Syst. Biol. 71(3):706–720, 2022

DOI:10.1093/sysbio/syab081
Advance Access publication October 4, 2021

Maximum Parsimony Inference of Phylogenetic Networks in the Presence of Polyploid Complexes

ZHI YAN[1], ZHEN CAO[1], YUSHU LIU[1], HUW A. OGILVIE[1], AND LUAY NAKHLEH[1,2,*]

[1]*Department of Computer Science, Rice University, Houston, 6100 Main Street, Houston, TX 77005, USA and;* [2]*Department of Biosciences, Rice University, Houston, 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 02 October 2020; reviews returned 26 September 2021; accepted 29 September 2021
Associate Editor: Mark Holder

Abstract.—Phylogenetic networks provide a powerful framework for modeling and analyzing reticulate evolutionary histories. While polyploidy has been shown to be prevalent not only in plants but also in other groups of eukaryotic species, most work done thus far on phylogenetic network inference assumes diploid hybridization. These inference methods have been applied, with varying degrees of success, to data sets with polyploid species, even though polyploidy violates the mathematical assumptions underlying these methods. Statistical methods were developed recently for handling specific types of polyploids and so were parsimony methods that could handle polyploidy more generally yet while excluding processes such as incomplete lineage sorting. In this article, we introduce a new method for inferring most parsimonious phylogenetic networks on data that include polyploid species. Taking gene tree topologies as input, the method seeks a phylogenetic network that minimizes deep coalescences while accounting for polyploidy. We demonstrate the performance of the method on both simulated and biological data. The inference method as well as a method for evaluating evolutionary hypotheses in the form of phylogenetic networks are implemented and publicly available in the PhyloNet software package. [Incomplete lineage sorting; minimizing deep coalescences; multilabeled trees; multispecies network coalescent; phylogenetic networks; polyploidy.]

Hybridization and polyploidization have long been recognized as crucial factors in speciation and genomic and phenotypic novelties (Oxelman et al. 2017; Blischak et al. 2018). While in homoploid hybridization, the hybrid has the same number of chromosome sets as the two parental species, allopolyploid hybrids receive both chromosome sets from the parents, thus increasing the size of the chromosome set as compared to the two parents. Both types of hybridization result in reticulate evolutionary histories of the whole genomes that are best modeled by phylogenetic networks. Autopolyploidy, on the other hand, is whole-genome duplication (WGD) that involves a single lineage and does not violate a treelike evolutionary history at the level of the species phylogeny.

Polyploidy is prevalent across the eukaryotic branch of the Tree of Life. Many extant flowering plants are neopolyploids, and for the remaining diploid species, one or more rounds of ancient WGD events can be traced (Masterson 1994; Jiao et al. 2011), thus indicating that they are paleopolyploids. Although hybrid and polyploid species are less commonly observed in animals than plants, presumably owing to their potential reduced fitness, fish and amphibians are known to have high incidence of polyploidy (Glasauer and Neuhauss 2014; Berthelot et al. 2014; Woods et al. 2005). Furthermore, it is believed that at least two rounds of ancient WGD occurred in the vertebrate lineage (the 2R hypothesis) (Ohno 2013; Muffato and Crollius 2008). Fungi also include polyploids, with evidence supporting that the ancestor of the baker's yeast *Saccharomyces cerevisiae* underwent WGD (Marcet-Houben and Gabaldón 2015).

Allopolyploidization has been attracting attention for decades as it reflects the joint effects of genome doubling and interspecific hybridization. Unlike autopolyploids, allopolyploids contain multiple divergent subgenomes, each derived from distinct parental species. The investigation of subgenome evolution has been propelled by recent advances in homoeology analyses (Glover et al. 2016; Sancho et al. 2021). From modeling and inference perspectives, allopolyploidy is of particular interest, as it results from hybridization of two species and gives rise to evolutionary histories in the form of phylogenetic networks. Although polyploidy could potentially be identified from the chromosome count, the task of determining its mode of origin is nontrivial, especially when the parental taxa are closely related. Moreover, other evolutionary processes, such as incomplete lineage sorting (ILS), complicate this task. Therefore, extending models such as the multispecies coalescent to account for polyploidy could provide a powerful approach inferring evolutionary histories of polyploids.

Yu et al. (2012, 2014) extended the multispecies coalescent to incorporate hybridization into the model, giving rise to the multispecies network coalescent (MSNC). Based on this generative process, PhyloNet (Than et al. 2008; Wen et al. 2018) implements a wide array of methods for inferring phylogenetic networks in the presence of incomplete lineage sorting and diploid hybrids, including parsimony methods (Yu et al. 2013), maximum likelihood and pseudolikelihood methods (Yu et al. 2014; Yu and Nakhleh 2015; Zhu and Nakhleh 2018), and Bayesian methods (Wen et al. 2016; Wen and Nakhleh 2018; Zhu et al. 2018). Since almost all these methods are computationally demanding (with the exception of Yu and Nakhleh (2015)), a divide-and-conquer approach was recently introduced to speed

them up (Zhu et al. 2019). Cao et al. (2019) illustrated the use of many of these inference methods, as well as network summarization methods, on data generated under the multispecies network coalescence, which assumes diploid hybridization.

Kamneva et al. (2017) evaluated various inference tools in PhyloNet (Than et al. 2008; Wen et al. 2018) for inferring phylogenetic networks of polyploid strawberries, *Fragaria* (Rosaceae), with species that ranged in ploidy from tetraploid to decaploid. As Blischak et al. (2018) correctly pointed out, "Though not specifically built for polyploids, PhyloNet can model multiple haplotypes in each lineage. It can hence infer hybridization in these species when the assumptions of its model (the coalescent) are not violated, making it most appropriate for recent allopolyploids."

While existing inference tools in PhyloNet were not designed for handling polyploidy, there are several existing methods that are designed specifically to model polyploidy events (Oxelman et al. 2017). Many of these methods have relied on multilabeled species trees, or MUL-trees, to model polyploids. As multiple copies of a locus could be present in the genome due to polyploidization, a MUL-tree extends standard phylogenetic trees by allowing multiple leaves to be labeled by the same taxon name (Fig. 1a–d). There is a straightforward connection between a phylogenetic network and a MUL-tree, as illustrated in Figure 1a,b. Indeed, in one of the earliest works in this area, Huber and Moulton (2006) provided an algorithm for converting a MUL-tree into a phylogenetic network, which was later implemented in the PADRE software (Lott et al. 2009). This connection between phylogenetic networks and MUL-trees was the basis for computations under the MSNC in Yu et al. (2012, 2014) before moving towards computations directly on the phylogenetic network in subsequent implements in PhyloNet. The software tool GRAMPA (Thomas et al. 2017) uses MUL-trees to reconcile a set of gene trees parsimoniously with a given species tree to postulate polyploidy events.

To the best of our knowledge, the only statistical inference methods specifically designed for handling allopolyploids are the AlloppNET method of Jones et al. (2013) and its extension (Oxelman et al. 2017). AlloppNET uses Bayesian Markov chain Monte Carlo (MCMC) to sample, using multilocus DNA sequence data, the posterior distribution of phylogenetic networks that contain diploid and allotetraploid species. The method allows for multiple individuals per species and samples, in addition to the phylogenetic network topology, parameters including divergence and hybridization times as well as population sizes. Like other statistical inference methods in PhyloNet, AlloppNET is computationally intensive, in particular as it employs reversible-jump MCMC to sample the transdimensional space of phylogenetic networks. Furthermore, as mentioned above, AlloppNET allows for only diploid and allotetraploid species.

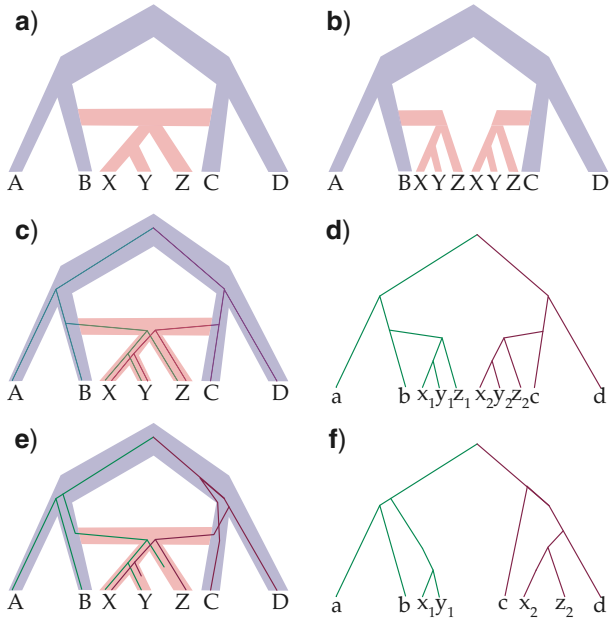

FIGURE 1. Allopolyploidy, phylogenetic networks, and MUL-trees. a) Phylogenetic network depicting allopolyploidization involving the ancestor of *X*, *Y*, and *Z*. b) MUL-tree representation of the network. c) Gene tree inside the branches of the phylogenetic network. d) Gene tree with two copies of the gene in each of the species *X*, *Y*, and *Z*. e) Gene tree inside the branches of the phylogenetic network in the presence of incomplete lineage sorting and gene loss. f) Gene tree where the signal for the allopolyploids is confounded by ILS and gene loss.

As described in Blischak et al. (2018), producing an all-encompassing stochastic model of polyploidization would be a massive undertaking due to the complexities of processes occurring during and after polyploidization events. In this article, we take a maximum parsimony approach to phylogenetic network inference in the presence of general allopolyploidy events. Recognizing that the parsimony method of Yu et al. (2013) as implemented in PhyloNet was not designed to handle allopolyploids, Oberprieler et al. 2017 coupled it with a permutation scheme where multiple analyses are conducted in each of which only two copies from the polyploid are mapped to one parent and the other copies are mapped to a second parent, and then reporting the optimal result over all these analyses. In this work, we extend the method of Yu et al. (2013) to properly handle polyploids without the need for a permutation approach. We implemented both an inference method (MPAllopp) and a scoring method in PhyloNet and assessed their performance on the simulated and biological data used in (Jones, 2017), (Marcussen et al., 2014), and (Joly et al., 2009). We report on the accuracy and running time of MPAllopp, and compare it to AlloppNET as well as existing methods in PhyloNet that were not designed to handle polyploid hybridization. We show that MPAllopp outperforms other methods in PhyloNet that were not designed to handle allopolyploidy and performs similarly to AlloppNET in terms of accuracy while being much faster.

