RESEARCH ARTICLE

Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent

Dingqiao Wen[1]*, Yun Yu[1], Luay Nakhleh[1,2]*

1 Computer Science, Rice University, Houston, Texas, United States of America, **2** BioSciences, Rice University, Houston, Texas, United States of America

* dw20@rice.edu (DW); nakhleh@rice.edu (LN)

Abstract

The multispecies coalescent (MSC) is a statistical framework that models how gene genealogies grow within the branches of a species tree. The field of computational phylogenetics has witnessed an explosion in the development of methods for species tree inference under MSC, owing mainly to the accumulating evidence of incomplete lineage sorting in phylogenomic analyses. However, the evolutionary history of a set of genomes, or species, could be reticulate due to the occurrence of evolutionary processes such as hybridization or horizontal gene transfer. We report on a novel method for Bayesian inference of genome and species phylogenies under the multispecies network coalescent (MSNC). This framework models gene evolution within the branches of a phylogenetic network, thus incorporating reticulate evolutionary processes, such as hybridization, in addition to incomplete lineage sorting. As phylogenetic networks with different numbers of reticulation events correspond to points of different dimensions in the space of models, we devise a reversible-jump Markov chain Monte Carlo (RJMCMC) technique for sampling the posterior distribution of phylogenetic networks under MSNC. We implemented the methods in the publicly available, open-source software package PhyloNet and studied their performance on simulated and biological data. The work extends the reach of Bayesian inference to phylogenetic networks and enables new evolutionary analyses that account for reticulation.

Author Summary

Trees have long formed in biology the basic structure with which to represent and understand evolutionary relationships. Mathematical models, computational methods, and software tools for inferring phylogenetic trees and studying their mathematical properties are currently the norm in biology. The availability of genomic data from closely related species, as well as from multiple individuals within species, have brought the two fields of phylogenetics and population genetics closer than ever. In particular, the last two decades have witnessed a great flourish in the development and implementation of phylogenetic methods based on the multispecies coalescent model to capture the intricate relationship between gene and genome evolution. However, when reticulation processes such as hybridization occur, the phylogenetic history is best represented by a network. In this

Citation: Wen D, Yu Y, Nakhleh L (2016) Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. PLoS Genet 12(5): e1006006. doi:10.1371/journal.pgen.1006006

Editor: Scott Edwards, Harvard University, UNITED STATES

Received: October 11, 2015

Accepted: April 4, 2016

Published: May 4, 2016

Data Availability Statement: The input files with the gene trees used in the data analyses have been deposited into Dryad under DOI: 10.5061/dryad.n2606 at http://dx.doi.org/10.5061/dryad.n2606. All other information is contained in the paper and its Supporting Information files.

Funding: This work was supported in part by National Science Foundation (http://www.nsf.gov) grants CCF-1302179 and DBI-1355998. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

work, we demonstrate how the multispecies coalescent model can be adapted to reticulate evolutionary histories and report on a Bayesian method for inference of such histories under this extended model. As networks subsume trees, the model and method provide a principled and unified statistical framework for inferring treelike and non-treelike evolutionary relationships.

Introduction

Species trees capture how species evolved and diverged from a common ancestor. These trees provide a framework for understanding how genes, genomes, and traits evolve [1, 2]. Consequently, accurate inference of species trees has been a major endeavor in evolutionary biology [3, 4]. With the availability of data from multiple genomic regions, and often whole genomes, modern inference techniques utilize all these data and employ the multispecies coalescent (MSC) model [5]. This model captures how gene (more generally, non-recombining genomic regions) genealogies grow within the branches of a species tree when extending the coalescent model [6] to multiple populations tied together by a phylogenetic tree (Fig 1**A**). MSC naturally models incomplete lineage sorting (ILS) and, when combined with models of sequence evolution, connects species trees with genomic sequence data and provides a statistical framework for species tree inference. Indeed, a wide array of methods have been devised for inferring species trees under the MSC model either directly from the sequence data [7–9] or from gene tree estimates [10, 11]; see [12–14] for recent reviews of species phylogeny inference methods.

It has long been acknowledged that the evolutionary histories of many groups of species, from across all domains of life, are reticulate. Horizontal gene transfer is ubiquitous in prokaryotic evolution [15, 16], and several bodies of work are pointing to much larger extent and role of hybridization in eukaryotic evolution than once thought [17–22]. Reticulate evolutionary histories are best modeled by *phylogenetic networks*. There are two categories of phylogenetic networks: data-display networks and evolutionary, or, explicit phylogenetic networks [23–25]. The former group is aimed at displaying pairwise relationships in the data that cannot be adequately captured by a single tree, yet not necessarily due to reticulation. The latter category provides an explicit phylogenetic model of evolutionary relationships that extends trees to allow

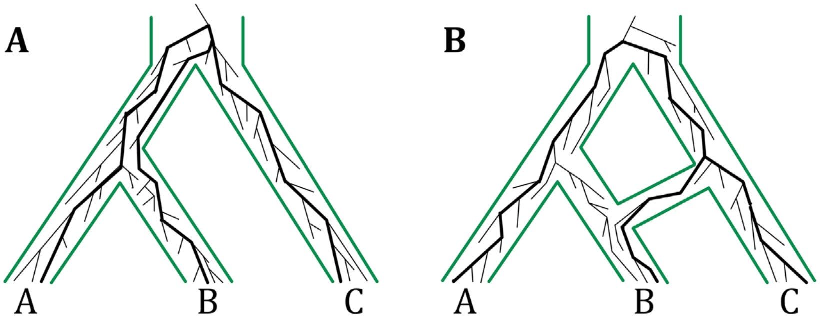


Fig 1. The multispecies coalescent on trees and networks. (**A**) The multispecies coalescent (MSC) links populations by a tree structure and allows for modeling gene genealogies within the branches of a species tree. The gene genealogy indicated by thick lines inside the species tree is incongruent with the species tree due to incomplete lineage sorting (ILS). (**B**) The multispecies network coalescent (MSNC) links populations by a network structure, thus allowing for reticulations events among populations. The gene genealogy indicated by thick lines inside the species network is involved in reticulation, e.g., hybridization. The gene genealogies in both panels have the same topologies, but have different probabilities under the MSC and MSNC models.

doi:10.1371/journal.pgen.1006006.g001

for reticulations. The work here concerns the inference of evolutionary phylogenetic networks, or phylogenetic networks as we shall refer to them hereafter.

A phylogenetic network is a rooted, directed, acyclic graph whose leaves are labeled uniquely by a set of taxa (see Methods for a formal definition). It extends a phylogenetic tree by allowing for nodes with two parents (called *reticulation nodes*) to capture reticulation. For example, in Fig 1**B**, the phylogenetic network captures a hybridization event between species B and C. Methods for inferring phylogenetic networks with the minimum number of reticulation nodes from a set of estimated gene trees were recently introduced [26–28]. These methods assume that incongruences among gene trees are solely due to reticulation and employ parsimony as the criterion for selecting the phylogenetic networks among all possible explanations. However, as was highlighted by several recent studies [29–33], ILS could very well be at play in data sets where reticulation is suspected. Therefore, it is important to devise a statistical framework that accounts *simultaneously* for ILS and reticulation and to develop models for inference of species evolutionary histories under this framework.

Nakhleh and colleagues [30, 34, 35] recently extended the MSC model to phylogenetic networks, a model that we now call the multispecies network coalescent, or MSNC (see Methods for a formal definition). Under this model, the growth of a gene genealogy is viewed backward in time (the time flows from the root toward the leaves) within the branches of a phylogenetic network (Fig 1**B**). When a reticulation node is encountered, the genealogy traces one of the two parental species with a certain probability that is dependent on the locus for that genealogy as well as the specific reticulation node encountered. A large divergence time between C and the MRCA of A and B or a small population size of the MRCA of A and B would be unlikely to give rise to the indicated gene genealogy. However, these same settings coupled with a scenario of hybridization between B and C could very well give rise to the same gene genealogy. Yu *et al.* [30] recently devised a local search heuristic for inferring phylogenetic networks under the MSNC model. The method’s good results notwithstanding, the analyses highlighted three major issues. First, knowledge about reticulation could not be readily incorporated into the likelihood model. Second, avoiding overfitting by extra reticulations needed to be handled in a principled way. Third, a point estimate of the maximum likelihood phylogenetic network was not adequate given the closeness in likelihood of other phylogenetic networks.

To address all three issues, we devise a Bayesian framework under the MSNC model and a Markov chain Monte Carlo (MCMC) sampler of the posterior distribution on phylogenetic networks. This framework allows for systematically incorporating knowledge about reticulations and penalizing for model complexity via appropriate prior distributions. Further, the MCMC technique allows for obtaining a sample of the posterior distribution of phylogenetic networks, rather than a point estimate. Phylogenetic networks on the same taxa yet with different numbers of reticulations correspond to different numbers of parameters. Thus, walking the space of phylogenetic networks is trans-dimensional, where the number of dimensions when a new sample is proposed could decrease (due to the removal of a reticulation), increase (due to the addition of a reticulation), or remain unchanged. To account of this issue, posterior sampling is done via reversible-jump MCMC, or RJMCMC [36].

While our Bayesian framework makes use of the likelihood functions that we had derived earlier [30, 34, 35, 37], our derivations for the RJMCMC here are inspired by two works. Some of the specifics of our RJMCMC implementation are inspired by the work of Lewis *et al.* [38], where RJMCMC was employed to walk the space of phylogenetic networks with and without polytomies. For the prior on the phylogenetic network topology, our derivation was inspired by the work of Bloomquist and Suchard [39]. However, our work differs significantly from these two works in that the work of [38] is focused on trees and does not handle networks, and the work of [39] is not based on the multispecies coalescent.

We have implemented our methods in the PhyloNet software package [40], which is publicly available in open source. We tested the accuracy of the method on several simulation data sets, where we varied the topology and branch lengths of the phylogenetic network, the amount of data used in the sampling, and the prior. Our results demonstrate a good performance of the method, including the desirable property that the prior has less of an effect as the amount of data increases. We then analyzed three biological data sets: The bread wheat data set of [29], the mosquito data set of [31], and the house mouse data set that we analyzed recently in [30]. A major difference between the house mouse data set and the other two is that the former consists of multiple individuals of the same species, whereas in the other two data sets each genome is obtained from a different species. This illustrates the applicability of our method to these two different scenarios. Nonetheless, our results demonstrate the challenges with analyzing such data, particularly in terms of detecting hybridization.

To the best of our knowledge, this is the first Bayesian approach to sampling the phylogenetic network posterior under the extended MSNC. Computationally, the major bottleneck stems from the likelihood calculations under the MSNC. The computational requirements notwithstanding, our results demonstrate that this is a very promising direction to pursue in terms of application to data analysis and development of new phylogenetic network inference methods.

Results

A Bayesian model of reticulate phylogenies

The data $S = \{S_1, \ldots, S_m\}$ consists of the sequence alignments of m loci that we assume to be independent and recombination-free (S_i is the sequence alignment that corresponds to locus i). Our model consists of Ψ, the phylogenetic network (topology and branch lengths), and Γ, the inheritance probabilities matrix (see Methods). The posterior of the model is then given by

$$p(\Psi,\Gamma|S) \propto p(S|\Psi,\Gamma)p(\Psi)p(\Gamma) = p(\Psi)p(\Gamma)\prod_{i=1}^{m}\int_{G} p(S_i|g)p(g|\Psi,\Gamma)dg \tag{1}$$

where the integration is taken over all possible gene trees, $p(S_i|g)$ is the probability of the sequence alignment S_i given a particular gene tree g [41], and $p(g|\Psi, \Gamma)$ is the density of the gene tree (topologies and branch lengths) given the model parameters [30] (see S1 Text for details of the density function).

If the gene tree estimates (or, a posterior distribution thereof) of the individual loci are also of interest, the formulation above can be modified to co-estimate gene trees, in addition, as follows:

$$p(\Psi,\Gamma,G|S) \propto p(\Psi)p(\Gamma)\prod_{i=1}^{m} p(S_i|g_i)p(g_i|\Psi,\Gamma) \tag{2}$$

where g_i is the estimated gene tree for locus i.

Indeed, this co-estimation approach is adopted by two popular species tree inference methods, *BEAST [8] and BEST [7], with the major difference being that in these methods Ψ is a tree and Γ is therefore redundant.

Inference based on Eq (1) is computationally infeasible due to the integration over all gene trees. Even Monte Carlo integration techniques would fail at estimating the integral, except for very small data sets. While sampling gene trees, as is done in *BEAST and BEST, has been shown to yield very good estimates of species and gene trees, these methods are computationally prohibitive for large data sets. Consequently, a wide array of methods for inferring species trees from gene tree estimates, rather than sequence alignments, have been introduced. In the case of networks, maximum likelihood inference of networks that uses gene tree estimates has been shown to provide good results as well [30]. If we assume that a set G of gene trees has been estimated for the m loci, then we get

$$p(\Psi,\Gamma|G) \propto p(G|\Psi,\Gamma)p(\Psi)p(\Gamma) = p(\Psi)p(\Gamma)\prod_{i=1}^{m} p(g_i|\Psi,\Gamma) \quad (3)$$

where g_i is the estimated gene tree for locus i (with or without branch lengths). While inference from sequences directly accounts naturally for gene tree uncertainty, one has to account for this uncertainty explicitly when using gene tree estimates. Here, we will adopt the same strategy as in [30], where for each locus i, a set of gene tree estimates are obtained (e.g., the set of gene trees in a bootstrap analysis or the set of gene trees obtained from sampling the posterior of trees for that locus). Furthermore, while the method is applicable to data that consist of gene tree with branch lengths, we focus here on gene tree topologies alone (see results and discussion below for more on this point). The mass function for gene tree topologies given a phylogenetic network was derived in [35, 37] and is given in S1 Text.

To fully specify the model given by Eq (1), we need two priors $p(\Psi)$ and $p(\Gamma)$. For the phylogenetic network, we define a prior that is similar to that defined on ancestral recombination graphs in [39]. Given a phylogenetic network Ψ, we denote by Ψ_{top}, Ψ_{λ}, and Ψ_{ret} the topology, branch lengths vector, and number of reticulation nodes, respectively, of Ψ. We have

$$p(\Psi|\nu,\delta,\eta) = p(\Psi_{ret}|\nu) \times p(\Psi_{\lambda}|\delta) \times p(\Psi_{top}|\Psi_{ret},\Psi_{\lambda}). \quad (4)$$

where $p(\Psi_{ret}|\nu) \sim \frac{1}{T_{n,m}}\text{Poisson}(\nu)$, where $T_{n,\,m}$ is the number of phylogenetic network topologies with n leaves and m reticulation nodes, $p(\Psi_{\lambda}|\delta) \sim \text{Exp}(\delta)$, and $p(\Psi_{top}|\Psi_{ret},\Psi_{\lambda}) \sim \text{Exp}(\eta)$. For the inheritance probabilities Γ, we use a uniform prior on [0, 1], though a Beta distribution would also be appropriate in general cases (see S1 Text for full details).

A reversible-jump Markov chain Monte Carlo sampler

As computing the posterior distribution given by Eq (3) is computationally intractable, we implement an Markov chain Monte Carlo (MCMC) Metropolis-Hastings algorithm. While we introduced the inheritance probabilities Γ as one parameter per reticulation node and locus, in practice, this results in a scenario where the number of parameters grows with the number of loci. Therefore, we make the simplifying assumption that there is one inheritance probability per reticulation node that is the same across all loci. In this case, Γ is a vector of length k, where k is the number of reticulation nodes in Ψ.

The description given hereafter assumes that the gene trees in the input are given by their topologies and their branch lengths are ignored. When branch lengths of the gene trees are taken into account, they pose temporal constraints on the phylogenetic network and change the moves allowed, as well as some of the quantities computed in the algorithm below. However, it is important to note that, in practice, coalescent times tend to be underestimated and that this underestimation results in biased phylogenetic estimates when sampling of loci is increased [42]. For the three biological data sets we consider below, we computed the branch lengths of the gene trees and plotted their distributions. In agreement with [42], the estimated branch lengths are very low and would result in phylogenetic networks all of whose nodes are roughly at the same level as that of the leaves. We further discuss this issue below.

