Syst. Biol. 00(0):1–7, 2016
© The Author(s) 2016. Published by Oxford University Press, on behalf of the Society of Systematic Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syw064

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

Zhi Yan[1], Peng Du[1], Matthew W. Hahn[2], AND Luay Nakhleh[1,3,*]

[1]*Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA;* [2]*Department of Biology and Department of Computer Science, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA;* [3]*Department of BioSciences, Rice University, 6100 Main Street, Houston, TX 77005, USA;*
[*]*Correspondence to be sent to: Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA; E-mail: nakhleh@rice.edu.*

Received ...; reviews returned ...; accepted ...

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 that 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 provide very accurate results, either by treating all gene copies within a family as alleles from multiple individuals or by randomly selecting one copy per species. Our results have significant implications for the use of MSC-based phylogenomic analyses, demonstrating that they do not have to be restricted to single-copy loci, thus greatly increasing the amount of data that can be used. [Multispecies coalescent; incomplete lineage sorting; gene duplication and loss; orthology; paralogy.]

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. The reasons for this incongruence include biological causes such as incomplete lineage sorting (ILS) and horizontal gene transfer (broadly interpreted to include all biological processes involving gene flow), as well as technical causes such as the misidentification of paralogs as orthologs ("hidden paralogy"; (Maddison, 1997)).

Inference of species phylogenies in light of incongruence relies on a model of how gene trees evolve 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 genealogies. 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. Indeed, most phylogenetic methods require the identification of orthologs. Therefore, before such inference methods are 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 within larger gene families, 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 it is likely that the MSC induces a distribution that differs from that arising from gene duplication and loss (GDL). An obvious way to handle data sets where ILS and gene duplication/loss could have simultaneously acted on 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 (Boussau *et al.*, 2013; 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 three species tree inference methods, all of which use gene trees as the input data: The maximum pseudo-likelihood method of Yu and Nakhleh (2015) as implemented by the function `InferNetwork_MPL` in PhyloNet (Wen *et al.*, 2018), ASTRAL-III (Zhang *et al.*, 2018), and NJ_{st} (Liu and Yu, 2011). It is important to point out that we ran `InferNetwork_MPL` restricting the search to the tree speace (i.e., not allowing reticulations), which amounts to running the species tree inference method MP-EST (Liu *et al.*, 2010).

For the sake of conclusions that we draw from this study, it maybe helpful to highlight the difference between `InferNetwork_MPL` on the one hand and ASTRAL and NJ_{st} on the other hand. The former optimizes a pseudo-likelihood function that is derived based on the assumptions of the MSC. This function is 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, its 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 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 either 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. We find that gene tree estimation error affects the methods' performances at a similar, or even higher, level than ILS.

METHODS

Species tree inference methods

For species tree inference, we use three methods that are statistically consistent under the MSC:

- 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.
- ASTRAL-III (Zhang *et al.*, 2018), Version 5.6.3.
- NJ_{st} (Liu and Yu, 2011).

While the maximum likelihood method of Yu *et al.* (2014) as implemented by the InferNetwork_ML function in PhyloNet (Wen *et al.*, 2018) is relevant here, it is much more computationally demanding than maximum pseudo-likelihood, and we chose not to run it.

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 eliminate, 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 lactis*

(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).

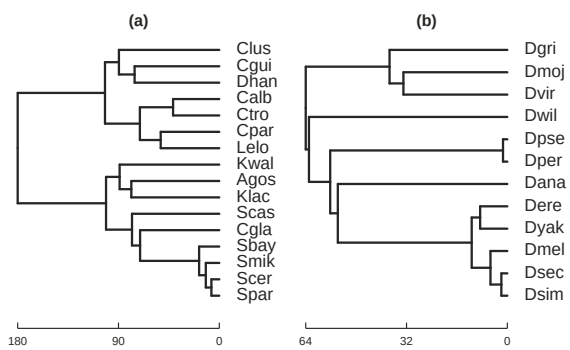

FIGURE 1. The species trees reported in (Rasmussen and Kellis, 2012) for the yeast and fly species, which are also used as the true species trees in the simulations. (a) The species tree of 16 fungi species. (b) The species tree of 12 fly species. The species tree topologies and their branch lengths in units of million years are reported on http://compbio.mit.edu/dlcoal/.

For the fungal 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 (2011). 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 0.9 year per generation 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.

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 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} found in (Schrider *et al.*, 2013).

To generate gene trees while allowing for ILS and GDL, 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. 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 (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 its 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, trees without estimation error) of a given size, we fed the data set as input to `InferNetwork_MPL`, ASTRAL, and NJ_{st} 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.