METHODS

Minimizing Deep Coalescences in the Presence of Allopolyploidy

Maddison (1997) proposed minimizing deep coalescences (MDC) as a parsimony criterion for reconciling a gene tree with a given species tree, as well as for inferring a species tree from a collection of gene trees, both under the assumption that gene tree heterogeneity is caused by ILS. Maddison and Knowles (2006) later implemented and tested a heuristic for inferring a species tree under the MDC criterion. Than and Nakhleh (2009) provided a mathematical characterization of the number of deep coalescences given a clade in the species tree (without having the species tree itself), which allowed for developing exact algorithms for species tree inference under the MDC criterion. To account for hybridization and introgression simultaneously with ILS, Yu et al. (2013) introduced the MDC criterion for inferring phylogenetic networks from a collection of gene trees whose heterogeneity is assumed to be caused by ILS and introgression. All these works on the MDC criterion naturally allow for including multiple individuals per species.

While the method of Yu et al. (2013) could in practice be used on data with allopolyploid species while treating multiple gene copies as alleles from different individuals, the criterion is not mathematically designed for handling polyploids. Let us illustrate this issue with the scenario depicted in Figure 1c. The gene tree shown inside the branches of the phylogenetic network has two copies from each of the three taxa X, Y, and Z, due to the polyploid hybridization event involving taxa B and C. If each of the two copies are treated as alleles from two different individuals, then this gene history will be heavily penalized by the MDC criterion as lineages failed to coalesce on the four branches connected to X, to Y, to Z, and to the most recent common ancestor (MRCA) of X and Y. Indeed, the MDC score of this gene tree given the phylogenetic network is 4 in this case. However, when considering that this is a polyploid hybridization event and the two lineages correspond to two different copies of the gene, these lineages should not be expected to coalesce on these four branches, and the true score of this phylogenetic network/gene tree reconciliation should be 0; that is, this is a gene tree that "perfectly" fits the species evolutionary history with no deep coalescence events. It is this issue that led Oberprieler et al. (2017) to justifiably avoid the method of Yu et al. (2013), rather applying a permutation scheme to it so as to sample gene copies and not leave them in the analysis and treat them as alleles from different individuals of the species. We now describe a modification to the MDC criterion so that it properly handles polyploid hybridization events. This extension can be viewed as a generalization of Thomas et al. (2017) by allowing for ILS.

Using the notation of (Huber et al., 2016), we denote by $U(\Psi)$ a MUL-tree representation of phylogenetic network Ψ, and we denote by $F(T)$ a phylogenetic network that corresponds to MUL-tree T. As Huber et al. (2016) showed, neither $U(\Psi)$ nor $F(T)$ are unique in general, though they are unique for special classes of phylogenetic networks. While the leaves of a gene tree are uniquely mapped to the leaves of a phylogenetic network (since each species labels exactly one leaf in the network), this is not the case for MUL-trees. For example, for the gene tree in Figure 1d and the MUL-tree of Figure 1b, each of the two copies x_1 and x_2 could map to either of the two leaves labeled by X in the MUL-tree, leading to 64 possible allele mappings in total (here, multiple alleles correspond to multiple copies of a gene obtained from the same individual): x_1 and x_2 could both map to the same X in the MUL-tree (there are two choices) or each to a different X (there are two choices, and the same for the Y and Z alleles, resulting in $4\times4\times4=64$ possible mappings. However, the number of possible allele mappings could be greatly reduced if information is given on the subgenome assignment of each gene copy. Figure 2a shows all possible allele mappings for these gene trees and MUL-tree when the allopolyploid alleles are known to come from distinct subgenomes. For example, in this case, it is assumed that x_1 and x_2 did not come from the same subgenome, and similarly for the Y and Z alleles. It is worth noting here, though, that delineating subgenomes could be challenging in ancient polyploids due to extensive homeologous exchanges (Edger et al. 2018) and in groups of species whose evolutionary histories involve deep coalescence. Below we describe our new method that is implemented to run in either of two modes: assuming a known subgenome assignment or inferring such an assignment during phylogenetic inference.

Given the set $\mathcal{F}$ of all allele mappings of a gene tree g to a MUL-tree T, the number of extra lineages of g given T is

$$XL(T,g)=\min_{f\in\mathcal{F}} XL(T,g,f). \tag{1}$$

Here, $XL(T,g,f)$ is the number of extra lineages given a specific allele mapping f, which is the sum, over all branches of the MUL-tree, of the number of extra lineages on each branch given the allele mapping f. The number of extra lineages resulting from each of the allele mappings in Figure 2a is shown in the same panel, and illustrations of how these quantities are computed for two of the allele mappings are shown in Figure 2b,c.

Given a collection $\mathcal{G}$ of gene trees, inferring a phylogenetic network Ψ^* under the criterion of minimizing the number of extra lineages amounts to computing

$$\Psi^* \leftarrow \text{argmin}_{\Psi}\left(\sum_{g\in\mathcal{G}} XL(U(\Psi),g)\right). \tag{2}$$

Based on this formulation, the inference is done by walking the space of phylogenetic networks using the implementation of Yu et al. (2013), while evaluating the number of extra lineages on the MUL-tree representation of each network, using Equation (1). The

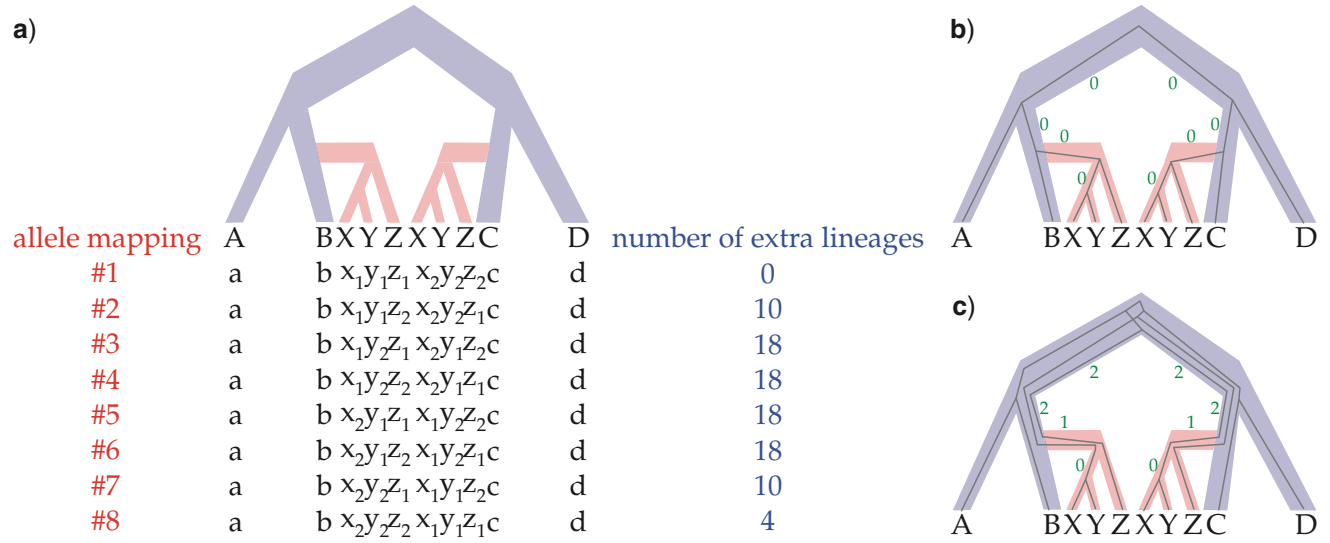

FIGURE 2. Allele mappings from a gene tree to a MUL-tree and the number of extra lineages. a) The set of all eight possible allele mappings for the gene tree of Figure 1d and the MUL-tree of Figure 1b given that each copy of allopolyploids came from distinct subgenome, along with the number of extra lineages that results from each of the allele mappings. b–c) The reconciliations that correspond to allele mappings #1 and #7, respectively, of the gene tree within the branches of the MUL-tree. Shown next to each of the MUL-tree branches is the number of extra lineages on that branch, which is the number of lineages that persist without coalescing on that branch minus 1.

limitations of the method as formulated are that it cannot detect homoploid hybridization and assumes no recombination between parental subgenomes.

PhyloNet Implementation

We implemented the inference based on Equation (2) in PhyloNet (Than et al. 2008; Wen et al. 2018) as a new command called `InferNetwork_MP_Allopp`. This method takes as input a set of gene trees and reports the top-k species networks and their MUL-trees based on the MUL-tree score. If during the search there are multiple networks with the same scores encountered, those networks will be reported if their scores are within the top-k scores, although the total number of networks returned will be truncated at k, which by default is set to 5.

- The subgenome assignment. This method makes use of the known subgenome identity for each sampled allopolyploid allele to determine the (reduced) set of candidate allele mappings to use during the phylogenetic network inference. We refer to the new method that makes use of this information as MPAllopp ("MP" is for maximum parsimony). In the absence of information about the assignment of gene sequences to subgenomes, MPAllopp considers all possible subgenome assignment per-locus. However, this requires the addition of heuristics due to the exponential growth in the number of ways alleles may be mapped to subgenomes (i.e., we may have to consider an extraordinarily large number of sets of assignments) when computing the parsimony score.
- The maximum number of reticulations allowed during the search. If this number is set at 0, then the method searches the space of genome trees (Thomas et al. 2017; Oxelman et al. 2017) only.
- Whether the hybrid species are known. If the hybrid species are known and specified by the user, and the maximum number of reticulations allowed equals the number of specified hybrid species, the method searches only networks that have the specified hybrids; that is, it does not detect any other potential hybrids. If the maximum number of reticulations allowed is greater than the number of specified hybrid species, the method could detect additional hybrids. If the hybrids are not specified, then the method identifies hybrids.