In each iteration of the sampling, a new state (Ψ_i, Γ_i) is proposed and either accepted or rejected according to the Metropolis-Hastings ratio r, which is composed of the likelihoods, priors, and Hastings ratio. When the new sample changes dimensionality with respect to the current sample (which occurs only under two moves: adding a new reticulation or removing an existing reticulation), the Jacobian is also taken into account in the ratio, which results in a reversible-jump MCMC, or RJMCMC [43]. To compute the Hastings ratio, we follow the technique of [36] and illustrated by [44] for phylogenetic trees. Using this technique, given the current state $\mathbf{x}$, a set of random numbers $\mathbf{u}$ is generated using a probability distribution with the joint probability density $g(\mathbf{u})$. A deterministic function generates the new proposed state $\mathbf{x}' = h(\mathbf{x}, \mathbf{u})$. To calculate the Hastings ratio, we need to account for the move that would reverse the effects of the move $\mathbf{x} \rightarrow \mathbf{x}'$. To propose $\mathbf{x}$ given state $\mathbf{x}'$, a new set of random numbers, $\mathbf{u}'$ is generated according to a distribution with density $g'(\mathbf{u}')$. Then, $\mathbf{x} = h'(\mathbf{x}', \mathbf{u}')$ where h' is another deterministic function. Green [36] replaced the Hastings ratio by

$$\frac{g'(\mathbf{u}')}{g(\mathbf{u})}|J|,$$

where J is the Jacobian of the transformation from $\{\mathbf{x}, \mathbf{u}\}$ to $\{\mathbf{x}', \mathbf{u}'\}$.

Our algorithm employs seven moves to propose a new state of the Markov chain, illustrated in Fig 2 and the Methods section, along with their respective Hastings ratios whose full derivation is given in S1 Text.

We implemented our method in PhyloNet [40], which is a publicly available, open-source software package for phylogenetic network inference and analysis. We studied the performance of the method on simulated data and three biological data sets.

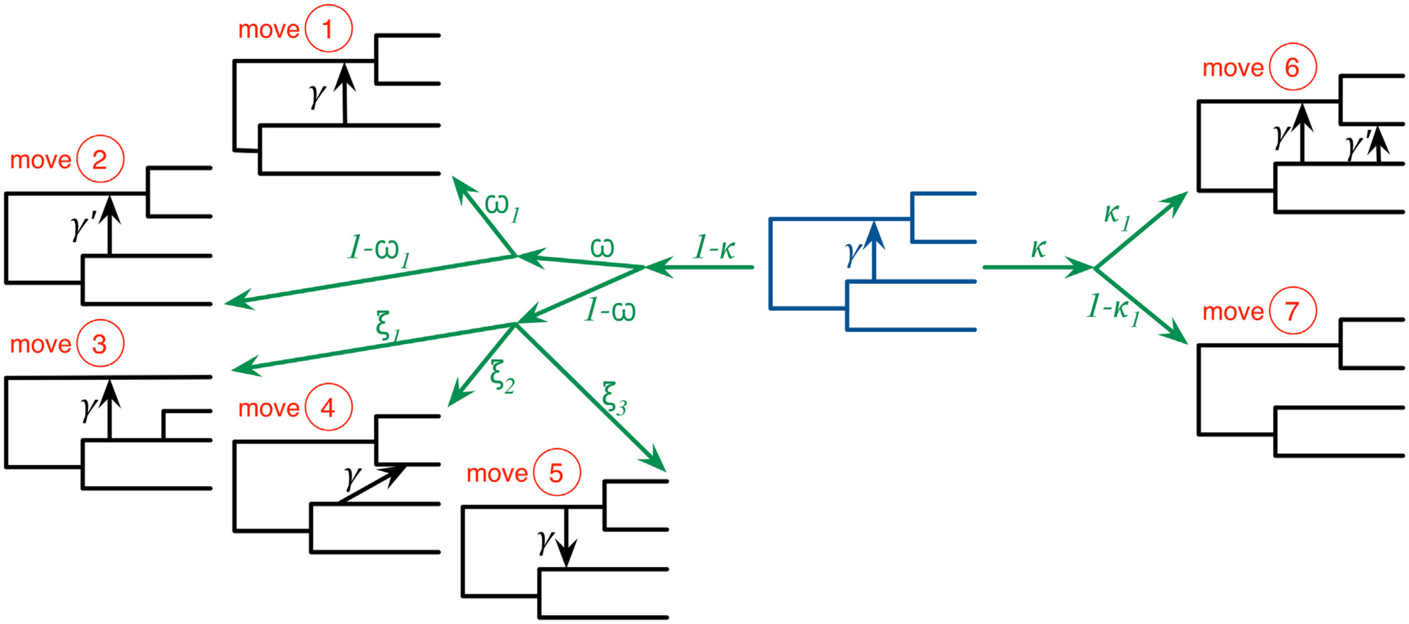


Fig 2. The seven moves that the MCMC sampler utilizes can be classified into ones that do not modify the topology of the phylogenetic network (moves 1 and 2), ones that modify the topology but do not change the model's dimensions (moves 3, 4, and 5), and ones that modify the topology and model's dimensions (moves 6 and 7). The current-state phylogenetic network is shown at the center in dark blue, and the resulting next-state phylogenetic network after each moves is show in black lines. Moves 1 and 2 modify branch lengths and inheritance probabilities, respectively. Moves 3–5 relocate one of the children of a tree node, relocate the head of a reticulation edge, and reverse the direction of a reticulation edge, respectively. Moves 6 and 7 add and remove a reticulation edge, respectively. The probabilities κ and ω determine which of the three groups of moves is selected in an iteration. Within each group, an edge is selected and a move is selected uniformly at random among all the ones that are applicable to the selected edge within that group.

doi:10.1371/journal.pgen.1006006.g002

Performance on simulated data

For the simulations, we used three model phylogenetic networks whose topologies and branch lengths were inspired by the estimated phylogenetic networks of the mosquito data set in [31]. Each of these networks has seven taxa, one of which is designated as an outgroup. Unfortunately, due to the prohibitive running times of computing likelihoods of networks, we currently cannot experiment with much larger (in terms of the number of taxa and/or number of reticulations) networks. The branch lengths vary from very short (about 0.5 coalescent units) to longer ones (about 1.5 coalescent units). The networks differ in the numbers of reticulations they posses (1, 2, and 3), as well as in the inheritance probabilities associated with them.

We generated gene trees for varying numbers of loci (128, 320, 800, and 2000) within each of the three networks under the multispecies coalescent process. Then, using a population mutation rate $\theta = 0.036$, we simulated 1000-nucleotide sequences on the generated gene trees under the general time-reversible (GTR) model. Finally, for each generated alignment, we inferred 100 bootstrap trees under maximum likelihood and used those estimated gene trees as the data for our method. It is important to note that the estimated gene trees differed from the true gene trees on average in about 10% of their branches, with a standard deviation of about 10% as well. See S1 Text for the exact details of the model phylogenetic networks as well as the generated data.

Our results show that for the 1-reticulation model phylogenetic network, the 95% credible set consists of only one topology that is identical to the model network, regardless of the number of loci used. For the 2-reticulation model phylogenetic network, the 95% credible set on 128 and 320 loci consists of a single network that differs from the true network only in missing one of the two reticulations, whereas the 95% credible set on 800 and 2000 loci consists of the true phylogenetic network alone. For the 3-reticulation model network, using 128 and 320 loci resulted in 95% credible sets with one and two reticulations, respectively, of the true set of three reticulations. For 800 and 2000 loci, the 95% credible set consists of three different phylogenetic networks. However, these three networks are indistinguishable based on likelihood using the data, in the sense that their branch lengths and inheritance probabilities could be optimized to yield the same gene tree distributions. In this case, the differences among the topologies stem from different temporal orderings of the reticulation events involving the same pair of taxa. To summarize these results, the method performs very well in terms of recovering the true evolutionary history, including the number and placement of the reticulation events. For smaller numbers of loci, the method obtained networks that are missing one or two of the reticulations, but the rest of the evolutionary history was correct. In other words, for these smaller numbers of loci, the false positive rate was effectively 0.

In terms of runtime, the method took about 2.8 hours to run for 5 million iterations on the smallest data set (one reticulation and 128 loci) and about 9.2 hours for the same number of iterations on the largest data set (three reticulations and 2000 loci). The bottleneck in these computations comes from the likelihood calculations. The phylogenetic network topology and gene tree topology both play a role in a large variations in likelihood computation times, as reflected by large standard deviations when averaging times over the different distinct gene trees.

Full results of the performance of the method on the simulated data in terms of phylogenetic network quality and runtimes are given in S1 Text.

Reticulate phylogenies of wheat, mosquito, and mouse genome data sets

In addition to the synthetic data, we analyzed a bread wheat genome data set from [29], a mosquito genome data set from [31], and a house mouse genome data set from [30]. It is important

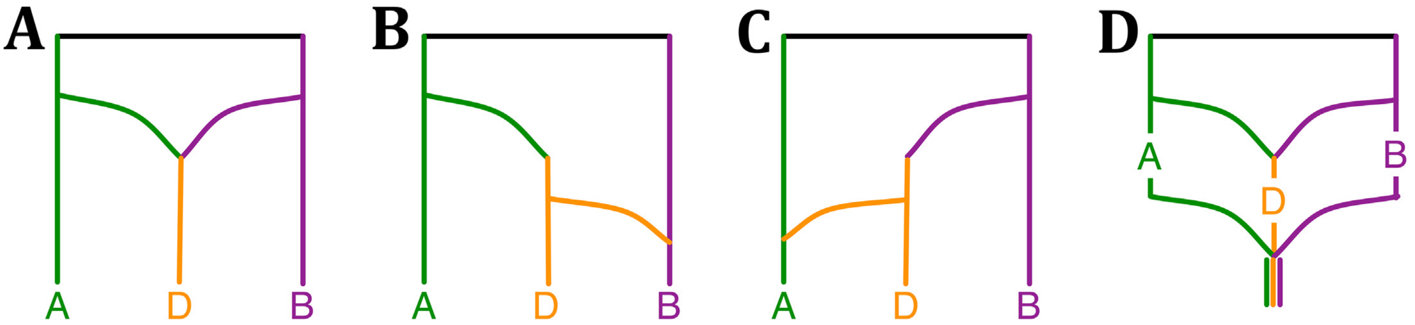


Fig 3. Phylogenetic history of the bread wheat. (**A**–**C**) The three phylogenetic networks that comprise the 95% credible set, (**D**) and a plausible summary of the three networks that is consistent with the model of phylogenetic history of bread wheat (Fig 3 in [29]).

doi:10.1371/journal.pgen.1006006.g003

to note that the wheat and mosquito data sets consist of the genomes of different species, whereas the house mouse data set consists of multiple genomes from the same species. Thus, these analyses highlight the applicability of our method to the detection of intra- and inter-specific hybridization, as well as the challenges that arise in the different evolutionary scenarios.

The bread wheat data set consists of three subgenomes of wheat: *Triticum aestivum*, TaA (A subgenome), TaB (B subgenome) and TaD (D subgenome). Marcussen *et al.* found that each of the A and B lineages is more closely related to D than to each other, as represented by the phylogenetic network in Fig 3**A** that they inferred using the parsimony approach of [45]. Based on this network, they proposed an evolutionary history of *Triticum aestivum*, where the D genome originated from the A and B genome lineages, AABB originated from AA and BB, finally AABB and DD led to origination of AABBDD by polyploidizations and hybridizations, as shown in Fig 3 in [29].

To analyze this data set, we constructed bootstrap trees from the sequences of 2269 genes provided in [29]. The MCMC chains converged fast within a short period of burn-in, as indicated by the trace plot. The three phylogenetic network topologies in the 95% credible set are shown in Fig 3**A**–3**C**. A plausible summary of the three networks is shown in Fig 3**D**, which is consistent with the model of bread wheat proposed by Marcussen *et al.* In terms of runtime, the five million iterations of the MCMC sampling took about 2.2 hours.

While we only used gene tree topologies here, we also inferred gene trees with branch lengths under maximum likelihood. We then estimated for every pair of species the coalescence times based on all 2269 gene trees. We observed that the median pairwise distance for each pair of taxa was around 0.025, and with minimum distances of 0. Since each pairwise distance poses an upper bound on the time of the most recent common ancestor (MRCA) of that pair of species (considering the time at the leaves to be 0), then inference of phylogenetic networks using the likelihood function employed here would result in sampling only phylogenetic networks all of whose nodes have time 0. In other words, using gene tree branch lengths here would result in uninformative phylogenetic networks.

The mosquito data set [31] consists of the four autosomes and X chromosome of six species from the *An. gambiae* complex: *An. gambiae* (G), *An. coluzzii* (C), *An. arabiensis* (A), *An. quadriannulatus* (Q), *An. merus* (R) and *An. melas* (L). This data set was collected and analyzed by Fontaine *et al.* [31]. In that study, the authors inferred a species tree based on the X chromosome and postulated two major hybridization events to explain the extensive introgression. More recently, Wen *et al.* [33] reanalyzed the data set using the maximum likelihood method of [30] while restricting the number of reticulations to three, due to computational

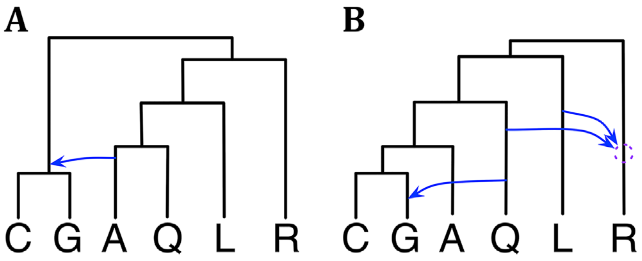


Fig 4. Phylogenetic history of the six mosquito genomes. (**A**) The single phylogenetic network in the 95% credible set sampled on the X chromosome data. (**B**) A summary of the three phylogenetic networks in the 95% credible set sampled on the autosome data. The dotted circle indicates the temporal order of the two reticulation events involving R cannot be discerned with confidence from the data.

doi:10.1371/journal.pgen.1006006.g004

requirements. Their results corroborated parts of the evolutionary history presented in [31] and provided a different scenario for other parts.

We reanalyzed this data set using our Bayesian method. We first analyzed the X chromosome data alone. The 95% credible set consists of a single phylogenetic network, shown in Fig 4A, that agrees with [31, 33]. We then analyzed the autosome data. The 95% credible set consists of three phylogenetic networks that have the same three reticulation events but differ in terms of their temporal orders. As discussed above, these three are indistinguishable under our model, and their summary is given in Fig 4B. The result is in agreement with that in [33]. However, an important point here is that in this analysis, we did not bound the number of reticulations. This number was inferred as a function of the data used and the prior setting. In contrast, in [33], the number of reticulations was bounded by three, for computational reasons. In terms of runtime, this analysis took about 7.65 hours.

Given that the data set was larger than the wheat data set, we also experimented here with the prior and amount of data. In particular, we tried three values for the Poisson prior mean: 0.1, 1.0, and 10. Furthermore, while the full data set used here consists of 2791 loci, we also sampled 311 and 931 loci to create two additional data sets with smaller numbers of loci. We then conducted sampling under each combination of prior setting and data set size, our hypothesis being that as the amount of data increases, the effect of the prior setting would diminish. Indeed, we found that as the number of loci increases, the 95% credible set becomes the same regardless of the Poisson prior mean value. For the smallest data of 311 loci, a mean value of 0.1, resulted in a 95% credible set with one phylogenetic network. Yet, when the mean value was changed to 1.0 or 10, the 95% credible set contained 4 phylogenetic networks. These results further demonstrate a desirable behavior of the method.

It is worth mentioning that when setting the Poisson prior mean value to 10, the runtime increased significantly. For example, on the data set with 2791 loci, it took about 30 hours for the 5 million iterations. This is because the chain samples in this case networks with larger numbers of reticulations and whose likelihood computation time is large.

Just as in the case of the wheat data set, we also computed the pairwise distances among species based on estimated gene tree branch lengths. Once again, the results point to minimum pairwise estimates that are very close to 0 (they equal 0 for some pairs). In this case, inference the uses our likelihood formulation and gene tree branch lengths would result in an uninformative network.

Finally, the mouse data set consists of individuals sampled from five *Mus musculus* populations: two samples of *M. m. domesticus* from France (DF) and Germany (DG), and three samples of *M. m. musculus* from the Czech Republic (MZ), Kazakhstan (MK), and China (MC). For the gene genealogies, 20,639 trees were inferred for sampled loci; see [30] for details. Yu *et al.* found two main introgressions by maximum likelihood regularized by cross-validation. One involves the MRCA of {DF, DG} as a recipient population and MK, MC, or their MRCA as the donor population. The other involves MZ as a recipient population and DF, DG, or their MRCA as the donor population.

