Syst. Biol. 67(2):285–303, 2018

DOI:10.1093/sysbio/syx077
Advance Access publication September 15, 2017

To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods

ERIN K. MOLLOY AND TANDY WARNOW*

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
**Correspondence to be sent to: Department of Computer Science, Thomas M. Siebel Center for Computer Science, 201 North Goodwin Avenue, Urbana, IL 61801-2302, USA; E-mail: warnow@illinois.edu.*

Received 8 December 2016; reviews returned 8 September 2017; accepted 13 September 2017
Associate Editor: Matthew Hahn

Abstract.—With the increasing availability of whole genome data, many species trees are being constructed from hundreds to thousands of loci. Although concatenation analysis using maximum likelihood is a standard approach for estimating species trees, it does not account for gene tree heterogeneity, which can occur due to many biological processes, such as incomplete lineage sorting. Coalescent species tree estimation methods, many of which are statistically consistent in the presence of incomplete lineage sorting, include Bayesian methods that coestimate the gene trees and the species tree, summary methods that compute the species tree by combining estimated gene trees, and site-based methods that infer the species tree from site patterns in the alignments of different loci. Due to concerns that poor quality loci will reduce the accuracy of estimated species trees, many recent phylogenomic studies have removed or filtered genes on the basis of phylogenetic signal and/or missing data prior to inferring species trees; little is known about the performance of species tree estimation methods when gene filtering is performed. We examine how incomplete lineage sorting, phylogenetic signal of individual loci, and missing data affect the absolute and the relative accuracy of species tree estimation methods and show how these properties affect methods' responses to gene filtering strategies. In particular, summary methods (ASTRAL-II, ASTRID, and MP-EST), a site-based coalescent method (SVDquartets within PAUP*), and an unpartitioned concatenation analysis using maximum likelihood (RAxML) were evaluated on a heterogeneous collection of simulated multilocus data sets, and the following trends were observed. Filtering genes based on gene tree estimation error improved the accuracy of the summary methods when levels of incomplete lineage sorting were low to moderate but did not benefit the summary methods under higher levels of incomplete lineage sorting, unless gene tree estimation error was also extremely high (a model condition with few replicates). Neither SVDquartets nor concatenation analysis using RAxML benefited from filtering genes on the basis of gene tree estimation error. Finally, filtering genes based on missing data was either neutral (i.e., did not impact accuracy) or else reduced the accuracy of all five methods. By providing insight into the consequences of gene filtering, we offer recommendations for estimating species tree in the presence of incomplete lineage sorting and reconcile seemingly conflicting observations made in prior studies regarding the impact of gene filtering. [Gene tree estimation error; incomplete lineage sorting; missing data; multispecies coalescent; species tree estimation.]

Species tree estimation is greatly enabled through the use of multiple loci, and increased access to genomic data over the last decade has opened up the possibility of improving our understanding of how life has evolved on earth (Posada 2016). The traditional approach to multilocus species tree estimation is concatenation analysis, where the alignments for different loci are combined into a single supermatrix to which a phylogeny estimation method, such as maximum likelihood, is applied. Although numerical parameters may be optimized separately for each locus, this approach assumes that all sites in the concatenated alignment evolve down a common tree topology. However, this assumption does not always hold, as biological processes, such as gene duplication and loss, horizontal gene transfer, and incomplete lineage sorting, result in different genomic regions having different evolutionary histories (Ohno 1970; Syvanen 1985; Maddison 1997). Gene tree heterogeneity due to incomplete lineage sorting (ILS), which is modeled by the multispecies coalescent (MSC) model (Pamilo and Nei 1988; Maddison 1997; Rannala and Yang 2003), is likely to occur with high frequency in the presence of rapid radiations; for example, ILS is expected to have impacted many major groups, including birds (Jarvis et al. 2014), land plants (Wickett et al. 2014), lizards (Linkem et al. 2016), and placental mammals (McCormack et al. 2012). Hence, species tree estimation in the presence of ILS is receiving considerable attention (Degnan and Rosenberg 2009; Edwards 2009).

Simulations under the MSC model have shown that concatenation analysis using maximum likelihood (CA-ML) can have poor accuracy in the presence of gene tree heterogeneity due to ILS (Kubatko and Degnan 2007), leading to the conjecture (later proven in Roch and Steel 2015) that CA-ML is statistically inconsistent under the MSC model. In fact, CA-ML has been proven to converge to a tree other than the species tree as the number of genes increases under some conditions with high levels of ILS (Roch and Steel 2015); in other words, when data are generated under the MSC model, CA-ML can be positively misleading. However, recent results have shown that the model conditions under which CA-ML is statistically inconsistent are not restricted to the anomaly zone (i.e., when the most probable gene tree topology does not match the species tree topology; Degnan and Rosenberg 2006), and that CA-ML can even be statistically consistent under some model conditions in the anomaly zone (Mendes and Hahn 2017). Thus, the conditions under which CA-ML can be relied upon to provide accurate species trees are not fully understood, and there is substantial interest in statistically

consistent methods for inferring species trees under the MSC model.

Bayesian coestimation of the gene trees and the species tree is widely considered to be the most promising approach to species tree estimation; coestimation methods include BEST (Edwards et al. 2007; Liu 2008), *BEAST (Heled and Drummond 2010), and StarBeast2 (Ogilvie et al. 2017). Although simulation studies have demonstrated that these methods can offer substantial improvements in accuracy over other coalescent methods (Edwards 2009; Leaché and Rannala 2010; Bayzid and Warnow 2013), one of the most popular coestimation methods, *BEAST, does not converge in practical amounts of time on data sets with much more than 25 species and 100 genes (McCormack et al. 2009; Zimmermann et al. 2014; Leavitt et al. 2016). StarBeast2 (Ogilvie et al. 2017) is an improved version of *BEAST that may scale to somewhat larger data sets. BBCA (Zimmermann et al. 2014) is an approach for scaling coestimation methods to large numbers of genes but does not improve scalability to large numbers of species.

Coalescent methods that combine estimated gene trees into a species tree, referred to as "summary methods," are able to analyze data sets with both large numbers of species and large numbers of loci; summary methods include STAR (Liu et al. 2009), STEM (Kubatko et al. 2009), MP-EST (Liu et al. 2010), NJst (Liu and Yu 2011), iGLASS (Jewett and Rosenberg 2012), ASTRAL (Mirarab et al. 2014b), ASTRAL-II (Mirarab and Warnow 2015), and a modification of NJst, called ASTRID, designed for greater scalability and ability to handle missing data (Vachaspati and Warnow 2015). Although summary methods all use estimated gene trees as inputs, they differ in their approaches to species tree estimation. For example, NJst computes the average leaf-to-leaf distances from the input gene trees and then applies Neighbor Joining (Saitou and Nei 1987) to the resulting distance matrix. In contrast, ASTRAL and ASTRAL-II solve a constrained optimization problem based on the frequency with which four-leaf trees, called quartet trees, appear in the input gene trees. Many different summary methods are statistically consistent under the MSC model and have excellent accuracy when given a sufficient number of highly accurate gene trees (Liu et al. 2010; Liu and Yu 2011; Mirarab et al. 2014b; Mirarab and Warnow 2015; Vachaspati and Warnow 2015). Comparisons of summary methods and CA-ML on simulated data sets have suggested that CA-ML is typically more accurate than summary methods when ILS is sufficiently low and that summary methods are typically more accurate than CA-ML when ILS is sufficiently high (Leaché and Rannala 2010; Bayzid and Warnow 2013; Patel et al. 2013; Mirarab et al. 2014a; Bayzid et al. 2015; Chou et al. 2015; Mirarab et al. 2016). Hence, summary methods are popular methods for species tree estimation when ILS is high.

However, as discussed in Roch and Warnow (2015), the proofs of statistical consistency for standard summary methods assume true gene trees (without any missing data) based on recombination-free genomic regions, and so the statistically consistency of summary methods has not yet been established for more general conditions. Furthermore, many simulations have shown that gene tree estimation error (GTEE) reduces the accuracy of summary methods (Huang et al. 2010; Bayzid and Warnow 2013; Patel et al. 2013; DeGiorgio and Degnan 2014; Mirarab et al. 2014a; Lanier and Knowles 2015; Mirarab and Warnow 2015; Xi et al. 2015), suggesting that summary methods may be inappropriate methods when gene trees cannot be estimated with high accuracy. This raises potential concerns, since low bootstrap support values for gene trees, which are suggestive of high GTEE, have been reported for empirical data sets. For example, the average bootstrap support for gene trees computed for the Thousand Plant Transcriptome project was ~50% (Wickett et al. 2014); the Avian Phylogenomics Project reported average bootstrap values for its gene trees that ranged from as low as ~25% for the exons to as high as ~50% for the introns, with intermediate values of 40% for the UCEs (Jarvis et al. 2014). Our analyses of gene trees estimated on the UCE data set from Hosner et al. (2016) and the exon data set from Blom et al. (2017) show average bootstrap support values below 30% (Table 1). While low bootstrap support can have many causes, a common explanation is low phylogenetic signal resulting from insufficient sequence lengths or low rates of evolution (Hosner et al. 2016; Blom et al. 2017). Finally, low phylogenetic signal in individual genes is known to result in high GTEE, suggesting that GTEE is likely to be a common problem for some types of phylogenomic data sets, such as UCEs and RADseq data sets.

Missing data is another common challenge to species tree estimation, as many (or perhaps even most) genes will have some degree of missing data if full genomes are to be utilized (see Driskell et al. 2004; Streicher et al. 2016; Xi et al. 2016 for an entry into this literature). Simulations have shown that the accuracy of summary methods can degrade when genes are missing taxa, especially when data sets have limited numbers of genes (Hovmöller et al. 2013; Vachaspati and Warnow 2015; Xi et al. 2016) or when the distribution of missing data is biased (Xi et al. 2016). Because GTEE and missing data affect both the theoretical guarantees and the empirical accuracy of summary methods, some researchers have substantial concern about the validity of estimating species trees using summary methods under many biologically realistic conditions (Gatesy and Springer 2014; Springer and Gatesy 2016)—with some groups deciding not to use summary methods for species tree estimation on their multilocus data sets (e.g., Leaché et al. 2015; de Oca et al. 2017).

