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

Peng Du[1], Matthew W. Hahn[2], and Luay Nakhleh[*,1,3]
[1]Department of Computer Science
[2]Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN
[3]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 ILS-aware inference methods, whether based directly on the MSC or proven to be statistically consistent under it (collectively referred to here as MSC-based methods), to exclude all loci that are present in more than a single copy in any of the studied genomes. Furthermore, such analyses entail 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 such species tree inference methods 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 such methods 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. Our results have two significant implications. First, MSC-based phylogenomic analyses do not have to be restricted to single-copy loci, thus maximizing the amount of data to be used. Second, when an MSC-based inference method infers an accurate species tree, this does not necessarily imply that ILS was the main 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 can 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; Hudson, 1983; 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 (Koonin, 2005). 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 arising from gene duplication and loss. An obvious way to handle data sets where ILS and gene duplication/loss could have simultaneously acted on the gene families 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 infeasible 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 four species tree inference methods, all of which use gene trees as the input data: The maximum likelihood (and pseudo-likelihood) method of Yu and Nakhleh (2015); Yu *et al.* (2014) as implemented by the functions `InferNetwork_ML` and `InferNetwork_MPL` in PhyloNet (Wen *et al.*, 2018), while restricting the search to the tree space (i.e., not allowing reticulations), ASTRAL-III (Zhang *et al.*, 2018), and NJst (Liu and Yu, 2011). It is important to point out that `InferNetwork_MPL` restricted to the tree space amounts to running the species tree inference method MP-EST (Liu *et al.*, 2010). Furthermore, for the simulated data, since gene tree branch lengths are available, we use the maximum likelihood method `InferNetwork_ML`. However, for the two biological data sets, we use `InferNetwork_MPL` since it scales better computationally to large data sets when only gene tree topologies are considered.

For the sake of conclusions that we draw from this study, it is important to highlight the difference between `InferNetwork_ML` and `InferNetwork_MPL` on the one hand and ASTRAL and NJst on the other hand. The former two optimize likelihood and pseudo-likelihood functions that are derived based on the assumptions of the MSC. Such likelihood functions are very different, for example, from a likelihood function that is based on a birth-death model of gene family evolution (Arvestad *et al.*, 2009). Therefore, their accuracy in inferring species trees from data with paralogs reflects directly on the performance of MSC-based methods on such data. The latter two methods make no direct use of the MSC, but have been shown to be statistically consistent under the MSC. Therefore, their accuracy on data with paralogs reflects on the suitability of these methods, rather than the MSC itself, for analyzing such data.

As these methods assume only orthologs are used for each locus—and to facilitate their use on data with paralogs—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 the single-copy genes are orthologs (in fact, some would not be with very high probability). Our results show that inferences made by all methods are very accurate, and are mostly identical to the accuracy of their inferences when using only genes that are present in exactly one copy in each of the species. Particularly striking is the finding that these methods infer very accurate species trees when all gene tree incongruence is due to GDL, and ILS is not a factor.

Materials and Methods

Species Tree Inference Methods

For species tree inference, we use four methods that are statistically consistent under the MSC. The first two methods are maximum likelihood inference from gene trees. On the simulated data sets, we used the `InferNetwork_ML` function in PhyloNet (Wen *et al.*, 2018), which implements the maximum likelihood inference method of Yu *et al.* (2014), while setting the maximum number of reticulations to 0. This effectively searches for the maximum likelihood species tree under the MSC by making use of the gene trees along with their branch lengths. However, in the case of biological data sets, we used the maximum pseudo-likelihood inference function `InferNetwork_MPL` in PhyloNet, which implements the method of Yu and Nakhleh (2015) and amounts to running MP-EST (Liu *et al.*, 2010) when restricted to trees with no reticulations. There are two reasons as to why we switched to maximum pseudo-likelihood in the case of biological data sets. On the one hand, we use gene trees that were estimated previously, and accurate branch lengths are not available. On the other hand, while the method of Yu *et al.* (2014) can be run on gene tree topologies alone, the computation becomes prohibitive when the gene trees are large (in terms of the number of tips). The other two methods are ASTRAL-III (Zhang *et al.*, 2018), Version 5.6.3, and NJst (Liu and Yu, 2011), neither of which uses the MSC directly, but both have been shown to yield accurate estimates of species trees from limited amount of gene tree data where ILS is the sole (biological) cause of incongruence.

Given a collection of trees corresponding to gene families (one tree per gene family), we generated three types of input to each of the methods:

- 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, as it reduces, though not necessarily completely eliminates, the potential of hidden paralogy—a term that was introduced by Daubin and Gouy (2001); Gribaldo and Philippe (2002) to denote cases where single-copy paralogs within a gene family are mistaken

for orthologs. Unless GDL is not involved, 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. The 16 fungi species are: *Candida alibicans* (calb), *Candida tropicalis* (ctro), *Candida parapsilosis* (cpar), *Lodderomyces elongisporus* (lelo), *Candida guilliermondii* (cgui), *Debaryomyces hansenii* (dhan), *Candida lusitaniae* (clus), *Saccharomyces cerevisiae* (scer), *Saccharomyces paradoxus* (spar), *Saccharomyces mikatae* (smik), *Saccharomyces bayanus* (sbay), *Candida glabrata* (cgla), *Saccharomyces castellii* (scas), *Kluyveromyces latics* (klac), *Ashbya gossypii* (agos), and *Kluyveromyces waltii* (kwal).

The 12 fly species are: *Drosophila melanogaster* (dmel), *Drosophila simulans* (dsim), *Drosophila sechellia* (dsec), *Drosophila erecta* (dere), *Drosophila yakuba* (dyak), *Drosophila ananassae* (dana), *Drosophila pseudoobscura* (dpse), *Drosophila persimilis* (dper), *Drosophila willistoni* (dwil), *Drosophila mojavensis* (dmoj), *Drosophila virilis* (dvir), and *Drosophila grimshawi* (dgri).

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 ILS, but not GDL, acted on the gene families), 1×10^{-10}, 2×10^{-10}, and 5×10^{-10} per generation where 1×10^{-10} is similar to the duplication rate of 7.32×10^{-11} and loss rate of 8.59×10^{-11} 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 nucleotide 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. We also assumed that gene conversion occurred and set the gene conversion rate to be 50X the mutation rate, similar to that reported in (Mansai *et al.*, 2011), and which is a conservative value in terms of its effect on gene tree inference.