This data set differs from the previous two data sets in a significant way: The five genomes are obtained from individuals of the same species (*M. musculus*). That is, this is a data set with very low divergence. Indeed, when estimating gene trees with branch lengths, we found that all pairwise distances among the five individuals were 0; in this case, even the medians were mostly 0. Since the number of loci is very large in this case, we first analyzed the estimated gene trees for resolution. For each locus, we computed the majority-rule consensus of the 100 bootstrap trees of that locus, and counted the number of internal branches in the resulting tree. 11,457 loci had fully-resolved majority-rule consensus trees. Within this set of trees, and out of all the 105 possible binary trees on 5 taxa, 3 trees appeared with a frequency greater than 2000 each, 7 trees appeared with a frequency in the range [200, 399] each, and 11 trees appeared with a frequency in the range [50–199] each. The other 84 binary trees on 5 taxa appeared with a frequency smaller than 50 (7 trees did not appear at all in this set). In our reanalysis of this data set, we used the first set of 3 + 7 + 11 = 21 trees. In total, this amounted to using 10,575 loci.

Since the taxa correspond to individuals from the same species, more extensive gene flow is expected. Indeed, the analysis resulted in a 95% credible set with six different phylogenetic network topologies. A plausible summary of these six networks is shown in Fig 5 (the actual six networks are shown in S1 Text). This network indicates significant hybridization involving MZ. Furthermore, MC and DG are not involved in any of the hybridization events, while their ancestors and sibling taxa are involved. The network gives rise to a hypothesis of hybridization involving the ancestors of {MC,MK} and {DF, DG}, which are two different subspecies, *M. m. musculus* and *M. m. domesticus*, respectively.

In terms of computational runtime, this analysis too about 45 hours, which is much longer than the other two data sets. This is a reflection of the larger size of the data, and the complexity of the networks visited during the MCMC walk.

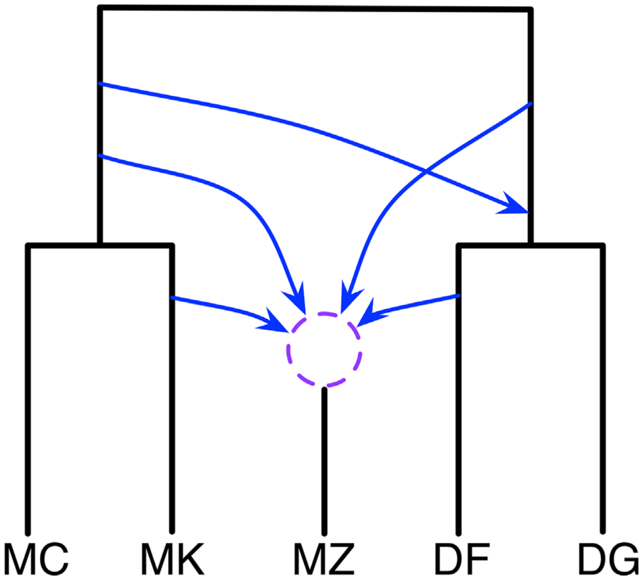


Fig 5. Phylogenetic history of the five mouse genomes. A summary of the six phylogenetic networks that comprise the 95% credible set. The dashed circle indicates that the different networks in the 95% credible set resolve the order of the hybridization events involving MZ in different ways.

doi:10.1371/journal.pgen.1006006.g005

Discussion

To conclude, we have devised a Bayesian framework for phylogenetic networks under the multispecies network coalescent (MSNC). In this work, the prior on the network topology allows for controlling for overfitting, as estimated phylogenetic networks could get arbitrarily complex otherwise. To enable sampling the posterior distribution of phylogenetic networks, we devised a reversible jump Markov chain Monte Carlo (RJMCMC) Metropolis-Hastings algorithm that employs a set of moves for sampling the states of the Markov chain. Implementation of the algorithm is available in the open-source software package PhyloNet [40]. We demonstrated the utility of our method on simulated data and two biological data sets. Despite its expensive computational requirements, Bayesian inference has been used extensively in the context of phylogenetic tree inference as implemented in popular software tools and programs such as MrBayes [46], *BEAST [8], BEST [7], and SNAPP [9]. Our work provides the first framework for Bayesian inference under the MSNC. The most relevant work is that of [39]. However, that work differs from ours in several significant ways.

First, the inference was performed on an ultrametric model called ancestral recombination graphs (ARGs) and the likelihood and prior computation are different. Second, the designs of the RJMCMC, mainly the moves and hasting ratios, are very different. Third, the work here assumes independent loci and inferred gene trees (while accounting for uncertainty), where the work of Bloomquist and Suchard applies to sequences and delineates recombination breakpoints. Last but not least, an implementation of the method of [39] is not publicly available which makes it hard to assess and evaluate. In particular, we employed mixing and convergence tests here, whereas it is unclear how the method of [39] performed in terms of these criteria.

Bayesian inference, particularly in the context of sampling the posterior distribution rather than obtaining a single maximum a posteriori (MAP) estimate, has become commonplace in phylogenetics. Here, we demonstrated how to extend this framework to phylogenomic analyses that account simultaneously for reticulation and incomplete lineage sorting. While the work provides a significant step in that direction, we identify several directions for further improvements. First, the full likelihood computation is very computationally extensive. It is in fact prohibitive except for data sets with small numbers of taxa and reticulation. Pseudo-likelihood-based computation was recently introduced [47, 48]. However, phylogenetic networks are not identifiable by sets of rooted triplets; hence, networks estimated based on pseudo-likelihood might differ from the true ones. Developing efficient algorithms and techniques with theoretical guarantees for estimating the likelihood of a phylogenetic network is imperative. Second, while proper priors have been developed for species trees based on models such as the birth-death model (e.g., employed by [8]), no such priors currently exist for phylogenetic networks. The prior we introduced here is a first step, but a more principled prior that captures the growth of networks just like a birth-death model captures the growth of a tree is needed. Third, while we penalized against model complexity via a Poisson distribution on the number of reticulations, devising other regularization terms would provide an alternative approach. Last but not least, while there exist standard techniques for summarizing trees sampled from the posterior distribution, such as strict or majority-rule consensus, no methods for summarizing phylogenetic networks exist. Developing methods that summarize phylogenetic networks would have impact beyond Bayesian inference.

As we discussed above, we focused here on gene tree topologies alone. The inference method can be extended to work on gene trees with their estimated branch lengths by modifying the set of operations and using the probability density function of [30]. However, under the formulation we gave here and that in [30], the coalescence times estimated for the gene trees constrain the speciation and hybridization times associated with the nodes in the phylogenetic network.

For example, if at least one gene tree gives an estimate of zero for coalescence time between A and B, then either the divergence time between A and B must be set at zero in the network, or a contemporary hybridization between these two taxa must be invoked. DeGiorgio and Degnan [42] studied the effect of branch length underestimation in gene trees and its effect on species tree inference. As the study focused on gene tree estimation independently of the species phylogeny, similar issues carry over to the domain of phylogenetic network inference. Indeed, we showed for all three biological data sets studied here that divergence times based on estimated gene tree branch lengths would be problematic in terms of the network inference.

Our work is, to our knowledge, the first Bayesian approach for phylogenetic network inference under the multispecies network coalescent. The developed method allows for sampling the posterior of phylogenetic networks from multi-locus data sets while accounting for incomplete lineage sorting and hybridization. We demonstrated through analyses of simulated and biological data sets that the method performs well in practice and that it provides a powerful analytical tool for phylogenomic analyses. In particular, as the role and extent of hybridization and subsequent introgression in eukaryotic genomes continue to be investigated, we believe our method will provide a means for such a systematic investigation.

Materials and Methods

Phylogenetic networks

A reticulate, i.e., non-treelike, evolutionary history that arises in the presence of processes such as hybridization and horizontal gene transfer is best represented by a phylogenetic network.

Definition 1 *A* phylogenetic $\mathscr{X}$-network, *or* $\mathscr{X}$*-network for short*, Ψ *is a directed, acyclic graph (DAG) with* $V = \{r\} \cup V_L \cup V_T \cup V_N$, *where*

- $indeg(r) = 0$ (*r is the root of* Ψ);
- $\forall v \in V_L$, $indeg(v) = 1$ *and* $outdeg(v) = 0$ (V_L *are the* external tree nodes, *or* leaves, *of* Ψ);
- $\forall v \in V_T$, $indeg(v) = 1$ *and* $outdeg(v) \geq 2$ (V_T *are the* internal tree nodes *of* Ψ); *and,*
- $\forall v \in V_N$, $indeg(v) = 2$ *and* $outdeg(v) = 1$ (V_N *are the* reticulation nodes *of* Ψ),

$E \subseteq V \times V$ *are the network's edges, including*reticulation edges *whose heads are reticulation nodes, and* tree edges *whose heads are tree nodes., and* ℓ: $V_L \rightarrow \mathscr{X}$ *is the* leaf-labeling *function, which is a bijection from* V_L *to* $\mathscr{X}$.

We use $V(\Psi)$ and $E(\Psi)$ to denote the set of nodes and edges of phylogenetic network Ψ respectively. In addition to the topology of a phylogenetic network Ψ, each edge $b = (u, v)$ in $E(\Psi)$ has a length λ_b measured in coalescent units, which is the number of generations divided by effective population size on that branch.

The multispecies network coalescent (MSNC)

As an orthologous genomic region from a set $\mathscr{X}$ of species evolves within the branches of the species phylogeny of $\mathscr{X}$, the genealogy of this region, also called *gene tree*, can be viewed as a discrete random variable whose values are all possible gene tree topologies on the set of genomic regions. When the gene tree branch lengths are also taken into account, the random variable becomes continuous. Yu *et al.* [35] gave the probability mass function (pmf) for this discrete random variable given the phylogenetic network Ψ and an additional parameter Γ that contains the inheritance probabilities associated with reticulation nodes, which we now describe briefly.

Let $E_R \subseteq E(\Psi)$ be the set of reticulation edges, and $\rho = |E_R|$, and consider a data set that consists of m independent loci. The inheritance probabilities are given by a $\rho \times m$ matrix Γ such that for every $1 \leq j \leq m$:

1. $\Gamma[b, j] \in [0, 1]$ for every $b \in E_R$, and
2. $\Gamma[b, j] + \Gamma[b', j] = 1$ for every distinct pair $b, b' \in E_R$ such that b and b' are incident into the same reticulation node.

For an edge b incident into node v in Ψ, the entry $\Gamma[b, j]$ denotes the probability that a sample from locus i tracks branch b when "entering" the population represented by node v.

The mass and density functions of gene trees given a phylogenetic network, its branch lengths, and inheritance probabilities were derived in [30, 34, 35, 37]; see S1 Text for a brief discussion of these two functions. Furthermore, Yu *et al.* discussed unidentifiability issues of phylogenetic networks and their parameters from gene tree topologies [35].

Inference using RJMCMC

The general form of the Metropolis-Hastings algorithm that we implement is given in Algorithm 1. While the algorithm is described in a way that all accepted samples are returned, in the actual implementation all samples from a burn-in period are discarded, and only a small percentage of the samples beyond that are stored (the burn-in period and percentage of samples to store are both user-defined in our implementation).

```
Algorithm 1: Metropolis-Hastings.
Input: A set of gene trees $G$ and the number of iterations $N$.
Output: A collection $S$ of $(\Psi, \Gamma)$ samples.
Initialize $\Psi_0, \Gamma_0$;
for $i \leftarrow 1$ to $N$ do
  $\Psi_i, \Gamma_i \leftarrow propose(\Psi_{i-1}, \Gamma_{i-1})$;
  if proposal does not change dimensionality then
    $r \leftarrow \begin{pmatrix}\text{likelihood}\\ \text{ratio}\end{pmatrix} \times \begin{pmatrix}\text{prior}\\ \text{ratio}\end{pmatrix} \times \begin{pmatrix}\text{Hastings}\\ \text{ratio}\end{pmatrix}$;
  end
  else
    $r \leftarrow \begin{pmatrix}\text{likelihood}\\ \text{ratio}\end{pmatrix} \times \begin{pmatrix}\text{prior}\\ \text{ratio}\end{pmatrix} \times \begin{pmatrix}\text{Hastings}\\ \text{ratio}\end{pmatrix} \times (\text{Jacobian})$;
  end
  if $r < random(0, 1)$ then
    $\Psi_i \leftarrow \Psi_{i-1}$;
    $\Gamma_i \leftarrow \Gamma_{i-1}$;
  end
  $S \leftarrow S \cup \{(\Psi_i, \Gamma_i)\}$;
end
return $S$;
```

The function $propose(\Psi, \Gamma)$ proposes a sample based on the current sample and set of predefined moves that are listed in Table 1 and illustrated in Fig 2 above. As we described above, the moves might generate a phylogenetic network topology that deviates "slightly" from the conditions of Definition 1. What we mean by "slightly" is that the network could violate Definition 1 in one of the following ways:

- the proposed network topology has a cycle (moves 3–5 in Table 1 might cause this);
- the proposed network topology has two edges with the same tail and head (moves 3 and 6 in Table 1 might cause this); or,
- the proposed network has more than a single node of in-degree 0 (move 3 in Table 1 might cause this).

Table 1. The 7 moves that the Metropolis-Hastings algorithm employs.

1. Change-Length:	Modifies the length of a randomly selected edge
2. Change-Inheritance:	Modifies the inheritance probability of a randomly selected reticulation edge
3. Move-Tail:	Modifies the tail of a randomly selected edge whose tail is a tree node
4. Move-Head:	Modifies the head of a randomly selected edge whose head is a reticulation node
5. Flip-Reticulation:	Reverses the direction of a randomly selected reticulation edge
6. Add-Reticulation:	Adds a reticulation edge between two randomly selected edges
7. Delete-Reticulation:	Deletes a randomly selected reticulation edge

The function *propose* in Algorithm 1 selects one of these randomly and applies it to the current sample to generate a new one. Moves 1 and 2 do not change the model dimension or the network's topology. Moves 3–5 change the network's topology but not the model dimension. Moves 6 and 7 change the network's topology and model dimension.

doi:10.1371/journal.pgen.1006006.t001

Therefore, in computing the Metropolis-Hastings ratio, our implementation explicitly tests whether the proposed network topology has any of these properties; if it does, the phylogenetic network prior is set to 0, and otherwise, the prior is set based on Eq 4.

The function *propose* selects the move as follows:

- With probability κ, one of the two dimension-changing moves (moves 6 and 7 in Table 1) is selected. If the current network has at least one reticulation edge, then both moves 6 and 7 are possible. Add-Reticulation is selected with probability κ_1 (and Delete-Reticulation is selected with probability $1 - \kappa_1$). If the current network has no reticulation edges (i.e., it is a tree), then Add-Reticulation is selected.
- With probability $1 - \kappa$, a non-dimension-changing move (moves 1–5 in Table 1) is selected.
    - With probability ω a non-topology-changing move (moves 1 and 2 in Table 1) is selected. If the current network has at least one reticulation edge, then Change-Length is selected with probability ω_1 and Change-Inheritance is selected with probability $1 - \omega_1$. If the current network has no reticulation edges, then Change-Length is selected.
    - With probability $1 - \omega$ a topology-changing move (moves 3–5 in Table 1) is selected. Moves 3, 4, and 5 are selected with probabilities ξ_1, ξ_2, and ξ_3, respectively, where $\xi_1 + \xi_2 + \xi_3 = 1$.

Full derivation of the Hastings ratios for all seven moves is given in S1 Text.

Supporting Information

S1 Text. Supporting information file that contains method details and additional results.
(PDF)

Acknowledgments

The authors would like to thank Guest Editor, Dr. Scott Edwards, and the two anonymous reviewers for the very helpful comments on an earlier draft of this manuscript that helped us improve the method and manuscript.

Author Contributions

Conceived and designed the experiments: DW YY LN. Performed the experiments: DW. Analyzed the data: DW YY LN. Contributed reagents/materials/analysis tools: DW YY. Wrote the paper: DW YY LN. Designed the software used in analysis: DW YY.