Site-based methods, such as SNAPP (Bryant et al. 2012), SVDquartets (Chifman and Kubatko 2014, 2015), SMRT-ML (DeGiorgio and Degnan 2010), and METAL (Dasarathy et al. 2015, 2017), bypass gene tree estimation, and thus, they are expected to be more accurate than summary methods when individual loci have few phylogenetically informative sites. SVDquartets, for example, estimates quartet trees by applying techniques from statistical linear algebra to the concatenated gene

TABLE 1. Empirical statistics for biological data sets

Study	Data set type	Number of taxa	Number of loci	Locus length	Number of informative sites	Mean bootstrap support	SMC versus CA-ML FP	SMC versus CA-ML FN
Blom et al. (2017)	exon	29	1361	384 ± 251	12 ± 10	0.28±0.11	0.45±0.26	0.87±0.09
Hosner et al. (2016)	UCE	91	4817	462 ± 369	37 ± 42	0.28±0.12	0.37±0.25	0.83±0.17
Jarvis et al. (2014)	exon	48	8251	1612 ± 1308	453 ± 418	0.26±0.07	0.22±0.25	0.85±0.08
Jarvis et al. (2014)	intron	48	2516	7654 ± 8539	4315 ± 4654	0.47±0.12	0.20±0.13	0.68±0.09
Jarvis et al. (2014)	UCE	48	3679	2509 ± 164	1062 ± 278	0.40±0.05	0.11±0.11	0.71±0.03
Streicher et al. (2016)	UCE	29	4784	NA	NA	0.39±0.09	0.36±0.26	0.78±0.12

Notes: Means and standard deviations across all loci are reported for the individual locus length, the number of parsimony informative sites, the mean bootstrap support, and the distance between a gene tree and the species tree estimated via concatenation analysis using maximum likelihood (CA-ML). Empirical statistics were computed using the loci alignments, the bootstrap gene trees, and the CA-ML trees provided by these studies (see Supplementary Materials available on Dryad for details). In particular, we used the bootstrap gene trees for each locus to build a greedy consensus tree and a (strict) majority consensus (SMC) tree (i.e., a tree with all branches having greater than 50% support). The greedy consensus tree was used to compute mean bootstrap support by averaging the bootstrap support values across all branches. The SMC tree was used to compute the distance between a gene tree and the species tree computed using CA-ML. The distance between the SMC tree and the estimated species tree is separated into False Positive (FP) rate and False Negative (FN) rate. FN rate is the fraction of branches in the estimated species tree that are missing from the SMC; FP rate is the fraction of branches in the SMC that are missing from the estimated species tree. Streicher et al. (2016) provided concatenated alignments but not alignments for the individual loci in their Dryad repository, and so the mean locus length and the mean number of parsimony informative sites per locus could not be computed for this table.

alignments; then a quartet amalgamation method (e.g., Snir and Rao 2012; Reaz et al. 2014) is used to assemble these estimated quartet trees into a species tree on the full taxon set. A preliminary study (Chou et al. 2015) comparing SVDquartets to CA-ML and two summary methods (ASTRAL-II and NJst) demonstrated that SVDquartets was more accurate than the summary methods under some conditions with very short loci—but was not more accurate than CA-ML under these conditions. For longer loci, Chou et al. (2015) found that ASTRAL-II and NJst were more accurate than SVDquartets, likely due to greater phylogenetic signal across individual loci. Hence, summary methods and CA-ML remain important tools for species trees estimation.

Because poor gene tree quality reduces the accuracy of summary methods, gene filtering (where genes are removed based on predetermined criteria prior to species tree estimation) is an increasingly explored aspect of experimental design. Information-theoretic arguments using the classical Data Processing Inequality (Kinney and Atwal 2014) would seem to suggest that phylogenetic estimation methods should benefit from more data, and hence, gene filtering would not be beneficial. However, the proof that more data are never detrimental to phylogeny estimation has only been established for maximum likelihood (Steel and Székely 2002) under some conditions, including the condition that the data do not violate model assumptions (e.g., gene sequences evolve down the same model tree). Consequently, applying the intuition that more data are always better for species tree estimation can be problematic; for example, summary methods that are statistically inconsistent under conditions without missing data *can* be statistically consistent when enough data are missing (see Appendix).

Filtering data (both sites and genes) has a long history in phylogenetics (see Wiens and Morrill 2011; Chen et al. 2015; Streicher et al. 2016 for an entry into this literature). Many of these prior gene filtering studies have been restricted to concatenation analyses (often based on maximum likelihood) and/or data sets simulated without gene tree heterogeneity (e.g., Cho et al. 2011; Wiens and Morrill 2011; Salichos and Rokas 2013; Betancur-R et al. 2014; Dornburg et al. 2014; Jiang et al. 2014; Salichos et al. 2014; Streicher and Wiens 2016; Dornburg et al. 2017), and so are not directly applicable to understanding the effect of gene filtering on coalescent species tree estimation methods and on summary methods in particular.

Although some studies have examined the impact of gene filtering strategies on coalescent methods (e.g., Chen et al. 2015; Liu et al. 2015b; Xi et al. 2015; Hosner et al. 2016; Huang and Knowles 2016; Meiklejohn et al. 2016; Simmons et al. 2016; Streicher et al. 2016; Blom et al. 2017; Longo et al. 2017; Lanier et al. 2014), many of these studies used empirical data sets, and so the true species trees were unknown. Therefore, species tree accuracy on empirical data sets was assessed using various criteria, including overall bootstrap support, similarity to another species tree estimated (typically using concatenation analysis) on the same data set, and recovery and bootstrap support of well-established clades. Some studies have also examined stability by comparing species trees estimated on subsets of the loci to the species tree estimated on the full set of loci. These empirical studies have come to contradictory conclusions: some found filtering to be beneficial while others found filtering to be detrimental—making it difficult to draw any general guidelines from these studies. One difficulty in interpreting these results is that simulations have shown species tree estimation methods can produce highly supported false positive branches under some model conditions (see **?** for an example with CA-ML in the presence of high ILS, and see Bayzid et al. 2015 for examples with summary methods in the presence of high GTEE). Therefore, high similarity to an estimated species tree or high bootstrap

support may not be reliable indicators of topological accuracy.

To the best of our knowledge, only three prior simulation studies (Lanier et al. 2014; Liu et al. 2015b; Huang and Knowles 2016) have directly or indirectly examined the impact of gene filtering on coalescent methods. Huang and Knowles (2016) filtered genes with missing data when using the shallowest divergence method (Takahata, 1989) on datasets with eight species; Liu et al. (2015b) added genes with lower bootstrap support when using MP-EST on datasets with six species; Lanier et al. (2014) added low-variation genes when using STEM on datasets with eight species. None of these studies found filtering to be beneficial to coalescent methods. However, many other coalescent methods are now in active use, and the effect of gene filtering likely depends on the method itself as well as the model condition, including the number of species, the number of genes, the level of ILS, the degree of phylogenetic signal or GTEE, and the amount of deviation from the strict molecular clock (Liu et al. 2009). Hence, a thorough evaluation of gene filtering is needed, especially to examine its effect on the relative performance of some of the leading species tree estimation methods.

We present the results of a simulation study examining the impact of phylogenetic signal (of individual loci), missing data, and gene filtering strategies on species tree estimation methods, all in the context of gene tree heterogeneity due to ILS. In particular, we used 26-taxon, 1000-gene data sets simulated under a wide range of model conditions to evaluate several of the leading coalescent species tree estimation methods (ASTRAL-II, ASTRID, MP-EST, and SVDquartets) as well as unpartitioned CA-ML using RAxML (Stamatakis 2014). Summary methods (ASTRAL-II, ASTRID, and MP-EST) were more accurate than CA-ML, provided that the level of ILS was sufficiently high and the level of GTEE was sufficiently low. When GTEE was sufficiently high, SVDquartets was more accurate than the summary methods, but otherwise it was often among the least accurate methods. CA-ML was competitive with (and often outperformed) the other methods under many conditions, including when the levels of ILS and GTEE were both extremely high. In general, the relative performance of different species tree estimation methods was unaffected by the use of gene filtering based on either GTEE or missing data. SVDquartets and CA-ML did not benefit from either type of filtering, and filtering based on missing data generally reduced the accuracy of all methods examined. Filtering genes based on GTEE typically improved the accuracy of summary methods when the level of ILS was sufficiently low but otherwise tended to reduce accuracy. Exceptions to this trend occurred when the level of GTEE was extremely high, in which case filtering based on GTEE often improved summary methods. However, the exceptions occurred for model conditions with only a few replicates: 2 replicates with low/moderate ILS, 5 replicates with high ILS, and 17 replicates with very high ILS.

Materials and Methods

Overview

Our study evaluated three summary methods (ASTRAL-II, ASTRID, and MP-EST), a site-based coalescent method (SVDquartets using PAUP*; Swofford 2002, 2016), and unpartitioned CA-ML (using RAxML) on a collection of data sets originally simulated by Mirarab and Warnow (2015). We modified these data sets for this study to produce a range of model conditions from the relatively easy (i.e., moderate GTEE and low/moderate levels of ILS) to the very challenging (i.e., high levels of ILS and GTEE as well as missing data). Genes were removed from these data sets (based on GTEE or amount of missing data) to explore the impact of gene filtering on species tree estimation methods. All distances between trees were measured with the normalized Robinson–Foulds (RF) distance (Robinson and Foulds 1981) using Dendropy (Sukumaran and Holder 2010).

Simulated Data Sets

We give a brief overview of the simulation protocol used by Mirarab and Warnow (2015), describe our modifications to their data sets, and report empirical statistics about the simulated data sets that we explored.

Model species trees and gene trees.—Mirarab and Warnow (2015) used SimPhy (Mallo et al. 2016) to simulate 200-taxon species trees and gene trees under three levels of ILS with deep or recent speciation events. Because MP-EST is computationally intensive on data sets with 50 species (Bayzid et al. 2014; Mirarab and Warnow 2015), we restricted the data sets to 26 species (the outgroup taxon and 25 randomly selected taxa) and used only 20 (out of the original 50) replicates in our study.

After the data sets were restricted to 26 species, we computed several empirical statistics for each model condition that reflect the level of ILS: the average distance (AD) between the true species tree and the true gene trees, the percentage of replicates with species trees in the anomaly zone (Degnan and Rosenberg 2006), and the number of different tree topologies that appeared in the set of 1000 true gene trees. The mean AD ($\pm$standard deviation) across replicates was $12\pm2\%$ for the *low/moderate* ILS condition, $41\pm6\%$ for the *high* ILS condition, and $75\pm1\%$ for the *very high* ILS condition. Under the low/moderate ILS condition, 20% of the replicates with speciation towards the leaves and 60% of the replicates with speciation towards the root were in the anomaly zone. The number of distinct gene tree topologies was also high for the low/moderate ILS condition: there were 344–442 different topologies (across all 1000 true gene trees) for replicates with recent speciation events and 509–824 different topologies (across all 1000 true gene trees) for replicates with deep speciation events. Hence, while the AD was only 12%, the gene tree heterogeneity in this "low/moderate" ILS

condition was still substantial; this condition represents cases where the species tree is a mixture of short and long edges, with perhaps a few rapid radiations creating the anomaly zone. In the other model conditions (high and extremely high ILS), 100% of the replicates were in the anomaly zone, and each replicate had 999 or 1000 different gene tree topologies. These high and extremely high levels of ILS are representative of a single clade that has undergone a rapid radiation, so that nearly every edge is short. Thus, all model conditions explored have a high incidence of species trees in the anomaly zone but differ in the fraction of branches in the species tree that are short enough for coalescence to be likely to occur. We evaluated species tree estimation methods on this entire range of data sets in order to examine conditions that were similar to those from prior simulation studies, including those that mainly focused on estimating species trees with mostly short internal branches (e.g., Liu et al. 2010) and others that focused on estimating species trees with a mixture of short and long branches (e.g., Mirarab and Warnow 2015). Both types of species trees occur in phylogenomic data sets and so are relevant to systematics.

Gene sequence data.—Mirarab and Warnow (2015) used Indelible (Fletcher and Yang 2009) to simulate the evolution of sequences under the Generalized Time Reversible or GTR (Tavaré 1986) model (with gamma-distributed rates across sites) down gene trees with branch lengths deviating from a strict molecular clock. These sequences had variable lengths (300–1500 sites) and no insertions/deletions. We modified these gene sequence alignments to have shorter lengths (100 sites) and/or missing data. The data sets with sequences truncated to the first 100 sites were intended to produce conditions with fewer phylogenetically informative sites and higher GTEE; such conditions may be characteristic of data sets where the mean bootstrap support of gene trees is low (e.g., Jarvis et al. 2014; Wickett et al. 2014; Hosner et al. 2016; Blom et al. 2017) or data sets where gene sequences are shortened to avoid recombination (e.g., Hobolth et al. 2011). In the data set from Hosner et al. (2016), the degree of missing data varied across genes but was uncorrelated with evolutionary rate (Supplementary Table S1 and Fig. S1 available on Dryad at http://dx.doi.org/10.5061/dryad.km24v), and thus, we deleted species from gene sequence alignments using a protocol (see Supplementary Materials available on Dryad for details) designed to produce data sets with missing data biased towards random subsets of genes. The resulting data sets had the following pattern of missing data: 250 genes missing between 13 and 19 sequences (i.e., 50–73%), 250 genes missing between 7 and 12 sequences (i.e., 27–46%), 250 missing between 3 and 6 sequences (i.e., 12–23%), 150 genes missing 2 sequences (i.e., 8%), 50 genes missing 1 sequence (i.e., 4%), and 50 genes missing no sequences (Supplementary Table S2 available on Dryad). The total amount of missing data was approximately 30% for all data sets.