In addition to phylogenetic network inference, we provide the user with the option to evaluate, rather than infer, competing hypotheses. Given a phylogenetic network Ψ and a set of gene trees $\mathcal{G}$, the parsimony score of the Ψ is given by

$$\left(\sum_{g\in\mathcal{G}} XL(U(\Psi),g)\right).$$

This analysis is enabled by the new PhyloNet command `DeepCoalCount_AlloppNet`. We envision this command used in at least two contexts. First, if the user has two or more evolutionary hypotheses (in the form of phylogenetic networks), this command can be used to assess which of these hypotheses has the best score. Second, if the inference method above returns a phylogenetic network that does not match some biological knowledge, the user can manipulate

the inferred network to obtain one that matches the biological knowledge and compare the two. In other words, this command can be used in an exploratory mode.

Last but not least, a brief explanation of the nature of heuristic searches is in order. Inferring the optimal phylogenetic network according to Equation 2 is computationally very demanding. Therefore, the algorithm implemented by the command `InferNetwork_MP_Allopp` performs a random walk in the space of phylogenetic networks, evaluates phylogenetic networks encountered during the walk, and returns the best (i.e., the one with the lowest parsimony score) network among all those encountered. For each run, the search starts from either a random tree or a starting tree quickly estimated from the data using MDC (Than and Nakhleh 2009) or NJ_{st} (Liu and Yu 2011). Then the search explores the network space by modifying the current topology. Six moves were employed to alter the network topology: i) reticulation edge addition, which creates a new hybridization event; ii) reticulation edge deletion, which removes one hybridization event; iii) reticulation edge tail relocation, which modifies a parent lineage of a hybridization event; iv) reticulation edge direction flip, which reverses the direction of the "gene flow" underlying the hybridization event; v) reticulation edge replacement, which replaces an existing hybridization event by a completely different one; and vi) reticulation edge head relocation, which modifies the hybrid lineage resulting from the hybridization event. A new network is proposed by applying a randomly chosen operation from the six, proportional to their associate weights, to the current network. The score of the resulting network is calculated, and it is accepted or rejected based on this score. If the score of the proposed topology is worse, we reject the proposal, and the search continues from the current network. If it is accepted, the resulting network replaces the current one and the search proceeds from it. This type of heuristics is not guaranteed to find the optimal network, which is why it is recommended to run the command multiple times and return the optimal solution among all the runs (in the simulation experiments below, we ran the method 30 times on each data set). The search heuristic computes the parsimony score of each network candidate encountered during the search by considering the set of all allele mappings that satisfy the known subgenome assignment.

If the subgenome assignment is not known, we found that it was only possible to estimate this assignment and the genome tree with the addition of further heuristics in more complex scenarios, due to the previously mentioned combinatorial explosion. We implemented a similar heuristic for reducing the number of mappings as in Thomas et al. (2017) to efficiently evaluate allele mappings without exhaustively considering all of them as their number could be prohibitive even for a small number of species. This heuristic works as follows for each gene tree:

1. Determine the set of clades $C_1, \ldots, C_k$ that include all leaf species with polyploid alleles and exclude all species with diploid alleles.

2. For each clade C_i

   - If the sister node D_i to the root node of C_i is diploid, then the leaves of C_i will be mapped to the subgenome most related to the genome of D_i
   - If the leaves of C_i come from different polyploid species, then they will be mapped to the same subgenome

Simulations

We used the AlloppDT simulator (Jones 2012) and parameter settings of Jones et al. (2013) and Jones (2017) to simulate a collection of data sets. The data were generated under four different evolutionary scenarios with 1, 2, 3 and 3 reticulations (Fig. 3), where for each scenario the mutation rate, the number of individuals per species, and the number of genes were varied. In addition, we increased the number of genes to 100 to emulate the kinds of data sets enabled by next-generation sequencing. In total, there were 116 model conditions, each with 10 replicate data sets (Table 1). Since the gene tree-based inference methods considered in this study require rooted gene trees, an outgroup species was added for the purpose of rooting the estimated gene trees Table 1 and was excluded in the species phylogeny reconstruction. It would be possible for molecular clock methods to root the gene trees by finding the tree with the best ultrametric fit, but this may be unreliable when the strict molecular clock assumption is violated.

For each gene tree, DNA sequences of length 500 were generated using Seq-Gen Version 1.3.2 (Rambaut and Grass 1997) under the HKY model. Then gene trees were reconstructed from the simulated sequence alignments using IQ-TREE Version 2.1.2 (Minh et al. 2020), and rooted by the outgroup.

TABLE 1. AlloppDT parameters for data sets with allopolyploidy. Based on the species phylogenies in Fig. 3, we varied the parameters and obtained 116 conditions. For each condition, 10 replicates were simulated.

series Scenario	Parameters
D E F	G (Number of genes) $\in \{1,3,10,100\}$ N (Individuals per species) $\in \{1,3,9\}$ T (Mutation rates) $\in \{4e-9, 2e-8, 1e-7\}$ substitutions per site per generation H (Root height)=0.02 substitutions per site
J	G (Number of genes) $\in \{1,3,10,100\}$ N (Individuals per species) $\in \{1,3\}$ T (Mutation rates) $=2e-8$ substitutions per site per generation H (Root height)=0.045 substitutions per site

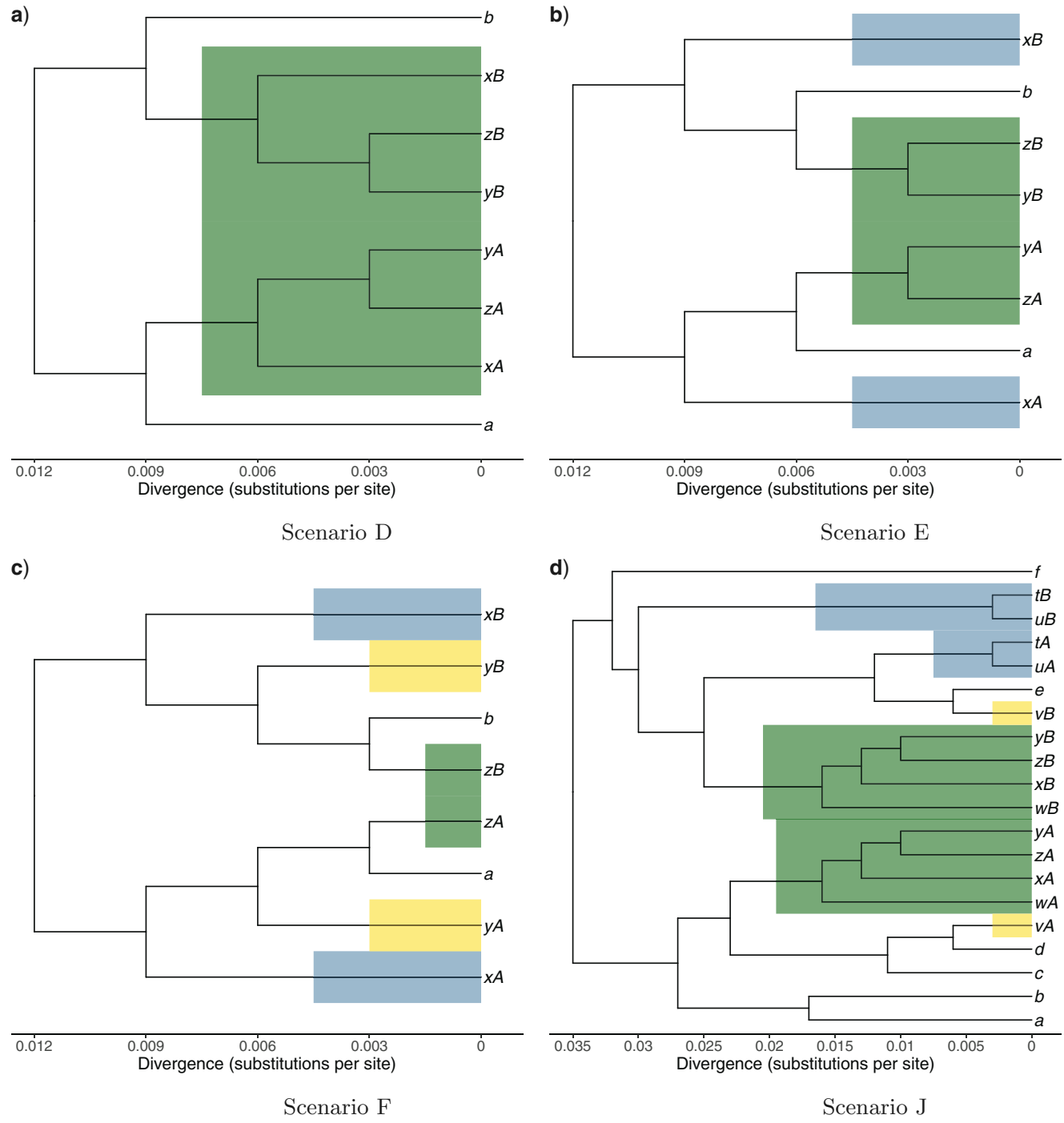

FIGURE 3. Multilabeled species trees used in allopolyploid simulations (taken from Jones (2017)). a) Scenario D with 5 species has one hybridization event that gave rise to the allopolyploid clade {x,y,z}. b) Scenario E with five species has two hybridization events that led to the allopolyploid clades {x} and {y,z}. c) Scenario F with five species has three hybridization events that resulted in the allopolyploid clades {x}, {y}, and {z}. Scenario J with 13 species has three hybridization events that produced the allopolyploid clades {v}, {t,u}, and {x,y,z,w}.

Comparison with other methods—The AlloppNET input XML files were generated by AlloppDT with the same model parameters and MCMC settings as used in Jones (2017). This includes coestimating the assignment of alleles to subgenomes.

We tested MPAllopp (with known subgenome assignment) and compared it to three methods:

- `InferNetwork_MPL` in PhyloNet, which implements the maximum pseudolikelihood method of Yu and Nakhleh (2015) (this is labeled MPL below).