References

1. Pamilo P, Nei M. Relationships between gene trees and species trees. Molecular biology and evolution. 1988; 5(5):568–583. PMID: 3193878
2. Maddison WP. Gene trees in species trees. Systematic biology. 1997; 46(3):523–536. doi: 10.1093/sysbio/46.3.523
3. Page RD, Charleston MA. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Molecular phylogenetics and evolution. 1997; 7(2):231–240. doi: 10.1006/mpev.1996.0390 PMID: 9126565
4. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics. 2005; 6(5):361–375. doi: 10.1038/nrg1603 PMID: 15861208
5. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in ecology & evolution. 2009; 24(6):332–340. doi: 10.1016/j.tree.2009.01.009
6. Hudson RR, et al. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology. 1990; 7(1):44.
7. Liu L. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 2008; 24(21):2542–2543. doi: 10.1093/bioinformatics/btn484 PMID: 18799483
8. Heled J, Drummond AJ. Bayesian inference of species trees from multilocus data. Molecular biology and evolution. 2010; 27(3):570–580. doi: 10.1093/molbev/msp274 PMID: 19906793
9. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Molecular biology and evolution. 2012; 29(8):1917–1932. doi: 10.1093/molbev/mss086 PMID: 22422763
10. Kubatko LS, Carstens BC, Knowles LL. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 2009; 25(7):971–973. doi: 10.1093/bioinformatics/btp079 PMID: 19211573
11. Wu Y. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution. 2012; 66(3):763–775. doi: 10.1111/j.1558-5646.2011.01476.x PMID: 22380439
12. Liu L, Yu LL, Kubatko L, Pearl DK, Edwards SV. Coalescent methods for estimating phylogenetic trees. Molecular Phylogenetics and Evolution. 2009; 53:320–328. doi: 10.1016/j.ympev.2009.05.033 PMID: 19501178
13. Nakhleh L. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends in Ecology & Evolution. 2013; 28(12):719–728. doi: 10.1016/j.tree.2013.09.004
14. Liu L, Xi Z, Wu S, Davis CC, Edwards SV. Estimating phylogenetic trees from genome-scale data. Annals of the New York Academy of Sciences. 2015; 1360(1):36–53. doi: 10.1111/nyas.12747 PMID: 25873435
15. Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. Molecular biology and evolution. 2002; 19(12):2226–2238. doi: 10.1093/oxfordjournals.molbev.a004046 PMID: 12446813
16. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification 1. Annual Reviews in Microbiology. 2001; 55(1):709–742. doi: 10.1146/annurev.micro.55.1.709
17. Arnold ML. Natural Hybridization and Evolution. Oxford: Oxford University Press; 1997.
18. Barton NH. The role of hybridization in evolution. Molecular Ecology. 2001; 10(3):551–568. doi: 10.1046/j.1365-294x.2001.01216.x PMID: 11298968
19. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. 2005; 20(5):229–237. doi: 10.1016/j.tree.2005.02.010 PMID: 16701374
20. Mallet J. Hybrid speciation. Nature. 2007; 446:279–283. doi: 10.1038/nature05706 PMID: 17361174
21. Rieseberg LH. Hybrid origins of plant species. Annu Rev Ecol Syst. 1997; 28:359–389. doi: 10.1146/annurev.ecolsys.28.1.359
22. Mallet J, Besansky N, Hahn MW. How reticulated are species? BioEssays. 2016; 38(2):140–149. doi: 10.1002/bies.201500149 PMID: 26709836

23. Morrison DA. Networks in phylogenetic analysis: new tools for population biology. International journal for parasitology. 2005; 35(5):567–582. doi: 10.1016/j.ijpara.2005.02.007 PMID: 15826648

24. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Molecular biology and evolution. 2006; 23(2):254–267. doi: 10.1093/molbev/msj030 PMID: 162218396

25. Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, et al. Networks: expanding evolutionary thinking. Trends in Genetics. 2013; 29(8):439–441. doi: 10.1016/j.tig.2013.05.007 PMID: 23764187

26. Park H, Jin G, Nakhleh L. Algorithmic strategies for estimating the amount of reticulation from a collection of gene trees. In: Proceedings of the 9th Annual International Conference on Computational Systems Biology; 2010. p. 114–123.

27. Wu Y. Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. Bioinformatics. 2010; 26(12):i140–i148. doi: 10.1093/bioinformatics/btq198 PMID: 20529899

28. Albrecht B, Scornavacca C, Cenci A, Huson DH. Fast computation of minimum hybridization networks. Bioinformatics. 2012; 28(2):191–197. doi: 10.1093/bioinformatics/btr618 PMID: 22072387

29. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, et al. Ancient hybridizations among the ancestral genomes of bread wheat. Science. 2014; 345(6194):1250092. doi: 10.1126/science.1250092 PMID: 25035499

30. Yu Y, Dong J, Liu KJ, Nakhleh L. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences. 2014; 111(46):16448–16453. doi: 10.1073/pnas.1407950111

31. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 2015; 347 (6217):1258524. doi: 10.1126/science.1258524 PMID: 25431491

32. Clark AG, Messer PW. Conundrum of jumbled mosquito genomes. Science. 2015; 347(6217):27–28. doi: 10.1126/science.aaa3600 PMID: 25554775

33. Wen D, Yu Y, Hahn MW, Nakhleh L. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Molecular Ecology. 2016;. doi: 10.1111/mec.13544 PMID: 26808290

34. Yu Y, Than C, Degnan JH, Nakhleh L. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Systematic Biology. 2011; 60(2):138–149. doi: 10.1093/sysbio/syq084 PMID: 21248369

35. Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS genetics. 2012; 8(4):e1002660. doi: 10.1371/journal.pgen.1002660 PMID: 22536161

36. Green PJ. Trans-dimensional Markov chain Monte Carlo. In: Green PJ, Hjort NL, Richardson S, editors. Highly Structured Stochastic Processes. Oxford, UK: Oxford University Press; 2003. p. 179–198.

37. Yu Y, Ristic N, Nakhleh L. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. BMC Bioinformatics. 2013; 14(Suppl 15):S6. doi: 10.1186/1471-2105-14-S15-S6 PMID: 24564257

38. Lewis PO, Holder MT, Holsinger KE. Polytomies and Bayesian phylogenetic inference. Systematic Biology. 2005; 54(2):241–253. doi: 10.1080/10635150590924208 PMID: 16012095

39. Bloomquist EW, Suchard MA. Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. Systematic Biology. 2010; 59(1):27–41. doi: 10.1093/sysbio/syp076 PMID: 20525618

40. Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC bioinformatics. 2008; 9(1):322. doi: 10.1186/1471-2105-9-322 PMID: 18662388

41. Felsenstein J. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution. 1981; 35:1229–1242. doi: 10.2307/2408134

42. DeGiorgio M, Degnan JH. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Systematic biology. 2014; 63(1):66–82. doi: 10.1093/sysbio/syt059 PMID: 23988674

43. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995; 82(4):711–732. doi: 10.1093/biomet/82.4.711

44. Holder MT, Lewis PO, Swofford DL, Larget B. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. Systematic biology. 2005; 54(6):961–965. doi: 10.1080/10635150500354670 PMID: 16385776

45. Yu Y, Barnett RM, Nakhleh L. Parsimonious Inference of Hybridization in the Presence of Incomplete Lineage Sorting. Systematic Biology. 2013; 62(5):738–751. doi: 10.1093/sysbio/syt037 PMID: 23736104

46. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19(12):1572–1574. doi: 10.1093/bioinformatics/btg180 PMID: 12912839

47. Yu Y, Nakhleh L. A Maximum Pseudo-likelihood Approach for Phylogenetic Networks. BMC Genomics. 2015; 16(Suppl 10):S10. doi: 10.1186/1471-2164-16-S10-S10 PMID: 26450642

48. Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genetics. 2016; 12(3):e1005896. doi: 10.1371/journal.pgen.1005896 PMID: 26950302

Supplementary Material: Bayesian Inference of Reticulate Phylogenies Under the Multispecies Network Coalescent

Dingqiao Wen[1,*], Yun Yu[1], Luay Nakhleh[1,2,*]

1 Computer Science, Rice University, Houston, TX, USA

2 BioSciences, Rice University, Houston, TX, USA

$*$ **E-mail: {dw20,nakhleh}@rice.edu.**

Contents

1 A Bayesian formulation

1.1 The likelihood

As described in the main text, the likelihood of a phylogenetic network and inheritance probabilities is based on gene trees and assumes the trees are estimated from independent loci. The likelihood formulation and computations were derived fully in [1–3]. Since loci are assumed to be independent, the likelihood of a phylogenetic network and vector of inheritance probabilities is given in terms of the mass or density of the independent gene trees. We reproduce the probability mass function (pmf) and probability density function (pdf) of gene trees here for the sake of readability and emphasize that these functions and their computations are not a contribution of this work.

1.2 The pmf of gene tree topologies

Given a phylogenetic network Ψ, we denote by Ψ_u the set of nodes that are reachable from the root of Ψ via at least one path that goes through node $u \in V(\Psi)$. Then given a phylogenetic network Ψ and a gene tree G for some locus j, a coalescent history is a function $h : V(G) \rightarrow E(\Psi)$ such that the following two conditions hold:

- if v is a leaf in G, then $h(v) = (x, y)$ where y is the leaf in Ψ with the label of the species from which the allele labeling leaf v in G is sampled;
- if v is a node in G_u, and $h(u) = (p, q)$, then $h(v) = (x, y)$ where $y \in \Psi_q$.

Given a phylogenetic network Ψ and a gene tree G for locus j, we denote by $H_\Psi(G)$ the set of all coalescent histories of G within the branches of Ψ. Then the pmf of the gene tree is given by

$$\mathbf{P}(G|\Psi, \Gamma) = \sum_{h \in H_\Psi(G)} \mathbf{P}(h|\Psi, \Gamma), \tag{1}$$

where Γ is the inheritance probabilities matrix (see the main text) and $\mathbf{P}(h|\Psi, \Gamma)$ gives the pmf of the coalescent history random variable, which can be computed as

$$\mathbf{P}(h|\Psi, \Gamma) = \frac{w(h)}{d(h)} \prod_{b \in E(\Psi)} \frac{w_b(h)}{d_b(h)} \Gamma[b, j]^{u_b(h)} p_{u_b(h) v_b(h)}(\lambda_b). \tag{2}$$

In this equation, $u_b(h)$ and $v_b(h)$ denote the number of lineages enter and exit edge b of Ψ under coalescent history h. The term $p_{u_b(h) v_b(h)}(\lambda_b)$ is the probability of $u_b(h)$ gene lineages coalescing into $v_b(h)$ during time λ_b [**?**]. And $w_b(h)/d_b(h)$ is the proportion of all coalescent scenarios resulting from $u_b(h) - v_b(h)$ coalescent events that agree with the topology of the gene tree. This quantity without the b subscript corresponds to the root of Ψ.

1.3 The pdf of gene trees with branch lengths

We use $\tau_\Psi(v)$ to denote the height of node v in phylogeny Ψ with branch lengths λ. Given a gene tree G whose branch lengths are given by λ' and a phylogenetic network Ψ whose branch lengths are given by λ, we define a coalescent history with respect to coalescence times to be a function $h : V(G) \rightarrow E(\Psi)$, such that the following condition holds:

- for $h \in H_\Psi(G)$, if $h(v) = (x, y)$ and $\tau_\Psi(x) > \tau_G(v) \geq \tau_\Psi(y)$, then $h(v) = (x, y)$.

The quantity $\tau_G(v)$ indicates at which point of branch (x, y) coalescent event v happens. We denote the set of coalescent histories with respect to coalescence times for gene tree G and phylogenetic network Ψ by $H_\Psi(G)$. Clearly, in this case, the set H depends on λ and λ'.

Given a phylogenetic network Ψ, the pdf of the gene tree (topology and branch lengths) random variable is given by

$$p(G|\Psi, \Gamma) = \sum_{h \in H_\Psi(G)} \mathbf{P}(h|\Psi, \Gamma), \tag{3}$$

where $p(h|\Psi, \Gamma)$ gives the pdf of the coalescent history (with respect to coalescence times) random variable.

Consider a locus j, whose gene tree is G and an arbitrary $h \in H_\Psi(G)$. For an edge $b = (x, y) \in E(\Psi)$, we define $T_b(h)$ to be a vector of the elements in the set $\{\tau_G(w) : w \in h^{-1}(b)\} \cup \{\tau_\Psi(y)\}$ in increasing order. We denote by $T_b(h)[i]$ the i-th element of the vector. Furthermore, we denote by $u_b(h)$ the number of gene lineages entering edge b and $v_b(h)$ the number of gene lineages leaving edge b under h. Then we have

$$p(h|\Psi, \Gamma) = \prod_{b \in E(\Psi)} \left[\prod_{i=1}^{|T_b(h)|-1} e^{-\binom{u_b(h)-i+1}{2}(T_b(h)_{i+1} - T_b(h)_i)} \right] \times e^{-\binom{v_b(h)}{2}(\tau_\Psi(b) - T_b(h)_{|T_b(h)|})} \times \Gamma[b, j]^{u_b(h)}. \tag{4}$$

1.4 Prior distributions

Prior on the phylogenetic network. We define a prior that is similar to that defined on ancestral recombination graphs in [4]. We have

$$p(\Psi|\nu, \delta, \eta) = p(\Psi_{ret}|\nu) \times p(\Psi_\lambda|\delta) \times p(\Psi_{top}|\Psi_{ret}, \Psi_\lambda, \eta). \tag{5}$$

It is important to note here that if Ψ_{top} does not follow the phylogenetic network definition (Main Text), then $p(\Psi|\nu, \delta, \eta) = 0$. This is very important since in the MCMC kernels we describe below, we allow the moves to produce directed graphs that slightly deviate from the definition; in this case, having the prior be 0 guarantees that the proposal is rejected. Using the strategy, rather than defining only "legal" moves simplifies the calculation of the Hastings ratios. See more details below.

We assume a Poisson distribution with hyperparameter ν on the number of reticulation nodes in Ψ, weighted by 1 over the number of networks with that number of reticulations. More specifically, the Poisson prior gives a probability of $\frac{\nu^m e^{-\nu}}{m!}$ for a network having m reticulation nodes. The weight is $1/T_{n,m}$, where n is the number of leaves in the network, and $T_{n,m}$ is the number of networks that have n leaves and m reticulation. Putting these two together, for a phylogenetic network Ψ with n leaves and m reticulation nodes, we have

$$p(\Psi_{ret}|\nu) = \frac{1}{T_{n,m}} \text{Poisson}(m, \nu).$$

- For all moves except Add-Reticulation and Delete-Reticulation, the prior ratio on Ψ_{ret} between the next- and current-state networks is 1.

- For the Delete-Reticulation move, the prior ratio is

$$\frac{T_{n,m}}{T_{n,m-1}} \cdot \frac{\text{Poisson}(m-1, \nu)}{\text{Poisson}(m, \nu)}.$$

- For the Add-Reticulation move, the prior ratio is

$$\frac{T_{n,m}}{T_{n,m+1}} \cdot \frac{\text{Poisson}(m+1, \nu)}{\text{Poisson}(m, \nu)}.$$

To the best of our knowledge, there is no known closed formula or algorithm for calculating $T_{n,m}$ for general values of n and m. However, if k is the number of edges in a phylogenetic network Ψ with m reticulations, then the number of ways to add an additional reticulation edge to Ψ is bounded by $k(k-1)$. Based on this observation, we make use of the recurrence

$$T_{n,m} = T_{n,m-1} \cdot k \cdot (k-1)$$

where $k = 2(n-1) + 3(m-1)$ and $T_{n,0} = c$, where c is, in theory, the number of rooted phylogenetic networks, but in practice can be any non-zero value since in the prior ratio, this value cancels out. In our implementation, for the prior ratios of Delete-Reticulation and Add-Reticulation moves, we evaluate this recurrence explicitly for the numerator and denominator.

This prior penalizes against adding many reticulations. From a biological perspective, it is not unreasonable to have a prior belief of a small number of reticulations. From a computational perspective, computing the likelihood of a network is prohibitively slow. The computational requirements of this step are affected heavily by the number of reticulations and their configurations (that is, how they are placed in the network) [2]. Penalizing against very complex networks helps with the feasibility of these computations.

We assume an exponential distribution on the branch lengths so that every branch length is $\sim \text{Exp}(\delta)$.

While [4] assumed a uniform distribution on all topologies with the same number of reticulation nodes, a reasonable prior for phylogenetic networks is one that favors reticulations between closely related species. This would make a difference particularly in cases where the number of taxa is very large and the species form groups with small divergences within and large divergences across those groups.

Consider a phylogenetic network Ψ and inheritance probabilities Γ, and let x be a reticulation node in Ψ whose two parents in Ψ are v_1 and v_2. Let T be the tree that results from Ψ by removing every edge that has inheritance probability < 0.5 (if both edges incoming a reticulation node have inheritance probability of exactly 0.5, then both edges are removed one at a time and the one that results in the smaller diameter is chosen). Then, the diameter of x, denoted by $d(x)$, is defined as

$$d(x) = l_{r \rightsquigarrow v_1} + l_{r \rightsquigarrow v_2} + \lambda_{(v_1,x)} + \lambda_{(v_2,x)},$$

where r is the MRCA (most recent common ancestor) of v_1 and v_2 in T, $l_{r \rightsquigarrow v_1}$ and $l_{r \rightsquigarrow v_2}$ are the lengths of the paths from r to v_1 and v_2, respectively, in T, and $\lambda_{(v_1,x)}$ and $\lambda_{(v_2,x)}$ are the lengths of the two edges (v_1, x) and (v_2, x), respectively, in Ψ. A figure illustrating the measurement of diameter is shown in Fig. 1. We now assume an exponential distribution on the diameter of a reticulation nodes in Ψ, so that $d(x_i) \sim \text{Exp}(\eta)$ for each reticulation nodes x_i, and we treat the reticulation nodes in the network as independent.