Gene Tree Estimation

We estimated maximum likelihood gene trees using RAxML v8.2.8 with a single tree search under the GTRGAMMA model of evolution (see Supplementary Materials available on Dryad for details). The mean GTEE of a replicate (with 1000 genes) is the normalized RF distance between the true and estimated gene trees, averaged across all genes. Replicates were partitioned based on their mean GTEE. Replicates in the full length sequence data sets (300–1500 sites) were partitioned into *low/moderate* GTEE (i.e., mean GTEE between 0% and 20%) and *moderate/high* GTEE (i.e., mean GTEE between 20% and 50%). The mean GTEE averaged across these full-length replicates (±standard deviation) was 16±2% and 35±8% for *low/moderate* and *moderate/high* GTEE, respectively (Supplementary Tables S3 and S4 available on Dryad). Replicates in the truncated sequence data sets (100 sites) had higher GTEE, and so were partitioned into *very high* GTEE (i.e., mean GTEE within 50–80%) and *extremely high* GTEE (i.e., mean GTEE within 80–100%). The mean GTEE averaged across these truncated sequence data set replicates (± standard deviation) was 69±8% and 86±5% for *very high* and *extremely high* GTEE, respectively (Supplementary Tables S5 and S6 available on Dryad).

Gene Filtering Experiments

We evaluated the impact of gene filtering by removing 0%, 25%, 50%, 75%, 90%, and 95% of the genes, thus producing data sets that varied in the number of genes retained for species tree inference. To filter genes by GTEE, gene trees were sorted based on GTEE, and then 25%, 50%, 75%, 90%, and 95% of genes with the highest GTEE were removed prior to species tree estimation. To filter genes by missing data, gene trees were sorted based on the amount of missing data (i.e., the fraction of species deleted from the gene sequence alignment) and genes missing at least 50%, 25%, 10%, 5% and 1% of species were removed prior to species tree estimation. Given the protocol for producing data sets with missing data, these thresholds for filtering resulted in the same number of genes being removed for each of the two filtering experiments, making them comparable.

Species Tree Estimation

We explored the performance of five methods for species tree estimation, including, three summary methods (ASTRAL-II (i.e., ASTRAL v4.10.5), ASTRID v1.1, and MP-EST v1.5), a site-based method SVDquartets using PAUP* v4a152, (Swofford 2016), and unpartitioned CA-ML (RAxML v8.2.8) under the GTRGAMMA model of evolution (Supplementary Table S9 available on Dryad). Summary methods were run on the best found maximum likelihood gene trees (rather than on the bootstrap gene trees), which has been shown to improve species tree accuracy for sufficiently large numbers of loci (Mirarab et al. 2016). ASTRAL-II and ASTRID

were run in default mode on unrooted gene trees. Since MP-EST requires rooted gene trees, estimated gene trees were rooted at the outgroup when available and otherwise rooted at the midpoint of the longest leaf-to-leaf path using Dendropy v4.1.0 (Sukumaran and Holder 2010). For MP-EST, the best pseudolikelihood scoring species tree was taken from 10 independent runs. We used the local branch support technique from ASTRAL-II [rather than the more common approach of multilocus bootstrapping (MLBS)] to compute branch support for the species trees estimated using ASTRAL-II. The local branch support implemented in ASTRAL-II has been shown to be a better predictor of topological accuracy than support calculated using MLBS (Sayyari and Mirarab 2016).

Results

We present the results of four experiments. The first and second experiments evaluated the impact of ILS and phylogenetic signal per gene (which impacts GTEE) on five species tree estimation methods (ASTRID, ASTRAL-II, MP-EST, SVDquartets, and CA-ML) with and without missing data, respectively. The third and fourth experiments evaluated the impact of gene filtering based on GTEE and missing data, respectively, on the five methods.

Effects of Incomplete Lineage Sorting and Phylogenetic Signal

In this experiment, species trees were estimated on data sets without missing data and without gene filtering to show how ILS and phylogenetic signal per gene affect method accuracy. The phylogenetic signal per gene was reduced by truncating gene sequence alignments from their original lengths (300–1500 sites) down to 100 sites. Because this modification decreases the number of sites per gene as well as the total number of sites in the concatenated alignment, the resulting decrease in phylogenetic signal has the potential to impact summary methods (ASTRAL, ASTRID, and MP-EST) as well as site-based methods (SVDquartets and CA-ML). We quantify the average phylogenetic signal per gene by reporting the mean GTEE, noting that high mean GTEE corresponds to low phylogenetic signal per gene and conversely low mean GTEE corresponds to moderate to high phylogenetic signal per gene.

We found that species tree error increased for all methods as levels of ILS and/or GTEE increased (Fig. 1) and that the relative performance between methods depended on both ILS and GTEE. For the low/moderate ILS condition (12% AD), CA-ML was the most accurate method for all levels of GTEE (Fig. 1a). All five methods had good accuracy with mean species tree error below 7% when mean GTEE was less than 50%. When mean GTEE was between 80% and 85%, the mean species tree error rate for CA-ML was 5%, the mean error rates for summary methods ranged from 16% (ASTRAL-II) to 19% (MP-EST), and the mean error rate for SVDquartets was 20%. Although the differences between methods were noteworthy, this model condition had only five replicates.

For the high ILS condition (41% AD), the relative performance between methods changed dramatically with GTEE (Fig. 1b). The three summary methods outperformed SVDquartets and CA-ML when GTEE was low to moderate (i.e., mean GTEE $< 50\%$), but SVDquartets and CA-ML outperformed the summary methods when GTEE was extremely high (Fig. 1b,c). CA-ML produced more accurate species trees than SVDquartets under the high ILS condition except for the highest GTEE condition (a model condition with only four replicates), where they had similar accuracy.

For the very high ILS condition (75% AD), results were similar but more pronounced than those observed for the high ILS condition (Fig. 1c). Under lower levels of GTEE, CA-ML and SVDquartets were distinctly worse than the summary methods but still provided reasonable accuracy. All methods decreased in accuracy as the level of GTEE increased, but the accuracy of SVDquartets and CA-ML decreased more gradually than that of the summary methods. When mean GTEE was at least 90% (a model condition with only four replicates), the differences between methods were dramatic: the mean species tree error rates for summary methods were all greater than 90%, while the mean error rates for CA-ML and SVDquartets were much lower at 30% and 37%, respectively.

Thus, both ILS and GTEE had strong effects on the absolute and the relative performance of methods. The summary methods typically dominated or else matched CA-ML when GTEE was sufficiently low but were less accurate when GTEE was high. SVDquartets was typically less accurate than the other methods but was dramatically more accurate than the summary methods under the most difficult conditions (very high GTEE and very high ILS). Finally, CA-ML nearly always outperformed SVDquartets and also outperformed summary methods under the lowest level of ILS as well as under higher levels of GTEE.

Effects of Missing Data

The relative performance of different methods was typically similar for data sets with and without missing data (Supplementary Fig. S4 available on Dryad). Although deleting species from gene sequence alignments nearly always reduced accuracy (Fig. 2), this reduction tended to be fairly small (typically below 5%). Methods differed somewhat in their response to missing data. ASTRAL-II and ASTRID were quite robust to missing data, with mean species tree estimation error never increasing by more than 6% (and most increases in error were much smaller). Missing data resulted in larger increases in species tree estimation error for the other methods, especially under very high levels of ILS or GTEE. Interestingly, the accuracy of some summary

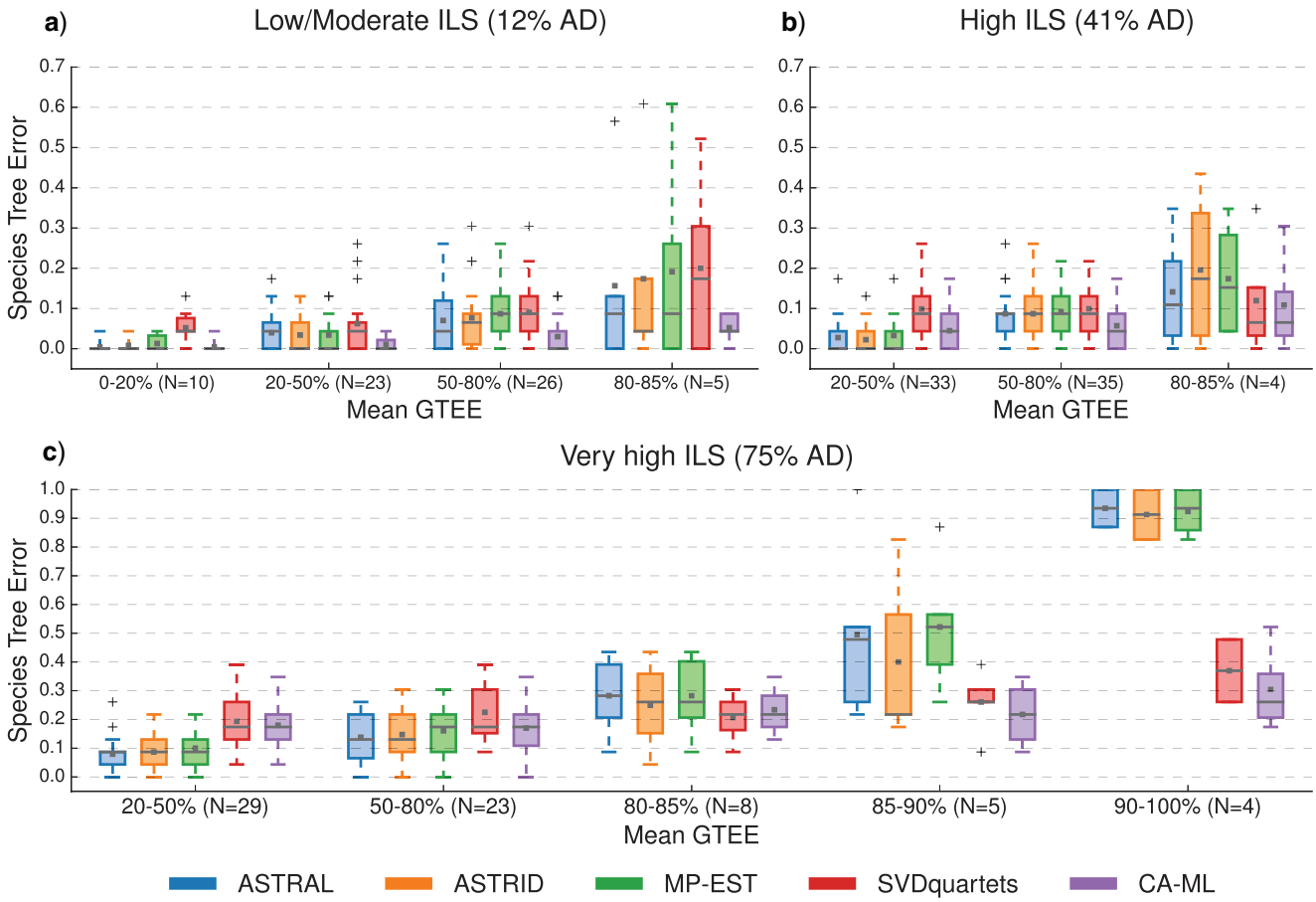

FIGURE 1. The impact of gene tree estimation error (GTEE) and incomplete lineage sorting (ILS) on species tree error is shown for five methods: ASTRAL-II (blue), ASTRID (orange), MP-EST (green), SVDquartets (red), and unpartitioned concatenation analysis using maximum likelihood (CA-ML), specifically RAxML (purple). Species tree error is the normalized Robinson–Foulds (RF) distance between the true and estimated species trees. Subplots a, b, and c show three levels of increasing ILS, where AD is the normalized RF distance between the true species and true gene trees averaged across all genes. The mean GTEE range and the number of replicates (N) for that model condition are given on the x-axis. Means and medians are denoted by the gray dot and bar, respectively. Box plots are defined by quartiles, e.g., boxes extend from the first to the third quartiles. Greater levels of ILS and/or GTEE increased species tree error rates for all methods, and the relative performance of methods depended on both ILS and GTEE. Under low to moderate ILS, CA-ML tended to have better accuracy than the coalescent methods. Under higher levels of ILS, summary methods were typically more accurate than CA-ML and SVDquartets except for conditions with high GTEE.

methods *improved* on some data sets with incomplete genes (Fig. 2c) when the level of ILS was very high and the mean GTEE was greater than 85%; however, this model condition had only nine replicates.

