The Impact of Missing Data on Species Tree Estimation

Zhenxiang Xi,[1] Liang Liu,[2,3] and Charles C. Davis*[,1]
[1]Department of Organismic and Evolutionary Biology, Harvard University
[2]Department of Statistics, University of Georgia
[3]Institute of Bioinformatics, University of Georgia
***Corresponding author:** E-mail: cdavis@oeb.harvard.edu.
Associate editor: Claudia Russo

Abstract

Phylogeneticists are increasingly assembling genome-scale data sets that include hundreds of genes to resolve their focal clades. Although these data sets commonly include a moderate to high amount of missing data, there remains no consensus on their impact to species tree estimation. Here, using several simulated and empirical data sets, we assess the effects of missing data on species tree estimation under varying degrees of incomplete lineage sorting (ILS) and gene rate heterogeneity. We demonstrate that concatenation (RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (matrix representation with parsimony [MRP]) methods perform reliably, so long as missing data are randomly distributed (by gene and/or by species) and that a sufficiently large number of genes are sampled. When data sets are indecisive sensu Sanderson et al. (2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol Biol. 10:155) and/or ILS is high, however, high amounts of missing data that are randomly distributed require exhaustive levels of gene sampling, likely exceeding most empirical studies to date. Moreover, missing data become especially problematic when they are nonrandomly distributed. We demonstrate that STAR produces inconsistent results when the amount of nonrandom missing data is high, regardless of the degree of ILS and gene rate heterogeneity. Similarly, concatenation methods using maximum likelihood can be misled by nonrandom missing data in the presence of gene rate heterogeneity, which becomes further exacerbated when combined with high ILS. In contrast, ASTRAL, MP-EST, and MRP are more robust under all of these scenarios. These results underscore the importance of understanding the influence of missing data in the phylogenomics era.

Key words: **coalescent methods, concatenation methods, gene rate heterogeneity, incomplete lineage sorting, missing data, species tree estimation**

Article

Introduction

Over the past decade, the effects of missing data on phylogenetic analyses have been extensively explored using both simulated and empirical data sets. Numerous studies have indicated that phylogenetic reconstruction is not sensitive to missing data, as long as the overall number of characters is large (e.g., Philippe et al. 2004; Fulton and Strobeck 2006; Wiens and Moen 2008; de la Torre-Bárcena et al. 2009; Thomson and Shaffer 2010; Wiens and Morrill 2011; Jiang et al. 2014). Similarly, it has been concluded that adding taxa, even with vast amounts of missing data should generally increase the accuracy of phylogenetic inference (e.g., Wiens 2003, 2005; Cho et al. 2011; Wiens and Tiu 2012). This conclusion is further exemplified by numerous efforts to build so-called mega-phylogenies, which include hundreds or even thousands of species using very sparse data matrices (e.g., Driskell et al. 2004; McMahon and Sanderson 2006; Edwards and Smith 2010; Thomson and Shaffer 2010; Pyron and Wiens 2011; Smith et al. 2011; Zanne et al. 2014). These data sets are typically assembled by mining sequences from resources such as GenBank (Benson et al. 2015) to build a single, enormous concatenated matrix. Such matrices have the advantage of including many species in a single analysis, but suffer from a high amount of missing data (i.e., more than 90% in many cases). As a counterpoint to these findings, other studies have demonstrated that phylogenetic reconstruction may be compromised by missing data, especially when they are nonrandomly distributed (e.g., Agnarsson and May-Collado 2008; Hartmann and Vision 2008; Lemmon et al. 2009; Kupczok et al. 2010; Simmons 2012a, 2012b, 2014; Kvist and Siddall 2013; Xia 2014). Additionally, recent studies by Sanderson et al. (2010) and Steel and Sanderson (2010) have demonstrated that data sets with incomplete taxon coverage—whereby sequences from some partitions are missing for some taxa—can be phylogenetically indecisive. Indecisive data sets can result in a vast terrace of phylogenetic trees that have different topologies but the same optimality score (Sanderson et al. 2011).

This dueling viewpoint on the effects of missing data obviously necessitates further investigation, and is especially timely owing to two main developments in the field of phylogenetics. The first are the technological advances in next-generation sequencing (Mardis 2013), which have facilitated the utilization of genome-scale data to resolve major branches in the tree of life. This is perhaps best exemplified by the recent effort to understand the evolutionary history of modern birds using whole-genome data from 48 species (Jarvis et al. 2014). The second is the shifting emphasis in

phylogenetic studies from gene tree to species tree estimation (Edwards 2009), which leads to the consideration of how best to reconstruct the species tree from a cloud of gene tree histories (Maddison 1997). Along these lines, traditional concatenation methods (William and Ballard 1996; de Queiroz and Gatesy 2007) have been commonly employed for this purpose, which implicitly assume that all genes have the same or very similar evolutionary histories. The utility of concatenation methods is further advocated by several recent simulation and empirical studies (e.g., Gatesy and Springer 2014; Springer and Gatesy 2014; Tonini et al. 2015). In contrast, coalescent-based methods permit gene trees to have different evolutionary histories (Liu, Yu, Kubatko, et al. 2009). 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 presently scale up for genome-level analyses (Leaché and Rannala 2011; Bayzid and Warnow 2013; Mirarab, Bayzid, and Warnow 2014). Thus, they are not the focus of our study. 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, Reaz, Bayzid, et al. 2014; which constructs species tree from quartets [Bryant and Steel 2001]), 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 the latter consensus methods are not strictly coalescent-based, they can accommodate gene tree discordance due to incomplete lineage sorting (ILS), and have been shown to be statistically consistent under the multispecies coalescent model (Liu, Yu, Pearl, et al. 2009; Liu and Yu 2011; Mirarab, Reaz, Bayzid, et al. 2014). For simplicity, we also refer to these consensus methods as gene-tree-based coalescent methods. Moreover, a recent review on the topic of coalescent methods indicates that they are likely to be more computationally manageable than concatenation methods, especially for data sets with many taxa and thousands of genes (Liu, Xi, Wu, et al. 2015).

Until now, the evaluation and comparison of coalescent methods have been conducted primarily using simulated data under conditions where missing data are absent (e.g., Leaché and Rannala 2011; Mirarab, Bayzid, and Warnow 2014; Mirarab, Reaz, Bayzid, et al. 2014; Liu, Xi, and Davis 2015). However, this is seldom the case for large-scale phylogenomic data, and the amount of missing data varies greatly across data sets (in this article, we specifically refer missing data to missing sequences in genes or loci), for example, 6% missing data for the 2,320-gene mammal data set by Tsagkogeorga et al. (2013), 28% for the 310-gene vascular plant data set by Xi et al. (2014), 34% for the 852-gene land pant data set by Wickett et al. (2014), 55% for the 150-gene animal data set by Dunn et al. (2008), and 81% for the 1,487-gene animal data set by Hejnol et al. (2009). The presence of missing data in phylogenomic data is attributed to a variety of factors that may be either methodological or biological. For example, 1) insufficient sequencing coverage in next-generation sequencing experiments; 2) degraded DNA or RNA, which is especially prominent for taxa acquired from historical materials (e.g., older museum specimens and fossilized bones); 3) bias in the pattern of gene loss due to variation in functional constraints across clades; and 4) species with greatly elevated substitution rates, which may result in highly variable gene sequences that can be especially problematic for target enrichment methods when universal primers or probes are used across a broad swath of the tree of life. In some cases, these issues may result in randomly distributed missing data (e.g., low sequencing coverage); in others, however, missing data may be nonrandomly distributed across species and/or genes (e.g., parasites are often associated with significant genome reduction by gene loss [Wolfe et al. 1992; Katinka et al. 2001; Sakharkar et al. 2004; de Koning and Keeling 2006; McNeal et al. 2007; Molina et al. 2014]).

The effects of missing data on phylogenetic inference are beginning to be addressed more explicitly using large-scale simulated or empirical data (e.g., Hartmann and Vision 2008; Hovmöller et al. 2013; Kvist and Siddall 2013; Roure et al. 2013; Jiang et al. 2014). A recent study by Hovmöller et al. (2013) in particular examined the effects of missing data on coalescent analyses (*BEAST and STEM). In their simulations, missing data were randomly distributed among species or concentrated in certain species. The latter case might result from species sampled using degraded DNA as we outline above. The authors concluded that the amount of missing data (up to 50%) had a negligible effect on coalescent analyses of data sets with 25–100 genes. These results are reassuring and follow a long history of well-cited studies arguing that missing data are generally not problematic for phylogenetic inference. Here, we take this opportunity to further explore the effects of missing data on species tree estimation using both simulated and previously published empirical data. Specifically, our study focuses on a comparison of concatenation, gene-tree-based coalescent, and supertree methods, and seeks to explore 1) how missing data affect species tree estimation under varying degrees of ILS and gene rate heterogeneity, and 2) the circumstances under which missing data may mislead species tree estimation.

Results and Discussion

Our analyses demonstrate that missing data can indeed influence species tree estimation, but this influence depends on a variety of factors. As a roadmap to our results and discussion below, we summarize results from 17- and 5-taxon simulation analyses in figure 1. In general, all methods we investigated here—concatenation (RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (matrix representation with parsimony [MRP]) methods—perform reliably, as long as missing data are randomly distributed (by gene and/or by species) and that a sufficiently large number of genes are sampled. When data sets are indecisive

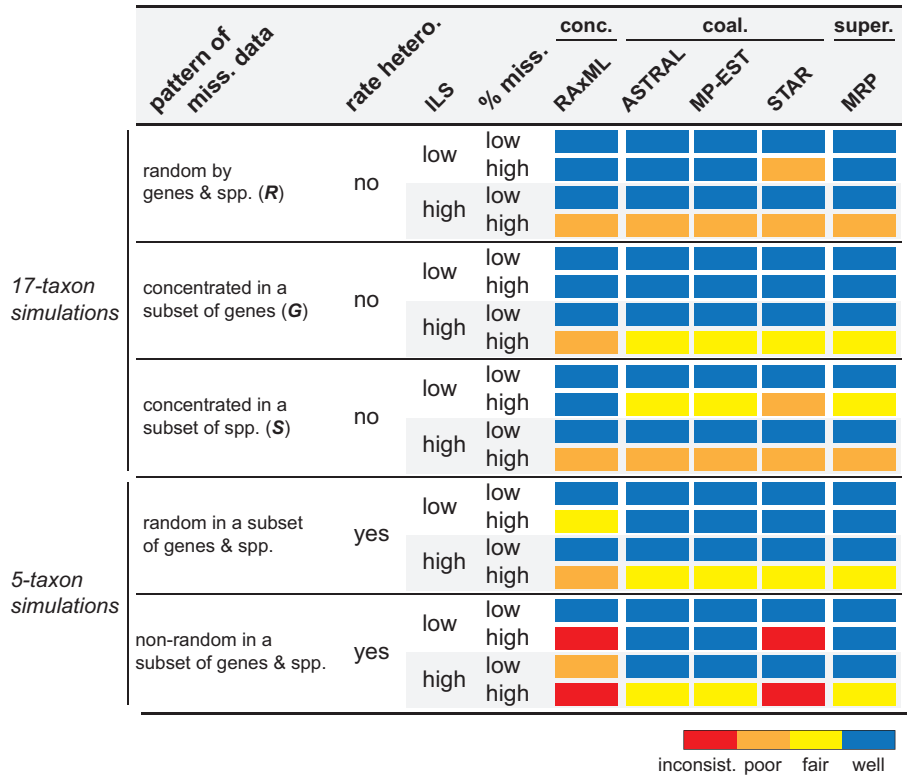


Fig. 1. Summary of the impact of missing data on species tree estimation under varying degrees of ILS and gene rate heterogeneity. The performance of concatenation (RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods was evaluated using data sets simulated on 17- and 5-taxon species trees. Various amounts of missing data were generated using five different patterns. Colored cells indicate the performance of species tree estimation methods under a particular situation. Here, blue, yellow, orange, and red represent that the method performs well, fair, poor, and inconsistently, respectively. For 17-taxon simulations, we present the mean RF distance, in which case, a smaller value equals increased phylogenetic accuracy. Here, the performance of a particular method is assigned as fair and poor if the mean RF distance between the true species tree and those estimated from 200-gene data sets is larger than 0.2 and 0.4, respectively. In contrast, for 5-taxon simulations, we present the proportion of simulations in which a particular method recovers the true species tree. Here, a bigger value equals increased phylogenetic accuracy, and the performance is assigned as fair and poor if the proportion is less than 0.8 and 0.6, respectively (50- and 200-gene data sets for low and high ILS, respectively).

sensu Sanderson et al. (2010) and/or ILS is high, however, high amounts of missing data that are randomly distributed require exhaustive levels of gene sampling, likely exceeding most empirical studies to date. Moreover, missing data become especially problematic when they are nonrandomly distributed. STAR, in particular, produces inconsistent results when the amount of nonrandom missing data is high, regardless of the degree of ILS and gene rate heterogeneity. Similarly, concatenation methods using maximum likelihood (ML) can be misled by nonrandom missing data in the presence of gene rate heterogeneity, which becomes further exacerbated when combined with high ILS. We discuss these results in detail below, and juxtapose our findings with corroborating evidence from empirical data sets.

