Syst. Biol. 67(3):439–457, 2018

DOI:10.1093/sysbio/syx085
Advance Access publication October 27, 2017

Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data

DINGQIAO WEN[1] AND LUAY NAKHLEH[1,2,*]

[1]*Department of Computer Science and* [2]*Department of BioSciences, Rice University, 6100 Main Street, Houston, TX 77005, USA*
**Correspondence to be sent to: Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA; E-mail: nakhleh@rice.edu.*

Received 24 April 2017; reviews returned 17 October 2017; accepted 24 October 2017
Associate Editor: Laura Kubatko

Abstract.—The multispecies network coalescent (MSNC) is a stochastic process that captures how gene trees grow within the branches of a phylogenetic network. Coupling the MSNC with a stochastic mutational process that operates along the branches of the gene trees gives rise to a generative model of how multiple loci from within and across species evolve in the presence of both incomplete lineage sorting (ILS) and reticulation (e.g., hybridization). We report on a Bayesian method for sampling the parameters of this generative model, including the species phylogeny, gene trees, divergence times, and population sizes, from DNA sequences of multiple independent loci. We demonstrate the utility of our method by analyzing simulated data and reanalyzing an empirical data set. Our results demonstrate the significance of not only coestimating species phylogenies and gene trees, but also accounting for reticulation and ILS simultaneously. In particular, we show that when gene flow occurs, our method accurately estimates the evolutionary histories, coalescence times, and divergence times. Tree inference methods, on the other hand, underestimate divergence times and overestimate coalescence times when the evolutionary history is reticulate. While the MSNC corresponds to an abstract model of "intermixture," we study the performance of the model and method on simulated data generated under a gene flow model. We show that the method accurately infers the most recent time at which gene flow occurs. Finally, we demonstrate the application of the new method to a 106-locus yeast data set. [Bayesian inference; incomplete lineage sorting; multispecies network coalescent; phylogenetic network; reticulation; RJMCMC.]

The availability of sequence data from multiple loci across the genomes of species and individuals within species is enabling accurate estimates of gene and species evolutionary histories, as well as parameters such as divergence times and ancestral population sizes (Rannala and Yang 2003). Several statistical methods have been developed for obtaining such estimates (Rannala and Yang 2003; Edwards et al. 2007; Heled and Drummond 2010; Bouckaert et al. 2014). All these methods employ the *multispecies coalescent* (Degnan and Rosenberg 2009) as the stochastic process that captures the relationship between species trees and gene genealogies.

As evidence of hybridization (admixture between different populations of the same species or across different species) continues to accumulate (Arnold 1997; Rieseberg 1997; Barton 2001; Koonin et al. 2001; Gogarten et al. 2002; Mallet 2005, 2007), there is a pressing need for statistical methods that infer species phylogenies, gene trees, and their associated parameters in the presence of hybridization. We recently introduced for this purpose the *multispecies network coalescent* (MSNC) along with a maximum likelihood search heuristic (Yu et al. 2014) and a Bayesian sampling technique (Wen et al. 2016a). However, these methods use gene tree estimates as input. Using these estimates, instead of using the sequence data directly, has at least three drawbacks. First, the sequence data allows for learning more about the model than gene tree estimates (Rannala and Yang 2003). Second, gene tree estimates could well include erroneous information, resulting in wrong inferences (DeGiorgio and Degnan 2014; Wen et al. 2016a). Third, coestimating the species phylogeny and gene trees results in better estimates of the gene trees themselves (Bayzid and Warnow 2013; DeGiorgio and Degnan 2014).

We report here on a Bayesian method for coestimating species (or, population) phylogenies and gene trees along with parameters such as ancestral population sizes and divergence times using DNA sequence alignments from multiple independent loci. Our method utilizes a two-step generative process (Fig. 1) that links, via latent variables that correspond to local gene genealogies, the sequences of multiple, unlinked loci from across a set of genomes to the phylogenetic network (Nakhleh 2010a) that models the evolution of the genomes themselves.

Our method consists of a reversible-jump Markov chain Monte Carlo (RJMCMC) sampler of the posterior distribution of this generative process. In particular, our method coestimates, in the form of posterior distribution samples, the phylogenetic network and its associated parameters for the genomes as well as the local genealogies for the individual loci. We demonstrate the performance of our method on simulated data. Furthermore, we analyze an empirical data set, and discuss the insights afforded by our method. In particular, we find that methods that do not account, wrongly, for admixture in the data tend to underestimate divergence times of the species or populations and overestimate the coalescent times of individual gene genealogies. Our method, on the other hand, estimates both the divergence times and coalescent times with high accuracy. Furthermore, we demonstrate that coalescent times are much more accurately estimated when the estimation is done simultaneously with the phylogenetic network than when the estimation is done in isolation.

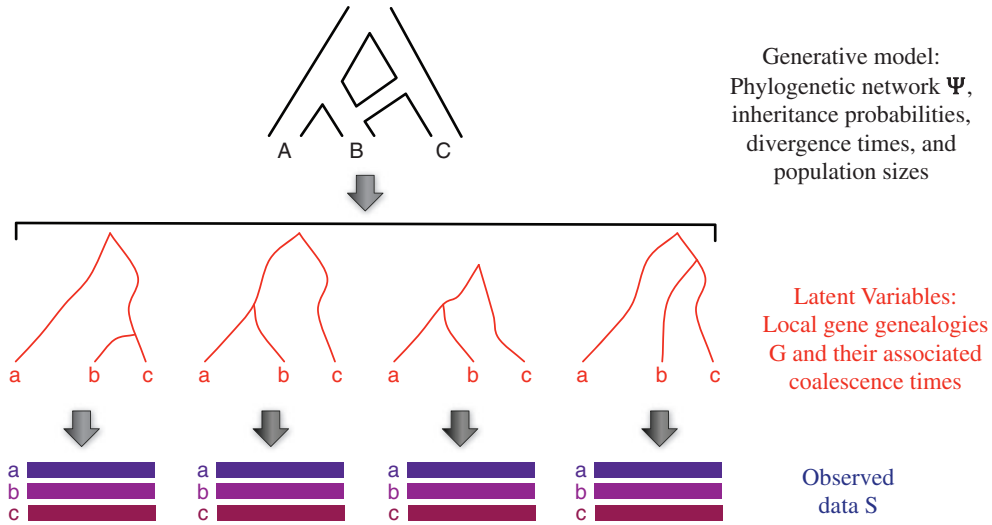

FIGURE 1. From a phylogenetic network to multilocus sequences via latent gene genealogies. The multispecies network coalescent (Yu et al. 2014) is a stochastic process that defines a probability distribution on gene genealogies along with their coalescent times. The parameters of the process consist of a phylogenetic network topology, inheritance probabilities, divergence times, and population sizes. Each gene genealogy, when coupled with model of sequence evolution, defines a probability distribution on sequence alignments.

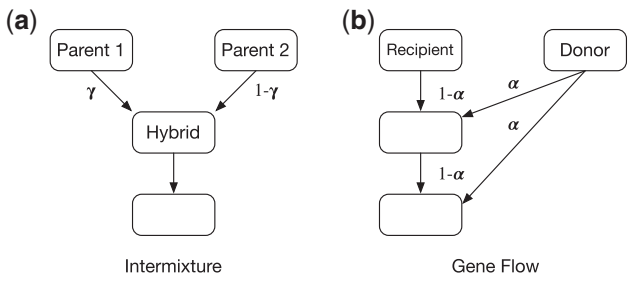

FIGURE 2. Two admixture models for a hybrid population (Long 1991). a) The hybrid population is formed by a single intermixture event between two parental populations, where γ is the inheritance probability measuring the proportion of the parental populations. b) The hybrid population (recipient) receives gene flow from a donor population, where α is the migration rate.

An important contribution of this manuscript is also to study the performance of the MSNC on data generated under gene flow scenarios. In particular, the population genetics community has developed models of reticulate evolution (i.e., admixture) at the population level. An important question is: How do phylogenetic network methods perform on data generated under such scenarios? To answer this question, it is important to highlight the difference in abstraction employed in the MSNC model as opposed to a gene flow model. It turns out that this difference was well articulated in (Long 1991), where two models of admixture were presented: the intermixture model and the gene flow model (Fig. 2). The MSNC employs the intermixture model, whereas the population genetics community mostly uses the gene flow model (Slatkin and Maddison 1989; Whitlock and Mccauley 1999; Hey and Nielsen 2004; Hey and Nielsen 2007; Strasburg and Rieseberg 2010; Gronau et al. 2011; Leaché et al. 2013). Note that the intermixture model also underlies the admixture graph model of Reich et al. (2009) and Pickrell and Pritchard (2012) where γ is the admixture proportion. In the admixture graph model, the branch lengths correspond to genetic drift values that measure variation in allele frequency corresponding to random sampling of alleles from generation to generation in a finite-size population.

Hudson's ms program (Hudson 2002) allows for generating data under each of the two admixture models—intermixture and gene flow. In this article, we generate data under both models and study the performance of inference under the MSNC in both cases.

For an empirical data set, we analyzed the yeast data set of Rokas et al. (2003), which consists of 106 loci from seven Saccharomyces species, and contrasted our results to those obtained from the method of Wen et al. (2016a) on gene tree estimates.

Finally, as the model underlying our method extends the multispecies coalescent to cases that include admixture, our method is applicable to data from different subpopulations, not only different species, and to data where more than one individual per species or subpopulation is sampled. The method is implemented and publicly available in the PhyloNet software package (Than et al. 2008). We point out that after the initial submission of this manuscript, another method for coestimating gene trees and phylogenetic networks was developed (Zhang et al. 2017). The method has similarity to ours in that it employs the MSNC model and uses the sequence data directly (as opposed to gene tree estimates). However, the MCMC sampler differs from the one we present here.

METHODS

Phylogenetic Networks and Their Parameters

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

- indeg(s)=0 and outdeg(s)=1 (s is a special node, that is the parent of the root node, r);
- indeg(r)=1 and outdeg(r)=2 (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).

The network's edges, $E(\Psi)\subseteq V\times V$, consist of *reticulation edges*, whose heads are reticulation nodes, *tree edges*, whose heads are tree nodes, and special edge $(s,r)\in E$. Furthermore, $\ell: V_L\rightarrow\mathscr{X}$ is the *leaf-labeling* function, which is a bijection from V_L to $\mathscr{X}$. Each node in $V(\Psi)$ has a species divergence time parameter and each edge in $E(\Psi)$ has an associated population size parameter. The edge $er(\Psi)=(s,r)$ is infinite in length so that all lineages that enter it coalesce on it eventually. Finally, for every pair of reticulation edges e_1 and e_2 that share the same reticulation node, we associate an inheritance probability, γ, such that $\gamma_{e_1},\gamma_{e_2}\in[0,1]$ with $\gamma_{e_1}+\gamma_{e_2}=1$. We denote by Γ the vector of inheritance probabilities corresponding to all the reticulation nodes in the phylogenetic network (for each reticulation node, Γ has the value for one of the two incoming edges only).

Given a phylogenetic network Ψ, we use the following notation:

- Ψ_{top}: The leaf-labeled topology of Ψ; that is, the pair (V,E) along with the leaf-labeling ℓ.
- Ψ_{ret}: The number of reticulation nodes in Ψ. $\Psi_{ret}=0$ when Ψ is a phylogenetic tree.
- Ψ_τ: The species divergence time parameters of Ψ. $\Psi_\tau\in(\mathbb{R}^+)^{|V(\Psi)|}$.
- Ψ_θ: The population size parameters of Ψ. $\Psi_\theta\in(\mathbb{R}^+)^{|E(\Psi)|}$. For each branch, the corresponding θ equals $4N_e\mu$, where N_e is the effective population size for that branch, and μ is the mutation rate per site per generation.

We use Ψ to refer to the topology, species divergence times and population size parameters of the phylogenetic network.

It is often the case that divergence times associated with nodes in the phylogenetic network are measured in units of years, generations, or coalescent units. On the other hand, branch lengths in gene trees are often in units of expected number of mutations per site. We convert estimates back and forth between units as follows:

- Given divergence time in units of expected number of mutations per site τ, mutation rate per site per generation μ and the number of generations per year g, $\tau/\mu g$ represents divergence times in units of years.
- Given population size parameter in units of population mutation rate per site θ, $2\tau/\theta$ represents divergence times in coalescent units.

Bayesian Formulation and Inference

The data in our case is a set $\mathscr{S}=\{S_1,\ldots,S_m\}$ where S_i is a DNA sequence alignment from locus i (the bottom part in Fig. 1). A major assumption is that there is no recombination within any of the m loci, yet there is free recombination between loci. The model $\mathscr{M}$ consists of a phylogenetic network Ψ (the topology, divergence times, and population sizes) and a vector of inheritance probabilities Γ (the top part in Fig. 1).

The posterior distribution of the model is given by

$$\begin{aligned} p(\mathscr{M}|\mathscr{S}) &\propto p(\mathscr{S}|\mathscr{M})p(\mathscr{M}) \\ &= p(\mathscr{M})\prod_{i=1}^{m}\int_G p(S_i|g)p(g|\mathscr{M})dg, \end{aligned} \tag{1}$$

where the integration is taken over all possible gene trees (the middle part in Fig. 1). The term $p(S_i|g)$ gives the gene tree likelihood, which is computed using Felsenstein's algorithm (Felsenstein 1981) assuming a model of sequence evolution, and $p(g|\mathscr{M})$ is the probability density function for the gene trees, which was derived for the cases of species tree and species network in Rannala and Yang (2003) and Yu et al. (2014), respectively.

The integration in Eq. (1) is computationally infeasible except for very small data sets. Furthermore, in many analyses, the gene trees for the individual loci are themselves a quantity of interest. Therefore, to obtain gene trees, we sample from the posterior distribution as given by

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

where $G=(g_1,\ldots,g_m)$ is a vector of gene trees, one for each of the m loci. This coestimation approach is adopted by the two popular Bayesian methods *BEAST (Heled and Drummond 2010) and BEST (Liu 2008), both of which coestimate species trees (hybridization is not accounted for) and gene trees.

The Likelihood Function

Felsenstein (1981) introduced a pruning algorithm that efficiently calculates the likelihood of gene tree g and

DNA evolution model parameters Φ as

$$p(S|g,\Phi)=\prod_{i=1}^{l}p(s_i|g,\Phi),$$

where s_i is i-th site in S, l is the sequence length, and

$$p(s_i|g,\Phi)=p(s_i|g_{\text{top}},g_\tau,\pi,q,\mu).$$

Here, g_{top} is the tree topology, g_τ is the divergence times of the gene tree, $\pi=\{\pi_A,\pi_T,\pi_C,\pi_G\}$ is a vector of equilibrium frequencies of the four nucleotides, $q=\{q_{AT},q_{AC},q_{AG},q_{TC},q_{TG},q_{CG}\}$ is a vector of substitution rates between pairs of nucleotides, and μ is the mutation rate. Over a branch j whose length (in expected number of mutations per site) is t_j, the transition probability is calculated as $e^{\mu q t_j}$. In the implementation, we use the BEAGLE library (Ayres et al. 2011) for more efficient implementation of Felsenstein's algorithm.

Yu et al. (2012, 2013a, 2014) fully derived the mass and density functions of gene trees under the multispecies network coalescence, where the lengths of a phylogenetic network's branches are given in coalescent units. Here, we derive the probability density function (pdf) of gene trees for a phylogenetic network given by its topology, divergence/migration times and population size parameters following (Rannala and Yang 2003; Yu et al. 2014). Coalescence times in the (sampled) gene trees posit temporal constraints on the divergence and migration times of the phylogenetic network.

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

- if $(x,y)\in E(\Psi)$ and $\tau_\Psi(x)>\tau'_g(v)\geq\tau_\Psi(y)$, then $h(v)=(x,y)$.
- if r is the root of Ψ and $\tau'_g(v)\geq\tau_\Psi(r)$, then $h(v)=er(\Psi)$.

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 $H_\Psi(g)$.

Given a phylogenetic network Ψ, the pdf of the gene tree random variable is given by

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

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

Consider gene tree g for locus j 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

$$\begin{aligned}p(h|\Psi,\Gamma)&=\prod_{b\in E(\Psi)}\left[\prod_{i=1}^{|T_b(h)|-1}\frac{2}{\theta_b}e^{-\left(\frac{2}{\theta_b}\right)\binom{u_b(h)-i+1}{2}(T_b(h)_{i+1}-T_b(h)_i)}\right]\\&\quad\times e^{-\left(\frac{2}{\theta_b}\right)\binom{v_b(h)}{2}(\tau_\Psi(x_b)-T_b(h)_{|T_b(h)|})}\times\Gamma_b^{u_b(h)},\end{aligned} \tag{4}$$

where x_b is the source node of edge b, $\theta_b=4N_b\mu$ and N_b is the population size corresponding to branch b, μ is the mutation rate per-site per-generation, and Γ_b is the inheritance probability associated with branch b.

Prior Distributions

We extended the prior of phylogenetic network composed of topology and branch lengths in Wen et al. (2016a) to phylogenetic networks composed of topology, divergence times and population sizes, as given by Eq. (5),

$$\begin{aligned}p(\Psi|\nu,\delta,\eta,\psi)&=p(\Psi_{ret}|\nu)\times p(\Psi_d|\Psi_{\text{top}},\Psi_\tau,\eta)\\&\quad\times p(\Psi_\tau|\delta)\times p(\Psi_\theta|\psi)\end{aligned} \tag{5}$$

where $p(\Psi_{\text{ret}}|\nu)$, the prior on the number of reticulation nodes, and $p(\Psi_d|\Psi_{\text{top}},\Psi_\tau,\eta)$, the prior on the diameters of reticulation nodes, were defined in Wen et al. (2016a).

It is important to note here that if Ψ_{top} does not follow the phylogenetic network definition, then $p(\Psi|\nu,\delta,\eta,\psi)=0$. This is crucial 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.

Rannala and Yang used independent Gamma distributions for time intervals (branch lengths) instead of divergence times. However, in the absence of any information on the number of edges of the species network as well as the time intervals, it is computationally intensive to infer the hyperparameters of independent Gamma distributions. Currently, we use a uniform distribution (as in BEST, Liu 2008).

We assume one population size per edge, including the edge above the root. Population size parameters are Gamma distributed, $\theta_b\sim\Gamma(2,\psi)$, with a mean 2ψ and a shape parameter of 2. In the absence of any information on the population size, we use the noninformative prior $P_\psi(x)=1/x$ for hyperparameter ψ (Heled and Drummond 2010). The number of elements in θ is $|E(\Psi)|+1$. To simplify inference, our implementation also supports a constant population size across all branches, in which case θ contains only one element.

For the prior on the inheritance probabilities, we use $\Gamma_b \sim \text{Beta}(\alpha, \beta)$. Unless there is some specific knowledge on the inheritance probabilities, a uniform prior on $[0,1]$ is adopted by setting $\alpha = \beta = 1$. If the amount of introgressed genomic data is suspected to be small in the genome, the hyperparameters α and β can be appropriately set to bias the inheritance probabilities to values close to 0 and 1 (a U-shaped distribution).

The RJMCMC Sampler

As computing the posterior distribution given by Eq. (2) is computationally intractable, we implement a Markov chain Monte Carlo (MCMC) sampling procedure based on the Metropolis–Hastings algorithm. In each iteration of the sampling, a new state (Ψ', Γ', G') is proposed and either accepted or rejected based on the Metropolis–Hastings ratio r that is composed of the likelihood, prior, and Hastings ratios. When the proposal changes the dimensionality of the sample by adding a new reticulation to or removing an existing reticulation from the phylogenetic network, the absolute value of the determinant of the Jacobian matrix is also taken into account, which results in a reversible-jump MCMC, or RJMCMC (Green 1995, 2003).

