Species Tree Inference under the Multispecies Coalescent on Data with Paralogs is Accurate

Peng Du[1] and Luay Nakhleh[*,1,2]
[1]Department of Computer Science
[2]Department of BioSciences, Rice University, 6100 Main Street, Houston, TX 77005
***Corresponding author:** E-mail: nakhleh@rice.edu.
Associate Editor: ...

Abstract

The multispecies coalescent (MSC) has emerged as a powerful and desirable framework for species tree inference in phylogenomic studies. Under this framework, the data for each locus is assumed to consist of orthologous, single-copy genes, and heterogeneity across loci is assumed to be due to incomplete lineage sorting (ILS). These assumptions have led biologists who use MSC-based inference methods to exclude from their analyses all loci that are present in more than a single copy in any of the genomes under consideration. Furthermore, such analyses involve the task of orthology assignment to avoid the potential of hidden paralogy in the data. The question we seek to answer in this study is: What happens if one runs an MSC-based species tree inference method on data where paralogy is present, in addition to or without ILS being present? Through simulation studies and analyses of two biological data sets, we show that running an MSC-based species tree inference method on data with paralogs, while treating all gene copies within a family as alleles from multiple individuals or by randomly selecting one copy per species, provide very accurate results that are almost identical to results obtained from single-copy genes. Our results have two significant implications. The operational implication is that MSC-based phylogenomic analyses do not have to be restricted to single-copy loci, thus maximizing the amount of data to be used. The interpretative implication is that when an MSC-based inference method infers an accurate species tree, this does not mean that ILS was much of a factor in the heterogeneity across loci.

Key words: Multispecies coalescent, incomplete lineage sorting, gene duplication and loss, orthology, paralogy.

Article

Introduction

Species phylogeny inference from genome-wide data entails accounting for the fact that the evolutionary histories of different loci could disagree with each other, as well as with the phylogeny of the species. Causes of such incongruence include incomplete lineage sorting (ILS), horizontal gene transfer (broadly interpreted to include all biological processes involving gene flow), and hidden paralogy caused by gene duplication/loss (GDL) (Maddison, 1997).

As inference of species phylogenies in light of incongruence relies on a model of how gene trees

grow within the branches of the species phylogeny (in addition to probabilistic models of sequence evolution on the gene trees), the multispecies coalescent (MSC) (Degnan and Rosenberg, 2009; Rannala and Yang, 2003; Takahata, 1989) has emerged as the most commonly employed model of such gene genealogy growth. Indeed, in the last two decades, a wide array of methods and computer programs have been developed for species tree inference under the MSC; see (Knowles and Kubatko, 2011; Liu *et al.*, 2009, 2015; Nakhleh, 2013) for recent reviews and surveys of these methods.

MSC-based inference of species trees assumes that the data consist of only orthologous sequences obtained from multiple loci across the genomes of the species under study. Therefore, before such an inference method is applied to a phylogenomic data set, paralogs must be identified and removed from the data. One way to accomplish this is by using orthology detection tools, and another is to restrict the data to only loci that are present in exactly (and, in some cases, at most) one copy in each of the genomes. Neither of these two approaches guarantees that the resulting data set includes only orthologous sequences. Furthermore, restricting the data to single-copy genes, which is by far the most common practice in the community, means much of the data must be excluded from the analysis. In particular, as more species are sampled, the frequency of genes that are present in single copy across all species is expected to decrease.

MSC-based inference relies on the distribution of gene trees as the signal for inference, and one could argue that the MSC induces a distribution that differs from that arisen from gene duplication and loss. While this could very well be true, large-scale phylogenomic studies are repeatedly highlighting scenarios where no gene tree in the entire data set agrees with the species tree; e.g., (Fontaine *et al.*, 2015; Jarvis *et al.*, 2014; Salichos and Rokas, 2013).

An obvious way to handle data sets where ILS and gene duplication/loss could both be suspected as being at play is to employ models of gene evolution that go beyond the MSC in order to incorporate GDL as well. Indeed, such models are beginning to emerge (Rasmussen and Kellis, 2012). However, the more complex the models of gene family evolution, the more computationally prohibitive statistical inference under these models becomes (Du and Nakhleh, 2018), rendering their applicability infeasibility except for very small data sets in terms of the number of species and gene families.

Given that much progress in terms of accuracy and computational efficiency has been made on MSC-based species tree inference methods, we ask in this paper the following question: Is MSC-based species tree inference robust to the presence of paralogs in the data? If the answer to this question is positive, then the reach of MSC-based

inference methods is significantly extended and the exclusion of loci from phylogenomic data sets is deemed unnecessary, thus providing more signal for the inference task. To answer this question, we study the performance of an accurate and efficient MSC-based species tree inference method, namely ASTRAL-III (Zhang *et al.*, 2018), on simulated and biological data in the presence of GDL. To facilitate this, we treat multiple copies of a gene in the same genome as multiple alleles from different individuals or randomly sample a single copy per genome. Of course, the former violates the mathematical assumptions of the MSC, and the latter does not guarantee that paralogs are not present (in fact, they would be present with a very high probability). Our results show that ASTRAL's inferences are very accurate, and are mostly identical to those made by ASTRAL when using only genes that are present in exactly one copy in each of the species.