The Impact of Missing Data on Species Tree Estimation in the Presence of ILS

For each of the data sets simulated on species trees T1–T6 (fig. 2), we examined the impact of missing data on species tree estimation using three patterns (fig. 3), that is, missing data that were randomly distributed across genes and species (***R***), missing data that were concentrated in a subset of randomly chosen genes (***G***), or concentrated in a subset of randomly chosen species (***S***). Importantly, the overall percentages of missing data were comparable across these three patterns (i.e., 35%, 53%, and 70%). For each of these data sets, we used the metric of phylogenetic decisiveness sensu Sanderson et al. (2010) to characterize the pattern of incomplete taxon coverage induced by missing data. Here, we referred to a data set as decisive if the true species tree was uniquely defined by subtrees determined by each separate gene. These subtrees were induced from the true species tree (i.e., species trees T1–T6) by pruning away any taxa that had missing data for each gene. For the pattern ***G***, there were always some genes including sequences from all 17 species. Thus, these data sets were decisive as defined by Sanderson et al. (2010), that is, the pattern of incomplete taxon coverage uniquely defined the true species tree. This applied even when the amount of missing data was high (i.e., ***G70%***). For the pattern ***R***, incomplete taxon coverage led to indecisive data sets in regard to the true species tree only when the amount of missing data was high (i.e., ***R70%***) and the number of genes was $\leq$200 (fig. 4*a*). Thus, when missing data were randomly distributed across genes and species, increasing the number of genes eventually led to decisive data sets, which

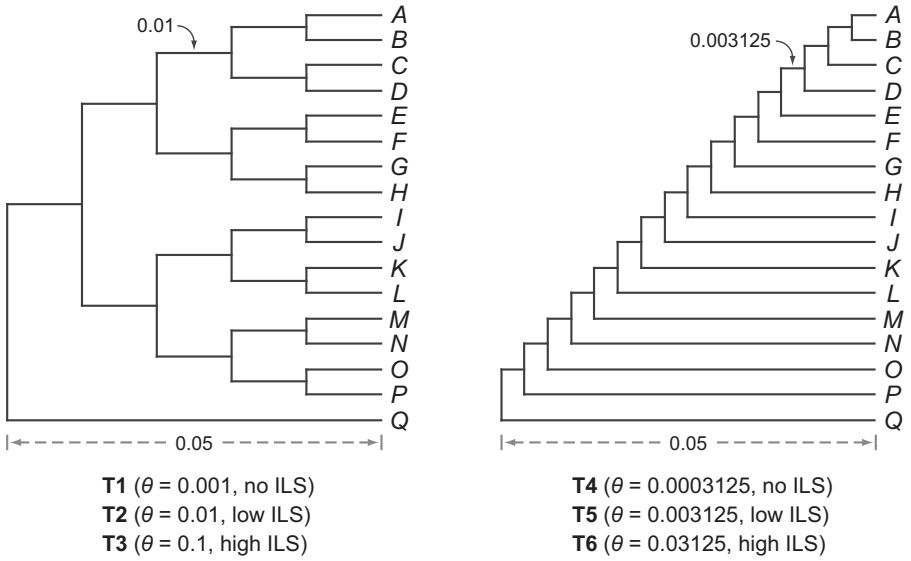


Fig. 2. DNA simulations using 17-taxon species trees to investigate the impact of missing data in the presence of ILS. DNA sequences were simulated on ultrametric species trees T1–T6 under the multispecies coalescent model (Rannala and Yang 2003). The heights of these species trees are 0.05 (branch lengths are in mutation units). To achieve similar tree heights, the internal branch lengths of the symmetrical (T1–T3) and pectinate (T4–T6) species trees were set to be 0.01 and 0.003125, respectively. 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.

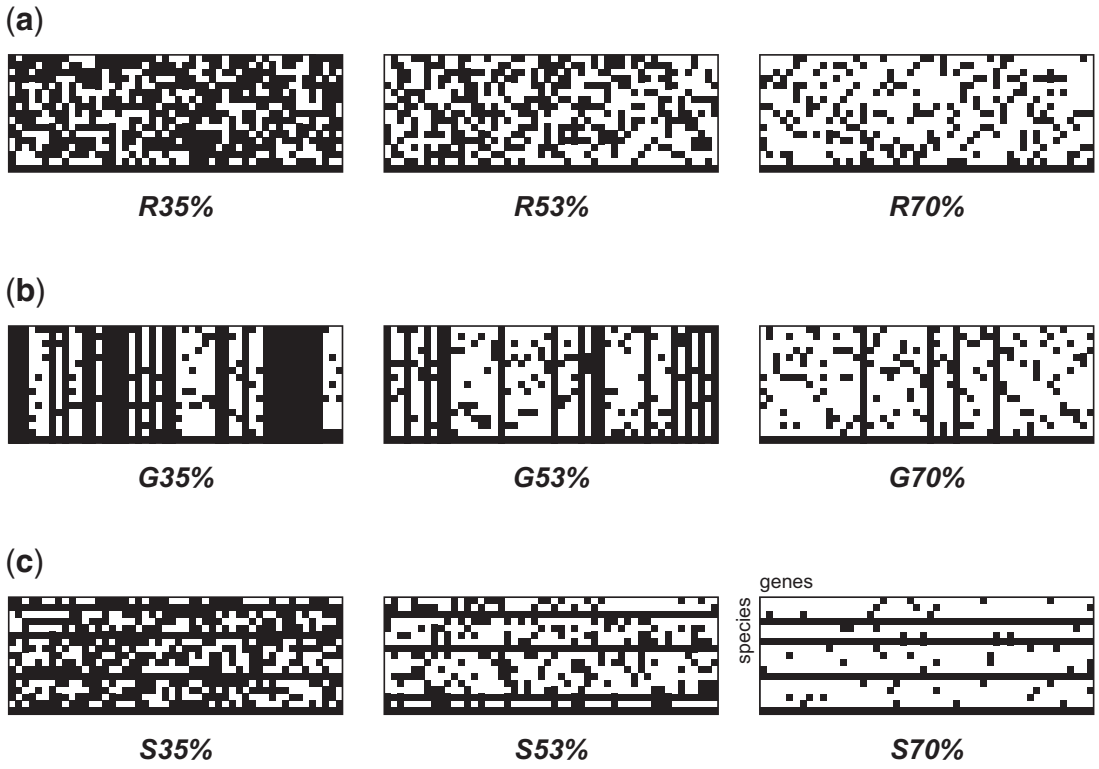


Fig. 3. Examples of three patterns (***R, G,*** and ***S***) used to simulate missing data. For each data set, 35%, 53%, or 70% of the total gene sequences were removed. Present and absent gene sequences are shown in black and white, respectively. (*a*) For the pattern ***R***, missing data are randomly distributed across ingroup species for all genes. (*b*) For the pattern ***G***, missing data are randomly distributed across ingroup species but concentrated in a subset of randomly chosen genes. (*c*) For the pattern ***S***, missing data are randomly distributed across all genes but concentrated in a subset of randomly chosen species.

was consistent with Sanderson et al. (2010). In contrast, the incomplete taxon coverage was especially problematic for the pattern ***S*** with regard to data decisiveness. When the amount of missing data was high (i.e., ***S70%***), 23–34% of our simulated data sets were indecisive in regard to the true species tree even when the number of genes increased to 2,000 (fig. 4*b*). Under these circumstances, the pattern of incomplete taxon coverage likely results in a vast terrace of phylogenetic trees

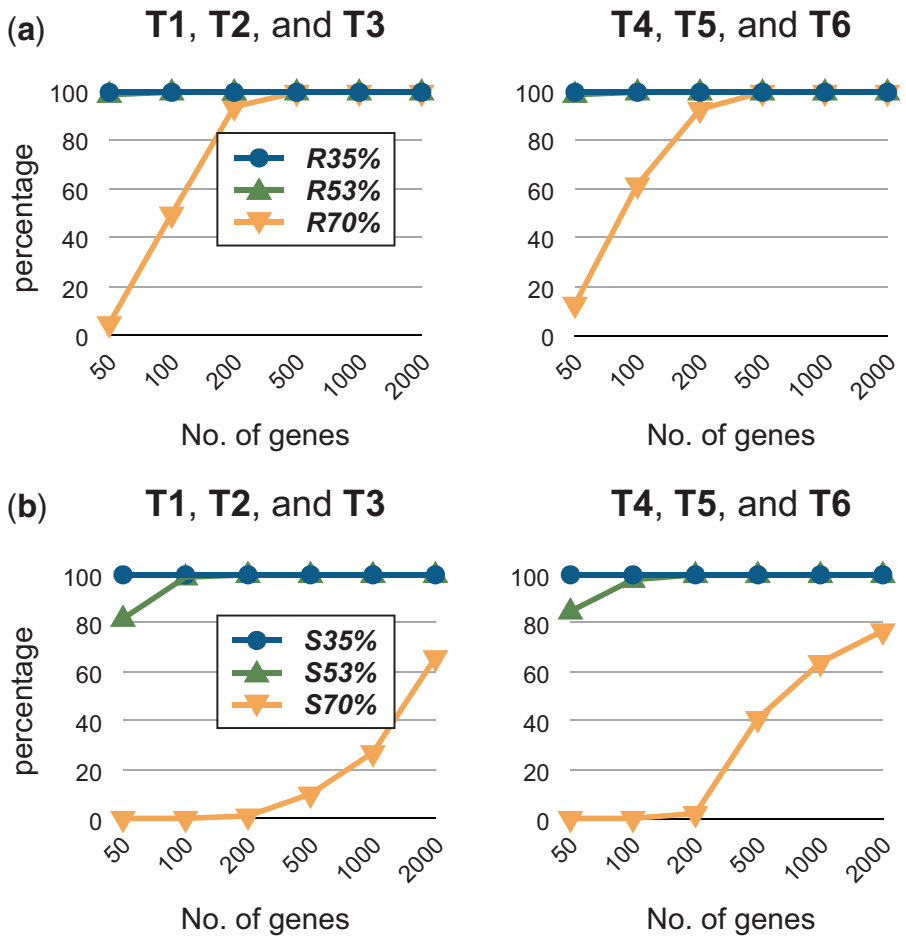


FIG. 4. Percentages of simulated 17-taxon data sets that were phylogenetically decisive sensu Sanderson et al. (2010). DNA sequences were simulated on symmetrical (T1–T3) and pectinate (T4–T6) species trees (fig. 2), and missing data were then generated on each of the 50-, 100-, 200-, 500-, 1,000-, and 2,000-gene data sets. All data sets with missing data concentrated in a subset of randomly chosen genes (***G***) were decisive and thus not presented here. (*a*) Missing data are randomly distributed across ingroup species and all genes (***R***). (*b*) Missing data are randomly distributed across all genes but concentrated in a subset of randomly chosen species (***S***).

that have different topologies but the same optimality score (Sanderson et al. 2011). These results indicate that high amounts of missing data are more likely to create indecisive data sets, especially when missing data are concentrated in a subset of randomly chosen species.

Simulation analyses of our 17-taxon species trees T1–T6 demonstrated that when θ was low (i.e., 0.001 and 0.0003125 for species trees T1 and T4, respectively), all simulated gene trees (when rooted with species Q) were congruent with the species tree topology. When θ increased (i.e., 0.01 and 0.003125 for species trees T2 and T5, respectively), on average 26% of the simulated gene trees were congruent with the species tree topology. When θ was high (i.e., 0.1 and 0.03125 for species trees T3 and T6), the topologies of simulated gene trees were highly variable, and none of these gene trees matched the species tree topology.

When the degree of ILS was low (i.e., species trees T1, T2, T4, and T5), missing data had a minimal effect on the accuracy of species tree estimation, as long as the data sets were decisive. Under these circumstances, the mean Robinson–Foulds (RF) distances between the true species tree and those estimated by the concatenation method (RAxML; very similar results were observed in partitioned RAxML analyses [supplementary fig. S1, Supplementary Material online]), two gene-tree-based coalescent methods (ASTRAL and MP-EST; but not STAR, see below), and the supertree method (MRP) were less than 0.013 as the number of genes increased to 100 (fig. 5). However, the accuracy of species tree estimation appeared to be adversely affected by missing data when data sets were indecisive, which was especially notable for gene-tree-based coalescent and supertree methods. For example, for the pattern ***G70%***, the mean RF distance between the species tree T5 and those estimated by ASTRAL was small (0.024) as the number of genes increased to 50 (fig. 5). In contrast, when data sets possessed the same amount of missing data (i.e., 70% for the 50-gene data sets), the mean RF distance between the species tree T5 and those estimated by ASTRAL increased to 0.262 and 0.549 for patterns ***R70%*** and ***S70%***, respectively (fig. 5). Under these circumstances, the concatenation method showed higher accuracy with a mean RF distance of 0.042 and 0.224 for patterns ***R70%*** and ***S70%***, respectively (fig. 5). These results indicate that under a low degree of ILS, the concatenation method is more robust to missing data, even when the data

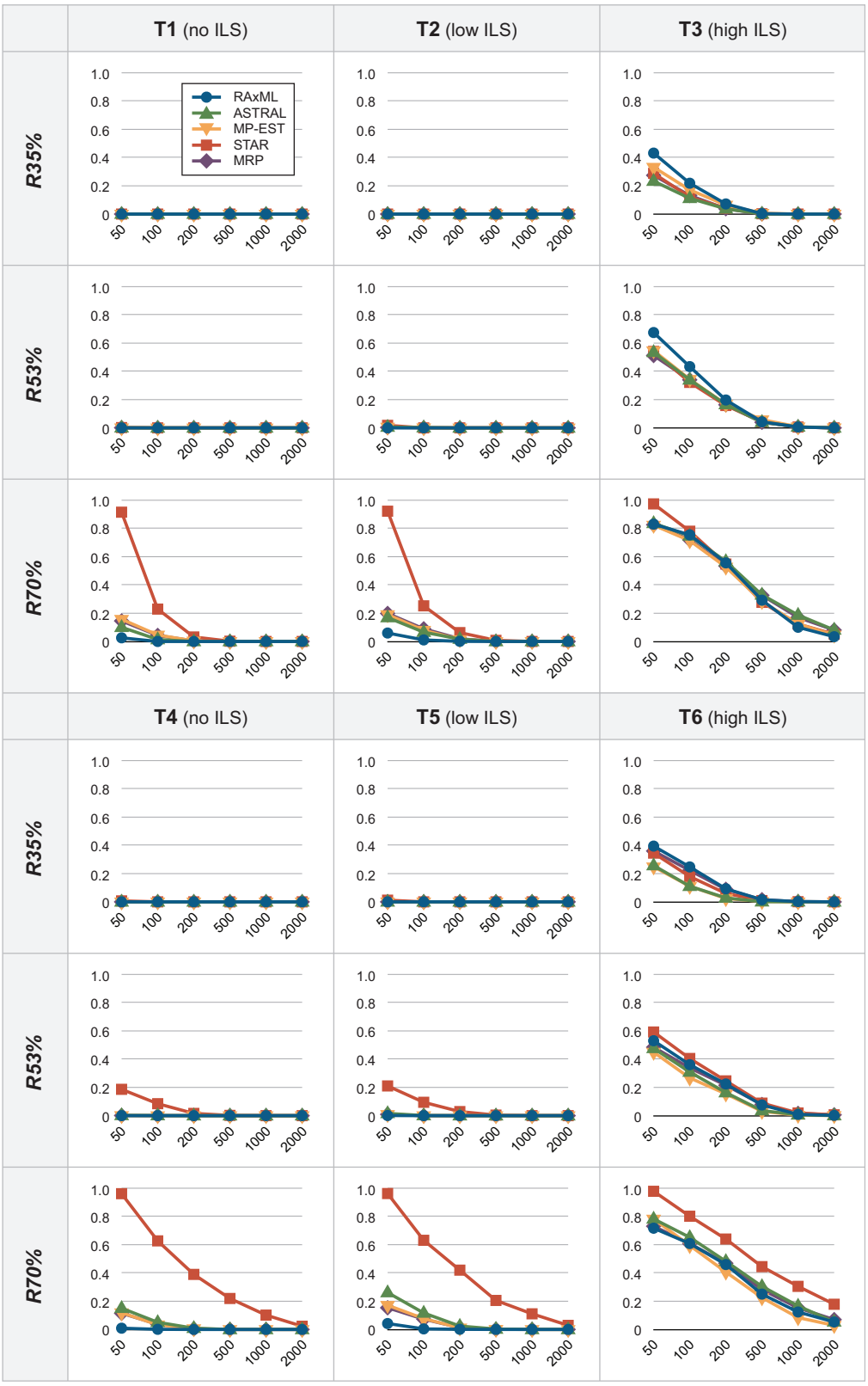


FIG. 5. The mean RF distances between the true species trees T1–T6 and those estimated from data sets with various amounts of missing data. DNA sequences were simulated on species trees T1–T6 (fig. 2), and missing data were then generated on each of the data sets using one of the three patterns, ***R***, ***G***, and ***S*** as described in the main text and in figure 3. Species trees were estimated from 50-, 100-, 200-, 500-, 1,000-, and 2,000-gene data sets using concatenation (unpartitioned RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods.