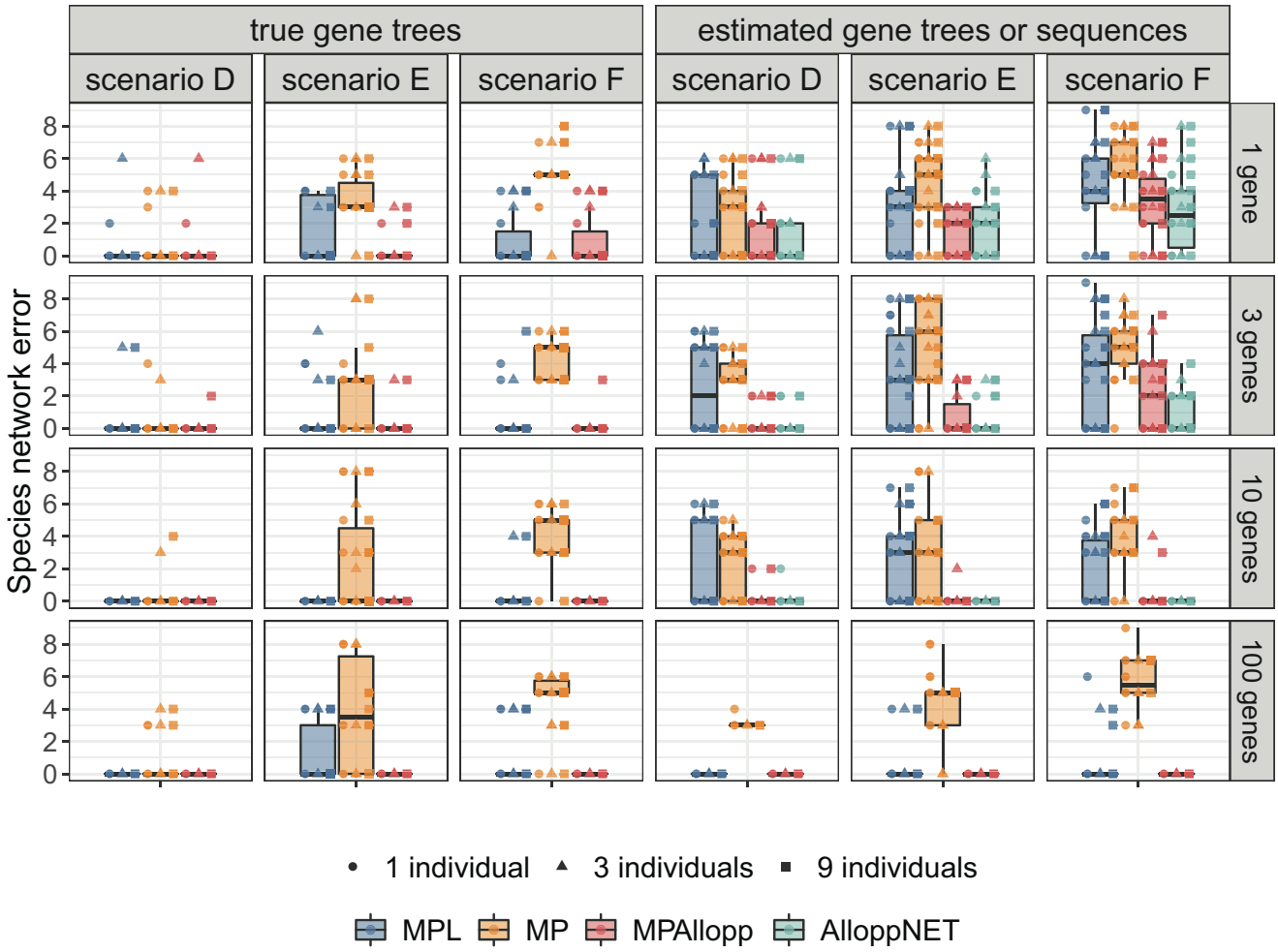

FIGURE 4. Boxplots of inference accuracy on simulated data with low ILS (mutation rate of 4×10^{-9}) over 10 replicates. Data sets with 100 genes were too large for AlloppNET to converge within a reasonable amount of time, so they are omitted from the 100- gene results shown here. Furthermore, scenario J is not shown since only one mutation rate (2×10^{-8}) was used.

- `InferNetwork_MP` in PhyloNet, which implements that MDC method of Yu et al. (2013) (this is labeled MP below).
- AlloppNET (Jones et al. 2013; Oxelman et al. 2017).

When running MPL and MP, multiple gene copies were treated as different alleles from the same species, since these methods do not incorporate any association between gene copies and specific individuals.

Due to the nondeterministic nature of the search heuristics underlying all gene tree-based methods used here, each method (except for AlloppNET) was run 30 times on each replicate data set and the optimal solution across all 30 runs was returned.

Evaluating inferences—We evaluated both the accuracy and the runtime of methods. The accuracy was measured in terms of the error in the inferred network compared to the true network using the metric of Nakhleh (2010). The experiments were performed on a server equipped with 2.2GHz Intel(R) Xeon(R) Gold 5220R CPUs, each with 48 cores. All the methods were executed using single thread, and the execution time was measured in CPU time spent in user mode.

RESULTS AND DISCUSSION

Results on Simulated Data

Treating Gene Copies as Alleles From Different Individuals: Performance of MPL and MP—MPL and MP are not intended to model the evolutionary relationship between subgenomes, and hence, unlike our new methods do not include any mechanism for assigning alleles to subgenomes. Despite this, we wished to analyze whether MPL and MP could accurately model the relationship between species, by treating alleles from different subgenomes as being randomly sampled from a species without subgenome structure.

The accuracy results of the methods on data sets with low, moderate, and high levels of ILS are shown

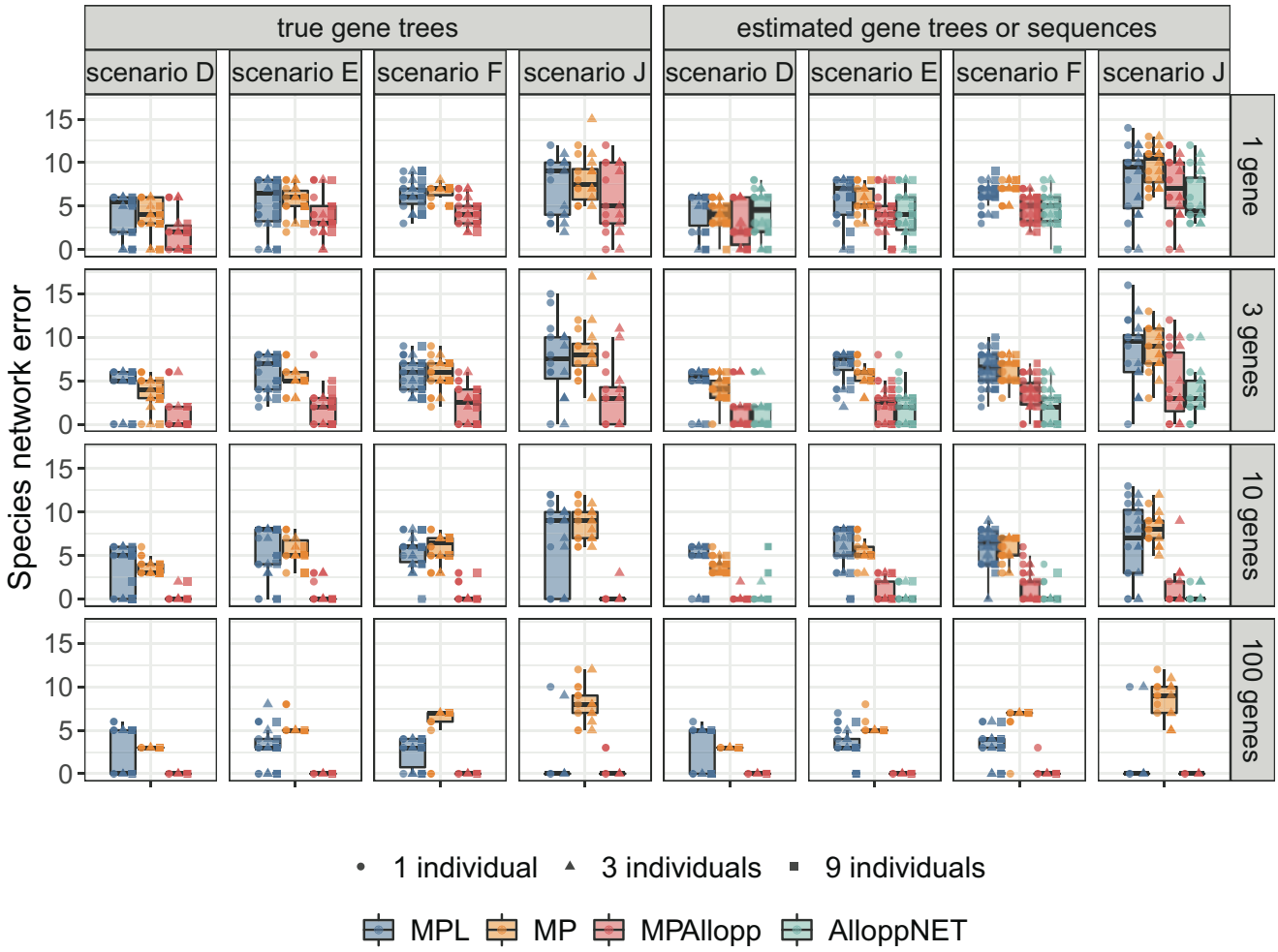

FIGURE 5. Boxplots of inference accuracy on simulated data with moderate ILS (mutation rate of 2×10^{-8}) over 10 replicates. Data sets with 100 genes were too large for AlloppNET to converge within a reasonable amount of time, so they are omitted from the 100-gene results shown here.

in Figures 4, 5, and 6, respectively, where the species network error was measured as the distance between the true and estimated networks according to the metric of Nakhleh (2010). It is important to note that AlloppNET uses sequence data as input, therefore it was compared against other gene tree-based methods using estimated gene trees but not the true gene trees.