Our findings have at least two major implications. First, orthology detection and restricting data to single-copy genes as a requirement for employing MSC-based inference can be mostly eliminated, thus making use of as much of the data as possible. In particular, for very large data sets (in terms of the number of species), eliminating all but single-copy genes might leave too few loci for the species tree to be inferred accurately. Our findings show that this data exclusion could be an unnecessary practice. The second implication has to do with invoking ILS as an explanation of the extensive incongruences observed in empirical data sets. Needless to say, the implication that ILS is a cause of the incongruence because MSC-based inference yielded a "correct" species tree is a logical fallacy. Our results described below clearly show that MSC-based inference obtains correct species trees even when ILS is not a factor at all. While our study does not answer the question "how extensive ILS really is in empirical data sets," it calls for caution about invoking ILS as a cause of incongruence because MSC-based inference produced a tree that "makes sense." Furthermore, if GDL is a cause of the incongruence and MSC-based inference is used, the estimated branch lengths of the species tree could be much smaller than their true values, and then, based on the inference, ILS is invoked due to the *estimated* short branches. This is a circular argument and should also be avoided. After all, GDL could be much more ubiquitous in practice than ILS, even though MSC-based inference yields accurate estimates of the species tree topology.

Results

Methods

For MCS-based species tree inference, we used ASTRAL-III (Zhang *et al.*, 2018), Version 5.6.3. ASTRAL takes as input gene trees. Given a collection of trees corresponding to gene families (one per gene family), we generated three types of input to ASTRAL:

- ONLY: The input consists of trees of *only* gene families that are present in exactly one copy in each of the species.
- ONE: The input consists of trees of *all* gene families, but where a single copy per species per gene family is selected at random and the remaining copies are removed. If a gene family has no copies at all for some species, then the resulting tree of that gene family also has no copies for that species.
- ALL: The input consists of trees of *all* gene families, but where all copies of a gene in a species are treated as multiple alleles from different individuals within the species. Similar to ONE, if a gene family has no copies at all for some species, then the resulting tree of that gene family also has no copies for that species.

ONLY corresponds to the practice that is followed in almost all phylogenomic studies that employ MSC-based inference, as it reduces, though not necessarily completely eliminates, the potential of hidden paralogy in the data. Unless GDL is not at play, ONE is almost always guaranteed to have hidden paralogs in the input. By construction, ALL has all orthologs and paralogs in the input, but all are effectively labeled as orthologs, with the (wrong) implied assumption that multiple individuals are sampled per species.

Simulation Setup

For model species trees, we used the trees on 16 fungi species and 12 fly species reported in (Rasmussen and Kellis, 2012) and shown in Fig. 1. Furthermore, to obtain 4-taxon and 8-taxon species trees, we restricted each of the two species trees to 4 and 8 species, respectively, as shown in Fig. 1.

For the fungi simulated data sets, we used four different duplication and loss rates (assuming duplication and loss rates are equal): 0 (to investigate the performance when only ILS is at play), 6×10^{-10}, 2×10^{-9}, and 6×10^{-9} per generation. The highest rate is similar to the duplication rate of 7.32×10^{-9} and loss rate of 8.59×10^{-9} used by Rasmussen and Kellis (2010). We used two effective population sizes of 10^7 and 5×10^7, where the former was also used by Rasmussen and Kellis (2012) as the true population size. We assumed 10 generations per year as in (Rasmussen and Kellis, 2012) and used 4×10^{-10} as the mutation rate per generation per site, similar to the rates of 3.3×10^{-10} and 3.8×10^{-10} used by Zhang and Wu (2017) and Lang and Murray (2008), respectively.

For the fly simulated data sets, we used four different duplication and loss rates (assuming duplication and loss rates are equal): 0, 1×10^{-10}, 2×10^{-10}, and 5×10^{-10} per generation. A GDL rate of 1.2×10^{-10} was used in (Rasmussen and Kellis, 2012; Zhang and Wu, 2017) and reported by Hahn *et al.* (2007). We used two effective population sizes of 10^6 and 5×10^6, similar to the values used in (Rasmussen and Kellis, 2012) and true value of 1.15×10^6 reported in (Pollard *et al.*,