(continued)

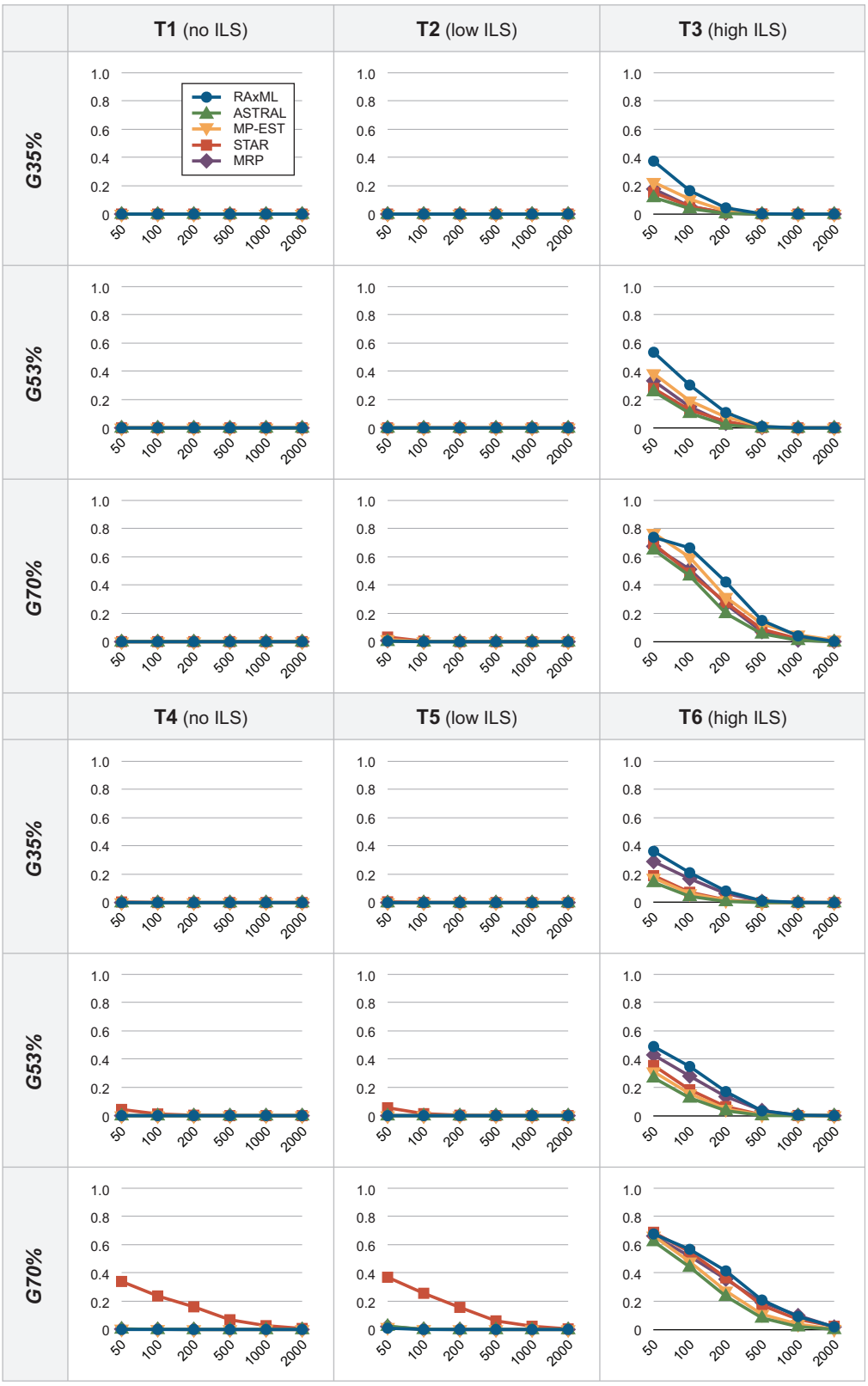


FIG. 5. (Continued)

set is indecisive. However, the performance of gene-tree-based coalescent and supertree methods can be greatly improved by sampling more genes. For example, as the number of genes increased to 2,000, the mean RF distance between the species tree T5 and those estimated by ASTRAL decreased to 0.021 for the pattern ***S70%*** (fig. 5). These results corroborate findings from previous studies in which adding incompletely sampled genes generally increases the accuracy of phylogenomic analyses, especially when the amount of missing data is high (Hovmöller et al. 2013; Jiang et al. 2014).

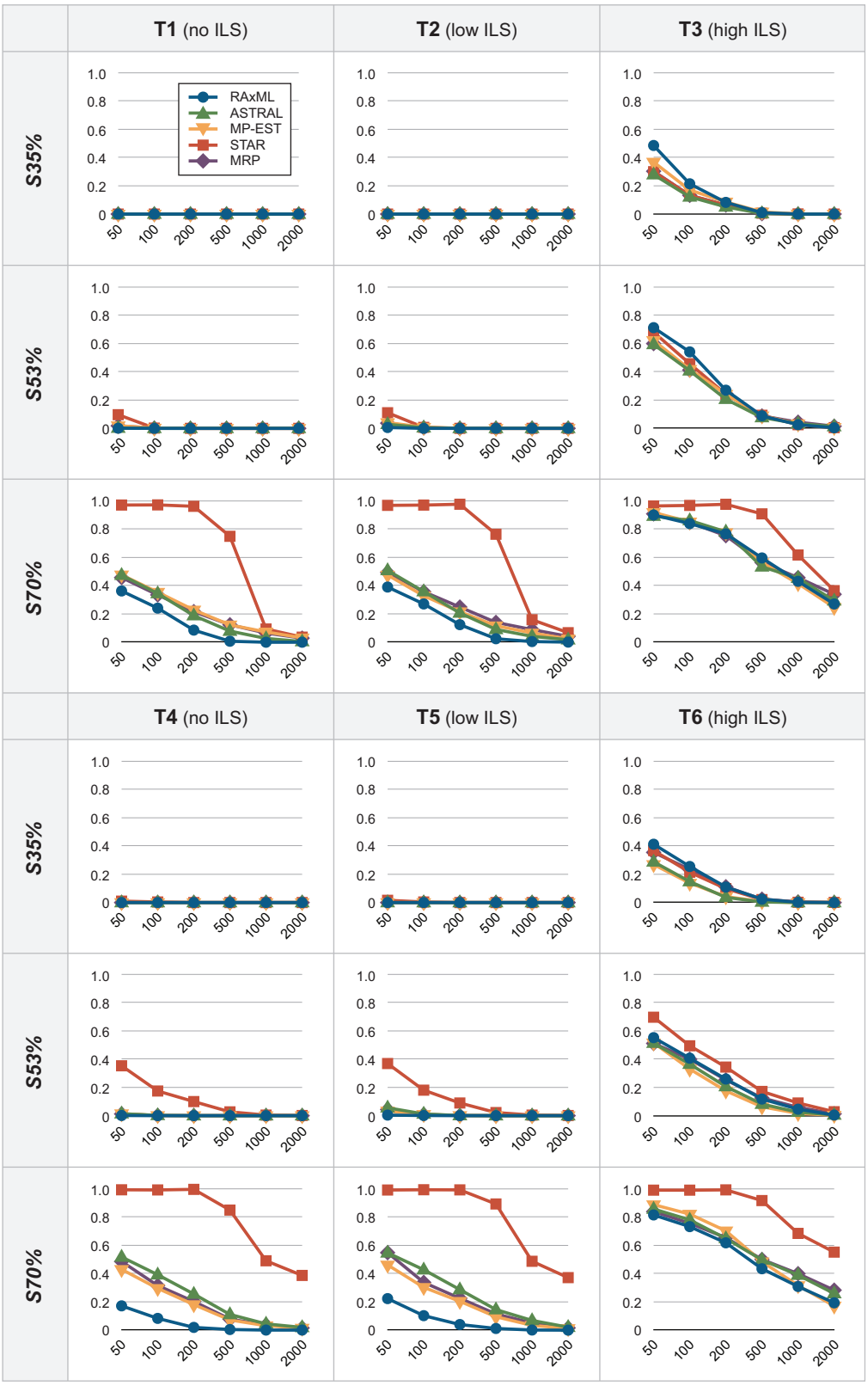


FIG. 5. (Continued)

In addition, our simulations show that compared with other gene-tree-based coalescent methods (ASTRAL and MP-EST), missing data had a more adverse effect on STAR when data sets were indecisive. For example, for the pattern ***S70%***, the mean RF distances between the species tree T1 and those estimated by ASTRAL and MP-EST were 0.189 and 0.229, respectively, as the number of genes increased to 200 (fig. 5). In contrast, the mean RF distance between the species

tree T1 and those estimated by STAR was 0.965 (fig. 5). As the accuracy of gene tree estimation was very high for the pattern ***S70%***, that is, on average more than 99% of the estimated gene trees matched the gene trees simulated on the species tree T1, the negative effects of missing data on STAR observed here could not be attributed to gene tree estimation error. Furthermore, even when data sets were decisive, the performance of STAR was greatly compromised by missing data if the true species tree was pectinate (i.e., species trees T4 and T5). For example, for the pattern ***G70%***, the mean RF distance between the pectinate species tree T4 and those estimated by STAR was 0.341 as the number of genes increased to 50 (fig. 5). In contrast, STAR consistently recovered the symmetrical species tree T1 even for the pattern ***G70%*** (i.e., the mean RF distance = 0; fig. 5).

What may explain this discrepancy between STAR and other gene-tree-based coalescent methods? By default, STAR first produces a distance matrix by counting the ranks between all pairs of taxa for each of the rooted gene trees, and then constructs the species tree from this distance matrix using neighbor-joining (NJ) (Saitou and Nei 1987). As the ranks depend on the number of taxa in individual gene trees, gene trees possessing a large number of missing taxa likely bias the estimation of ranks (Zhong et al. 2014). In addition, the accuracy of NJ declines dramatically when the amount of missing data is high (Wiens 2003; Hartmann and Vision 2008), which appears to be exacerbated if data sets are indecisive. In contrast, ASTRAL finds the species tree that maximizes the total number of quartet trees induced by gene trees, and MP-EST utilizes rooted triples in gene trees to estimate the species tree. These two methods appear to be less sensitive to missing data likely because the distributions of triples and quartet trees are invariant to missing taxa in the gene trees. Furthermore, for species trees with the same number of taxa, the pectinate topology (e.g., species tree T4) possesses many more nested ranks than the symmetrical one (e.g., species tree T1). Thus, the performance of STAR appears to be greatly compromised by the high amount of missing data, which consistently estimates incorrect species trees especially when data sets are indecisive. Under these circumstances, the performance of STAR can still be improved by sampling more genes, but the number of genes required may be larger than most empirical studies to date. For example, for the pattern ***S70%***, the mean RF distance between the species tree T4 and those estimated by STAR was 0.388 even as the number of genes increased to 2,000 (fig. 5).

When the degree of ILS was high (i.e., species trees T3 and T6), the accuracy of species tree estimation declined, sometimes dramatically, as the amount of missing data increased. For example, using complete data sets, the mean RF distance between the species tree T3 and those estimated by the concatenation method was 0.060 as the number of genes increased to 100 (fig. 6). In contrast, the mean RF distance increased to 0.166, 0.304, and 0.666 for patterns ***G35***%, ***G53***%, and ***G70***%, respectively (fig. 5). Under these circumstances, ASTRAL showed higher accuracy with a mean RF distance of 0.039, 0.104, and 0.468 for patterns ***G35***%, ***G53***%, and ***G70***%, respectively (fig. 5). These results indicate that in the presence of high ILS, gene-tree-based coalescent methods (ASTRAL and MP-EST) perform better than the concatenation method when the number of genes is relatively small. In addition, the adverse effects of missing data were more pronounced for patterns ***R*** and ***S*** when the amount of missing data was high (i.e., $\geq$53%). For example, for the pattern ***G53%***, the mean RF distance between the species tree T3 and those estimated by ASTRAL was 0.021 as the number of genes increased to 200 (fig. 5). In contrast, the mean RF distance increased to 0.166 and 0.207 for patterns ***R53%*** and ***S53%***, respectively (fig. 5), despite the same amount of missing data (i.e., 53% for the 200-gene data sets). Moreover, although the performance of species tree estimation was greatly compromised by missing data when ILS was high, the mean RF distances between estimated species trees and

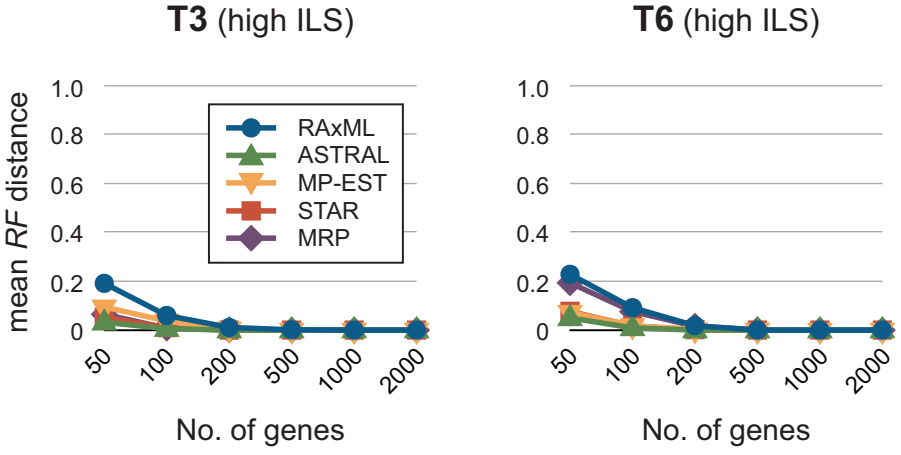


Fig. 6. The mean RF distances between the true species trees and those estimated from complete data sets. DNA sequences were simulated on species trees T3 and T6, respectively (fig. 2). Species trees were estimated from 50-, 100-, 200-, 500-, 1,000-, and 2,000-gene data sets using concatenation (unpartitioned RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods.