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

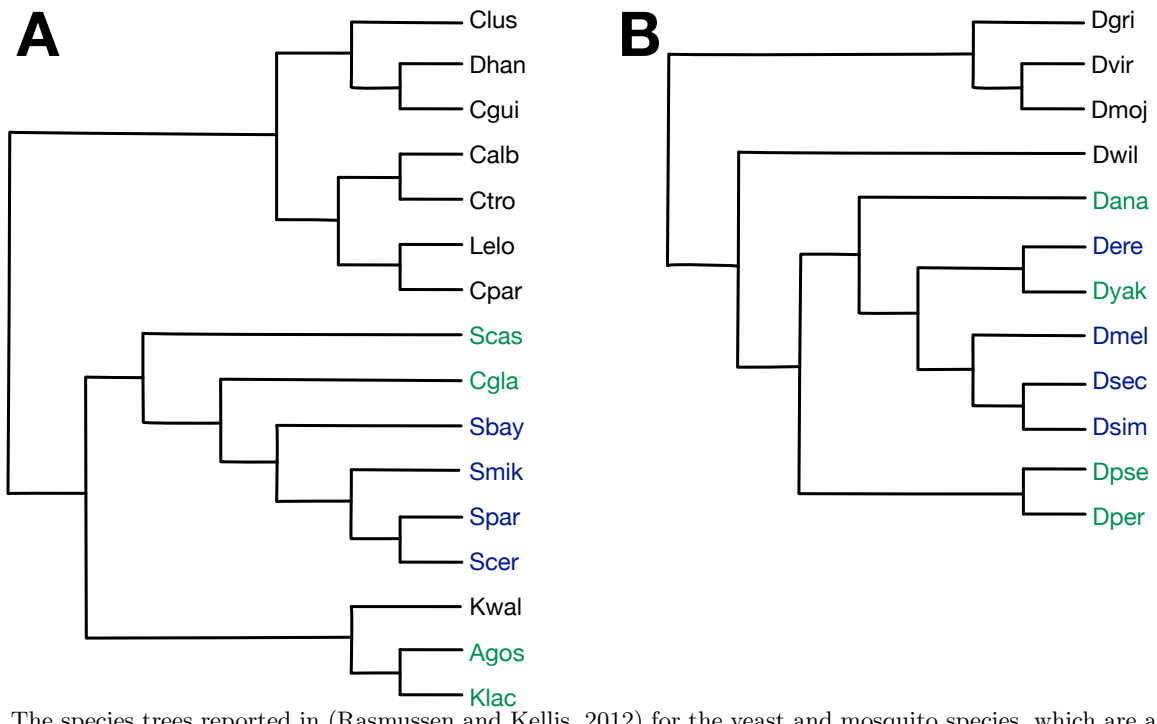

FIG. 1. The species trees reported in (Rasmussen and Kellis, 2012) for the yeast and mosquito species, which are also used as the true species trees in the simulations. (A) The species tree of 16 fungi species. (B) The species tree of 12 fly species. For both species trees, trees on 4 and 8 species were obtained by restricting the tree to the 4 leaves colored in blue, and the 8 leaves colored in blue and green. The species tree topologies and their branch lengths in units of million years are reported on http://compbio.mit.edu/dlcoal/.

population sizes of 10^6 and 5×10^6, similar to the values used in (Rasmussen and Kellis, 2012) and estimated value of 1.15×10^6 reported in (Pollard *et al.*, 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 (Schrider *et al.*, 2013). We also assumed that gene conversion occurred and set its rate to be 50X the mutation rate, similar to that reported in (Arguello *et al.*, 2006; Thornton and Long, 2004).

To generate gene trees while allowing for ILS, GDL, and gene conversion, we used SimPhy (Mallo *et al.*, 2015) with the parameters specified above (assuming all species are diploid). SimPhy uses the three-tree model developed in (Rasmussen and Kellis, 2012) to simulate data. 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 and gene conversion. 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 and gene conversion (but not ILS). Setting the rates of GDL to 0.0 amounts to generating gene trees where all incongruence is solely due to ILS. 10,000

gene families (each containing a locus tree and corresponding gene tree) 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 study the effect of using data sets of varying sizes, for each of the 10 data sets, 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). Finally, for each data set of trees (true gene trees or true locus trees, that is, tree without estimation error) of a given size, we fed the data set as input to InferNetwork_ML, ASTRAL, and NJst 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.

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, NJst, and maximum pseudo-likelihood with these gene trees as input.

Given the large computational requirements (time and memory) of species tree inference under maximum likelihood from gene tree topologies alone, we used maximum pseudo-likelihood inference, as implemented by the function InferNetwork_MPL in PhyloNet (Wen *et al.*, 2018; Yu and Nakhleh, 2015), while setting the maximum number of reticulations to 0 so that the method searches the tree space only. As pointed out above, this amounts to optimizing the same function like MP-EST (Liu *et al.*, 2010). Notice that in the case of the simulated data sets, it was computationally feasible to infer species trees under the full likelihood function (InferNetwork_ML) because gene tree branch lengths were used and that makes the likelihood calculations orders of magnitude faster than when only gene tree topologies are used. However, it has been shown that methods that rely on gene tree branch length estimates have poorer performance than those that make use of only gene tree topologies (DeGiorgio and Degnan, 2013), which is why we use InferNetwork_MPL on the gene tree topologies alone rather than InferNetwork_ML on the gene trees with estimated branch lengths.

Furthermore, InferNetwork_MPL requires that the input gene trees are rooted. Rooting gene trees is a challenging task in the presence of paralogs, in particular, when gene duplication events occurred

prior to the most recent common ancestor of all species under study. In this study, we rooted each gene tree with respect to the species tree of Fig. 1A so as to minimize deep coalescences (Maddison, 1997; Than and Nakhleh, 2009) using the method of Yu *et al.* (2011a, b), as implemented by the function ProcessGT in PhyloNet (Wen *et al.*, 2018).

Results

Characteristics of the Simulated 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. Figure 2 shows data on the average size (number of copies) of gene families 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 larger the gene families. The figure also shows that the median size of a gene family is roughly equal to the number of species, with the largest gene families ranging between 60 and 120 for the different GDL rate settings of the 16-taxon data sets, and between 30 and 55 for the different GDL rate settings of the 12-taxon data sets. 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 1.

The table shows 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, 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. 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 2,000 single-copy gene families (for the largest GDL rates considered). However, when considering the full 16-species data set, the full extent of complexity emerges, as almost *all* of these 2,000 gene families underwent duplication and loss, and almost none of them is present in single copy.

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.