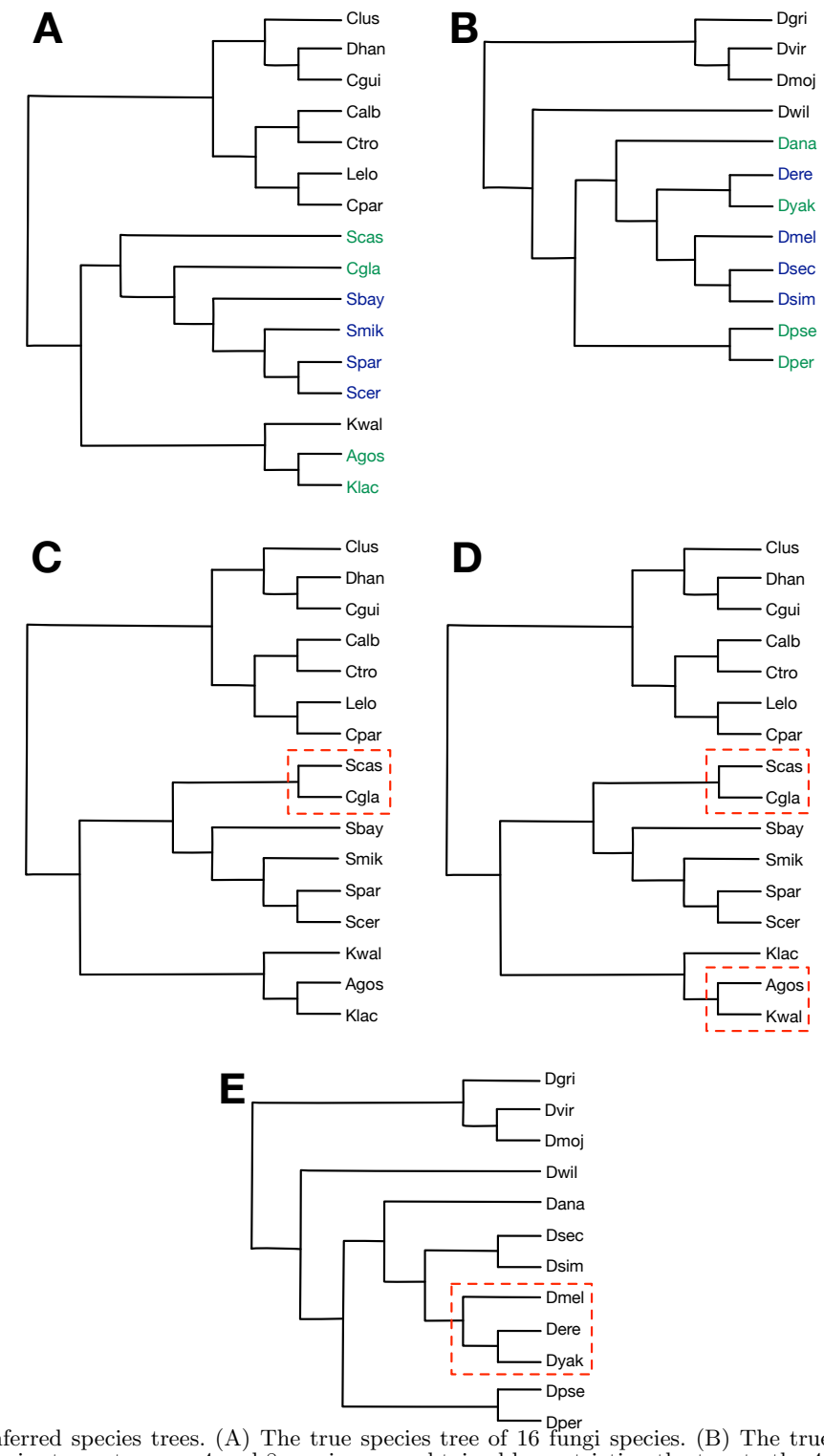

FIG. 1. True and inferred species trees. (A) The true species tree of 16 fungi species. (B) The true species tree of 12 fly species. For both species trees, trees on 4 and 8 species were obtained by restricting the tree to the 4 leaves colored in blue, and the 8 leaves colored in blue or green. The species tree topologies and their branch lengths in units of million years are reported on http://compbio.mit.edu/dlcoal/. For the fungi species tree, the length of the branch above the root was set to 25 million years, and for the fly species tree, it was set to 2.5 million years. (C) The yeast species tree inferred by ASTRAL(ONLY). (D) The yeast species tree inferred by ASTRAL(ALL) and ASTRAL(ONE). (E) The fly species tree inferred by ASTRAL(ONE).

2006; Sawyer and Hartl, 1992). We assumed 10 generations per year as in (Rasmussen and Kellis, 2012; Zhang and Wu, 2017) and used 3×10^{-9} as the mutation rate per generation per site, similar

to the rate of 5×10^{-9} used in (Zhang and Wu, 2017).

To generate gene trees, we used the three-tree model of Rasmussen and Kellis (2012) using the aforementioned parameters. In this model, a *locus tree* is simulated within the branches of the species tree. All incongruence between the locus tree and the species tree is due to GDL. Then, a gene tree is simulated within the branches of the locus tree, where all incongruence between the locus tree and the gene tree is due to ILS. The resulting gene tree differs from the species tree due to a combination of ILS and GDL. Using the locus trees as input to an inference method amounts to using data where all incongruence is solely due to GDL. Setting the rates of GDL to 0.0 amounts to generating gene trees where all incongruence is solely due to ILS. We then used seq-gen (Rambaut and Grass, 1997) to generate DNA sequences of length 1000 under the JC69 model (Jukes *et al.*, 1969) on the simulated gene trees (whose branch lengths are in units of mutation per site). 10,000 gene families (each containing a locus tree and corresponding gene tree and DNA sequences) were simulated in this fashion as one data set for each combination of GDL rate and population size. Ten such data sets, each with 10,000 gene families, were generated to obtain statistically significant results.

To study the effect of using data sets of varying sizes, for each of the 10 data set, we randomly sampled 10, 50, 100, and 250 gene families from the 10,000 gene families under the ALL, ONE, and ONLY scenarios. In case the number of available gene families that fits ONLY is smaller than the desired size, that number of gene families is used (e.g., when only 6 gene family trees are available when data sets of size 10 are desired, the 6 trees are used as input). We inferred gene trees from the DNA sequences using RAxML (Stamatakis, 2014), Version 8, on the DNA sequences. Finally, for each data set of trees (true gene trees or inferred ones and locus trees) of a given size, we fed the data set as input to ASTRAL and computed the Robinson-Foulds distance (Robinson and Foulds, 1981), normalized by the number of internal branches in the (unrooted) species tree to obtain a value between 0 and 1, between the true and inferred species trees.

Characteristics of the Input Data

Before we describe the inference results, we discuss the characteristics of the simulated gene trees. First, we inspect the effects of gene duplication and loss on the number of gene copies per species in each gene family. Table 1 shows the average number of gene copies per species in the 16-taxon and 12-taxon data sets under the various settings of effective population sizes and duplication and loss rates. Clearly, the higher the GDL rates, the more gene copies are present per species. We then counted the average (over the 10 data sets per setting) number of gene families for each setting that have exactly one copy per species. The results are shown in Table 2. The table shows

Table 1. Average number of gene copies per species per gene family for the 10 data sets of 10,000 gene families each, per setting of effective population size, duplication and loss rates.