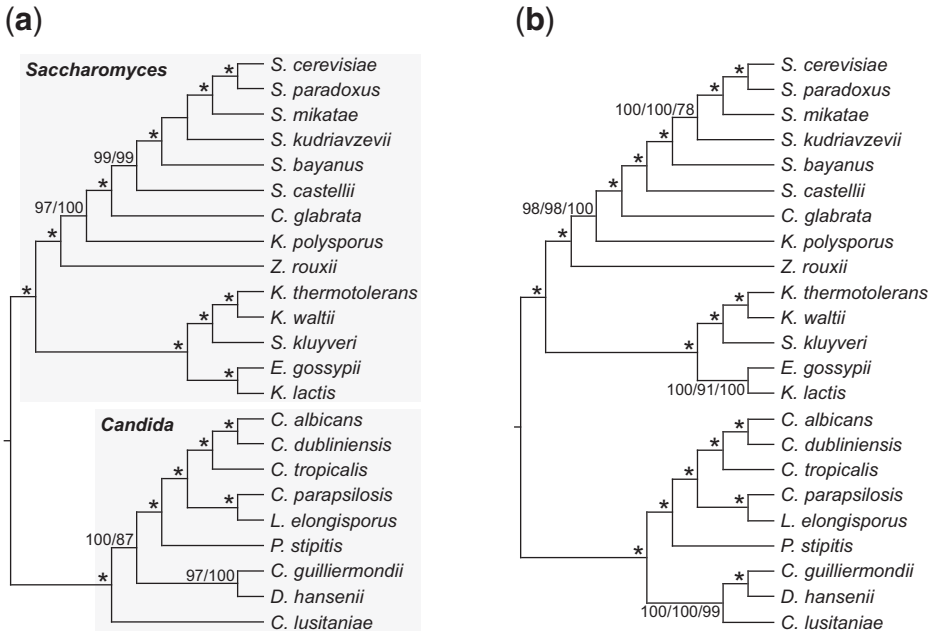


Fig. 7. Species trees of 23 yeasts inferred from the complete 502-gene data set. (*a*) The species tree inferred using concatenation (RAxML) and supertree (MRP) methods. BPs from RAxML/MRP are indicated for each branch, and an asterisk indicates that the branch is supported by 100 BPs from both RAxML and MRP. (*b*) The species tree inferred using gene-tree-based coalescent methods (ASTRAL, MP-EST, and STAR). BPs from ASTRAL/MP-EST/STAR are indicated for each branch, and an asterisk indicates that the branch is supported by 100 BPs from ASTRAL, MP-EST, and STAR. The genera abbreviations are as follows: *C.*, *Candida*; *D.*, *Debaryomyces*; *E.*, *Eremothecium*; *K.*, *Kluyveromyces*; *L.*, *Lodderomyces*; *P.*, *Pichia*; *S.*, *Saccharomyces*; and *Z.*, *Zygosaccharomyces*.

the true species tree decreased as the number of genes increased. For example, as the number of genes increased to 1,000, the mean RF distance between the species tree T3 and those estimated by ASTRAL decreased to 0.007 and 0.033 for patterns ***R53%*** and ***S53%***, respectively (fig. 5). These results mirror our results above for indecisive data sets involving high amounts of missing data. Here, we show that although high amounts of missing data are problematic for all these methods when ILS is high, the accuracy of species tree estimation can be improved by sampling more genes. Under extreme circumstances, however, the number of genes required may exceed most empirical studies to date.

We additionally explored the effects of missing data in the presence of ILS using the yeast data set assembled by Salichos and Rokas (2013). After removing poorly aligned amino acid sequences and ambiguously aligned sites, the yeast data set included 502 genes from 23 species, and the average number of amino acid sites for each gene was 473. The species tree inferred from the complete 502-gene data set using the concatenation method was strongly supported (i.e., ≥97 bootstrap percentage [BP]; fig. 7*a*), and congruent with the species tree inferred by Salichos and Rokas (2013). Gene-tree-based coalescent and supertree methods similarly produced well-resolved species trees. The only topological differences between these methods were in the placement of *Candida lusitaniae*. Here, MRP supported the same relationships as those identified using the concatenation method, that is, *C. lusitaniae* as sister to all other species in the *Candida* clade with 87 BP (fig. 7*a*). In contrast, ASTRAL, MP-EST, and STAR placed *C. lusitaniae* as sister to *C. guilliermondii* plus *Debaryomyces hansenii* with 100, 100, and 99 BP, respectively (fig. 7*b*). Thus, for downstream assessments we considered the species tree shown in figure 7*a* as the accepted topology for the concatenation and supertree analyses; whereas the species tree shown in figure 7*b* was the accepted topology for all gene-tree-based coalescent analyses. In addition, phylogenetic analyses of the 502 genes produced 502 distinct gene trees, and none of which matched the two species trees. This mirrors previous findings by Salichos and Rokas (2013), and suggests that there is likely a high degree of ILS in the yeast data set.

When generating missing data on this 502-gene data set using three patterns (***R***, ***G***, and ***S***; fig. 3), phylogenomic analyses of these incomplete data sets generally corroborated results using simulated data sets described above. As expected, in the presence of ILS, the accuracy of species tree estimation declined as the amount of missing data increased, and was especially pronounced for patterns ***R*** and ***S*** when the amount of missing data was high (i.e., 70%). Here, the concatenation method, ASTRAL, and MP-EST were more robust to missing data: The mean RF distances between species trees inferred from the complete data set and those from data sets with 35% or 53% missing data were less than 0.040, and increased up to 0.204 only for the pattern ***S70%*** (fig. 8). In addition, the adverse effects of missing data were more pronounced for STAR when data sets were indecisive. For example, the mean RF distance between the species tree inferred from the complete data set by STAR and those from the pattern ***G70%*** was only 0.066, whereas the mean RF distance increased to 0.366 for the pattern

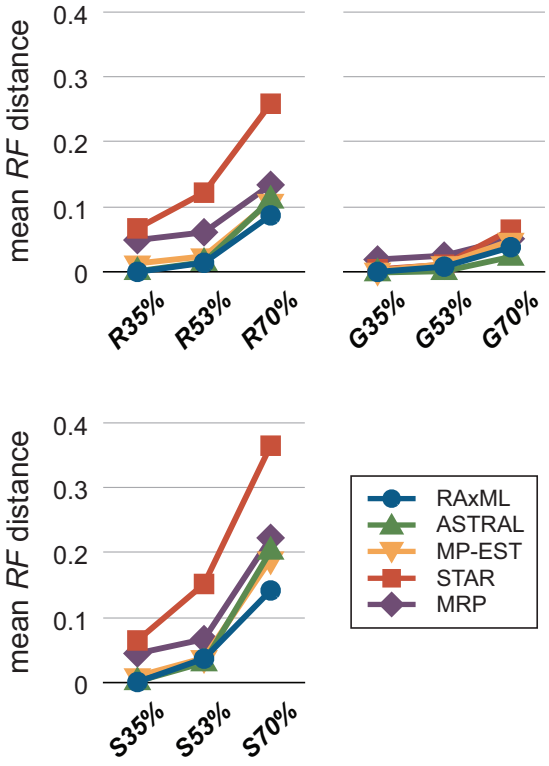


Fig. 8. The mean RF distances between species trees inferred from the complete yeast data set versus those inferred from data sets with various amounts of simulated missing data. Missing data were generated using one of three patterns, ***R***, ***G***, and ***S*** as described in the main text and in figure 3. Species trees were inferred using concatenation (RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods.

S70% even though the amount of missing data was the same (fig. 8).

The Impact of Missing Data in the Presence of Both ILS and Gene Rate Heterogeneity

Simulation analyses of our 5-taxon species trees S1–S4 (fig. 9*a*) demonstrated that when θ was low (i.e., 0.001 for species trees S1 and S2), on average 83% of the simulated gene trees (when rooted with species *E*) were congruent with the species tree topology. When θ was high (i.e., 0.01 for species trees S3 and S4), the topologies of simulated gene trees were highly variable. Under these circumstances, on average only 16% of the simulated gene trees were congruent with the species tree topology. Importantly, despite the highly discordant topologies among gene trees, the most probable gene tree still matched the species tree topology. Thus, species trees S3 and S4 are not in the anomaly zone (Degnan and Rosenberg 2006). In addition, each of these simulated data sets included both slow- (i.e., gene sequences simulated on the species tree S1 or S3) and fast-evolving (i.e., gene sequences simulated on the species tree S2 or S4 that possessed long external branches leading to species *A*–*C*) genes (fig. 9*b*). This allowed us to evaluate the effects of missing data in the presence of both ILS and gene rate heterogeneity.

Despite the fact that these data sets included different ratios of slow- versus fast-evolving genes (i.e., 9:1, 7:3, 5:5, or 3:7), concatenation (RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods consistently recovered the true species tree from complete data sets as the number of genes increased. When ILS was low, the proportion of simulations in which all five methods recovered the true species tree increased to 1.0 as the number of genes increased to 100 (fig. 10). When ILS was high, all methods similarly recovered the true species tree with a proportion of $\geq$0.99 as the number of genes increased to 500 (fig. 10). These results suggest that in the presence of both ILS and gene rate heterogeneity, all methods perform reliably when data sets are complete and a sufficient number of genes are sampled.

We generated missing data that were concentrated in fast-evolving genes using one of two patterns, "random" versus "nonrandom" (fig. 9*b*). Here, we use quotation marks to designate these patterns because in both cases missing data were concentrated in fast-evolving genes and certain species. Thus, they necessarily exhibit some inherent bias. For random missing data, a single gene sequence from one of ingroup species *A*–*C* was randomly removed in each of the fast-evolving genes; for nonrandom missing data, gene sequences from only species *C* were removed in all fast-evolving genes. For these two patterns, the amount of missing data was linked to the percentage of fast-evolving genes in the data set. For example, when 30% of the genes were simulated on the species tree S2, 30% of total gene sequences in species *C* were removed for the nonrandom missing data (fig. 9*b*).

In the case of random missing data, all five methods accurately recovered the true species tree as the number of genes increased. This is true even when the percentage of fast-evolving genes increased to 70% (i.e., 70% of the total genes contain a single missing sequence). Here, the proportion of simulations in which the concatenation method using ML recovered the true species tree was high ($\geq$0.97) as the number of genes increased to 500 and 2,000 for low and high ILS, respectively (fig. 10). Gene-tree-based coalescent methods similarly recovered the true species tree with a high proportion ($\geq$0.99), but using only 50 and 500 genes for low and high ILS, respectively (fig. 10). Thus, in the presence of gene rate heterogeneity and missing data, gene-tree-based coalescent methods are more likely to recover the true species tree when the number of genes is relatively small.

For nonrandom missing data, all methods accurately recovered the true species tree as the number of genes increased, as long as the percentage of fast-evolving genes was low (i.e., $\leq$30% and $\leq$10% for low and high ILS, respectively). Here, the proportion of simulations in which the concatenation method using ML recovered the true species tree increased to $\geq$0.88 as the number of genes increased to 50 and 2,000 for low and high ILS, respectively (fig. 10). In contrast, gene-tree-based coalescent methods recovered the true species with a higher proportion ($\geq$0.95) as the number of genes increased to 50 and 200 for low and high ILS, respectively (fig. 10). Thus, in the presence of ILS and a low amount of nonrandom missing data, the concatenation method using