Results on Simulated Data

We are now in position to describe the inference results. We first ran the three inference methods, InferNetwork_ML, ASTRAL, and NJst, 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 of InferNetwork_ML on the full 16-taxon and 12-taxon data sets are shown in Fig. 3 and Fig. 4, respectively. Results of ASTRAL

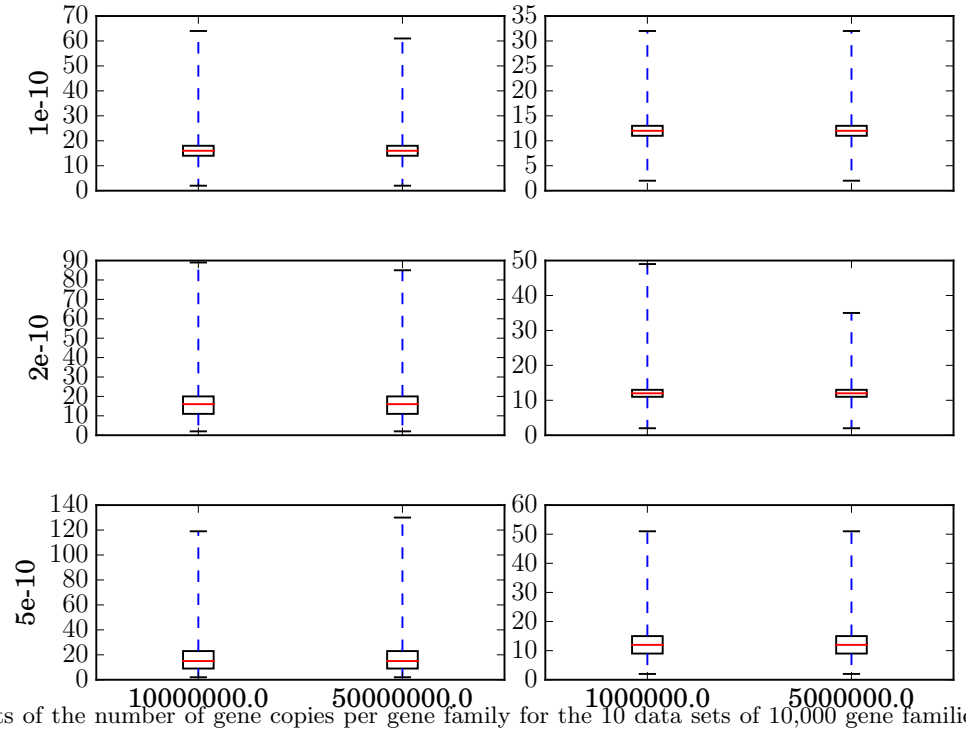

FIG. 2. Box plots of the number of gene copies per gene family for the 10 data sets of 10,000 gene families each, for each setting of effective population size, duplication and loss rates. From top to bottom: GDL rates 1×10^{-10}, 2×10^{-10}, and 5×10^{-10}, respectively. Left and right columns correspond to the 16-taxon and 12-taxon data sets, respectively. The two values on the x-axis in each panel correspond to the two effective population sizes considered.

Table 1. 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 → GDL rate ↓	10^7	5×10^7	10^7	5×10^7	10^7	5×10^7	N_e → GDL rate ↓	10^6	5×10^6	10^6	5×10^6	10^6	5×10^6
1×10^{-10}	894.9	889.1	3518.4	3493.9	6662.1	6696.7	1×10^{-10}	4601.7	4575.5	6735.3	6714.2	8502.9	8498.5
2×10^{-10}	92.0	94.3	1292.6	1327.4	4658.7	4694.0	2×10^{-10}	2180.6	2176.4	4621.7	4604.0	7284.2	7279.6
5×10^{-10}	0.4	0.1	104.1	107.3	1970.9	1951.8	5×10^{-10}	271.4	267.6	1611.5	1601.6	4735.5	4703.4

and NJst for the full 16-taxon and 12-taxon data sets are shown in Fig. 5 and Fig. 6, respectively. Note that in all 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, InferNetwork_ML always infers the correct species tree (Fig. 4). In the case of the simulated yeast data, InferNetwork_ML infers the correct species tree in all cases for the smaller of the two population sizes (except for one case under the highest GDL rate and when only 10 gene family trees are used) and consistently infers the correct species tree under the larger population size when 50 or more gene family trees are used (Fig. 3).

In the case of the simulated yeast data, ASTRAL and NJst have higher error rates than those of InferNetwork_ML. In particular, when

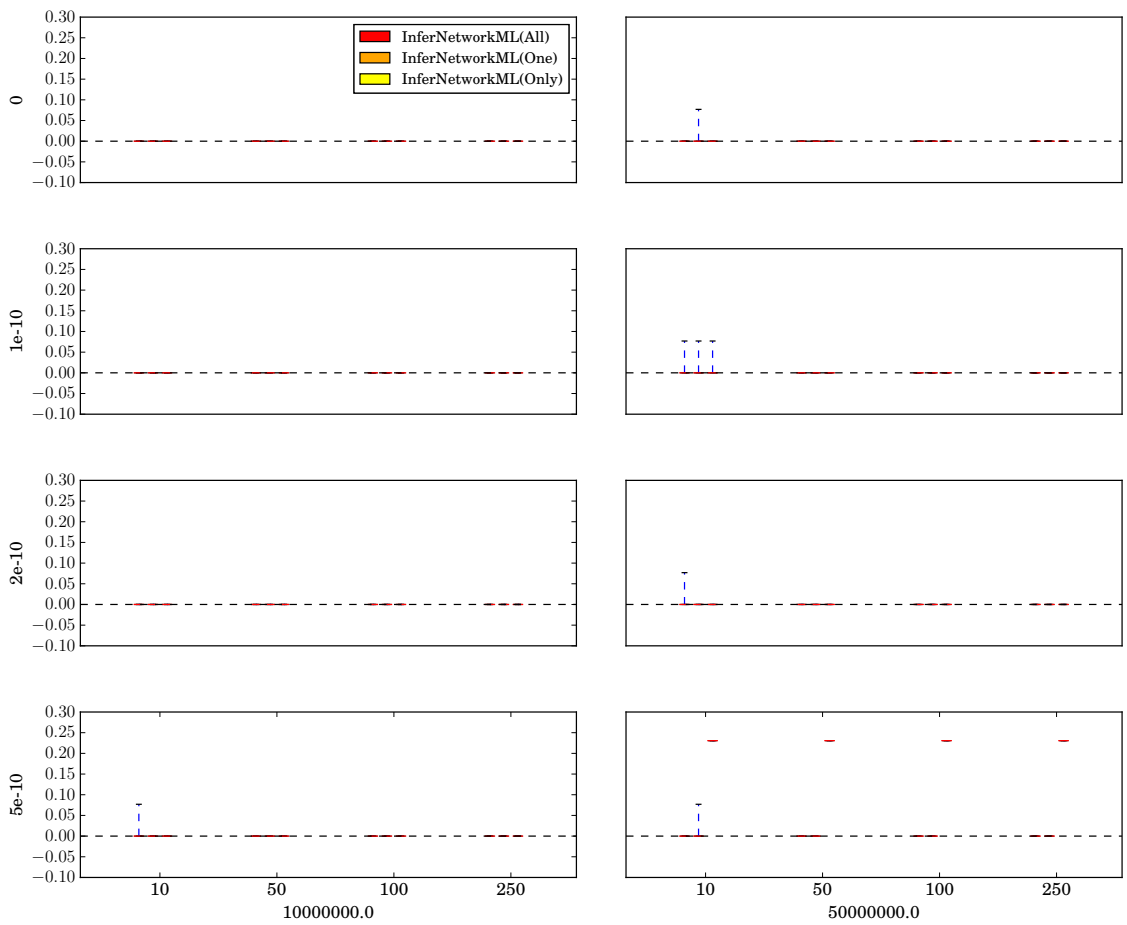

FIG. 3. The RF distance between the true species tree and the ones inferred by maximum likelihood from true gene trees of the 16-taxon simulated yeast gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

only 10 gene family trees are used as input, both methods have over 5% error under the larger of the two population sizes (Fig. 5). As more gene family trees are used, the performance improves significantly. In the case of the simulated fly data, ASTRAL and NJst almost always infer the correct species tree, regardless of how the input gene families were selected (ALL, ONLY, or ONE), except when using only 10 gene families as input under the highest GDL rates used (Fig. 6).