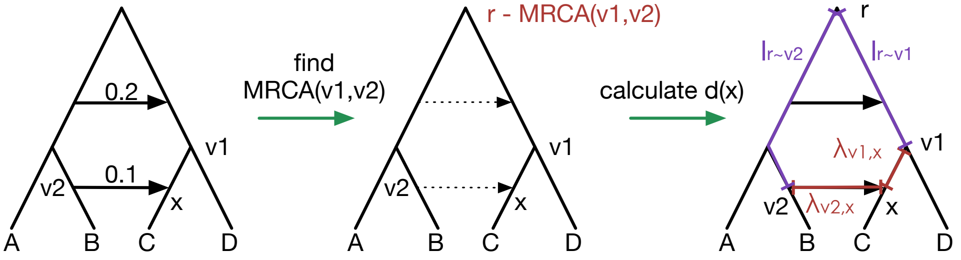


Figure 1. An illustration of the diameter of a reticulation node x.

In the analyses we report in the main manuscript and below, we used a uniform prior on the diameter since the number of taxa was very small and the taxa themselves are very closely related.

Prior on the inheritance probabilities. As discussed above, for each reticulation node, there are two edges incoming into it, b and b'. For every locus i, we associate values $\Gamma[b,i]$ and $\Gamma[b',i]$ such that $\Gamma[b,i] + \Gamma[b',i] = 1$. We propose $\Gamma[b,i] \sim \text{Beta}(\alpha,\beta)$ for a prior. In the absence of any information on the inheritance probabilities, setting $\alpha = \beta = 1$ amounts to using a uniform prior on $[0,1]$. If the amount of introgressed genomic data is suspected to be small in the genome, the hyper-parameters α and β can be appropriately set to bias the inheritance probabilities to values close to 0 and 1 (a U-shaped distribution).

1.5 Sampling the posterior using MCMC

For each of the 7 moves (see Main Text), we now describe how it is implemented and the Hastings ratio (and the Jacobian wherever relevant).

Change-Length. An edge is selected uniformly at random and the branch length ℓ of the edge is modified into ℓ' using the proposal (similar to [5])

$$\ell' = \ell e^{\sigma(u-0.5)}$$

where σ is a tuning parameter and $u \sim \text{Uniform}(0,1)$. The Hastings ratio is $\frac{\ell'}{\ell}$ (derived in [5]). It is important to note that if a single individual is sampled per species (or, per taxon that labels a leaf in the network), then modifying the length of an external branch (a branch that is incident with a leaf) does not affect the likelihood of the network. The same observation holds for the lengths of reticulation edges whose head is the parent of a leaf in the network; see Fig. 2. In this case, edges to which Change-Length is applied exclude the external tree edges and reticulation edges whose head is the parent of a leaf.[1]

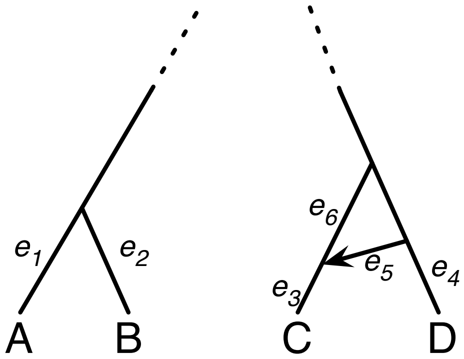


Figure 2. If a single individual is sampled from each of the four taxa A, B, C, D, then the lengths of branches $e_1, \ldots, e_6$ are not identifiable when gene tree topologies are used to infer the phylogenetic network. If, for example, two or more individuals are sampled from taxon C, then the lengths of branches e_3, e_5, and e_6 might be possible to estimate from gene tree topologies. Beyond setting the lengths of such branches immediately upon their creation during the MCMC sampling, their lengths are not sampled; that is, Change-Length is not applied to such branches, unless at least two individuals are sampled "below" the branch.

[1] Note that other branches in the network might have an unidentifiable length (alternatively, modifying their length does not affect the likelihood). Characterizing those is not a simple task, and we focus here on the two types of edges that we have listed, namely external tree branches and reticulation branches whose head is the parent of a leaf.

Change-Inheritance. A reticulation edge is selected uniformly at random from the set of all reticulation edges and the inheritance probability γ associated with it is modified into γ' using the proposal

$$\gamma' = \begin{cases} \gamma + u & \text{if} \quad 0 \leq \gamma + u \leq 1 \\ -(\gamma + u) & \text{if} \quad \gamma + u < 0 \\ 2 - (\gamma + u) & \text{if} \quad \gamma + u > 1 \end{cases}$$

where $u \sim \text{Uniform}(-0.1, +0.1)$. The value 0.1 can be replaced by a tuning parameter for a more general setting. Under this setting, the Hastings ratio is $\frac{p(\gamma|\gamma')}{p(\gamma'|\gamma)} = 1$.

Move-Tail. An edge (x, y_1) is selected uniformly at random from the set of all edges whose tail is a tree node. Let w be the parent of x (if x is the root node, then w does not exist) and y_2 be the second child of x (in addition to y_1). Let v_1 be a node such that $v_1 \notin \{w, x, y_1, y_2\}$ (in particular, v_1 could be the root if neither x nor w is the root). The following operations are performed:

1. If v_1 is not the root, let u_1 be a parent of v_1. Then,
   (a) two new edges are added: (u_1, x) and (x, v_1);
   (b) if x is not the root of the network (node w exists), then (w, y_2) is also added;
   (c) $\lambda_{(u_1,x)} + \lambda_{(x,v_1)} = \lambda_{(u_1,v_1)} = \ell_1$, $\lambda_{(u_1,x)} \sim \text{Uniform}(0, \ell_1)$;
   (d) $\lambda_{(w,y_2)} = \lambda_{(w,x)} + \lambda_{(x,y_2)} = \ell_2$ (if w exists);
   (e) If v_1 was a reticulation node before the move, then $\gamma_{(x,v_1)} = \gamma_{(u_1,v_1)}$;
   (f) If y_2 was a reticulation node before the move, then $\gamma_{(w,y_2)} = \gamma_{(x,y_2)}$; and,
   (g) Finally, delete the edges (along with their parameters): (w, x), (x, y_2), and (u_1, v_1).
2. If v_1 is the root:
   (a) two new edges are added: (x, v_1) and (w, y_2);
   (b) $\lambda_{(x,v_1)} = \ell_r = -\frac{1}{\delta} \ln(1 - w_1)$ where $w_1 \sim \text{Uniform}(0, 1)$;
   (c) $\lambda_{(w,y_2)} = \lambda_{(w,x)} + \lambda_{(x,y_2)}$;
   (d) If y_2 was a reticulation node before the move, then $\gamma_{(w,y_2)} = \gamma_{(x,y_2)}$; and,
   (e) Finally, delete the edges (along with their parameters): (w, x) and (x, y_2).

It is important to note here that we do not allow a selection where x is the root of the network and y_2 is a reticulation node whose parents are x and y_1, since in this case applying this move would result in y_2 becoming a tree node and, consequently, modify the dimension of the model. If the nodes are selected with this configuration, the move is nullified and a new proposal is made.

Hereafter, we use Δt to represent an infinitesimally small region near the proposed point in a distribution. The Hastings ratio for this move is $\frac{\Delta t \cdot 1/\ell_2}{\Delta t \cdot 1/\ell_1} = \frac{\ell_1}{\ell_2}$ when x is not the root before or after proposal, $\frac{\Delta t \cdot 1/\ell_2}{\Delta t \cdot \delta e^{-\delta \ell_r}} = \frac{1}{\ell_2 \cdot \delta e^{-\delta \ell_r}}$ when x is the root after proposal, and $\frac{\Delta t \cdot \delta e^{-\delta \ell_r}}{\Delta t \cdot 1/\ell_1} = \ell_1 \cdot \delta e^{-\delta \ell_r}$ when x is the root before proposal.

Move-Head. A reticulation edge $e = (x, y)$ is selected uniformly at random from the set of all reticulation edges. Let u_1 be the other parent of y (in addition to x) and v_1 be the child of y. The two edges (u_1, y) and (y, v_1) are deleted (along with their parameters) and replaced by a new edge $e_1 = (u_1, v_1)$ whose length is the sum of the two original lengths $\lambda_{e_1} = \lambda_{(u_1,y)} + \lambda_{(y,v_1)} = \ell_1$. Then, a new edge $e_2 = (u_2, v_2)$, with $e_2 \neq e_1$, is selected uniformly at random, deleted, and replaced by two new edges (u_2, y) and (y, v_2) whose branch lengths satisfy the conditions $\lambda_{(u_2,y)} + \lambda_{(y,v_2)} = \lambda_{(u_2,v_2)} = \ell_2$, $\lambda_{(u_2,y)} \sim \text{Uniform}(0, \ell_2)$. The length and inheritance probability of the original reticulation edge e are unchanged (and an inheritance probability of $1 - \gamma_e$ is assigned to (u_2, y)). The Hastings ratio in this case is $\frac{\Delta t \cdot 1/\ell_1}{\Delta t \cdot 1/\ell_2} = \frac{\ell_2}{\ell_1}$.

Flip-Reticulation. Let $e = (x, y)$ be the randomly selected reticulation edge. Let u_1 be the other parent of y (in addition to x) and v_1 be the child of y. Let u_2 be the parent of x and v_2 be the other child of x (in addition to y). The two edges (u_1, y) and (y, v_1) are deleted (along with their parameters) and replaced by two edges (u_1, x') and (x', v_1) under the condition that $\lambda_{(u_1,x')}+\lambda_{(x',v_1)} = \lambda_{(u_1,y)}+\lambda_{(y,v_1)} = \ell_1$, $\lambda_{(u_1,x')} \sim \text{Uniform}(0, \ell_1)$. The two edges (u_2, x) and (x, v_2) are deleted (along with their parameters) and replaced by two edges (u_2, y') and (y', v_2) under the condition that $\lambda_{(u_2,y')}+\lambda_{(y',v_2)} = \lambda_{(u_2,x)}+\lambda_{(x,v_2)} = \ell_2$, $\lambda_{(u_2,y')} \sim \text{Uniform}(0, \ell_2)$. The edge (x, y) is deleted and replaced with a new edge (x', y'). The inheritance probability of edge (u_2, y') and (x', y') are $1 - \gamma_e$ and γ_e respectively. The Hastings ratio in this case is $\frac{\Delta t \cdot 1/\ell_1 \cdot \Delta t \cdot 1/\ell_2}{\Delta t \cdot 1/\ell_2 \cdot \Delta t \cdot 1/\ell_1} = 1.0$.

Add-Reticulation. Two edges $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ are selected uniformly at random from the set of all edges in the network. Edge e_1 is replaced by two edges $e_{11} = (u_1, x_1)$ and $e_{12} = (x_1, v_1)$, where x_1 is a new node. The length

$$\lambda_{e_{11}} = \lambda_{e_1} w_1$$
$$\lambda_{e_{12}} = \lambda_{e_1}(1 - w_1)$$

where $w_1 \sim \text{Uniform}(0, 1)$. Similarly, edge e_2 is replaced by two edges $e_{21} = (u_2, x_2)$ and $e_{22} = (x_2, v_2)$, where x_2 is a new node. The length

$$\lambda_{e_{21}} = \lambda_{e_2} w_2$$
$$\lambda_{e_{22}} = \lambda_{e_2}(1 - w_2)$$

where $w_2 \sim \text{Uniform}(0, 1)$. Finally, a new edge $e_r = (x_1, x_2)$ is added with length

$$\ell_r = -\frac{1}{\delta} \ln(1 - w_3)$$

where $w_3 \sim \text{Uniform}(0, 1)$ and inheritance probability

$$\gamma_r = w_4$$

where $w_4 \sim \text{Uniform}(0, 1)$ (in which case the inheritance probability of edge (u_2, x_2) is set to $1 - \gamma_{e_r}$).

Since removing this reticulation edge does not require setting a length and inheritance probability, the Jacobian in this case involves:

- $\frac{\partial \lambda_{e_{11}}}{\partial \lambda_{e_1}} = w_1$, $\frac{\partial \lambda_{e_{11}}}{\partial w_1} = \lambda_{e_1}$, $\frac{\partial \lambda_{e_{12}}}{\partial \lambda_{e_1}} = 1 - w_1$, $\frac{\partial \lambda_{e_{12}}}{\partial w_1} = -\lambda_{e_1}$
- $\frac{\partial \lambda_{e_{21}}}{\partial \lambda_{e_2}} = w_2$, $\frac{\partial \lambda_{e_{21}}}{\partial w_2} = \lambda_{e_2}$, $\frac{\partial \lambda_{e_{22}}}{\partial \lambda_{e_2}} = 1 - w_2$, $\frac{\partial \lambda_{e_{22}}}{\partial w_2} = -\lambda_{e_2}$
- $\frac{\partial \ell_r}{\partial w_3} = \frac{1}{\delta(1-w_3)} = \frac{1}{\delta e^{-\delta \ell_r}}$.
- $\frac{\partial \gamma_r}{\partial w_4} = 1$.
- The value of the rest partial derivatives in Jacobian is zero.

Therefore, $J = \lambda_{e_1} \lambda_{e_2} \frac{1}{\delta e^{-\delta \ell_r}}$.

We now derive the Hastings ratio.

- Let re be the number of reticulation edges in the new proposed network. The probability of selecting the same edge to remove is $(1/re)$. The probability of choosing Delete-Reticulation operation from the two dimension-changing operations is $1 - \kappa_1$. (κ_1 is the defined in Materials and Methods of the main manuscript).

- To propose adding the reticulation edge, let k be the number of edges in the current network (not the newly sampled one). The probability of adding this edge is $d(1/k)(1/(k-1))(1)(1)$, where the two 1 terms correspond to the uniform density with which the length and inheritance probability of the newly added edge are chosen. If the current network has no reticulations, then $d = 1$, since it is the only one of the two dimension-changing moves that can be performed; otherwise, $d = \kappa_1$.

In summary, the product of the Hastings ratio and $|J|$ for this move is

$$\frac{(1-\kappa_1)(1/re)}{d(1/k)(1/(k-1))}\lambda_{e_1}\lambda_{e_2}\frac{1}{\delta e^{-\delta \ell_r}}.$$

Delete-Reticulation. A reticulation edge $e = (x, y)$ is selected uniformly at random from the set of all reticulation edges in the network and is removed along with its length and inheritance probability. A forced contraction is performed on nodes x and y to remove nodes of in- and out-degree 1. That is, if $e_{11} = (u_1, x)$ and $e_{12} = (x, v_1)$ are edges in the network, then both are removed and replaced by the single edge (u_1, v_1) whose length is $\lambda_{e_{11}} + \lambda_{e_{12}}$. A similar operation is applied to the two edges incoming and outgoing of y. We now derive the Hastings ratio:

- There is probability $1/re$ of selecting the reticulation edge e to remove. The probability of choosing this operation from the two dimension-changing operations is $1 - \kappa_1$.
- The probability of adding the same reticulation to add is

  $$d(1/k')(1/(k'-1))(1)(1)$$

  where $d = \kappa_1$ if the the proposed network has at least one reticulation and $d = 1$ if the proposed network has no reticulations. Here, k' is the number of edges in the proposed network (after removing a reticulation edge).

Since the proposal merges four edges into two and removes two parameters (the length and inheritance probabilities) of an edge, the Jacobian can be derived in terms of the reverse proposal:

- $\frac{\partial \lambda_{e_1}}{\partial \lambda_{e_{11}}} = \frac{1}{w_1}$, $\frac{\partial w_1}{\partial \lambda_{e_{11}}} = \frac{1}{\lambda_{e_1}}$, $\frac{\partial \lambda_{e_1}}{\partial \lambda_{e_{12}}} = \frac{1}{1-w_1}$, $\frac{\partial w_1}{\partial \lambda_{e_{12}}} = -\frac{1}{\lambda_{e_1}}$
- $\frac{\partial \lambda_{e_2}}{\partial \lambda_{e_{21}}} = \frac{1}{w_2}$, $\frac{\partial w_2}{\partial \lambda_{e_{21}}} = \frac{1}{\lambda_{e_2}}$, $\frac{\partial \lambda_{e_2}}{\partial \lambda_{e_{22}}} = \frac{1}{1-w_2}$, $\frac{\partial w_2}{\partial \lambda_{e_{22}}} = -\frac{1}{\lambda_{e_2}}$
- $\frac{\partial w_3}{\partial \ell_r} = \delta(1 - w_3) = \delta e^{-\delta \ell_r}$.
- $\frac{\partial w_4}{\partial \gamma_r} = 1$.
- The value of the rest partial derivatives in Jacobian is zero.