Our sampling algorithm employs three categories of moves: One for sampling the phylogenetic network and its parameters (divergence times and population mutation rates), one for sampling the inheritance probabilities, and one for sampling the gene trees (topologies and coalescence times). To propose a new state of the Markov chain, one element from $(\Psi, \gamma_1, \ldots, \gamma_{\Psi_{\text{ret}}}, g_1, \ldots, g_m)$ is selected at random, then a move from the corresponding category is applied. The workflow, design and full derivation of the Hastings ratios of the moves are given in Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.3h185.

We implemented our method in PhyloNet (Than et al. 2008), a publicly available, open-source software package for phylogenetic network inference and analysis.

Results

*Our Method and *BEAST Perform Similarly in Cases of No Reticulation*

∗BEAST (Heled and Drummond 2010) is the most commonly used software tool for Bayesian inference of species trees from multilocus data. In our first experiment, we set out to study how our method performs compared to this well-established software tool on simulated data whose evolutionary history is treelike. To accomplish this task, we used the phylogenetic tree shown in Fig. 3 as the model species phylogeny. Using the program ms (Hudson 2002), we simulated 20 data sets each consisting of 10 conditionally independent gene trees with the command

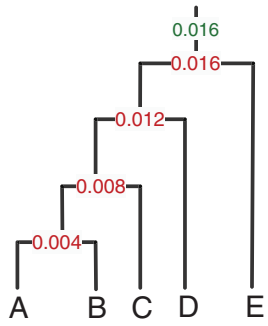

FIGURE 3. A model species tree used to generate multilocus data sets. The divergence times in units of expected number of mutations per site are shown at the internal nodes of the tree, and the population size parameter in units of population mutation rate per site is shown at the branch above the root (0.016). The population mutation rate was assumed to be constant across all branches of the tree.

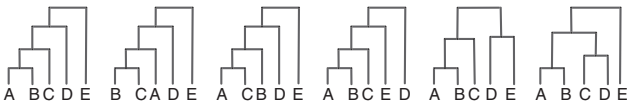

FIGURE 4. The trees that constitute the 95% credible set of each of our method and ∗BEAST. The proportions of these trees from left to right as sampled by our method were 77.7%, 5.7%, 5.0%, 3.0%, 3.0%, and 2.8%, respectively, and as sampled by ∗BEAST were 70.7%, 6.0%, 6.7%, 4.7%, 4.5%, and 3.6%, respectively.

```
ms 5 10 -T -I 5 1 1 1 1 1 -ej 0.25 3 2 -ej 0.5 4 2 -ej 0.75 5 2 -ej
1.0 2 1
```

We then used the program Seq-gen (Rambaut and Grassly 1997) to simulate the evolution of 1000-site sequences under the Jukes-Cantor model of evolution (Jukes and Cantor 1969) with the command

```
seq-gen -m HKY -l 1000 -s 0.008
```

For each of the 20 10-locus data sets, we ran two MCMC chains, each with 5×10^5 iterations and 5×10^4 burn-in, using our method as well as ∗BEAST. One sample was collected from every 500 iterations, resulting in 900 collected samples per data set and a total of 18,000 collected samples from all 20 data sets. In comparing the two tools, we used all 18,000 collected samples to evaluate the estimates obtained for the various parameters of interest: population size parameter, divergence times, and the topology of the inferred species phylogeny.

Both our method and ∗BEAST inferred exactly the same 95% credible set, which consists of the six topologies shown in Fig. 4. Our method sampled the true phylogeny with higher frequency than ∗BEAST.

Figure 5 shows histograms of the estimates obtained for the divergence times at each node of the maximum *a posteriori* (MAP) species tree estimate of our method (based on $0.777 \times 18{,}000$ observations) and ∗BEAST (based on $0.707 \times 18{,}000$ observations), which was identical in both cases to the true species tree. The histograms of both methods are very similar. In fact, the histograms obtained by our method have peaks that

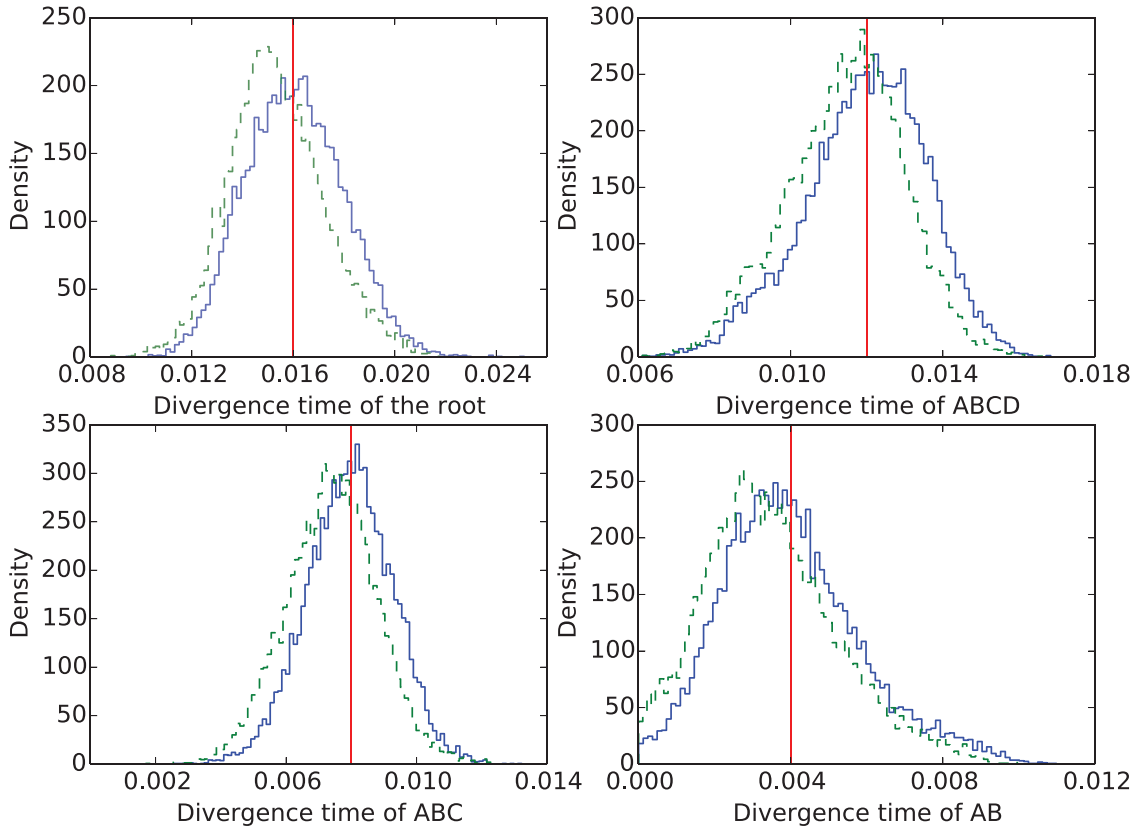

FIGURE 5. Histograms of divergence times of each node of the true species phylogeny as estimated by our method (solid line) and ∗BEAST (dashed line). The vertical lines indicate the true divergence times.

are closer to the true divergence time values than those obtained by ∗BEAST.

Figure 6 shows the histograms of the population mutation rate (one value across all branches of the species tree was assumed) estimated by the two methods. As in the case of divergence time estimates, the two methods obtain similar results in the case of population mutation rate estimates. However, we observe here a histogram of our method with a single peak around the true value, whereas we observe a bimodal histogram obtained by ∗BEAST.

All the results reported above were obtained by running the code on Night Owls Time-Sharing Service (NOTS), which is a batch scheduled High-Throughput Computing (HTC) cluster. We used two cores, with two threads per core running at 2.6GHz, and 1G RAM per thread. The runtime for ∗BEAST is around 28 ± 1 s for each data set, while our method takes longer time: 185 ± 7 s per data set. This can be explained by the fact that ∗BEAST has been under continued development for several years now, while our implementation hardly has any optimization components yet.

When we ran ∗BEAST on multilocus sequence data simulated under species phylogenies with reticulations, we found that ∗BEAST overestimated the coalescence times in individual loci and underestimated the divergence times of the species phylogeny. We report these results in Supplementary Materials available on Dryad as ∗BEAST is not intended for evolutionary analyses with gene flow. Furthermore, there are existing, extensive studies on the impact of gene flow on the inference of species trees (Leaché et al. 2013; Solís-Lemus et al. 2016).

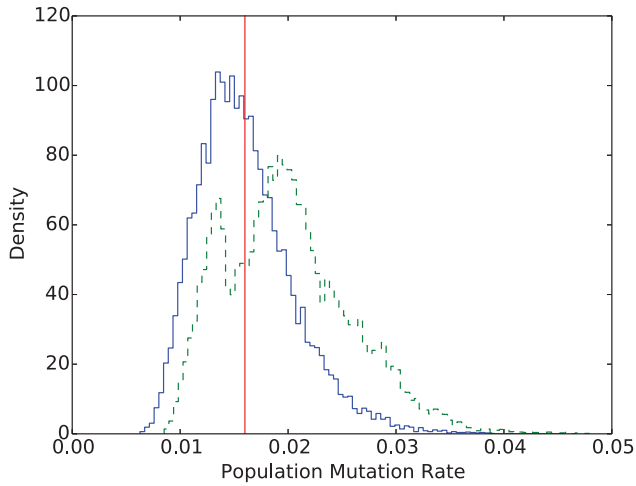

FIGURE 6. Population mutation rate estimated by our method (solid line) and ∗BEAST (dashed line). The vertical line indicates the true population mutation rate.

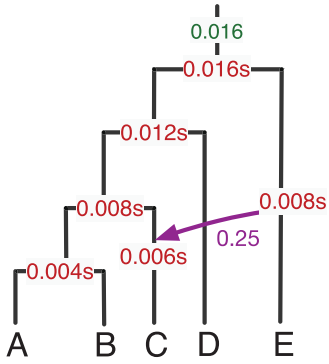

FIGURE 7. A model phylogenetic network used to generate simulated data. The divergence times in units of expected number of mutations per site are shown at the internal nodes of the network. The population size parameter in units of population mutation rate per site (0.016) is shown at the branch above the root. The inheritance probability, shown on the arrow, is 0.25. Parameter s is used to scale the divergence times.

Our Method Provides Accurate Estimates of the Network and Its Associated Parameters

We used the phylogenetic network shown in Fig. 7 as the model species phylogeny. The scale parameter of the divergence times s was varied to take on values in the set $\{0.1, 0.25, 0.5, 1.0\}$. Setting $s=0.1$ results in very short branches and, consequently, the hardest data sets on which to estimate parameters. Setting $s=1.0$ results in longer branches and higher signal for a more accurate estimate of the parameter values. It is important to note that the topology, reticulation event, divergence times (with $s=1.0$) and population size are inspired by the species phylogeny recovered from the Anopheles mosquitoes data set (Fontaine et al. 2015; Wen et al. 2016b).

For the four settings of s values, 0.1, 0.25, 0.5, and 1.0, we used the program ms (Hudson 2002) to simulate 20 data sets each with 128 gene trees of conditionally independent loci with the four following commands respectively:

- `ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.025 4 3 -es 0.0375 1 0.3 -ej 0.05 6 3 -ej 0.05 2 1 -ej 0.075 5 3 -ej 0.1 3 1`
- `ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.0625 4 3 -es 0.09375 1 0.3 -ej 0.125 6 3 -ej 0.125 2 1 -ej 0.1875 5 3 -ej 0.25 3 1`
- `ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.125 4 3 -es 0.1875 1 0.3 -ej 0.25 6 3 -ej 0.25 2 1 -ej 0.375 5 3 -ej 0.5 3 1`
- `ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.25 4 3 -es 0.375 1 0.3 -ej 0.5 6 3 -ej 0.5 2 1 -ej 0.75 5 3 -ej 1.0 3 1`

The program Seq-gen (Rambaut and Grassly 1997) was used to generate sequence alignments down the gene trees under the Jukes Cantor model (Jukes and Cantor 1969) with lengths *seqLen* in $\{250, 500, 1000\}$ using the command

seq-gen -m HKY -l *seqLen* -s 0.008

To vary the number of loci used in the inference, we produced data sets with 32, 64, and 128 loci by sampling loci without replacement from the full data set of 128 loci. Each of these sequence data sets was then used as input to the inference method.

To assess the signal in the sequence data sets we obtained, we quantified the percentage of variable sites for each setting, averaged over all 20 replicates for that setting. The percentages of variable sites in the generated alignments for $s=0.1, 0.25, 0.5, 1.0$ (varying the sequence length had negligible effect for the same scaling factor s) are $\sim 0.039 \pm 0.02$, $\sim 0.048 \pm 0.02$, $\sim 0.061 \pm 0.02$, and $\sim 0.088 \pm 0.02$, respectively.

For each data set, we ran an MCMC chain of 8×10^6 iterations with 1×10^6 burn-in. One sample was collected from every 5000 iterations, resulting in a total of 1400 collected samples. We summarized the results based on 28,000 samples from 20 replicates for each of the 36 simulation settings (four values of s, three sequence lengths, and three numbers of loci). In Figs. 8 and 9 below, the five bars from bottom to top correspond to the minimum, first-, second-, third-quantile, and the maximum, respectively, from the 20 replicates for each setting. In Figs. 11, 12, 13, 15, and 16, the error bars correspond to standard deviations calculated from the 20 replicates for each setting.

In assessing the performance of our method, we evaluated the estimates obtained for the various parameters of interest: divergence times, population mutation rates, the number of reticulations, and the topology of the inferred species phylogeny. Figure 8 shows the estimates obtained for the divergence time at the root of the network. Three observations are in order. First, for any combination of sequence length and scaling parameter value, the divergence time estimate converges to the true value as the number of loci increases. Second, for any combination of number of loci and scaling parameter value, the divergence time estimate converges to the true value as the sequence length increases. Third, the estimates are relatively poor only under the extreme settings of scaling parameter value 0.1 and sequence length 250. In this case, the signal in the sequence data is too weak to obtain good estimates. However, it is worth noting that even under this setting, using 128 loci produces a very accurate estimate of the divergence time.

Figure 9 shows the estimates obtained for the population mutation rate parameter (one value across all branches of the species network was assumed). The results show very similar trends to those obtained for the divergence time estimates, with the main difference being that the estimates now are very accurate even for the hardest of cases: $s=0.1$ and sequence length 250, regardless of the number of loci used.

The results are quite different when it comes to estimating the number of reticulations and the topology of the phylogenetic network itself. Figure 10 shows the estimates of the number of reticulations under different settings. As the figure clearly shows, under the case of extremely short branches ($s=0.1$), the method recovers a tree (not necessarily the same tree in all cases); that is, it estimates the number of reticulations to be 0,

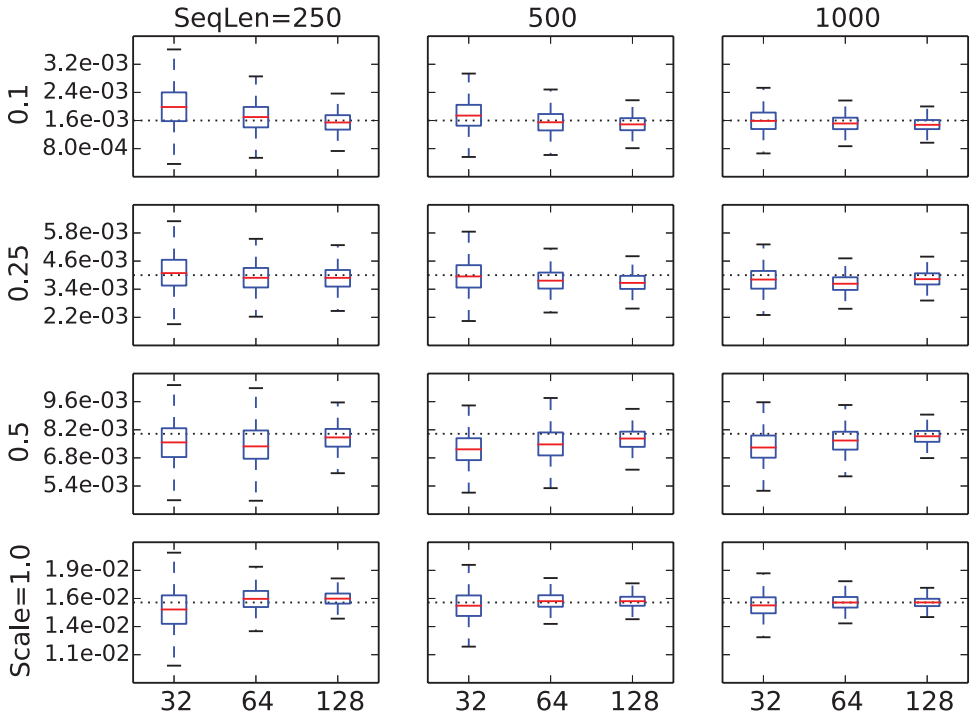

FIGURE 8. Divergence time estimates at the root under different values of the scaling parameter s (different rows), sequence lengths (different columns), and numbers of loci (three values within each panel). The dashed horizontal lines indicate the true value in the model network.

