Systematic Biology Advance Access published June 4, 2013

Parsimonious Inference of Hybridization in the Presence of Incomplete Lineage Sorting

Yun Yu*† R. Matthew Barnett* Luay Nakhleh*†

*Dept. of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA.
†Corresponding author. Email: {yy9,nakhleh}@rice.edu; fax (713) 349-5930.

Abstract

Hybridization plays an important evolutionary role in several groups of organisms. A phylogenetic approach to detect hybridization entails sequencing multiple loci across the genomes of a group of species of interest, reconstructing their gene trees, and taking their differences as indicators of hybridization. However, methods that follow this approach mostly ignore population effects, such as incomplete lineage sorting (ILS). Given that hybridization occurs between closely related organisms, ILS may very well be at play and, hence, must be accounted for in the analysis framework. To address this issue, we present a parsimony criterion for reconciling gene trees within the branches of a phylogenetic network, and a local search heuristic for inferring phylogenetic networks from collections of gene-tree topologies under this criterion. This framework enables phylogenetic analyses while accounting for both hybridization and ILS. Further, we propose two techniques for incorporating information about uncertainty in gene-tree estimates. Our simulation studies demonstrate the good performance of our framework in terms of identifying the location of hybridization events, as well as estimating the proportions of genes that underwent hybridization. Also, our framework shows good performance in terms of efficiency on handling large data sets in our experiments. Further, in analyzing a yeast data set, we demonstrate issues that arise when analyzing real data sets. While a probabilistic approach was recently introduced for this problem, and while parsimonious reconciliations have accuracy issues under certain settings, our parsimony framework provides a much more computationally efficient technique for this type of analysis. Our framework now allows for genome-wide scans for hybridization, while also accounting for ILS.

Key words: phylogenetic networks; hybridization; incomplete lineage sorting; coalescent; multi-labeled trees.

Hybridization is believed to be an important evolutionary mechanism for several

groups of eukaryotic organisms (Arnold, 1997; Barton, 2001; Mallet, 2005, 2007; Rieseberg, 1997). Evolutionary histories of species and genomes that involve hybridization are best modeled by *phylogenetic networks*, which account for both vertical and non-vertical evolutionary events (Nakhleh, 2010). Additionally, trees that trace the evolution of different segments of the genome, also known as gene trees, grow within the branches of a phylogenetic network (Maddison, 1997). This intertwined relationship between phylogenetic networks and the trees they contain naturally gave rise to a phylogeny-based approach to inferring phylogenetic networks from gene trees. In this approach, gene trees are compared, typically using a metric such as the *subtree prune and redraft* (SPR) distance, and the differences are taken as proxies for the amount and location of hybridization events (Nakhleh, 2010).

However, in addition to hybridization, the incongruence among gene trees may be partly caused by incomplete lineage sorting (ILS), or deep coalescence events (Maddison, 1997). Ignoring the presence of incomplete lineage sorting could result in an over- or under-estimation of the amount of hybridization events and/or wrong inference of the location of these events . Recent studies have documented large extents of incomplete lineage sorting in groups of organisms across the Tree of Life (Syring et al., 2005; Pollard et al., 2006; Than et al., 2008b; Kuo et al., 2008; Cranston et al., 2009; White et al., 2009; Hobolth et al., 2011; Takuno et al., 2012). A wide array of methods have been developed for species-tree inference from gene-tree topologies when all incongruence is assumed to be due to incomplete lineage sorting (see (Degnan and Rosenberg, 2009; Liu et al., 2009; Rannala and Yang, 2008) for recent surveys of such methods).

Relevant to this study are methods for inference under hybridization alone and under ILS alone that follow a parsimony approach: inferring the phylogenetic network with minimum number of reticulations in the former case, and inferring the phylogenetic tree

that minimizes the amount of ILS in the latter case. This approach in both cases was proposed by Maddison (1997) and much progress has been made on developing methods for parsimonious reconciliations ever since, both in the case of hybridization (Bordewich and Semple, 2005; Nakhleh et al., 2005; MacLeod et al., 2005; Beiko and Hamilton, 2006) and ILS (Maddison and Knowles, 2006; Than and Nakhleh, 2009, 2010; Yu et al., 2011b,c). Nonetheless, the first class of methods does not account for ILS, and the latter does not account for hybridization. Accounting for both kinds of events is a very challenging task (Mallet, 2005). Several attempts have been made in the last five years to handle both reticulation and incomplete lineage sorting. Than *et al.* introduced a stochastic framework for computing the probability of a gene tree given a species tree under the coalescent, and in the presence of a single horizontal gene transfer event (Than et al., 2007). Meng and Kubatko introduced methods for estimating the contribution of hybridization using a model that allows for both hybridization and incomplete lineage sorting (Meng and Kubatko, 2009). Kubatko further proposed using model selection with standard information criteria to identify hybridization in the presence of incomplete lineage sorting (Kubatko, 2009). Joly *et al.* introduced a statistical approach for the same task based on genetic distances between sequences (Joly et al., 2009). Yu *et al.* proposed extending the MDC (Minimize Deep Coalescences) criterion of (Maddison, 1997) to detect hybridization despite incomplete lineage sorting (Yu et al., 2011a). However, these methods all focused on very limited cases: fewer than 5 taxa, one or two hybridization events, and a single allele samples per species.

It is important to note that another ubiquitous cause of gene tree incongruence is *gene duplication and loss*. Recent efforts have emerged for combining gene duplication/loss with ILS (Rasmussen and Kellis, 2012) and for combining gene duplication/loss with horizontal gene transfer (Bansal et al., 2012), but incorporating duplication/loss with ILS and hybridization is beyond the scope of this work.

Most recently, Yu *et al.* proposed a method for computing the probability of gene-tree topologies given a phylogenetic network that is applicable to arbitrary numbers of taxa, arbitrary configurations of hybridization events, and any number of alleles sampled per species (Yu et al., 2012). While this general framework allows for inference of hybridization in the presence of ILS, it currently suffers from two issues. First, to turn the work of Yu *et al.* into an inference method, there is a need to develop methods for searching the phylogenetic network space and optimizing branch lengths and inheritance probabilities. Second, the method is computationally very expensive; developing new algorithmic techniques to achieve scalability is imperative.

In this paper we present a parsimony framework for inferring hybridization in the presence of ILS that extends Maddison's proposal (Maddison, 1997) to phylogenetic networks, and extend in novel ways the work of (Yu et al., 2011a) to general networks. The computational contribution of this work is two-fold: A parsimony criterion for reconciling a gene tree within the branches of a phylogenetic network so as to account for both hybridization and ILS, and a phylogenetic network search heuristic to enable inference of evolutionary histories from sets of gene-tree topologies. The framework only assumes knowledge of gene-tree topologies, which can be inferred by the method of choice of the practitioner, and infers a phylogenetic network with inheritance probabilities that correspond to proportions of genes involved in each of the hybridization events inferred. Our framework is general enough that it allows for multiple hybridization (in any configuration), multiple alleles sampled per species, arbitrary divergence patterns following hybridizations, and methodologically no bounds on the numbers of leaves in the gene trees.

We demonstrate the performance of our framework in terms of estimating the hybridization events and inheritance probabilities on simulated data under different evolutionary settings. For most cases, the framework exhibits very good performance from a small number of loci. Further, we reanalyze a yeast data set and show the performance of

the framework on biological data. We highlight two important issues: how to deal with uncertainty in the input gene trees, and the model selection problem that naturally arises when inferring phylogenetic networks. The speed of this parsimony framework makes it a good candidate for unrestricted analyses of multi-locus data sets, where hybridization is suspected, at least in order to obtain a first approximation to the true evolutionary history. While parsimonious reconciliation of species/gene trees and inference under such a criterion is known to have consistency issues under certain settings (Than and Rosenberg, 2011), parsimony remains a powerful approach in this domain, given its speed and good accuracy in many cases. We believe that our framework here can help in identifying good evolutionary hypotheses, which can be further analyzed with more detailed approaches such as the one of (Yu et al., 2012).

We have implemented our method and made it available in open-source form in the software package PhyloNet (Than et al., 2008a). The software package, as well as supporting documentation and a tutorial on its use, can be accesses at: http://bioinfo.cs.rice.edu/PhyloNet.

Methods

Here we describe our parsimony criterion for reconciling gene trees within the branches of phylogenetic networks, and our heuristic search for inferring phylogenetic networks and inheritance probabilities under this criterion.

Phylogenetic Networks and Gene Trees

The coalescent model (Kingman, 1982) views the evolution of multiple alleles of a locus backward in time. The multispecies coalescent generalizes the model to a phylogenetic tree that captures the evolution of multiple populations (Degnan and Rosenberg, 2009). Under

this model, a gene tree may disagree with a species tree due to incomplete lineage sorting (Fig. 1A). Here, each gene tree models the evolution of a set of alleles at a single locus in multiple species, and all incongruence is assumed to be due to incomplete lineage sorting.