Effects of Filtering Based on Gene Tree Estimation Error

The impact of filtering genes based on GTEE depended on the method and also on the levels of ILS and GTEE; for example, gene filtering based on GTEE made summary methods more accurate when the level of ILS was sufficiently low (12% AD) and the level of GTEE was moderate to very high, provided that the number of retained genes was not too small (Fig. 3a,b). For example, the removal of ~75% of the genes based on GTEE resulted in ~2–3% improvement in species tree accuracy for the three summary methods under the lowest level of ILS. Under higher ILS conditions, filtering based on GTEE was at best neutral and typically increased species tree error (Fig. 3c–f). However, when GTEE was extremely high (mean GTEE > 85%), filtering based on GTEE improved the summary methods for all ILS levels (Table 2, Supplementary Tables S10 and S11 available on Dryad); the replicates with extremely high GTEE were limited to 2 replicates with low/moderate ILS, 5 replicates with high ILS, and 17 replicates with very high ILS. SVDquartets and CA-ML decreased in topological accuracy when the sequence alignments from genes with high GTEE were removed (Supplementary Fig. S5, Tables S12 and S13 available on Dryad). Although summary methods could become more accurate with gene filtering based on GTEE for the low/moderate ILS condition we examined (12% AD), CA-ML was still the most accurate method under this ILS condition. Finally, gene filtering based on GTEE had minimal impact on ASTRAL-II's local branch support but typically decreased the mean support of the true branches recovered by ASTRAL-II and increased the number of true branches with support

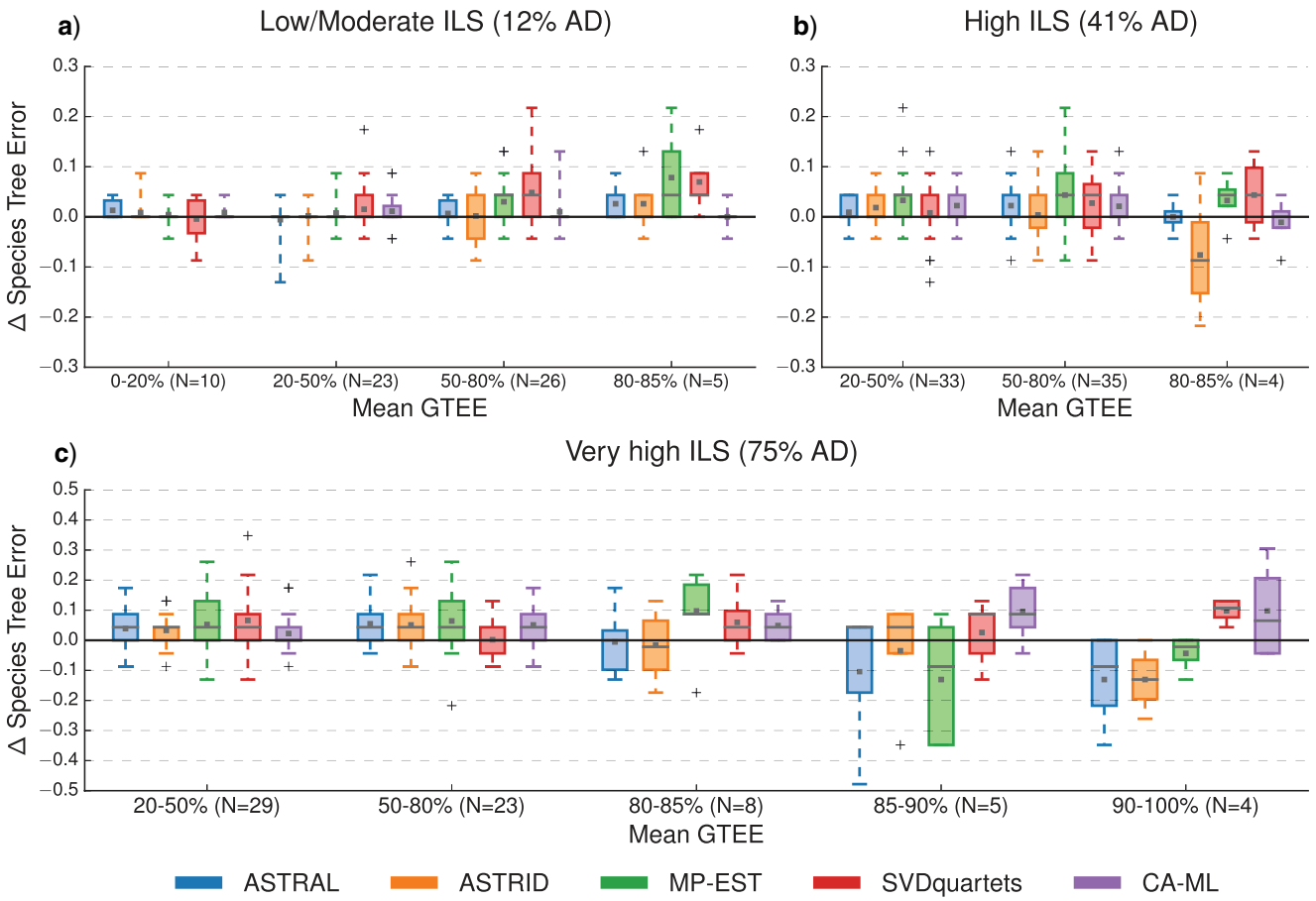

FIGURE 2. Differences in species tree error between data sets with no missing data and data sets with approximately 30% missing data are shown for five methods: ASTRAL-II (blue), ASTRID (orange), MP-EST (green), SVDquartets (red), and unpartitioned concatenation analysis using maximum likelihood (CA-ML), specifically RAxML (purple). Positive values indicate increases in error, whereas negative values indicate reductions in error. Subplots a, b, and c show three levels of increasing incomplete lineage sorting (ILS), where AD is the normalized RF distance between the true species and true gene trees averaged across all genes. The mean gene tree estimation error (GTEE) range and the number of replicates (N) for that model condition are given on the x-axis. Means and medians are denoted by the gray dot and bar, respectively. Box plots are defined by quartiles, e.g., boxes extend from the first to the third quartiles.

less than 75% (Supplementary Figs. S6 and S7 available on Dryad).

Filtering genes based on GTEE reduced the mean GTEE among the set of retained genes as compared to the original set of genes (Supplementary Fig. S8 available on Dryad). For example, when GTEE was moderate/high, the mean GTEE of the unfiltered data sets was 35–40%; after the removal of 75% of the genes, the mean GTEE was ~20%, corresponding to a 15–20% reduction in mean GTEE (Supplementary Fig. S8a,c,e available on Dryad). Despite the substantial reduction in mean GTEE, species tree error tended to increase except for the lowest ILS condition.

Effects of Filtering Based on Missing Data

Filtering genes based on missing data typically reduced the accuracy of all methods, but the extent of this reduction depended on the levels of ILS and GTEE as well as the number of genes remaining after filtering (Fig. 4, Supplementary Tables S14–S18 available on Dryad). For all the experiments shown, deleting only half the genes (and so retaining 500 genes of the original 1000) had a negligible impact on accuracy. When more genes were deleted, the error rates increased for all methods under all conditions examined. Under the easiest conditions (i.e., for low/moderate ILS and moderate GTEE), the impact of filtering 95% of the genes was relatively small for all methods (i.e., species tree error rates increased by approximately 5% (Fig. 4a). At the other extreme, when the levels of ILS and GTEE were both very high (Fig. 4f), filtering 95% of the genes (corresponding to removing all genes with missing data) decreased accuracy by approximately 25% for all methods. Finally, filtering genes based on missing data could decrease the mean branch support of the true branches recovered by ASTRAL-II, resulting in a higher frequency of true branches with support less than 75% (Supplementary Figs. S9 and S10 available on Dryad).