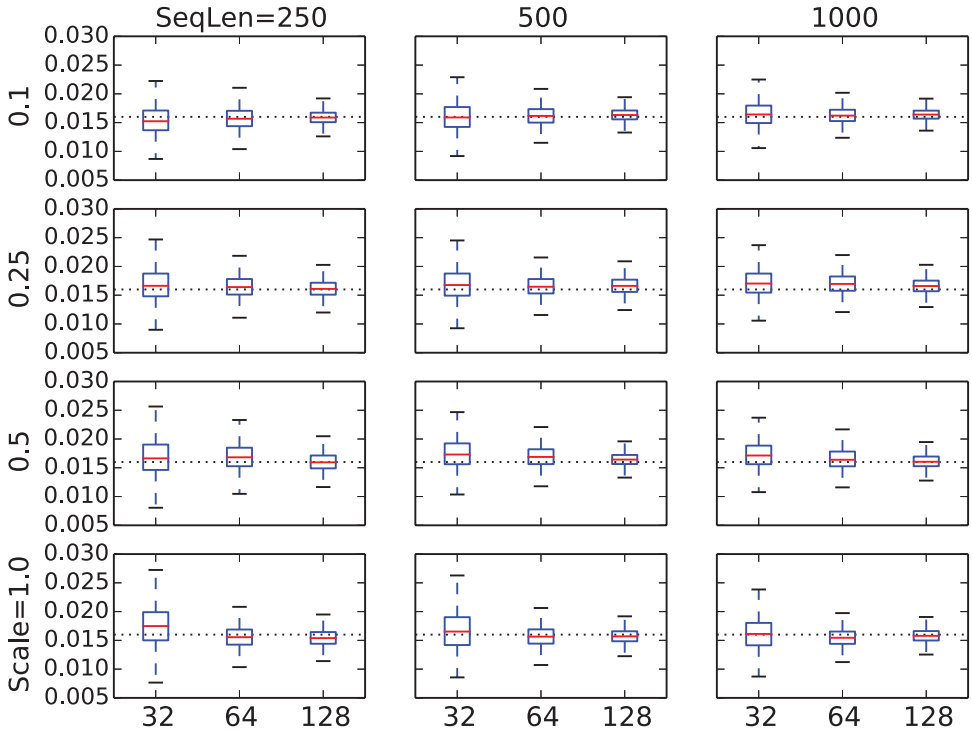

FIGURE 9. Population mutation rate estimates under different values of the scaling parameter s (different rows), sequence lengths (different columns), and numbers of loci (three values within each panel). The dashed horizontal lines indicate the true value in the model network.

regardless of the number of loci or sequence length used. Here, the signal is too weak to recover any reticulation. In the case of slightly longer branches ($s=0.25$), the estimate of the number of reticulations becomes slightly more accurate when the sequences are long and 128 loci are used. Given the observed trend, the method could recover the true number of reticulations if a thousand or so loci are used. However, it is important

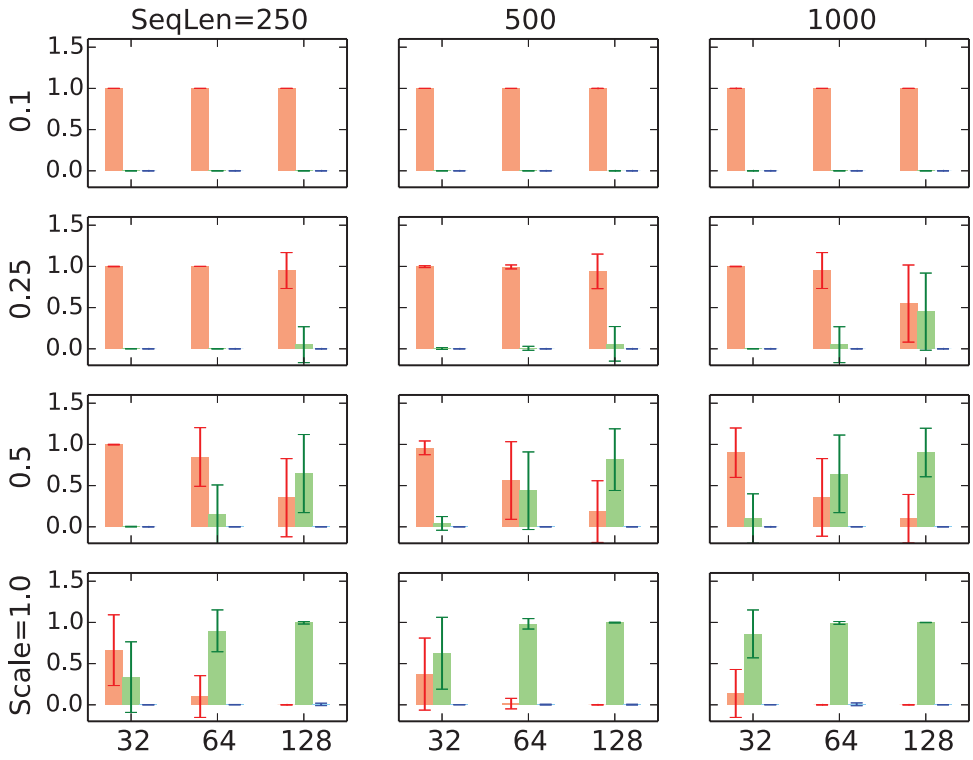

FIGURE 10. Proportions of different types of networks inferred under different simulation conditions. For each number of loci in each panel, the three bars from left to right correspond to proportions of inferred trees, inferred 1-reticulation networks, and inferred 2-reticulations networks, respectively. The model network has a single reticulation.

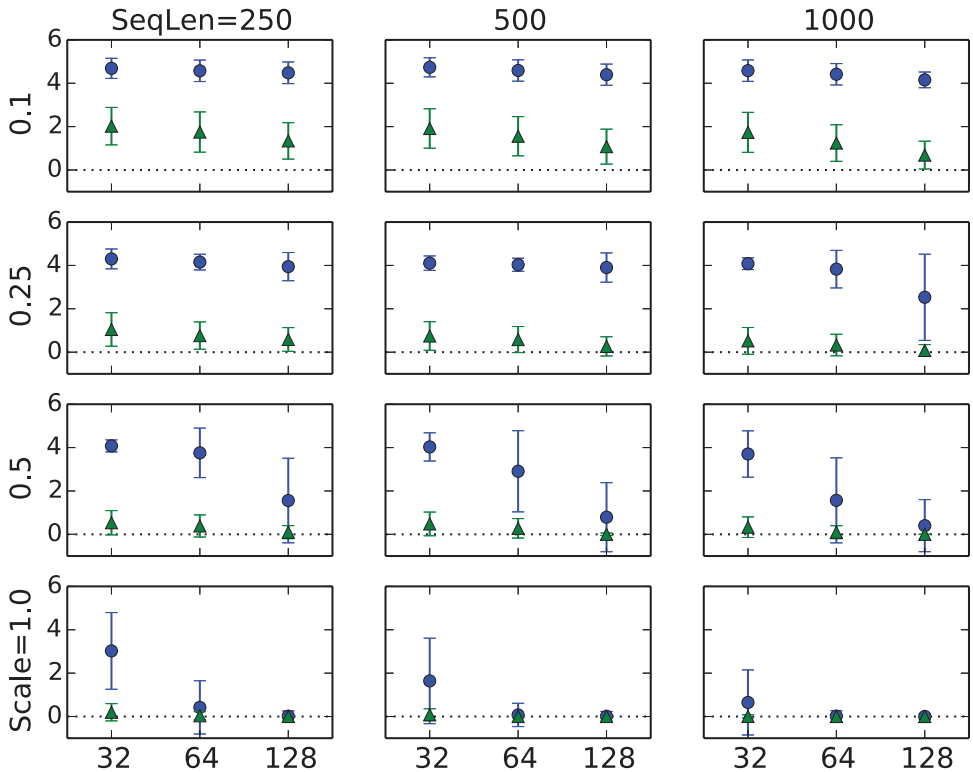

FIGURE 11. The topological difference between the true and inferred networks (circles) and the Robinson-Foulds distance between the inferred tree (if a network is inferred, this case is not included) and the backbone tree of the true network (triangles).

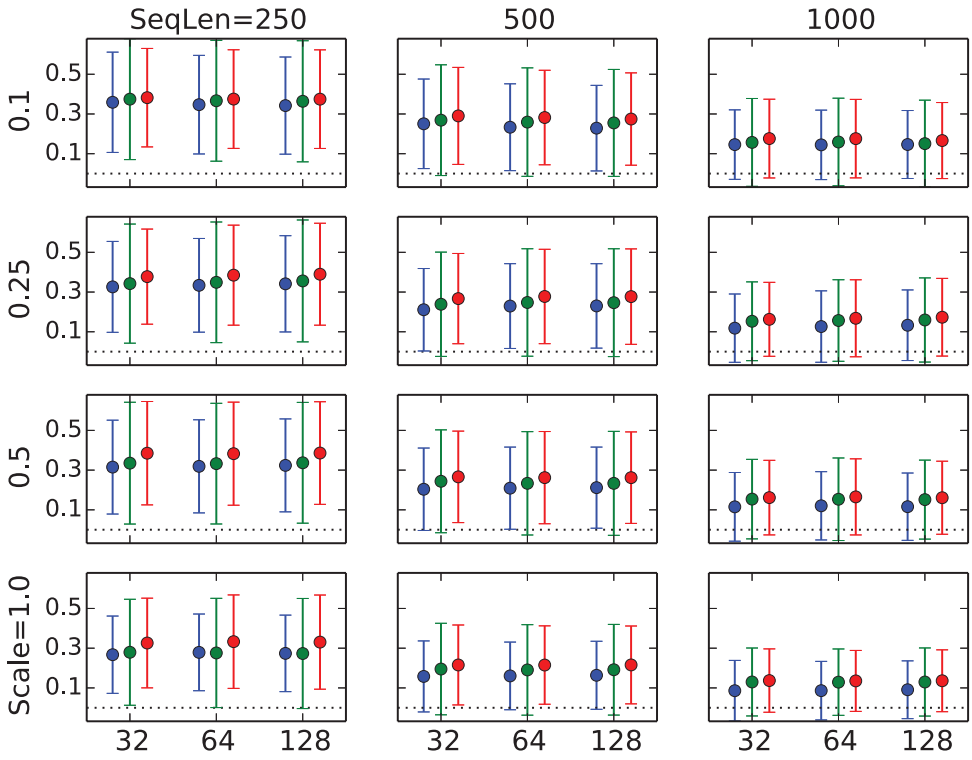

FIGURE 12. The Robinson-Foulds distances (RF) between pairs of trees. For each number of loci, the three circles and their corresponding error bars from left to right correspond to RF between true trees and those estimated by our method, RF between true trees and those estimated by RAxML, and RF between the gene tree topologies estimated by our method and those estimated by RAxML, respectively.

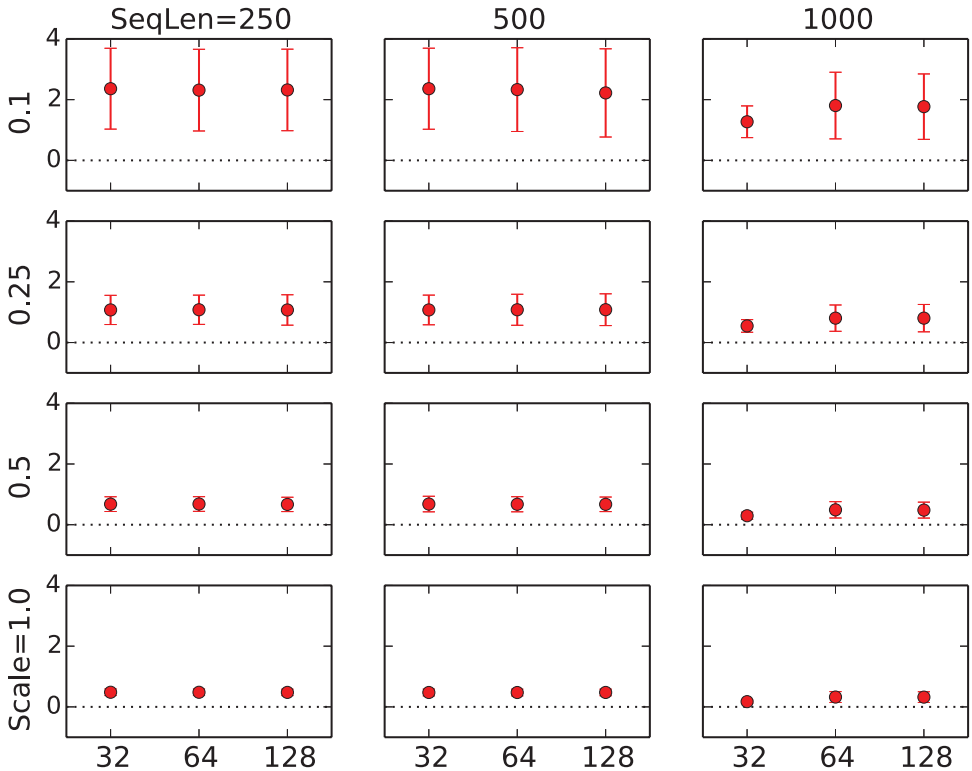

FIGURE 13. The Normalized Rooted Branch Score (NRBS) (Heled and Drummond 2010) between the true gene trees and those estimated by our method. The branch lengths are scaled in coalescent units and divided by their corresponding scale parameter 0.1, 0.25, 0.5, 1.0 for better comparison.

FIGURE 14. The top five topologies sampled using the method (Wen et al. 2016a) on the true gene trees, as well as the gene tree estimates. The leftmost topology is the true network topology and the second from left is the backbone tree of the true network topology. See the main text for details on the 95% credible sets in terms of these five topologies for the different data sets used.

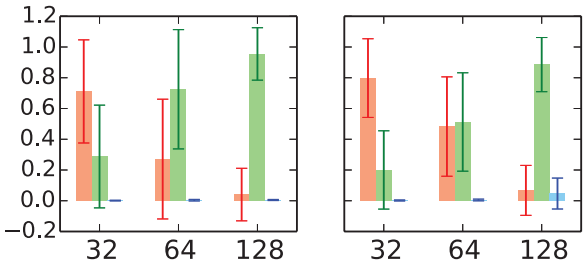

FIGURE 15. Proportions of different types of networks in the 95% credible sets sampled by the method of (Wen et al. 2016a) on data sets with 32, 64, and 128 loci. For each number of loci in each panel, the three bars from left to right correspond to proportions of inferred trees, inferred 1-reticulation networks, and inferred 2-reticulations networks, respectively. Left: The true gene tree topologies are used as the input data. Right: The gene tree estimates (using RAxML) are used as the input data.

to point out that using a large number of loci also requires a very large number of iterations and would add significantly to the complexity of sampling the posterior distribution. In the case of $s=0.5$, a fast convergence towards the true number is observed as the number of loci increases. It is worth pointing out that, in the case of $s=0.5$, increasing the number of loci, even when the sequences are very short, is much more advantageous than increasing the sequence lengths of the individual loci. It is also important to note here that in analyzing biological data sets, one cannot use longer sequences without risking violating the recombination-free loci assumption. In the case of $s=1.0$, the method does very well at estimating the number of reticulations. Finally, observe that the method almost never overestimates the number of reticulations on these data sets.

In assessing the quality of the estimated network topology itself, we analyzed the recovered networks in two ways. First, we compared the inferred network to the true network using a topological dissimilarity measure (Nakhleh 2010b). Second, when the method infers a tree, rather than a network, we compared the tree to the "backbone tree" of the true network (the tree resulting from removing the arrow in Fig. 7) using the Robinson–Foulds metric (Robinson and Foulds 1981). The latter comparison allows us to answer the question: When the method estimates the species phylogeny to be a tree, how does this tree compare to the backbone tree of the true network? It is important to note, though, that the relationship of a phylogenetic network and its constituent trees can become too complex to be captured

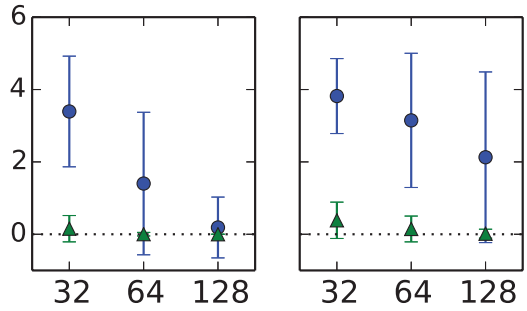

FIGURE 16. The topological difference between the true and inferred networks (circles) and the Robinson–Foulds distance between the inferred tree (if a network is inferred, this case is not included) and the backbone tree of the true network (triangles). Left: The true gene tree topologies are used as the input data. Right: The gene tree estimates (using RAxML) are used as the input data.

by a backbone tree in the presence of incomplete lineage sorting (Zhu et al. 2016). Figure 11 shows the results. The results in terms of the topological difference between the inferred and true networks parallel those that we discussed above in terms of the estimates of the number of reticulations: Poor accuracy and no sign of convergence to the true network in cases of very small values of the scaling parameter, and very good accuracy and fast convergence to accurate estimates in cases of larger values of the scaling parameter. However, the topological difference between the inferred trees (in the cases where trees were inferred) and the backbone tree reveal an important insight: When the method fails to recover the true network, it does a very good job at recovering the backbone tree of the true network.

Finally, quantifying and assessing the quality of the inheritance probability estimates is very challenging. The reason is that some runs result in network topology estimates that differ from the true network topology. In our inspection of cases where the true network was recovered, the inheritance probability estimates are very accurate.

Our Method Provides Accurate Estimates of the Gene Trees

Thus far, we have analyzed the accuracy of the inferred networks and their associated parameters. While MCMC methods in this context are deployed to approximate the integration over gene trees in a simulated manner, the methods do provide the sampled gene trees (topologies and coalescence times). The accuracy of those sampled gene trees is important for at least two reasons. First, their accuracy directly impacts and explains the accuracy of the networks. Second, the gene trees themselves are a quantity of interest in many applications.

It is important to note here two relevant studies that have addressed the issue of gene tree accuracy in the context of species tree estimation. First, (Bayzid and Warnow 2013) showed that ∗BEAST yields more accurate gene trees than would be estimated by RAxML, attributing the higher accuracy to the coestimation nature of the former method. Second, (DeGiorgio and Degnan 2014) found that methods for estimating gene trees do a better job at estimating the topologies than

the coalescence times and that this leads to more accurate species tree estimates when using gene tree topologies alone as opposed to using coalescence times as well. While both studies were conducted in the context of species trees, our goal here is not to reproduce these extensive studies in the context of phylogenetic networks, but rather to demonstrate that the main conclusions still hold even when the species phylogeny is reticulate.

In Fig. 12, we report the Robinson–Foulds distances between the true gene tree topologies and those sampled by our method, as well as the distance between the true gene tree topologies and those estimated by RAxML. The results demonstrate that the coestimated gene tree topologies are, on average, slightly closer to the true gene tree topologies than those estimated in a standalone manner using RAxML. Nonetheless, it is worth pointing out that the error bars of our method are smaller than those pertaining to the RAxML gene trees. Both methods obtained improved accuracy as the sequence length increased.

As the results in the next section show, the networks inferred from sequences directly are more accurate than those inferred from gene tree estimates. The question is: What is causing this difference if the gene tree topologies estimated by both our method and RAxML are not that different? One interesting observation we make is that while both our method and RAxML infer gene tree topologies that, on average, are of equal distance from the true gene tree topologies, the two methods return different trees, as shown in Fig. 12. That is, under the Robinson–Foulds distance, both methods infer gene trees whose topologies could be considered to be, roughly, equally good. However, the topologies are not the same. This difference could explain, at least in part, the increased accuracy of the networks and their associated parameters when inferred from sequences as opposed to gene tree estimates.

To further investigate this question, we turned our attention to the accuracy of the coalescence times estimated by our method. Figure 13 shows the Normalized Rooted Branch Score (NRBS) (Heled and Drummond 2010) between the gene trees estimated by our method and the true gene trees. This measure takes into account the branch lengths of the gene trees and not only the topologies. These results clearly show that, except for the hardest case of 0.1 scaling factor, the method performs very well in terms of estimating the coalescence times, not only in terms of the mean value but also in terms of the very small standard deviations.

It is important to comment on a seeming discrepancy between Figs. 12 and 13. For example, in the case of scaling factor 1.0, Fig 12 shows a Robinson–Foulds distance of 0.3, yet Fig. 13 shows an NRBS value close to 0. Given that the number of taxa is 5, a Robinson–Foulds value of 0.3 amounts, roughly, to a single incorrect branch in the gene tree. However, while the true and estimated gene tree differ by one branch, the difference in coalescence times between the two trees could be negligible, which explains the small NRBS values.

Next, we show the effect of errors in gene tree estimates on the accuracy of and data requirement for phylogenetic network estimates.

Inference from Gene Tree Estimates Requires More Data Than Inference from Sequences

We also set out to compare the performance of our method to that of the method we developed earlier for Bayesian inference of phylogenetic networks from gene tree data (Wen et al. 2016a). This method is also implemented in PhyloNet (Than et al. 2008) and executed via the command `MCMC_GT`. The goal here is to assess the gains one obtains by using the sequence data directly rather than first estimating gene trees and then using those as the data for species phylogeny inference.

For the purpose of this experiment we used the subset of the data sets described above and simulated on the phylogenetic network of Fig. 7 under the settings of $s=1.0$, sequence length 250, and 32, 64, and 128 loci. When using the method of (Wen et al. 2016a) we ran it once on the true gene trees and again using the gene tree estimates obtained by RAxML (Stamatakis 2014).

We ran the method of (Wen et al. 2016a) for $1,100,000$ iterations with $100,000$ burn-in and sampled every 1000 iterations. The top five topologies sampled are shown in Fig. 14 (they were the same top topologies when either the true gene trees or gene tree estimates were used).