16-taxon data			12-taxon data		
$N_e \rightarrow$ GDL rate $\downarrow$	10^7	5×10^7	$N_e \rightarrow$ GDL rate $\downarrow$	10^6	5×10^6
6×10^{-10}	1.318	1.314	1×10^{-10}	1.003	1.002
2×10^{-9}	2.516	2.495	2×10^{-10}	1.006	1.007
6×10^{-9}	6.155	5.837	5×10^{-10}	1.027	1.026

two expected trends. First, as the GDL rates increase, the number of single-copy gene families decreases. In particular, for the 16-taxon data set and its 8- taxon subset, there are almost no single-copy gene families. Second, when restricting the data to smaller numbers of taxa, more single-copy gene families are available. These two trends highlight a very important point that is often ignored in empirical analyses: The availability of single-copy genes in a small set of species does not necessarily mean these genes did not undergo gene duplication and loss, nor does it mean that paralogy is not hidden in the data. When considering this same set of species in the context of a larger clade from which they are sampled, the extent of duplication and loss emerges. For example, when focusing on the gene families in the 4-taxon subset of the 16-species data set, there are about 1,600 single-copy gene families (for the smallest GDL rates considered). However, when considering the full 16-species data set, the full extent of complexity emerges, as *all* of these 1,600 gene families underwent duplication and loss, and almost none of them is present in single copy. This pattern of hidden paralogy is a result of taxon sampling, and its extent is fully revealed only when a larger data set of species is considered.

Finally, for the higher population sizes, the amount of ILS is larger. For the higher GDL rates, the amount of incongruence and multiple-copy genes become larger. And, for the larger gene families, the inferred gene trees are less accurate. These results are summarized in Supplementary Material.

Results on Simulated Data

We are now in position to describe the inference results. We first ran ASTRAL on the true gene trees for all three input scenarios: ALL, ONE, and ONLY. In this case, gene tree estimation error is not a cause of gene tree incongruence. Instead, all incongruence is due to a combination of ILS and GDL. Results for the full 16-taxon and 12-taxon data sets are shown in Fig. 2 and Fig. 3, respectively. Observe that in both cases, using input data with GDL rates of 0 amounts to inferring a species tree from gene trees whose incongruence is solely due to ILS. In the case of the simulated fly data, ASTRAL almost always infers the correct species tree, regardless of how the input gene families were selected (ALL, ONLY, or ONE). In the case of the simulated yeast data, ASTRAL almost always infers the correct species tree when the GDL rates are 0, and converges to the true species as the number of gene families increases when one copy per species is randomly

Table 2. The average number of gene families that fit the ONLY setting out of the 10,000 gene families for 10 data sets. For each of the 16- and 12-taxon data sets, numbers are shown for the full data set as well as for the 8- and 4-taxon subsets.

16-taxon data							12-taxon data						
	16 taxa		8 taxa		4 taxa			12 taxa		8 taxa		4 taxa	
$N_e \rightarrow$ GDL rate $\downarrow$	10^7	5×10^7	10^7	5×10^7	10^7	5×10^7	$N_e \rightarrow$ GDL rate $\downarrow$	10^6	5×10^6	10^6	5×10^6	10^6	5×10^6
6×10^{-10}	0	0.1	48.6	50.4	1564.3	1564.3	1×10^{-10}	4631.3	4562.8	6743.0	6730.5	8504.2	8489.6
2×10^{-9}	0	0	0.3	0	173.3	176.3	2×10^{-10}	2166.3	2154.3	4609.5	4605.0	7279.1	7289.4
6×10^{-9}	0	0	0	0	6.5	7.8	5×10^{-10}	275.1	281.8	1616.5	1615.3	4760.1	4742.5

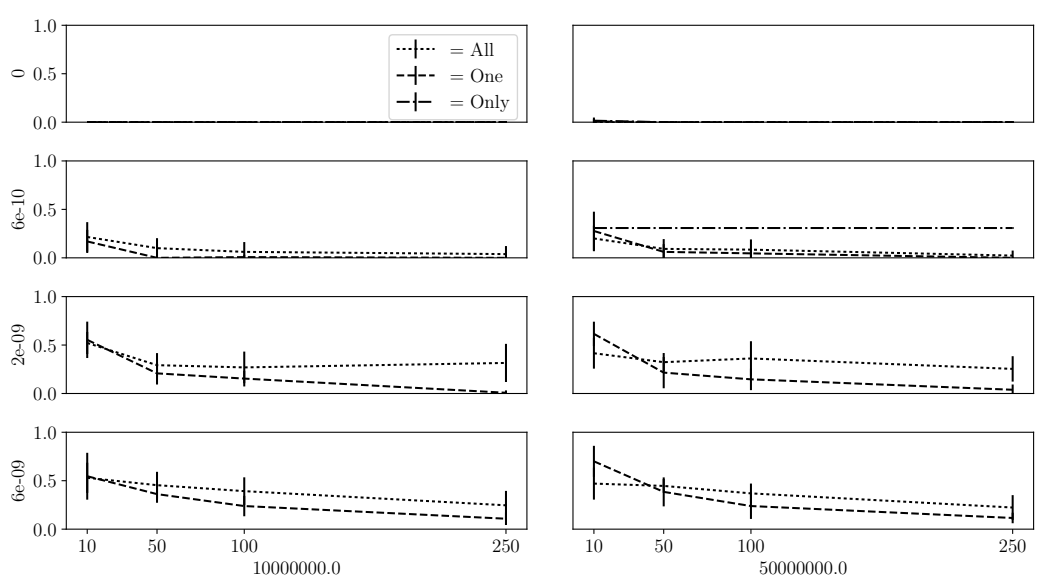

FIG. 2. The RF distance between the true species tree and the inferred ones from true gene trees of the 16-taxon simulated yeast gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

chosen for each gene family. The results are very comparable, as well, when all copies are treated as multiple individuals. It is unsurprising that the results based on ONLY are poor, given how few single-copy gene families are available in this data set (Table 2).