As the results show, both MPL and MP infer very accurate networks when using the true gene trees of scenario D with low levels of ILS, which is the simplest of all four scenarios considered, as it contains a single hybridization event. However, this observation changes significantly for the same scenario when the level of ILS increases and/or when gene tree estimates as used. In particular, while still using the true gene trees, as the level of ILS increases, the error rates of both methods increase, with MP performing slightly better. We hypothesize that MP performs slightly better because it is a simple summary method that does not rely on the mathematical assumptions of the coalescent model, which are violated here and are used explicitly by MPL. Specifically, coalescent methods assume that alleles can be inherited at random from any individual chromosome in the previous generation, but when subgenomes are nonrecombining, alleles can only be inherited from chromosomes belonging to the same subgenome. For low levels of ILS, gene tree estimation error significantly impacts the performance of the methods; however, that gene tree estimation error does not make the performance much worse in the presence of higher levels of ILS. We determined that the difference in error rates is somewhat artifactual, as it is being caused by the inability to resolve relationships between individuals at each locus within each extant species, due to very short coalescent times when ILS is low (Supplementary Fig. S1 available on Dryad at http://dx.doi.org/10.5061/dryad.4xgxd25b0.).

Scenario E has two hybridization events, Scenario F has three hybridization events, and Scenario J has three hybridization events and is much larger. Therefore, these scenarios are more challenging for the methods than Scenario D, and they increase in complexity from

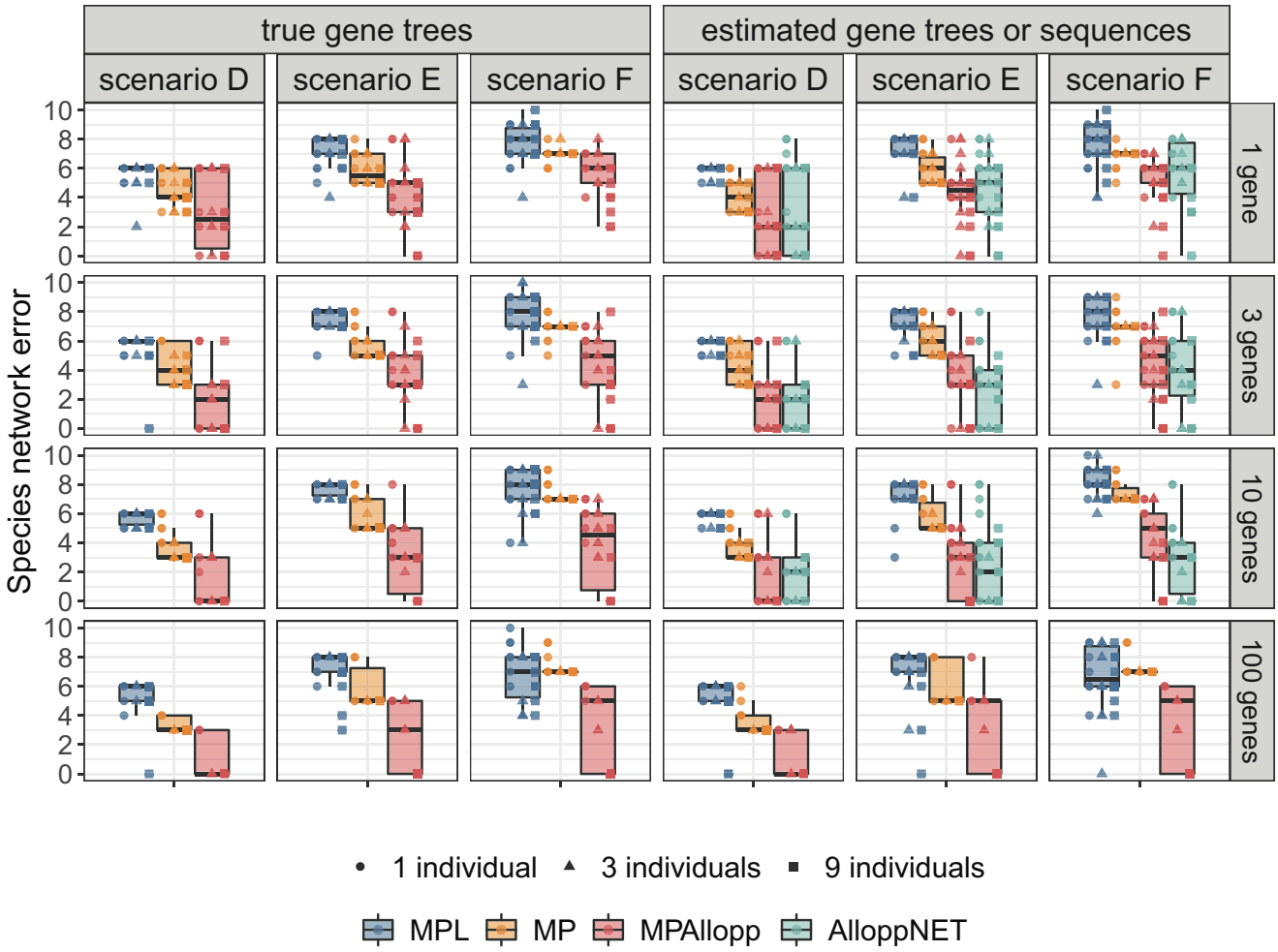

FIGURE 6. Boxplots of inference accuracy on simulated data with high ILS (mutation rate of 1×10^{-7}) over 10 replicates. Data sets with 100 genes were too large for AlloppNET to converge within a reasonable amount of time, so they are omitted from the 100-gene results shown here. Furthermore, scenario J is not shown since only one mutation rate (2×10^{-8}) was used.

Scenario E to F to J. We observe the impact of the increase in complexity on the performance of MPL and MP, as both infer less accurate networks on these scenarios than on the data of Scenario D.

Finally, it worth noting that increasing the numbers of genes does not seem to affect the performance of these two methods, with the only exception observed when using 100 gene trees on the low-ILS conditions of Scenarios D and E, where the performance improved significantly over smaller numbers of genes when using gene tree estimates.

While inference under the MSNC using MPL could result in the correct phylogenetic network topology, several of the branch lengths could be inferred incorrectly. Consider the scenario of Figure 1c. Since inference under the MSNC assumes that x_1 and x_2 are two alleles of species X, and the same for the pair (y_1,y_2), and the pair (z_1,z_2), the lengths of all four branches in the subtree ((X,Y),Z) would be underestimated to account for the absence of coalescence events on these branches. In other words, while MPL happened to provide good results in some of these simulations in terms of the phylogenetic network topology, it did so at the expense of the branch lengths. For example, for scenario F (see Fig. 3c), the coalescent times of the homeologous alleles from the tetraploid z have to be more ancient than the divergence time between species a and b. That is to say, no homoeologs could coalesce along the two horizontal edges or the branch connected to z. So, even when the tetraploid z was correctly inferred as the hybrid species, MPL worked by forcing the age of the hybridization to be zero.

Accounting for Allopolyploidy Explicitly: Performance of MPAllopp and AlloppNET—For low levels of ILS, MPAllopp outperforms both MPL and MP in particular when using gene tree estimates. Furthermore, increasing the number of genes has a noticeable positive impact on the performance of MPAllopp. For moderate and high levels of ILS, MPAllopp outperforms both methods significantly when using both the true or estimated gene trees, across all scenarios. For low levels of ILS,

AlloppNET and MPAllopp performed almost identically on Scenario D. For Scenarios E and F, AlloppNET outperformed MPAllopp slightly when using 1 or 3 genes, but they both had almost identical performance when using 10 genes. For higher levels of ILS, AlloppNET could outperform MPAllopp, especially when using a larger number of genes. However, increasing the number of genes also increases the computational requirements of AlloppNET. For 100-gene data sets, 29/30 AlloppNET chains converged (effective sample size for log-posterior density samples $\geq$ 200) for scenarios D, E, and F given a single individual and high ILS. But for other model conditions only 26/240 chains converged after at least 46 h of running time, so we excluded AlloppNET from our analysis of 100-gene data sets.

Overall, counting the number of times AlloppNET and MPAllopp inferred the correct network, MPAllopp successfully recovered the true network in 44% of replicates across all combinations that are being compared of scenario, mutation rate, and number of individuals (56% of the replicates when using the true gene trees). AlloppNET, on the other hand, performed slightly better, inferring the true network in 48% of the cases. We also noted that MPAllopp had an advantage over AlloppNET when the number of individuals per species increased to 9 under high ILS, with a mean error of 1.57 as compared to 1.71.

As mentioned above, AlloppNET is more powerful than the other methods studied here in that it estimates other evolutionary parameters (beside the network topology) including sampling gene trees from the posterior distribution. Normally AlloppNET is initialized with user-supplied assignments of alleles to subgenomes, but will coestimate these assignments along with the gene and species phylogenies. To assess the quality of gene trees estimated by AlloppNET, we quantified their accuracy and compared it to that of gene trees estimated from the sequence alignments using IQ-TREE. As AlloppNET returns a sample of gene trees per locus, we computed for each locus the maximum clade credibility (MCC) tree. As Supplementary Figure S1 available on Dryad shows, AlloppNET infers more accurate gene trees, owing mainly to the fact that AlloppNET samples the gene trees and phylogenetic networks simultaneously. The average gene tree error of data sets with low, moderate, and high ILS were 0.39, 0.27, and 0.23, respectively for IQ-TREE gene trees, whereas they were 0.27, 0.21, and 0.15, respectively for AlloppNET.

We also evaluated the performance of MPAllopp when the subgenome assignment is unknown, a method we call MPAllopp*. Based on this evaluation, estimates of the species network are much higher in error when the subgenome assignment must be estimated using the heuristic we describe in Methods (Supplementary Fig. S2 available on Dryad). Running time was generally unaffected, except for scenario J where it increased by up to 20 minutes (Supplementary Fig. S3 available on Dryad).

Running Times of the Methods—We recorded the computational time in CPU minutes for each method on each data set; the results for low, moderate, and high levels of ILS are shown in Figures 7, 8, and 9, respectively.

As expected, AlloppNET was in general the slowest method (71 min on average), and MPL was the second slowest (25 min on average). The parsimony methods typically took 1–2 min to complete; however, for MP, the variation in running time was noticeable when gene tree estimation error was involved, especially when more individuals were included. In fact, MP failed to complete on 2 and 28 data sets of gene tree estimates with 3 and 9 individuals per species, respectively.