When using the true gene tree topologies as input data, the results were as follows:

- For the 32-locus data set, the 95% credible set contains 16.4% the true network, 59.6% the backbone tree, 12.5% other 1-reticulation networks, and 11.5% other trees.
- For the 64-locus data set, the 95% credible set contains 66.0% the true network, 27.1% the backbone tree, and 3.8% the 1-reticulation network resulting for the backbone tree with reticulation edge $C \rightarrow E$ (the network in the middle of Fig. 14).
- For the 128-locus data set, the 95% credible set contains 91.7% the true network, and 4.4% the backbone tree.

When using the gene tree topology estimates as input data, the results were as follows:

- For the 32-locus data set, the 95% credible set contains 6.1% the true network, 47.3% the backbone tree, 14.1% other 1-reticulation networks, and 32.5% other trees.
- For the 64-locus data set, the 95% credible set contains 24.7% the true network, 40.5% the backbone tree, and 8.6% the 1-reticulation network resulting for the backbone tree with reticulation edge $C \rightarrow E$, 18.4% other 1-reticulation networks, and 7.8% other trees.

- For the 128-locus data set, the 95% credible set contains 49.9% the true network, 19.1 the 1-reticulation network resulting for the backbone tree with reticulation edge $C \to E$, 5.7% the backbone tree, and 35.2% other 1-reticulation networks.

More comprehensively, Fig. 15 shows the proportions of 0- (tree), 1-, and 2-reticulation networks in the 95% credible sets on each of the data sets when different numbers of loci are used and when the method of (Wen et al. 2016a) is run on true and estimated gene tree topologies.

We also assessed the quality of the inferred network/tree topologies by comparing them to the true network using the topological dissimilarity measure (Nakhleh 2010b). When the method infers a tree, rather than a network, we compared the tree to the backbone tree of the true network using the Robinson–Foulds metric (Robinson and Foulds 1981). The results are in Fig. 16.

Clearly, the results indicate the method's performance in terms of phylogenetic inference improves as the number of loci increases, and, unsurprisingly, the method has a much better performance when the true gene trees are used as input. However, for empirical data sets, the "true" gene trees are never known, and their estimates must be used for methods that utilize gene trees as data.

Contrast these results to those obtained by our method when it is run on the sequence data as input (bottom left panel in Fig. 11). Estimation from sequence data outperforms inference from gene tree topologies, even when using the true gene tree topologies. This is mainly due to the fact that the gene tree topology does not capture all the information that the sequence data do. In particular, we observe that inference from sequence data requires a much smaller number of loci than that required to achieve a similar accuracy when making inferences from gene tree topology estimates.

Intermixture Versus Gene Flow: Comparing the Method's Performance on Data Under Both Models

As we discussed above and illustrated in Fig. 2, intermixture and gene flow provide two different abstract models of reticulation. Furthermore, the program ms (Hudson 2002) allows for generating data under both models. While the MSNC is based on an intermixture model, we study here how it performs on data simulated under a gene flow model. We set up the experiment so that data are generated under the same phylogenetic networks and their parameters, yet under the scenarios of intermixture and gene flow separately. Furthermore, in this part, we assess the performance when multiple reticulation events occur between the same pair of species—a very realistic scenario in practice. Figure 17 shows the six phylogenetic networks we used to generate data.

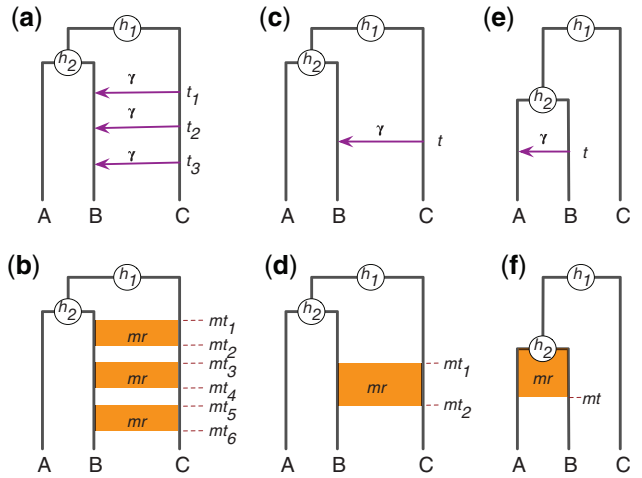

FIGURE 17. True phylogenetic histories with intermixture and gene flow models. Recurrent reticulations between nonsister taxa (a,b), a single reticulation between nonsister taxa (c,d), and a single reticulation between sister taxa (e,f) are captured under both the intermixture model (top) and gene flow model (bottom). Parameters h_1 and h_2 denote divergence times (in coalescent units), t_i parameters denote intermixture times, mt_i parameters denote start/end of migration epochs, γ is the inheritance probability, and mr is the population migration rate (see main text).

For each simulation setting, we simulated 20 data sets with 200 1-kb loci (in this part, we did not vary the sequence lengths and numbers of loci). We set the population mutation rate at 0.02 across all the branches. Furthermore we set the inheritance probability γ and the migration rate mr each to 0.20 (here, $mr = 2Nm$, where N is the effective population size, and m is the fraction of the recipient population that is made up of migrants from the donor population in each generation). We set $h_1 = 9, h_2 = 6$. For the intermixture model (Fig. 17(a)), we set $t_2 = 3$, and varied (t_1, t_3) to take on the values $(4,2)$, $(5,1)$, and $(6,0)$ so that the elapsed time, denoted by Δt, between subsequent reticulation events is 1, 2, or 3. For the gene flow model (Fig. 17(b)), we set $(mt_1, \ldots, mt_6)$ to $(6, 5, 3.5, 2.5, 1, 0)$, so that the duration of each gene flow epoch is 1 and the time elapsed between between two consecutive epochs, denoted by Δmt, is 1.5. The commands for the ms and Seq-gen programs are given in Supplementary Materials.

For each data set, we ran an MCMC chain of 8×10^6 iterations with 1×10^6 burn-in. One sample was collected from every 5000 iterations, resulting in a total of 1400 collected samples. We summarized the results based on 28,000 samples from 20 replicates for each parameter setting.

Table 1 shows the population mutation rates, divergence times, and numbers of reticulations estimated by our method on data generated under the models of Fig. 17(a) and (b). As the results show, the method performs very well in terms of estimating the divergence times and population mutation rates, regardless of whether the data were generated under an intermixture model or a gene flow model. Furthermore, for these two parameters, the estimates are stable

TABLE 1. Estimated population mutation rates (θ), divergence times (h_1 and h_2), and numbers of reticulations (#reti) as a function of varying Δt in the model of Fig. 17(a) and Δmt in the model of Fig. 17(b)

Case	θ	h_1	h_2	#reti
$\Delta t=1$	0.022 ± 0.002	8.9 ± 0.1	5.9 ± 0.1	1.2 ± 0.4
$\Delta t=2$	0.022 ± 0.002	8.9 ± 0.1	5.9 ± 0.1	2.0 ± 0.0
$\Delta t=3$	0.021 ± 0.003	9.0 ± 0.1	6.0 ± 0.1	2.6 ± 0.5
$\Delta mt=1.5$	0.023 ± 0.003	8.9 ± 0.1	6.0 ± 0.1	2.1 ± 0.3

The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2=0.01$.

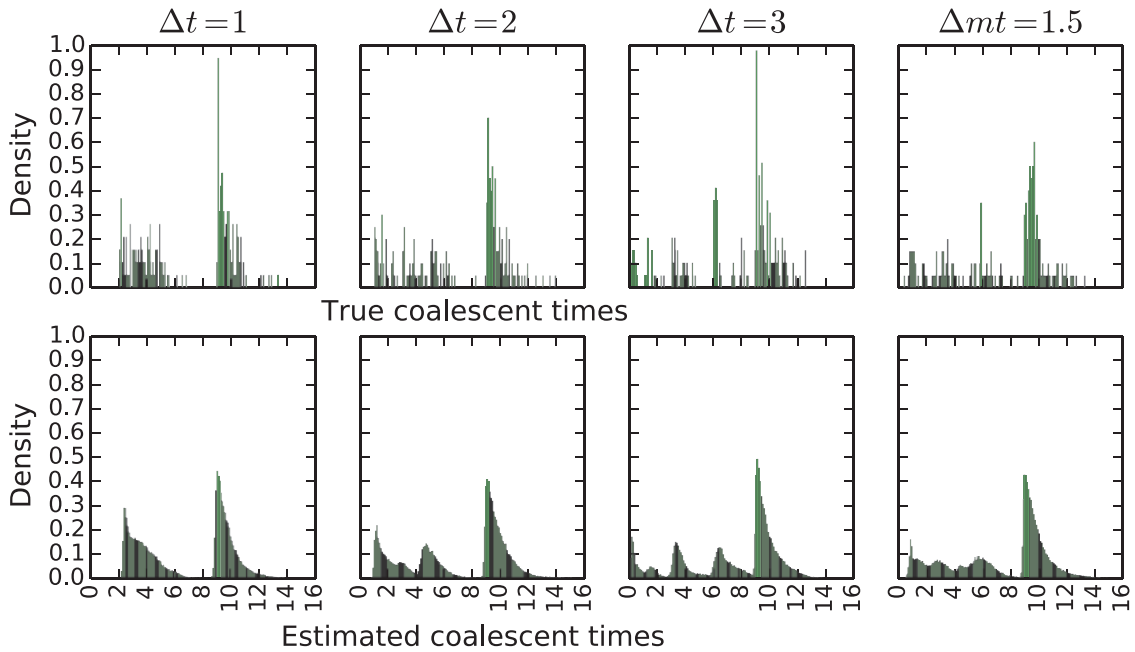

FIGURE 18. Histograms of the true (top) and estimated (bottom) coalescence times (in coalescent units) of the MRCA of alleles from B and C on data generated under the models of Fig. 17(a) and (b).

while varying the elapsed times between consecutive reticulation events.

As for the estimated number of reticulations, it becomes more accurate as the elapsed times between consecutive reticulations is larger. To better understand the factors that affect the detectability of reticulations, we plotted histograms of the true and estimated coalescence times of the most recent common ancestor (MRCA) of alleles from B and C in Fig. 18. Here, the true coalescence times are obtained from the true gene tree generated by the program `ms`. The estimated coalescence times are sampled by our method along with the gene tree topologies. For the estimated coalescence times, we plot them based on all the collected samples, which is why the histograms of estimated coalescence times are smoother than those of the true ones.

As Fig. 17(a) and (b) show, the coalescence times of alleles from B and C would form a mixture of four distributions: three due to the three reticulation events, and one above the root of the phylogenetic network. As the left three columns of panels in Fig 18 show, under an intermixture model, as Δt increases, the signal for a mixture of four distributions of (A,B) coalescence times becomes much stronger, thus pointing to three reticulations in addition to the coalescence events above the root of the phylogeny. This is why, under the intermixture model, the method's performance in terms of the estimated number of reticulations improves as Δt increases. However, on data simulated the under the gene flow model (the rightmost column of panels in Fig. 18), the signal of the mixture of four distributions of (A,B) coalescence times is surprisingly stronger than that under the intermixture model with the comparable $\Delta t=1$ and $\Delta t=2$.

Figure 19 shows results similar to those reported in Fig. 18, with the only difference being that these are the coalescence times from all 4000 loci generated from the 20 data sets of 200 loci each. Effectively, this is the signal in a data set of 4000 independent loci. Clearly, the signal is much stronger than in data sets of 200 loci, and all reticulations would be recoverable under the intermixture model for $\Delta t=2,3$ and for the gene flow model.

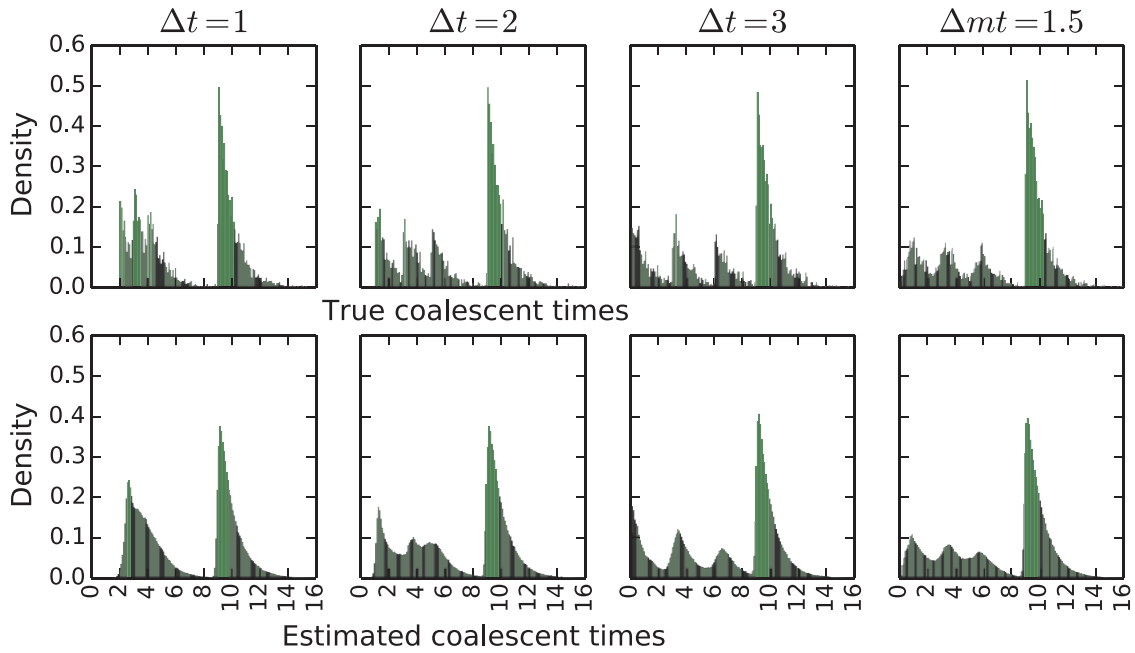

FIGURE 19. Histograms of the true (top) and estimated (bottom) coalescence times (in coalescent units) of the MRCA of alleles from *B* and *C* on 4,000 loci generated under the models of Fig. 17(a) and (b).

TABLE 2. Estimated population mutation rates (θ), divergence times (h_1 and h_2), inheritance/migration rates, and numbers of reticulations (#reti) as a function of varying t in the model of Fig. 17(c) and mt_2 in the model of Fig. 17(d)

Case	θ	h_1	h_2	γ (mr)	#reti
$t=1$	0.020±0.002	2.5±0.1	1.5±0.1	0.20±0.05	1.0±0.0
$t=0$	0.020±0.002	2.5±0.1	1.5±0.1	0.21±0.04	1.0±0.0
$mt_2=1$	0.020±0.002	2.5±0.1	1.5±0.1	0.18±0.05	1.0±0.0
$mt_2=0$	0.022±0.002	2.5±0.1	1.5±0.1	0.17±0.04	1.0±0.0

The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2=0.01$.

We also ran simulations where we varied the number of individuals sampled from species B (we sampled 1, 3, and 5 individuals). The results improve as the number of individuals increases from 1 to 3, but no discernible improvement is achieved under our simulation settings when the number of individual is increased to 5. Results are given in the Supplementary Material available on Dryad.

To assess the performance of our method on the simpler case of a single reticulation event, we considered the networks in Fig. 17(c) and (d), set $h_1=2.5$, $h_2=1.5$, and $mt_1=h_2$, and varied $t, mt_2 \in \{1,0\}$. As the results in Table 2 demonstrate, our method estimated the population mutation rate θ, the divergence times h_1 and h_2, and the inheritance probability/migration rate very accurately under all cases. The method did very well also in terms of estimating t and mt_2; results in Supplementary Materials available on Dryad.

A single reticulation was detected for all cases of intermixture and gene flow. We plotted the histograms of the true and estimated coalescence times of the MRCA of alleles from *B* and *C* in Fig. 20. As the figure shows, the distributions of estimated coalescence times match the distributions of true coalescence times very well. Furthermore, when using 4,000 loci, the signal becomes even stronger; results in Supplementary Materials available on Dryad.

Finally, we assessed the performance of our method on cases where the reticulation event involves sister taxa. Fig. 17(e) and (f) show the cases we considered, with setting $h_1=2.5$ and $h_2=1.5$, and varying $t, mt \in \{1,0\}$.

As the results in Table 3 demonstrate, our method obtained very accurate estimates of the various parameters under $t=0$ and $mt=0$. Under the cases of intermixture with $t=1$ and gene flow with $mt=1$, our method did not detect the reticulation, which resulted in an underestimation of h_2. In the case of $mt=0$, the migration rate was severely underestimated, most likely due to the short time interval between the migration and divergence events between *A* and *B*. The method did very well also in terms of estimating t and mt; results in Supplementary Materials available on Dryad.

We plotted the histograms of the true and estimated coalescence times of the MRCA of alleles from *A* and *B* in

TABLE 3. Estimated population mutation rates (θ), divergence times (h_1 and h_2), inheritance/migration rates, and numbers of reticulations (#reti) as a function of varying t in the model of Fig. 17(e) and mt in the model of Fig. 17(f)

Case	θ	h_1	h_2	γ	#reti
$t=1$	0.020 ± 0.002	2.5 ± 0.1	1.3 ± 0.1	NA	0.0 ± 0.0
$t=0$	0.020 ± 0.002	2.5 ± 0.1	1.5 ± 0.0	0.21 ± 0.06	1.0 ± 0.0
$mt=1$	0.020 ± 0.002	2.5 ± 0.1	1.4 ± 0.1	NA	0.0 ± 0.0
$mt=0$	0.022 ± 0.002	2.5 ± 0.1	1.5 ± 0.1	0.11 ± 0.06	1.0 ± 0.0

The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2=0.01$.

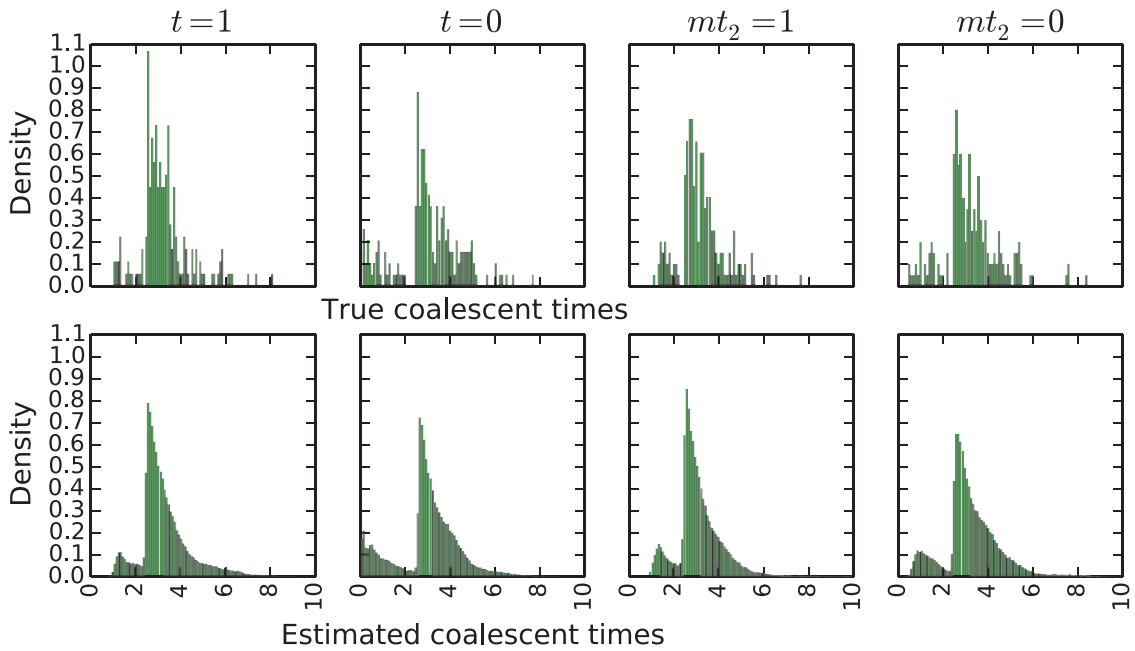

FIGURE 20. Histograms of the true (top) and estimated (bottom) coalescence times (in coalescent units) of the MRCA of alleles from B and C on data generated under the models of Fig. 17(c) and (d).