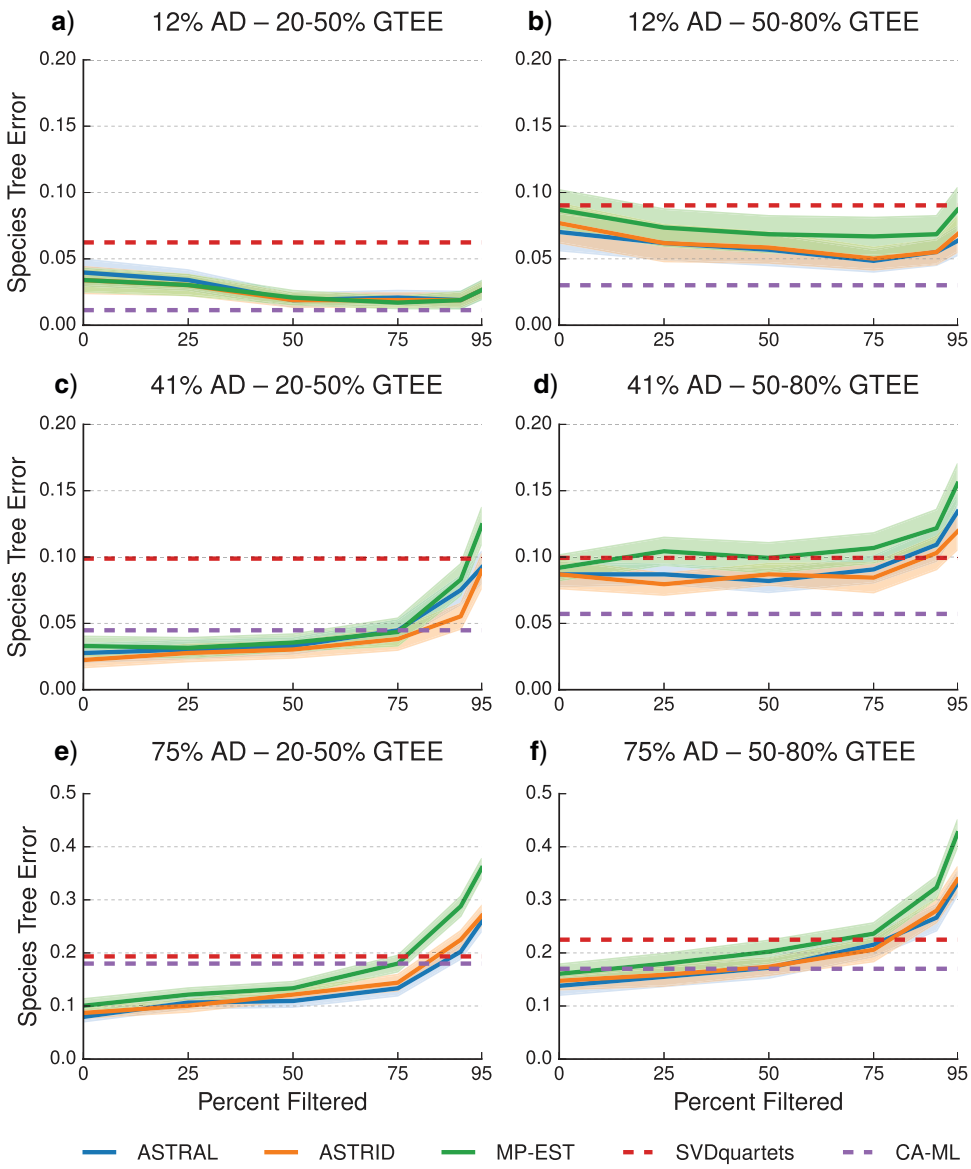

FIGURE 3. The impact of filtering genes by gene tree estimation error (GTEE) on species tree error is shown for three gene tree summary methods: ASTRAL-II (solid blue), ASTRID (solid orange), and MP-EST (solid green). Genes were filtered by removing the top 25%, 50%, 75%, 90%, and 95% of genes with the highest GTEE. SVDquartets (dashed red) and unpartitioned concatenation analysis using maximum likelihood (CA-ML), specifically RAxML (dashed purple), are shown without filtering. Species tree error is the normalized Robinson–Foulds (RF) distance between the true and estimated species trees. Lines indicate the mean across all replicates, and filled regions indicate the standard error. Rows show three levels of increasing incomplete lineage sorting (ILS), where the AD is defined as the normalized RF distance between the true species tree and true gene trees averaged across all genes. Columns show two levels of GTEE. When ILS was sufficiently low, gene filtering (up to 75% of genes) increased the accuracy of gene tree summary methods (a–b). When ILS was high to very high, gene filtering had little impact on the accuracy of gene tree summary methods or else reduced summary method accuracy (d–f).

DISCUSSION

The data sets used in this study cover a broad range of model conditions with varying levels of ILS and GTEE, both with and without missing data. Nearly all replicates were in the anomaly zone; therefore, this study considered a very wide range of model conditions where coalescent species tree estimation methods are relevant. However, this study was constrained to five methods and to data sets with 26 species and 1000 genes (unless gene filtering was performed). Therefore, the trends reported in this study may not generalize to other methods or to data sets with much smaller or much larger numbers of species and/or genes.

Accuracy of Species Tree Estimation Methods

Effects of incomplete lineage sorting and phylogenetic signal.—In this study, both ILS and GTEE affected the relative accuracy of species tree estimation methods. CA-ML had the best accuracy of all methods under

TABLE 2. Proportions of replicates for which filtering based on GTEE increased or decreased the accuracy of ASTRAL-II

Mean GTEE	Number of replicates	Number of informative sites	Proportion of replicates affected by filtering (increased/decreased accuracy) when the following percentages of genes were removed				
			25%	50%	75%	90%	95%
Low/moderate ILS (12% AD)							
0–20%	10	596 ± 224	0.00/0.00	0.00/0.00	0.00/**0.10**	0.00/0.00	0.00/**0.20**
20–50%	23	464 ± 276	**0.09**/0.04	**0.26**/0.00	**0.35**/0.09	**0.30**/0.04	**0.26**/0.13
50–80%	26	63 ± 14	**0.23**/0.04	**0.23**/0.04	**0.31**/0.12	**0.35**/0.23	**0.35**/0.31
80–85%	5	40 ± 28	0.40/**0.60**	0.40/0.40	0.40/**0.60**	0.20/**0.60**	0.20/**0.80**
85–100%	2	5 ± 3	**0.50**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
High ILS (41% AD)							
0–20%	2	487 ± 9	0.00/**0.50**	0.00/**0.50**	0.00/**1.00**	0.00/0.00	0.00/**0.50**
20–50%	33	349 ± 208	0.03/**0.09**	0.15/**0.27**	0.15/**0.33**	0.03/**0.55**	0.03/**0.70**
50–80%	35	42 ± 17	0.11/**0.17**	**0.26**/0.17	0.26/**0.34**	0.14/**0.49**	0.17/**0.60**
80–85%	4	22 ± 10	**0.75**/0.00	**0.50**/0.00	**0.25**/0.00	0.25/0.25	0.25/**0.50**
85–100%	1	6 ± 0	0.00/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00	**1.00**/0.00
Very high ILS (75% AD)							
0–20	0	NA	NA	NA	NA	NA	NA
20–50%	29	213 ± 123	0.07/**0.45**	0.07/**0.59**	0.14/**0.59**	0.00/**0.93**	0.00/**1.00**
50–80%	23	30 ± 9	0.22/**0.35**	0.17/**0.61**	0.09/**0.70**	0.04/**0.87**	0.00/**1.00**
80–85%	8	17 ± 10	**0.75**/0.00	**0.62**/0.25	**0.62**/0.25	**0.50**/0.38	0.00/**0.75**
85–100%	9	6 ± 6	**0.56**/0.00	**0.67**/0.00	**0.78**/0.00	**0.78**/0.11	**0.78**/0.22

Notes: The proportion of replicates for which filtering based on gene tree estimation error (GTEE) increased and decreased the accuracy of ASTRAL-II is given on the left and right of the forward slash, respectively. The larger of these two values is in bold. If these two fractions do not sum to one, then the remainder is the proportion of replicates for which filtering did not impact accuracy. The number of replicates as well as the mean (± standard deviation) of parsimony informative sites is given for each model condition, specified by the level of incomplete lineage sorting (ILS) and the range of mean GTEE.

the low/moderate ILS condition, even though many of the replicates were shown to be in the anomaly zone. Interestingly, the improvement of CA-ML over coalescent methods occurred under some conditions with very high ILS, specifically when mean GTEE was greater than 85% (a model condition with only nine replicates). The differences in accuracy between CA-ML and the summary methods were usually small except when GTEE was sufficiently high. Summary methods were typically more accurate than CA-ML when the level of ILS was not too low and GTEE was not too high (mean <50%). Hence, summary methods performed close to best (and sometimes best) under many conditions—but always provided that GTEE was not too high.

Prior simulation studies evaluating the performance of species tree estimation methods on multilocus data sets without missing data found similar trends with respect to the relative accuracy of ASTRAL-II, ASTRID, MP-EST, and CA-ML (Mirarab et al. 2014a,b; Bayzid et al. 2015; Chou et al. 2015; Davidson et al. 2015; Vachaspati and Warnow 2015; Mirarab et al. 2016); these trends are also consistent with earlier simulation studies evaluating other coalescent methods (Leaché and Rannala 2010; Liu and Yu 2011; Bayzid and Warnow 2013; Patel et al. 2013; Liu et al. 2015a). The improvement of CA-ML over summary methods has been noted before for high levels of ILS (e.g., Mirarab and Warnow 2015) but not (to our knowledge) for conditions with very high ILS (75% AD), as was observed in this study. Finally, the good performance of ASTRID in this study is consistent with prior simulation studies comparing ASTRID or NJst to other species tree estimation methods (Liu et al. 2015a; Vachaspati and Warnow 2015).

In general, SVDquartets was not among the best methods. Although it was dramatically more accurate than the summary methods under the highest ILS and GTEE condition (a model condition with only nine replicates), CA-ML was at least as accurate as SVDquartets (and usually more accurate), even under the highest levels of ILS, on which SVDquartets would be expected to have an advantage.

Relatively little is known about the performance of SVDquartets. A prior simulation study evaluating SVDquartets in comparison to other species tree estimation methods (Chou et al. 2015) also found that SVDquartets was less accurate than CA-ML and was typically less accurate than the summary methods examined. Our results agree with the overall trends in Chou et al. (2015), except that Chou et al. (2015) observed that SVDquartets was (slightly) more accurate than ASTRAL-II in a few cases. It is likely that some differences in model conditions between the two studies produced this small change in the relative performance between SVDquartets and summary methods.

Effects of missing data.—In this study, missing data typically resulted in a slight reduction in accuracy, a trend that has also been noted in prior studies (Hovmöller et al. 2013; Vachaspati and Warnow 2015; Xi et al. 2016). The few cases (Fig. 2c) where missing data improved the accuracy of some summary methods are worth examining more carefully. When missing data

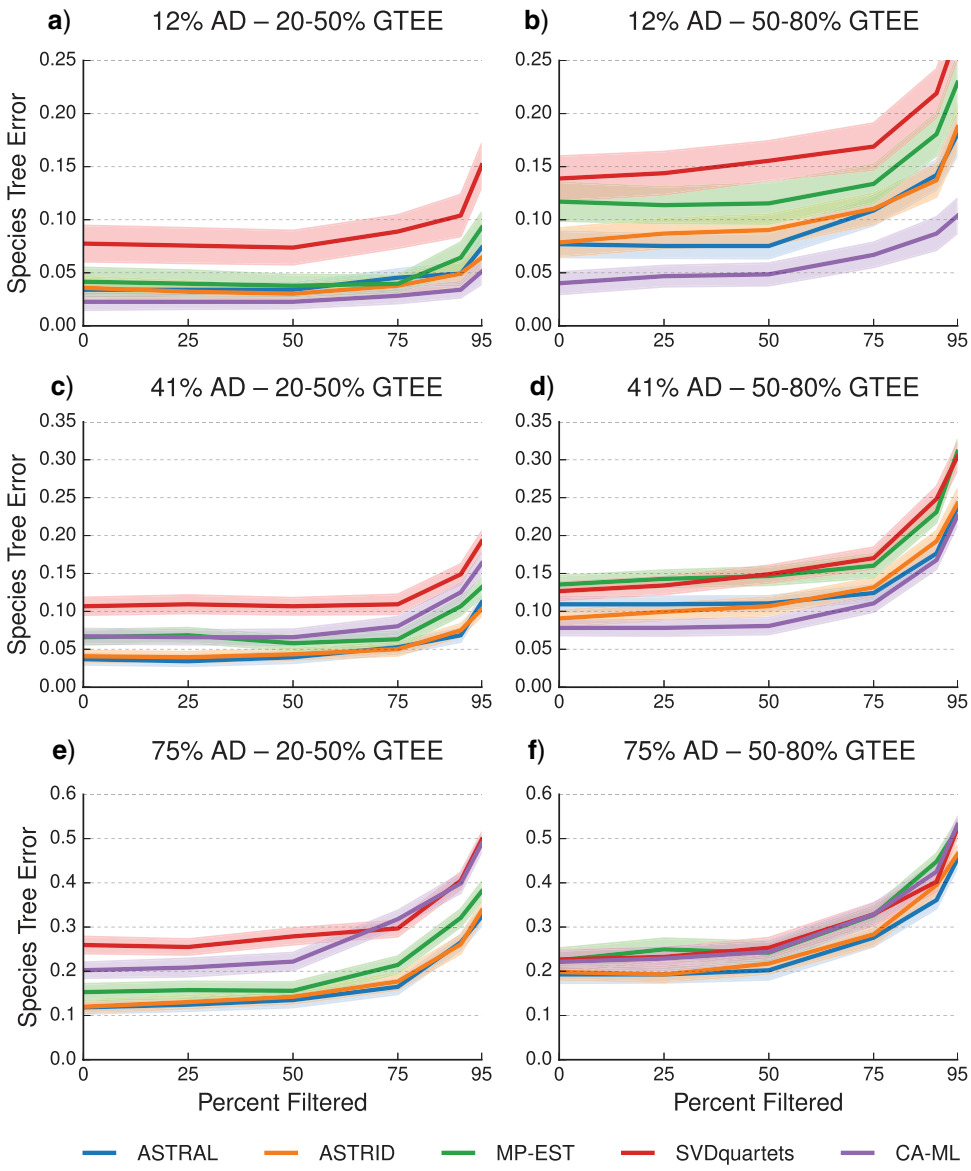

FIGURE 4. The impact of filtering genes by missing data on species tree error is shown for five methods: ASTRAL-II (blue), ASTRID (orange), MP-EST (green), SVDquartets (red), and unpartitioned concatenation analysis using maximum likelihood (CA-ML), specifically RAxML (purple). Genes were filtered by removing the top 25%, 50%, 75%, 90%, and 95% of genes with the highest fractions of missing data; this resulted in removing genes missing at least 50%, 25%, 10%, 5%, and 1% of taxa. Species tree error is the normalized Robinson–Foulds (RF) distance between the true and estimated species trees. Lines indicate the mean across all replicates, and filled regions indicate the standard error. Rows show three levels of increasing incomplete lineage sorting (ILS), where the AD is defined as the normalized RF distance between the true species tree and true gene trees averaged across all genes. Columns show two levels of gene tree estimation error (GTEE). Gene filtering based on missing data was at best neutral but often reduced the accuracy of species tree estimation methods.

improved accuracy, deleting species from gene sequence alignments typically resulted in more accurate gene trees; this was especially true for model conditions characterized by very high ILS (75% AD) and extremely high GTEE (>85%). For this model condition (with only 9 replicates), missing data reduced GTEE as follows: on average 835 genes (out of the 950 genes with missing data) had lower GTEE after deleting taxa, with an average reduction in error of 3.6%. One would expect many of these true gene trees to have short branches (because species trees with very high ILS have very short branches), and hence, a possible explanation for our observation is that the random deletion of taxa increased some branch lengths in the gene trees, making them easier to estimate.