To study the effect of error in the gene tree estimates, we simulated the evolution of sequences of length 500 on all gene trees under the HKY model, using Seq-gen (Rambaut and Grassly, 1997). We then inferred gene trees from the simulated sequence data using IQTREE (Nguyen *et al.*, 2014). `InferNetwork_MPL` assumes that the input gene trees are rooted. In this study, we rooted the gene tree estimates by minimizing deep coalescences (Maddison, 1997; Than and Nakhleh, 2009); that is, we rooted each gene tree in a way that minimizes the number of extra lineages when reconciled with the true species tree. Furthermore, to study the effect of error in the locus tree estimates, we treated the true locus tree as a gene tree and simulated the evolution of sequences of length 500 on all locus trees under the HKY model, using Seq-gen, and inferred locus trees from the simulated sequence data using IQTREE. It is important to note that in practice only gene trees, but not locus trees, are inferable, as the locus tree is an artifact of the three-tree model (Rasmussen and Kellis, 2012) and not a biological entity. However, conducting analysis using inferred locus trees

gives a picture of the performance when all incongruence is due to GDL and gene tree error only.

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, as well as 250 gene trees estimated with RAxML (Stamatakis, 2006) reported in http://www.treefam.org for the 12 fly genomes for the ALL, ONE and ONLY settings. We then estimated the species trees using ASTRAL, NJ_{st}, and maximum pseudo-likelihood with these gene trees as input.

We again rooted each gene tree in the empricial data with respect to the species tree of Fig. 1 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).contrast

RESULTS

Characteristics of the simulated data

Before we describe the inference results, we discuss the characteristics of the simulated data. First, we inspect the effects of gene duplication and loss on the number of gene copies per species in each gene family. Fig. 2(a-b) and Fig. S1(a-b) show data on the sizes (numbers of copies) of gene families in the 16-taxon and 12-taxon data sets, respectively, under the various settings of effective population sizes and duplication and loss rates.

Clearly, the higher the GDL rates, the larger the variance in size of gene families. The figure also shows that the average size of a gene family is roughly equal to the number of species, with the largest gene families having 51 copies for the 16-taxon data sets, and 58 copies for 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

TABLE 1. The average number of gene families that fit the ONLY setting out of the 10,000 gene families for 10 data sets.

	16-taxon data		12-taxon data	
GDL rate N_e ↓ →	10^7	5×10^7	10^6	5×10^6
1×10^{-10}	7619	7585	4591	4584
2×10^{-10}	5794	5787	2197	2176
5×10^{-10}	2554	2538	268	266

shows that as the GDL rates increase, the number of single-copy gene families decreases.

We then set out to assess the extent of incongruence in the gene trees due to GDL and ILS. For every pair of true species tree and true locus tree, we computed the number of extra lineages (Maddison, 1997) using the DeepCoalCount_tree command in PhyloNet (Than and Nakhleh, 2009; Wen *et al.*, 2018) as a proxy for the amount of incongruence in the data. Here, we treated all gene copies from the same species as different individuals. Zero extra lineages mean there is no incongruence between the two trees, and the higher the value, the more incongruence there is. In particular, no incongruence means that all gene copies from the same species are monophyletic in the locus tree, and when restricted to a single arbitrary copy per species, the locus tree and species tree have identical topologies.

Fig. 2(c-d) and Fig. S1(c-d) show data on the number of extra lineages in the 16-taxon and 12-taxon data sets, respectively, under the various settings of effective population sizes and duplication and loss rates. It is important to note that all incongruence in this case is exclusively due to GDL (ILS is not a factor in the results in these two panels). The panels do not have results for the GDL rate of 0x, because in such cases there is no incongruence at all between the locus tree and the species tree, and thus there are zero extra lineages. The results show that, unsurprisingly, there is much more incongruence for the ALL scenario than the ONE scenario. For the ONLY scenario, there is very little incongruence in the data, and such incongruence is relatively more noticeable in the 12-taxon data set, as all nonzero levels of GDL rates have nonzero number of extra lineages. In constrast, for the 16-taxon data set only the largest GDL rate causes incongruence between locus tree and species tree. In both data sets, the incongruence indicates the phenomenon of hidden paralogy: single-copy genes are paralogs, and their gene trees therefore do not always agree with the species tree in terms of topology.

Finally, we computed the number of extra lineages when reconciling the true gene trees with the true locus trees. Here, incongruence is exclusively due to ILS. Fig. 2(e-f) and Fig. S1(e-f) show data on the number of extra lineages in the 16-taxon and 12-taxon data sets, respectively, under the various settings of effective population sizes and duplication and loss rates. When the gene tree topology is identical to the locus tree topology, the number of extra lineages is 0, and the larger the number of extra lineages, the more ILS in the data. The figure shows that, as expected, the amount of ILS is larger for larger population sizes. Furthermore, there is much more ILS in the 16-taxon data set than in the 12-taxon data set. One other trend to observe is that, on average, the amount of incongruence due to ILS increases with the increase in the GDL rate. This is a reflection of the fact that for higher GDL rates, the locus trees are larger (more leaves and internal branches) and this naturally results in more ILS.