Fig. 21. When $t=1$ and $mt=1$, the signal of reticulation is very low, which explains the failure of our method to detect it. In the cases of $t=0$ and $mt=0$, the distributions of estimated coalescence times match those of true coalescence times very well. When using 4,000 loci, the signal becomes even stronger; results in Supplementary Materials.

Analysis of a 106-Locus Yeast Data Set

The yeast data set of (Rokas et al. 2003) consists of 106 loci from seven Saccharomyces species, *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae* (Smik), *S. kudriavzevii* (Skud), *S. bayanus* (Sbay), *S. castellii* (Scas), and *S. kluyveri* (Sklu). Rokas *et al.* (Rokas et al. 2003) reported on extensive incongruence of single-gene phylogenies and revealed the species tree from concatenation method (Fig. 22(a)). Edwards *et al.* (Edwards et al. 2007) reported as the two main species trees and gene tree topologies sampled from BEST (Liu 2008) the two trees shown in Fig. 22(a-b). The other gene tree topologies (Fig. 22(c)) exhibited weak phylogenetic signals among Sklu, Scas and the other species. Bloomquist and Suchard (2010) reanalyzed the data set without Sklu since it added too much noise to their analysis. Their analysis resulted in many horizontal events between Scas and the rest of the species because the Scas lineage-specific rate variation is much stronger than that of the other species. Yu et al. (2013b) analyzed the 106-locus data set restricted to the five species Scer, Spar, Smik, Skud, and Sbay and identified a maximum parsimony network that supports a hybridization from Skud to Sbay with inheritance probability of 0.38.

Analyzing the 106-locus data set using our method, the 95% credible set contains many topologies with similar hybridization patterns; the MAP network is shown in Fig. 22(d). All the previous findings are encompassed by the networks inferred by our method. The two hybridizations between Sklu and Scas (green edges in 22(d)) indicate the weak phylogenetic signals among Sklu, Scas and the rest of the species. The hybridization from Scas to the other species except for Sklu (red

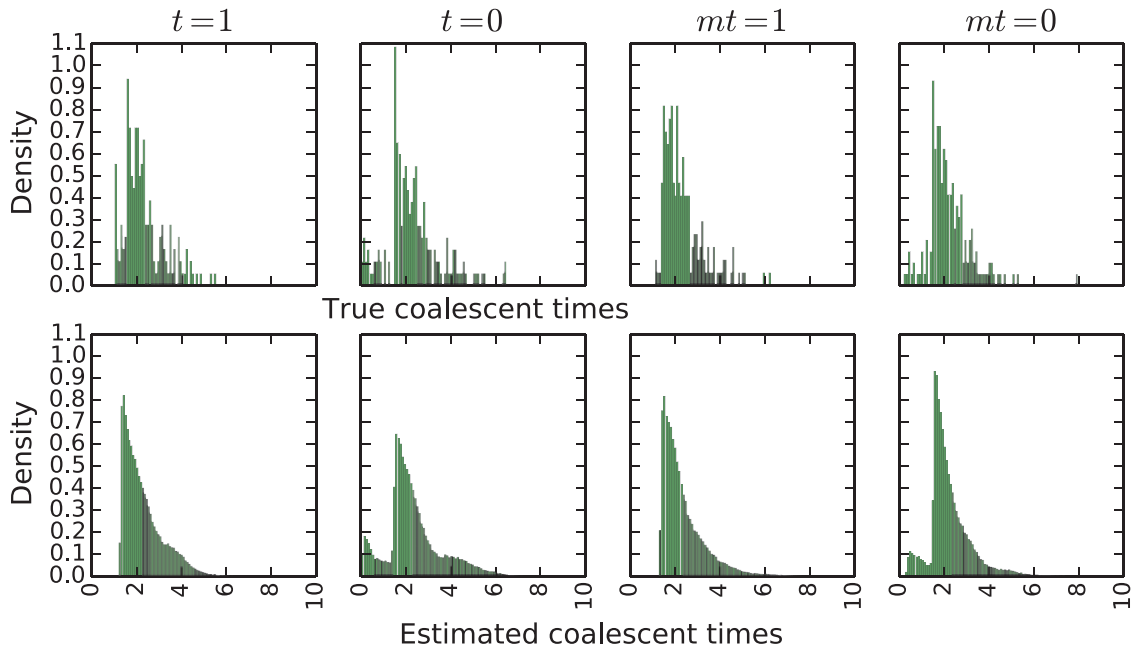

FIGURE 21. Histograms of the true (top) and estimated (bottom) coalescence times (in coalescent units) of the MRCA of alleles from *A* and *B* on data generated under the models of Fig. 17(e) and (f).

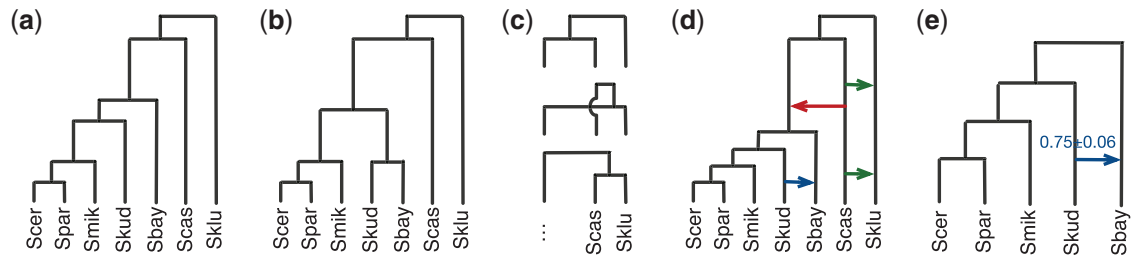

FIGURE 22. Results on the yeast data set of (Rokas et al. 2003). (a) The species tree inferred using the concatenation method (Rokas et al. 2003) and the main species tree and gene tree topology sampled using BEST (Edwards et al. 2007). (b) The second most frequently sampled species and gene tree topology by BEST (Edwards et al. 2007). (c) Many other gene tree topologies were sampled by BEST (Edwards et al. 2007), indicating weak phylogenetic signals among Sklu, Scas, and the rest of the species. (d) The MAP phylogenetic network inferred by our method on all 106 loci. (e) The single phylogenetic network inferred using all 106 loci from the five species Scer, Spar, Smik, Skud, Sbay.

edge in 22(d)) captures the stronger lineage-specific rate variation in Scas. Finally, the hybridization from Skud to Sbay (blue edge in 22(d)) resolves the incongruence between the two main species tree topologies in 22(a–b).

We then analyzed the 106-locus data set restricted to the five species Scer, Spar, Smik, Skud, and Sbay. The phylogenetic signal in this data set is very strong—the consensus trees of 99 out of the 106 loci contain two internal branches. The MAP phylogenetic network in Fig. 22(f) contains the hybridization from Skud to Sbay, which is identical to the subnetwork in Fig. 22(d). See Supplementary Materials for full details. In summary, analysis of the yeast data set demonstrates the effect of phylogenetic signal in the individual loci on the inference and the care that must be taken when selecting loci of analysis of reticulate evolutionary histories.

We compared these analyses to ones obtained by the method of (Wen et al. 2016a) when the input data consist of gene tree estimates. When the gene tree estimates on all seven Saccharomyces species are used, the 95% credible set consisted of a single network that is shown in Fig. 22(d), yet with only the single reticulation from Skud to Sbay. When the gene tree estimates on the subset of five species were used as input, the 95% credible set consisted of a single network that is shown in Fig. 22(e), in agreement with the results based on coestimation from the sequence data directly.

Finally, we quantified the Robinson–Foulds distances between the locus-specific gene tree estimates obtained by our method and by RAxML. The distances were 0.33 ± 0.19 for the 7-taxon data set, and 0.33 ± 0.16 for the 5-taxon data set. It is worth noting that these distances

are very similar to those observed in Fig. 12 above. Full details and further results for this data set are given in Supplementary Materials.

Discussion

To conclude, we have devised a Bayesian framework for sampling the parameters of the MSNC model, including the species phylogeny, gene trees, divergence times, and population sizes, from sequences of multiple independent loci. Our work provides the first general framework for Bayesian phylogenomic inference from sequence data in the presence of hybridization. The method is publicly available in the open-source software package PhyloNet (Than et al. 2008). We demonstrate the utility of our method on simulated data and an empirical data set.

Our results demonstrate several important aspects. First, ignoring hybridization when it had occurred results in underestimating the divergence times of species and overestimating the coalescence times of individual loci. Second, coestimation of species phylogeny and gene trees results in more accurate gene tree estimates than the inferences of gene trees from sequences directly. Third, comparing to existing phylogenetic network inference methods (Yu et al. 2014; Wen et al. 2016a) that use gene tree estimates as input, our method not only estimates more parameters, such as divergence times and population sizes, but also estimates more accurate phylogenetic networks from fewer loci. Further, we assessed the performance of our model and method on simulated data generated under a gene flow model. Our method performed very well on such data. However, given the nature of our abstract phylogenetic network model, a gene flow epoch is estimated as a single reticulation event. Finally, we analyzed a 106-locus yeast data set and demonstrated for empirical data the differences in results one obtains when coestimating the gene and species phylogenies when compared to inferences from gene tree estimates.

Finally, we identify several directions for further improvements of our proposed approach. First, while priors on species trees, such as the birth-death model, have been developed and employed by inference methods, similar prior distributions on phylogenetic networks are currently lacking. Second, while techniques such as the majority-rule consensus exist for summarizing the trees sampled from the posterior distribution, principled methods for summarizing sampled networks are needed. Our software package, PhyloNet, as well as the software package of (Solís-Lemus et al. 2017) currently have initial implementations of network summarization. Third, the complexity of computing the likelihood of a phylogenetic network could be several orders of magnitude than that of computing the likelihood of a species tree with the same number of taxa. Developing techniques for efficient likelihood computations on phylogenetic networks are essential for these methods to scale up to larger data sets. Last but not least, the sequence data used here, and in almost all phylogenomic analyses, consist of haploid sequences of randomly phased diploid genomes. The effect of random phasing on inferences in general needs to be studied in detail. Furthermore, the model could be extended to work directly on unphased data by integrating over possible phasings (Gronau et al. 2011).

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.3h185.

Funding

This work was supported by the National Science Foundation (CCF-1302179, CCF-1514177, and DBI-1062463) to L.N.; National Science Foundation grants OCI-0959097 (Data Analysis and Visualization Cyberinfrastructure), and CNS-1338099 (Big-Data Private-Cloud Research Cyberinfrastructure), in part.

References

Arnold M.L. 1997. Natural hybridization and evolution. Oxford: Oxford University Press.

Ayres D.L., Darling A., Zwickl D.J., Beerli P., Holder M.T., Lewis P.O., Huelsenbeck J.P., Ronquist F., Swofford D.L., Cummings M.P., Rambaut A., Suchard M.A. 2011. Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* 61:170–173.

Barton N. 2001. The role of hybridization in evolution. Mol. Ecol. 10:551–568.

Bayzid M.S., Warnow T. 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29:2277–2284.

Bloomquist E., Suchard M. 2010. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. Syst. Biol. 59:27–41.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10:e1003537.

DeGiorgio M., Degnan J.H. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst. Biol. 63:66–82.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. 104(14): 5936–5941.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y-C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

Gogarten J.P., Doolittle W.F., Lawrence J.G. 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19:2226–2238.

Green P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Green P.J. 2003. Trans-dimensional Markov chain Monte Carlo. In: Green P., Hjort N., Richardson S., editors. Highly structured

stochastic processes. Oxford, UK: Oxford University Press. p. 179–198.
Gronau I., Hubisz M.J., Gulko B., Danko C.G., Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genetics 43:1031–1034.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Hey J., Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760.
Hey J., Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. 104(8): 2785–2790.
Hudson R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Koonin E.V., Makarova K.S., Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification 1. Annu. Rev. Microbiol. 55:709–742.
Leaché, A.D., Harris R.B., Rannala B., Yang Z. 2013. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.
Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543.
Long J.C. 1991. The genetic structure of admixed populations. Genetics 127:417–428.
Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.
Mallet J. 2007. Hybrid speciation. Nature 446: 279–283.
Nakhleh L. 2010a. Evolutionary phylogenetic networks: models and issues. In: Heath L., Ramakrishnan N., editors. The problem solving handbook for computational biology and bioinformatics. New York: Springer. p. 125–158.
Nakhleh L. 2010b. A metric on the space of reduced phylogenetic networks. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 7:218–222.
Pickrell J.K., Pritchard J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8:e1002967.
Rambaut A., Grassly N.C. 1997. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13: 235–238.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Reich D., Thangaraj K., Patterson N., Price A.L., Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.
Rieseberg L. 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28: 359–389.
Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. Math. Biosci. 53: 131–147.
Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Slatkin M., Maddison W.P. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123: 603–613.
Solís-Lemus C., Yang M., Ané, C. 2016. Inconsistency of species tree methods under gene flow. Syst. Biol. 65:843–851.
Solís-Lemus C., Bastide P., Ané, C. 2017. Phylonetworks: a package for phylogenetic networks. Mol. Biol. Evol. (In press)
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30:1312–1313.
Strasburg J.L., Rieseberg L.H. 2010. How robust are "isolation with migration" analyses to violations of the IM model? A simulation study. Mol. Biol. Evol. 27:297–310.
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9:322.
Wen D., Yu Y., Nakhleh L. 2016a. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genetics 12:e1006006.
Wen D., Yu Y., Hahn M.W., Nakhleh L. 2016b. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol. Ecol. 25: 2361–2372.
Whitlock M.C., Mccauley D.E. 1999. Indirect measures of gene flow and migration: Fst≠ 1/(4nm+ 1). Heredity 82:117–125.
Yu Y., Degnan J.H., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genetics 8:e1002660.
Yu Y., Ristic N., Nakhleh L. 2013a. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. BMC Bioinformatics 14(Suppl 15): S6.
Yu Y., Barnett R.M., Nakhleh L. 2013b. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Syst. Biol. 62:738–751.
Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. 111(46): 16448–16453.
Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2017. Bayesian inference of species networks from multilocus sequence data. bioRxiv, pp. 124982. (https://doi.org/10.1101/124982)
Zhu J., Yu Y., Nakhleh L. 2016. In the light of deep coalescence: revisiting trees within networks. BMC Bioinformatics, 17:415.

Supplementary Information: Co-estimating Reticulate Phylogenies and Gene Trees from Multi-locus Sequence Data

Dingqiao Wen[1] and Luay Nakhleh[1,2,*]

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

[2]BioSciences, Rice University, Houston, TX, USA

[*]nakhleh@rice.edu

Contents

1 Sampling from the Posterior using RJMCMC **3**

1.1 Moves for the phylogenetic network and inheritance probabilities 5

1.1.1 Change-Parameters 5

1.1.2 Change-Topology 7

1.1.3 Change-Dimension 10

1.2 Convergence diagnostics 14

2 Our Method vs. ∗BEAST on Data with No Reticulations **15**

2.1 Simulation settings 15

2.2 Results 16

3 Our Method vs. ∗BEAST on Data with Reticulations **21**

3.1 Simulations settings 21

3.2 Our method provides accurate estimates of the phylogenetic networks, gene trees, and their parameters 22

3.3 ∗BEAST underestimates divergence times and overestimates coalescent times when the evolutionary history is reticulate 28

3.4 Simultaneous inference of phylogenetic networks and gene trees provides more accurate gene trees than gene trees estimated from individual loci . . 33

3.5 Inference from gene tree estimates requires more data than inference from sequences directly . . . 33

4 Simulations under Intermixture/Gene Flow Models 35

4.1 MCMC settings . . . 36

4.1.1 The effect of the number of individuals . . . 37

4.2 Paraphyletic intermixture/gene flow . . . 38

4.3 Isolation-migration between sister species . . . 38

5 Analysis of a Yeast Data Set 41

5.1 MCMC settings . . . 42

5.2 Data preprocessing . . . 42

5.3 Results for the full data set . . . 43

5.4 Results for the data set of 106 loci from five Saccharomyces species . . . 46

6 Runtimes 48

6.1 Simulations . . . 48

6.2 Biological data sets . . . 50

7 PhyloNet Implementation and Usage 51

8 References 52

1 Sampling from the Posterior using RJMCMC

We have implemented a reversible-jump MCMC, or RJMCMC, (6) algorithm to sample from the posterior distribution as given by Eq (2) in the main text. In each iteration of the sampling, a new state (Ψ', Γ', G') is proposed and either accepted or rejected based on the Metropolis-Hastings ratio r, which is composed of the likelihood, prior, and Hastings ratios. When the proposal changes the dimensionality of the sample by adding a new reticulation or removing an existing reticulation in the phylogenetic network, the absolute value of the determinant of the Jacobian is also taken into account.

Table S1: **The six moves that the RJMCMC algorithm employs for gene trees.** These moves are randomly selected and applied to a randomly selected gene tree to generate a new one. All the moves are adapted from BEAST2 (3). Note that the moves are restricted by the temporal constraints of phylogenetic network.

Move	Description	BEAST2 operation
1. TreeScaler:	Scales all the coalescent times by a random scale factor	ScaleOperator
2. TreeNodeReheight:	Modifies the time of a randomly selected internal node	Uniform
3. SubtreeSlide:	Modifies the time of the root of a randomly selected subtree, moves the subtree towards its ancestors/descendants based on the time if necessary	SubtreeSlide
4. WilsonBalding:	Prunes a randomly selected subtree and attaches it to a random location	WilsonBalding
5. NarrowNNI:	Swaps the parents of a randomly selected node and its parent's sibling	Exchange.narrow
6. WildNNI:	Swaps the parents of two randomly selected nodes	Exchange.wide

We describe the proposal workflow as follows:

- With probability ζ, gene tree g_i is selected from $G = \{g_1, \ldots, g_m\}$.
  - One of the moves 1-6 in Table S1 is selected and applied to g_i with probabilities $\xi_1, \xi_2, ..., \xi_6$, respectively, where $\sum_{i=1}^{6} \xi_i = 1$.
- With probability $1 - \zeta$, one of the moves for phylogenetic network Ψ and inheritance probabilities Γ from Moves 1-12 in Table S2 is applied.

Table S2: **The 12 moves that the RJMCMC algorithm employs for phylogenetic network and inheritance probabilities.** These moves are randomly selected and applied to the current phylogenetic network or inheritance probabilities. Moves 1–5 do not change the model dimension or the topology of the phylogenetic network. Moves 6–10 change the topology but not the model dimension. Moves 11 and 12 change the topology and model dimension. Note that Moves 4–10 and 12 may violate the temporal constraints of gene trees, if so, undo the move.