When hybridization between two populations occurs, the evolutionary history of the species takes the form of a *phylogenetic network*, rather than a tree, so as to capture the contributions of genetic material from two parents (Fig. 1B). A *phylogenetic network* is a rooted, directed, acyclic graph whose leaves are labeled uniquely by a set of species. A phylogenetic network contains a unique node of in-degree 0 and out-degree 2 (the root), a set of nodes of in-degree 1 and out-degree 0 (the leaves), a set of nodes of in-degree 1 and out-degree 2 (the tree nodes), and a set of nodes of in-degree 2 and out-degree 1 (the reticulation nodes). Associated with every pair of reticulation edges (e_1,e_2) that are incident into a reticulation node are two real numbers γ_{e_1} and γ_{e_2}, respectively, such that $\gamma_{e_1} + \gamma_{e_2} = 1$. These parameters are interpreted as the *inheritance probabilities*: Given a lineage x at a reticulation node, it is inherited from one parent with probability γ_{e_1} and from other parent with probability γ_{e_2} (see Section 1 in Supplementary Material for more details; doi:10.5061/dryad.sr534). If $\gamma_{e_1} = 0$ for a reticulation edge e_1 incident into node x, this indicates that e_1 is redundant and that no hybridization involves x.

The evolution of a gene within the branches of a phylogenetic network can be viewed backward in time, such that whenever a reticulation node is encountered, the gene traces one of the two parents with a certain probability (the *inheritance probability*). Fig. 1 shows examples of a phylogenetic network on three species A, B, and C, and a gene tree with one allele sampled from A, two alleles sampled from B, and one allele sampled from C. An inheritance probability that is estimated at a value different from 0 and 1 indicates hybridization at the reticulation node, whereas a value of 0 or 1 imply that the reticulation node is redundant and can be replaced by a tree node attached to one of the two parents only.

Coalescent Histories and the MDC Criterion

We denote by $V(g)$ and $E(g)$ the sets of nodes and edges, respectively, of graph g, and by $C_t(v)$ the subtree of tree t that is rooted at node v. For a phylogenetic network N, we denote by $C_N(v)$ the subgraph of N that is induced by all the nodes reachable (or, "under") from v.

Given gene tree gt and species phylogeny ST (tree or network), a *coalescent history* is a function $f : V(gt) \rightarrow V(ST)$ such that the following conditions hold: (1) if w is a leaf in gt that is labeled by an allele from species x, then $f(w)$ is the leaf in ST labeled with x; and, (2) if w is a node in $C_{gt}(v)$, then $f(w)$ is a node in $C_{ST}(f(v))$. Fig. 1A shows a coalescent history of the gene tree in Fig. 1C within the branches of a species tree, whereas Fig. 1B shows a coalescent history of the same gene tree yet within the branches of a species network.

Given a gene tree gt and a species phylogeny ST, and given a function f defining a coalescent history of gt within ST, the *number of lineages* in each branch in ST can be computed by inspection. For example, in Fig. 1A, the number of lineages in the branch leading directly to taxon B is 2, whereas the number of lineages in the branch leading directly to C is 1. Given a coalescent history of a gene tree within the branches of a species tree, the number of extra lineages on a branch of the species tree is the number of lineages "exiting" the branch minus one. For example, the number of extra lineages on the branch incident with species B in Fig. 1A is 1, since two lineages exit the branch. In fact, given the gene tree in Fig. 1C and species tree in Fig. 1A, the reconciliation given in panel A is the one with the smallest number of extra lineages (for that fixed species tree). Given a set of coalescent histories for a set of gene trees, the total number of extra lineages is obtained by summing the number of extra lineages over all gene trees. More formally, let us denote by $XL(ST, gt)$ the number of extra lineages within an optimal coalescent history of gt

within ST. For a set $\mathcal{G}$ of gene trees, we have

$$XL(ST, \mathcal{G}) = \sum_{g \in \mathcal{G}} XL(ST, g). \quad (1)$$

Under parsimony, a reconciliation of the set of gene trees within the branches of a species tree that minimizes the total number of extra lineages over all gene trees provides an optimal evolutionary history of the gene genealogies for the given species tree; this is the *minimizing deep coalescences* (MDC) criterion proposed in (Maddison, 1997). Given a collection $\mathcal{G}$ of gene trees, the MDC (Minimizing Deep Coalescences) criterion (Maddison, 1997) seeks the species tree ST^* where

$$ST^* = \text{argmin}_{ST} XL(ST, \mathcal{G}).$$

For the inference problem, a species tree is sought so as to minimize deep coalescences over all possible (species) tree candidates. Efficient algorithms for solving this inference problem were recently introduced (Than and Nakhleh, 2009, 2010; Yu et al., 2011b,c).

ILS and Hybridization: MDC on Phylogenetic Networks

While Maddison defined this criterion for species trees, we extend it naturally to phylogenetic networks, given that we defined the concept of coalescent histories on phylogenetic networks above. Than and Nakhleh (2009) defined a mapping between a species tree and a gene tree that yields the optimal coalescent history that results in the minimum number of extra lineages . The key principle is to let the gene lineages coalesce as "low" as possible as long as they are consistent with the topologies of the gene and species trees. However, a similar idea for obtaining the optimal coalescent history of a gene tree within the branches of a phylogenetic network does not work; we illustrate this issue in

Section 2 in Supplementary Material. Further, Than and Nakhleh (2009) devised exact algorithms for inferring species trees from collections of rooted, binary gene trees under the MDC criterion, which were later extended to handle cases of unrooted or non-binary gene trees and with arbitrary numbers of alleles sampled per species (Than and Nakhleh, 2009; Yu et al., 2011b,c). However, none of these algorithms apply directly to the case where the species phylogeny is a network with at least one reticulation node.

We recently introduced an approach for reconciling a gene tree within the branches of a species network based on the concept of a *multi-labeled tree*, or MUL-tree (Yu et al., 2012). A phylogenetic network can be converted to a MUL-tree by proceeding in a bottom-up fashion (leaves to root), replicating the subtree at a reticulation node every time such a node is encountered. Upon termination of this process (when the root is reached), the resulting structure is a rooted tree whose leaves are not necessarily uniquely labeled. Each of the four panels in Fig. 2 shows the single MUL-tree that corresponds to the phylogenetic network in Fig. 1B. Once the MUL-tree is obtained, the evolution of a gene tree is modeled by mapping the alleles it contains to the respective species from which they were sampled. Fig. 2 shows the four possible allele mappings for the gene tree in Fig. 1C. Given an allele mapping, the multispecies coalescent then proceeds in the standard manner within the branches of the MUL-tree. Every coalescent history of a set of alleles on a phylogenetic network corresponds to a coalescent history of an allele mapping on the corresponding MUL-tree. Consequently, the optimal number of extra lineages arising from reconciling a gene tree within the branches of a species network can be computed using a slightly modified application of the MDC criterion on the MUL-tree and the set of allele mappings (Yu et al., 2012). Our method is illustrated in Fig. 3 and its full details are given in Section 3 in the Supplementary Material.

Given a collection $\mathcal{G}$ of gene trees, once the optimal coalescent histories for all of them are computed within the branches of a phylogenetic network N (using the MUL-tree

approach), the inheritance probabilities associated with the reticulation nodes are estimated as follows. Let x be a reticulation node in N. Given the optimal coalescent histories computed, let l_x be the number of lineages that trace the left parent in all the coalescent histories, and let r_x be the number of lineages that trace the right parent in all the coalescent histories. Then, the probability associated with the left reticulation edge incident with x is $l_x/(l_x + r_x)$ and the probability associated with the right reticulation edge incident with x is $r_x/(l_x + r_x)$ (see Section 4 in Supplementary Material for details of special cases).

Searching the Network Space

The space of phylogenetic networks is very large, and it is infeasible to enumerate all networks in order to identify the optimal one under the MDC score. Instead, we employ a local search heuristic that searches the space of phylogenetic networks, while scoring them based on Eq. (1). We denote by $\Omega(n, k)$ the space of phylogenetic networks that contain n taxa and k reticulation nodes. Suppose an optimal phylogenetic network with at most m reticulation nodes is sought. Our search strategy first searches the space $\Omega(n, 0)$ until some (potentially local) optimum is reached. The search then proceeds to $\Omega(n, 1)$, searches in that space until an optimum is reached, and then jumps to $\Omega(n, 2)$. This strategy continues until either an optimal network is reached in $\Omega(n, m)$, or the locally optimal score in $\Omega(n, k + 1)$ is not better than that in $\Omega(n, k)$ for some $k < m$.