Results on Simulated Data

We are now in position to describe the inference results. We show figures for the 16-taxon data sets in the main text, while figures for the 12-taxon data sets Figs. S8 to S11 are all in the Supplementary Materials. The results for the 12-taxon data sets are consistently better in terms of accuracy, so we chose to focus here on the less-optimal results.

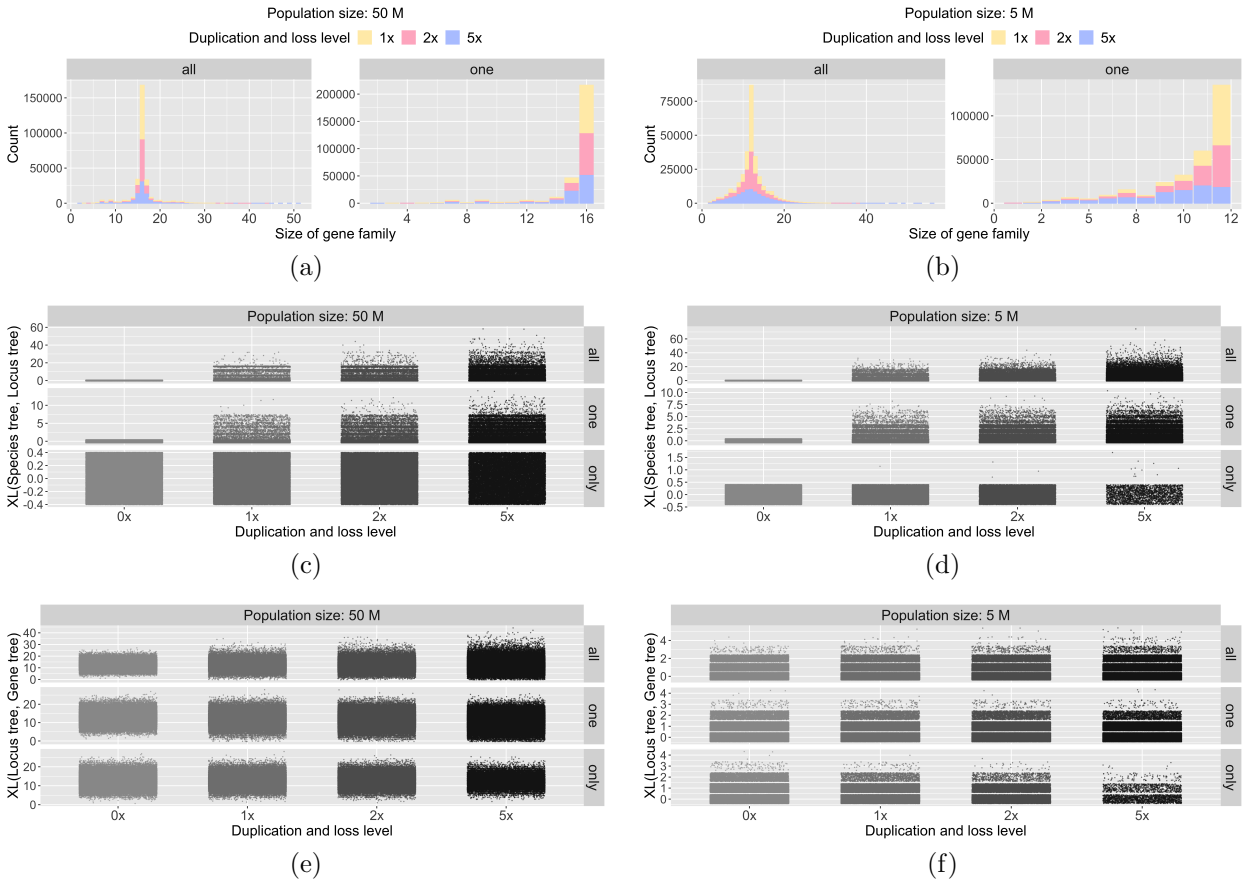

FIGURE 2. Characteristics of the simulated data under different settings of the duplication/loss rates and effective population sizes using all 10 replicates. The duplication/loss rates are denoted by the rate multiplier (0x, 1x, 2x and 5x), where 1x is the rate used by (Rasmussen and Kellis, 2011, 2012; Zhang and Wu, 2017). (a-b) Distribution of the total number of gene copies in individual gene families in the 16-taxon and 12-taxon data sets, respectively. (c-d) Scatter plots of XL(Species tree, Locus tree), the number of extra lineages when reconciling the true locus trees with the true species tree, for the 16-taxon and 12-taxon data sets, respectively. (e-f) Scatter plots of XL(Locus tree, Gene tree), the number of extra lineages when reconciling the true gene trees with the true locus tree, for the 16-taxon and 12-taxon data sets, respectively.