As the results show, Scenario D in general takes the least amount of time for the methods to analyze since it is the simplest in terms of the number of hybridization and the location of the hybridization events. Scenarios E and F are more complex and take more time, and Scenario J, which is the largest, takes the most time. Furthermore, in many cases of 3 or more genes, AlloppNET takes time that is two order of magnitude larger than that of MPAllopp. For 100 gene trees, and as mentioned above, AlloppNET did not converge within 46 h. Finally, MP had the largest variability in running time over data sets with the same parameter settings, a trend that is clearest on data sets with high ILS levels.

In summary, the simulation results show that AlloppNET and MPAllopp have comparable performance in terms of accuracy of the phylogenetic network topology, yet MPAllopp is much faster than AlloppNET.

Analysis of Biological Data Sets

A Triticum (Poaceae) data set—We used the bread wheat data set with 275 gene trees from Marcussen et al. (2014). The data set includes the allohexaploid *Triticum aestivum* (Ta), its five diploid relatives *Triticum monococcum* (Tm), *Triticum urartu* (Tu), *Aegilops sharonensis* (Ash), *Aegilops speltoides* (Asp), and *Aegilops tauschii* (At), together with three outgroup species *Hordeum vulgare* (Hv), *Brachypodium distachyon* (Bd), and *Oryza sativa* (Os). AlloppNET was not run on this data set since it only deals with diploids and tetraploids. Given that the bread wheat contains six sets of chromosomes, we set the maximum number of allowable reticulations to 2 and 3. Using our method with the genome identities being assigned to the alleles of the allopolyploid bread wheat, we obtained the optimal results after 30 runs of search, as shown in Figure 10.

Although there are three equally parsimonious networks inferred, they share the same underlying MUL-tree. As described above, MPAllopp searches the network space while evaluating the network based on its MUL-tree representation. In this case, three networks correspond to the same MUL-tree. This nonuniqueness of the network nothwithstanding, the three networks point to an evolutionary hypothesis with

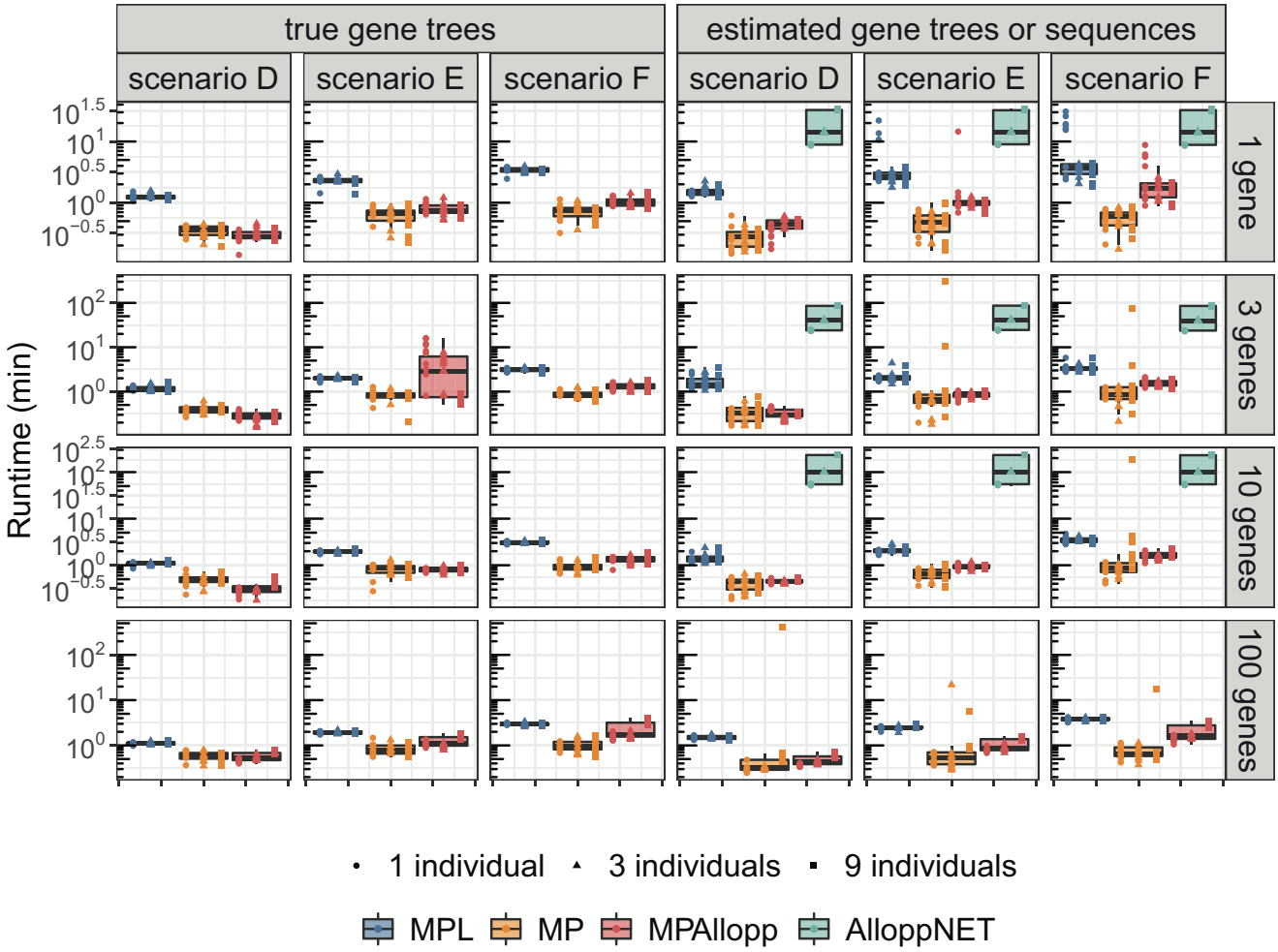

FIGURE 7. Boxplots of running time (in CPU minutes) on simulated data with low ILS (mutation rate of 4×10^{-9}) over 10 replicates. Data sets with 100 genes were too large for AlloppNET to converge within a reasonable amount of time, so they are omitted from the 100-gene results shown here. Furthermore, scenario J is not shown since only one mutation rate (2×10^{-8}) was used.

two hybridization events leading to the formation of *T. aestivum*. Furthermore, the identified putative diploid progenitors, namely *Ae. speltoides*, *Ae. tauschii* and *T. urartu*, were consistent with previous studies (Marcussen et al. 2014). The three networks differ in the order in which these three species hybridized to give rise to the two hybridization events shown.

A Pachycladon (Brassicaceae) data set— We reanalyzed the *Pachycladon* data set with five nuclear single-copy genes from Joly et al. (2009). We included eight tetraploid species of the *Pachycladon* genus, coupled with two diploids *Arabidopsis thaliana* and *Lepidium apelatum* as in Jones et al. (2013). For each gene, BEAST Version 2.6.4 was employed to infer a sample of trees, based on which we computed the maximum clade credibility tree by the utility program TreeAnnotator in BEAST. This produced five gene tree estimates in total (Supplementary Fig. S4 available on Dryad). We ran MPAllopp supplied with the genome identities of the allele copies from the *Pachycladon* species, varying from one to three reticulations. Results are shown in Figure 11 and Supplementary Figure S5 available on Dryad.

The main difference between these two phylogenetic networks is the placement of *P. wallii*: In Figure 11a, *P. wallii* is placed as basal to the clade formed by *P. latisiliqua*, *P. enysii*, *P. fastigiata*, and *P. stellata*, whereas in Figure 11b it is grouped with *P. latisiliqua* as sisters. The latter is in full agreement with the result reported in Jones et al. (2013) using the same data. Interestingly, *P. wallii* is traditionally regarded as a sister species to *P. novaezelandiae* considering that they both live in the southern South Island of New Zealand. Such conflicts were probably caused by missing data as *P. novaezelandiae* only has sequences for the CHS gene in this data set, and one of the gene copy was grouped within the *P. latisiliqua* et al. clade while the other was a sister to *P. novaezelandiae*. It is also worth mentioning that though six networks with equal score were returned, there are only four unique underlying