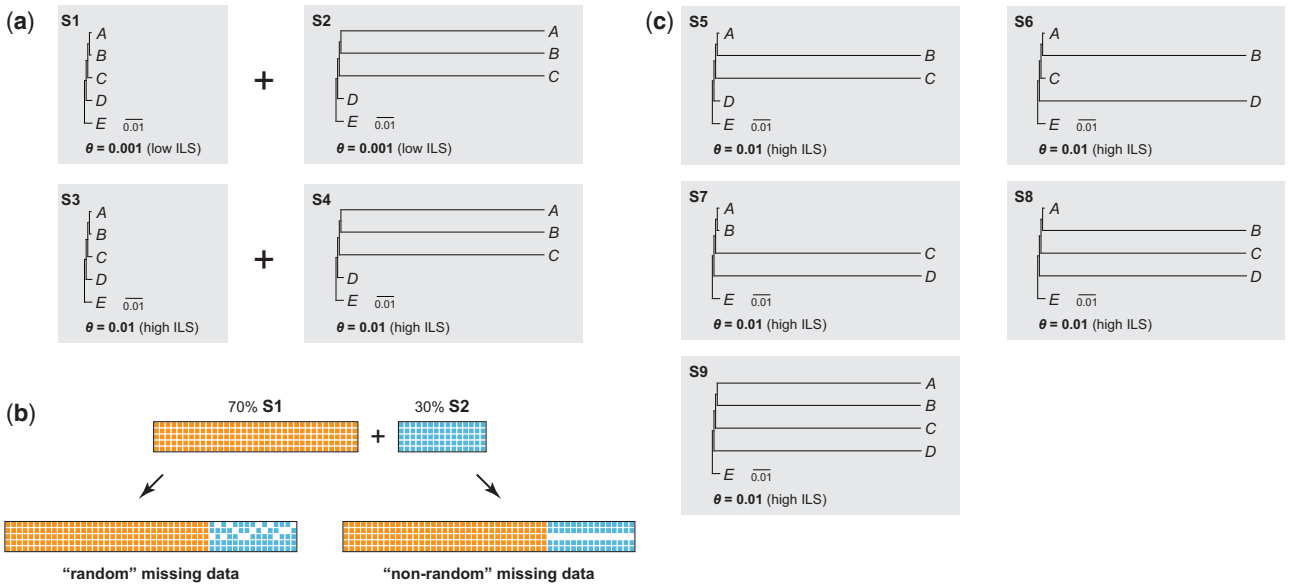


FIG. 9. DNA simulations using 5-taxon species trees to investigate the impact of missing data in the presence of both ILS and gene rate heterogeneity. DNA sequences were simulated on species trees S1–S9 under the multispecies coalescent model (Rannala and Yang 2003). The lengths of all internal branches are 0.001 (branch lengths are in mutation units), and the length of the external branch leading to outgroup species *E* is 0.004 for species trees S1–S9. 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. (*a*) Species trees used to simulate slow- and fast-evolving genes. For species trees S1 and S3, the external branches leading to species *A*, *B*, and *C* are 0.001, 0.001, and 0.002, respectively, whereas for species trees S2 and S4, these three external branches are 0.101, 0.101, and 0.102, respectively. Thus, DNA sequences simulated on species trees S1 and S3 represent slow-evolving genes, whereas DNA sequences simulated on species trees S2 and S4 represent fast-evolving genes. (*b*) Examples of the random and nonrandom missing data. For each data set, "*X*" percent of the total genes (where "*X*" ranges from 90 to 30 in decrements of 20) were simulated on the species tree S1 or S3 (slow-evolving genes), and the remaining genes were simulated on the species trees S2 or S4 (fast-evolving genes). To generate random missing data, a single gene sequence from one of the species that possessed long external branches (i.e., species *A*, *B*, or *C*) was randomly removed for each of the fast-evolving genes. To generate nonrandom missing data, gene sequences only from species *C* were removed for all fast-evolving genes. (*c*) Additional species trees with varying numbers and placements of the long external branches to simulate fast-evolving genes. For the species tree S5, the external branches leading to species *B* and *C* are 0.101 and 0.102, respectively; for the species tree S6, the external branches leading to species *B* and *D* are 0.101 and 0.103, respectively; for the species tree S7, the external branches leading to species *C* and *D* are 0.102 and 0.103, respectively; for the species tree S8, the external branches leading to species *B*, *C*, and *D* are 0.101, 0.102, and 0.103, respectively; and for the species tree S9, the external branches leading to species *A*, *B*, *C*, and *D* are 0.101, 0.101, 0.102, and 0.103, respectively.

ML performs worse than gene-tree-based coalescent methods, but still identifies the true species tree as the number of genes increases.

When increasing the percentage of fast-evolving genes in the data sets, thus simultaneously increasing the amount of missing data in species *C*, these methods differed sharply in their ability to recover the true species tree. Under these circumstances, ASTRAL, MP-EST, and MRP similarly recovered the true species tree with a high proportion ($\geq$0.96) as the number of genes increased to 50 and 1,000 for low and high ILS, respectively (fig. 10). In contrast, the proportion of simulations in which the concatenation method using ML recovered the true species tree decreased as the amount of nonrandom missing data increased. When ILS was low, the proportion of simulations in which the concatenation method using ML recovered the true species tree dropped to approximately 0.60 as the percentage of fast-evolving genes increased to 50% (fig. 10). Here, the concatenation method using ML inferred an incorrect species tree (topology II in fig. 11*a*) with a proportion of 0.40 as the number of genes increased to 2,000. When further increasing the percentage of fast-evolving genes to 70% this became even worse: The proportion of simulations in which the concatenation method using ML recovered the true species tree declined to zero as the number of genes increased to 2,000 (fig. 10). In these cases, the concatenation method using ML consistently inferred two incorrect species trees as the number of genes increased (fig. 11*a*). When the number of genes increased to 2,000, the BP values for these two incorrect relationships, that is, species *C* as sister to species *A* (topology I) or species *B* (topology II), ranged from 51 to 100 with a median of 79 and 90, respectively (fig. 11*b*). In addition, when data sets included only slow-evolving (i.e., gene sequences simulated on the species tree S1) or only fast-evolving (i.e., gene sequences simulated on the species tree S2) genes, the concatenation method using ML accurately recovered the true species tree as the number of genes increased, even if 70% of total genes were missing in species *C* (fig. 10).

Recent theoretical and simulation studies have demonstrated that phylogenetic analyses using ML can be biased by the combination of missing data and among-site rate variation (Lemmon et al. 2009; Simmons 2012b; Xia 2014). Our

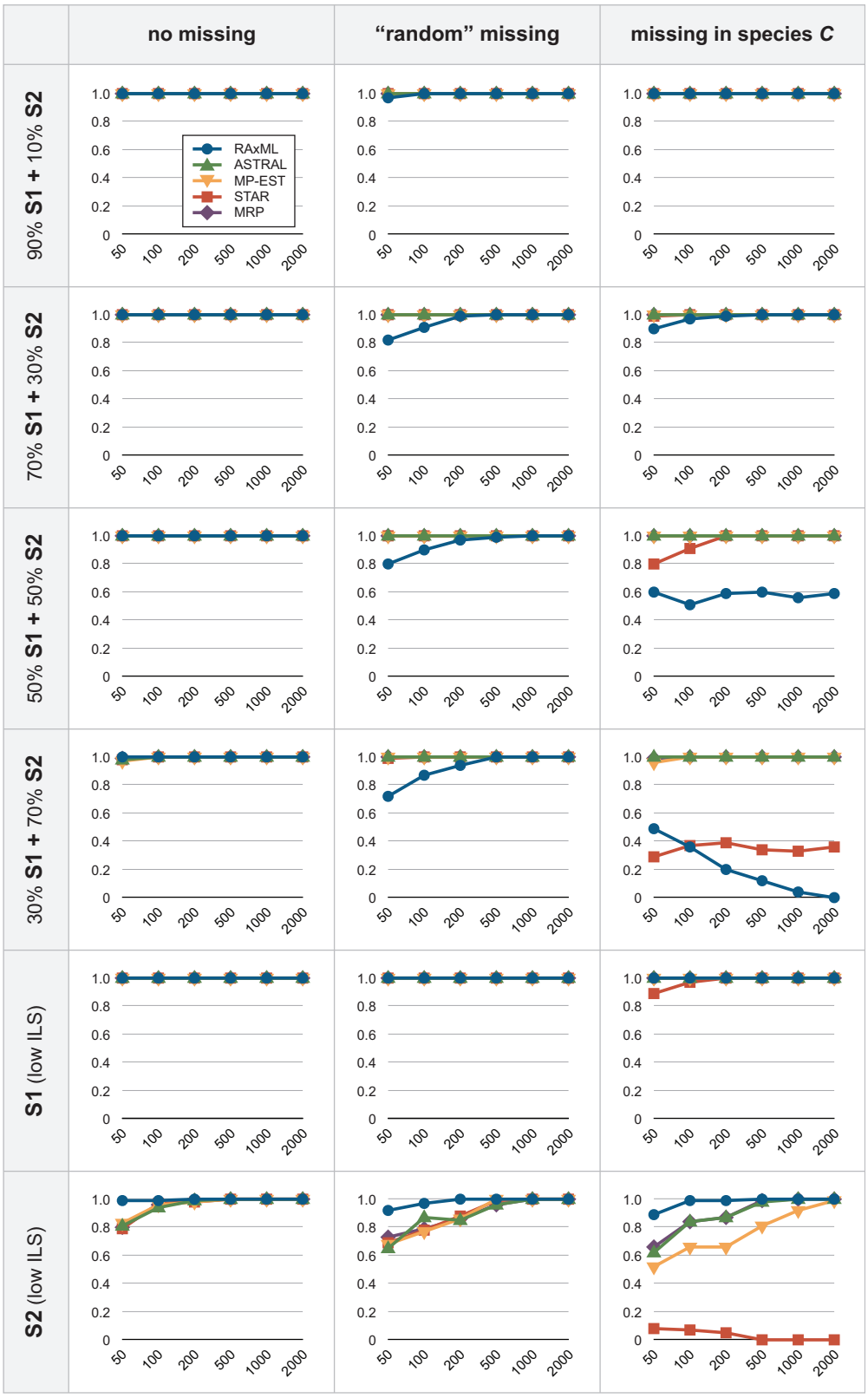


Fig. 10. Proportions of simulations in which the true five-taxon species tree was recovered from 50-, 100-, 200-, 500-, 1,000-, and 2,000-gene data sets using concatenation (unpartitioned RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods. For each data set, slow-evolving genes were simulated on the species tree S1 or S3, whereas fast-evolving genes were simulated on the species tree S2 or S4 (fig. 9*a*). Missing data were generated in fast-evolving genes using the random or nonrandom pattern as described in the main text and in figure 9*b*. For comparison, we additionally simulated data sets that included only slow-evolving (species tree S1 or S3) or only fast-evolving (species tree S2 or S4) genes.

(continued)

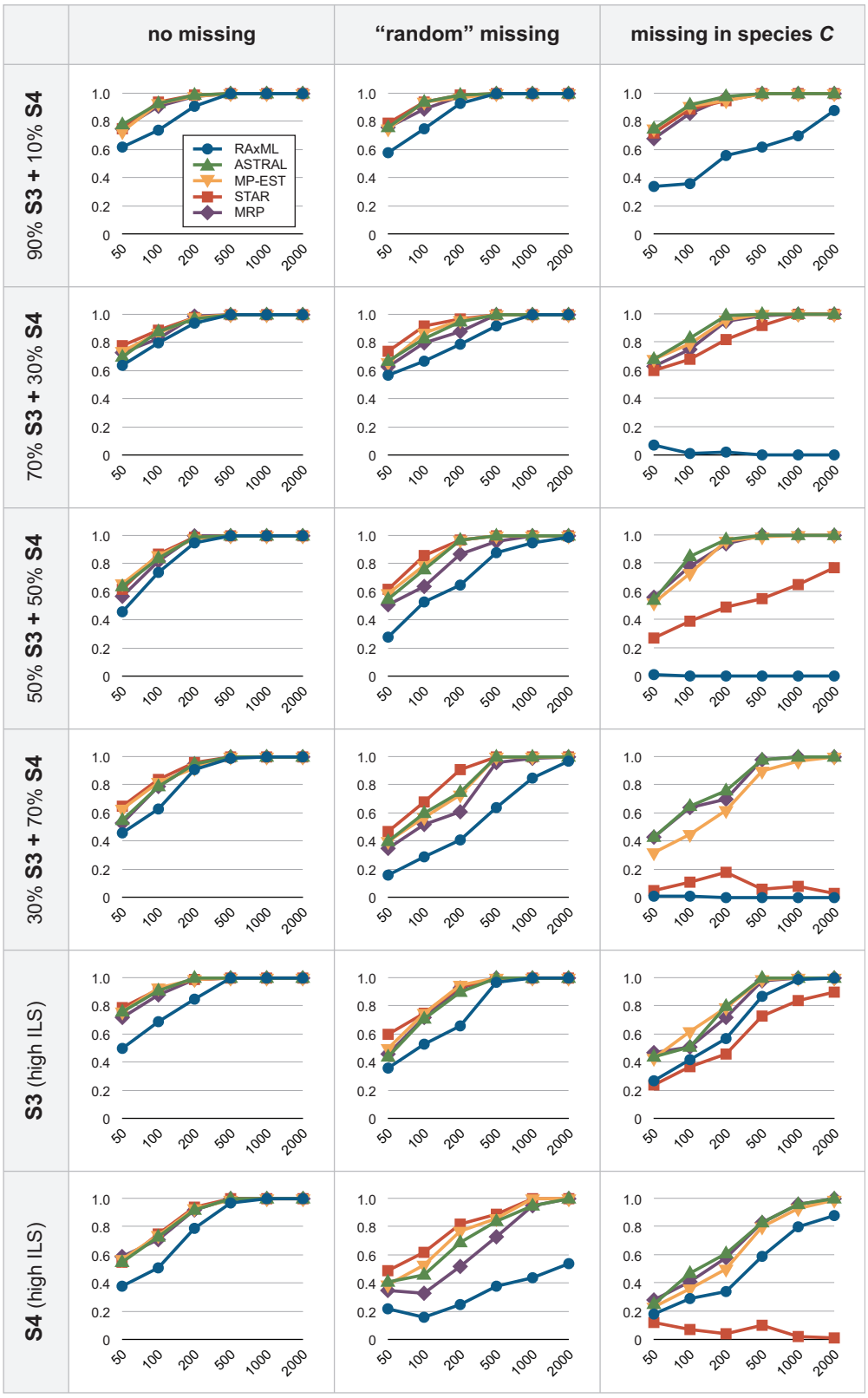


FIG. 10. (Continued)

results demonstrate that in the presence of gene rate heterogeneity, nonrandom missing data can similarly mislead the concatenation method using ML even when data sets are decisive. Under these circumstances, we demonstrate that adding genes with nonrandom missing data decreases the performance of the concatenation method using ML, which consistently infers the incorrect species trees. Moreover, when ILS was high, the proportion of simulations in which the concatenation method using ML recovered the true species tree declined to zero as the percentage of fast-