We then ran `InferNetwork_ML`, ASTRAL and NJst on the locus trees as input to infer species trees. By the three-tree model, this amounts to feeding these methods gene trees whose incongruence is solely due to GDL and gene conversion; that is, ILS plays no role in incongruence here. The results of the `InferNetwork_ML` inferences on the full 16-taxon and 12-taxon data sets are shown in Fig. 7 and Fig. 8, respectively. The results of the ASTRAL

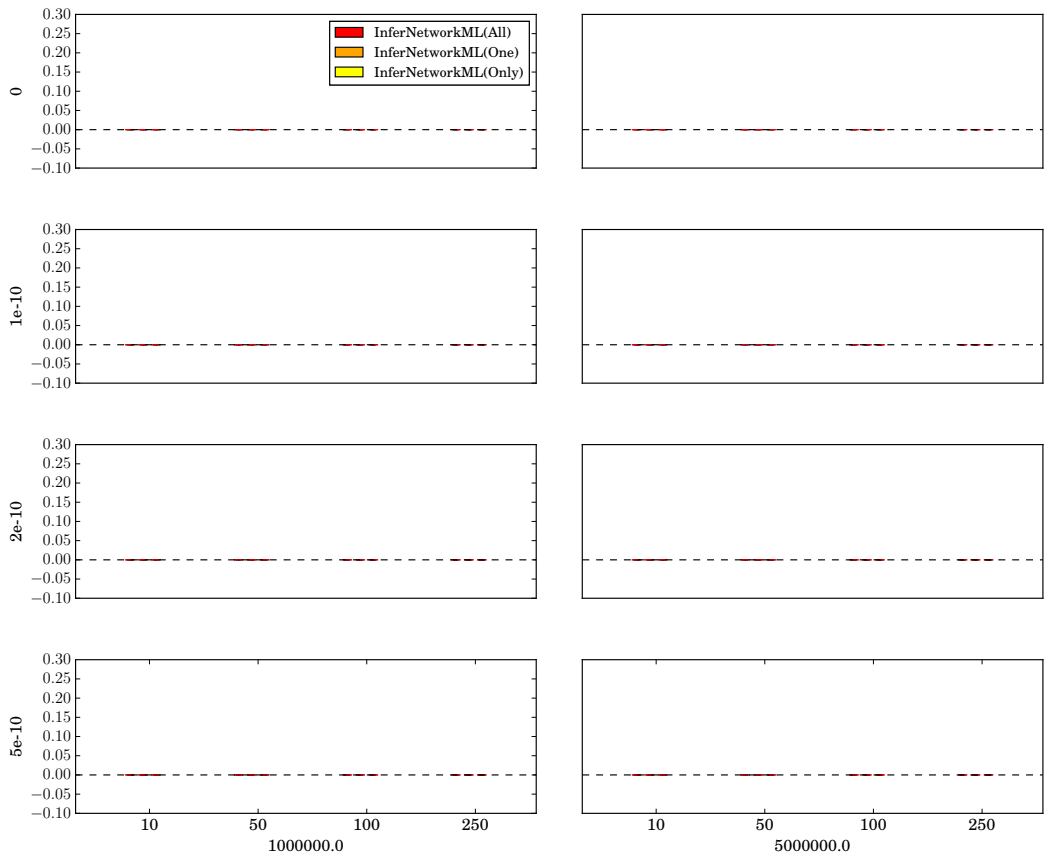

FIG. 4. The RF distance between the true species tree and the ones inferred by maximum likelihood from true gene trees of the 12-taxon simulated fly gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

and NJst inferences on the full 16-taxon and 12-taxon data sets are shown in Fig. 9 and Fig. 10, respectively. The results are very striking. The MSC method InferNetwork_ML infers the correct species tree in all cases on the fly data sets (Fig. 8) and all except for two cases of 10 gene families and the highest GDL rates on the yeast data sets (Fig. 7) . For the larger of the two population sizes on the yeast data sets, ASTRAL and NJst have higher accuracies than in the case where both ILS and GDL were involved (Fig. 9). In particular, when 100 or 250 gene family trees were used, both methods inferred the correct species tree almost on all data sets. similar performance in this case. In the case of the 12-taxon data, the methods' performances become slightly worse in the case of the largest GDL rates (Fig. 10).

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

Results on Biological Data

We ran the three methods, InferNetwork_MPL, ASTRAL, and NJst, on the two biological data sets, and described above the reason for using