Effects of Gene Filtering

The impact of gene filtering on the accuracy of species tree estimation depended on the filtering criterion, the method, and the model condition; however, filtering based on either GTEE or missing data always had a

negative affect on all species tree estimation methods when the number of retained genes became too small.

Filtering based on gene tree estimation error.—Filtering based on GTEE typically increased the accuracy of summary methods under the low/moderate ILS condition but tended to reduce accuracy under higher levels of ILS. Regardless of the ILS condition, the average GTEE of the retained genes was substantially reduced by filtering based on GTEE (Supplementary Fig. S8 available on Dryad); thus, even when filtering improved the quality of the input gene trees, this did not always offset the negative impact of reducing the total amount of data via filtering.

Differences due to ILS level could be explained as follows. When ILS is sufficiently low, a few highly accurate gene trees are sufficient to estimate the true species tree (e.g., one perfectly estimated gene tree is identical to the species tree in the no-ILS condition). However, a large sample of gene trees is necessary to accurately estimate the species tree under higher levels of ILS. Hence, filtering genes will be detrimental unless a sufficiently large number of genes is retained after filtering—and this number of genes will vary with the level of ILS. This analysis is consistent with recent mathematical results showing that the number of true gene trees required for ASTRAL to recover the true species tree with high probability grows in proportion to the shortest branch in the true species tree and thus ILS (Shekhar et al. 2017). Based on this explanation, one would expect gene filtering to have a particularly negative impact under high ILS conditions. Yet when the level of GTEE was extremely high (mean > 80%), filtering based on GTEE could improve summary methods, including under the higher levels of ILS. However, the number of replicates with extremely high GTEE was relatively small (only 5 replicates with high ILS and 17 replicates with very high ILS), and further investigation of this condition would be helpful.

Finally, filtering based on GTEE affected CA-ML and SVDquartets quite differently: filtering decreased the accuracy of CA-ML and SVDquartets, even when GTEE was very high. Equivalently, CA-ML and SVDquartets benefited from the additional loci, even when the added loci had very low signal. In bypassing gene tree estimation, CA-ML and SVDquartets are more robust to the quality of the loci, and may reliably improve with additional data regardless of the amount of phylogenetic signal per locus.

To the best of our knowledge, only two prior simulation studies (Lanier et al. 2014; Liu et al. 2015b) have examined how filtering genes based on GTEE or its proxies affects coalescent methods. Liu et al. (2015b) performed a simulation on 6-taxon model trees with high ILS (50% AD, Liang Liu, personal communication) in which there were two types of genes: "strong genes" (which had 1000 sites) and "weak genes" (which had 100 sites). Gene trees computed using maximum likelihood on the weak genes had average bootstrap support below 40%, while maximum likelihood trees computed on the strong genes had average bootstrap support greater than 80%, suggesting that there was low GTEE for the strong genes and moderate/high GTEE for the weak genes. Species tree were inferred from sets of the estimated gene trees using MP-EST. Liu et al. (2015b) observed that adding 60 weak genes (in increments of 10) to a set of 30 strong genes increased the fraction of replicates in which the true species tree was recovered from 33% to 50%; however, the improvement was not monotone (i.e., as weak genes were added the accuracy of MP-EST sometimes decreased). Based on the ILS level and number of genes, we would predict that accuracy would improve by including the 30 weak genes, and so the results in Liu et al. (2015b) are consistent with our study. Lanier et al. (2014) performed a simulation on 8-taxon model trees with two levels of ILS. Although Lanier et al. (2014) found that adding up to 50 low-variation genes to a single variable gene had little impact on STEM, each gene was represented by a majority-rule consensus tree from MrBayes (Huelsenbeck and Ronquist 2001) that may not have been fully resolved due to insufficient phylogenetic signal. Our study used fully resolved maximum likelihood gene trees, and so it is difficult to compare our results to those of Lanier et al. (2014).

Filtering based on missing data.—On average, filtering based on missing data did not improve the accuracy of any method under any model condition. Low amounts of filtering generally did not affect method accuracy, but large amounts of filtering resulted in increased species tree estimation error for all methods. Unlike filtering based on GTEE, filtering based on missing data did not substantially lower the average GTEE in the retained genes for most model conditions (Supplementary Fig. S11 available on Dryad).

To the best of our knowledge, only one prior simulation study (Huang and Knowles 2016) has explicitly examined how gene filtering based on missing data affects coalescent methods. Huang and Knowles (2016) simulated 8-taxon data sets using a protocol where gene trees differ from the species tree due to ILS and where the pattern of missing data was similar to those generated by RADtag protocols (Baird et al. 2008). This simulation design resulted in a correlation between the genes with missing data and the genes with higher rates of evolution. Huang and Knowles (2016) noted that these deleted genes were the ones that provided resolution at difficult nodes, so that deleting these genes decreased phylogenetic signal, and the species trees estimated using the shallowest divergence method (a site-based coalescent method) had higher error on the filtered data sets than on the full data sets. A likely explanation for why Huang and Knowles (2016) found gene filtering based on missing data to substantially reduce accuracy is that filtering *decreased* the amount of phylogenetic signal. In other words, gene filtering based on missing data can be doubly detrimental if it reduces the average signal

per gene as well as reduces the total number of genes, an observation that is consistent with our study.

Data quality versus data quantity.—By definition, filtering reduces the amount of data available, and hence should generally reduce species tree accuracy. However, sometimes filtering based on GTEE improved species tree estimation using summary methods. An examination of the conditions when filtering improved the accuracy of the summary methods shows that the average gene tree accuracy also improved substantially without removing too many genes. Hence, although there was a reduction in data quantity (number of genes), there was an increase in data quality (accuracy of gene trees). Similarly, when filtering reduced accuracy, either the gene tree quality did not improve by filtering (e.g., when filtering is based on missing data) or the gene tree quality improved but not by enough to offset the reduction in quantity. In other words, the impact of filtering is fundamentally a question of data quality versus data quantity.

Local branch support.—Gene filtering also impacted the local branch support (as estimated by ASTRAL-II); however, these results are somewhat difficult to interpret as the branches recovered in the estimated species trees were also impacted by gene filtering. Overall, gene filtering was either neutral or else decreased the mean local support of true positive and false positive branches in the ASTRAL-II species trees (Supplementary Figs. S6 and S9 available on Dryad).

When filtering was based on GTEE, the extent of this decrease in accuracy depended on the level of ILS. For example, it had very little impact on the mean support of true positive branches when the level of ILS was low/moderate (Supplementary Fig. S6a,b available on Dryad) but decreased by nearly 10% under very high ILS conditions (Supplementary Fig. S6e,f available on Dryad). Conversely, when the level of ILS was low/moderate, the mean local branch support of false positive branches decreased by over 10% but was not affected when the level of ILS was very high, an observation that may be at least partially explained by the recovery of fewer false branches.

Differences in local branch support due to filtering based on missing data in general did not seem to depend on the model conditions (Supplementary Figs. S9 and S10 available on Dryad). The exception to this was that the support of false branches increased from ~40% to ~50% under high ILS and moderate GTEE (Supplementary Fig. S10c available on Dryad). In this case, the average local support of true positive branches was still high (~90%), suggesting that the local support would still be useful in separating true and false branches.

Prior Empirical Studies

Several recent studies evaluated the impact of filtering on coalescent methods using empirical data sets: four studies evaluated filtering based on missing data (Chen et al. 2015; Hosner et al. 2016; Streicher et al. 2016; Longo et al. 2017) and six evaluated filtering based on proxies for GTEE or related criteria (Chen et al. 2015; Hosner et al. 2016; Meiklejohn et al. 2016; Simmons et al. 2016; Blom et al. 2017; Longo et al. 2017). These studies differ in many ways, including the type of data sets, the filtering criteria, the methods used, and the evaluation of species tree quality; however, all these studies used MLBS to estimate branch support values for the species trees computed using summary methods.

Two empirical studies observed that deleting genes based on the degree of missing data typically did not improve the quality of estimated species trees, and sometimes even reduced quality, as measured using appearance of unlikely clades (Chen et al. 2015; Hosner et al. 2016). The other two studies evaluating this type of filtering (Streicher et al. 2016; Longo et al. 2017) observed that deleting genes with missing data *could* increase branch support—provided that the correct filtering threshold (i.e., degree of missing data) was used. However, the best threshold differed between the studies, and selecting the wrong threshold could reduce branch support in the estimated species tree, indicating not only that selecting the threshold is important but also that the optimal filtering threshold based on missing data may depend on the species tree estimation method and the data set properties in ways that are difficult to ascertain. Given the difficulties in interpreting branch support in species trees computed using coalescent-based methods (especially when based on MLBS, which is how these studies computed branch support) and given the lack of evidence that filtering this way improves accuracy (Chen et al. 2015; Hosner et al. 2016), filtering genes because many species are missing seems to be undesirable. An interesting counterpart to this line of analysis is the observation in Hosner et al. (2016) that filtering genes that have fragmentary sequences (a different kind of missing data situation that they refer to as "type-II" missing data) can improve accuracy. The explanation is likely that type-II missing data increased GTEE, which in turn impacted the summary methods.

Five empirical studies have examined how species trees computed using summary methods are impacted by gene filtering based on proxies for GTEE; three of these (Hosner et al. 2016; Meiklejohn et al. 2016; Longo et al. 2017) recommended filtering and two (Chen et al. 2015; Blom et al. 2017) did not recommend filtering. A sixth study (Simmons et al. 2016) explored a related filtering strategy that identified and removed outlier genes (i.e., genes whose gene trees are topologically very distant from other gene trees) but did not recommend this type of filtering as a way to improve species tree estimation. The five studies that used proxies for GTEE to evaluate species tree quality did so in similar ways (usually evaluating the similarity to a CA-ML tree, the appearance of unlikely clades, or the branch support of the estimated species tree) and yet came to different conclusions. In particular, Meiklejohn et al.

(2016) noted that different summary methods were not all impacted identically by genes with potentially high GTEE (resulting from, for example, low numbers of parsimony informative sites), and so filtering may benefit some methods substantially and not benefit others as much (or at all). Thus, the impact of filtering based on proxies for GTEE depends on the specific method for estimating species trees, and most likely also depends on properties of the data set.