The optimality scoring is done using the MUL-tree technique discussed above, and we now describe the topological operations that we employ to search the phylogenetic network space. For every phylogenetic network N, we define two disjoint neighborhoods: $\Delta(N)$, which contains networks with the same number of reticulation nodes as that in N, and $\Delta_{+1}(N)$, which contains networks with one more reticulation node than that in N. Given a phylogenetic network N, a neighbor $N' \in \Delta(N)$ is obtained by either relocating

the source of one edge in N or relocating the destination of one reticulation edge in N. Relocating the source of one edge in N follows three steps:

1. Choose two distinct edges (u_1, v_1) and (u_2, v_2) in N such that u_1 is neither a reticulation node nor a predecessor of v_2.

2. Delete node u_1 and the four edges (u_1, v_1), (u_2, v_2), (w, u_1) and (u_1, z), where w is the parent node of u_1 and z is a child node of u_1 other than v_1.

3. Add a new node x and four new edges (u_2, x), (x, v_2), (x, v_1) and (w, z) to the network.

Relocating the destination of one reticulation edge in N follows three steps:

1. Choose two distinct edges (u_1, v_1) and (u_2, v_2) in N such that v_1 is a reticulation node and v_2 is not a predecessor of u_1.

2. Delete node v_1 and the four edges (u_1, v_1), (u_2, v_2), (w, v_1) and (v_1, z), where w is a parent node of v_1 other than u_1 and z is the child node of v_1.

3. Add new node x and four new edges (u_2, x), (x, v_2), (u_1, x) and (w, z) to the network.

Given a phylogenetic network N, a neighbor $N' \in \Delta_{+1}(N)$ is obtained from N by adding a single edge to form a new reticulation node using the following three steps:

1. Choose two distinct edges (u_1, v_1) and (u_2, v_2) in N such that v_2 is not a predecessor of u_1.

2. Delete both (u_1, v_1) and (u_2, v_2).

3. Add two new nodes x_1 and x_2 and five new edges (u_1, x_1), (x_1, v_1), (u_2, x_2), (x_2, v_2) and (x_1, x_2).

Given a collection $\mathcal{G}$ of gene trees, we search in $\Omega(n,k)$ as follows. Assume the current optimal network in the search is $N \in \Omega(n,k)$ and we search for the next optimal network in $\Omega(n,k)$. We compute $\min_{N' \in \Delta(N)} XL(N', \mathcal{G})$ and compare this value to $XL(N, \mathcal{G})$. If the latter is larger, we replace the current network N by the new optimal one and continue the search in $\Omega(n,k)$ from the new network; otherwise, we stop the search in $\Omega(n,k)$ since the local optimum has been reached. If the search has stopped and k has reached a pre-specified upper bound of the number of reticulation nodes, the entire search terminates and the current network is returned as the inferred optimal network. If the pre-specified upper bound is not reached, the search moves up to $\Omega(n,k+1)$ by computing $\min_{N' \in \Delta_{+1}(N)} XL(N', \mathcal{G})$ and compare this value to $XL(N, \mathcal{G})$. If the latter value is larger, we replace the current network N by the new optimal one and continue the search in $\Omega(n,k+1)$ from the new network N using $\Delta(N)$; otherwise, the search terminates and N is returned as the optimal phylogenetic network inferred. It is important to note that since the optimal network in $\Omega(n,0)$ is the optimal species tree under the MDC criterion, the globally optimal network in this sub-space can be found efficiently (without search) using the method of Than and Nakhleh (2009).

Handling Gene Tree Uncertainty

When analyzing biological data sets, gene tree topologies are estimated from sequence data. Consequently, these gene tree estimates may have uncertainty associated with them. We handle this uncertainty in two different ways. First, consider a case where for each gene, a non-binary tree is obtained, such as in an analysis involving bootstrapping followed by contraction of all branches with low support or in an analysis that considers the strict consensus of all optimal trees under a maximum parsimony analysis. In this case, for each

gene we have a tree g that is not necessarily binary, and we replace Eq. (1) by

$$XL(ST, \mathcal{G}) = \sum_{g \in \mathcal{G}} \min_{g' \in b(g)} XL(ST, g') \tag{2}$$

where $b(g)$ is the set of all binary refinements of g. Of course, if g contains nodes of very high degrees, this approach is computationally infeasible if done in a brute-force fashion (explicitly considering all possible refinements). However, using our MUL-tree conversion technique, the efficient algorithms for (Yu et al., 2011b,c) apply directly and achieve this computation in polynomial time (in the size of the MUL-tree), as opposed to the exponential time (in the size of the MUL-tree) of the brute-force approach.

The second way of dealing with gene tree uncertainty is by incorporating the posterior probabilities computed by a Bayesian inference of the gene tree topologies. For each locus i, let $g_1^i, \ldots, g_q^i$ be the set of gene trees along with their posterior probabilities $p_1^i, \ldots, p_q^i$. For a gene tree topology g, let p_g be the sum of posterior probabilities associated with all gene trees that have the same topology as g over all loci. Then, we replace Eq. (1) by

$$XL(ST, \mathcal{G}) = \sum_{g \in \mathcal{G}} [XL(ST, g) \times p_g] \tag{3}$$

where $\mathcal{G}$ the set of all distinct gene tree topologies computed over all loci.

Results

Evaluating Inference on Simulated Data

To study the performance of our criterion and method in terms of the phylogenetic network they infer and the inheritance probabilities they estimate, we first used simulated data. We considered four phylogenetic networks (Fig. 4) depicting evolutionary scenarios that

present different challenges. The phylogenetic network in Scenario **I** includes speciation after hybridization. Scenario **II** presents two independent hybridization events involving terminal taxa (leaves). Scenario **III** includes a hybrid species that further speciates, and then the two sister taxa hybridize again. Scenario **IV** includes two hybridization events the more recent of which involves a descendant and a descendant of a parent of the earlier hybrid. These different phylogenetic networks allow us to examine how combinations of speciation and hybridization affect the detectability of hybridization in particular, and the inference of phylogenetic networks in general. Further, we varied the inheritance probabilities associated with the hybridization events in the phylogenetic networks. For Scenario **I**, we considered $\alpha \in \{0.0, 0.3, 0.5\}$, and for Scenario **II** and **III** we considered $(\alpha, \beta) \in \{(0.0, 0.5), (0.3, 0.3), (0.5, 0.5)\}$. Since the hybridization events in Scenario **IV** are overlapping, we considered $(\alpha, \beta) \in \{(0.0, 0.5), (0.3, 0.3), (0.5, 0.0), (0.5, 0.5), (0.5, 1.0)\}$ in this case. The rationale for selecting the three values 0.0, 0.3, and 0.5 is that they represent no hybridization, "skewed" hybridization (different genetic contributions of the two parents to the hybrid), and perfect hybridization (equal genetic contributions of the two parents to the hybrid). Finally, to vary the extent of deep coalescence within each of the four evolutionary histories, we considered two settings for the branch lengths $t_1, \ldots, t_4$ (as measured in coalescent units): setting 1, in which $t1 = t2 = t3 = t4 = 1.0$, and setting 2, in which $t1 = t2 = t3 = t4 = 2.0$. As the extent of ILS increases as branches become shorter, we expect setting 1 to provide more challenging data for the method.

Using each combination of phylogenetic network, inheritance probabilities, and branch length setting, we used the `ms` program (Hudson, 2002) to generate 10, 25, 50, 100, 500, 1000 and 2000 gene trees within the branches of the phylogenetic networks. To obtain statistically significant results, we generated 100 data sets per parameter setting and evaluated the performance as averaged over these 100 data sets, for each point in the parameter space. In these experiments, a single allele per species per gene was sampled.

Using the input sets of gene tree topologies, we inferred phylogenetic networks along with inheritance probabilities. In this experiments, we started our search from the optimal species tree under MDC by the exact method of Than and Nakhleh (2009). In this section, we assume knowledge of the true number of hybridization events and made inference with these (known) numbers of hybridization events. More specifically, for data sets corresponding to Scenario **I**, we inferred phylogenetic networks with single hybridization events, and for the other three scenarios, we inferred phylogenetic networks with two hybridization events. We discuss later the issues arising when we do not control for the number of hybridization events. We compared each inferred phylogenetic network against the (known) true phylogenetic network in terms of the topology and estimated inheritance probability. For comparing the topologies of two phylogenetic networks, we used the dissimilarity measure of (Nakhleh et al., 2004; Than et al., 2008a) which computes the symmetric difference between the two sets of taxa clusters induced by the two networks. Results of the application of our methods to gene trees under Scenarios **I**, **II**, and **III** are given in Fig. 5.