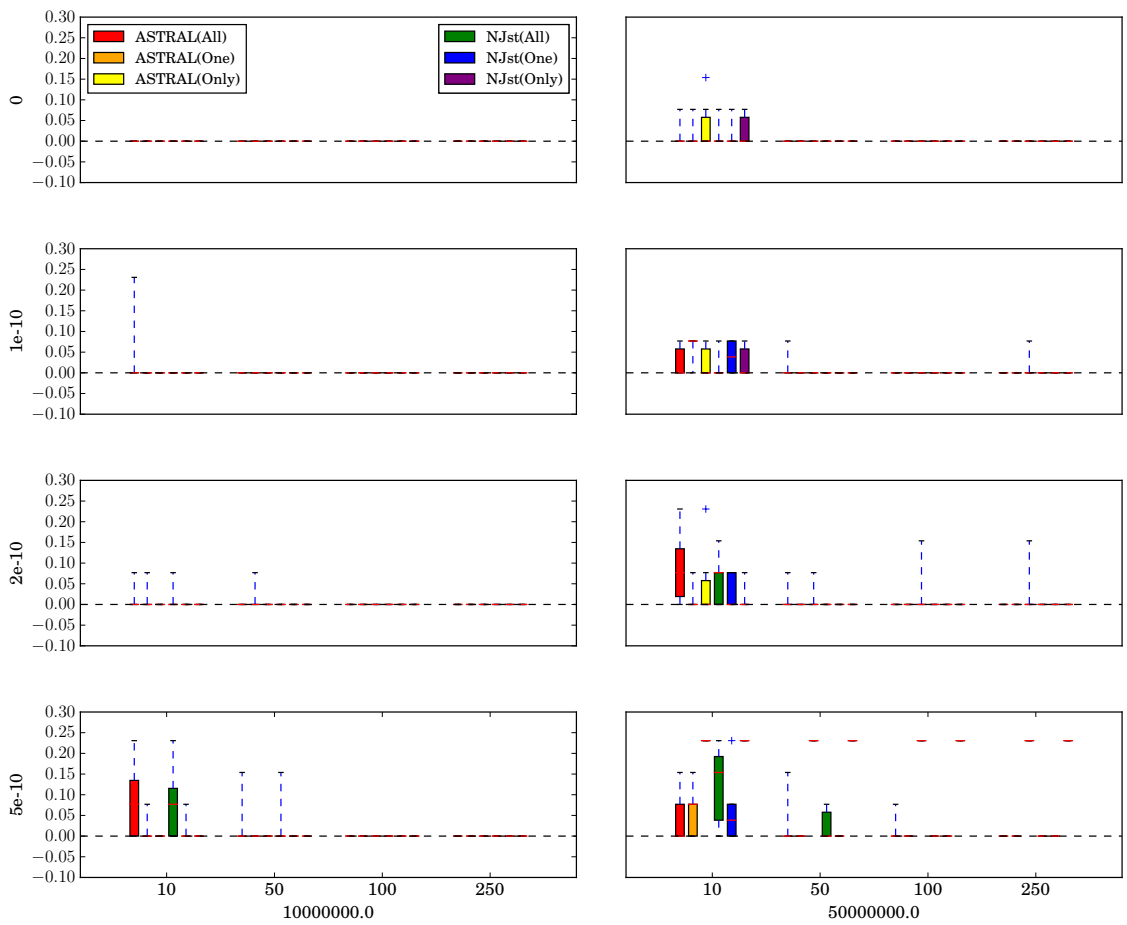

FIG. 5. The RF distance between the true species tree and the ones inferred by ASTRAL and NJst from true gene trees of the 16-taxon simulated yeast gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

InferNetwork_MPL instead of InferNetwork_ML in this case.

For the 16 fungi genome empirical data set, InferNetwork_MPL(ALL), InferNetwork_MPL(ONE), InferNetwork_MPL(ONLY), ASTRAL(ALL), ASTRAL(ONE), and ASTRAL(ONLY), all inferred the same tree, which differs from the tree shown in Fig. 1A in the grouping of *Ashbya gossypii*, *Kluyveromyces waltii*, and *Kluyveromyces latics* and the grouping of *Candida glabrata*, *Saccharomyces castellii*, and *Saccharomyces bayanus*, as shown in Fig. 11A. NJst, on the other hand, inferred an identical tree to that of Fig. 1A under all three input scenarios.

For the 12 fly genome empirical data set, ASTRAL(ONLY) inferred the exact same tree as the species tree shown in Fig. 1B. The trees inferred by ASTRAL(ALL) and ASTRAL(ONE) differed from the true tree of Fig. 1B in terms of the placement of *Drosophila melanogaster*, as shown in Fig. 11B and Fig. 11C, respectively.

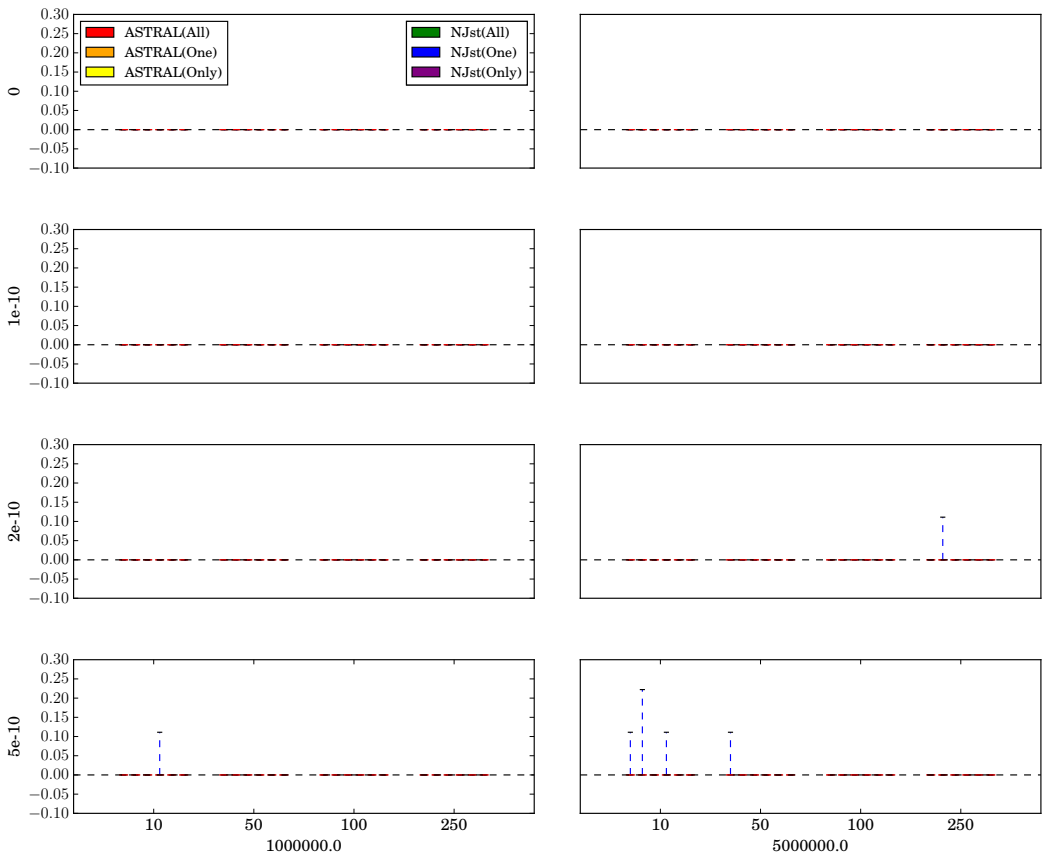

FIG. 6. The RF distance between the true species tree and the ones inferred by ASTRAL and NJst from true gene trees of the 12-taxon simulated fly gene families under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

InferNetwork_MPL and NJst inferred an identical tree to that of Fig. 1B under all three input scenarios.

Discussion

On simulated data sets where only ILS acted on gene families, and GDL was not a factor, all methods had the expected performance: Very accurate species tree estimates that were almost always identical to the true species tree. In the cases where both ILS and GDL acted on gene families, maximum likelihood inference always produced correct estimates of the species tree, and ASTRAL and NJst almost always inferred the correct species trees. These results imply that running these methods on data with paralogs (while treating them as alleles from different individuals) is a safe practice when the desired outcome is the species tree topology. This is especially important in light of the fact that very few single-copy gene family trees are present for high GDL rates and/or large data sets in terms of the number of species.