It is worth examining in some detail the study that showed the largest favorable impact of filtering (Meiklejohn et al. 2016) as the authors provided fairly convincing evidence that filtering reduced the appearance of unlikely clades in trees computed by MP-EST. Meiklejohn et al. (2016) noted that many of the gene trees in their data set—especially ones with few informative sites—had highly supported GTEE. Their evidence for this assertion is that these gene trees had strong support for clades that conflicted with an established clade in the species tree and that this established clade was separated from the rest of the species by a long branch. Since ILS is unlikely to occur on long branches, this means that most true gene trees would be expected to display this established clade. Therefore, the condition in which Meiklejohn et al. (2016) observed that gene filtering based on GTEE provided an improvement in species tree accuracy (by reducing the frequency of unlikely clades) is characterized by a particular kind of GTEE—strong support for a false branch, rather than low support on the false branches (which is what we would expect if given a sequence alignment without any phylogenetically informative sites). Meiklejohn et al. (2016) note that removing genes with few informative sites simultaneously removed many of these genes that had highly supported GTEE.

Even though the empirical studies did not come to any clear agreement regarding the impact of filtering, several conclusions can be drawn. First, as noted, not all methods responded identically, so improvements resulting from filtering for one method do not imply that other methods—even of the same type—will respond in the same way. Second, when there was improvement, it was often because the genes that were deleted had high GTEE—with potentially the biggest improvement occurring when the GTEE was highly supported. Third, interpreting increased similarity to a CA-ML tree as an improvement in accuracy is a bit tricky, as it depends on the accuracy of CA-ML. However, our study showed that CA-ML was frequently the most accurate of the species trees estimation methods, even under conditions with localized regions of high ILS (so that overall the AD value was not too high). We also showed that filtering genes based on GTEE reduced species tree error when species trees have a mixture of long and short branches (so that overall ILS is not large) or when the level of GTEE is extremely high. These two conditions seem likely to be true of many empirical studies, and in particular high GTEE seems to be a very general problem for modern phylogenomic data sets (Table 1).

Conclusions

Method Selection

Methods should be chosen based on how each method responds to data quality (per locus) and data quantity (number of loci). Therefore, given a collection of loci, many of which have only a few informative sites, we may prefer to use CA-ML over a summary method, as CA-ML is likely to be at least as accurate as the summary methods we examined under model conditions with high GTEE. However, given a collection of loci, many of which appear to produce high quality gene trees, we may prefer to use one of the summary methods over CA-ML. Whether we trust the results of the summary method over CA-ML and whether we decide to use gene filtering in conjunction with the summary method both depend on the level of ILS and GTEE. Similarly, although SVDquartets was not generally among the most accurate methods, this study showed that SVDquartets could be more accurate than summary methods when GTEE was sufficiently high but was not generally better than CA-ML. However, the conditions for which SVDquartets may be more accurate than summary methods and/or CA-ML (e.g., high to very high ILS and very large number of loci) require further exploration. Regardless, method selection depends on the data set and in particular on the levels of ILS and GTEE, making the estimation of these quantities important methodological issues in phylogenomics.

The computational cost of each method is also an important consideration. While it is certainly true that some approaches to coalescent-based species tree estimation are substantially more expensive than CA-ML (e.g., Bayesian coestimation methods), many coalescent methods can be faster than CA-ML for large multilocus data sets. For example, the main computational effort in using (most) summary methods is the estimation of the individual gene trees; however, this step is embarrassingly parallel, enabling the estimation of gene trees under more complex statistical models that (for computational reasons) cannot be applied to long concatenated alignments. In contrast, the running time for CA-ML is significantly impacted by the number of loci, and only some maximum likelihood methods (e.g., ExaML; Kozlov et al. 2015) are fast enough to be used on very long concatenated alignments. However, even ExaML is expensive to use when the number of species is large; the concatenated analysis using ExaML for 48 species and approximately 14,000 loci in the Avian Phylogenomics Project (Jarvis et al. 2014) took more than 250 CPU years and 1 TB of shared memory. In contrast, MP-EST took only 5 CPU years to analyze the same data set, and most of that time was spent estimating the gene trees with bootstrapping; hence, the same data set could be analyzed using single best ML gene trees and a faster summary method, and still use only a small fraction of the time. SVDquartets and other site-based methods are also very efficient for large numbers of loci, and improved versions of these methods may scale to very large numbers of species.

Marker Selection

With respect to marker selection, perhaps the most significant outcome of this study is the general observation that GTEE (resulting from low phylogenetic signal per locus) has a substantial negative impact on summary methods. This impact on summary methods can be so great that they may not be appropriate methods for analyzing UCEs and other commonly used types of phylogenomic data where the loci have few informative sites. In contrast, CA-ML seems to do well with these low-signal loci, as long as there are enough of them. Thus, this study and other empirical studies suggest that the type of marker that works well for CA-ML may not be the type that works well for summary methods; more concretely, the optimal markers for species tree estimation using summary methods may have higher rates of evolution than the optimal markers for CA-ML. We would recommend therefore that prior to collecting data, researchers should consider whether the expected level of ILS is high enough that a summary method would be beneficial, and then select markers based on this assessment.

Moving Forward

The trends in this study support the conclusion that ILS and GTEE affect method selection and use of gene filtering based on GTEE (if summary methods are used). This emphasizes the need to accurately predict ILS and GTEE (or, alternatively, how to modify the gene trees to reduce estimation error) in empirical data sets. However, both ILS and GTEE are very challenging to estimate. One way to estimate ILS is to examine the species tree for rapid radiations (i.e., successions of short branches), but this may depend on having a very good estimate of the species tree, which is not always available. Another way is to examine the heterogeneity between estimated gene trees; however, GTEE itself adds to the observed heterogeneity, and it can be difficult to distinguish between heterogeneity due to GTEE and heterogeneity due to ILS. Hence, estimating ILS generally also depends on the ability to evaluate error in estimated gene trees.

Thus, the estimation of GTEE is a fundamental problem in phylogenomics that impacts method selection and the interpretation of estimated species trees. Closely related to estimating GTEE is the estimation of the probability that a given branch in an estimated gene tree is correct (Anisimova et al. 2011). The most popular such technique for this is probably non-parametric bootstrapping (Felsenstein 1985), which can be reasonably reliable when performed correctly and the tree estimation method is statistically consistent (Efron et al. 1996; Holmes 2003, 2005; Susko 2009). Fortunately, other mathematical and statistical approaches for computing the probability of a branch in a tree (or even the entire tree) being correct have also been developed (Holmes 2005; Fischer and Steel 2009; Townsend et al. 2012; Susko and Roger 2012; Salichos et al. 2014), and may provide better indicators of GTEE than the usual non-parametric bootstrapping approach.

Empirical arguments can also be used to provide evidence of GTEE. For example, Meiklejohn et al. (2016) provided an argument for GTEE in their data set by demonstrating that a large number of the estimated gene trees had strongly supported branches that conflicted with an established clade in the species tree that was separated by a long branch from all the remaining species. Other evidence for GTEE can be provided by demonstrating model misspecification or by showing that the tree topology is not stable. Another approach is to identify gene trees that are topologically very distant to the other gene trees; such an approach was explored by Simmons et al. (2016), who found it useful for potentially identifying GTEE as well as other types of errors in phylogenomic data sets. Finally, very short sequences with low phylogenetic signal are inherently likely to produce high GTEE, and can even produce highly supported GTEE, which may be the most problematic of all conditions.

Finally, Chen et al. (2015) make an interesting point about gene selection strategies that is relevant to the question of gene filtering. They argue that node-specific strategies (Salichos and Rokas 2013) should be used to select genes, rather than "nonspecific" strategies that select genes based on overall phylogenetic signal and/or assessment of gene tree accuracy. Chen et al. (2015) showed that the selection of genes to help resolve specific phylogenetic questions (e.g., *Is Amborella the sister of land plants?*) is more likely to result in an answer with high support than selecting genes based on generic signal. For example, Chen et al. (2015) wrote, "In some extreme cases, these nonspecific data sets can correctly resolve some difficult nodes but result in high support for erroneous relationships for other nodes." They concluded that "One possible explanation for this phenomenon is that each gene has a different resolving power on different time scales and on different evolutionary scenarios... Nonspecific data sets may produce a well-resolved relationship for an ancient divergence event but do not have enough phylogenetic signal to recover accurate phylogeny for a recent radiation or vice versa." The observations by Chen et al. (2015) are very similar to earlier observations made by Townsend and Leuenberger (2011), who noted that "characters that are highly informative early in history rapidly become sources of phylogenetic noise due to multiple hits for deeper divergences." Thus, Townsend and Leuenberger (2011) and Chen et al. (2015) support the hypothesis that phylogenomic species tree estimation is likely to benefit from a mixture of genes aiming to resolve different parts of the tree, rather than selecting genes on the basis of overall high signal, and potentially suggest the possibility that adding genes rather than deleting genes may be the most promising direction.

This study also has consequences for method developers. In particular, because many phylogenomic data sets are based on UCEs and other markers that are

highly conserved, the performance of methods should be explored under conditions where individual genes have low phylogenetic signal and so gene trees have high estimation error. We also note that the published species trees for many phylogenomic data sets have a combination of short and long branches, and so fall into the "low/moderate" ILS condition that we explored. Thus, a common type of phylogenomic data set is one where the individual markers have low phylogenetic signal and the level of ILS, even if high enough to be in the anomaly zone, is not too high. However, many key evolutionary questions remain unanswered because the current coalescent-based approaches do not exceed the accuracy of CA-ML on these data sets. In particular, this study suggests that GTEE seems to be at least as influential as ILS in terms of its impact on species tree estimation, at least when using summary methods. Therefore, this study suggests that future research should carefully examine conditions where many of the loci tend to have low phylogenetic signal, and also suggests that method developers should design species tree estimation methods that can provide high accuracy under high GTEE conditions.

Fortunately, species tree estimation is a fast moving field with many new theoretical results established and methods created in just the last few years (e.g., Boussau et al. 2013; Roch and Snir 2013; Yu et al. 2014; Solís-Lemus and Ané 2016; Solís-Lemus et al. 2016; Wen et al. 2016; Dasarathy et al. 2017; Shekhar et al. 2017; Zhu et al. 2017). With the increased attention that species tree (and phylogenetic network) estimation is receiving, we are optimistic that, over the next few years, new methods may be developed that will provide even better accuracy and scalability to large data sets.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.km24v.

Funding

This work was supported by the U.S. National Science Foundation [Grant Number DGE-1144245 to E.K.M. and Grant Number CCF-1535977 to T.W.]. This research was part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation [Grant Numbers OCI-0725070 and ACI-1238993] and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This research made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program in conjunction with the National Center for Supercomputing Applications and which is supported by funds from the University of Illinois at Urbana-Champaign.

Acknowledgments

We thank Sébastien Roch, Andrew Roger, Mike Steel, and Ed Susko for helpful discussions regarding the theoretical literature on experimental design in phylogenomics, and we thank Mike Steel for the suggestion that summary methods can improve with missing data, described in the Appendix. We also thank Associate Editor Matthew Hahn and two anonymous reviewers for their very helpful comments that improved the quality of this work.

Appendix

We present an example of a summary method (suggested by Mike Steel) that actually improves in accuracy when enough data are missing. Consider a set of unrooted gene trees leaf-labelled by the same set S of species. For each possible subset A of four or more species in S, an unrooted tree t_A (called the "dominant tree" on A) is constructed as follows. From all gene trees that contain all species in A, the most frequently observed induced tree on A becomes the dominant tree t_A. If the set of dominant trees $\{t_A : A \subseteq S, |A| \geq 4\}$ is compatible, then a compatibility tree (i.e., the tree that agrees with all the dominant trees) is returned by the summary method; otherwise, a star tree (i.e., the tree with all leaves adjacent to a single central vertex) is returned by the summary method. If the species tree is in the anomaly zone (Degnan and Rosenberg 2006), then by definition the dominant tree t_S will be different from the true species tree on S with probability going to 1 as the number of genes increases. But because there are no anomalous unrooted quartet trees, the dominant tree t_A for $|A|=4$ will be identical to the species tree on A with probability converging to 1 as the number of genes increases (Allman et al. 2011). Hence, whenever a species tree with $|S|>4$ is in the anomaly zone, the set of dominant trees will be incompatible, and the summary method will return a star tree.