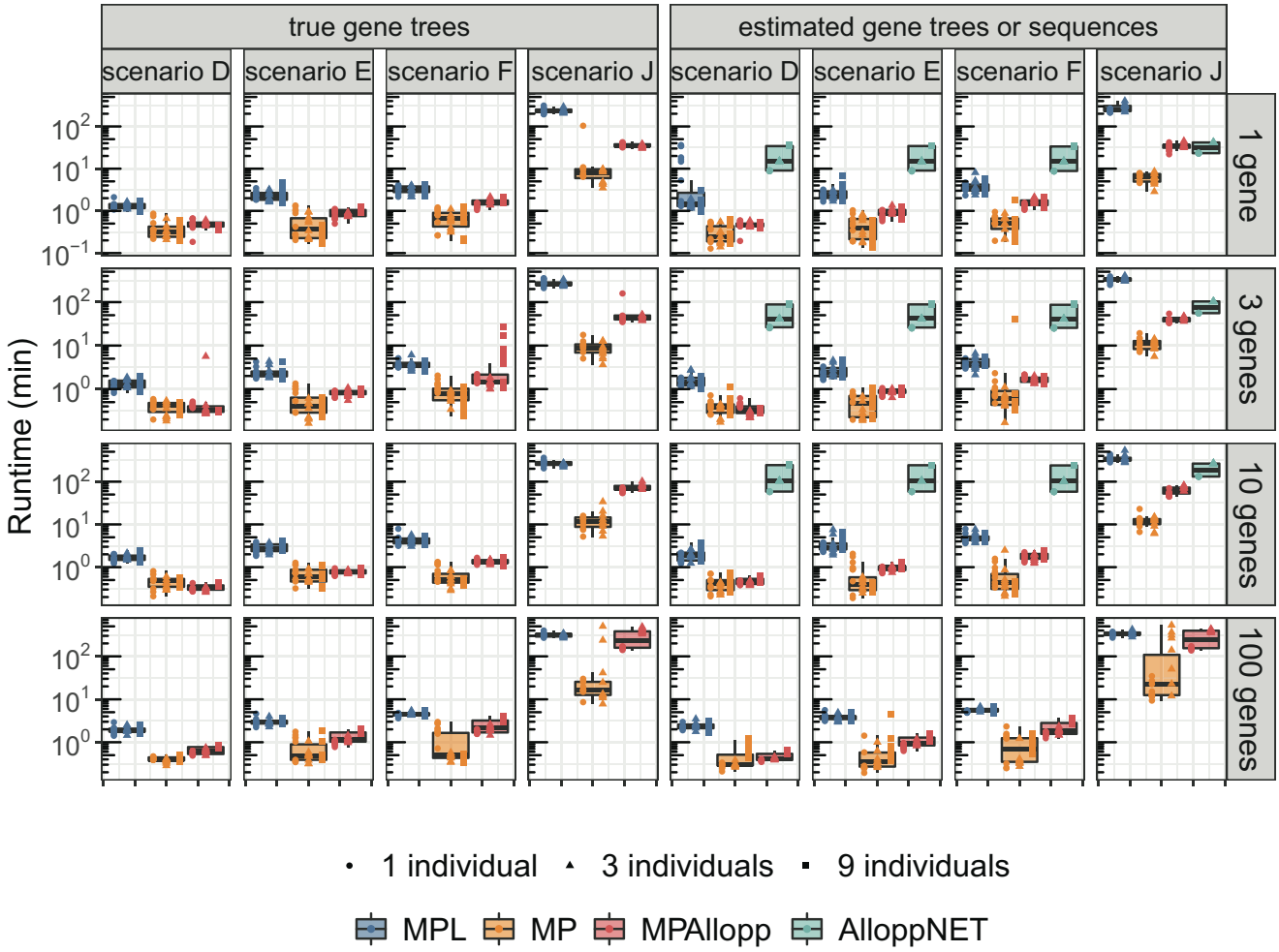

FIGURE 8. Boxplots of running time (in CPU minutes) on simulated data with moderate ILS (mutation rate of 2×10^{-8}) over 10 replicates. Data sets with 100 genes were too large for AlloppNET to converge within a reasonable amount of time so they are omitted from the 100-gene results shown here.

MUL-trees: Figure 11a and Supplementary Figure S5a available on Dryad are associated with the same MUL-tree despite different number of reticulations, so are Figure 11b and Supplementary Figure S5b available on Dryad. This again illustrates the identifiability issue for the conversion between MUL-trees and networks.

A potentially more interesting hypothesis based on the networks in Figure 11 is that this evolutionary history includes an autopolyploidization event at the most recent common ancestor of the in-group after the split from *A. thaliana*. That is, an alternative evolutionary scenario here is a species tree with no interspecific hybridization. These two alternative hypotheses have an equal score under the criterion optimized by MPAllopp and, thus, cannot be distinguished based on the data provided.

CONCLUSIONS

In this article, we introduced a new maximum parsimony method, MPAllopp, for inferring phylogenetic networks from gene tree topologies while accounting for polyploidy and incomplete lineage sorting simultaneously. The method employs a heuristic search for walking the network space while evaluating the parsimony score on the MUL-tree representation of the network.

The lack of a one-to-one mapping between MUL-trees and phylogenetic networks notwithstanding (Huber et al. 2016; Zhu et al. 2016), "seeing" the polyploid hybridization events in a MUL-tree is possible only for simplistic scenarios: a small number of taxa, a small number of hybridization events, a small number of gene extinctions, and, most importantly, the absence of confounding factors such as ILS (Fig. 1). Indeed, the parsimony algorithms and methods of Huber and Moulton (2006) and Thomas et al. (2017) do not account for ILS. Identifying the hybridization events computationally is the task of turning the MUL-tree into a phylogenetic network after a MUL-tree is inferred from the gene trees. Therefore, our method searches the phylogenetic network space directly by applying

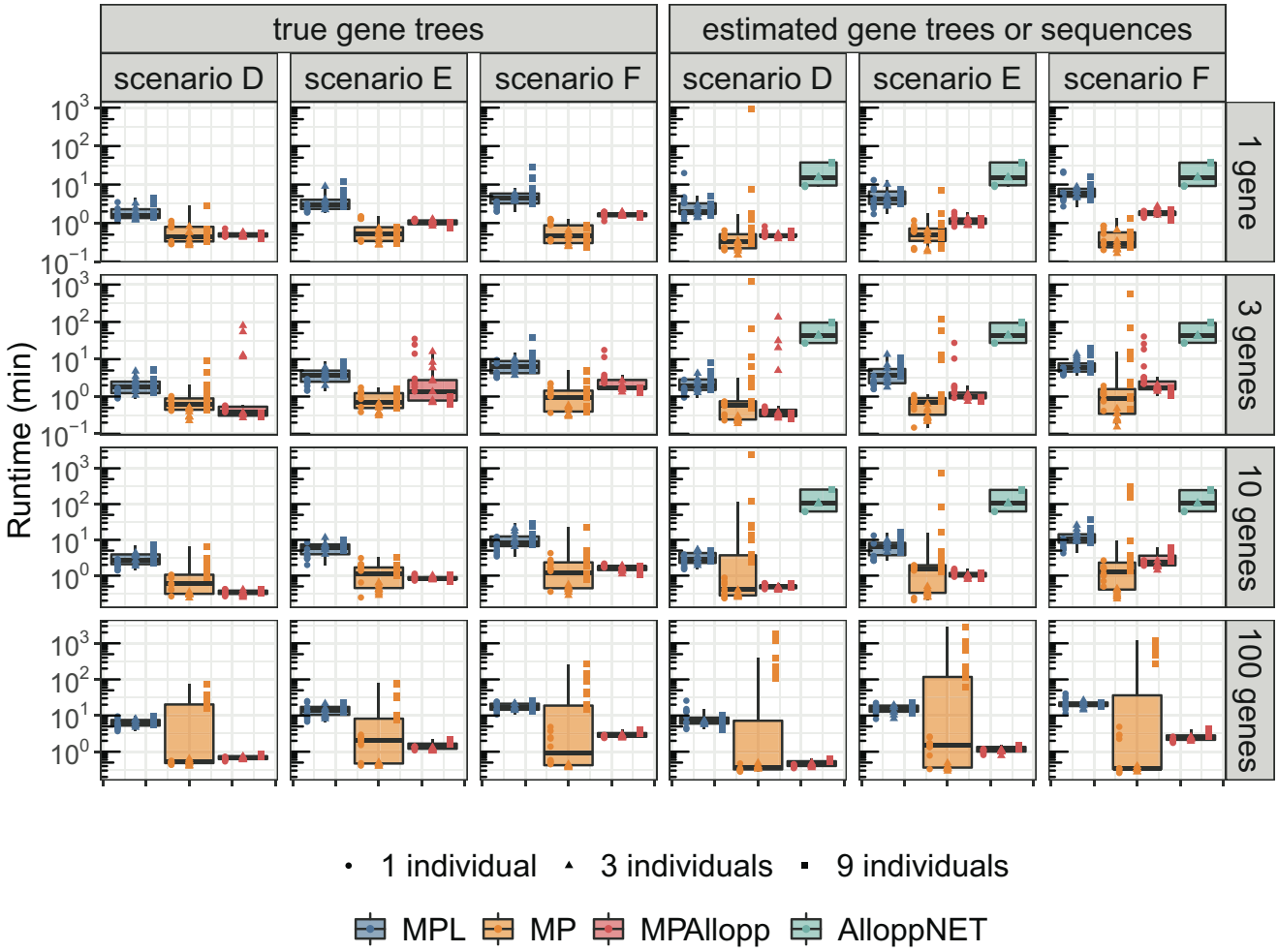

FIGURE 9. Boxplots of running time (in CPU minutes) on simulated data with high ILS (mutation rate of 1×10^{-7}) over 10 replicates. Data sets with 100 genes were too large for AlloppNET to converge within a reasonable amount of time so they are omitted from the 100-gene results shown here. Furthermore, scenario J is not shown since only one mutation rate (2×10^{-8}) was used.

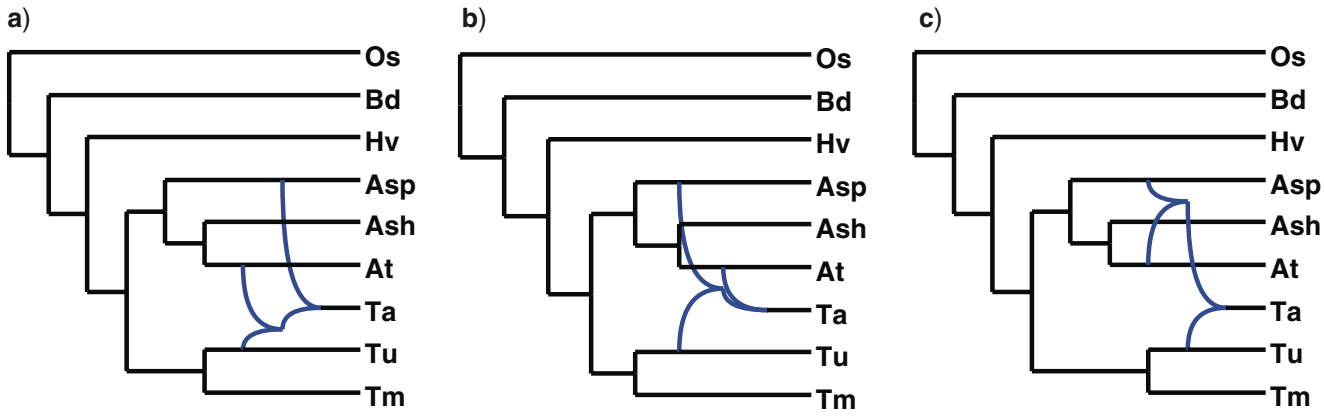

FIGURE 10. Results on bread wheat data. Species networks inferred by MPAllopp from the 9-taxon wheat data set with 275 reconstructed gene trees from (Marcussen et al., 2014). The three networks have the same score under the criterion optimized by MPAllopp.

subnetwork transfer operations on networks, so that the inference result is a network, rather than a MUL-tree.