In terms of the accuracy of the inferred phylogenetic network topology, we observe that as the number of gene trees used increases, the error in the estimated network decreases. For all three evolutionary scenarios, using about 50 gene trees under time setting 2 for branch lengths results in phylogenetic network inferences with 0 error. However, the performance is different under time setting 1, which incorporates larger extents of incomplete lineage sorting. Here, we see that using about 50 gene trees results in correct network inference only under Scenario **II**, which is the least challenging for all scenarios considered. When we consider Scenario **I**, which adds to Scenario **II** the complexity of divergence after hybridization, we observe that the number of genes required to obtain accurate phylogenetic networks increases significantly (by an order of magnitude). For Scenario **III**, we observe that even with 2000 gene trees, the search

heuristic fails to identify the true phylogenetic network. It is important to note here that we must distinguish between the performance of the optimality criterion and that of the search heuristic employed for inference. In this case, our search heuristic begins with a species tree that minimizes the number of extra lineages (or, deep coalescences) over all possible tree candidates, given the set of gene tree. Using this tree, the search proceeds in a hill descent fashion, each time exploring all neighboring topologies of the current optimal network, and continuing with the best found. An artifact of this search heuristic is that if the true network cannot be obtained from the starting tree in any possible way, then this search heuristic would not converge to the true network. Of course, this problem could be ameliorated by random restarts of the search heuristic or by exhaustively starting from all possible trees. While the former is also not guaranteed to result in convergence to the true network, the latter is prohibitive but for data sets with very small numbers of taxa, given the exponentially large size of the tree space. Nevertheless, we have inspected the cases pertaining to Scenario **III** and verified that the reason behind the lack of convergence to 0-error networks is the criterion: The number of extra lineages in the optimal network that the heuristic infers is *smaller* than that in the true network. This is not surprising, since parsimonious reconciliation and inference is known to have consistency issues, even when ILS is the only event at play (Than and Nakhleh, 2009, 2010; Than and Rosenberg, 2011). Finally, we observe that the performance is better for inheritance probabilities that are closer to 0.5. This is due to the fact that under these settings the contributions of the two parents to the genetic makeup of a hybrid species is more balanced, providing more phylogenetic signal for the method to infer the correct evolutionary history.

In terms of estimating the inheritance probabilities, the results show that our search heuristic makes very good estimates, regardless of the evolutionary scenario and branch length setting. Even though branch length setting 2 yields slightly more accurate estimates, which is expected, it is important to note that the method produces very good

estimates even for the shorter branch lengths, where the extent of ILS is much larger. Further, it is worth emphasizing that these good estimates are obtained even with the smallest data sets (in terms of the number gene trees). This is a strength of the method.

More Loci or More Alleles?

Given the finite resources associated with any phylogenomic analysis, a natural question to ask is: In order to obtain more accurate inferences of phylogenetic networks and inheritance probabilities, should one sample more loci across the genomes or more alleles per locus? To explore this question, we used the above simulation procedure to generate gene trees under evolutionary Scenario **IV**, where 1, 2, 4 and 8 alleles per locus per species were sampled. The multi-allele gene trees were then used as input in the inference procedure. The results of this experiment are shown in Fig. 6.

Several observations are in order. First, in the case of this evolutionary scenario, the ability of the method to infer the correct topology of the phylogenetic network is not affected much by the branch length settings, unlike the performance on the other three scenarios. However, in this case, the method always overestimates the inheritance probability (by about 5% hybridization), more so in the case of time setting 1. Second, in this case, the estimates of the probability β of the lower (closer to the leaves) hybridization are more accurate than that of the estimates of α, which is unlike Scenarios **II** and **III**, where we did not observe any differences in the quality of the estimates of the two hybridization events. The reason for this is that in this scenario, some lineages, or alleles, from species D that trace different parents at the hybridization event undergo a further hybridization event, affecting the coalescence patterns towards the root. Regarding the benefit obtained by increasing the number of alleles, none are observed in terms of the inheritance probability, and some are observed in terms of the phylogenetic network accuracy under time setting 1. That is, if the branches are very short, sampling two alleles,

instead of one, improves the quality of the inferred network significantly. However, adding alleles beyond that does not seem to add more power, or signal, to the method. Under the other three scenarios, a single allele was already sufficient to provide highly accurate estimates. In summary, given the experimental settings we used here, there does not seem to be much benefit in sampling many alleles per species. Rather, sampling more loci per genome, particularly when the number of loci afforded is smaller than 100, provides more benefit. It is worth mentioning that the probabilistic method of (Yu et al., 2012) yields very accurate estimates of the inheritance probabilities under this evolutionary scenario, even when a single allele is sampled per species (see supplementary material of (Yu et al., 2012)).

Evaluating Inference on a Yeast Data Set

To study the performance of our framework on biological data, we reanalyzed the yeast data set of (Rokas et al., 2003). This data set consists of 106 loci, each present in exactly a single copy in each of seven Saccharomyces species, *S. cerevisiae (Scer)*, *S. paradoxus (Spar)*, *S. mikatae (Smik)*, *S. kudriavzevii (Skud)*, *S. bayanus (Sbay)*, *S. castellii (Scas)*, *S. kluyveri (Sklu)*, and the outgroup fungus *Candida albicans (Calb)*. We reconstructed gene trees from sequence data using maximum parsimony in PAUP* (Swofford, 1996) and Bayesian inference in MrBayes (Huelsenbeck and Ronquist, 2001). In each of 106 gene trees, the genes from the five species *Scer*, *Spar*, *Smik*, *Skud* and *Sbay* formed a monophyletic group. From a parsimony perspective, all coalescent events involving genes from these five species occur at or below their most recent common ancestor. Therefore, in our analysis, we only focused on the evolutionary history of these five species.

It is important to note that the gene trees used in the analysis here are not all binary. In the case where the gene trees were inferred by maximum parsimony, we used the strict consensus of all optimal trees found for each gene, which resulted in non-binary trees.

In the case of Bayesian inference, we used each gene tree with its posterior probability. See Methods for how we accounted for uncertainty in gene trees using these two approaches.

Using our method, we inferred the optimal species networks containing 0, 1 and 2 reticulation nodes. The resulting species networks inferred from gene trees reconstructed by maximum parsimony are shown in Fig. 7 along with inheritance probabilities and total number of extra lineages. The optimal species tree in Fig. 7A has been reported by several studies (Edwards et al., 2007; Rokas et al., 2003; Than and Nakhleh, 2009). The optimal species network containing one reticulation node in Fig. 7B has also been proposed as an alternative evolutionary history under the stochastic framework of (Bloomquist and Suchard, 2010), the parsimony framework of (Than and Nakhleh, 2009) and the likelihood framework of (Yu et al., 2012). It is worth mentioning that the inheritance probability inferred by our method is almost the same as that inferred by the probabilistic approach of (Yu et al., 2012). The optimal species network with two reticulation nodes in Fig. 7C was not reported in any of the aforementioned studies.

For gene trees reconstructed using MrBayes, the inferred species networks are shown in Fig. 8. The optimal species tree in Fig. 8A has been reported as a very close candidate (Edwards et al., 2007; Than and Nakhleh, 2009). The optimal species network containing one reticulation node in Fig. 8B has the same topology as the one inferred from gene trees reconstructed by maximum parsimony in Fig. 7B, but with a slightly higher inheritance probability.

The Model Selection Problem

A major confounding issue that arises when inferring phylogenetic network topologies is that of determining the correct number of reticulation events (Nakhleh, 2010). As we observed in the yeast data set analysis, adding a single reticulation node to the optimal species tree reduces the number of extra lineages by about 70%. Further, adding an

additional reticulation node to the optimal species network with a single reticulation node reduces the number of extra lineages by about a half. This is the classical model selection problem arising in the domain of phylogenetic networks: Increasing the complexity of the phylogenetic network topology by adding more reticulation nodes to it mostly improves the fit of the data. Simply minimizing the sum of the number of hybridization events and deep coalescence events does not solve the problem. Further, minimizing a weighted sum of these two numbers raises the questions of how to weight them and whether weights are data-dependent or not.

As we pointed out above, when analyzing the simulated data, we assumed knowledge of the true number of reticulation nodes. To understand the performance of the method when this assumption is removed, we inferred phylogenetic networks with up to 4 reticulation nodes from the data we generated, and explored the number of extra lineages in these inferred networks as a function of the number of reticulation nodes. The results for Scenario **III** are shown in Fig. 9; similar results were observed under the other scenarios.

As the figure shows, the number of extra lineages of the optimal species networks keeps decreasing as more reticulation nodes are added. Thus, using the minimization of the number of extra lineages as the optimality criterion, without penalizing complexity, may result in gross overestimation of the amount of reticulation in the data.