When the methods were run on the locus tree data, where ILS did not play a role

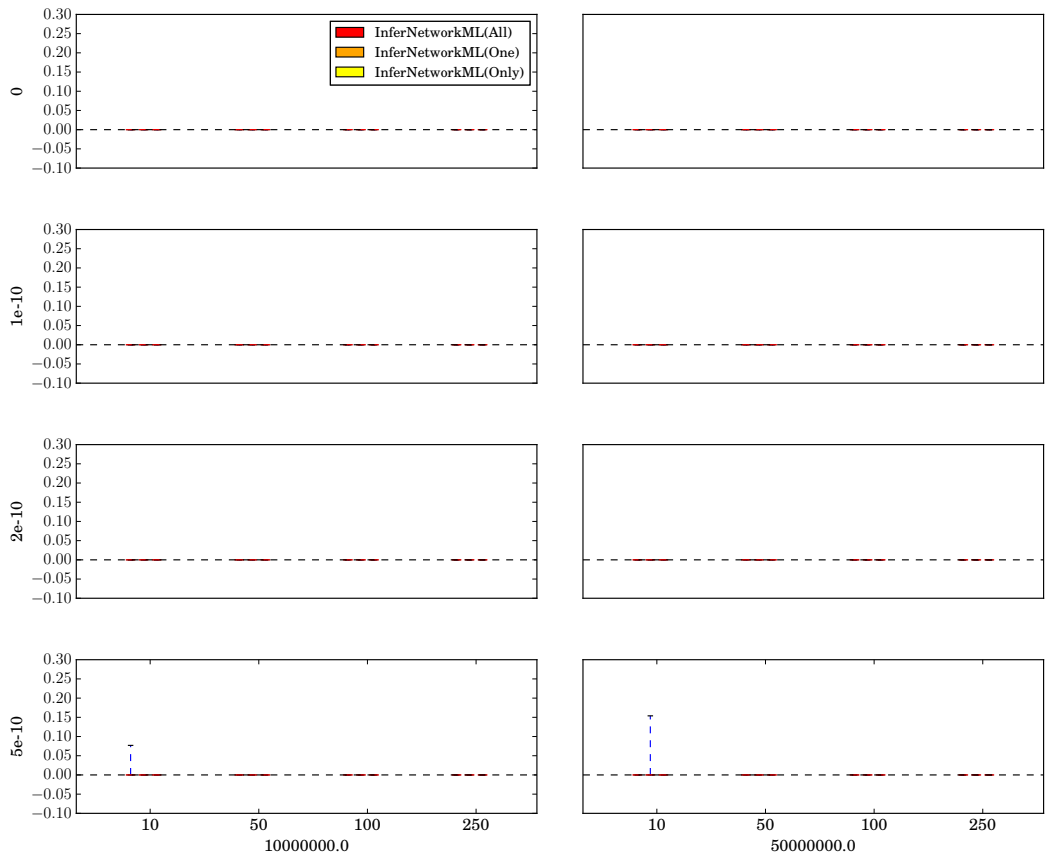

FIG. 7. The RF distance between the true species tree and the ones inferred by maximum likelihood from true locus trees of the 16 yeast genomes under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

and the data consisted of many gene families with multiple copies, the methods obtained very accurate species trees, with maximum likelihood inference almost always inferring the correct species tree, and the other two methods very close in terms accuracy (and their accuracy improved significantly, reaching almost 100% accuracy as 250 gene family trees were used as input).

While at first it may be surprising that these methods have done very well in terms of accuracy, the majority of signal in the input gene trees reflects species relationships. Gene duplication—if random across the species tree—simply adds noise to the data, while at the same time often doubling the amount of information on the relationships among species. Nevertheless, upon close inspection, some of these results are not intuitive, especially for the maximum likelihood inference. InferNetwork_ML makes direct use of the MSC, whose assumptions are clearly violated in all data sets except when the GDL rates are set to 0, whereas ASTRAL and NJst are summary methods that make no direct use

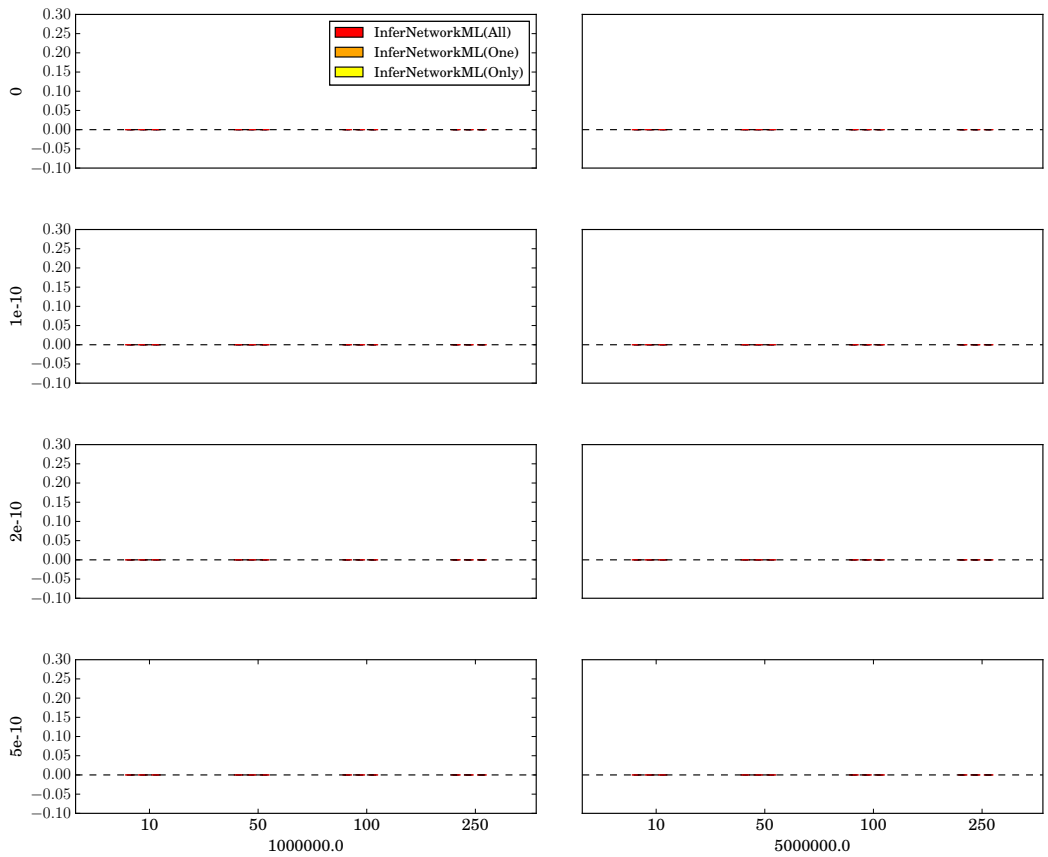

FIG. 8. The RF distance between the true species tree and the ones inferred by maximum likelihood from true locus trees of the 12 fly genomes under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

of the MSC. Consequently, one would have expected that InferNetwork_ML would be very sensitive to the presence of paralogs in the data, while ASTRAL and NJst less so. A potential explanation for this observation could be the fact that InferNetwork_ML in our results above makes use of the gene tree coalescent times, and those times provide strong signal for the speciation patterns even when GDL is present. ASTRAL and NJst, on the other hand, make use of only the gene tree topologies.