In addition, the method proposed here assumes allopolyploidization. However, in nature, homoploid hybrids, allopolyploids, and autopolyploids could

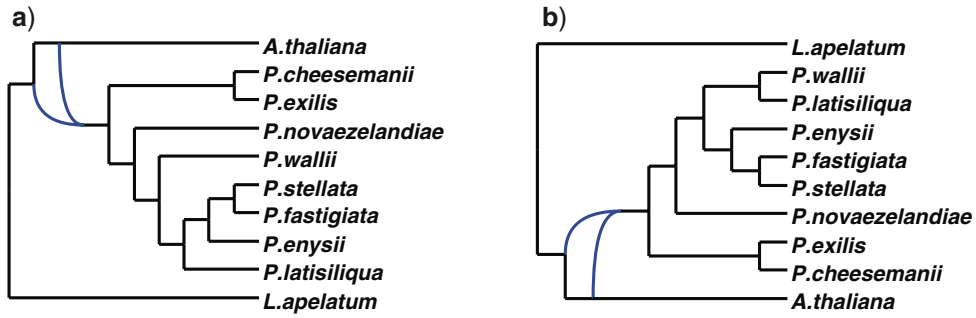

FIGURE 11. Results on *Pachycladon* data. Species network reconstructed from the 10-taxon *Pachycladon* data set with five genes from (Joly et al., 2009). The two networks have the same score under the criterion optimized by MPAllopp.

coexist, which would necessitate the development of methods incorporating various types of hybridization and polyploidization events. However, as we illustrated in the case of the Pachycladon data set above, some patterns in the phylogenetic network could potentially signal an autopolyploidization event. Currently, our method cannot infer autopolyploidy and requires the diploid progenitors to be sampled, as we have not implemented support for a data structure that can represent WGD in the absence of diploid progenitor species, or if those species never existed. We plan on implementing it in a future version.

As discussed in Blischak et al. (2018), statistical modeling of phylogenetic networks with polyploid hybridization is very complex. We believe that devising stochastic models and inference methods for restricted classes of polyploids, as in Jones et al. (2013); Oxelman et al. (2017), is most likely the way to make progress in this area.

In the last several years, there has been work on combining the multispecies coalescent model with a birth–death model of gene duplication and loss (Wu et al. 2014; Rasmussen and Kellis 2012; Du and Nakhleh 2018; Li et al. 2020). While these works do not consider hybridization or allelic introgression, Du et al. (2019) introduced a model that unifies the multispecies network coalescent and a birth–death model thus allowing for simultaneous modeling of incomplete lineage sorting, gene duplication and loss, and diploid hybridization. These works could be relevant for further advances in modeling polyploid hybridization in phylogenomic inference.

Statistical inference of phylogenetic networks is computationally much more demanding than inference of trees, severely limiting the sizes of data sets that can be analyzed with phylogenetic network methods. One approach to handling larger data sets is to analyze smaller subsets of the data (subsets in terms of taxa). This approach could be automated, as in Zhu et al. (2019) for example, but this requires developing methods for accurately estimating small networks with their evolutionary parameters and for merging these small networks into a network on the full data set. We view this as an essential direction for future research for phylogenetic network inference on large data sets involving polyploids to become feasible. Another approach is to combine the speed of methods like MPAllopp with the capabilities of methods like AlloppNET. For example, given the accuracy of MPAllopp, a phylogenetic network topology is first inferred using MPAllopp, and then its associated evolutionary parameters are estimated using statistical methods such as AlloppNET.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.4xgxd25b0.

FUNDING

This work was supported by the National Science Foundation [CCF-1514177 and CCF-1800723 to L.N.].

REFERENCES

Berthelot C., Brunet F., Chalopin D., Juanchich A., Bernard M., Noël B., Bento P., Da Silva C., Labadie K., Alberti A., Aury J.M., Louis A., Dehais P., Bardou P., Montfort J., Klopp C., Cabau C., Gaspin C., Thorgaard G.H., Boussaha M., Quillet E., Guyomard R., Galiana D., Bobe J., Volff J.N., Genêt C., Wincker P., Jaillon O., Roest Crollius H., Guiguen Y. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nature Commun. 5:1–10.

Blischak P.D., Mabry M.E., Conant G.C., Pires J.C. 2018. Integrating networks, phylogenomics, and population genomics for the study of polyploidy. Annu.Rev. Ecol. Evol. Syst. 49:253–278.

Cao Z., Liu X., Ogilvie H.A., Yan Z., Nakhleh L. 2019. Practical aspects of phylogenetic network analysis using phylonet. bioRxiv.

Du P., Nakhleh L. 2018. Species tree and reconciliation estimation under a duplication-loss-coalescence model. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; Washington, DC. New York (NY): Association for Computing Machinery. p. 376–385.

Du P., Ogilvie H.A., Nakhleh L. 2019. Unifying gene duplication, loss, and coalescence on phylogenetic networks. In: Cai Z., Skums P., Li M., editors. International Symposium on Bioinformatics Research and Applications., LNBI 11490. Switzerland: Springer Nature. p. 40–51.

Edger P.P., McKain M.R., Bird K.A., VanBuren R. 2018. Subgenome assignment in allopolyploids: challenges and future directions. Curr. Opin. Plant Biol. 42:76–80.

Glasauer S.M., Neuhauss S.C. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol. Genet. Genomics 289:1045–1060.

Glover N.M., Redestig H., Dessimoz C. 2016. Homoeologs: what are they and how do we infer them? Trends Plant Sci. 21:609–621.
Huber K.T., Moulton V. 2006. Phylogenetic networks from multi-labelled trees. J. Math. Biol. 52:613–632.
Huber K.T., Moulton V., Steel M., Wu T. 2016. Folding and unfolding phylogenetic trees and networks. J. Math. Biol. 73:1761–1780.
Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.
Joly S., Heenan P.B., Lockhart P.J. 2009. A pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. Mol. Phylogenet. Evol. 51:365–372.
Jones G. 2012. Simulations for allopolyploid networks: AlloppDT. http://www.indriid.com/goteborg/2012-09-01-simulations.pdf.
Jones G. 2017. Bayesian phylogenetic analysis for diploid and allotetraploid species networks. bioRxiv.
Jones G., Sagitov S., Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62:467–478.
Kamneva O.K., Syring J., Liston A., Rosenberg N.A. 2017. Evaluating allopolyploid origins in strawberries (Fragaria) using haplotypes generated from target capture sequencing. BMC Evol. Biol. 17:180.
Li, Q., Scornavacca C., Galtier N., Chan Y.-B. 2020. The multilocus multispecies coalescent: a flexible new model of gene family evolution. Syst. Biol. 70:822–837.
Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.
Lott M., Spillner A., Huber K.T., Petri A., Oxelman B., Moulton V. 2009. Inferring polyploid phylogenies from multiply-labeled gene trees. BMC Evol. Biol. 9:216.
Maddison W. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.
Marcet-Houben M., Gabaldón T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. PLoS Biol. 13:e1002220.
Marcussen T., Sandve S.R., Heier L., Spannagl M., Pfeifer M., Jakobsen K.S., Wulff B.B., Steuernagel B., Mayer K.F., Olsen O.-A., et al. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:1250092.
Masterson, J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science 264:421–424.
Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Von Haeseler A., Lanfear R. 2020. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.
Muffato M., Crollius H.R. 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. Bioessays 30:122–134.
Nakhleh L. 2010. A metric on the space of reduced phylogenetic networks. IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB) 7:218–222.
Oberprieler C., Wagner F., Tomasello S., Konowalik K. 2017. A permutation approach for inferring species networks from gene trees in polyploid complexes by minimising deep coalescences. Methods Ecol. Evol. 8:835–849.
Ohno S. 2013. Evolution by gene duplication. Berlin, Heidelberg: Springer.
Oxelman B., Brysting A.K., Jones G.R., Marcussen T., Oberprieler C., Pfeil B.E. 2017. Phylogenetics of allopolyploids. Annu. Rev. Ecol. Evol. Syst. 48:543–557.
Rambaut A., Grass N.C. 1997. Seq-gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13:235–238.
Rasmussen M.D., Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22:755–765.
Sancho R., Inda L.A., Diaz-Perez A.J., Des Marais D.L., Gordon S., Vogel J., Lusinska J., Hasterok R., Contreras-Moreira B., Catalan P. 2021. Tracking the ancestry of known and 'ghost' homeologous subgenomes in model grass brachypodium polyploids. bioRxiv.
Than C., Nakhleh L. 2009. Species tree inference by minimizing deep coalescences. PLoS Comput. Biol. 5:e1000501.
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9:322.
Thomas G.W., Ather S.H., Hahn M.W. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst. Biol. 66:1007–1018.
Wen D., Nakhleh L. 2018. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. Syst. Biol. 67: 439–457.
Wen D., Yu Y., Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genetics 12:e1006006.
Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst. Biol. 67:735–740.
Woods I.G., Wilson C., Friedlander B., P. Chang, D.K. Reyes, R. Nix, P.D. Kelly, F. Chu, J.H. Postlethwait, W.S. Talbot. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res. 15:1307–1314.
Wu Y.-C., Rasmussen M.D., Bansal M.S., Kellis M. 2014. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Res. 24:475–486.
Yu Y., Barnett R., Nakhleh L. 2013. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Syst. Biol. 62:738–751.
Yu Y., Degnan J., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genetics 8:e1002660.
Yu Y., Dong J., Liu K., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. USA 111:16448–6453.
Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics 16:S10.
Zhu J., Liu X., Ogilvie H.A., Nakhleh L.K. 2019. A divide-and-conquer method for scalable phylogenetic network inference from multilocus data. Bioinformatics 35:i370–i378.
Zhu J., Nakhleh L. 2018. Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. Bioinformatics 34: i376–i385.
Zhu J., Wen D., Yu Y., Meudt H.M., Nakhleh L. 2018. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. PLoS Comput. Biol. 14:1–32.
Zhu J., Yu Y., Nakhleh L. 2016. In the light of deep coalescence: revisiting trees within networks. BMC Bioinformatics 17:415.