Performance on large data sets

We recently proposed another exact method for computing the number of extra lineages of a phylogenetic network and showed that it is much faster than the MUL-tree based one (Yu and Nakhleh, 2012). Since both methods are exact, substituting one for the other does not affect the inference method. Still, the method based on MUL-trees has its advantages in that it is applicable efficiently to unrooted and non-binary gene trees, as discussed above. Since we seek to evaluate the performance of parsimonious inference of phylogenetic

networks, we employ the method of Yu and Nakhleh (2012) for scoring phylogenetic networks.

We conducted experiments on simulated data sets that are much larger than the ones used above. We first generated 100 random species trees with 10, 20 and 40 taxa using PhyloGen Rambaut (2012) and set the total heights of those species trees to 8 N_e, 16 N_e and 32 N_e, respectively. From each species tree, we then generated random species networks with 1, 2, 3, 4 and 5 reticulation nodes respectively. When expanding a species network with k reticulation nodes to a species network with $k+1$ reticulation nodes, we randomly selected two existing edges in the species network and connected their midpoints from the higher one to the lower one and then the lower one becomes a new reticulation node. For every reticulation node, we assigned random values from 0 to 1 as its inheritance probability. Finally, we simulated 25, 50, 100 and 200 gene trees respectively within the branches of each species network using the `ms` program Hudson (2002).

Using the input sets of gene tree topologies, we inferred phylogenetic networks using the search procedure described above, and assuming knowledge of the true number of reticulation nodes. The running times of the method are shown in Fig. 10. It is not surprising that the running time increases with the increase in the numbers of taxa and reticulation nodes. But overall our method is able to finish the computations on all data sets in a reasonable amount of time. For the largest data set which has 40 taxa and 5 reticulation nodes, 75% of the computations finished within 24 hours. The outliers in the figure indicate that some data sets took much more time than others, especially for larger data sets. This occurs because the topology of the phylogenetic network and gene trees affect the running time, even when keeping the numbers of taxa and reticulation nodes fixed.

In addition to the running times, we also investigated the topological accuracy of the inferred phylogenetic network. Since for each run we know the true network N_m and

the inferred network N_i, the two networks can be compared using the normalized symmetric difference of the two; that is, by calculating the number of clusters that appear in one but not both of the networks, and dividing the number by twice the number of clusters in N_m. This measure was first introduced in (Nakhleh et al., 2003) and implemented in PhyloNet (Than et al., 2008a). The values of this measure for the pairs of phylogenetic networks we consider here range between 0, indicating identical networks, and 1, indicating the pair of networks disagree on every cluster. Results on the accuracy of the inferred phylogenetic networks are given in Fig. 11. For a fixed number of taxa, the error of network inference increases with the number of reticulation nodes. It is expected because the addition of reticulation nodes increases the complexity of the phylogenetic networks. On the other hand, for a fixed number of reticulation nodes, the error of network inference decreases as the number of taxa increases. This happens because for a network with larger number of taxa, the randomly added reticulation nodes may have a higher chance to be independent of each other, which actually makes the inference easier. Last but not least, as the number of gene trees sampled increases, the accuracy improves, albeit slightly. This may be due to issues with the search strategy, issues with the MDC criterion, or both.

Discussion

In this study, we extended the MDC criterion (Maddison, 1997; Than and Nakhleh, 2009) in order to define a parsimonious reconciliation of a gene tree topology within the branches of a phylogenetic network. By doing so, the resulting reconciliation accounts simultaneously for incomplete lineage sorting and hybridization. Further, we devised a local search heuristic for searching the phylogenetic network space to identify optimal ones under the new criterion. We applied our criterion and search heuristic to simulated data and a biological data set, and demonstrated the quality of inferences.

A central technique that we use in our study entails converting a phylogenetic network into its corresponding multi-labeled tree, or MUL-tree. This technique enables applying existing, tree-based criteria and methods to phylogenetic networks by employing them on the tree representation of the network. Indeed, in (Yu et al., 2012), we showed how to apply standard coalescent-based probabilistic computations to MUL-trees, and in this study we demonstrated how to extend parsimony-based tree reconciliations to phylogenetic networks by working indirectly on the MUL-tree representations. A further potential use of MUL-trees might be in facilitating phylogenetic network space search to enable efficient inference techniques.

When inferring phylogenetic networks, the classical model selection problem arises: More complex networks (that is, ones with more hybridization) may be found to fit the data better than less complex ones. Even though there is a quadratic bound on the maximum number of reticulation events in terms of the number of leaves in the phylogenetic network, methods that do not account for this issue would result in gross over-estimations of hybridization. This is one of the major problems with parsimonious reconciliations.

While various biological events can cause gene trees to disagree with each other as well as with the species phylogeny, a major confounding factor that must be accounted for when conducting analyses is uncertainty in gene trees. As gene trees are estimated from sequence data using computational methods, not all branching patterns in these trees can be inferred with certainty. Therefore, it is very important that criteria and methods account for this issue. We showed two ways of doing so, by allowing for non-binary gene trees and by considering posterior distributions.

Just as parsimony approaches can have consistency issues when inferring trees from sequences (Felsenstein, 1978), their counterparts for species tree inference from gene trees suffer from similar issues (Than and Rosenberg, 2011). We expect that this issue would arise also in the case of parsimonious inference of phylogenetic networks. Nevertheless, we

showed in this study that for many cases of hybridization and ILS, parsimony obtains the same results as a probabilistic framework, within a fraction of the time that the latter approach takes. We believe one of the best uses of parsimony would be to quickly obtain a good initial network that can be used to seed searches for phylogenetic networks under probabilistic approaches.

One major simplifying assumption that we made here is ignoring other discord factors, such as gene duplication and loss. We will explore ways of incorporating duplication and loss into our framework, potentially along the lines of combining the idea of MUL-tree with that of the *locus tree* (Rasmussen and Kellis, 2012). Further, while we specifically address hybridization in this study, the framework is applicable in theory to horizontal gene transfer (HGT) in general. However, when only a single gene or very few genes are transferred, a large extent of ILS might overwhelm the signal for HGT.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.sr534.

Funding

This work was supported in part by NSF grants DBI-1062463 and CCF-1302179, grant R01LM009494 from the National Library of Medicine, an Alfred P. Sloan Research Fellowship, and a Guggenheim Fellowship to L.N. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NSF, National Library of Medicine, the National Institutes of Health, the Alfred P. Sloan Foundation, or the John Simon Guggenheim Memorial Foundation.

References

Arnold, M. L. 1997. Natural Hybridization and Evolution. Oxford University Press, Oxford.

Bansal, M., E. Alm, and M. Kellis. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics 28:i283–i291.

Barton, N. 2001. The role of hybridization in evolution. Molecular Ecology 10:551–568.

Beiko, R. and N. Hamilton. 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evolutionary Biology 6.

Bloomquist, E. and M. Suchard. 2010. Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. Systematic Biology 59:27–41.

Bordewich, M. and C. Semple. 2005. On the computational complexity of the rooted subtree prune and regraft distance. Annals of Combinatorics Pages 1–15.

Cranston, K. A., B. Hurwitz, D. Ware, L. Stein, and R. A. Wing. 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58:489–500.

Degnan, J. and N. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution 24:332–340.

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

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

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

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

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

Joly, S., P. A. McLenachan, and P. J. Lockhart. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. Am. Nat. 174:E54–E70.

Kingman, J. F. C. 1982. The coalescent. Stochast. Proc. Appl. 13:235–248.

Kubatko, L. S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.

Kuo, C.-H., J. P. Wares, and J. C. Kissinger. 2008. The Apicomplexan whole-genome phylogeny: An analysis of incongurence among gene trees. Mol. Biol. Evol. 25:2689–2698.

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

MacLeod, D., R. Charlebois, F. Doolittle, and E. Bapteste. 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. BMC Evolutionary Biology 5.

Maddison, W. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maddison, W. P. and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.

Mallet, J. 2007. Hybrid speciation. Nature 446:279–283.

Meng, C. and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theor. Popul. Biol. 75:35–45.

Nakhleh, L. 2010. Evolutionary phylogenetic networks: models and issues. Pages 125–158 *in* The Problem Solving Handbook for Computational Biology and Bioinformatics (L. Heath and N. Ramakrishnan, eds.). Springer, New York.

Nakhleh, L., D. Ruths, and L. Wang. 2005. RIATA-HGT: A fast and accurate heuristic for reconstrucing horizontal gene transfer. Pages 84–93 *in* Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05) (L. Wang, ed.) lNCS #3595.

Nakhleh, L., J. Sun, T. Warnow, C. Linder, B. Moret, and A. Tholse. 2003. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. Pages 315–326 *in* Proceedings of the Eighth Pacific Symposium on Biocomputing vol. 8.

Nakhleh, L., T. Warnow, and C. Linder. 2004. Reconstructing reticulate evolution in species–theory and practice. Pages 337–346 *in* Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB04).

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

Rambaut, A. 2012. Phylogen v1.1. Http://treebioedacuk/software/phylogen/.