We then inferred species trees using ASTRAL from the gene tree estimates. Results for the full 16-taxon and 12-taxon data sets are shown in Fig. 4 and Fig. 5, respectively. As the results show, for the 16-taxon data, there were hardly any differences in the accuracy of the species trees inferred by ASTRAL when compared to those obtained from the true gene trees. However, for the 12-taxon data, the species trees inferred from gene tree estimates are less accurate than those inferred from true gene trees. The reason for this is that the gene tree estimates had about 50% error (Supplementary Material), and this error has confounded the signal for ASTRAL. In other words, almost all error in the species tree estimates in this case have to do with the gene

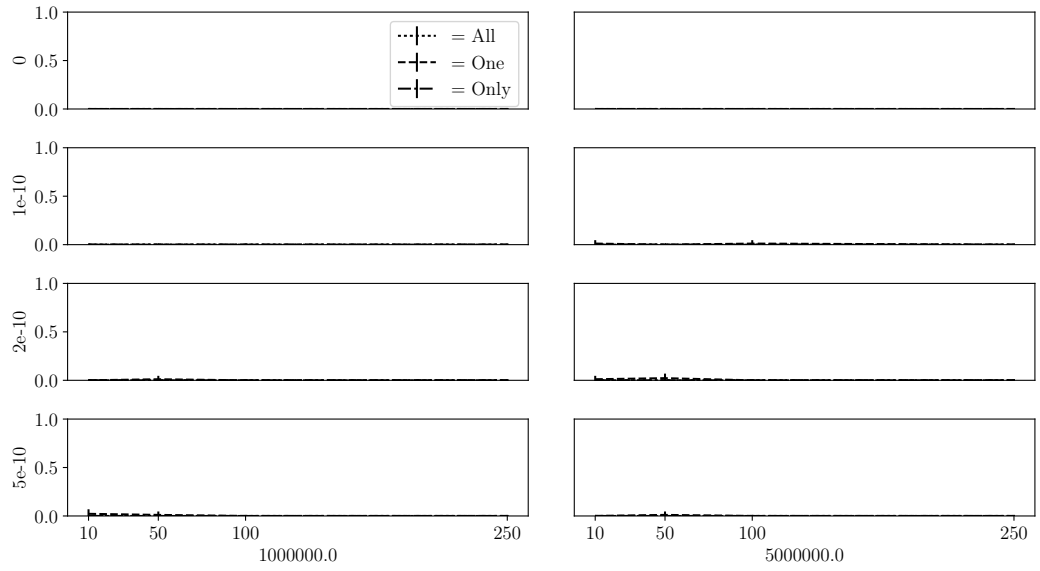

FIG. 3. The RF distance between the true species tree and the inferred ones from true gene trees of the 12-taxon simulated fly gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

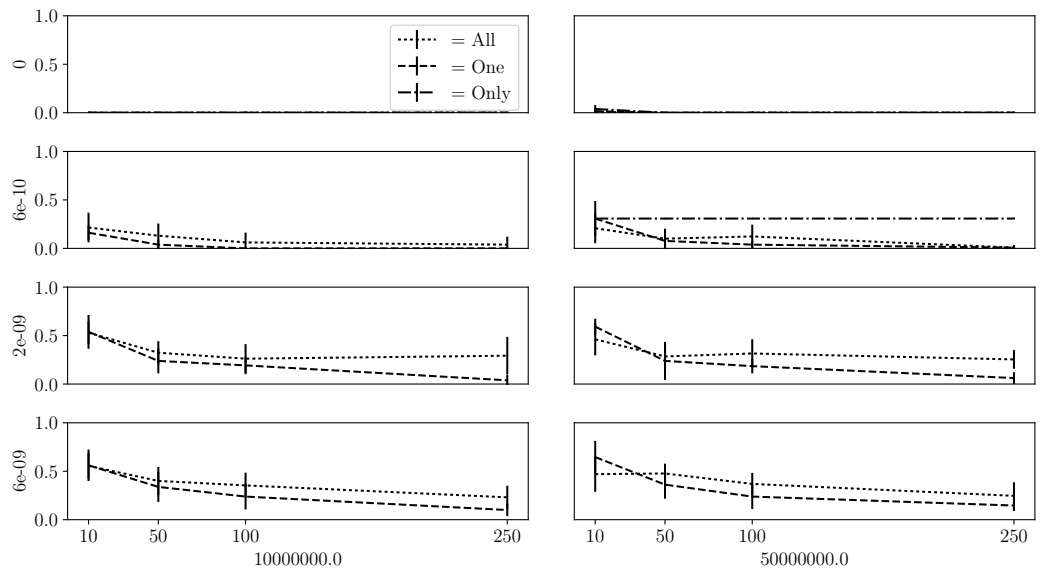

FIG. 4. The RF distance between the true species tree and the inferred ones from gene tree estimate of the 16-taxon simulated yeast gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

tree estimation error, and not hidden paralogy. In the case of the 16-taxon data, the error rate in the gene tree estimates was about 10%.

Finally, we ran ASTRAL on the locus trees as input to infer species trees. By the three-tree model, this amounts to feeding ASTRAL gene trees whose incongruence is solely due to GDL; that is, ILS plays no role in incongruence here. The results or the full 16-taxon and 12-taxon data sets are shown in Fig. 6 and 7, respectively.