1. Scale-PopSize:	Scale all the population sizes by a random scale factor
2. Change-PopSize:	Modifies the population size of a randomly selected edge
3. Change-Inheritance:	Modifies the inheritance probability of a randomly selected reticulation edge
4. Scale-Time:	Scale all the times by a random scale factor
5. Change-Time:	Modifies the time of a randomly selected internal node
6. Swap-Nodes:	Swap the parents of two randomly selected nodes
7. Flip-Reticulation:	Reverses the direction of a randomly selected reticulation edge
8. Slide-SubNet:	Modifies the time of the root of a randomly selected subnetwork whose tail is a tree node
9. Move-Tail:	Modifies the tail of a randomly selected edge whose tail is a tree node
10. Move-Head:	Modifies the head of a randomly selected edge whose head is a reticulation node
11. Add-Reticulation:	Adds a reticulation edge between two randomly selected edges
12. Delete-Reticulation:	Deletes a randomly selected reticulation edge

- With probability κ, one of the two dimension-changing moves, Moves 11–12 in Table S2, is selected. Add-Reticulation (Move 11) is selected with probability κ_1 and Delete-Reticulation (Move 12) is selected with probability $1-\kappa_1$. If the current network has at least one reticulation edge, then both moves are possible; otherwise, Add-Reticulation is selected.
- With probability $1-\kappa$, a non-dimension-changing move (Moves 1–10 in Table S2) is selected.
  - With probability ω a non-topology-changing move (Moves 1–5 in Table S2) is selected.If the current network has no reticulation edges, Change-Inheritance (Move 3) would not be selected.
  - With probability $1-\omega$ a topology-changing move (Moves 6-10 in Table S2) is selected. If the current network has no reticulation edges, Flip-

Reticulation (Move 6) and Move-Head (Move 10) would not be selected.

1.1 Moves for the phylogenetic network and inheritance probabilities

Since all the moves for gene trees are adapted from BEAST2 (3), we only describe the moves for phylogenetic network and inheritance probabilities below. Here, $|V|$, $|E|$, $|R|$, $|T|$, $|\theta|$ denote the number of nodes, the number of edges, the number of reticulation nodes, the number of taxa in the phylogenetic network, and the number of elements in the population size vector, respectively.

Note that these moves might

1. generate a phylogenetic network topology that violates the definition (given in the main text) in one of the following ways:
   - the proposed topology contains a cycle, or
   - the proposed topology is disconnected
2. generate a phylogenetic network that violates the temporal constraints of the gene trees.

Therefore, in computing the Metropolis-Hastings ratio, our implementation explicitly tests whether the proposed network has any of these violations; if it does, we either set the phylogenetic network prior to 0 if the topology violates the definition (given in the main text), or nullify the move if the divergence times are out of bounds.

1.1.1 Change-Parameters

Scale-PopSize. All the $|\theta|$ elements in θ are scaled by a scale factor $u \sim \mathrm{Uniform}(f, \frac{1}{f})$ where $f \in (0, 1)$ is a tuning parameter, resulting in $\theta' = u\theta$. Moving between (θ, u) and (θ', u') requires that $u' = \frac{1}{u}$, so the Hastings ratio is

$$\frac{g(u')}{g(u)}\left|\frac{\partial(\theta', u')}{\partial(\theta, u)}\right| = \frac{1}{\frac{1}{f} - f} \Big/ \frac{1}{\frac{1}{f} - f} \begin{vmatrix} \partial\theta'/\partial\theta & \partial\theta'/\partial u \\ \partial(1/u)/\partial\theta & \partial(1/u)/\partial u \end{vmatrix} = \begin{vmatrix} u\mathbf{I} & \theta \\ 0 & u^{-2} \end{vmatrix} = u^{|\theta|-2}.$$

Change-PopSize. One population size θ_b is selected uniformly at random from θ and modified into θ_b' using the proposal

$$\theta_b' = \begin{cases} \theta_b + u & \text{if} \quad \theta_b + u \geq 0 \\ -(\theta_b + u) & \text{if} \quad \theta_b + u < 0 \end{cases}$$

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

Change-Inheritance. A reticulation edge is selected uniformly at random from the list of 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 $u \sim \mathrm{Uniform}(-0.1, +0.1)$. The value 0.1 can be replaced by a tuning parameter for a more general setting. Under this setting, the Hastings ratio is $\frac{p(\gamma|\gamma')}{p(\gamma'|\gamma)} = 1$.

Scale-Time. The divergence times τ of all the internal nodes (root included) are scaled by a scale factor u and modified into $\tau' = u\tau$. u is drawn from $\mathrm{Uniform}(f, \frac{1}{f})$ where $f \in (0, 1)$ is a tuning parameter. Moving between (τ, u) and (τ', u') requires that $u' = \frac{1}{u}$, so the Hastings ratio is

$$\frac{g(u')}{g(u)}\left|\frac{\partial(\tau', u')}{\partial(\tau, u)}\right| = \frac{1}{\frac{1}{f} - f} \Big/ \frac{1}{\frac{1}{f} - f} = \begin{vmatrix} \partial\tau'/\partial\tau & \partial\tau'/\partial u \\ \partial(1/u)/\partial\tau & \partial(1/u)/\partial u \end{vmatrix} = \begin{vmatrix} u\mathbf{I} & \tau \\ 0 & u^{-2} \end{vmatrix} = u^{|V|-|T|-2}.$$

Change-Time. An internal node (root is excluded) v is selected uniformly at random and the time τ of the node is modified into $\tau_v' \sim \mathrm{Uniform}(l, h)$, where l and h are the lower and higher bound of time τ_v respectively. The lower bound should not exceed the times of the children of v (or child if v is a reticulation node). The higher bound is restricted by the times of the parents of v. Since this move is symmetric and acts uniformly at all steps, the Hastings ratio is $\frac{p(\tau_v|\tau_v')}{p(\tau_v'|\tau)} = 1$.

1.1.2 Change-Topology

Swap-Nodes. This move is adapted from ARG Swap Kernel in (2). An internal node v_1 is selected uniformly at random. If v_1 is a tree node, v_2 is selected uniformly at random from its two children and v_3 is the other child; otherwise, v_2 represents the only child of v_1 and v_3 is null. An edge $e_3 = (v_4, v_5)$ is selected uniformly at random from the edges that exist at the time of τ_{v_1}. Note that e_3 cannot be $e_1 = (v_1, v_2)$ or $e_2 = (v_1, v_3)$ if v_3 exists. There are two cases for the final step:

1. If v_2 is a reticulation node and v_4 is the other parent of v_2, or v_5 is a reticulation node and v_1 is the other parent of v_5, no action would be performed, and the Hastings ratio is set to $-\infty$.

2. Otherwise, the two edges e_1 and e_3 are removed and replaced with $e_1' = (v_1, v_5)$ and $e_3' = (v_4, v_2)$. Since the move is symmetric and acts uniformly at all steps, the Hastings ratio is 1.

Flip-Reticulation. A reticulation edge $e_1 = (v_1, v_2)$ is randomly selected from the list of reticulation edges, where v_2 is a network node.

1. If v_1 is a reticulation node as well, then this edge cannot be flipped. No action would be performed and the Hastings ratio is set to $-\infty$.

2. Let v_3 be the parent of v_1, v_4 be the other child of v_1, v_5 be the other parent of v_2, v_6 be the only child of v_2. If $\tau_{v_4} > \tau_{v_5}$, this edge cannot be flipped. The Hastings ratio is set to $-\infty$.

3. Otherwise, the edge $e_1 = (v_1, v_2)$ is replaced with the new edge $e_1 = (v_2, v_1)$. The new time τ_{v_2}', τ_{v_1}' are drawn from $\mathrm{Uniform}(\tau_{low} = \max(\tau_{v_6}, \tau_{v_4}), \tau_{v_5})$ and $\mathrm{Uniform}(\tau_{v_4}, \tau_{high} = \min(\tau_{v_3}, \tau_{v_5}))$ respectively. If $\tau_{v_1}' > \tau_{v_2}'$ (this case only happens when $\tau_{low} < \tau_{v_1}', \tau_{v_2}' < \tau_{high}$), the two times are exchanged. For the parameters, $\gamma_{(v_5,v_2)}$ is deleted and the value is assigned to (v_3, v_1), $\gamma_{e_1'} = \gamma_{e_1}$, $\theta_{e_1'} = \theta_{e_1}$. The Hastings ratio in this case is

$\frac{p(e_1|e_1')}{p(e_1'|e_1)}$ where

$$p(e_1'|e_1) = \begin{cases} \dfrac{\Delta\tau}{\tau_{v_5} - \tau_{low}} \times \dfrac{\Delta\tau}{\tau_{high} - \tau_{v_4}} & \text{if } \tau'_{v_2} \geq \tau_{high} \text{ or } \tau'_{v_1} \leq \tau_{low} \\ 2 \times \dfrac{\Delta\tau}{\tau_{v_5} - \tau_{low}} \times \dfrac{\Delta\tau}{\tau_{high} - \tau_{v_4}} & \text{if } \tau_{low} < \tau'_{v_1}, \tau'_{v_2} < \tau_{high} \end{cases}$$

and similarly,

$$p(e_1|e_1') = \begin{cases} \dfrac{\Delta\tau}{\tau_{v_3} - \tau_{low}} \times \dfrac{\Delta\tau}{\tau_{high} - \tau_{v_6}} & \text{if } \tau_{v_1} \geq \tau_{high} \text{ or } \tau_{v_2} \leq \tau_{low} \\ 2 \times \dfrac{\Delta\tau}{\tau_{v_3} - \tau_{low}} \times \dfrac{\Delta\tau}{\tau_{high} - \tau_{v_6}} & \text{if } \tau_{low} < \tau_{v_1}, \tau_{v_2} < \tau_{high} \end{cases}$$

Slide-SubNet. A tree node v_1 is randomly selected from the list of internal tree nodes (including the root r). v_2 is a child of v_1 selected at random. Let v_3 be the parent of v_1 (null if $v_1 = r$) and v_4 be the other child of v_1. A new time $\tau'_{v_1} = \tau_{v_1} + \Delta$ is proposed, where $\Delta \sim \text{Uniform}(-c, +c)$ and c is a tuning parameter.

1. If $\max(\tau_{v_2}, \tau_{v_4}) \leq \tau'_{v_1} \leq \tau_{v_3}$ ($\tau_{v_3} = \infty$ when $v_1 = r$), the topology stays the same. The time τ_{v_1} is modified into τ'_{v_1}. Since v_1 and τ'_{v_1} are both selected uniformly, the Hastings ratio is $\frac{p(\tau|\tau')}{p(\tau'|\tau)} = 1$.

2. If v_3 is already a parent of v_4, then v_1 cannot be removed from v_3 and v_4 (otherwise v_4 will become a non-binary node). No action would be performed, and the Hastings ratio is set to $-\infty$.

3. If $\tau'_{v_1} < \tau_{v_2}$, v_2 can no longer be a child of v_1. No action would be performed, and the Hastings ratio is set to $-\infty$.

4. If $\tau'_{v_1} > \tau_{v_3}$, we trace back from v_3 to its ancestors. Similarly, if $\tau'_{v_1} < \tau_{v_4}$, we trace downwards from v_4 to its descendants. During the search, all the edges $e = (x, y)$ satisfying the condition that $\tau_y \leq \tau'_{v_1} \leq \tau_x$ and $y \neq v_2$ are collected to the edge list $\mathcal{L}'$. Note that if $\tau'_{v_1} > \tau_r$, there would be only one edge $(null, r)$ in $\mathcal{L}'$. If no edge is collected, no action would be performed, and the Hastings ratio is set to $-\infty$.

5. An edge (v_5, v_6) is randomly selected from $\mathcal{L}'$.

(a) If $v_3 \neq null$ and $v_5 \neq null$, the two edges (v_3, v_1) and (v_1, v_4) are deleted and replaced by a new edge (v_3, v_4). The edge (v_5, v_6) is then deleted, and replaced by two new edges (v_5, v_1) and (v_1, v_6).

(b) If $v_3 = null$ and $v_5 \neq r$, the edge (v_1, v_4) is deleted and v_4 becomes the new root. The edge (v_5, v_6) is then deleted and replaced by two new edges (v_5, v_1) and (v_1, v_6). The population size of the root is unchanged. The parameters of the edge (v_5, v_1) are assigned to the original parameters of the edge (v_1, v_4).

(c) If $v_3 \neq r$ and $v_5 = null$, the two edges (v_3, v_1) and (v_1, v_4) are deleted and replaced by a new edge (v_3, v_4). The edge (v_1, v_6) is then added and v_1 becomes the new root. The population size of the root is unchanged. The parameters of the edge (v_1, v_6) are assigned to the original parameters of the edge (v_3, v_1).

The time of τ_{v_1} is replaced by τ'_{v_1}. To calculate the Hastings ratio, we need to trace back or downwards from v_1 (after proposal) and collect all the edges $e = (x, y)$ satisfying the condition that $\tau_y \leq \tau_{v_1} \leq \tau_x$ into $\mathcal{L}$. The Hastings ratio in this case is $\frac{1}{|\mathcal{L}|} / \frac{1}{|\mathcal{L}'|} = \frac{|\mathcal{L}'|}{|\mathcal{L}|}$.

Move-Tail. A tree node v_1 is randomly selected from the list of internal tree nodes (root is excluded). v_2 is a child of v_1 chosen at random. Let v_3 be the parent of v_1 and v_4 be the other child of v_1.

1. If v_3 is already a parent of v_4, then v_1 cannot be removed from v_3 and v_4 (otherwise v_4 will become a non-binary node). No action would be performed, and the Hastings ratio is set to $-\infty$.

2. All the edges $e = (x, y)$ satisfying the conditions that $\tau_x > \tau_{v_2}$, $x \neq v_1$ and $y \notin \{v_1, v_2\}$ are collected. If no such edge is found, no action would be performed, and the Hastings ratio is set to $-\infty$.

3. An edge (v_5, v_6) is randomly selected from the edge list in the previous step. The two edges (v_3, v_1) and (v_1, v_4) are deleted and replaced by a new edge (v_3, v_4). The edge (v_5, v_6) is then deleted and replaced by two new edges (v_5, v_1) and (v_1, v_6). A

new time τ'_{v_1} is drawn from $\mathrm{Uniform}(\max(\tau_{v_2}, \tau_{v_6}), \tau_{v_5})$. The Hastings ratio in this case is $\frac{\Delta\tau}{\tau_{v_3} - \max(\tau_{v_2}, \tau_{v_4})} / \frac{\Delta\tau}{\tau_{v_5} - \max(\tau_{v_2}, \tau_{v_6})} = \frac{\tau_{v_5} - \max(\tau_{v_2}, \tau_{v_6})}{\tau_{v_3} - \max(\tau_{v_2}, \tau_{v_4})}$.

Move-Head. A reticulation edge (v_1, v_2) is randomly selected from the list of reticulation edges where v_2 is a network node. Let v_3 be the other parent of v_2 and v_4 be the only child of v_2.

1. If v_3 is already a parent of v_4, then v_2 cannot be removed from v_3 and v_4 (otherwise, v_4 will become a non-binary node). No action would be performed, and the Hastings ratio is set to $-\infty$.

2. The two edges (v_3, v_2) and (v_2, v_4) are deleted and replaced by a new edge (v_3, v_4). Then a new edge (v_5, v_6) is selected uniformly at random from the list of edges where each edge (x, y) satisfies $\tau_{\mathrm{y}} < \tau_{v_1}$, $y \neq v_2$ and $x \notin \{v_1, v_2\}$. The edge (v_5, v_6) is deleted and replaced by two new edges (v_5, v_2) and (v_2, v_6). For the parameters, $\gamma_{(v_3,v_2)}$ and $\theta_{(v_3,v_2)}$ no longer exist and the values are assigned to $\gamma_{(v_5,v_2)}$ and $\theta_{(v_5,v_2)}$ respectively. The new time τ'_{v_2} is drawn from $\mathrm{Uniform}(\tau_{v_6}, \min(\tau_{v_1}, \tau_{v_5}))$. The Hastings ratio in this case is $\frac{\Delta\tau}{\min(\tau_{v_1}, \tau_{v_3}) - \tau_{v_4}} / \frac{\Delta\tau}{\min(\tau_{v_1}, \tau_{v_5}) - \tau_{v_6}} = \frac{\min(\tau_{v_1}, \tau_{v_5}) - \tau_{v_6}}{\min(\tau_{v_1}, \tau_{v_3}) - \tau_{v_4}}$.

1.1.3 Change-Dimension

We first describe the Add-Reticulation and Delete-Reticulation moves, then derive the Hastings-ratios.

Add-Reticulation. Two edges $e_1 = (v_3, v_4)$ and $e_2 = (v_5, v_6)$ are selected uniformly at random from the list of edges in the network satisfying the condition that $e_2 \neq e_1$. Then e_1 is deleted and replaced by two edges $e_{11} = (v_3, v_1)$ and $e_{12} = (v_1, v_4)$. Similarly, e_2 is deleted and replaced by $e_{21} = (v_5, v_2)$ and $e_{22} = (v_2, v_6)$. The times of the two new nodes τ_{v_1} and τ_{v_2} are drawn from $\mathrm{Uniform}(\tau_{v_4}, \tau_{v_3})$ and $\mathrm{Uniform}(\tau_{v_6}, \tau_{v_5})$ respectively.

1. If $\tau_{v_1} > \tau_{v_2}$, a new edge $e_0 = (v_1, v_2)$ is added and v_2 becomes a reticulation node. γ_{e_0} is drawn from $\mathrm{Uniform}(0, 1)$ and $\gamma_{e_{21}}$ is assigned to $1 - \gamma_{e_0}$. The population sizes

$\theta_{e_{11}}, \theta_{e_{21}}$ and θ_{e_0} are drawn from $f(x)$ where

$$f(x) = \begin{cases} \frac{c}{a}x & \text{if} \quad x < a \\ c & \text{if} \quad a \leq x \leq b \\ ce^{-4c(x-b)} & \text{if} \quad x > b \end{cases}$$

$$a = \min(\theta_{e_1}, \theta_{e_2})$$

$$b = \max(\theta_{e_1}, \theta_{e_2})$$

$$c = \frac{3}{2(2b-a)}$$

is a distribution we used to sample population size (see Fig. S1). The cumulative distribution function $F(x)$ of f can be written as

$$F(x) = \begin{cases} \frac{1}{2} \cdot \frac{c}{a}x^2 & \text{if } x < a \\ \frac{1}{2} \cdot ca + c(x-a) = -\frac{1}{2}ca + cx & \text{if } a \leq x \leq b \\ \frac{1}{2} \cdot c(2b-a) + \frac{1}{4}(1 - e^{-4c(x-b)}) = 1 - \frac{1}{4}e^{-4c(x-b)} & \text{if } x > b \end{cases}$$

Let us then set a random number $u \sim \text{Uniform}(0, 1)$ equal to $F(x)$. We have

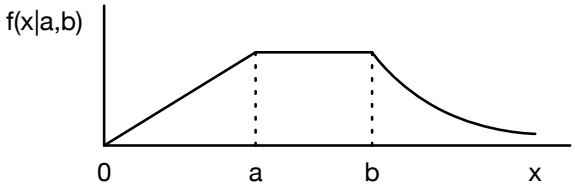

Fig. S1: **The three-piece distribution (Linear, Uniform, Exponential) for population size sampling.** Ideally, a new population size should be drawn from the prior $\Gamma(2, \psi)$. However, the inverse function Γ^{-1} cannot be solved directly. We designed the three-piece distribution as a replacement. Let the probability of sampling from $[0, b]$ be $p = 0.75$, we have $f(a|a, b) = f(b|a, b) = \frac{p}{b-0.5 \cdot a} = \frac{3}{4b-2a}$, and the Exponential parameter is $\frac{p}{(1-p)(b-0.5a)} = \frac{6}{2b-a}$.

$$x = h(u) = F^{-1}(u) = \begin{cases} \sqrt{\frac{2a}{c}u} & \text{if} \quad u < \frac{ca}{2} \\ \frac{u}{c} + \frac{a}{2} & \text{if} \quad \frac{ca}{2} \leq u \leq \frac{3}{4} \\ -\frac{1}{4c}\log(4(1-u)) + b & \text{if} \quad u > \frac{3}{4} \end{cases}$$

2. Otherwise, a new edge $e_0 = (v_2, v_1)$ is added and v_1 becomes a reticulation node. The parameter settings are similar to the previous step.