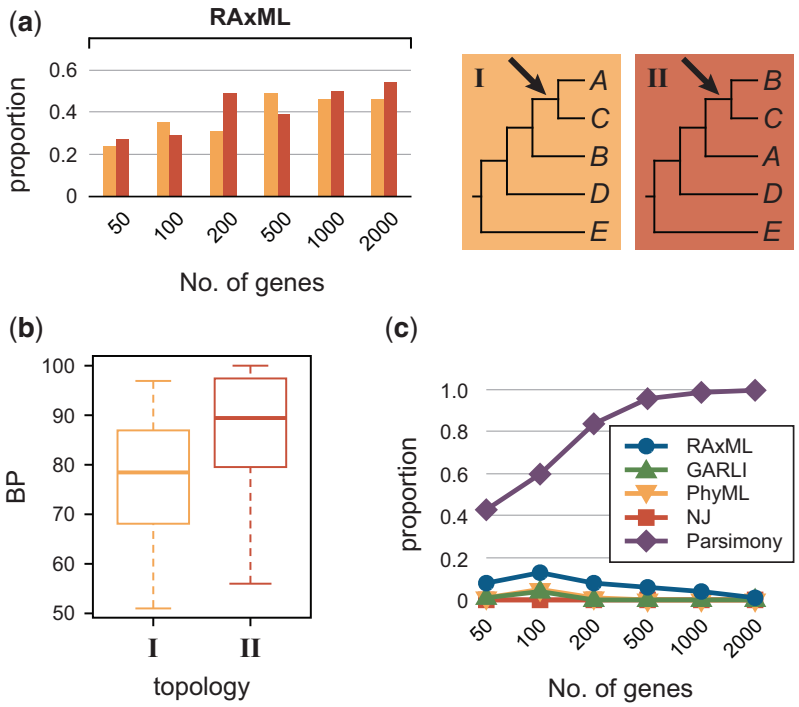


Fig. 11. (*a*) Proportions of simulations in which two incorrect five-taxon species trees were inferred from 50-, 100-, 200-, 500-, 1,000-, and 2,000-gene data sets using the concatenation method (unpartitioned RAxML). For each data set, 30% of genes were simulated on the species tree S1 (i.e., slow-evolving genes), and the remaining genes were simulated on the species tree S2 (i.e., fast-evolving genes; fig. 9*a*). Missing data were generated in fast-evolving genes using the nonrandom pattern as described in the main text and in figure 9*b*. The arrow highlights the branch of interest, and for which BPs are summarized in figure 11*b*. (*b*) Summary of BPs inferred from the 2,000-gene data sets using the concatenation method (unpartitioned RAxML). For each data set, 600 genes were simulated on the species tree S1 (i.e., slow-evolving genes), and the remaining 1,400 genes were simulated on the species tree S2 (i.e., fast-evolving genes; fig. 9*a*). Missing data were generated in fast-evolving genes using the nonrandom pattern as described in the main text and in figure 9*b*. (*c*) Proportions of simulations in which the true five-taxon species tree was recovered 50-, 100-, 200-, 500-, 1,000-, and 2,000-gene data sets using concatenation methods (partitioned RAxML, GARLI, PhyML, NJ, and parsimony). For each data set, 70% of genes were simulated on the species tree S3 (i.e., slow-evolving genes), and the remaining genes were simulated on the species tree S4 (i.e., fast-evolving genes; fig. 9*a*). Missing data were generated in fast-evolving genes using the nonrandom pattern as described in the main text and in figure 9*b*.

evolving genes increased to 30% (fig. 10), that is, only 30% of gene sequences were missing in species *C*. Under these circumstances, all ML programs implemented here (i.e., unpartitioned/partitioned RAxML, GARLI, and PhyML; fig. 11*c*) consistently inferred two incorrect species trees (fig. 11*a*), whereas the concatenation method using parsimony still accurately recovered the true species tree as the number of genes increased (fig. 11*c*). Thus, the negative effects of nonmissing data on the concatenation method using ML appear to be greatly exacerbated when gene rate heterogeneity is combined with a high degree of ILS. These results collectively indicate that the presence of nonrandom missing data, high ILS, and gene rate heterogeneity represent a triple threat to the concatenation method using ML.

STAR was also misled by nonrandom missing data when 70% of gene sequences were missing in the species *C*. Here, in the presence of gene rate heterogeneity, the proportion of simulations in which STAR recovered the true species tree decreased to 0.36 and 0.03 as the number of genes increased to 2,000 for low and high ILS, respectively (fig. 10). Similar to the concatenation method using ML, STAR consistently inferred two incorrect species trees (topologies I and II in fig. 11*a*) as the number of genes increased. In addition, for data sets that included only slow-evolving genes (i.e., gene sequences simulated on the species tree S1 or S3), STAR accurately recovered the true species tree even when 70% of total genes were missing in species *C* (fig. 10). However, when data sets included only fast-evolving genes (i.e., gene sequences simulated on the species tree S2 or S4), STAR consistently inferred two incorrect species trees (topologies I and II in fig. 11*a*) as the number of genes increased (fig. 10). These results indicate that a high amount of nonrandom missing data can also mislead the gene-tree-based coalescent method STAR. However, unlike the concatenation method using ML, it appears that the adverse effects of nonrandom missing data on STAR are not due to the presence of gene rate heterogeneity, but instead due to the presence of fast-evolving genes as simulated using a combination of long external and short internal branches in the species tree.

In addition to the species tree S4, we simulated fast-evolving genes under a high degree of ILS using species trees S5–S9 (fig. 9*c*), where we varied the number and placement of the long external branches. Here, DNA sequences from one of the ingroup species that possessed long external branches, for example, species *B* or *C* for the species tree S5, were removed in all fast-evolving genes. Thus, these data sets

allowed us to more thoroughly examine the effects of nonrandom missing data on species tree estimation in the presence of ILS and gene rate heterogeneity. Similar to species trees S3 and S4, the topologies of gene trees simulated on species trees S5–S9 were highly variable: On average only 16% of the simulated gene trees were congruent with the species tree topology. Despite the presence of a high amount of nonrandom missing data (i.e., 70% of gene sequences were missing in species *B*, *C*, or *D*), ASTRAL, MP-EST, and MRP consistently recovered the true species tree with a high proportion ($\geq$0.97) as the number of genes increased to 1,000 (fig. 12). This is true for all 14 data sets we examined here, suggesting that these methods are especially robust to nonrandom missing data. In contrast, STAR and the concatenation method using ML produced inconsistent results for most of these 14 data sets. For missing data concentrated in species *B*, STAR accurately recovered the true species tree with a high proportion ($\geq$0.99) as the number of genes increased to 500; whereas the concatenation method using ML consistently inferred an incorrect species tree (topology II in fig. 11*a*) as the number of genes increased to 2,000 when fast-evolving genes were simulated on species trees S5 and S8 (fig. 12). For missing data concentrated in species *C*, the proportion of simulations in which STAR recovered the true species dropped dramatically ($\leq$0.18) as the number of genes increased to 2,000 when fast-evolving genes were simulated on species trees S4 and S9 (fig. 12). The concatenation method using ML performed even worse under these circumstances: The proportion of simulations in which the true species tree was recovered decreased to zero as the number of genes increased to 2,000 for all species trees examined here (i.e., species trees S4, S5, S7, S8, and S9; fig. 12). For missing data concentrated in species *D*, the proportion of simulations in which STAR recovered the true species dropped to $\leq$0.04 as the number of genes increased to 2,000 for all species trees examined here (i.e., species trees S6–S9; fig. 12). In these cases, the proportion of simulations in which the concatenation method using ML recovered the true species was similarly low ($\leq$0.40) for species trees S7–S9, but increased to 0.95 as the number of genes increased to 2,000 for the species tree S6 (fig. 12). Thus, these results collectively suggest that in the presence of both ILS and gene rate heterogeneity, STAR and the concatenation method using unpartitioned or partitioned ML are more likely to be misled by nonrandom missing data. In contrast, ASTRAL, MP-EST, and MRP are more robust under these circumstances.

Finally, we explored the effects of nonrandom missing data in the presence of both ILS and gene rate heterogeneity using a subset of the mammal data set assembled by Tsagkogeorga et al. (2013). The six-taxon Scrotifera data set we analyzed here included 1,394 genes, and the average number of nucleotide sites for each gene was 1,078. Phylogenomic analyses of this data set produced a well-supported species tree (fig. 13*a*), which was congruent with the species trees inferred by Tsagkogeorga et al. (2013) and Liu, Xi, and Davis (2015). In addition, a clade consisting of Cetartiodactyla plus Perissodactyla was supported by all analyses using the complete data set, that is, 95/82, 72, 80, 81, 78, and 73 BP for the concatenation method using ML (unpartitioned/partitioned RAxML), the concatenation method using parsimony, ASTRAL, MP-EST, STAR, and MRP, respectively (fig. 13*c*). Thus, we interpret the monophyly of Cetartiodactyla plus Perissodactyla as the accepted relationship. As expected, the rapid radiation of the Laurasiatherian mammals (Zhou et al. 2012) produced a short internal branch separating the four orders in the inferred species tree (fig. 13*a*), and a high degree of discordance among individual gene trees (Liu, Xi, and Davis 2015). Moreover, the relative evolutionary rates of these 1,394 genes varied from 0.15 to 3.14, suggesting the presence of gene rate heterogeneity in the Scrotifera data set.

When increasing the amount of missing data that were concentrated in *Felis catus*, the clade of Cetartiodactyla plus Perissodactyla was still moderately to weakly supported by ASTRAL, MP-EST, and MRP. Here, when the amount of nonrandom missing data was high (i.e., 70% of gene sequences were missing in *F. catus*), the topology remained the same but BP values for this clade dropped to 56, 53, and 66 BP for ASTRAL, MP-EST, and MRP, respectively (fig. 13*c*). Similarly, the clade of Cetartiodactyla plus Perissodactyla was supported by the concatenation method using parsimony with 81 BP when 70% of gene sequences were missing in *F. catus* (fig. 13*c*). In contrast, the concatenation method using ML produced incongruent placements of Cetartiodactyla across data sets with various amounts of nonrandom missing data. Here, when the amount of nonrandom missing data was low (i.e., 10% of gene sequences were missing in *F. catus*), the concatenation method using ML supported the accepted placement of Cetartiodactyla as sister to Perissodactyla (67 and 59 BP for unpartitioned and partitioned RAxML, respectively; fig. 13*c*). When the amount of nonrandom missing data was high (i.e., 70% of gene sequences were missing in *F. catus*), however, the concatenation method using ML instead placed Cetartiodactyla as sister to Carnivora (fig. 13*b*) with moderate support (76 and 70 BP for unpartitioned and partitioned RAxML, respectively; fig. 13*c*). STAR similarly placed Cetartiodactyla incorrectly as sister to Carnivora when the amount of nonrandom missing data was high (i.e., 70% of gene sequences were missing in *F. catus*), but the BP value was lower (i.e., 58; fig. 13*c*). Thus, our analyses of this six-taxon Scrotifera data set corroborate our simulation results above, and suggest that in the presence of both ILS and gene rate heterogeneity, nonrandom missing data may positively mislead species tree estimation. This appears to be especially pronounced for STAR and the concatenation method using unpartitioned or partitioned ML, which consistently produce incorrect results when the amount of nonrandom missing data is high.

Conclusions

Our analyses identify that missing data can indeed influence species tree estimation under a variety of circumstances. We show that two gene-tree-based coalescent methods, ASTRAL and MP-EST, and the supertree method MRP are more robust even when the amount of missing data is high. To our knowledge, few studies have developed statistical metrics to assess

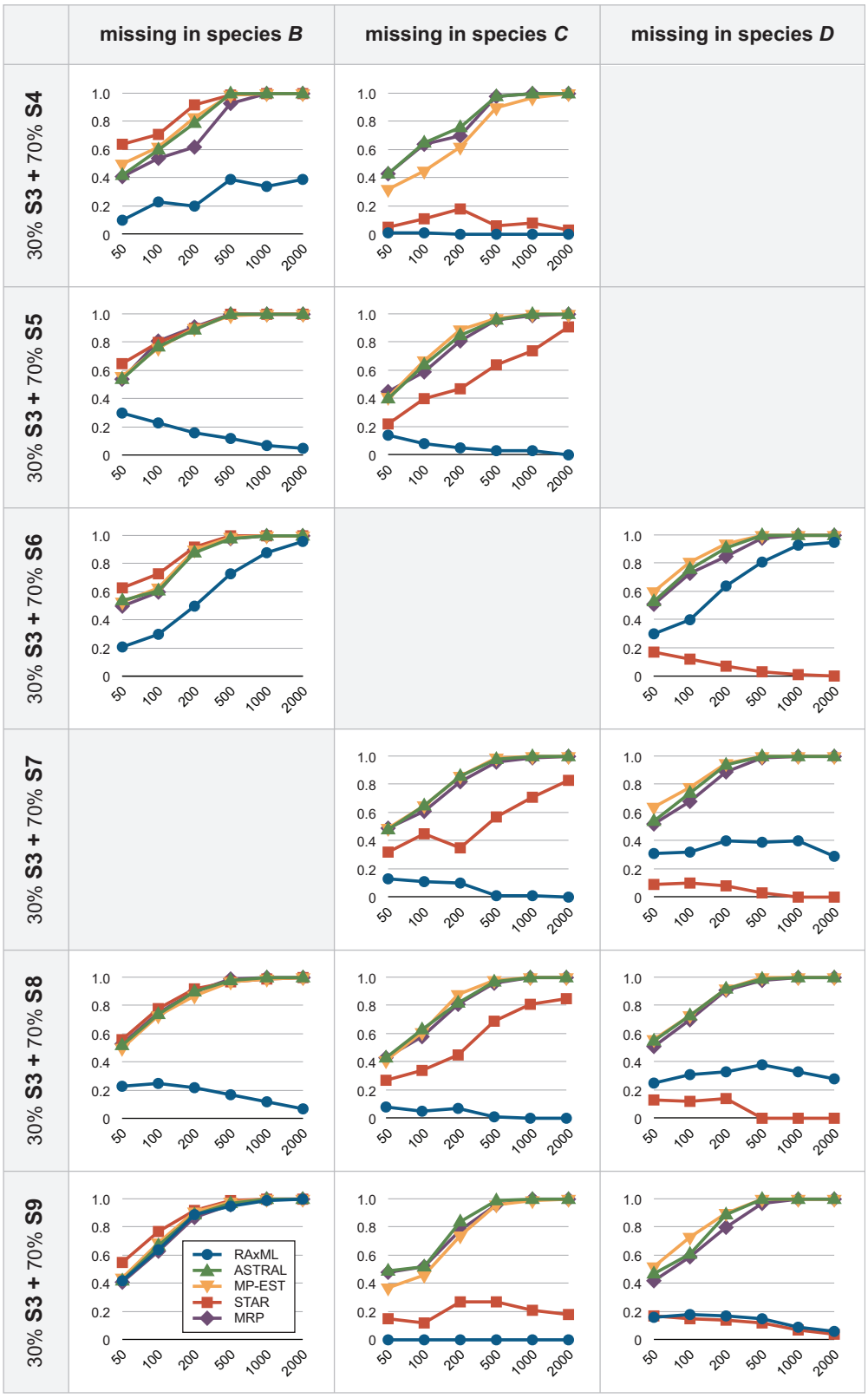
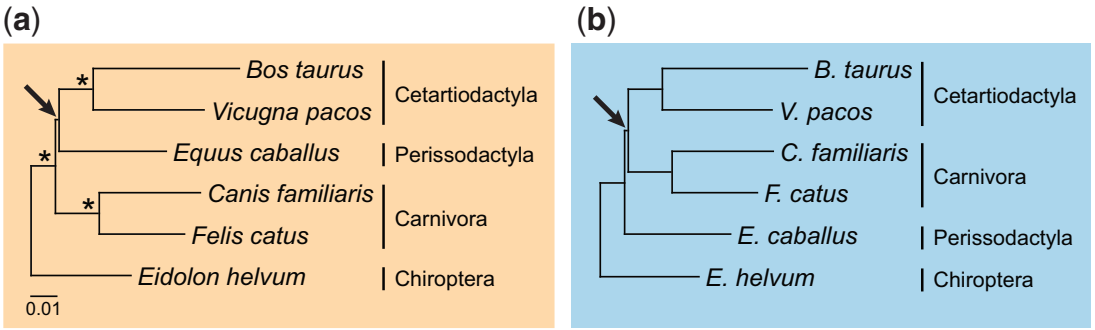


FIG. 12. Proportions of simulations in which the true five-taxon species tree was recovered from 50-, 100-, 200-, 500-, 1,000-, and 2,000-gene data sets using concatenation (unpartitioned RAxML), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods. For each of these data sets, 30% of the total genes were simulated on the species tree S3 (slow-evolving genes; fig. 9*a*) and the remaining genes were simulated on one of species trees S5–S9 (fast-evolving genes; fig. 9*c*). Missing data were generated in fast-evolving genes by removing gene sequences from one of the species that possessed long external branches (species *B*, *C*, or *D*).