Now suppose $|S|=5$ but a single species is selected at random for each gene and deleted. As every gene tree has only four remaining species, only subsets with $|A|=4$ can be defined, and the set of dominant trees will contain only unrooted quartet trees. Because there are no anomalous unrooted quartet trees, the dominant trees will match the species tree with probability converging to 1 as the number of genes increases (Allman et al. 2011). As the number of genes increases with probability converging to 1, the set of dominant trees will be compatible with each other, and the species tree will be a compatibility tree for the set of dominant trees. Finally, every quartet tree induced by the species tree will appear as a dominant tree with probability converging to 1 as the number of genes increases; hence, the species tree will be the unique compatibility tree for the set of dominant tree, and the summary method will return the true species tree. In other words, this summary method is statistically consistent under the MSC model when there

is enough missing data so that each gene has at most four species; however, this summary method is statistically inconsistent otherwise, including the case when no data are missing.

References

Allman E.S., Degnan J.H., Rhodes J.A. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62:833–862.

Anisimova M., Gil M., Dufayard J.-F., Dessimoz C., Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60:685–699.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C. Shiver A.L., Lewis Z.A., Selker W.A., Cresko E.U., Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

Bayzid M.S., Hunt T., Warnow T. 2014. Disk covering methods improve phylogenomic analyses. BMC Genomics 15:S7.

Bayzid M.S., Mirarab S., Boussau B., Warnow T. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS ONE 10:30129183.

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

Betancur-R R., Naylor G.J., Ortí G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst. Biol. 63:257–262.

Blom M.P.K., Bragg J.G., Potter S., Moritz C. 2017. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. Syst. Biol. 66:352–366.

Boussau B., Szöllősi G.J., Duret L., Gouy M., Tannier E., Daubin V. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23:323–330.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from Biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Chen M.-Y., Liang D., Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64:1104–1120.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Chifman J., Kubatko L. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J. Theoret. Biol. 374:35–47.

Cho S., Zwick A., Regier J.C., Mitter C., Cummings M.P., Yao J., Du Z., Zhao H., Kawahara A.Y., Weller S., Davis D.R. Baixeras J., Brown J.W. Parr C. 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol. 60:782–796.

Chou J., Gupta A., Yaduvanshi S., Davidson R., Nute M., Mirarab S., Warnow T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics 16:S2.

Dasarathy G., Mossel E., Nowak R., Roch S. 2017. Coalescent-based species tree estimation: a stochastic Farris transform. arXiv:1707.04300.

Dasarathy G., Nowak R., Roch S. 2015. Data requirement for phylogenetic inference from multiple loci: a new distance method. IEEE/ACM Trans. Comput. Biol. Bioinform. 12:422–432.

Davidson R., Vachaspati P., Mirarab S., Warnow T. 2015. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. BMC Genomics 16:S1.

de Oca A.N.-M., Barley A.J., Meza-Lázaro R.N., García-Vázquez U.O., Zamora-Abrego J.G., Thomson R.C., Leaché A.D. 2017. Phylogenomics and species delimitation in the knob-scaled lizards of the genus Xenosaurus (Squamata: Xenosauridae) using ddRADseq data reveal a substantial underestimation of diversity. Mol. Phylogenet. Evol. 106:241–253.

DeGiorgio M., Degnan J.H. 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. Mol. Biol. Evol. 27:552–569.

DeGiorgio M., Degnan J.H. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst. Biol. 63:66–82.

Degnan J., Rosenberg N. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics 2:762–768.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dornburg A., Townsend J.P. Brooks W., Spriggs E., Eytan R.I., Moore J.A., Wainwright P.C. Lemmon A., Lemmon E.M., Near T.J. 2017. New insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. Mol. Phylogenet. Evol. 110:27–38.

Dornburg A., Townsend J.P. Friedman M., Near T.J. 2014. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. BMC Evol. Biol. 14:169.

Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.

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

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104:5936–5941.

Efron B., Halloran E., Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:13429–13429.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Fischer M., Steel M. 2009. Sequence length bounds for resolving a deep phylogenetic divergence. J. Theoret. Biol. 256:247–252.

Fletcher W., Yang Z. 2009. INDELible: A flexible simulator of biological sequence Evolution. Mol. Biol. Evol. 26:1879–1888.

Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80:231–266.

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

Hobolth A., Dutheil J.Y. Hawks J., Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21:349–356.

Holmes S. 2003. Bootstrapping phylogenetic trees: theory and methods. Stat. Sci. 18:241–255.

Holmes S. 2005. Statistical approach to tests involving phylogenies. *Mathematics of Evolution and Phylogeny* (O. Gascuel ed.). Oxford, UK: Oxford University Press, pp. 91–120.

Hosner P.A., Faircloth B.C., Glenn T.C., Braun E.L., Kimball R.T. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol. Biol. Evol. 33:1110–1125.

Hovmöller R., Knowles L.L., Kubatko L.S. 2013. Effects of missing data on species tree estimation under the coalescent. Mol. Phylogenet. Evol. 69:1057–1062.

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

Huang H., Knowles L.L. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst. Biol. 65:357–365.

Huelsenbeck J.P., Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J.,

Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C. Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E. Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Jewett E.M., Rosenberg N.A. 2012. iGLASS: An Improvement to the GLASS method for estimating species trees from gene trees. J. Comput. Biol. 19:293–315.

Jiang W., Chen S.-Y., Wang H., Li D.-Z., Wiens J.J. 2014. Should genes with missing data be excluded from phylogenetic analyses? Mol. Phylogenet. Evol. 80:308–318.

Kinney J.B., Atwal G.S. 2014. Equitability, mutual information, and the maximal information coefficient. Proc. Natl. Acad. Sci. USA 111:3354–3359.

Kozlov A.M., Aberer A.J., Stamatakis A. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31:2577–2579.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Lanier H.C., Huang H., Knowles L.L. 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. Mol. Phylogenet. Evol. 70:112–119.

Lanier H.C., Knowles L.L. 2015. Applying species-tree analyses to deep phylogenetic histories: Challenges and potential suggested from a survey of empirical phylogenetic studies. Mol. Phylogenet. Evol. 83:191–199.

Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.

Leaché A.D., Rannala B. 2010. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60:126–137.

Leavitt S.D., Grewe F., Widhelm T., Muggia L., Wray B., Lumbsch H.T. 2016. Resolving evolutionary relationships in lichen-forming fungi using diverse phylogenomic datasets and analytical approaches. Sci. Rep. 6.

Linkem C.W., Minin V.N., Leaché A.D. 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (squamata: Scincidae). Syst. Biol. 65:465–477.

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

Liu L., Wu S., Yu L. 2015a. Coalescent methods for estimating species trees from phylogenomic data. J. Syst. Evol. 53:380–390.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015b. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360:36–53.

Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:1–18.

Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–477.

Longo S.J., Faircloth B.C. Meyer A., Westneat M.W., Alfaro M.E., Wainwright P.C. 2017. Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. Mol. Phylogenet. Evol. 113:33–48.

Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Mallo D., de Oliveira Martins L., Posada D. 2016. SimPhy: phylogenomic simulation of gene, locus and species trees. Syst. Biol. 65:334–344.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22:746–754.

McCormack J.E., Huang H., Knowles L.L. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58:501–508.

Meiklejohn K., Faircloth B., Glenn T., Kimball R., Braun E. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65:612–627.

Mendes F., Hahn M. 2017. Why concatenation fails near the anomaly zone. Syst. Biol. 67:158–169.

Mirarab S., Bayzid M.S. Boussau B., Warnow T. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:1250463.

Mirarab S., Bayzid M.S., Warnow T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65:366–380.

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

Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.

Ohno S. 1970. Evolution by gene duplication. New York, NY: Springer.

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

Patel S., Kimball R., Braun E. 2013. Error in phylogenetic estimation for bushes in the tree of life. J. Phylogenet. Evol. Biol. 1:110.

Posada D. 2016. Phylogenomics for systematic biology. Syst. Biol. 65:353–356.

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.

Reaz R., Bayzid M.S., Rahman M.S. 2014. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. PLoS ONE 9:e104008.

Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Roch S., Snir S. 2013. Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. J. Comput. Biol. 20:93–112.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of alignments can be statistically inconsistent. Theor. Popul. Biol. 100:56–62.

Roch S., Warnow T. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst. Biol. 64:663–676.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

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

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.

Shekhar S., Roch S., Mirarab S. 2017. Species tree estimation using ASTRAL: how many genes are enough? arXiv:1704.06831.

Simmons M.P., Sloan D.B., Gatesy J. 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. Mol. Phylogenet. Evol. 97:76–89.
Snir S., Rao S. 2012. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. Mol. phylogenet. Evol. 62:1–8.
Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12:e1005896.
Solís-Lemus C., Yang M., Ané C. 2016. Inconsistency of species tree methods under gene flow. Syst. Biol. 65:843–851.
Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94, Part A:1–33.
Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Steel M.A., Székely L.A. 2002. Inverting random functions II: Explicit bounds for discrete maximum likelihood estimation, with applications. SIAM J. Discrete Math. 15:562–575.
Streicher J.W., Schulte J.A., Wiens J.J. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? an empirical study in iguanian lizards. Syst. Biol. 65:128–145.
Streicher J.W., Wiens J.J. 2016. Phylogenomic analyses reveal novel relationships among snake families. Mol. Phylogenet. Evol. 100:160–169.
Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Susko E. 2009. Bootstrap support is not first-order correct. Syst. Biol. 58:211–223.
Susko E., Roger A.J. 2012. The probability of correctly resolving a split as an experimental design criterion in phylogenetics. Syst. Biol. 61:811–821.
Swofford D. 2002. Phylogenetic analysis using parsimony (* and other methods). version 4. Sunderland, MA: Sinauer Associates.
Swofford D.L. 2017. PAUP*: Phylogenetic analysis using parsimony (and other methods). version 4a152. Available at https://people.sc.fsu.edu/~dswofford/paup_test/.
Syvanen M. 1985. Cross-species gene transfer; implications for a new theory of evolution. J. Theoret. Biol. 112:333–343.
Takahata N. 1989. Gene genealogy in three related populations–consistency probability between gene and population trees. Genetics 122:967–966.
Tavaré S. 1986. *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences*. Lectures on Mathematics in the Life Sciences, vol. 17. American Mathematical Society, 57–86. Providence, RI.
Townsend J.P., Leuenberger C. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol. 60:358–365.
Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61:835–849.
Vachaspati P., Warnow T. 2015. ASTRID: Accurate Species TRees from Internode Distances. BMC Genomics 16:1–13.
Wen D., Yu Y., Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genet. 12:e1006006.
Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews, S., and Melkonian, M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., dePamphilis, C.W., Chen, T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack, J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. USA 111:E4859–E4868.
Wiens J.J., Morrill M.C. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.
Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92:63–71.
Xi Z., Liu L., Davis C.C. 2016. The impact of missing data on species tree estimation. Mol. Biol. Evol. 33:838–860.
Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. USA 111:16448–16453.
Zhu J., Wen D., Yu Y., Meudt H., Nakhleh L. 2017. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. bioRxiv: 143545.
Zimmermann T., Mirarab S., Warnow T. 2014. BBCA: Improving the scalability of *BEAST using random binning. BMC Genomics 15:S11.