We first ran the three inference methods ASTRAL, `InferNetwork_MPL`, and NJ_{st}, 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 on the full 16-taxon tree are shown in Fig. 3 and Fig. S4. Note that in all cases, using input data with GDL levels of 0 amounts to inferring a species tree from gene trees whose incongruence is solely due to ILS.

There are several observations based on the results. First, the accuracy of the inferred 16-taxon trees is much lower in general than that of the inferred 12-taxon trees. In particular, for the 12-taxon data sets, the species trees are perfectly estimated in almost all cases (Fig. S3), whereas the species tree estimation error is high, especially for the larger population sizes, for the 16-taxon data sets. As shown in Fig. 2 and Fig. S1, both data sets have similar gene family sizes, but differ significantly in terms of the amount of ILS in the data, with the 12-taxon data sets having very little ILS. Therefore, the straightforward explanation for the observed differences in the RF distances between the 16- and 12-taxon data sets is the higher level of ILS in the former. Given that the level of incongruence due to GDL is similar between the 16-taxon and 12-taxon data sets (Fig. 2(c-d) and Fig. S1(c-d)), these results point to the larger role that ILS plays in the methods' performances than GDL does.

Second, in the case of the 16-taxon data, the performance of all three methods improves as the number of gene families used as input to the method increases. Note also that the largest dataset used here consists of only 250 gene trees, which is much smaller than the number available in most phylogenomic data sets. While there is very little difference observed in the performance among the three methods on the 16-taxon data, ASTRAL and NJ_{st} are more similar to each other in terms of performance than either of them is to inference under maximum pseudo-likelihood. This makes sense as both ASTRAL and NJ_{st} are summary methods that make inference based on statistics derived from the

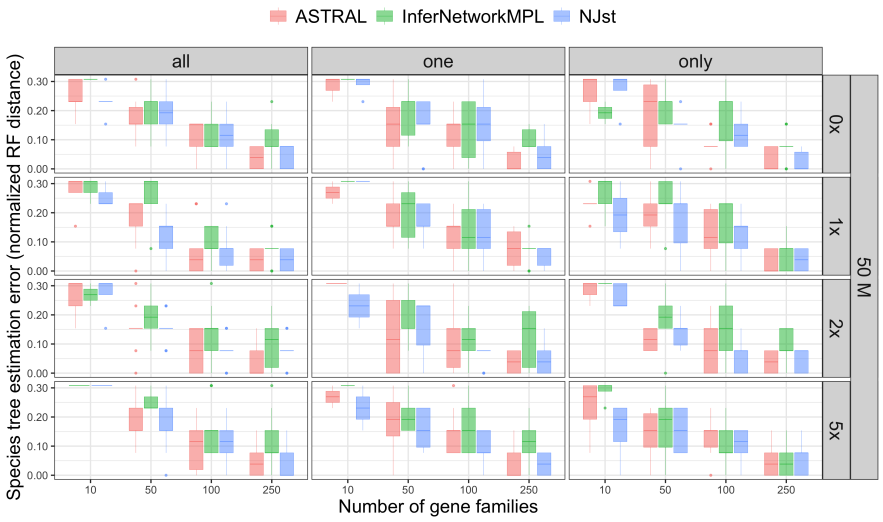

FIGURE 3. The normalized RF distances between the true species tree and the ones inferred by ASTRAL, `InferNetwork_MPL`, and NJ_{st} from true gene trees of the 16-taxon simulated yeast gene families under population size 5.0×10^7 and different GDL rates. The duplication/loss rates are denoted by the rate multiplier (0x, 1x, 2x and 5x), where 1x is the rate used by (Rasmussen and Kellis, 2011, 2012; Zhang and Wu, 2017). Each row corresponds to a combination of population size and GDL rates. The X axis in each panel represents the number of gene families used and the Y axis represents the normalized RF distance.

input gene trees, whereas maximum pseudo-likelihood uses calculations based on the multispecies coalescent directly.

Third, the level of ILS for a population size of 50M is higher than for a population size of 10M, and this results in lower accuracy of inferred species trees by all three methods in the former case. This behavior is expected for any MSC-based method, regardless of whether GDL is acting.

Last, but not least, we observe very little difference in the accuracy of inferred species trees across the three input scenarios: ALL, ONE, and ONLY. This implies that the presence of paralogs in the data, no matter how they are treated, does not have much of an effect on the performance of the three methods. In the case of the 12-taxon data, all three methods make perfect inferences in almost all cases (see Supplementary Materials Figs. S8 and S10).