Delete-Reticulation. A reticulation edge $e_0 = (v_1, v_2)$ is selected uniformly at random from all reticulation edges.

1. If v_1 is a reticulation node or v_1 is the root, the edge e_0 cannot be removed. No action would be performed, and the Hastings ratio is set to $-\infty$.

2. Let v_3 be the parent of v_1 and v_4 be the other child of v_1, if v_4 is also a network node and v_3 is the other parent of v_4, no action would be performed, and the Hastings ratio is set to $-\infty$.

3. Let v_5 be the other parent of v_2 and v_6 be the child of v_2, if v_6 is also a network node and v_5 is the other parent of v_6, no action would be performed, and the Hastings ratio is set to $-\infty$.

4. If $v_3 = v_5$ and $v_4 = v_6$, no action would be performed, and the Hastings ratio is set to $-\infty$.

5. The edge e_0 is deleted along with the parameters. Then the two edges (v_3, v_1) and (v_1, v_4) are deleted and replaced by a new edge (v_3, v_4). Similarly, the two edges (v_5, v_2) and (v_2, v_6) are deleted and replaced by a new edge (v_5, v_6).

Hastings ratios of Change-Dimension moves. Based on the two moves we described above, we have

- The probability of selecting Add-Reticulation p_a from Change-Dimension moves is 1 when the current topology is a tree, and κ_1 otherwise.
- In Add-Reticulation, the two edges e_1 and e_2 are selected with probability $\frac{1}{|E|(|E|-1)}$.
- The Jacobian matrix of Add-Reticulation is a diagonal matrix composed of
  - the time of τ_{v_1}. Generate $u_1 \sim \mathrm{Uniform}(0, 1)$. We have $\tau_{v_1} = (\tau_{v_3} - \tau_{v_4})u_1$. The partial derivative is $\partial\tau_{v_1}/\partial u_1 = \tau_{v_3} - \tau_{v_4}$.

- the time of τ_{v_2}. Generate $u_2 \sim \text{Uniform}(0,1)$. We have $\tau_{v_2} = (\tau_{v_5} - \tau_{v_6})u_2$. The partial derivative is $\partial\tau_{v_2}/\partial u_2 = \tau_{v_5} - \tau_{v_6}$.
- the inheritance probability of e_0. Generate $u_3 \sim \text{Uniform}(0,1)$. We have $\gamma_{e_0} = u_3$. The partial derivative is $\partial\gamma_{e_0}/\partial u_3 = 1$.
- the population size of e_0. Generate $u_4 \sim \text{Uniform}(0,1)$. We have $\theta_{e_0} = h(u_4)$. The partial derivative is $h'(u_4)$ where

$$h'(u) = \partial h(u)/\partial u = \begin{cases} \sqrt{\dfrac{a}{2cu}} & \text{if} \quad u < \dfrac{ca}{2} \\ \dfrac{1}{c} & \text{if} \quad \dfrac{ca}{2} \leq u \leq \dfrac{3}{4} \\ \dfrac{1}{4c(1-u)} & \text{if} \quad u > \dfrac{3}{4} \end{cases}$$

$$a = \min(\theta_{e_1}, \theta_{e_2})$$

$$b = \max(\theta_{e_1}, \theta_{e_2})$$

$$c = \frac{3}{2(2b-a)}$$

- the population size of e_{11}. Generate $u_5 \sim \text{Uniform}(0,1)$. We have $\theta_{e_{11}} = h(u_5)$. The partial derivative is $h'(u_5)$.
- the population size of e_{21}. Generate $u_6 \sim \text{Uniform}(0,1)$. We have $\theta_{e_{21}} = h(u_6)$. The partial derivative is $h'(u_6)$.

- In summary, $|J| = (\tau_{v_3} - \tau_{v_4})(\tau_{v_5} - \tau_{v_6})h'(u_4)h'(u_5)h'(u_6)$ for Add-Reticulation.
- The probability of selecting Delete-Reticulation p_d is $1 - \kappa_1$ when the current topology is a network.
- In Delete-Reticulation, the probability of selecting edge e_0 is $\frac{1}{2|R|}$.
- The Jacobian matrix of Delete-Reticulation is also a diagonal matrix composed of
  - $u_1 = \frac{\tau_{v_1}}{\tau_{v_3} - \tau_{v_4}}$. The partial derivative is $\partial u_1/\partial\tau_{v_1} = 1/(\tau_{v_3} - \tau_{v_4})$.
  - $u_2 = \frac{\tau_{v_2}}{\tau_{v_5} - \tau_{v_6}}$. The partial derivative is $\partial u_2/\partial\tau_{v_2} = 1/(\tau_{v_5} - \tau_{v_6})$.
  - $u_3 = \gamma_{e_0}$. The partial derivative is $\partial u_3/\partial\gamma_{e_0} = 1$.

- $u_4 = F(\theta_{e_0})$. The partial derivative is $F'(\theta_{e_0}) = f(\theta_{e_0})$.
- $u_5 = F(\theta_{e_{11}})$. The partial derivative is $f(\theta_{e_{11}})$
- $u_6 = F(\theta_{e_{21}})$. The partial derivative is $f(\theta_{e_{21}})$

- In summary, $|J| = \frac{1}{\tau_{v_3} - \tau_{v_4}} \cdot \frac{1}{\tau_{v_5} - \tau_{v_6}} \cdot f(\theta_{e_0}) f(\theta_{e_{11}}) f(\theta_{e_{21}})$ for Delete-Reticulation.

The Hastings ratio of Add-Reticulation is

$$\frac{p_d}{p_a} \cdot \frac{|E|(|E|-1)}{2|R'|} \cdot (\tau_{v_3} - \tau_{v_4})(\tau_{v_5} - \tau_{v_6}) h'(u_4) h'(u_5) h'(u_6)$$

where $|R'| = |R| + 1$ is the number of reticulation nodes in the proposed network, and

$$p_d/p_a = \begin{cases} (1-\kappa)/\kappa & \text{if} \quad |R| > 0 \\ 1 - \kappa & |R| = 0 \end{cases}$$

The Hastings ratio of Delete-Reticulation is

$$\frac{p_a}{p_d} \cdot \frac{2|R|}{|E'|(|E'|-1)} \cdot \frac{1}{\tau_{v_3} - \tau_{v_4}} \cdot \frac{1}{\tau_{v_5} - \tau_{v_6}} \cdot f(\theta_{e_0}) f(\theta_{e_{11}}) f(\theta_{e_{21}})$$

where $|E'| = |E| - 3$ is the number of edges in the proposed network.

Note that if one assumes a constant population size across all branches, there is no need to sample population size parameters, then the Hastings ratio of Add-Reticulation becomes

$$\frac{p_d}{p_a} \cdot \frac{|E|(|E|-1)}{2|R'|} \cdot (\tau_{v_3} - \tau_{v_4})(\tau_{v_5} - \tau_{v_6}).$$

Similarly, the Hastings ratio of Delete-Reticulation is simplified into

$$\frac{p_a}{p_d} \cdot \frac{2|R|}{|E'|(|E'|-1)} \cdot \frac{1}{\tau_{v_3} - \tau_{v_4}} \cdot \frac{1}{\tau_{v_5} - \tau_{v_6}}.$$

1.2 Convergence diagnostics

We make use of three commonly used diagnostics:

Trace plot. A trace plot is a plot of the sampled values of a variable in an MCMC chain as a function of the number of iterations. The variable can be the posterior, the prior, or any other parameters of interest.

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 posterior values and proportions for all topologies in the 95% credible set. Similar results across the chains are desired.

Effective Sample Size. Effective Sample Size, or ESS, is the number of effectively independent draws from the posterior distribution. Adequate ESS on the posterior demonstrates good mixing.

2 Our Method vs. ∗BEAST on Data with No Reticulations

Since phylogenetic networks generalize the phylogenetic tree model, we first compared the results obtained by our implementation to those obtained by ∗BEAST on simulated data that we generated on a species tree. Here we describe the results based of one experiment, which is in addition to a different one in the main text.

2.1 Simulation settings

True species tree. The true species tree we used to generate simulated data is shown in Fig. S2.

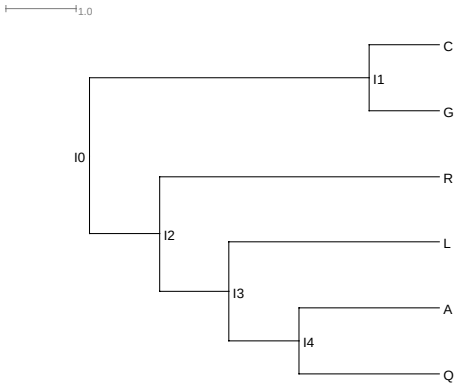

Fig. S2: **The true species tree used to generate a simulated data for testing the MCMC sampler.** The branch lengths of the species tree are measured in coalescent units.

True gene trees. We used the program ms (8) to simulate 128 gene trees given the true species tree. The command is:

ms 6 128 -T -I 6 1 1 1 1 1 1 -ej 0.5 6 5 -ej 1.0 2 1 -ej 1.5 3 1 -ej 2.0 4 1 -ej 2.5 5 1

Sequences. The program Seq-gen (11) was used to simulate sequence alignments from gene trees under the GTR model. The population mutation rate we used is $\theta = 0.036$. The length of sequences is 500. The command is:

seq-gen -m gtr -s0.018 -f0.2112, 0.2888, 0.2896, 0.2104
-r0.2173, 0.9798, 0.2575, 0.1038, 1, 0.2070 -l500

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.

Data sets. From the 128-locus data set we sampled subsets of 16, 32, 64, and 128 loci and used them in the analysis.

2.2 Results

We set the substitution model to GTR and applied the parameters we used for simulation. We assume a constant population size across all branches and the population size parameter θ is set to 0.036. Only gene trees and species tree were estimated.

Results from *BEAST. We first ran an MCMC chain of 6×10^7 iterations with 1×10^7 burn-in for each data set. One sample is collected from every $5,000$ iterations.

- 95% credible sets of species tree topologies. For all four data sets, the 95% credible sets of topologies only contain the true species tree.

- Convergence. The trace plots are shown in Fig. S3. We can see that the MCMC chains for the 16 and 32-locus data sets mix well. In the MCMC chain for the 64-locus data set, the posterior value keeps increasing until the end of the first 3×10^7

iterations. For the 128-locus data set, the posterior value keeps increasing, and the MCMC chain does not converge at the end of the chain.

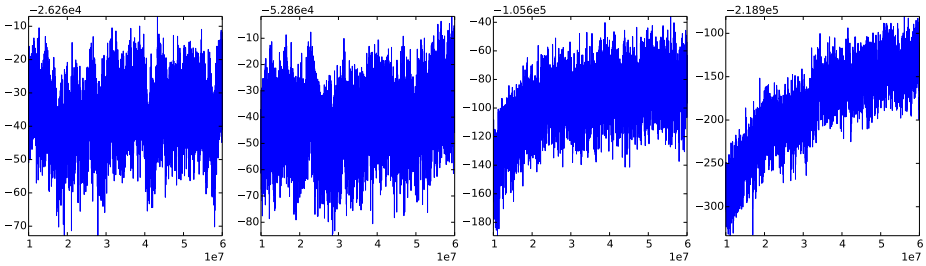

Fig. S3: **Trace plots of the MCMC chains using ∗BEAST given four data sets with 16, 32, 64, and128 loci, respectively (from left to right), simulated from the true species tree in Fig. S2.** The trace plots of the MCMC chains for 16, 32, 64-locus data sets indicate good mixing and convergence from 1×10^7, 1×10^7, and 3×10^7 iterations, respectively; however, the MCMC chain for the 128-locus data set barely converges at the end of the chain.

- The acceptance rates of the moves. ∗BEAST only implements one operation (NodeReheight move) for the species tree. The acceptance rates of that move are 0.0686, 0.0237, 0.0144, and 0.0093 for the 16, 32, 64, 128-locus data sets, respectively.

- Evaluation of gene tree and species tree samples. We used the Robinson-Foulds (RF) distance (12) to evaluate the similarity between two tree topologies. The Normalized Rooted Branch Score (nrBS) (7, 10) is used to measure the distance between the estimated tree and the true tree when accounting for both topology and divergence times.

  – We plotted the average RF distances and nrBS values between the sampled gene trees and the true gene trees for every iteration; see Fig. S4.

  – We plotted the nrBS between the sampled species tree and the true species tree for every iteration; see Fig. S5.

  The RF distances and nrBS values for both species tree and gene trees decrease as the data size increases, especially for the 128-locus data set, reflecting an improvement

in the quality of samples.

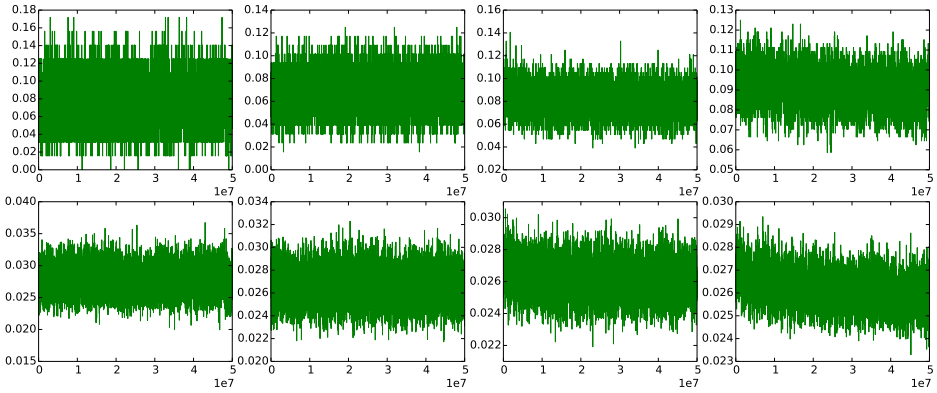

Fig. S4: **Plots of the RF distances (top row) and nrBSs (bottom row) between gene tree samples inferred by ∗BEAST and the true ones.** From left to right, the four plots correspond to the four data sets of 16, 32, 64, and 128 loci, respectively, simulated from the true species tree in Fig. S2. The RF distances and nrBS values become smaller by as the size of the data set increases, indicating more accurate estimates of topologies and divergence times.

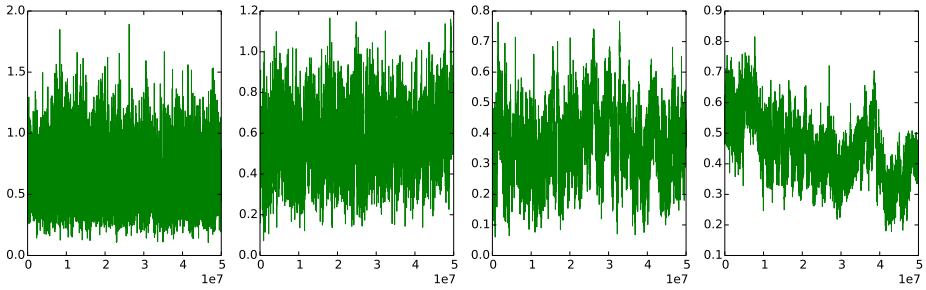

Fig. S5: **Plots of the nrBS values between the species tree sample inferred by ∗BEAST and the true species tree in Fig. S2.** The divergence times of the samples are converted to coalescent units. From left to right, the plots correspond to the four data sets (16, 32, 64, and 128 loci, respectively) simulated from the true species tree. The nrBS values become smaller as the data set size increases, indicating more accurate estimates of topologies and divergence times of the samples.

Results from our method. In this case, we did not allow adding reticulations, effectively limiting the sampling to the tree space. The settings are the same as ∗BEAST: we ran 6×10^7 iterations with 1×10^7 burn-in and collected 1 sample from every $5,000$ iterations.

- 95% credible sets of species tree topologies. For all four data sets, the 95% credible sets of topologies only contain the true species tree.
- Convergence. The trace plots are shown in Fig. S6. We can see that the MCMC chains mix well. Compared with the trace plots from *BEAST (Fig. S3), these plots are less jagged.

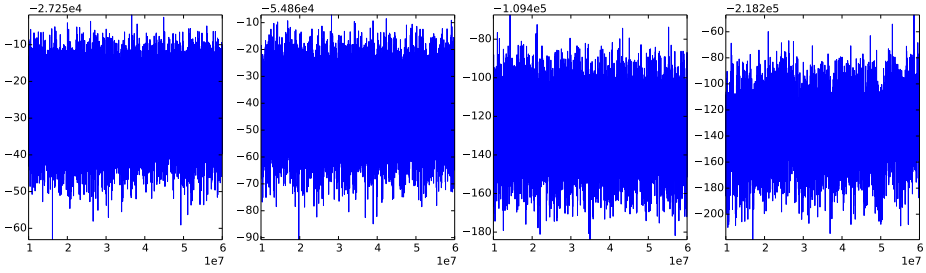

Fig. S6: **Trace plots of the MCMC chains using our method on the four data sets (16, 32, 64, and 128 loci, respectively, from left to right) simulated from the true species tree in Fig. S2.** The results indicate good mixing and convergence.

- Evaluation of gene tree and species tree samples.
  - We plotted the average RF distances and nrBS values between the sampled gene trees and the true gene trees for every iteration; see Fig. S7. The average distances for the four data sets are similar to the ones inferred by ∗BEAST (Fig. S4).
  - We plotted the nrBS values between the sampled species tree and the true species tree for every iteration; see Fig. S8. The average distances are smaller than the ones inferred by ∗BEAST (Fig. S5), especially when the data size is small.

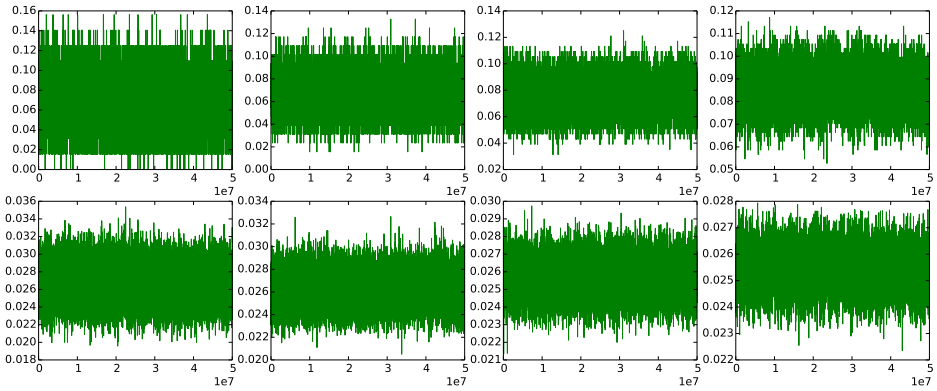

Fig. S7: **Plots of the RF distances (top row) and nrBS values (bottom row) between gene tree samples inferred by our method and the true ones.** From left to right, the plots correspond to the four data sets of 16, 32, 64, and 128 loci, respectively, simulated from the true species tree in Fig. S2. The RF distances and nrBS values become smaller as the size of the data set increases, indicating more accurate estimates of topologies and divergence times.