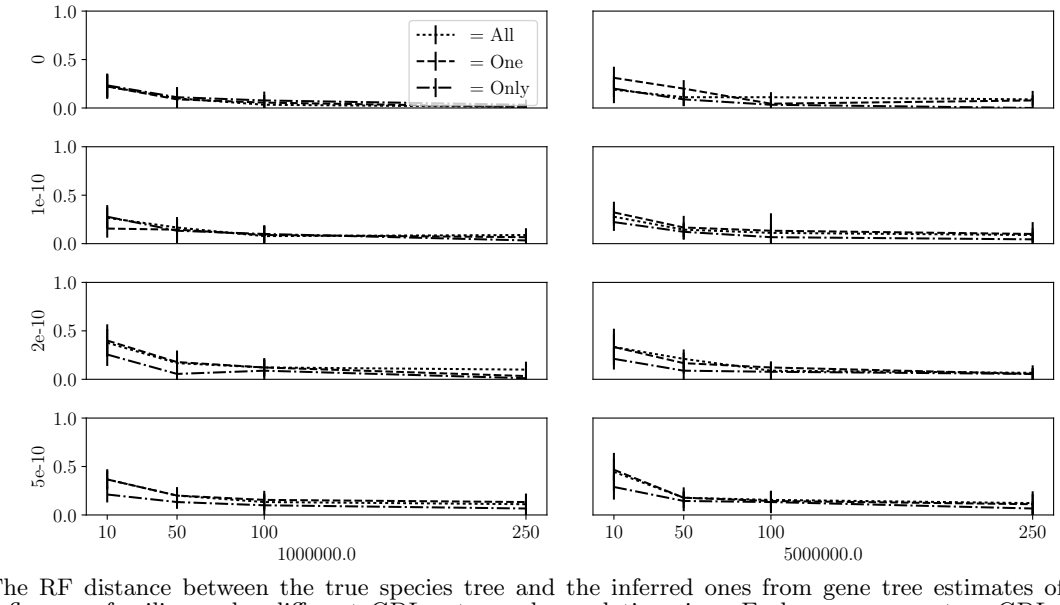

FIG. 5. The RF distance between the true species tree and the inferred ones from gene tree estimates of the 12-taxon simulated fly gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

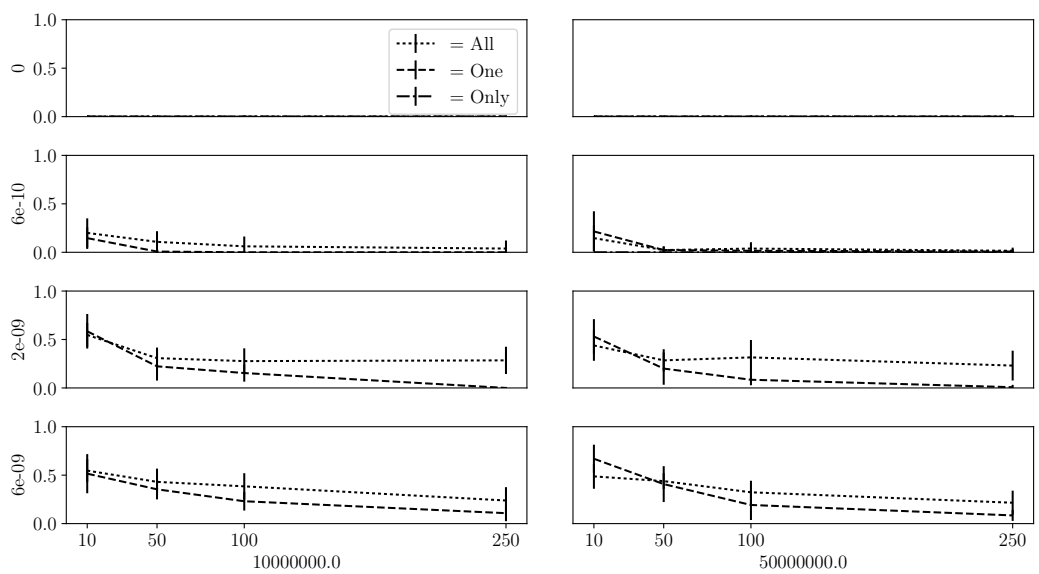

FIG. 6. The RF distance between the true species tree and the inferred ones from true locus trees of the 16 yeast genomes under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

The results are very striking. In the case of the 12-taxon data, ASTRAL almost always returns the correct species tree even though ILS is not present and paralogs are present in the data, and regardless of the population size and GDL rates. For the 16-taxon data, choosing one gene copy per species at random from each gene family results in very accurate species trees that converge onto the true one, as the number of gene family used increases. In fact, for the lowest non-zero

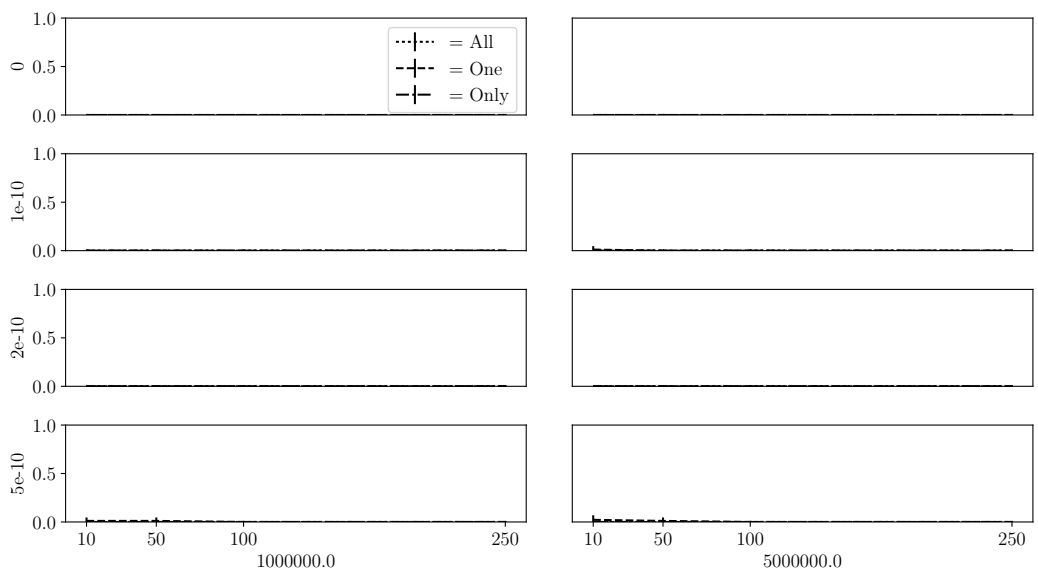

FIG. 7. The RF distance between the true species tree and the inferred ones from true locus trees of the 12 fly genomes under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

GDL rates, ASTRAL(ALL) and ASTRAL(ONE) almost always infer the correct species tree, except when only 10 gene families are used. As the GDL rates increase, ASTRAL(ONE) outperforms ASTRAL(ALL). It is important to highlight here that there were almost no single-copy genes in this data set, so ASTRAL(ONLY) had hardly any data in the input.