These results raise the important question: Does GDL have any effect on the performance of these three methods? To answer this question, we ran ASTRAL, `InferNetwork_MPL`, and NJ_{st} 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; that is, ILS plays no role in incongruence here. It is important to point out here that locus trees are mathematical constructs of the three-tree model; in practice, inferring a locus tree is not possible, unless the data has no ILS. We conducted this experiment to study the performance of methods when GDL, but not ILS, causes all incongruence. Results on the full 16-taxon data sets are shown in Fig. 4 and Fig. S5. As the results now show, all three methods infer the true species tree on almost all data sets regardless of the parameter settings and the input scenario. In other words, when these three methods—which have been developed based on the multispecies coalescent directly (`InferNetwork_MPL`) or "inspired" by it (ASTRAL and NJ_{st})—are applied to data that have no ILS but do have paralogs in them, they have almost perfect accuracy in terms of the species tree topology they infer, under the conditions of our simulations. Combined with the results summarized in Fig. 3 and Fig. S4, these results show, perhaps surprisingly, that methods developed to handle ILS but not GDL do much better in handling GDL than they do in handling ILS.

In practice, gene trees are unknown and inferred from the sequence data. To simulate more realistic data, we inferred gene trees and locus trees from simulated sequence data and fed these tree estimates as input to the three methods. In this case, gene tree error is a factor in the observed incongruences. Fig. S2 shows the extent of error of in the estimated gene and locus trees for the 16-taxon data. The gene tree estimation error is measured by the normalized RF distance between the true gene tree and the reconstructed gene tree. For the 12-taxon data set, the average gene tree estimation error ranges from 0.581 to 0.623, whereas the average locus tree estimation error is slightly lower, ranging from 0.57 to 0.612 (Fig. S3). For the 16-taxon data set, the average gene tree estimation error ranges between 0.101 to 0.106 while the average locus tree estimation error ranges from 0.0792 to 0.0848. In other words, there is much less gene tree estimation error in the 16-taxon data sets than in the 12-taxon data sets.

Results of species tree inference using the full 16-taxon data set based on estimated gene trees are shown in Fig. 5 and Fig. S6 and based on the locus tree estimates are shown in Fig. 6 and Fig. S7. These results should be contrasted with Fig. 3, Fig. S4, Fig. 4 and

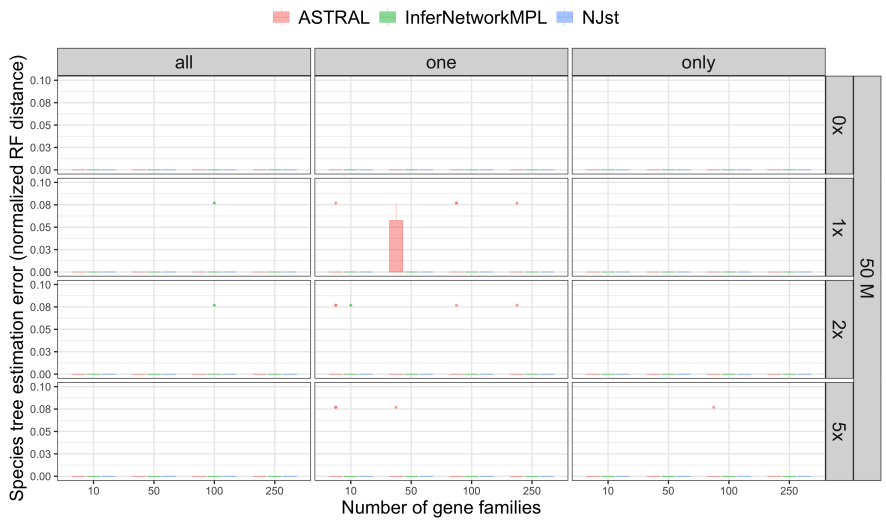

FIGURE 4. The normalized RF distances between the true species tree and the ones inferred by ASTRAL, `InferNetwork_MPL`, and NJ_{st} from true locus trees of the 16-taxon simulated yeast gene families under population size 5.0×10^7 and different GDL rates. The duplication/loss rates are denoted by the rate multiplier (0x, 1x, 2x and 5x), where 1x is the rate used by (Rasmussen and Kellis, 2011, 2012; Zhang and Wu, 2017). Each row corresponds to a combination of population size and GDL rates. The X axis in each panel represents the number of gene families used and the Y axis represents the normalized RF distance.