Therefore, $J = \frac{1}{\lambda_{e_1}\lambda_{e_2}}\delta e^{-\delta \ell_r}$.

In summary, the product of the Hastings ratio and $|J|$ for this move is

$$\frac{d(1/k')(1/(k'-1))}{(1-\kappa_1)(1/re)}\frac{1}{\lambda_{e_1}\lambda_{e_2}}\delta e^{-\delta \ell_r}.$$

1.6 Testing the MCMC sampler

To test our implementation of the sampler, we ran MCMC chains to sample from the prior distribution only; that is, we did not use any data here, so that the likelihood played no role. We ran two experiments:

- Experiment 1: We used the prior on the number of reticulations and branch lengths only.
- Experiment 2: We used the prior on the number of reticulations, branch lengths, and reticulation diameters.

For each of these two experiments, we ran the sampler for 2,020,000 iterations (first 20,000 iterations constitute the burn-in period) and collected 2,000 samples (1,000 iterations per sample) from each chain. Based on these 2,000 sampled, we plotted the distribution of the number of reticulations in the sampled networks. The results of Experiments 1 and 2 are shown in Fig. 3.

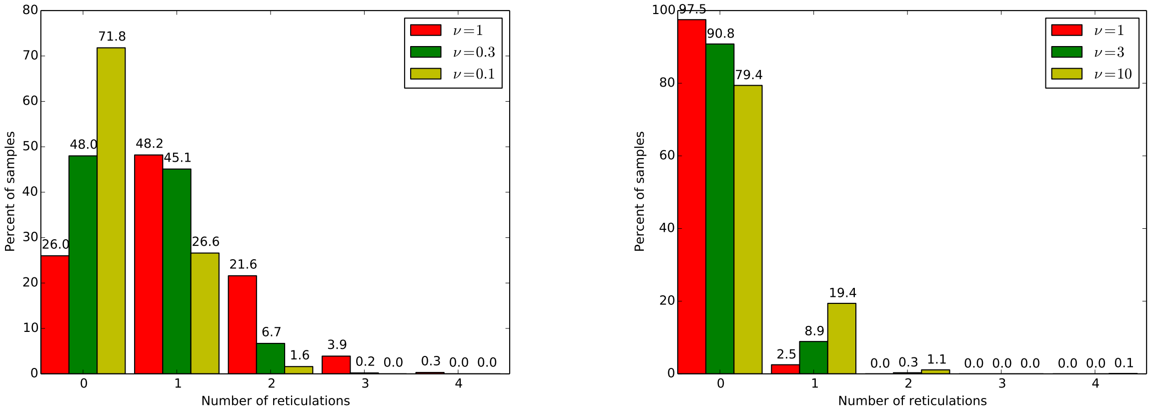


Figure 3. Distribution of the number of reticulations in networks sampled from the posterior when no data is used. Left: Priors on the number of reticulations and branch lengths are used. Right: Priors on the number of reticulations, branch lengths, and reticulation diameter are used. The different bars correspond to the three different Poisson distribution hyperparameters used.

We observe in the left panel of Fig. 3 that as the Poisson prior hyperparameter gets smaller, the number of reticulations in the sampled networks becomes smaller. This is expected since the Poisson distribution on the number of reticulations penalizes adding more reticulations more heavily when its parameter is smaller. For $\nu = 1$, the Poisson prior on 0 and 1 reticulations is equal. However, we observe that significantly more networks with 1 reticulation are sampled than networks with 0 reticulations (i.e., trees). However, in this case, the Hastings ratio plays a bigger role and favors adding the single reticulation. As the sampler go beyond 1 reticulation, the Poisson prior (and the penalty term from the number of networks) cancel out the effect of the Hastings ratio, which is why we observe a decrease again in the frequency of networks with larger numbers of reticulations.

When the prior on the reticulation diameter is added, we observe a significant bias towards trees and networks with single reticulations (the right panel of Fig. 3). The reason is that networks with more reticulations incur an added penalty based on the diameter, which reflects the prior on the reticulation diameter.

1.7 Convergence diagnostics

Mixing evaluation and convergence test are important in MCMC sampling since the samples gathered from a well mixed, converged MCMC chain are more reliable. In this work, we make use of two commonly used diagnostics:

Trace plot. A trace plot is a plot of the iterations versus the sampled value of a variable in an MCMC chain. The variable can be the posterior, the prior, or any other parameters of the distribution. A trace plot would tell us if the chain gets stuck in certain regions of the parameter space, which indicates bad mixing.

95% credible sets from multiple chains. To ensure that results are consistent among chains, we run multiple chains and maintain a 95% credible set of topologies for each chain. We then summarize the frequencies and the posterior probabilities for all topologies in the 95% credible set. Similar results across the chains is desired.

2 Simulations

2.1 Settings for the simulations

Phylogenetic networks. In order to test the performance of our method on varying number of reticulations, branch lengths and inheritance probabilities, we used three phylogenetic networks (see Fig. 4) whose topologies and branch lengths are inspired by a recent work on hybridization in mosquitos [6] and simulated data sets with varying numbers of loci from these three networks. *O* is the outgroup for rooting gene trees reconstructed from simulated sequences.

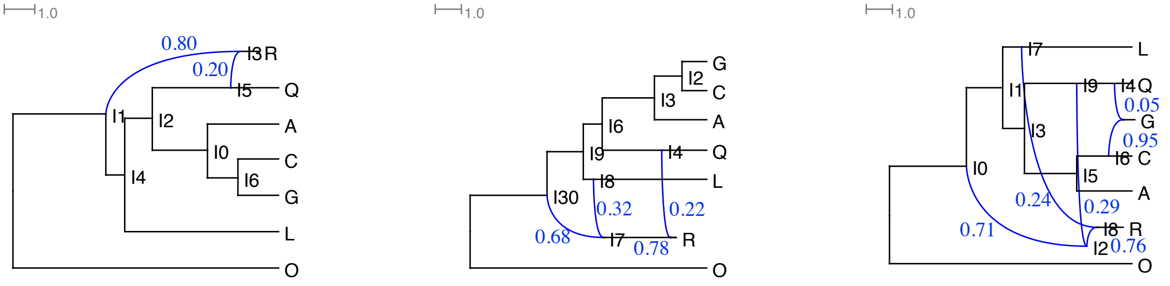


Figure 4. Three phylogenetic networks used to generate simulated data sets. Branch lengths are in coalescent units. The inheritance probabilities are marked in blue.

True gene trees. The program ms [7] was used to simulate gene trees (4 data sets with 128, 320, 800, and 2000 gene trees, respectively) within the branches of each of the phylogenetic networks. The command we used is:

```
ms 7 numLoci -T -I 7 1 1 1 1 1 1 1 -es 0.28 6 0.8 -ej 0.67 4 3 -ej 0.78 8 1 -ej 1.15 3 2 -ej 2.06 2 1 -ej 2.50
5 1 -ej 2.81 6 1 -ej 4.31 7 1
ms 7 numLoci -T -I 7 1 1 1 1 1 1 1 -es 0.10 3 0.78 -ej 0.48 5 4 -ej 0.87 8 2 -ej 0.99 6 4 -es 1.68 3 0.68 -ej
2.03 4 2 -ej 2.18 9 1 -ej 2.39 2 1 -ej 3.10 3 1 -ej 4.60 7 1
ms 7 numLoci -T -I 7 1 1 1 1 1 1 1 -es 0.21 3 0.95 -ej 0.33 4 8 -ej 0.45 5 3 -es 0.54 1 0.76 -es 0.83 1 0.29
-ej 1.06 6 3 -ej 1.06 1 8 -ej 2.10 3 8 -ej 2.15 2 9 -ej 2.52 8 9 -ej 3.23 9 10 -ej 4.73 7 10
```

Sequences. We used the simulated gene trees to simulate sequence alignments using the program Seq-gen [8] under the GTR model. The population mutation rate we used is $\theta = 0.036$. The length of sequences is 1000. The command is:

```
seq-gen -mgtr -s0.018 -f0.2112, 0.2888, 0.2896, 0.2104 -r0.2173, 0.9798, 0.2575, 0.1038, 1, 0.2070 -l1000
```

where $0.2112, 0.2888, 0.2896, 0.2104$ are the base frequencies of the nucleotides A, C, G and T, respectively, and $0.2173, 0.9798, 0.2575, 0.1038, 1, 0.2070$ are the relative rates of substitutions.

Estimated gene trees. We built 100 bootstrap gene tree topologies for each sequence alignment by RAxML8 [9] under GTR model. The command used was

```
raxmlHPC-PTHREADS -m GTRGAMMA -# 100 -o O
```

To assess the difference between the estimated bootstrap trees on each locus and the true tree for that locus, we summed the Robinson-Foulds distance [10] between each bootstrap tree and the true tree, normalized by the number of internal edges in the true tree, and divided the sum by 100. Finally, in each data set, we computed the mean and standard deviation of the normalized Robinson-Foulds distances across all loci in that data set. The results are given in Table 1.

Table 1. The mean and standard deviation of normalized Robinson-Foulds distances between the true gene tree and estimated gene trees for each locus within each simulated data set.

	128	320	800	2000
Data set 1	0.10 ± 0.13	0.11 ± 0.12	0.11 ± 0.12	0.11 ± 0.11
Data set 2	0.11 ± 0.10	0.09 ± 0.10	0.09 ± 0.10	0.09 ± 0.10
Data set 3	0.09 ± 0.11	0.08 ± 0.11	0.09 ± 0.10	0.08 ± 0.10

2.2 Experiments and Results

For each simulated data set, we performed Bayesian inference on the gene trees estimated from the sequence alignments.

MCMC settings.

- Total iterations: 5,050,000.
- Burn-in iterations: 50,000.
- Number of MCMC iterations per sample: 1,000.
- Number of samples sampled from one chain: 5,000.
- Prior: prior on the number of reticulations with Poisson parameter $\nu = 1.0$, exponential priors on branch lengths and reticulation diameters.

Table 2. Total elapsed time (hour) of MCMC chains on the simulate data sets.

	128	320	800	2000
Data set 1	2.77	3.46	4.31	5.02
Data set 2	2.64	3.17	4.78	5.74
Data set 3	2.83	5.06	6.53	9.20

Running times. All MCMC chains were run on NOTS (Night Owls Time-Sharing Service), a batch scheduled HTC cluster running on the Rice Big Research Data (BiRD) cloud infrastructure. We acquired 16 2.6GHz CPU, 1GB RAM per CPU for each task. The running times are given in Table 2.

Note that likelihood computations are the bottleneck in our Bayesian inference method. The likelihood computation times for a given data set are affected by the topology and the branch lengths of the phylogenetic network, the number of distinct gene tree topologies, and the topology of each gene tree. We calculated the mean and standard deviation of the likelihood computation across all distinct gene tree topologies given the true phylogenetic network for each data set. The results are given in Table 3. The large standard deviations clearly indicate the variability in likelihood computation times. In particular, the likelihood computation runtime increases when the number of reticulations in the phylogenetic network gets larger.

Table 3. The mean and standard deviation of the likelihood computation time (ms) for the distinct gene tree topologies within each simulated data set. The number of distinct gene tree topologies are reported in parentheses.

	128 loci	320 loci	800 loci	2000 loci
Data set 1	0.16 ± 0.37 (100)	0.15 ± 0.36 (170)	0.15 ± 0.36 (254)	0.14 ± 0.34 (320)
Data set 2	0.19 ± 0.39 (83)	0.18 ± 0.39 (142)	0.18 ± 0.39 (201)	0.18 ± 0.39 (256)
Data set 3	0.28 ± 0.45 (111)	0.28 ± 0.45 (140)	0.27 ± 0.44 (205)	0.26 ± 0.44 (269)

Results. In total, we ran 12 MCMC chains for the data sets simulated from the three phylogenetic networks and varying numbers of loci 128, 320, 800, and 2000. The trace plots are shown in Fig. 5.

Furthermore, we summarized the networks in the 95% credible sets. The results are shown in Fig. 6.

In particular, the results were as follows:

- For the true phylogenetic network with 1 reticulation node,
  - regardless of the number of loci used, all the 4 MCMC chains had a 95% credible set consisting of the true topology (1-reticulation) only (Fig. 6A).
- For the true phylogenetic network with 2 reticulation nodes,
  - on 128 and 320 loci, both MCMC chains had a 95% credible set consisting of the 1-reticulation topology only (Fig. 6A).
  - on 800 and 2000 loci, both MCMC chains had a 95% credible set consisting of the true topology (2-reticulation) only (Fig. 6B).
- For the true phylogenetic network with 3 reticulation nodes,

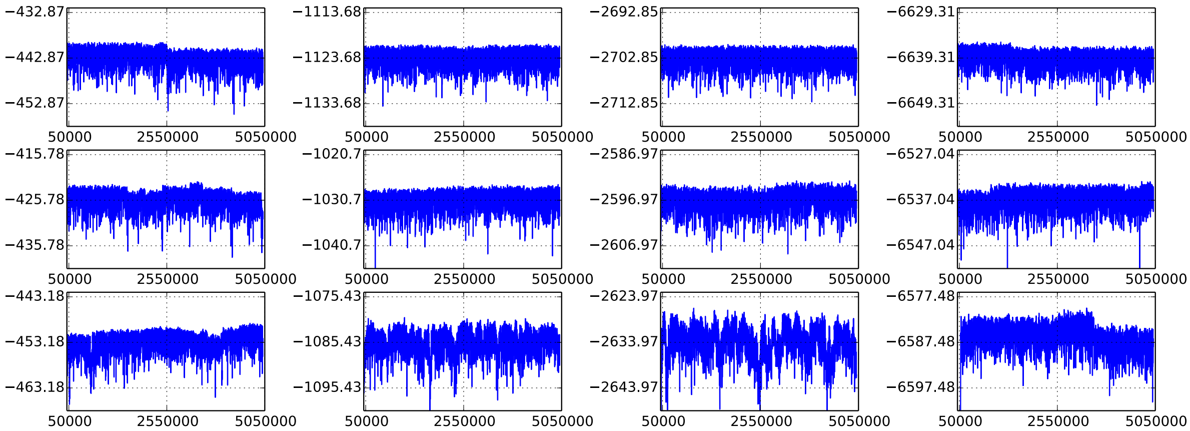


Figure 5. Trace plots of the MCMC samples from the simulated data sets. Rows from top to bottom correspond to data sets generated on the networks with 1, 2, and 3 reticulations, respectively. The columns from left to right correspond to data sets with 128, 320, 800, and 2000 gene trees, respectively. The burn-in iterations are excluded.

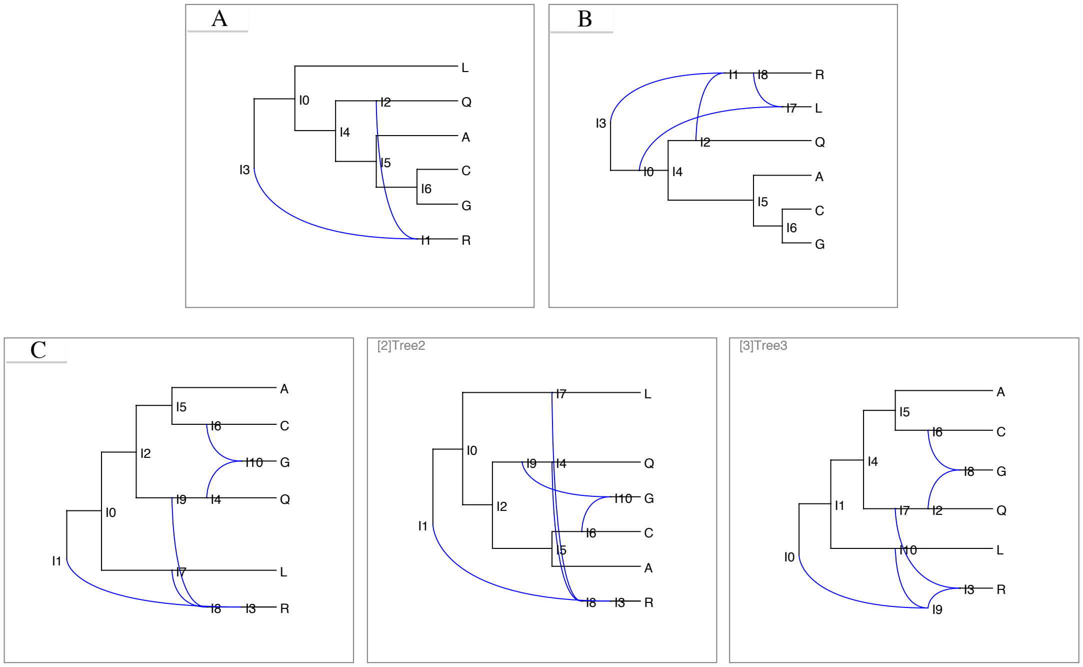


Figure 6. The phylogenetic network topologies in the 95% credible sets of the results using the simulated data.