Results on 4- and 8-taxon data sets are shown in Supplementary Material.

Results on Biological Data

We randomly selected 250 gene trees estimated with PhyML (Guindon and Gascuel, 2003) reported in http://compbio.mit.edu/dlcoal/ for the 16 fungi genomes and estimated with RAxML (Stamatakis, 2006) reported in http://www.treefam.org for the 12 fly genomes for the ALL, ONE and ONLY settings . Then we estimated the species trees using ASTRAL with these gene trees as input. For the 16 fungi genome empirical data set, using ONLY, ASTRAL inferred a tree that differs from the "true" tree in only the grouping of Cgla and Scas, as shown in Fig. 1C. When using ALL and ONE, ASTRAL inferred the same tree, which differed from the true tree in the Cgla and Scas grouping, as in the case of ONLY, and in the grouping of Agos and Kwal, as shown in Fig. 1D.

For the 12 fly genome empirical data set, using method ALL and ONLY, ASTRAL inferred the exact same tree as the true species tree. When using ONE, the tree inferred by ASTRAL grouped Dmel with (Dere,Dyak) instead of (Dsec,Dsim), as shown in Fig. 1E.

Discussion

On data sets where only ILS was at play, and GDL was not a factor, ASTRAL had the expected performance: Very accurate species tree

estimates that were almost always identical to the true species tree. The only exception was when ASTRAL was run on the gene tree estimates for the 12-taxon simulated fly gene families; in this case, the species tree error was due to a high error rate in the gene tree estimates.

In the cases where both ILS and GDL were at play, ASTRAL(ONE) consistently inferred a very accurate species tree, implying that randomly choosing a single copy per species per gene family and running an MSC-based species tree inference method is a good practice. This is especially important in light of the fact (supported by the results above) that very few, if any, single-copy gene family trees are present for high GDL rates and/or large data sets in terms of the number of species.

A similar striking result was obtained when ASTRAL was run on the locus tree data. In this case, ILS was not at play and the data consisted of many gene families with multiple copies. ASTRAL obtained very accurate species trees, and the accuracy improved significantly, reaching almost 100% accuracy as 250 gene family trees were used as input. In particular, ASTRAL(ONE) performed better than ASTRAL (ALL). Once again, there were hardly any single-copy gene family trees in this case.

In analyses of the two biological data sets, the same performance emerged: Running ASTRAL on ALL and ONE hardly made a difference as compared to running it on ONLY. All these results point to a clear message: Under the conditions of our simulations and on the two biological data we used, running MSC-based species tree inference on data sets with paralogs yields very accurate results. This conclusion is powerful for at least two reasons. First, it implies that orthology assignment and paralogy removal are not necessary for running MSC-based species tree inference; simply treating all copies as different individuals or randomly selecting a single copy would yield very accurate species tree topologies. Second, in many practical cases, too few single-copy gene families would be available to ensure good performance of species tree inference from those data alone.

While inspecting the branch lengths of the estimated species trees for the two biological data sets, no clear pattern emerged other than the three methods of selecting the input could yield wildely different branch lengths (as estimated by ASTRAL, which, by itself, is not ideal for estimating species tree branch lengths).

The results point to the following potential method for inferring accurate species trees (topologies and branch lengths):

1. Use all available gene family trees, while treating all copies from the same species within a family as multiple individuals from the same species, or while randomly selecting a copy per species.

2. Feed all gene trees to an MSC-method such as ASTRAL and obtain a species tree topology.

3. Using the full sequence data along with a statistical model that combines the MSC with duplication and loss (e.g., that of Rasmussen and Kellis (2012)) and fixing the species tree topology obtained from Step (2), sample or optimize the branch lengths of the species tree.

For Steps (1) and (2), one option is to infer a species tree based on ALL, and repeat the random sampling of single copies to obtain multiple versions of the input under ONE, and then score all the inferred species trees under some criterion that combines the MSC with a model of gene duplication/loss. This overcomes the issue of fixing a single species tree as input to Step (3), and avoids searching the species tree space while evaluating a likelihood function that is very complex and computationally very demanding to compute. Furthermore, Step (3) allows for paralogy detection and orthology assignment, conditional on the fixed species tree (or species trees), by using the more detailed evolutionary model and the full signal in the sequence data. For example, the orthology assignment could be "integrated out" or sampled, depending on the desired outcomes of the analysis.

Concluding Remarks

In this paper we set out to study how MSC-based species tree inference would perform on data with paralogs. The motivation for exploring this question was two-fold. First, as MSC-based species tree inference has become commonplace and as practitioners are almost never certain that their data contain no paralogs, it is important to understand the effect of such hidden paralogy on the quality of the inference. Second, as larger phylogenomic data sets become available, insistence on single-copy genes would mean throwing away most of the data and potentially keeping a number of loci that is severely inadequate for the data requirement for MSC-based inference methods to perform well. We investigated the question through a combination of simulations and biological data analyses. Our results show that MSC-based inference is robust to the presence of paralogs in the data, at least under the simulation conditions and on the empirical data sets we investigated.