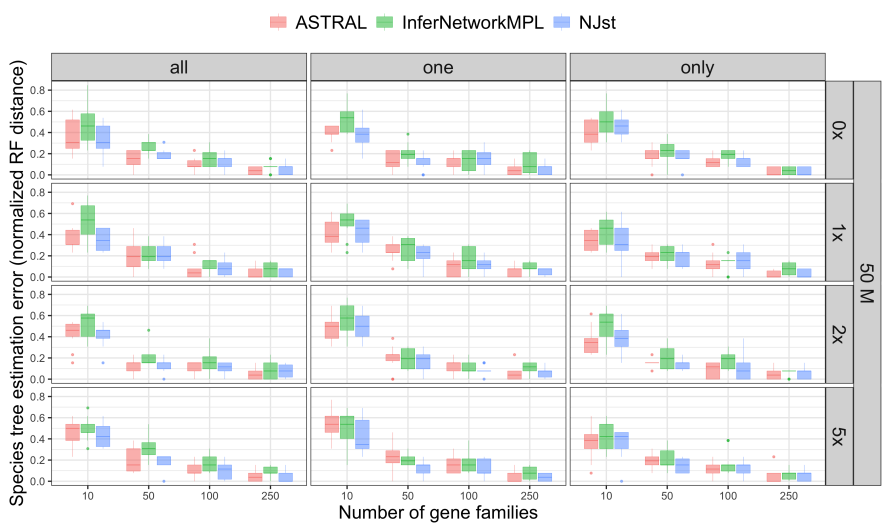

FIGURE 5. The normalized RF distances between the true species tree and the ones inferred by ASTRAL, `InferNetwork_MPL`, and NJ_{st} from estimated gene trees of the 16-taxon simulated yeast gene families under population size 5.0×10^7 and different GDL rates. The duplication/loss rates are denoted by the rate multiplier (0x, 1x, 2x and 5x), where 1x is the rate used by (Rasmussen and Kellis, 2011, 2012; Zhang and Wu, 2017). Each row corresponds to a combination of population size and GDL rates. The X axis in each panel represents the number of gene families used and the Y axis represents the normalized RF distance.

Fig. S5, respectively, to understand the effect of gene tree estimation error on the accuracy of the three species tree inference methods.

In the case of inferences using gene tree estimates where ILS, GDL, and gene tree estimation error are involved, the error rates of all three species tree inference methods went up, as expected, due to gene tree error (Fig. 5 and Fig. S6), but only slightly. The accuracy of the species trees improves as the number of gene families increases. As we discussed above, the error in gene tree estimates in the 16-taxon data sets is very low. Since gene tree estimation error in the 12-taxon data sets is much higher (higher substitution rates resulting in noisier sequence data), we now observe a larger impact on this error on the performance of methods on the 12-taxon data sets. While the methods had an almost perfect accuracy on true gene trees, species tree estimates now have as high as 50% error when 10 gene family trees are used, and close to 25% error when 250 gene family trees are used (Fig. S10). These results illustrate the big impact gene tree estimation error has on these methods. In the case of the 12-taxon data sets, the impact of gene

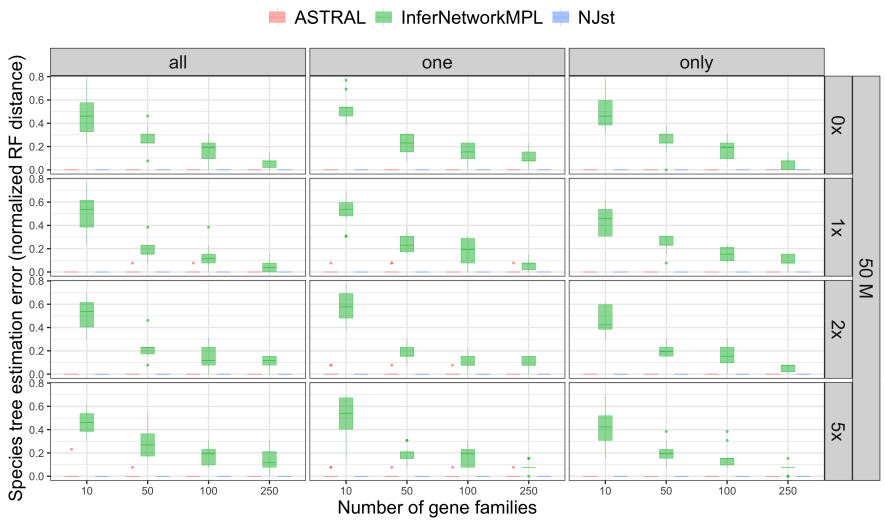

FIGURE 6. The normalized RF distances between the true species tree and the ones inferred by ASTRAL, `InferNetwork_MPL`, and NJ_{st} from estimated locus trees of the 16-taxon simulated yeast gene families under population size 5.0×10^7 different GDL rates. Each row corresponds to a combination of population size and GDL rates. The X axis in each panel represents the number of gene families used and the Y axis represents the normalized RF distance.

tree estimation error significantly outweighs that of ILS and/or GDL.

Fig. 6 demonstrates how GDL and gene tree inference error (but no ILS) impact species tree inference. As Fig. 4 and Fig. S5 show almost perfect performance of species tree inference from true locus trees (i.e., GDL and no ILS), all the error observed in Fig. 6 and Fig. S7 can be ascribed to error in the gene trees. The results demonstrate that in the absence of ILS, both ASTRAL and NJ_{st} are robust to gene tree estimation error, whereas `InferNetwork_MPL` is very sensitive to gene tree estimation error. In the case of the 12-taxon data sets, where locus tree estimation errors are much higher, the three species tree inference methods have comparable accuracies (Fig. S11). It is worth mentioning that as `InferNetwork_MPL` requires rooted gene trees, poorer performance of this method could be due to the incorrect rooting of the estimated gene trees. We rooted the gene tree estimates by minimizing the number of extra lineages when reconciling the gene/locus trees with the true species tree (according to the formulation given in (Yu *et al.*, 2011a, b)). This further highlights the challenge with identifying the root of gene trees when GDL occurs.

