and short internal branches occur simultaneously with high ILS (i.e., species tree R3). DNA sequences were simulated on each of the species trees R1–R3 as described above, and the species trees were inferred using PhyML, RAxML, MP-EST, and STAR as described above. Each simulation was repeated 100 times.

To investigate how varying the length of the short internal branches affects species tree estimation in the presence of long external branches and high ILS, we additionally simulated DNA sequences on species trees S1–S3 (fig. 3) under the multispecies coalescent model (Rannala and Yang 2003). In each of these three 5-taxon species trees, the lengths of the external branches leading to ingroup species *A*–*D* were held constant as in the species tree Q6 described above (i.e., 0.001 for the branch leading to species *A* and 0.101 for branches leading to species *B*–*D*). We varied the lengths of the three internal branches (i.e., 0.002, 0.005, and 0.01 for species trees S1–S3, respectively) and the external branch leading to outgroup species *E* (i.e., 0.005, 0.011, and 0.021 for species trees S1–S3, respectively). Moreover, to generate the same degree of ILS as in the species tree Q6, we variously set θ to 0.02, 0.05, and 0.1 for species trees S1–S3, respectively (fig. 3). DNA sequences were simulated on each of the species trees S1–S3 as described above, and the species trees were inferred using PhyML, RAxML, MP-EST, and STAR as described above. Each simulation was repeated 100 times.

To more thoroughly address if varying the number and placement of long external branches affects species tree estimation in the presence of short internal branches and high ILS, we further simulated DNA sequences on 16 species trees (fig. 4) with two different topologies (i.e., symmetrical species trees T1–T5, and pectinate species trees T6–T16). In each of these five-taxon species trees, the length of the internal branches was held constant (i.e., $x = 0.001$), and two values of θ (i.e., 0.0001 and 0.01) were applied to simulate varying degrees of ILS. Furthermore, in each of species trees T1–T16, various numbers (i.e., one to four) of the external branches leading to the ingroup species *A*–*D* were assigned as long (fig. 4). Here, the length of these long external branches varies as $y = (0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5)$. For each value of *y*, we simulated 2,000 gene trees using Phybase as described above, and these gene trees were then utilized to simulate DNA sequences of 1,000 bp using Seq-Gen with the JC69 model. The species trees were inferred using PhyML, RAxML, MP-EST, and STAR as described above. Each simulation was repeated 100 times.

Using Empirical Data to Examine the Impact of Long External and Short Internal Branches on Species Tree Estimation

In addition to our simulated data above, we explored the performance of concatenation and coalescent methods using empirical phylogenomic data in circumstances where long external and short internal branches are in close phylogenetic proximity. To investigate this we reanalyzed the data set from Tsagkogeorga et al. (2013), which included 2,320 coding DNA sequence alignments (subsequently referred as genes) from 22 mammals. We first created a submatrix by pruning the original data set to include seven Scrotifera mammal species from four orders (*B. taurus* [order Cetartiodactyla], *C. familiaris* [Carnivora], *E. helvum* [Chiroptera], *Eq. caballus* [Perissodactyla], *F. catus* [Carnivora], *Tursiops truncatus* [Cetartiodactyla], and *Vicugna pacos* [Cetartiodactyla]). We included only those genes containing DNA sequences from all seven species to alleviate concerns of missing data. These four orders were targeted because they exhibit a rapid radiation in the Late Cretaceous (Zhou et al. 2012), that is, short internal branches separating these orders in the inferred species tree (Tsagkogeorga et al. 2013). These compressed internal branches are where ILS is likely to be high. To demonstrate the degree of gene tree discordance, we calculated the distribution of estimated gene trees for the Scrotifera data sets, and compared them to the inferred species trees. Based on the phylogram estimated by Tsagkogeorga et al. (2013), we further exacerbated long external branches in the species tree by removing two of the three Cetartiodactyla species, *T. truncatus* and *V. pacos*, from our initial seven-taxon Scrotifera data set. This reduced five-taxon Scrotifera data set allowed us to better examine how the combination of long external and short internal branches will affect species tree estimation in this ancient rapid radiation.

Species trees were inferred from each of the seven- and five-taxon Scrotifera data sets using both concatenation and coalescent methods. For concatenation analyses, the species tree was estimated using PhyML and RAxML with the GTR+Γ model as described above. For coalescent analyses, gene trees were first inferred using PhyML with the GTR+Γ model and rooted with *E. helvum*. These estimated gene trees were then utilized to construct the species tree using MP-EST and STAR as described above. Bootstrap support was estimated using a multilocus bootstrapping approach (Seo 2008) with 100 replicates.

Acknowledgments

The authors thank Scott Edwards, Joshua Rest, and members of the Davis and Liu laboratories for helpful comments and discussion. They also thank Paul Edmon and Mike Ethier for technical support. Finally, they thank the two anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This work was supported

by the United States National Science Foundation (DMS-1222745 to L.L. and DEB-1120243 to C.C.D.).

References

Anderson FE, Swofford DL. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18 S rDNA. *Mol Phylogenet Evol.* 33:440–451.