Rannala, B. and Z. Yang. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

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

Rieseberg, L. 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28:359–389.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Swofford, D. L. 1996. PAUP*: Phylogenetic analysis using parsimony (and other methods). Sinauer Associates, Underland, Massachusetts, Version 4.0.

Syring, J., A. Willyard, R. Cronn, and A. Liston. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. American Journal of Botany 92:2086–2100.

Takuno, S., T. Kado, R. P. Sugino, L. Nakhleh, and H. Innan. 2012. Population genomics in bacteria: A case study of staphylococcus aureus. Molecular Biology and Evolution 29:797–809.

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

Than, C. and L. Nakhleh. 2010. Inference of parsimonious species phylogenies from multi-locus data by minimizing deep coalescences. Pages 79–98 *in* Estimating Species Trees: Practical and Theoretical Aspects (L. Knowles and L. Kubatko, eds.). Wiley-VCH.

Than, C. and N. Rosenberg. 2011. Consistency properties of species tree inference by minimizing deep coalescences. Journal of Computational Biology 18:1–15.

Than, C., D. Ruths, H. Innan, and L. Nakhleh. 2007. Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. J. Comput. Biol. 14:517–535.

Than, C., D. Ruths, and L. Nakhleh. 2008a. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9:322.

Than, C., R. Sugino, H. Innan, and L. Nakhleh. 2008b. Efficient inference of bacterial strain trees from genome-scale multi-locus data. Bioinformatics 24:i123–i131 proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB '08).

White, M. A., C. Ane, C. N. Dewey, B. R. Larget, and B. A. Payseur. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet 5:e1000729.

Yu, Y., J. Degnan, and L. Nakhleh. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genetics 8:e1002660.

Yu, Y. and L. Nakhleh. 2012. Fast algorithms for reconciliation under hybridization and incomplete lineage sorting. arXiv 1212.1909.

Yu, Y., C. Than, J. Degnan, and L. Nakhleh. 2011a. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Systematic Biology 60:138–149.

Yu, Y., T. Warnow, and L. Nakhleh. 2011b. Algorithms for MDC-based multi-locus phylogeny inference. Pages 531–545 *in* Proceedings of the 15th Annual International Conference on Research in Computational Molecular Biology (RECOMB) vol. 6577 of *Lecture Notes in Bioinformatics* Springer.

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

Figure Captions

Figure 1: **Gene trees within species trees and species networks.** (A) Under the multispecies coalescent model, a gene tree may be incongruent with the species tree due to incomplete lineage sorting (ILS). (B) When hybridization occurs between two species (or, populations), the species phylogeny takes the shape of a network, and a gene tree "grows" within the branches of the network. The variable γ corresponds to the probability of a lineage in the hybrid population being inherited from the "left" parent ($1 - \gamma$ is the probability of inheritance from the "right" parent). (C) ILS and hybridization can give rise to the same gene tree shape (or, topology).

Figure 2: **Phylogenetic networks and MUL-trees.** The MUL-tree that corresponds to the phylogenetic network in Fig. 1B. The four panels show the four possible allele mappings of the gene tree in Fig. 1C. The allele mapping in (B) corresponds to the coalescent history in Fig. 1B. Values of 1.0 and 0.0 for γ in panels A and D, respectively, indicate no support for hybridization in these two cases. The MUL-tree in panel B has the lowest number of extra lineages, and hence corresponds to the optimal reconciliation of the gene tree within the branches of the phylogenetic network.

Figure 3: **A schematic illustration of our method for computing an optimal coalescent history of a gene tree within the branches of a phylogenetic network.** The phylogenetic network is converted to a MUL-tree, and the alleles sampled are mapped in every possible way to the leaves of the MUL-tree. For each allele mapping, the number of extra lineages is computed on the MUL-tree, and the mapping that yields the minimum overall number corresponds to an optimal coalescent history.

Figure 4: **Phylogenetic networks depicting different hybridization/divergence/extinction scenarios.** The α and β parameters denote the proportions (or, probabilities) of alleles that are inherited from the "left" parents of the reticulation nodes ($1-\alpha$ and $1-\beta$ denote the proportions of the alleles that are inherited from the "right" parents of the nodes).

Figure 5: **Accuracy of the inferred phylogenetic networks and inheritance probabilities.** The three columns from left to right correspond to Scenarios **I**, **II**, and **III** in Fig. 4, respectively. One allele per gene per species is sampled.

Figure 6: **The effect of the number of alleles.** Accuracy of the phylogenetic networks and inheritance probabilities estimated from gene trees simulated under Scenario **IV**, with true inheritance probabilities $\alpha = \beta = 0.3$, where the number of alleles sampled per species also varies. Top and bottom rows correspond to time settings 1 and 2, respectively.

Figure 7: **Analysis of the yeast data set, where gene trees are reconstructed using MP.** Optimal species phylogenies, along with inheritance probabilities, inferred from gene trees reconstructed by maximum parsimony for the yeast data set of (Rokas et al., 2003). (A) The optimal species tree (network with 0 reticulation nodes). (B) The optimal species network containing one reticulation node. (C) The optimal species network containing two reticulation nodes. For each species phylogeny, the total number of extra lineages (XL) is computed using Eq. (2) and reported.

Figure 8: **Analysis of the yeast data set, where gene trees are reconstructed using Bayesian inference.** Optimal species phylogenies, along with inheritance probabilities, inferred from gene trees reconstructed by MrBayes for the yeast data set of (Rokas et al., 2003). (A) The optimal species tree (network with 0 reticulation nodes). (B) The optimal species network containing one reticulation node. (C) The optimal species network containing two reticulation nodes. For each species phylogeny, the total number of extra lineages (XL) is computed using Eq. (3) and reported.

Figure 9: **Network complexity and the number of extra lineages.** The decrease in the number of extra lineages in the inferred phylogenetic network as a function of the increase in number of hybridization events inferred. The results were obtained from data pertaining to Scenario **III** under two different settings of the inheritance probabilities and two different settings of the branch lengths.

Figure 10: **Running time of phylogenetic network inference.** The three columns from left to right correspond to data sets with 10, 20 and 40 taxa, respectively. The six rows from bottom to top correspond to data sets with 0, 1, 2, 3, 4 and 5 reticulation nodes, respectively. In each sub-figure, the x-axis is the number of gene trees sampled and the y-axis is the running time in seconds.

Figure 11: **Accuracy of inferred phylogenetic networks.** The three columns from left to right correspond to data sets with 10, 20 and 40 taxa, respectively.