All of these results combined point to a very small impact of GDL on the performance of the three studied species tree inference methods, regardless of how the paralogs are handled. Across all data sets it was evident that gene tree estimation error has noticeable impact on the methods' performances.

Results on Biological Data

We ran the three methods, `InferNetwork_MPL`, ASTRAL, and NJ_{st}, on two biological data sets. Again, each data set consists of only 250 gene trees, although many more trees are available for each.

For the 16 fungi genome empirical data set, NJ_{st} inferred an identical tree to that of Fig. 1(a) under all three input scenarios. Assuming this tree is the "true" tree, NJ_{st} has performed remarkably well. For the other two methods, for all three sampling schemes, the inferred tree differs from the tree shown in Fig. 1(a). In particular, the positions of *Kluyveromyces waltii* and *Kluyveromyces latics* have been switched, as have the positions of *Candida glabrata* and *Saccharomyces castellii* (Fig. 7(a)).

For the 12 fly genome empirical data set, ASTRAL(ONLY) inferred the exact same tree as the species tree shown in Fig. 1(b). The trees inferred by ASTRAL(ALL) and ASTRAL(ONE) differed from the true tree of Fig. 1(b) in terms of the placement of *Drosophila melanogaster*, as shown in Fig. 7(b) and Fig. 7(c), respectively. `InferNetwork_MPL` and NJ_{st} inferred an identical tree to that of Fig. 1(b) under all three input scenarios.

DISCUSSION

As phylogenomic data sets grow, our ability to use them within the bounds of current analysis paradigms shrinks. One of the main problems is the decreasing number of gene families that are single-copy as the number of sampled species increases (Emms and Kelly, 2018). Because most current phylogenetic methods assume that only single-copy orthologs are being used, this restriction means that such methods cannot be used for data sets with even several dozen taxa without severe downsampling or other *ad hoc* solutions (e.g., (Thomas *et al.*, 2020)). Here, we set out to ask whether MSC-based

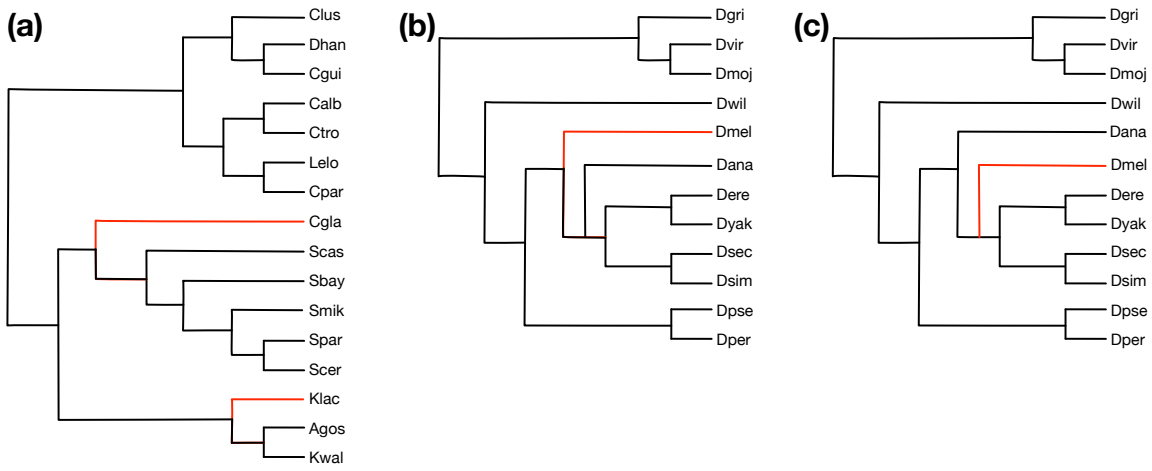

FIGURE 7. Inferred yeast and fly species trees. (a) The yeast species tree inferred by `InferNetwork_MPL` and ASTRAL. (b) The fly species tree inferred by ASTRAL(ALL). (c) The fly species tree inferred by ASTRAL(ONE). Differences between the inferred species trees and the trees of Fig. 1 are highlighted in red.

phylogenomic methods can be applied to data containing both orthologs and paralogs.

On simulated data sets where only ILS acted on gene families, and GDL was not a factor, all methods had the expected performance: accurate species tree estimates that improved as the number of gene family trees used increases. In the case where the level of ILS was very low (the 12-taxon data), the methods had perfect performance under almost all conditions, regardless of the number of gene trees used.