(c)

Data set	conc.		coal.			super.
	RAxML	Parsimony	ASTRAL	MP-EST	STAR	MRP
complete	95/82	72	80	81	78	73
10%	67/59	56	75	74	71	67
30%	62/62	52	74	73	52	59
50%	69/61	72	64	61		58
70%	76/70	81	56	53	58	66

FIG. 13. (*a*) The species tree of six mammals inferred from the complete 1,394-gene data set using concatenation (unpartitioned/partitioned RAxML and parsimony as implemented in PAUP*), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods. Branch lengths shown here (in mutation units) were estimated from the concatenated matrix using unpartitioned RAxML. BPs are indicated for each branch, and an asterisk indicates that the branch is supported by 100 BPs using all methods. The arrow highlights the branch of interest, and for which BPs are summarized in figure 13*c*. (*b*) The alternative species tree inferred from data sets with nonrandom missing data using concatenation (unpartitioned/partitioned RAxML) and gene-tree-based coalescent (STAR) methods. Branch lengths shown here (in mutation units) were estimated from the concatenated matrix using unpartitioned RAxML. The arrow highlights the branch of interest, and for which BPs are summarized in figure 13*c*. (*c*) Summary of BPs inferred from data sets with various amounts of missing data that were concentrated in *Felis catus*. We first sorted all 1,394 genes based on relative evolutionary rates estimated using the DistR method. DNA sequences from *F. catus* were then removed in fast-evolving genes, which corresponded to "*X*" percent of the total genes (where "*X*" ranges from 10 to 70 in increments of 20). The cell with hatching indicates bootstrap support is below 50 BP; colored cells (orange: Cetartiodactyla as sister to Perissodactyla, blue: Cetartiodactyla as sister to Carnivora) indicate relationships that received bootstrap support $\geq$50 BP.

the fitness of sequence data for phylogenomic analyses in the presence of missing data (e.g., Waddell 2005; Sanderson et al. 2010; Steel and Sanderson 2010). The recent establishment of phylogenetic decisiveness (Sanderson et al. 2010; Steel and Sanderson 2010) to characterize incomplete taxon coverage is a major advancement in this respect and merits wider usage. However, it does not completely address the potentially biased effects of nonrandom missing data we identify here. Owing to the prevalence of missing data in phylogenomic analyses, it is entirely possible that nonrandom missing data may be a more widespread phenomenon than is thought. Whether or not this is the case, and to what extent this could be exacerbated by gene rate heterogeneity and a high degree of ILS is a fertile ground for future investigation.

Finally, how should we as a community deal with missing data in empirical analyses? This depends on numerous factors, including the percentage of missing data, the pattern of missing data (e.g., random vs. nonrandom), the degree of ILS, and the sample size (e.g., the number of taxa and the number of genes). In general, under ideal circumstances, we recommend that practitioners not filter genes with missing data. This recommendation was also concluded recently by Streicher et al. (2015). In addition, we recommend choosing models and methods that are more robust to the adverse effect of missing data. Our study demonstrates the additional ways in which these models and methods could be better refined and applied to phylogenomic data sets.

Materials and Methods

Simulating Missing Data Using 17-Taxon Species Trees under Varying Degrees of ILS

To investigate the impact of missing data on species tree estimation in the presence of ILS, we simulated DNA sequences on six 17-taxon species trees with two different topologies (i.e., symmetrical species trees T1–T3 and pectinate species trees T4–T6) under the multispecies coalescent model (Rannala and Yang 2003). For each of the ultrametric species trees T1–T6 (fig. 2), species *Q* was designated as the outgroup and the height of the tree was held constant at 0.05 (lengths herein are reported in mutation units, i.e., the number of nucleotide substitutions per site). For each gene, one allele was sampled from each of the species *A*–*Q*. The lengths of the internal branches were held constant in species trees T1–T3 and T4–T6 (i.e., 0.01 and 0.003125, respectively). In addition, we applied different values of the population size parameter θ to simulate varying degrees of ILS (i.e., 0.001, 0.01, and 0.1 for species trees T1–T3, respectively, and 0.0003125, 0.003125, 0.03125 for species trees T4–T6, respectively; fig. 2).

For each of the species trees T1–T6, 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 whether our values of θ were comparable with empirical studies, we 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 T1–T6 (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).

We then simulated 50, 100, 200, 500, 1,000, and 2,000 gene trees on each of species trees T1–T6 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). Each simulation was repeated 100 times, which resulted in a total of 600 data sets for each of the species trees T1–T6.

Next, we generated missing data on each of our simulated data sets. The three patterns we used to generate missing data were designated as ***R***, ***G***, and ***S*** (fig. 3). These largely followed Hovmöller et al. (2013) and Roure et al. (2013), and were designed to mimic patterns of missing data from published phylogenomic data sets. For pattern ***R***, missing data were randomly distributed across ingroup species for all genes, that is, it was equally likely to remove sequences from any gene and any of the ingroup species *A*–*P*. Here, 35%, 53%, or 70% of the total gene sequences were removed in each data set, which corresponded to removing an average of 6, 9, or 12 sequences per gene, respectively. In addition, for each gene, a minimum of four taxa (three ingroups and one outgroup) was required to produce a meaningful gene tree. This emulates the pattern of missing data due to low coverage in next-generation sequencing. For pattern ***G***, missing data were randomly distributed across ingroup species but concentrated in a subset of randomly chosen genes. To do this, we first randomly selected a subset of genes (i.e., 46%, 69%, or 92% of the total genes to achieve 35%, 53%, or 70% missing data, respectively), and then for each selected gene, randomly removed sequences from ingroup species *A*–*P* until there were only three ingroup species remaining. This pattern of missing data represents the case where certain genes are more likely to be lost from the genome, but in which these losses are not concentrated in a subset of species. For pattern ***S***, missing data were randomly distributed among all genes but concentrated in a subset of randomly chosen ingroup species. To do this, we first randomly selected 13 ingroup species, and then for each selected species, randomly removed sequences from a predefined proportion of all genes (i.e., 0.46, 0.69, or 0.92 to achieve 35%, 53%, or 70% missing data, respectively). This pattern of missing data represents the case where only a few species are sampled completely (e.g., full genome sequences) and other species are sampled less completely (e.g., shallow sequencing or low-quality DNA). Finally, for each of the three patterns, we used the metric of phylogenetic decisiveness sensu Sanderson et al. (2010) to characterize the pattern of incomplete taxon coverage induced by missing data. Perl scripts from Sanderson et al. (2010) were used to determine whether a data set was decisive regarding the true species tree (i.e., species trees T1–T6).

Species trees were estimated using 1) the concatenation method as implemented in RAxML (Stamatakis 2014), 2) three commonly used gene-tree-based coalescent methods (i.e., ASTRAL, MP-EST, and STAR), and 3) a supertree method MRP (Baum 1992; Ragan 1992). For concatenation analyses, the best-scoring ML trees were estimated from concatenated gene sequences using both unpartitioned (i.e., a single GTR + Γ model) and partitioned (i.e., a separate GTR + Γ model for each gene) models. Optimal tree searches were conducted using RAxML v8.1.3 with five independent searches starting from random trees (-d -f o -m GTRGAMMAX —no-bfgs). For coalescent analyses, gene trees were first estimated using RAxML with the GTR + Γ model (-d -f o -m GTRGAMMAX —no-bfgs), and rooted with species *Q*. These estimated gene trees were then utilized to construct species trees using ASTRAL v4.7.1, MP-EST v1.4, and the STAR method as implemented in Phybase (default settings were used for ASTRAL, MP-EST, and STAR). For supertree analyses, the MRP was first computed from estimated gene trees following Mirarab, Reaz, Bayzid, et al. (2014), and species trees were then estimated using parsimony analyses as implemented in PAUP* v.4.0b10 (Swofford 2002) with the standard heuristic search (hsearch start = stepwise addseq = random nreps = 100 savereps = no swap = tbr hold = 1 multrees = yes). Topological differences between estimated species trees and their true species tree were measured using the normalized RF distance as implemented in RAxML (-f r). The normalized RF distance, or the RF distance (Kupczok et al. 2010), ranges between 0.0 and 1.0, and is calculated by dividing the RF metric (Robinson and Foulds 1981) by $2 \times (n - 3)$, where n is the number of species. If the estimated species tree matches the true species tree, the RF distance equals zero; if there is one split (i.e., a bipartition of a set of species) that is present in the estimated species tree but not in the true species tree, the RF distance equals 0.071 and 0.050 for a data set with 17 and 23 species, respectively. The mean RF distance was then calculated on the 100 data sets for each of the gene number categories (i.e., 50, 100, 200, 500, 1,000, and 2,000 genes).

Simulating Missing Data Using a Phylogenomic Data Set of 23 Yeasts

To further explore the impact of missing data on species tree estimation in the presence of ILS, we conducted analyses similar to the ones described above but instead using an empirical data set. For this objective, we obtained the amino acid sequences assembled by Salichos and Rokas (2013), which included 1,070 orthologs (referred to here as

genes) from 23 budding yeast genomes. For each gene, amino acid sequences were aligned using MUSCLE v3.8.31 (Edgar 2004) with the default settings. We first cleaned up the data by removing sequences from the alignment if they contained less than 70% of the total alignment length (Jiao et al. 2012). Poorly aligned amino acid sequences were further removed using trimAl v1.2rev59 (-resoverlap 0.75 -seqoverlap 80) (Capella-Gutiérrez et al. 2009), and ambiguously aligned sites were trimmed using trimAl with the heuristic automated method (-automated1). We included only those genes containing amino acid sequences from all 23 species. *Candida lusitaniae* was used to root the gene trees.

Missing data were similarly generated on the complete data set using one of three patterns described above (i.e., ***R***, ***G***, and ***S***; fig. 3). For the pattern ***R***, 35%, 53%, or 70% missing data were randomly distributed across ingroup species and all genes. For the pattern ***G***, we first randomly selected a subset of genes (i.e., 42%, 64%, or 85% of the total genes to achieve 35%, 53%, or 70% missing data, respectively), and then for each selected gene, randomly removed sequences from ingroup species until only three ingroup species remained. For the pattern ***S***, we first randomly selected 19 ingroup species, and then for each selected species, randomly removed sequences from a predefined proportion of all genes (i.e., 0.42, 0.64, or 0.85 to achieve 35%, 53%, or 70% missing data, respectively).

For concatenation analyses, the best-scoring ML trees were inferred from concatenated gene sequences using RAxML with a single WAG model (Whelan and Goldman 2001) following Salichos and Rokas (2013). Optimal tree searches were conducted using five independent searches starting from random trees (-d -f o -m PROTGAMMAWAGF —no-bfgs). For coalescent and supertree analyses, gene trees were inferred from each gene using RAxML with the WAG model (-d -f o -m PROTGAMMAWAG —no-bfgs), and rooted with *C. lusitaniae*. These gene trees were then utilized to construct species trees using ASTRAL, MP-EST, STAR, and MRP as described above. Each simulation of missing data was repeated 100 times, and the mean RF distances were calculated as described above to assess topological differences between species trees inferred from the complete data set and those from data sets with various amounts of missing data.

Simulating Missing Data Using 5-Taxon Species Trees in the Presence of Both ILS and Gene Rate Heterogeneity

Recent studies have shown that rate heterogeneity across sites can greatly affect species tree estimation. In particular, incongruence in the phylogenetic placement of key lineages has been attributed to fast-evolving sites (e.g., Goremykin et al. 2009, 2013; Zhong et al. 2011; Xi et al. 2013, 2014). Thus, we are specifically interested in the influence of missing data that are concentrated in fast-evolving genes. This could be particularly relevant to target enrichment methods (e.g., multiplex PCR and sequence capture) (Thomson et al. 2008; Turner et al. 2009; Lemmon and Lemmon 2013) when universal primers or probes are used. In this case, a probe set might not be designed to capture fast-evolving gene sequences, especially for species that exhibit elevated substitution rates.