Figures

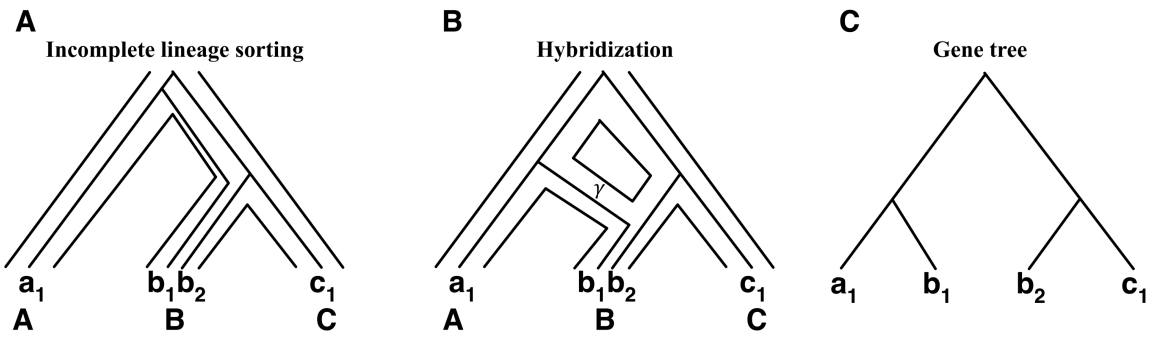

Figure 1

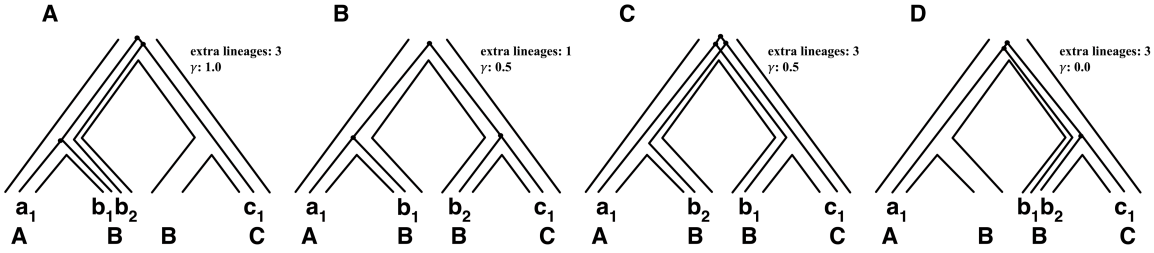

Figure 2

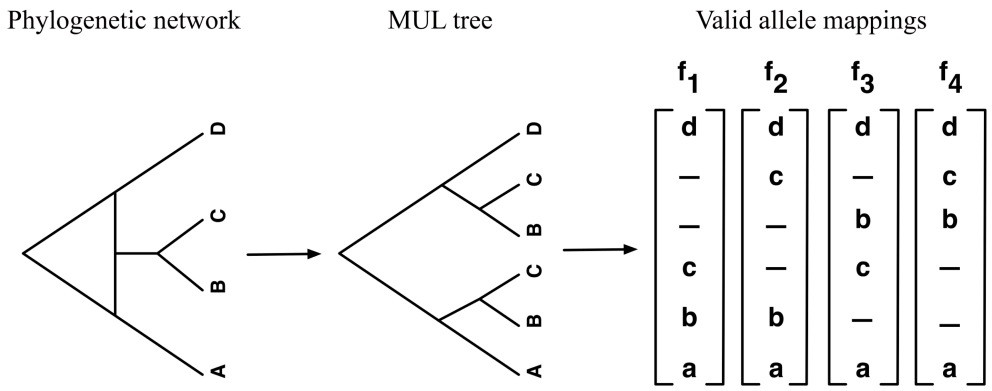

Figure 3

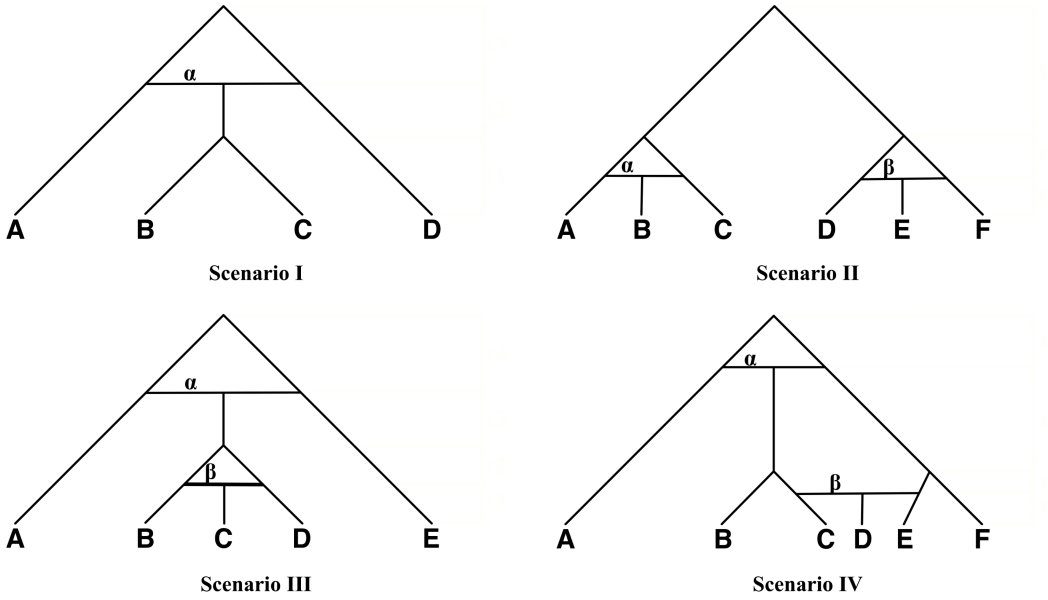

Figure 4

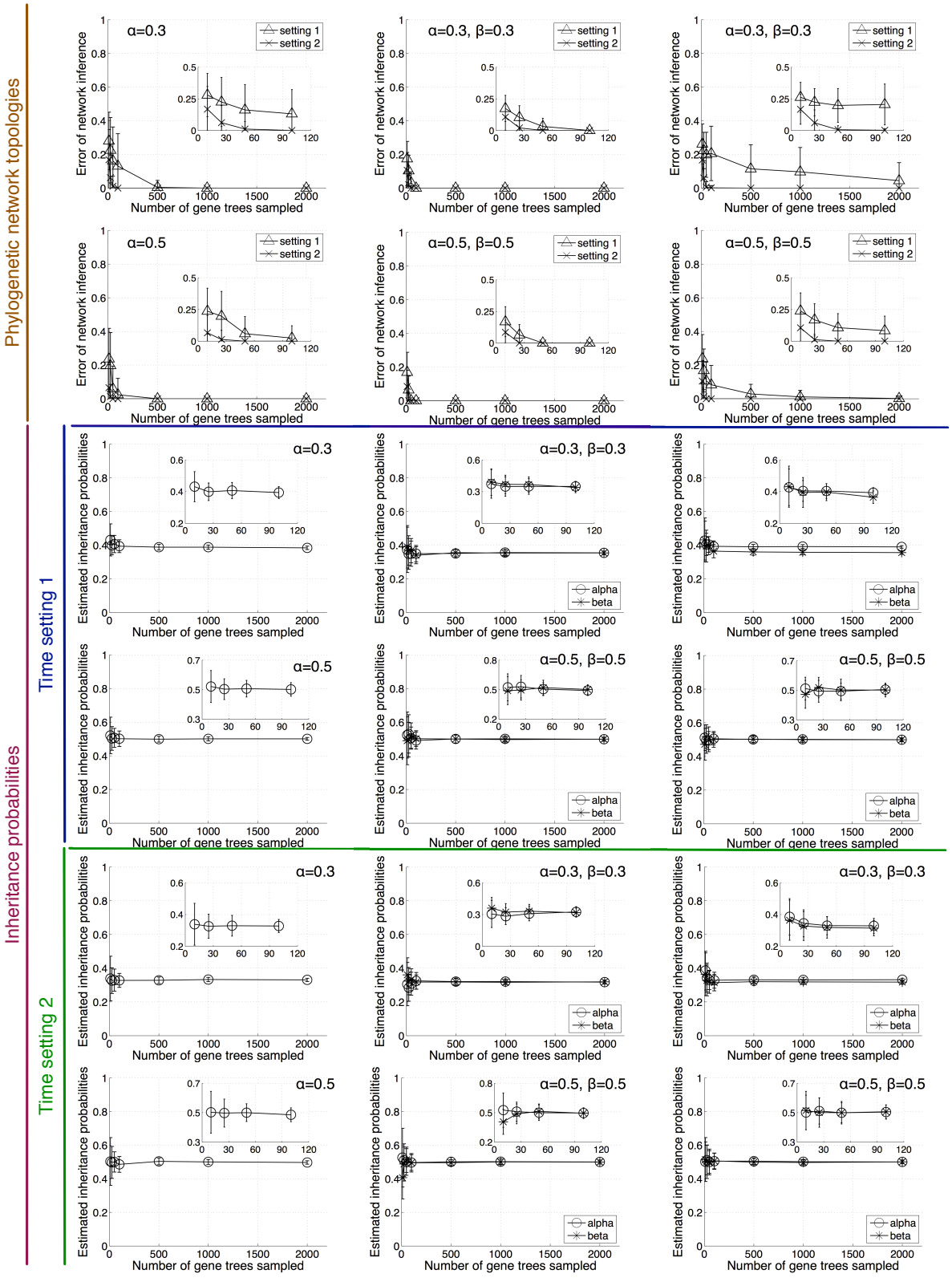

Figure 5

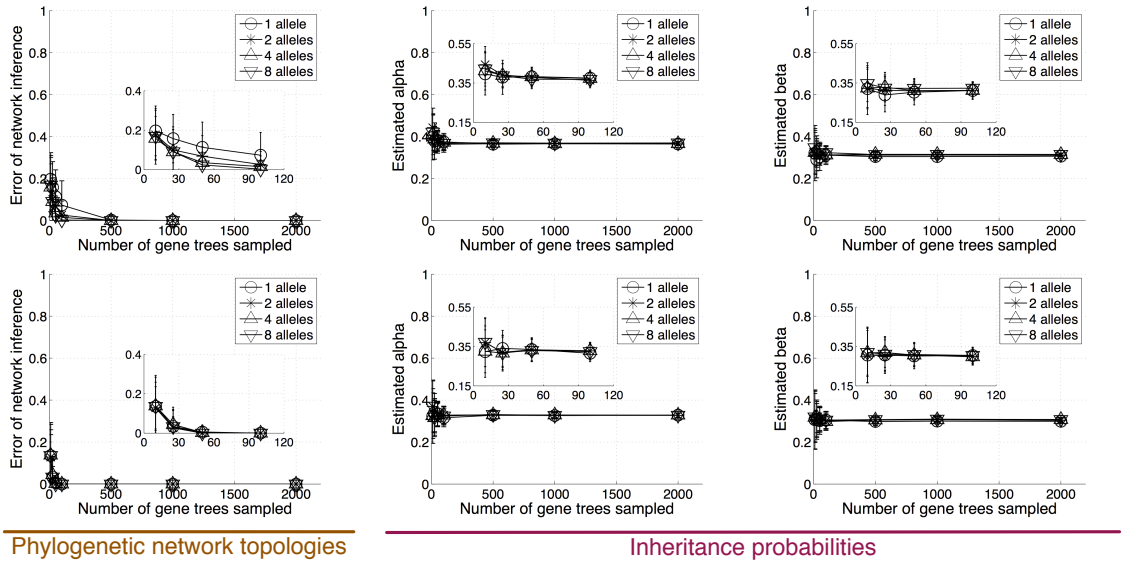