In the cases where both ILS and GDL acted on gene families, the performance of the three methods was hardly affected. These results imply that running these methods on data with paralogs—either by treating them as alleles from different individuals or by sampling a single copy at random—is a safe practice when the desired outcome is the species tree topology. This is especially important for data sets with a large number of species or high GDL rates.

When the methods were run on the locus tree data, where ILS did not play a role and the data consisted of many gene families with multiple copies, the methods obtained very accurate species trees. This was true even when the data consisted of only 10 gene family trees. This further demonstrates that GDL has very little effect on the performance of the three methods.

While at first it may be surprising that these methods have done very well in terms of accuracy, the majority of signal in any input gene tree 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 carrying an extra gene-copy. Similarly, gene loss does not positively mislead these methods, leading to accurate reconstructions of the species tree. Nevertheless, upon close inspection, some of these results are not intuitive, especially for the maximum pseudo-likelihood inference. `InferNetwork_MPL` 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 NJ_{st} are summary methods that make no direct use of the MSC. Consequently, one would have expected that `InferNetwork_MPL` would be very sensitive to the presence of paralogs in the data, while ASTRAL and NJ_{st} less so. However, we did not observe any such behavior of the methods.

In practice, gene trees are estimated from sequence data and can be erroneous. Error in the gene tree estimates, rather than ILS, could explain much of the heterogeneity observed in phylogenomic analyses, especially at deeper nodes in a species tree (Scornavacca and Galtier, 2017). We showed the gene tree estimation error can indeed impact species tree inference significantly, and the level of its impact is similar to that of ILS, if not larger.

In analyses of two biological data sets where a species tree has been inferred using hundreds or thousands of loci, we found high accuracy of methods using paralogs with only 250 gene trees. All methods accurately inferred the published fly species tree, except for ASTRAL(ALL) and ASTRAL(ONE), which only differed in terms of placing one species. For the yeast species tree, NJ_{st} inferred the published species tree, while maximum pseudo-likelihood and ASTRAL inferred trees that differed from the true one in only two groupings of taxa (around two very short branches in the species tree).

All these results point to a clear message: Under the conditions of our simulations and on the two biological data sets we used, running MSC-based or MSC-inspired species tree inference on gene trees 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 are available to ensure good performance of species tree inference from those data alone. In these cases, our results suggest a ready source of more phylogenetic signal. Summary methods that do not explicitly use the MSC model (i.e., ASTRAL and NJ_{st}) are expected to be more robust in the presence of GDL than methods that explicitly use the model.

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 this reason, branch lengths inferred by such methods should not be used. Instead, an alternative approach is needed.

The results of our analyses point to the following potential approach for inferring accurate species trees (topologies and branch lengths) by utilizing as much of the phylogenomic data as possible:

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

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

3. Using a smaller subset of truly single-copy genes, and fixing the species tree topology obtained from Step (2), 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. As an alternative to using only single-copy orthologs in Step (3), one could also use a statistical model that combines the MSC and GDL models (e.g., (Boussau *et al.*, 2013; Rasmussen and Kellis, 2012)). Such methods allow for paralogy detection and orthology assignment, conditional on the fixed species tree (or species trees), by using a 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.

CONCLUSIONS

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.

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. It is important to note however, that our results do not apply to concatenated analyses, and in such cases the presence of paralogs may indeed have a large, negative effect (Brown and Thomson, 2016). Species tree inference from a concatenation of the sequences with gene families is challenging in the presence of paralogs for at least two reasons. First, when gene families have different numbers of copies across species, the concatenated alignment will have very large gaps. Second, correct orthology detection is still required, so that orthologous gene copies are placed in correct correspondence across the multiple genomes in the concatenated alignment. This issue is very important to examine so as to avoid aligning non-orthologous sequences in the concatenated data set.

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 are also undoubtedly violated in most empirical data sets. 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.*, 2019), 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, including data files and online-only appendices, can be found in the Dryad data repository at .

FUNDING

Funding was provided by the National Science Foundation (DBI-1355998, CCF-1514177, CCF-1800723, and DMS-1547433) to L.N. and (DBI-1564611 and DEB-1936187) to M.W.H.

REFERENCES

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.

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. 2013. Genome-scale coestimation of species and gene trees. *Genome research*, 23(2): 323–330.

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.

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. 2019. Advances in computational methods for phylogenetic networks in the presence of hybridization. In T. Warnow, editor, *Bioinformatics and Phylogenetics*, pages 317–360. Springer.

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.

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.

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

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1): 268–274.

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 Grassly, 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. 2011. 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*, 65(5): 843–851.

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.

Thomas, G. W., Dohmen, E., Hughes, D. S., Murali, S. C., Poelchau, M., Glastad, K., Anstead, C. A., Ayoub, N. A., Batterham, P., Bellair, M., *et al.* 2020. Gene content evolution in the arthropods. *Genome Biology*, 21(1): 1–14.

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.