To explore the impact of missing data in the presence of both ILS and gene rate heterogeneity, we simulated DNA sequences on five-taxon species trees under the multispecies coalescent model (Rannala and Yang 2003). Here, we downsized our taxon sampling to reduce the computational burden associated with these analyses. For species trees S1–S4 (fig. 9*a*), species *E* was designated as the outgroup. For each gene, one allele was sampled from each of the species *A*–*E*. The lengths of the internal branches (i.e., 0.001) and the lengths of the external branches leading to species *D* and *E* (i.e., 0.003 and 0.004, respectively) were held constant in species trees S1–S4. In order to simulate various evolutionary rates along the same external branches among species trees, we varied branch lengths rather than mutation rates as one allele was sampled from each gene. For species trees S1 and S3, the external branches leading to species *A*–*C* are short (i.e., 0.001, 0.001, and 0.002, respectively), whereas for species trees S2 and S4, these three external branches are long (i.e., 0.101, 0.101, and 0.102, respectively). Our choice of long and short branches was guided by phylogenies of clades that have undergone an ancient rapid radiation, for example, the insect clade Neoptera (Kjer et al. 2006; Whitfield and Kjer 2008) and the plant clade Malpighiales (Davis et al. 2005; Xi et al. 2012). 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, gene trees simulated on species trees S1 and S3 effectively represent slow-evolving genes, whereas gene trees simulated on species trees S2 and S4 represent genes that evolve rapidly in species *A*–*C* (referred to here as fast-evolving genes). Furthermore, we applied different values of θ to simulate varying degrees of ILS (i.e., 0.001 for species trees S1 and S2, and 0.01 for species trees S3 and S4; fig. 9*a*).

We then simulated 50, 100, 200, 500, 1,000, and 2,000 genes on species trees S1 and S2 using Phybase and Seq-Gen as described above. For each data set (fig. 9*b*), "*X*" percent of the total genes (where "*X*" ranges from 90 to 30 in decrements of 20) were simulated on the species tree S1 (slow-evolving genes), and the remaining ones were simulated on the species trees S2 (fast-evolving genes). These simulations represent different ratios of slow- versus fast-evolving genes in the data sets. Next, we generated missing data that were concentrated in fast-evolving genes using one of the two patterns we designated as random and nonrandom (fig. 9*b*). To generate random missing data, a single gene sequence from one of the species that possessed long external branches (i.e., species *A*, *B*, or *C* for the species tree S2) was randomly removed for each of the fast-evolving genes. To generate nonrandom missing data, gene sequences only from species *C* were removed for all fast-evolving genes. We additionally simulated data sets that included only slow-evolving (species tree S1) or only fast-evolving (species tree S2) genes, which functioned as a control to assess the influence of missing data in the absence of gene rate heterogeneity. For each of these two data sets, we first randomly selected 70% of the total genes (i.e., the highest

percentage of fast-evolving genes in our simulated data sets), and then generated missing data that were concentrated in these genes using the random or nonrandom pattern as described above. Species trees were then estimated for each data set using RAxML, ASTRAL, MP-EST, STAR, and MRP as described above. In addition, we conducted bootstrap analyses for each of the concatenated 2,000-gene matrices using standard bootstrapping. To do this, we first created 100 bootstrapped matrices from the original concatenated matrix (-f j -m GTRGAMMAX -N 100 —no-bfgs), and then estimated the best-scoring ML trees from bootstrapped matrices using unpartitioned RAxML as described above. For comparison, we also estimated the species trees from concatenated gene sequences using parsimony as implemented in PAUP* (pset stepmatrix = allstates gapmode = missing; bandb), NJ as implemented in PAUP* (dset distance = jc; nj breakties = random), and two other commonly used ML programs, GARLI v2.01.1067 (ratematrix = 6rate statefrequencies = estimate ratehetmodel = gamma numratecats = 4 invariantsites = none streefname = random searchreps = 5 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 5).

Similarly, we simulated data sets with missing data using species trees S3 (slow-evolving genes) and S4 (fast-evolving genes), which possessed a higher degree of ILS compared with species trees S1 and S2. In addition to the species tree S4 (i.e., three long external branches leading to species *A*–*C*), we explored the effects of nonrandom missing data in the presence of both ILS and gene rate heterogeneity using species trees S5–S9 (fig. 9*c*). For these species trees, we varied the number and placement of long external branches. Here, the external branches leading to species *B* and *C* are long in species tree S5 (i.e., 0.101 and 0.102, respectively), the external branches leading to species *B* and *D* are long for the species tree S6 (i.e., 0.101 and 0.103, respectively), the external branches leading to species *C* and *D* are long for the species tree S7 (i.e., 0.102 and 0.103, respectively), the external branches leading to species *B*–*D* are long for the species tree S8 (i.e., 0.101, 0.102, and 0.103, respectively), and the external branches leading to species *A*–*D* are long for the species tree S9 (i.e., 0.101, 0.101, 0.102, and 0.103, respectively). For each of these data sets, 30% of the total genes were simulated on the species tree S3 (slow-evolving genes) and the remaining ones were simulated on one of species trees S5–S9 (fast-evolving genes). To generate nonrandom missing data, gene sequences from one of the species that possessed long external branches (i.e., species *B*, *C*, or *D*) were removed for all fast-evolving genes. Species trees were estimated using RAxML, ASTRAL, MP-EST, STAR, and MRP as described above.

Simulating Nonrandom Missing Data Using a Phylogenomic Data Set of Six Mammals

To further explore the impact of nonrandom missing data in the presence of both ILS and gene rate heterogeneity, we reanalyzed the data set assembled by Tsagkogeorga et al. (2013), which included 2,320 coding DNA sequence alignments (referred to here as genes) from 22 mammals. We first created a submatrix by pruning the original data set to include six species from the clade Scrotifera (*Bos taurus* [order Cetartiodactyla], *Canis familiaris* [Carnivora], *Eidolon helvum* [Chiroptera], *Equus caballus* [Perissodactyla], *F. catus* [Carnivora], and *Vicugna pacos* [Cetartiodactyla]), and *E. helvum* was assigned as the outgroup. These four orders were targeted because they exhibit a rapid radiation in the Late Cretaceous (Zhou et al. 2012), that is, short internal branches separate these orders in the inferred species tree (Tsagkogeorga et al. 2013). These compressed internal branches are where ILS is likely to be high.

To create a complete data set for subsequent analyses, we included only those genes containing DNA sequences from all six species. We additionally created four data sets with various amounts of missing data that were concentrated in *F. catus* (fig. 13*c*). To do this, genes were first sorted based on relative evolutionary rates estimated using the Distance Rates (DistR) method (Bevan et al. 2005). DNA sequences from *F. catus* were then removed in fast-evolving genes, which corresponded to "*X*" percent of the total genes (where "*X*" ranges from 10 to 70 in increments of 20). Species trees were estimated using concatenation (RAxML and parsimony as implemented in PAUP*), gene-tree-based coalescent (ASTRAL, MP-EST, and STAR), and supertree (MRP) methods as described above, and bootstrap support was estimated using a multilocus bootstrapping approach (Seo 2008) with 100 replicates.

Supplementary Material

Supplementary figure S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

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, and three reviewers for greatly improving the manuscript. This work was supported by the United States National Science Foundation (DEB-1120243 to C.C.D. and DMS-1222745 to L.L.).

References

Agnarsson I, May-Collado LJ. 2008. The phylogeny of Cetartiodactyla: the importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. *Mol Phylogenet Evol.* 48:964–985.

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.

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

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015. GenBank. *Nucleic Acids Res.* 43:D30–D35.

Bevan RB, Lang BF, Bryant D. 2005. Calculating the evolutionary rates of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. *Syst Biol.* 54:900–915.

Bryant D, Steel M. 2001. Constructing optimal trees from quartets. *J Algorithms.* 38: 237–259.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

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.
Cho S, Zwick A, Regier JC, Mitter C, Cummings MP, Yao JX, Du ZL, Zhao H, Kawahara AY, Weller S, et al. 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? *Syst Biol* 60:782–796.
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 Koning AP, Keeling PJ. 2006. The complete plastid genome sequence of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured. *BMC Biol.* 4:12.
de la Torre-Bárcena JE, Kolokotronis SO, Lee EK, Stevenson DW, Brenner ED, Katari MS, Coruzzi GM, DeSalle R. 2009. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. *PLoS One* 4:e5764.
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.
Driskell AC, Ane C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ. 2004. Prospects for building the tree of life from large sequence databases. *Science* 306:1172–1174.
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.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
Edwards EJ, Smith SA. 2010. Phylogenetic analyses reveal the shady history of C_4 grasses. *Proc Natl Acad Sci U S A.* 107:2532–2537.
Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19
Fulton TL, Strobeck C. 2006. Molecular phylogeny of the Arctoidea (Carnivora): effect of missing data on supertree and supermatrix analyses of multiple gene data sets. *Mol Phylogenet Evol.* 41:165–181.
Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogenet Evol.* 80:231–266.
Goremykin VV, Nikiforova SV, Biggs PJ, Zhong BJ, Delange P, Martin W, Woetzel S, Atherton RA, McLenachan PA, Lockhart PJ. 2013. The evolutionary root of flowering plants. *Syst Biol.* 62:50–61.
Goremykin VV, Viola R, Hellwig FH. 2009. Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. *J Mol Evol.* 68:197–204.
Hartmann S, Vision TJ. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol Biol* 8:95.
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.
Hovmöller R, Knowles LL, Kubatko LS. 2013. Effects of missing data on species tree estimation under the coalescent. *Mol Phylogenet Evol.* 69:1057–1062.
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.
Jiang W, Chen S-Y, Wang H, Li D-Z, Wiens JJ. 2014. Should genes with missing data be excluded from phylogenetic analyses? *Mol Phylogenet Evol* 80:308–318.
Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers J, McKain M, McNeal J, Rolf M, Ruzicka D, Wafula E, Wickett N, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13:R3.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414:450–453.
Kjer KM, Carle FL, Litman J, Ware J. 2006. A molecular phylogeny of Hexapoda. *Arthropod Syst Phylogeny.* 64:35–44.
Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.
Kupczok A, Schmidt HA, von Haeseler A. 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms Mol Biol.* 5:37.
Kvist S, Siddall ME. 2013. Phylogenomics of Annelida revisited: a cladistic approach using genome-wide expressed sequence tag data mining and examining the effects of missing data. *Cladistics* 29:435–448.
Leaché AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol.* 60:126–137.
Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst Biol.* 58:130–145.
Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst.* 44:99–121.
Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24:2542–2543.
Liu L, Xi Z, Davis CC. 2015. 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 CC, Edwards SV. 2015. Estimating phylogenetic trees from genome-scale data. *Ann N Y Acad Sci.* 1360: 36–53.
Liu L, Yu L. 2011. Estimating species trees from unrooted gene trees. *Syst Biol.* 60:661–667.
Liu L, Yu L. 2010. Phybase: an R package for species tree analysis. *Bioinformatics* 26:962–963.
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.
Maddison WP. 1997. Gene trees in species trees. *Syst Biol.* 46:523–536.
Mardis ER. 2013. Next-generation sequencing platforms. *Annu Rev Anal Chem.* 6:287–303.
McMahon MM, Sanderson MJ. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Syst Biol.* 55:818–836.
McNeal JR, Kuehl JV, Boore JL, de Pamphilis CW. 2007. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol.* 7:57.
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.
Molina J, Hazzouri KM, Nickrent D, Geisler M, Meyer RS, Pentony MM, Flowers JM, Pelser P, Barcelona J, Inovejas SA, et al. 2014. Possible loss

of the chloroplast genome in the parasitic flowering plant *Rafflesia lagascae* (Rafflesiaceae). *Mol Biol Evol.* 31:793–803.
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol.* 21:1740–1752.
Pyron RA, Wiens JJ. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Mol Phylogenet Evol.* 61:543–583.
Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol.* 1:53–58.
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.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53:131–147.
Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol.* 30:197–214.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.
Sakharkar KR, Dhar PK, Chow VTK. 2004. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int J Syst Evol Microbiol.* 54:1937–1941.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.
Sanderson MJ, McMahon MM, Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evol Biol.* 10:155.
Sanderson MJ, McMahon MM, Steel M. 2011. Terraces in phylogenetic tree space. *Science* 333:448–450.
Seo T-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol.* 25:960–971.
Simmons MP. 2012a. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. *Mol Phylogenet Evol.* 62:472–484.
Simmons MP. 2012b. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics* 28:208–222.
Simmons MP. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. *Mol Phylogenet Evol.* 80:267–280.
Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. *Am J Bot.* 98:404–414.
Springer MS, 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.
Steel M, Sanderson MJ. 2010. Characterizing phylogenetically decisive taxon coverage. *Appl Math Lett.* 23:82–86.
Streicher JW, Schulte JA, Wiens JJ. 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Syst Biol.* doi:10.1093/sysbio/syv058.
Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (and other methods) 4.0 beta. Sunderland (MA): Sinauer Associates.
Thomson RC, Shaffer HB. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Syst Biol.* 59:42–58.
Thomson RC, Shedlock AM, Edwards SV, Shaffer HB. 2008. Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. *Mol Phylogenet Evol.* 49:514–525.
Tonini J, Moore A, Stern D, Shcheglovitova M, Ortí G. 2015. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PLoS Curr.* doi:10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be.
Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. *Curr Biol.* 23:2262–2267.
Turner EH, Ng SB, Nickerson DA, Shendure J. 2009. Methods for genomic partitioning. *Annu Rev Genomics Hum Genet.* 10:263–284.
Waddell PJ. 2005. Measuring the fit of sequence data to phylogenetic model: allowing for missing data. *Mol Biol Evol.* 22:395–401.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18:691–699.
Whitfield JB, Kjer KM. 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu Rev Entomol.* 53:449–472.
Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A.* 111:E4859–E4868.
Wiens JJ. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst Biol.* 52:528–538.
Wiens JJ. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Syst Biol* 54:731–742.
Wiens JJ, Moen DS. 2008. Missing data and the accuracy of Bayesian phylogenetics. *J Syst Evol.* 46:307–314.
Wiens JJ, Morrill MC. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst Biol.* 60:719–731.
Wiens JJ, Tiu J. 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. *PLoS One* 7:e42925.
William J, Ballard O. 1996. Combining data in phylogenetic analysis. *Trends Ecol Evol.* 11:334.
Wolfe KH, Morden CW, Palmer JD. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc Natl Acad Sci U S A.* 89:10648–10652.
Wu Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66:763–775.
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.
Xia X. 2014. Phylogenetic bias in the likelihood method caused by missing data coupled with among-site rate variation: an analytical approach. In: Basu M, Pan Y, Wang J, editors. Bioinformatics research and applications. New York: Springer Publishing Company. p. 12–23.
Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, et al. 2014. Three keys to the radiation of angiosperms into freezing environments. *Nature* 506:89–92.
Zhong B, Deusch O, Goremykin VV, Penny D, Biggs PJ, Atherton RA, Nikiforova SV, Lockhart PJ. 2011. Systematic error in seed plant phylogenomics. *Genome Biol Evol.* 3:1340–1348.
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.
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.