Figure 6

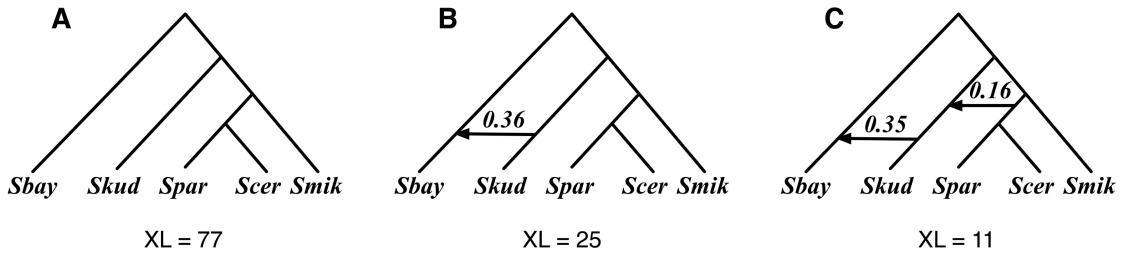

Figure 7

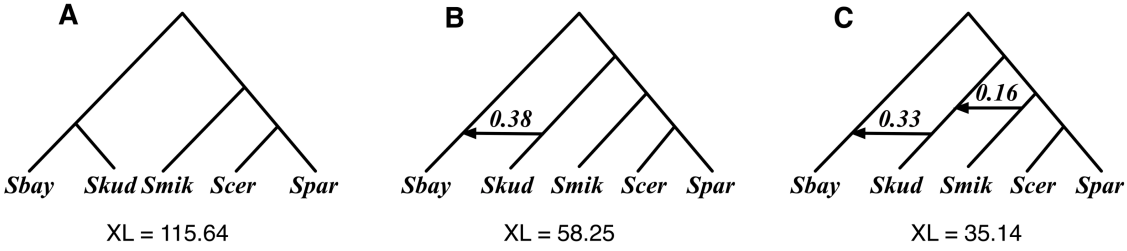

Figure 8

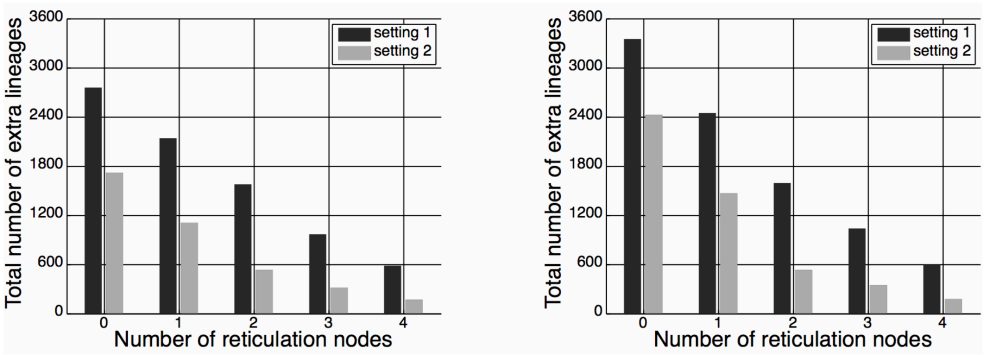

Figure 9

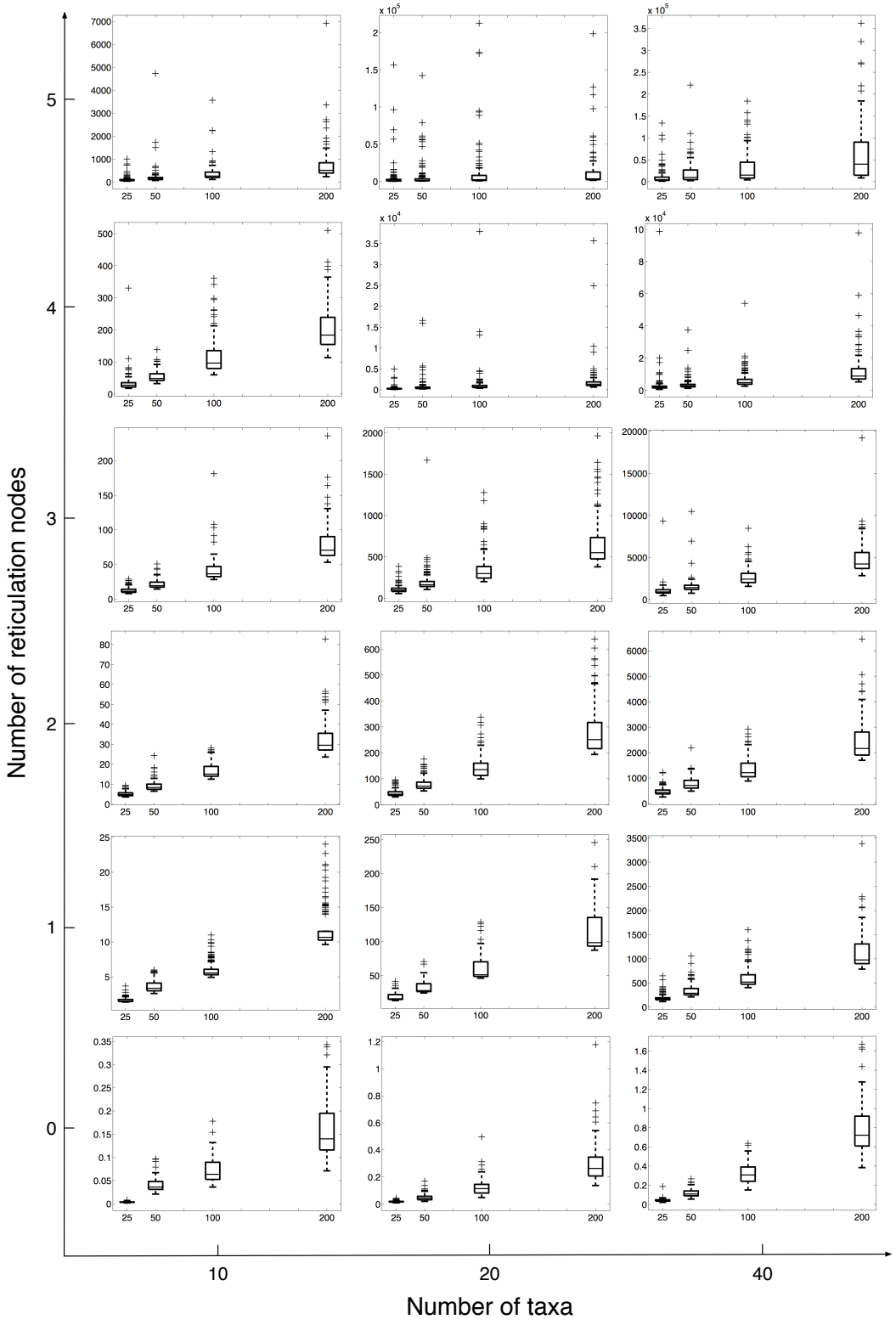

Figure 10

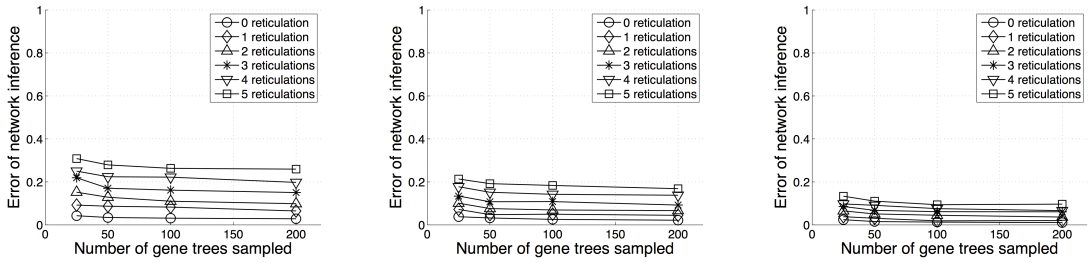

Figure 11