Bayzid MS, Warnow T. 2013. Naive binning improves phylogenomic analyses. *Bioinformatics* 29:2277–2284.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21: 163–193.

Brinkmann H, Van der Giezen M, Zhou Y, De Raucourt GP, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol.* 54:743–757.

Carling MD, Brumfield RT. 2008. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in Passerina buntings. *Genetics* 178:363–377.

Davis CC, Webb CO, Wurdack KJ, Jaramillo CA, Donoghue MJ. 2005. Explosive radiation of Malpighiales supports a mid-Cretaceous origin of modern tropical rain forests. *Am Nat.* 165:E36–E65.

de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol Evol.* 22:34–41.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol.* 24: 332–340.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19.

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

Finet C, Timme RE, Delwiche CF, Marlétaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr Biol.* 20:2217–2222.

Guindon S, Dufayard JF, 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.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol.* 46:239–257.

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc Lond B Biol Sci.* 276:4261–4270.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol.* 27:570–580.

Hess J, Goldman N. 2011. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. *PLoS One* 6: e22783.

Hillis DM, Huelsenbeck JP, Cunningham CW. 1994. Application and accuracy of molecular phylogenies. *Science* 264:671–677.

Huelsenbeck JP. 1995. Performance of phylogenetic methods in simulation. *Syst Biol.* 44:17–48.

Huelsenbeck JP, Hillis DM. 1993. Success of phylogenetic methods in the four-taxon case. *Syst Biol.* 42:247–264.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

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

Kolaczkowski B, Thornton JW. 2009. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS One* 4:e7891.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol.* 56:17–24.

Kück P, Mayer C, Wägele J-W, Misof B. 2012. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS One* 7:e36593.

Kutschera VE, Bidon T, Hailer F, Rodi JL, Fain SR, Janke A. 2014. Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Mol Biol Evol.* 31:2004–2017.

Lanfear R, Welch JJ, Bromham L. 2010. Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol Evol.* 25:495–503.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol.* 7(Suppl 1):S4.

Leaché AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol.* 60:126–137.

Lee EK, Cibrian-Jaramillo A, Kolokotronis SO, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, McCombie WR, et al. 2011. A functional phylogenomic view of the seed plants. *PLoS Genet.* 7: e1002411.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24:2542–2543.

Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. *Syst Biol.* 58:452–460.

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

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol.* 53:320–328.

Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst Biol.* 58:468–477.

Lyons-Weiler J, Hoelzer GA. 1997. Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. *Mol Phylogenet Evol.* 8: 375–384.

Mirarab S, Bayzid MS, Warnow T. 2014. Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol.* doi: 10.1093/sysbio/syu063.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol.* 5:568–583.

Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* 470:255–258.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Worheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9:e1000602.

Philippe H, Roure B. 2011. Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biol.* 9:91.

Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alié A, Morgenstern B, Manuel M, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol.* 27:1983–1987.

Pol D, Siddall ME. 2001. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. *Cladistics* 17:266–281.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* 2:e173.

Rambaut A, Grassly NC. 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.

Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463: 1079–1083.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol Biol Evol.* 19:101–109.

Sanderson MJ, Wojciechowski MF, Hu J-M, Khan TS, Brady SG. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Mol Biol Evol.* 17:782–797.

Seo TK. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol.* 25:960–971.

Siddall ME. 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone. *Cladistics* 14: 209–220.

Siddall ME, Whiting MF. 1999. Long-branch abstractions. *Cladistics* 15: 9–24.

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322:86–89.

Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* 480:364–367.

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A.* 109:14942–14947.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stefanović S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol Biol.* 4:35.

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol.* 50: 525–539.

Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One* 7:e29696.

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. *Curr Biol.* 23:2262–2267.

Wall JD, Kim SK, Luca F, Carbone L, Mootnick AR, de Jong PJ, Di Rienzo A. 2013. Incomplete lineage sorting is common in extant gibbon genera. *PLoS One* 8:e53682.

Whitfield JB, Kjer KM. 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu Rev Entomol.* 53:449–472.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. *Trends Ecol Evol.* 22:258–265.

Wiens JJ. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Syst Bio.l* 54:731–742.

William J, Ballard O. 1996. Combining data in phylogenetic analysis. *Trends Ecol Evol.* 11:334.

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

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462: 1056–1060.

Wu Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66:763–775.

Wurdack KJ, Davis CC. 2009. Malpighiales phylogenetics: gaining ground on one of the most recalcitrant clades in the angiosperm tree of life. *Am J Bot.* 96:1551–1570.

Xi Z, Liu L, Rest JS, Davis CC. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Syst Biol.* 63:919–932.

Xi Z, Rest JS, Davis CC. 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PLoS One* 8:e80870.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc Natl Acad Sci U S A.* 109:17519–17524.

Zhong B, Liu L, Yan Z, Penny D. 2013. Origin of land plants using the multispecies coalescent model. *Trends Plant Sci.* 18:492–495.

Zhou X, Xu S, Xu J, Chen B, Zhou K, Yang G. 2012. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the Laurasiatherian mammals. *Syst Biol.* 61:150–164.