In the simulation results described above, we focused on the methods' performances on true gene trees (i.e., gene trees with no estimation error). In practice, gene trees are estimated from sequence data and could be erroneous. Error in the gene tree estimates, rather than ILS, could explain much of the heterogeneity in phylogenomic analyses (Scornavacca and Galtier, 2017). Therefore, to keep this study focused on the performance of methods in the presence of paralogy and set aside the issue of gene tree estimation error, we did not address the latter.

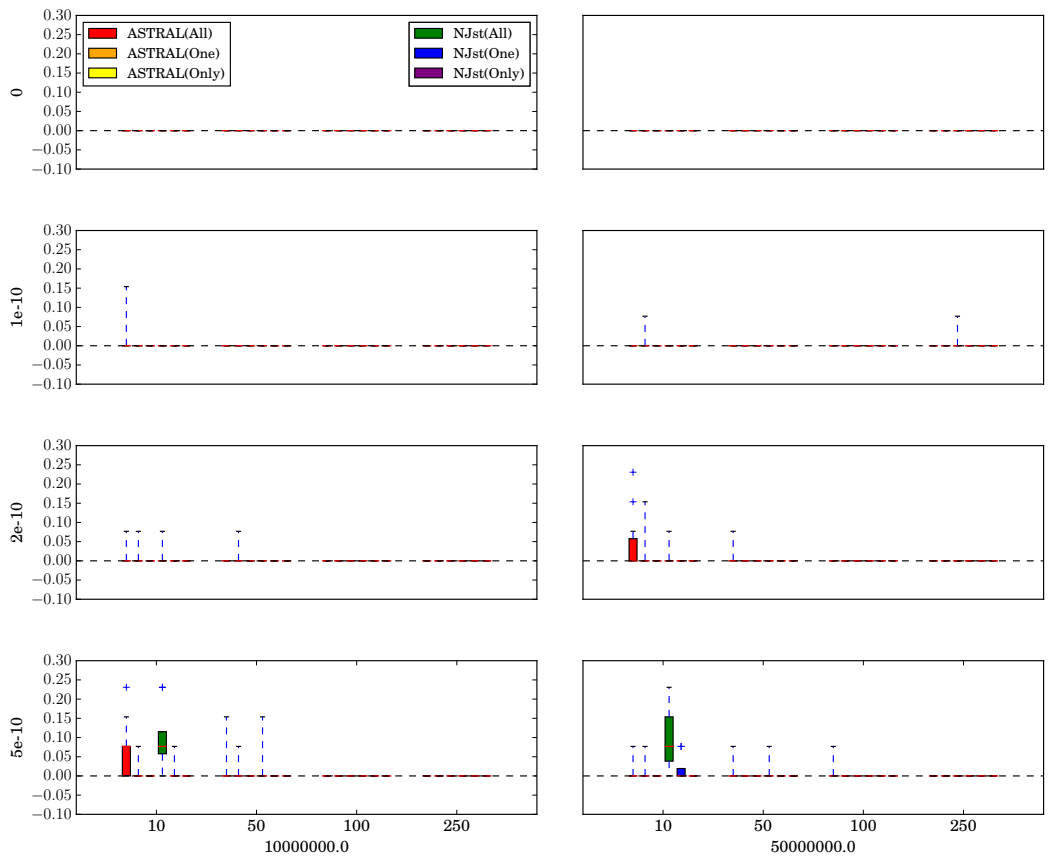

FIG. 9. The RF distance between the true species tree and the ones inferred by ASTRAL and NJst from true locus trees of the 16 yeast genomes under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

In analyses of the two biological data sets, and treating the species trees reported in the cited literature as the truth, all methods inferred the true fly species tree, except for ASTRAL(ALL) and ASTRAL(ONE), which only differed in terms of placing one species. For the yeast species tree, NJst inferred the true species tree, while maximum pseudo-likelihood and ASTRAL inferred trees that differed from the true one in only two groupings of taxa.

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

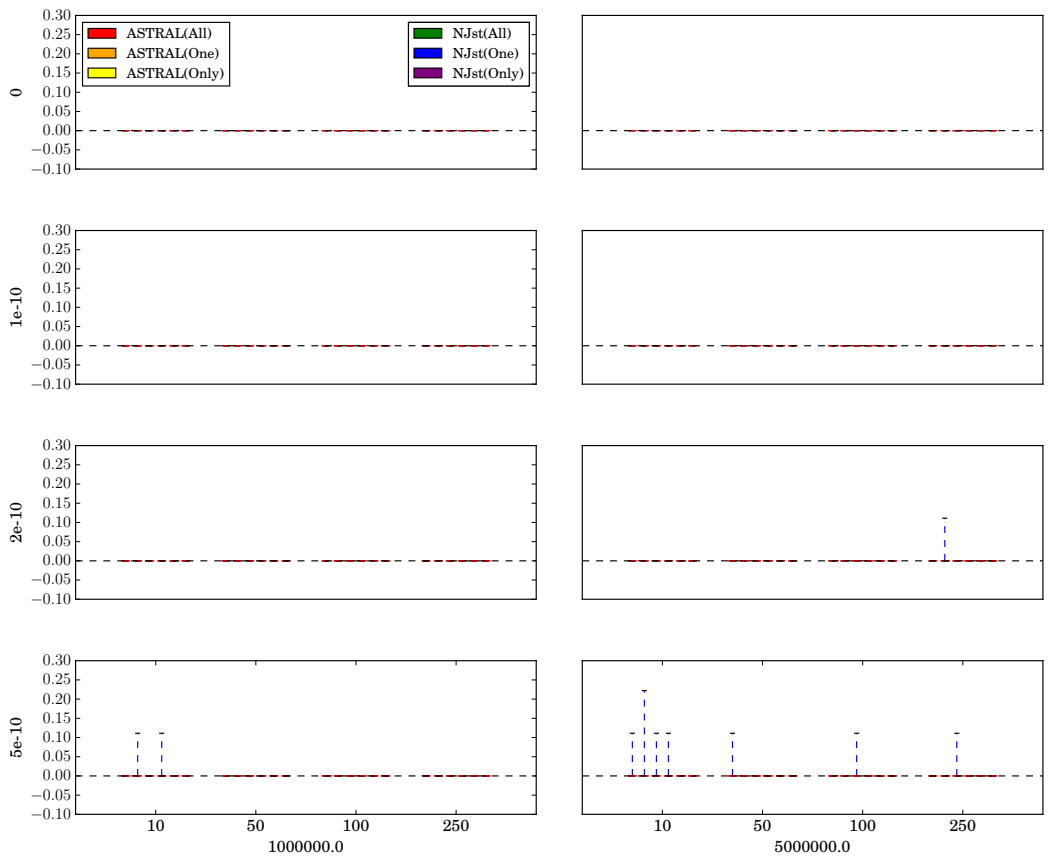

FIG. 10. The RF distance between the true species tree and the ones inferred by ASTRAL and NJst from true locus trees of the 12 fly genomes under different GDL rates and population sizes. Each row represents a GDL rate and each column represents a population size. The X axis represents the number of gene families used and the Y axis represents the RF distance.