– on 128 loci, the MCMC chain had a 95% credible set consisting of the 1-reticulation topology only (Fig. 6A).

– on 320 loci, the MCMC chain had a 95% credible set consisting of the 2-reticulation topology only (Fig. 6B).

- on 800 and 2000 loci, the MCMC chain had a 95% credible set consisting of three 3-reticulation topologies (Fig. 6C). Note that these topologies are indistinguishable using the gene tree topologies and employing our likelihood formulation, and thus can be viewed as equivalent to the true topology.

To summarize, the method shows very good performance on these simulated data sets.

3 Analysis of a bread wheat (*Triticum aestivum*) data set

Marcussen *et al.* [11] investigated ancient hybridization among the ancestral genomes of bread wheat by performing parsimonious inference of hybridization in the presence of ILS [12] implemented in PhyloNet [13]. 2269 gene trees were constructed from three subgenomes of wheat TaA (*T. aestivum* A subgenome), TaB (*T. aestivum* B subgenome), TaD (*T. aestivum* D subgenome). Using this data set, they inferred a species phylogeny, shown in Fig. 7. We reanalyzed this data set using our newly developed method.

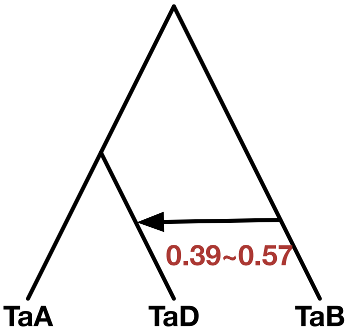


Figure 7. The species phylogeny reported in [11] and inferred from a data set of 2269 gene trees using parsimonious inference of hybridization in the presence of ILS implemented in PhyloNet [13].

3.1 Data preprocessing

We downloaded the sequence alignments of 2269 genes from Dryad Digital Repository (doi:10.5061/dryad.f6c34). Each alignment is composed of genes from TaA, TaB, TaD and three outgroups Bd (*Brachypodium distachyon*, Os (*Oryza sativa*) and Hv (*Hordeum vulgare*).

We built 100 bootstrap gene tree topologies for each alignment using RAxML8 [9] under GTR model. The command is

```
raxmlHPC-PTHREADS -m GTRGAMMA -# 100 -o Outgroup
```

3.2 MCMC settings

- Total iterations: 5,050,000.
- Burn-in iterations: 50,000.
- Number of MCMC iterations per sample: 1,000.
- Number of samples in one chain: 5,000.

- Prior: prior on the number of reticulations (Poisson parameter $\nu = 1.0$), prior on branch lengths, prior on reticulation diameter.

3.3 Runtime.

The elapsed time for this analysis is 2.20 hr.

3.4 Results

The three topologies composing the 95% credible set are shown in Fig. 8.

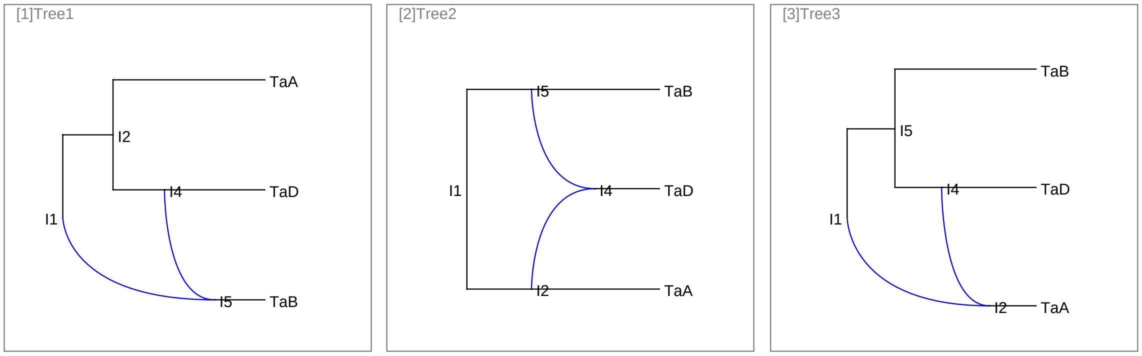


Figure 8. The three topologies composing the 95% credible set for the wheat data set.

The trace plot is displayed in Fig. 9.

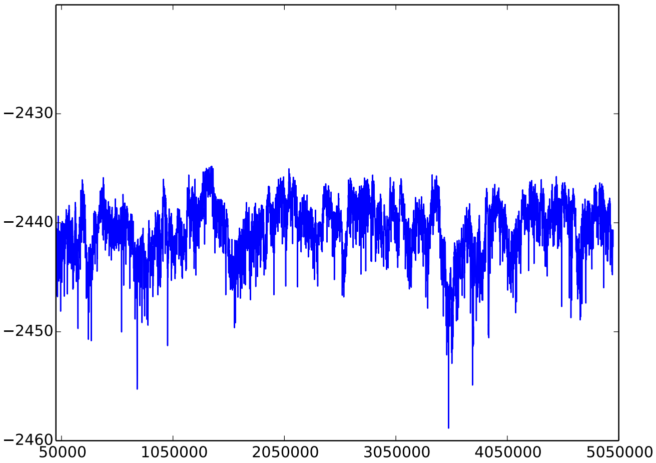


Figure 9. Trace plot of the MCMC samples from the bread wheat data set. The burn-in iterations are excluded.

3.5 Gene trees with branch lengths

Fig. 10 shows data on pairwise distances inferred from across all loci for all pairs of taxa. Since the smallest pairwise distance per pair of taxa is a upper bound on the speciation time of these taxa (according to the likelihood formulation used here), it is clear that using gene tree branch lengths in the inference step would bias the inferred network and produce erroneous results. A similar observation was observation was made in [14].

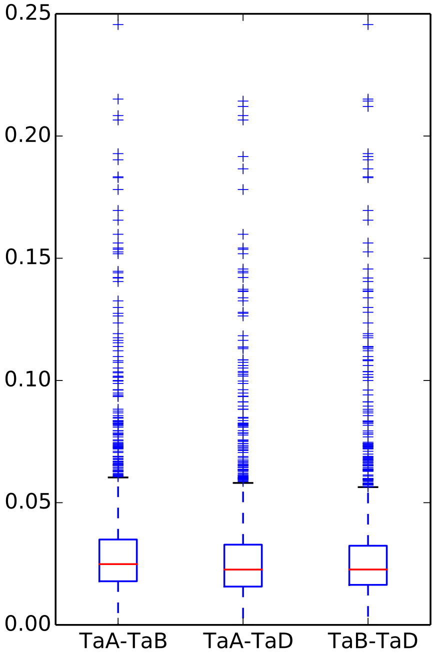


Figure 10. Whisker-box plot of all pairwise distances for each pair of species in the wheat data set. Gene trees and their branch lengths were estimated using maximum likelihood in PAUP*. A pairwise distance between two leaves in a gene tree is the sum of branch lengths on the path between the two leaves in the tree. The points on the x-axis correspond to all possible pairs of taxa, and the various pairwise distances per pair come from the loci that were used in the inference.

4 Analysis of Anopheles mosquitoes (*An. gambiae* complex) data set

Fontaine et al. recently reported on hybridization and extensive introgression in the *An. gambiae* complex [6]. The authors constructed a phylogenetic network by adding three major reticulation edges to the species tree constructed from X chromosome. They used gene tree analysis to detect the location of the reticulation edges. More recently, Wen *et al.* [15] applied systematic inference of phylogenetic networks to the data using the maximum likelihood method of [3]. They provided an a new view on the evolutionary history of the species. We reanalyzed the same data used in [15].

4.1 Data preprocessing

We downloaded the MAF genome alignment from high depth field samples from Dryad (doi:10.5061/dryad.f4114). The species we included in our analysis are *An. gambiae* (G), *An. coluzzii* (C), *An. arabiensis* (A), *An. quadriannulatus* (Q), *An. merus* (R) and *An. melas* (L). *An. christyi* serves as the outgroup for gene tree reconstruction and rooting. For each chromosome (2L, 2R, 3L 3R, X), we randomly sampled genome alignment chunks from the original dataset. The alignment chunks are at lease 64 kb far away from each other to minimize the likelihood of dependence among loci.

The total number of alignments we sampled for each chromosome is

2L	2R	3L	3R	X	all
669	849	564	709	228	3019

We built 100 bootstrap trees (topology only) for each alignment using RAxML8 [9] under GTR model. The command we used is

```
raxmlHPC-PTHREADS -m GTRGAMMA -# 100 -o Outgroup
```

4.2 MCMC settings for inference from the sex chromosome and autosomes

We separate the inference from the sex chromosome and autosomes since the effective population sizes differ between sex chromosomes and autosomes.

The settings for the MCMC chain used in inference is as follows.

- Total iterations: 5,050,000
- Burn-in iterations: 50,000
- Number of MCMC iterations per sample: 1,000
- Number of samples: 5,000
- Prior: prior on the number of reticulations, prior on branch lengths, prior on diameters of hybridizations

For the X chromosome, we used Poisson parameter $\nu = 1.0$. For the autosome data, we tried three different values, $\nu = 0.1, 1.0, 10.0$, in order to test the effect of the prior.

4.3 Results from the X chromosome data set

Only one topology exists in the 95% credible sets (Fig. 11).

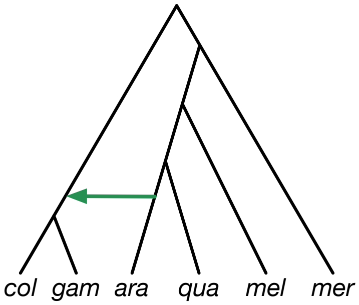


Figure 11. The single topology in the 95% credible set from the mosquito X chromosome data set.

The trace plot on the posterior distribution of the MCMC chain is displayed in Fig. 12. The elapsed time is 7.65 hr.

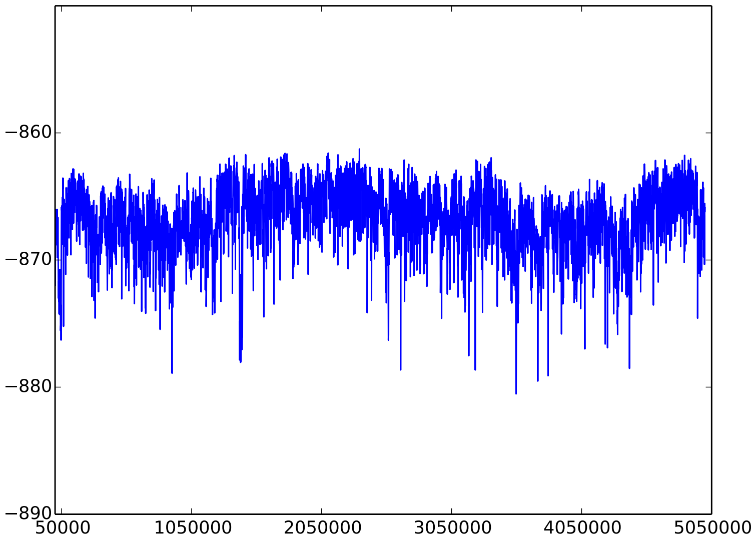


Figure 12. Trace plot of the MCMC samples from the mosquito X chromosome data set. The burn-in iterations are excluded.

4.4 Results from the autosome data set

Aside from reanalyzing the autosome dataset, we also studies the effect of priors on different data sizes. We sampled 311 and 931 loci from the full data set (2791 loci), and then performed Bayesian inference on different priors (Poisson parameter $\nu = 0.1, 1, 10$):

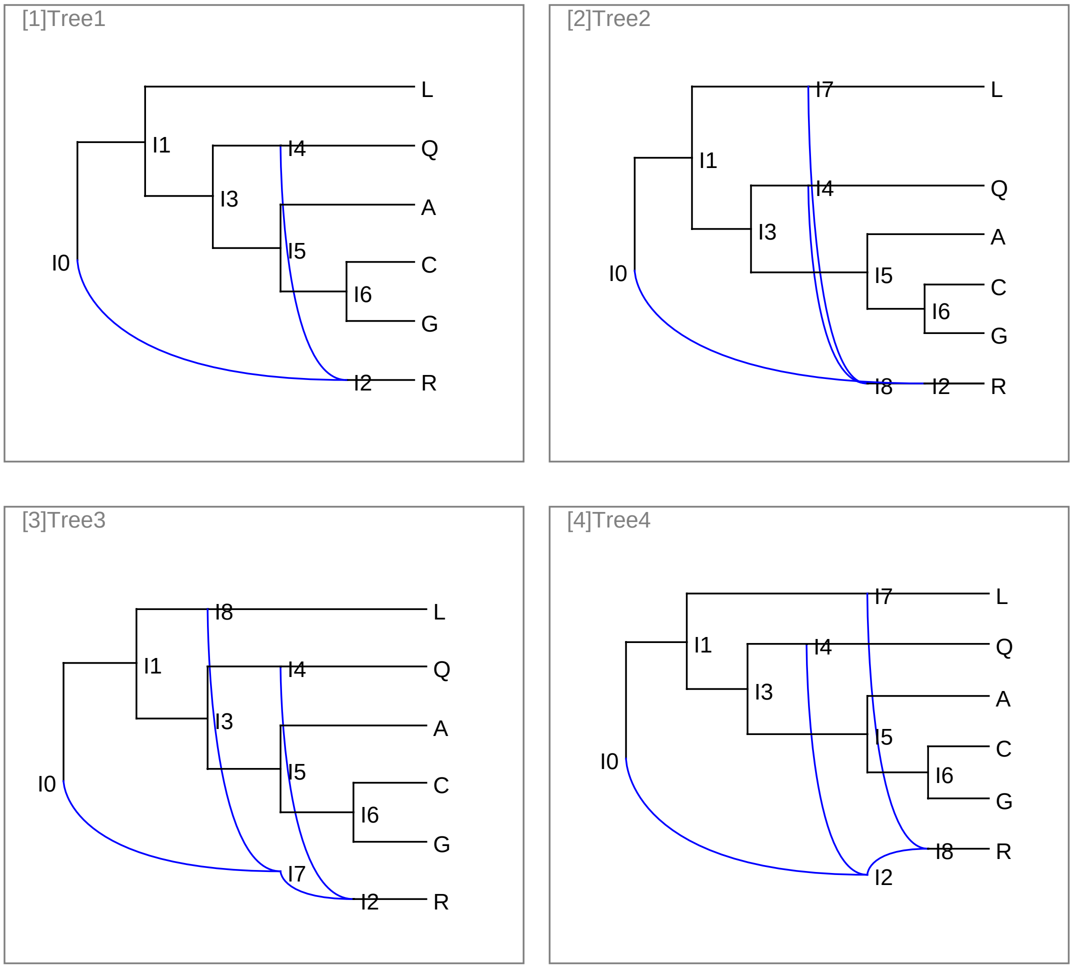
- On 311 loci,
  - for $\nu = 0.1$, only one topology exists in the 95% credible set (Fig. 13[1]), which is a network with one reticulation node.
  - for $\nu = 1.0$, the 95% credible set contains 4 topologies (Fig. 13[1]-[4]). The proportions of the 1-reticulation and 2-reticulation topologies are 75.3% and 22.5%, respectively.
  - for $\nu = 10$, the 95% credible set contains 4 topologies (Fig. 13[1]-[4]). The proportions of the 1-reticulation and 2-reticulation topologies are 27.7% and 63.2%, respectively.

  By increasing ν, the proportion of 2-reticulation topologies are increased.
- On 931 loci, varying ν did not affect the results—the 95% credible set contains 3 indistinguishable, 2-reticulation networks (Fig. 13[2]-[4]).
- On 2,791, varying ν did not affect the results—the 95% credible set contains 3 indistinguishable, 3-reticulation networks (Fig. 14).

These results demonstrate that when the data size is small, the prior could play an important role in the posterior distribution, and when the data size gets larger, varying the prior (the Poisson parameter in our case) could hardly affect the results. This further attests to the good performance of our method.