In this paper we used ASTRAL as the candidate for MSC-based species tree inference methods. ASTRAL has already been shown to be accurate and scalable (while the number of species we considered is relatively small, the number of taxa in gene trees can be much larger due to gene duplication). Furthermore, the results in our analyses above indicate that not much improvement can be gained by other methods as

in most cases ASTRAL obtains the correct species tree.

Our results highlight the issue that MSC-based inference could result in very accurate species trees even when ILS is not a factor or not the only factor. This observation calls for caution against invoking ILS as the source of heterogeneity simply because MSC-based inference yielded the correct species tree. Furthermore, one should avoid the circular argument of invoking extensive ILS based on the short branch lengths estimated by methods that only account for ILS.

In our simulations, we generated gene families under a neutral model and with GDL rates that were the same across all families. It is well known that the functional implications of gene duplication and the ways in which they are fixed and maintained in the genome result in much more complex scenarios (Innan and Kondrashov, 2010) than those captured in our simulations. However, analyses of the two biological data sets yield results with very similar trends to those observed in our simulations. Furthermore, the assumptions underlying the MSC model too are undoubtedly violated in most empirical phylogenomic analyses. Another direction for future research entails characterizing the conditions under which species tree inference under the MSC is robust to the presence of paralogs in the data.

Finally, while we did not discuss or incorporate gene flow in our study, it is possible that all three processes—ILS, GDL, and gene flow—are simultaneously at play in some phylogenomic analyses. Studies of the robustness of MSC-based species tree inference under some models of gene flow exist (Davidson *et al.*, 2015; Roch and Snir, 2012; Solís-Lemus *et al.*, 2016; Steel *et al.*, 2013; Zhu *et al.*, 2016), but, to the best of our knowledge, such studies under scenarios that incorporate all the aforementioned processes do not exist yet. It is important to highlight, as well, that great strides have been made in developing methods for phylogenetic network inference in the presence of ILS (Elworth *et al.*, 2018), but none currently incorporate gene duplication and loss. We believe methods along the lines described in the previous section could be promising for accurate and scalable phylogenomic inferences without sacrificing much of the data.

Supplementary Material

Supplementary material are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported in part by NSF grants DBI-1355998, CCF-1302179, CCF-1514177, CCF-1800723, and DMS-1547433.

References

Davidson, R., Vachaspati, P., Mirarab, S., and Warnow, T. 2015. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genomics*, 16(10): S1.

Degnan, J. H. and Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies

coalescent. *Trends in Ecology & Evolution*, 24(6): 332–340.

Du, P. and Nakhleh, L. 2018. Species tree and reconciliation estimation under a duplication-loss-coalescence model. *Proceedings of the 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 376–385.

Elworth, R. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. 2018. Advances in computational methods for phylogenetic networks in the presence of hybridization. *arXiv preprint arXiv:1808.08662*.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y.-C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W., and Besansky, N. J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217): 1258524.

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5): 696–704.

Hahn, M. W., Han, M. V., and Han, S.-G. 2007. Gene family evolution across 12 drosophila genomes. *PLoS genetics*, 3(11): e197.

Innan, H. and Kondrashov, F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2): 97.

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., and Zhang, G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215): 1320–1331.

Jukes, T. H., Cantor, C. R., *et al.* 1969. Evolution of protein molecules. *Mammalian protein metabolism*, 3(21): 132.

Knowles, L. L. and Kubatko, L. S. 2011. *Estimating species trees: practical and theoretical aspects*. John Wiley and Sons.

Lang, G. I. and Murray, A. W. 2008. Estimating the per-base-pair mutation rate in the yeast saccharomyces cerevisiae. *Genetics*, 178(1): 67–82.

Liu, L., Yu, L. L., Kubatko, L., Pearl, D. K., and Edwards, S. V. 2009. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.*, 53: 320–328.

Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. 2015. Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*, 1360(1): 36–53.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology*, 46(3): 523–536.

Nakhleh, L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology and Evolution*, 28(12): 719–728.

Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. 2006. Widespread discordance of gene trees with species tree in drosophila: evidence for incomplete lineage sorting. *PLoS genetics*, 2(10): e173.

Rambaut, A. and Grass, N. C. 1997. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3): 235–238.

Rannala, B. and Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4): 1645–1656.

Rasmussen, M. D. and Kellis, M. 2010. A bayesian approach for fast and accurate gene tree reconstruction. *Molecular Biology and Evolution*, 28(1): 273–290.

Rasmussen, M. D. and Kellis, M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4): 755–765.

Robinson, D. F. and Foulds, L. R. 1981. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2): 131–147.

Roch, S. and Snir, S. 2012. Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. *Journal of Computational Biology*, 20(2): 93–112.

Salichos, L. and Rokas, A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449): 327.

Sawyer, S. A. and Hartl, D. L. 1992. Population genetics of polymorphism and divergence. *Genetics*, 132(4): 1161–1176.

Solís-Lemus, C., Yang, M., and Ané, C. 2016. Inconsistency of species-tree methods under gene flow. *Systematic biology*. 10.1093/sysbio/syw030.

Stamatakis, A. 2006. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21): 2688–2690.

Stamatakis, A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

Steel, M., Linz, S., Huson, D., and Sanderson, M. 2013. Identifying a species tree subject to random lateral gene transfer. *Journal of Theoretical Biology*, 322: 81–93.

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122(4): 957–966.

Zhang, B. and Wu, Y.-C. 2017. Coestimation of gene trees and reconciliations under a duplication-loss-coalescence model. In *International Symposium on Bioinformatics Research and Applications*, pages 196–210. Springer.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. 2018. Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics*, 19(6): 153.

Zhu, J., Yu, Y., and Nakhleh, L. 2016. In the light of deep coalescence: revisiting trees within networks. *BMC Bioinformatics*, 17(14): 415.