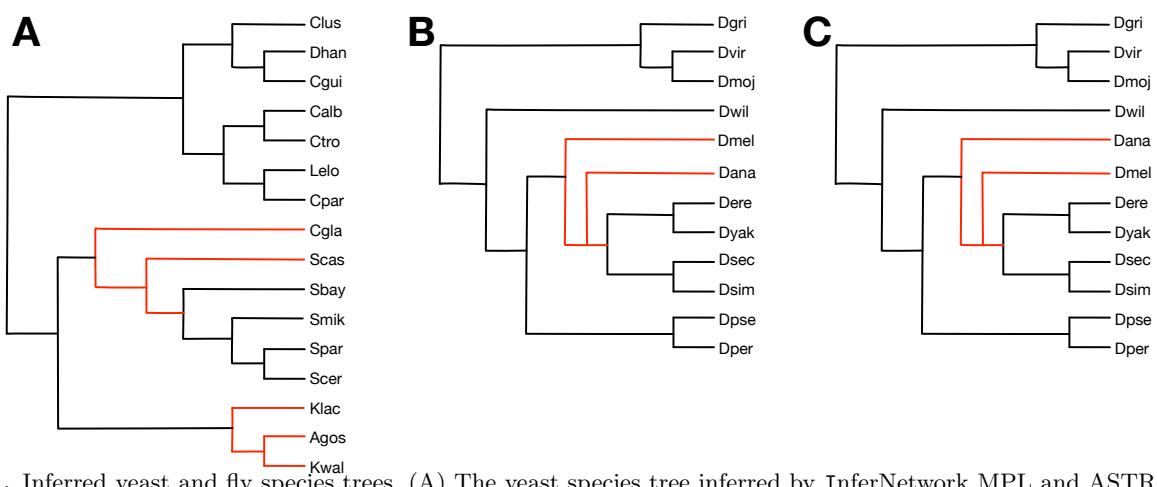

FIG. 11. Inferred yeast and fly species trees. (A) The yeast species tree inferred by InferNetwork_MPL and ASTRAL. (B) The fly species tree inferred by ASTRAL(ALL). (C) The fly species tree inferred by ASTRAL(ONE). Differences between the inferred species trees and the trees of Fig. 1 are highlighted in red.

tree inference from those data alone. In these cases, our results suggest a ready source of more phylogenetic signal.

While this study focused on the accuracy of the inferred species tree topology, treating paralogs as orthologs from multiple individuals would have

impact on the estimated branch lengths of the species tree. In particular, under the ALL setting, there could be much more incongruence due to the large number of lineages, and, consequently, MSC-based methods would estimate the branch lengths to be shorter than they truly are. For example, for the yeast data set, the lengths of the branch immediately above the MRCA of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* in coalescent units as estimated by InferNetwork_MPL on ALL, ONE, and ONLY were about 0.14, 0.68, and 1.52, respectively.

The results of our analyses point to the following potential method for inferring accurate species trees by utilizing as much of the phylogenomic data as possible (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 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. Species tree inference from a concatenation of the sequences in the gene families is challenging in the presence of paralogs for at least two reasons. First, when gene families have different copies across species, the concatenated sequences will have very large gaps. Second, correct orthology detection is required so that orthologous gene copies are placed in correct correspondence across the multiple genomes in the concatenation. If there is hidden paralogy in the data set, this issue is very important to examine so as to avoid aligning non-orthologous sequences in the concatenated data set.

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 finding implies that 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 (Emms and Kelly, 2018; Hellmuth *et al.*, 2015), though in certain cases it seems that paralogs can have undue influence on the species tree inferred (Brown and Thomson, 2016).

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 than those captured in our simulations (Hahn, 2009; Innan and Kondrashov, 2010). 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 involved 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 to LN and NSF grant DBI-1564611 to MWH.

References

Arguello, J. R., Chen, Y., Yang, S., Wang, W., and Long, M. 2006. Origination of an x-linked testes chimeric gene by illegitimate recombination in drosophila. *PLoS Genetics*, 2(5): e77.

Arvestad, L., Lagergren, J., and Sennblad, B. 2009. The gene evolution model and computing its associated probabilities. *Journal of the ACM (JACM)*, 56(2): 7.

Brown, J. M. and Thomson, R. C. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, 66(4): 517–530.

Daubin, V. and Gouy, M. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Informatics*, 12: 155–164.

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.

DeGiorgio, M. and Degnan, J. H. 2013. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Systematic Biology*, 63(1): 66–82.

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*.

Emms, D. and Kelly, S. 2018. STAG: Species tree inference from all genes. *bioRxiv*, page 267914.

Gribaldo, S. and Philippe, H. 2002. Ancient phylogenetic relationships. *Theoretical Population Biology*, 61(4): 391–408.

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. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, 100(5): 605–617.

Hahn, M. W., Han, M. V., and Han, S.-G. 2007. Gene family evolution across 12 drosophila genomes. *PLoS*

genetics, 3(11): e197.

Hellmuth, M., Wieseke, N., Lechner, M., Lenhof, H.-P., Middendorf, M., and Stadler, P. F. 2015. Phylogenomics with paralogs. *Proceedings of the National Academy of Sciences*, 112(7): 2058–2063.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37(1): 203–217.

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

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

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39: 309–338.

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. and Yu, L. 2011. Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5): 661–667.

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., Yu, L., and Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*, 10(1): 302.

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.

Mallo, D., de Oliveira Martins, L., and Posada, D. 2015. Simphy: phylogenomic simulation of gene, locus, and species trees. *Systematic Biology*, 65(2): 334–344.

Mansai, S. P., Kado, T., and Innan, H. 2011. The rate and tract length of gene conversion between duplicated genes. *Genes*, 2(2): 313–331.

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.

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.

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

Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. 2013. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. *Genetics*, pages genetics–113.

Scornavacca, C. and Galtier, N. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1): 112–120.

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.

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.

Than, C. and Nakhleh, L. 2009. Species tree inference by minimizing deep coalescences. *PLoS Computational Biology*, 5(9): e1000501.

Thornton, K. and Long, M. 2004. Excess of amino acid substitutions relative to polymorphism between x-linked duplications in drosophila melanogaster. *Molecular Biology and Evolution*, 22(2): 273–284.

Wen, D., Yun, Y., Zhu, J., and Nakhleh, L. 2018. Inferring phylogenetic networks using phylonet. *Systematic Biology*, 67(4): 735–740.

Yu, Y. and Nakhleh, L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10): S10.

Yu, Y., Warnow, T., and Nakhleh, L. 2011a. Algorithms for MDC-based multi-locus phylogeny inference. In *International Conference on Research in Computational Molecular Biology*, pages 531–545. Springer.

Yu, Y., Warnow, T., and Nakhleh, L. 2011b. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, 18(11): 1543–1559.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46): 16448–16453.

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.