The trace plot for the case of $\nu = 1.0$ and 2791 loci is shown in Fig. 15.

The runtimes are given in Table 4. We can see that when the data size and the expected number of reticulations (reflected by ν) are smaller, the sampling space is almost within the 1-reticulation network


[1]Tree1

[2]Tree2

[3]Tree3

[4]Tree4

Figure 13. The topologies in the 95% credible sets sampled from the 311 and 931 autosome loci. Note that the topologies [2]-[4] are indistinguishable using gene tree topologies under the likelihood function used here.

space and the likelihood computation is much faster. When the data size and the expected number of reticulations gets larger, the likelihood computation is more time demanding, thus the total running time is longer.

Table 4. Total elapsed time (hour) of MCMC chains of the mosquito data analyses.

	$\nu = 0.1$	$\nu = 1$	$\nu = 10$
311 loci	9.02	8.45	12.02
931 loci	12.48	13.56	15.50
2791 loci	23.71	24.15	29.78

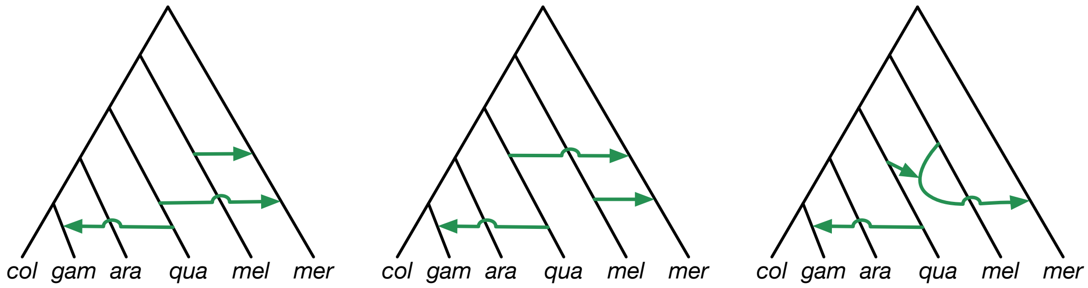


Figure 14. The topologies in the 95% credible sets sampled from the full autosome data set. Note that the topologies are indistinguishable using gene tree topologies under the likelihood function used here.

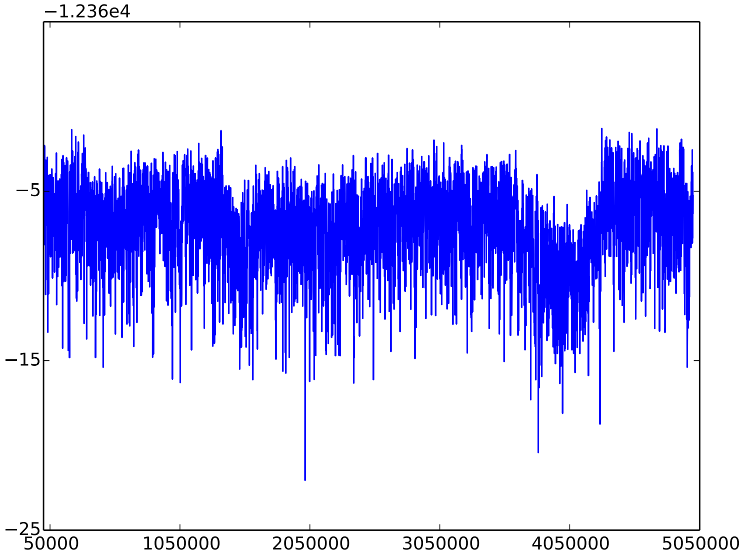


Figure 15. Trace plot of the MCMC samples from the mosquito autosome data set. The number of loci is 2791 and the Poisson hyperparameter value is $\nu = 1.0$. The burn-in iterations are excluded.

4.5 Gene trees with branch lengths

Fig. 16 shows data on pairwise distances inferred from across all loci for all pairs of taxa. Results are similar to those we observed above in the wheat data set and would have similar implications on inference from gene trees with branch lengths.

5 Analysis of a house mouse (*Mus musculus*) data set

The house mouse data set is composed of individuals sampled from five populations: *M. m. domesticus* from France (DF), *M. m. domesticus* from Germany (DG), *M. m. musculus* from the Czech Republic (MZ), *M. m. musculus* from Kazakhstan (MK), and *M. m. musculus* from China (MC). To satisfy the assumption of free recombination between loci, local phylogenies were sampled at 100 kb intervals. Local

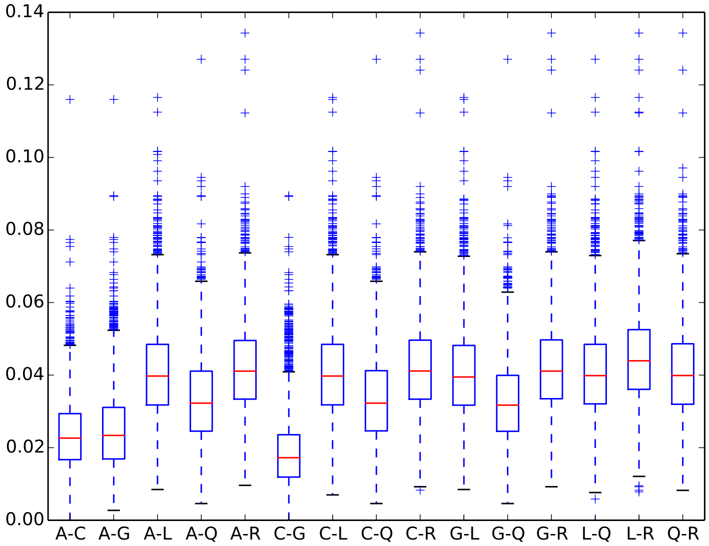


Figure 16. Whisker-box plot of all pairwise distances for each pair of species in the mosquito data set. Gene trees and their branch lengths were estimated using maximum likelihood in PAUP*. A pairwise distance between two leaves in a gene tree is the sum of branch lengths on the path between the two leaves in the tree. The points on the x-axis correspond to all possible pairs of taxa, and the various pairwise distances per pair come from the loci that were used in the inference.

phylogenies were rooted using *R. norvegicus* as an outgroup. In total, 20,639 local phylogenies were reconstructed.

Yu *et al.* [3] investigated the house mouse dataset using maximum likelihood and reported two main hybridization events. We reanalyzed the house mouse data set using our new Bayesian inference method.

5.1 Data Preprocessing

In the preliminary analysis, we found several 4-reticulation networks with hybridization near the root (between the two branches emanating from the root), which indicates poor signal in the data. This is not surprising, given that this data set differs from the other two in that it consists of individuals of the same species, rather than different species.

Since for each locus we inferred 100 bootstrap trees, we computed the majority-rule consensus of the 100 trees for each locus, and we analyzed the number of loci that have gene trees with 0-3 internal branches to assess the signal in the data. Gene trees with 0 internal branches are star phylogenies with no signal at all. Gene trees with 3 internal branches are fully resolved.

We found that 11,457 (55.5%) loci have gene trees with 3 internal branches. Among the 11,457 gene trees built on these loci, 98 distinct topologies were present. Note that for 5 taxa, there are only 105 possible gene tree topologies. In other words, for almost every fully-resolved gene tree there are some loci that support it. The distribution of the number of fully-resolved gene trees that are supported by different numbers of loci is given in

The distribution shows that three distinct gene tree topologies are supported by over 2000 loci each. On the other end of the distribution, 45 distinct gene trees are supported by between 1 and 9 loci only.

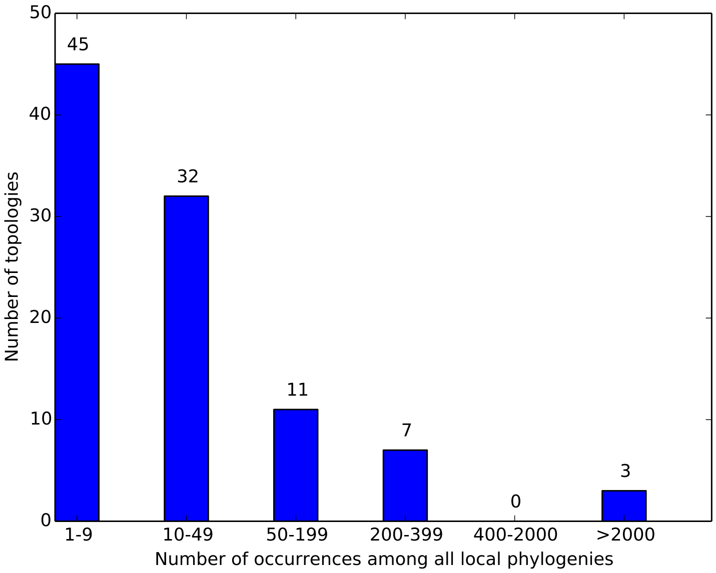


Figure 17. Distribution of topologies on the number of occurrences.

To capture the main hybridization events, we performed the inference on the fully resolved 21 gene tree topologies that are supported by at least 50 loci each. This reduced the size of the data set from 20,639 gene trees to 10,575 gene trees.

5.2 Settings

We employed Metropolis-coupled MCMC (MC^3) (described in [16]) to help the sampler traverse the posterior landscape.

The settings for MCMC inference are

- Total iterations: 4,050,000
- Burn-in iterations: 50,000
- Number of MCMC iterations per sample: 1,000
- Number of samples: 4,000
- Prior Parameter: Poisson parameter $\nu = 1.0$)
- (MC^3) Number of chains: 3 (1 cold chain, 2 heated chains)
- (MC^3) Temperature settings: 1 (for cold chain), 2, 4
- (MC^3) Swap frequency: swap two random chains once every 100 iteration
- (MC^3) Starting tree for each chain: a random tree from the input phylogenies

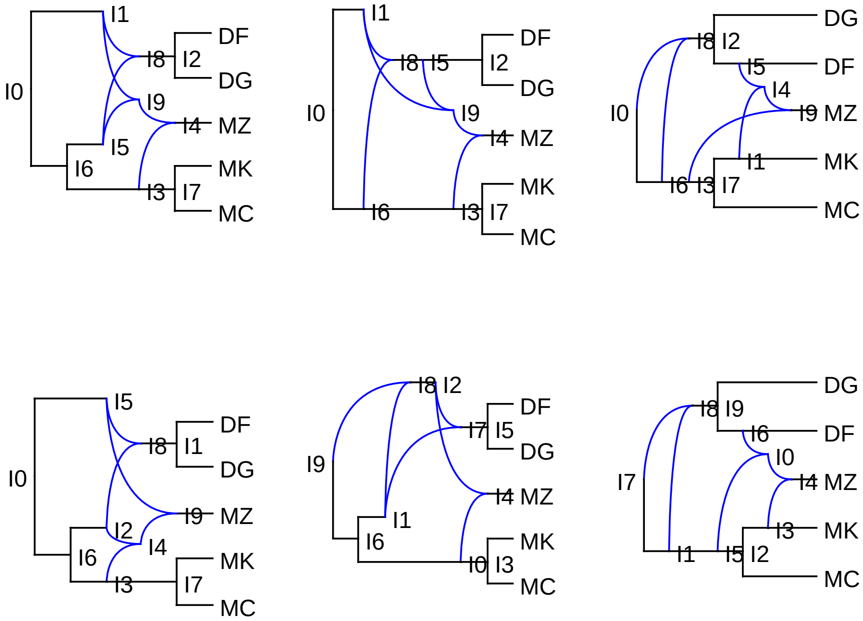


Figure 18. The top six topologies sampled in the 95% credible set from mouse data set.

5.3 Results

The topologies in the 95% credible set are shown in Fig. 18. Unlike the cases of the wheat and mosquito data sets, where multiple indistinguishable networks were sampled, several distinguishable (population) networks with similar posterior values were sampled from the mouse dataset. This issue emphasizes the applicability of the method to population data and, at the same time, the complexity in the inferred evolutionary history in this case due to extensive gene flow.

The trace plot is shown in Fig. 19.

The elapsed time is 44.65 hr, which is longer than the other datasets because we ran two heated chains along with the cold chain.

5.4 Gene trees with branch lengths

Fig. 20 shows data on pairwise distances inferred from across all loci for all pairs of taxa. Results are similar to those we observed above in the wheat data set and would have similar implications on inference from gene trees with branch lengths. In this case, however, the data consists of multiple individuals of the same species. That is, this is a data set of very low divergence levels. This is clearly obvious from the pairwise distances, and further highlights the complexity of analyzing such a data set, despite the applicability of the method.

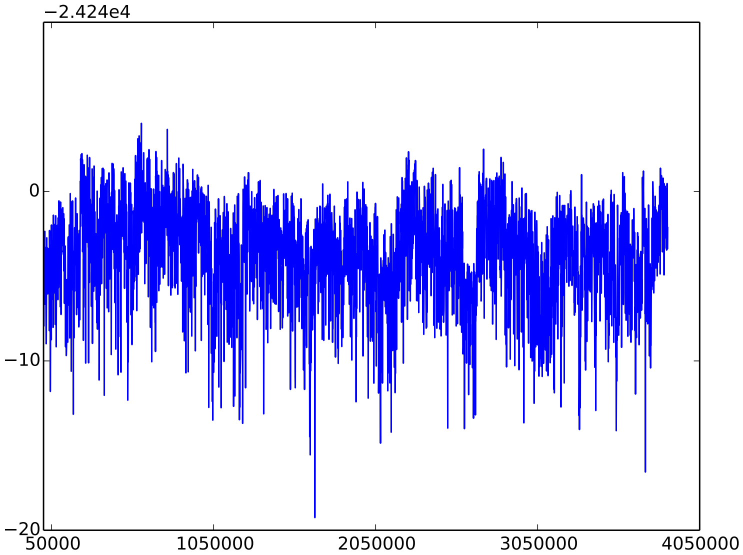


Figure 19. Trace plot of the MCMC samples from the mouse data set. The burn-in iterations are excluded.

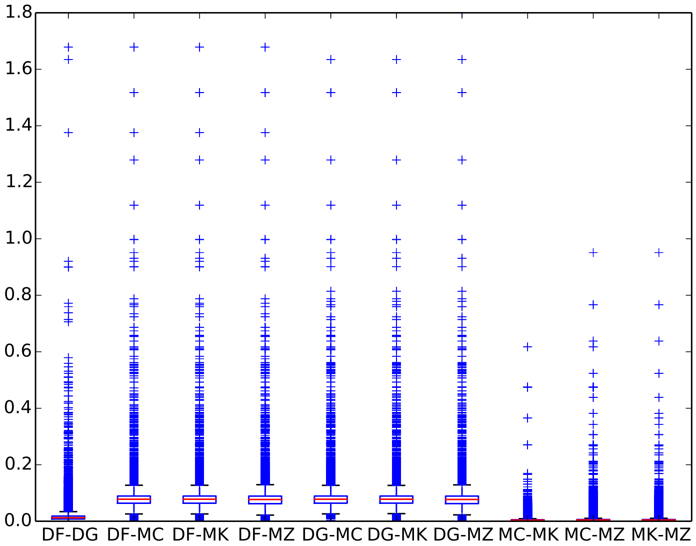


Figure 20. Whisker-box plot of all pairwise distances for each pair of species in the mouse data set. Gene trees and their branch lengths were estimated using maximum likelihood in PAUP*. A pairwise distance between two leaves in a gene tree is the sum of branch lengths on the path between the two leaves in the tree. The points on the x-axis correspond to all possible pairs of taxa, and the various pairwise distances per pair come from the loci that were used in the inference.

References

1. Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS genetics 8: e1002660.

2. Yu Y, Ristic N, Nakhleh L (2013) Fast algorithms and heuristics for phylogenomics under ILS and hybridization. BMC Bioinformatics 14: S6.

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

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

5. Lewis PO, Holder MT, Holsinger KE (2005) Polytomies and Bayesian phylogenetic inference. Systematic Biology 54: 241–253.

6. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, et al. (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347: 1258524.

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

8. Rambaut A, Grassly NC (1997) Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comp Appl Biosci 13: 235-238.

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

10. Robinson D, Foulds L (1981) Comparison of phylogenetic trees. Math Biosci 53: 131–147.

11. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, et al. (2014) Ancient hybridizations among the ancestral genomes of bread wheat. Science 345: 1250092.

12. Yu Y, Barnett RM, Nakhleh L (2013) Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Systematic Biology 62: 738-751.

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

14. DeGiorgio M, Degnan JH (2014) Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Systematic biology 63: 66–82.

15. Wen D, Yu Y, Hahn MW, Nakhleh L (2016) Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Molecular Ecology .

16. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20: 407–415.