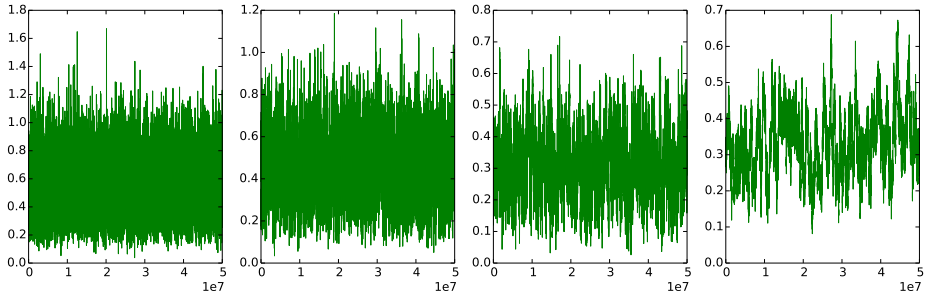

Fig. S8: **Plots of the nrBS values between the species tree sample inferred by our method and the true species tree in Fig. S2.** The divergence times of the samples are converted to coalescent units. From left to right, the plots correspond to the four data sets (16, 32, 64, 128 loci, respectively) simulated from the true species tree. The nrBS values become smaller as the data set size increases, indicating more accurate estimates of the topologies and divergence times of the samples.

3 Our Method vs. ∗BEAST on Data with Reticulations

3.1 Simulations settings

Model phylogenetic networks. We simulated data sets with 16, 32, 64, and 128 loci on each of the four phylogenetic networks shown in Fig. S9. The topologies and reticulation edges are inspired by the species phylogeny recovered from the Anopheles mosquitoes data set in (5).

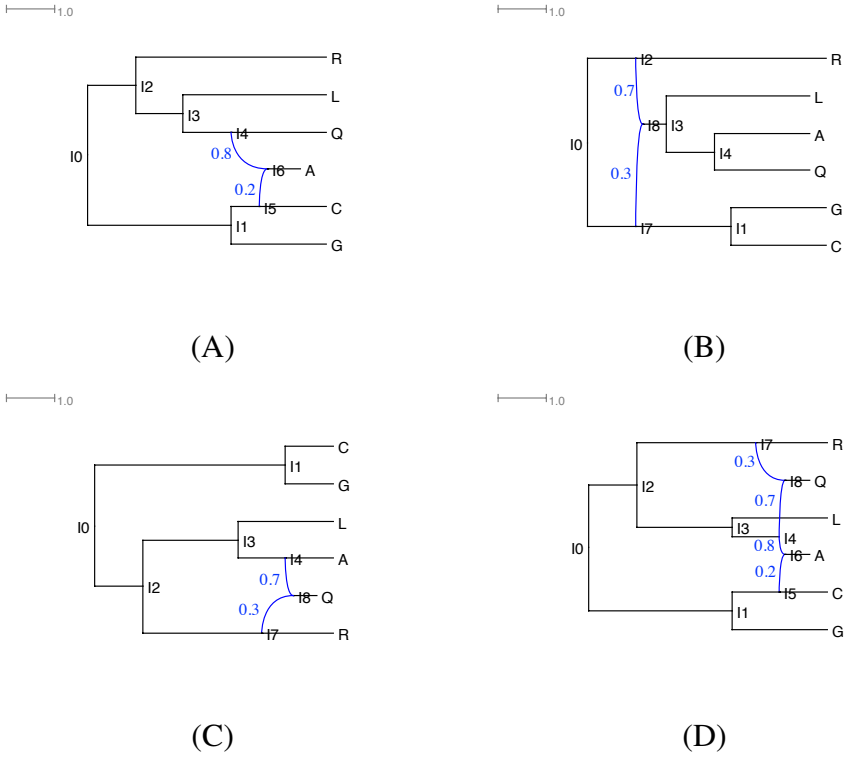

Fig. S9: **The four model phylogenetic networks used to generate the simulated data sets.** The branch lengths of the phylogenetic networks are measured in coalescent units. The inheritance probabilities are marked in blue.

Model gene trees. The program ms (8) was used to simulate 128 gene trees on each of the four model phylogenetic networks. The commands used for the phylogenetic networks in Fig. S9(A–D) are, respectively:

1. ms 6 128 -T -I 6 1 1 1 1 1 1 -es 0.35 1 0.8 -ej 0.7 6 7 -ej 1.0 7 5 -ej 1.0 2 1 -ej 1.5 3 1 -ej 2.0 4 1 -ej 2.5 5 1

2. ms 6 128 -T -I 6 1 1 1 1 1 1 -ej 1.0 6 5 -ej 1.0 2 1 -ej 1.5 3 1 -es 1.75 1 0.7 -ej 2.0 5 7 -ej 2.0 4 1 -ej 2.5 7 1

3. ms 6 128 -T -I 6 1 1 1 1 1 1 -es 0.25 1 0.7 -ej 0.5 6 5 -ej 0.5 2 1 -ej 0.75 4 7 -ej 1.0 3 1 -ej 2.0 7 1 -ej 2.5 5 1

4. ms 6 128 -T -I 6 1 1 1 1 1 1 -es 0.25 2 0.8 -es 0.25 1 0.7 -ej 0.5 5 7 -ej 0.5 2 1 -ej 0.75 4 8 -ej 1.0 6 7 -ej 1.0 3 1 -ej 2.0 8 1 -ej 2.5 7 1

Sequences. We used each of the true gene trees to simulate sequence alignments using the program Seq-gen (11) under the GTR model. We used $\theta = 0.036$ for the population mutation rate and 500 bps for the sequence length. The command is:

$$\text{seq-gen -mgtr -s}0.018 \text{ -f}0.2112, 0.2888, 0.2896, 0.2104$$
$$\text{-r}0.2173, 0.9798, 0.2575, 0.1038, 1, 0.2070 \text{ -l}500$$

where 0.2112, 0.2888, 0.2896, and 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, and 0.2070 are the relative rates of substitutions, respectively.

Data sets. For each of the phylogenetic networks, we created four sequence data sets by sampling (without replacement) randomly 16, 32, 64, and 128 loci of the full data set of 128 loci. Each of these data sets was then used as input to the methods.

3.2 Our method provides accurate estimates of the phylogenetic networks, gene trees, and their parameters

The phylogenetic network of Fig. S9(A). We ran MCMC chains of 6×10^7 iterations with 1×10^7 burn-in for the 16, 32, 64, and 128-locus data sets. One sample was collected from every 5,000 iterations.

- 95% credible sets. For all four data sets, the 95% credible sets of topologies data setonly contain the true species network.

- Convergence. The trace plots are shown in Fig. S10 and the MCMC chains mix well. All ESSs are much larger than 200, and the overall acceptance rates are in the range of $0.17 \sim 0.18$.

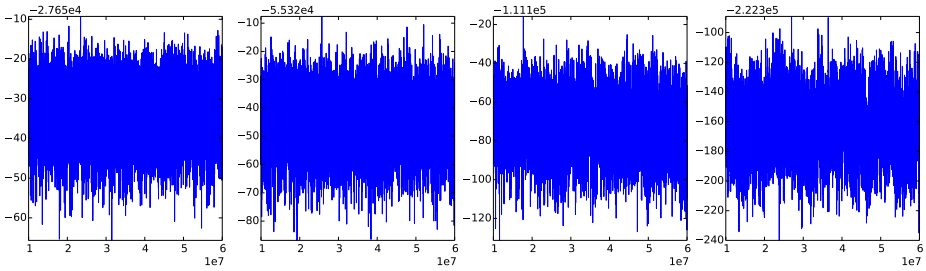

Fig. S10: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(A).** From left to right: 16, 32, 64, and 128 loci, respectively.

- Evaluation of gene tree and species tree samples.

  – We plotted the RF distances and nrBS values between the sampled gene trees and the true gene trees for every iteration in Fig. S11. As the data size increases, the average values of the RF distances and nrBS values almost stay the same, while the variations become smaller, which means the gene tree topologies and divergence times become more stable along the MCMC chain.

The phylogenetic network of Fig. S9(B). We ran 6×10^7 iterations with 1×10^7 burn-in iterations for all four data sets. One sample was collected from every 5,000 iterations.

- 95% credible sets.

  – For the 16 and 32-locus data sets, the 95% credible sets of topologies only contain the species tree backbone of the phylogenetic network (the tree shown in Fig. S2).

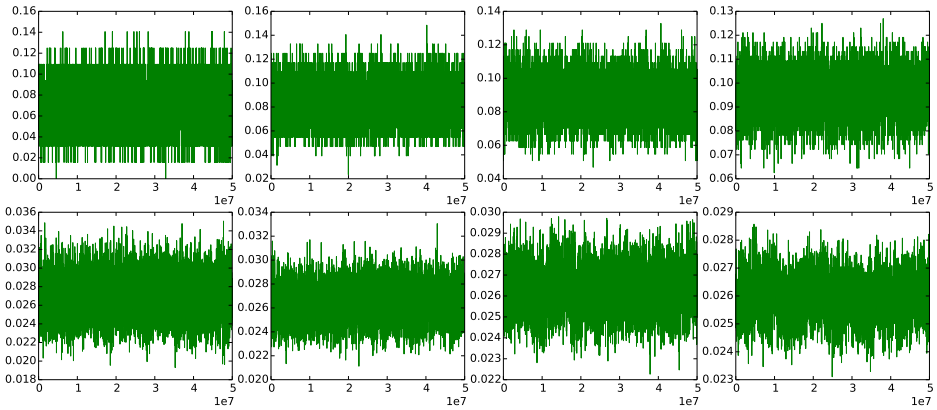

Fig. S11: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(A).** From left to right: 16, 32, 64, and 128 loci, respectively.

  - For the 64-locus data set, the first 62.2% samples are the species tree backbone and the remaining 37.8% samples are the true species network. The proportion would change if we increase the chain length.
  - For the 128-locus data set, the 95% credible set of topologies only contain the true species network.

- Convergence. The trace plots are shown in Fig. S12. All plots display good mixing except the one from the 64-locus data set. We can clearly see at around iteration 3×10^7, there is a jump in the posterior value; in fact, at iteration $28,905,000$ the chain started sampling the true network instead of the species tree backbone. All ESSs are much larger than 200 except for the one from the 64-locus data set. The overall acceptance rates are in the range of $0.15 \sim 0.18$.

- Evaluation of gene tree and species tree samples.
  - We plotted the RF distances and nrBS values between the sampled gene trees and the true gene trees for every iteration in Fig. S13. As the data size increases, the average values of the RF distances and nrBS values decrease, and the vari-

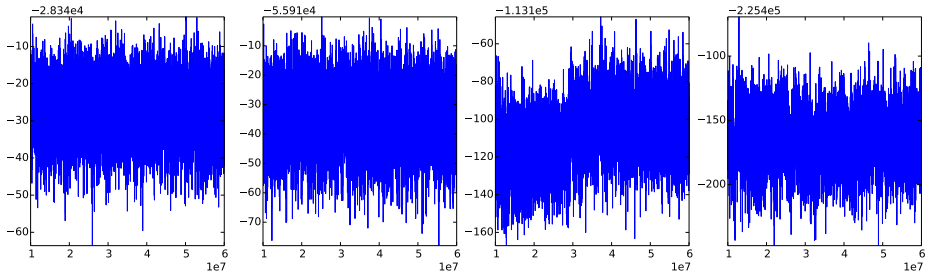

Fig. S12: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(B).** From left to right: 16, 32, 64, and 128 loci, respectively.

ations become smaller, which means the gene tree topologies and divergence times become more accurate and more stable along the MCMC chain.

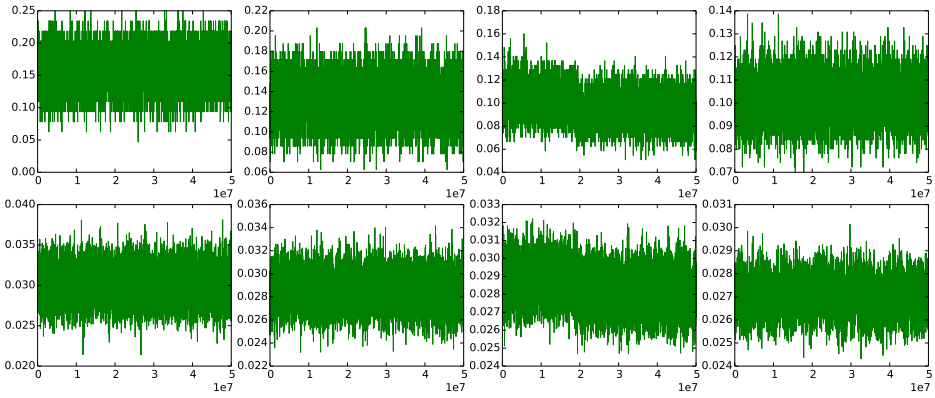

Fig. S13: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(B).** From left to right: 16, 32, 64, and 128 loci, respectively.

In this network, this introgression happened near the root of the phylogenetic network, so the likelihood of a model involving hybridization is not significantly better than that of a treelike model that explains all heterogeneity across loci in terms of incomplete lineage sorting, especially for smaller numbers of loci. In this case, detecting the hybridization event requires a larger number of loci.

The phylogenetic network of Fig. S9(C). We ran 6×10^7 iterations with 1×10^7 burn-in iterations for all four data sets. One sample was collected from every 5,000 iterations.

- 95% credible sets. For all four data sets, the main topology sampled is the true species network.

- Convergence. The trace plots are shown in Fig. S14. All ESSs are much larger than 200. The overall acceptance rates are in the range of $0.16 \sim 0.17$.

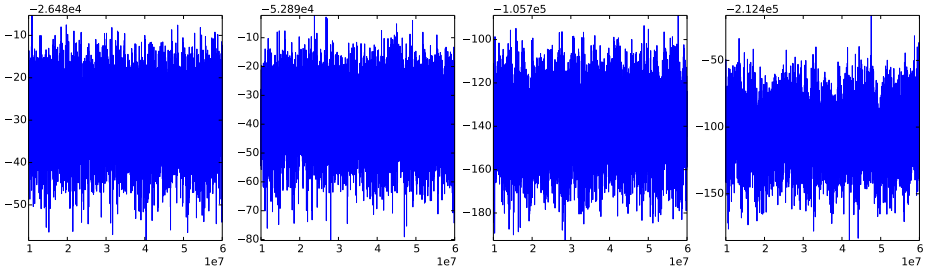

Fig. S14: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(C).** From left to right: 16, 32, 64, and 128 loci, respectively.

- Evaluation of gene tree and species tree samples.

  – We plotted the RF distances and nrBS values between the sampled gene trees and the true gene trees for every iteration in Fig. S15. As the data size increases, the average values of the RF distances and nrBS values decrease, and the variations become smaller, which means the gene tree topologies and divergence times become more accurate and more stable along the MCMC chain.

The phylogenetic network of Fig. S9(D). We ran 6×10^7 iterations with 1×10^7 burn-in iterations for all four data sets. One sample was collected from every 5,000 iterations.

- 95% credible sets. For all four data sets, the 95% credible sets of topologies only contain the true species network.

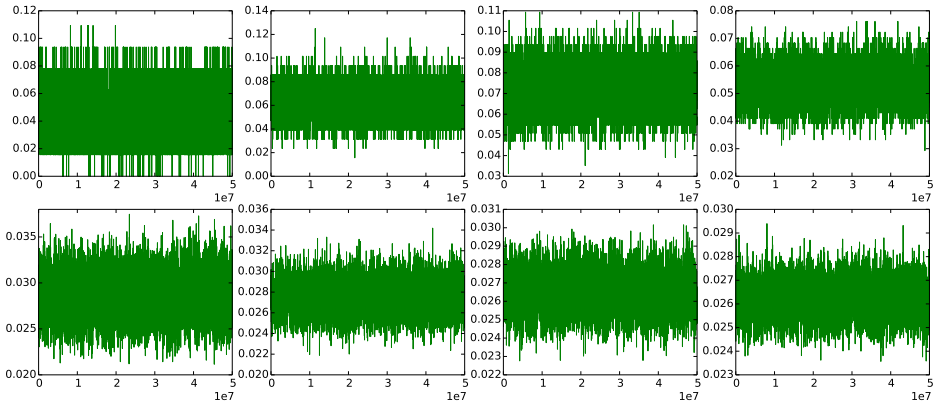

Fig. S15: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(C).** From left to right: 16, 32, 64, and 128 loci, respectively.

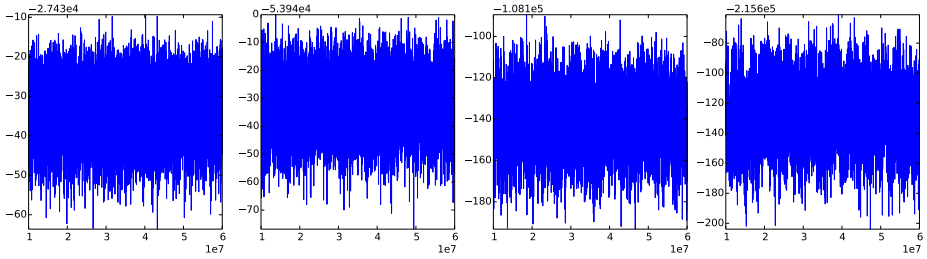

Fig. S16: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(D).** From left to right: 16, 32, 64, and 128 loci, respectively.

- Convergence. The trace plots are shown in Fig. S16. All ESSs are much larger than 200. The overall acceptance rates are in the range of $0.16 \sim 0.18$.

- Evaluation of gene tree and species tree samples.

  – We plotted the RF distances and nrBS values between the sampled gene trees and the true gene trees for every iteration in Fig. S15. As the data size increases, the average values of the RF distances and nrBS values decrease, and the variations become smaller, which means the gene tree topologies and divergence

times become more accurate and more stable along the MCMC chain.

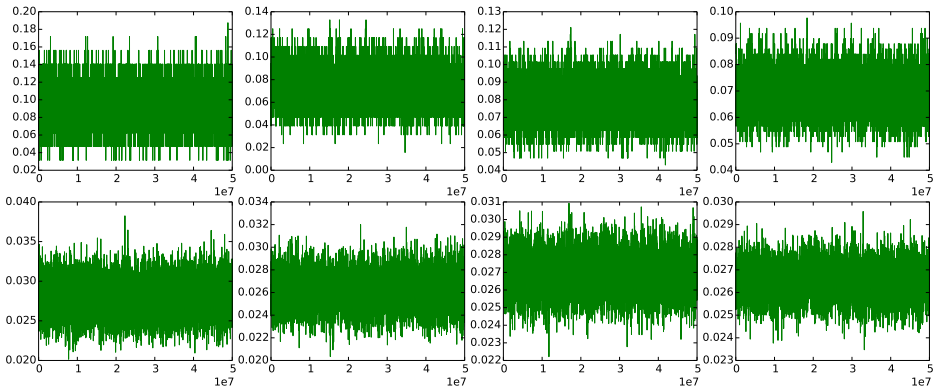

Fig. S17: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(D).** From left to right: 16, 32, 64, and 128 loci, respectively.

3.3 ∗BEAST underestimates divergence times and overestimates coalescent times when the evolutionary history is reticulate

We ran an MCMC chain of 6×10^7 iterations with 1×10^7 burn-in on the 128-locus data set simulated from the phylogenetic network of Fig. S9(D) using ∗BEAST. The resulting trace plot, shown in Fig. S18, indicates good convergence and mixing.

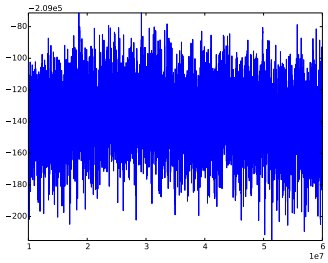

Fig. S18: **Trace plot of the MCMC chain using ∗BEAST on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).**

We considered two hypotheses:

1. the species tree topologies inferred by ∗BEAST are the ones embedded in the true network.

2. the gene trees inferred from our method are more accurate than the ones inferred from ∗BEAST since ∗BEAST forces the evolutionary history to be a tree.

To explore these hypotheses, we looked at multiple lines of evidence.

- The 95% set of species phylogenies. The 95% credible set inferred by ∗BEAST contains two topologies (Fig. S19) with proportions 94% and 4%. The MAP (maxi-

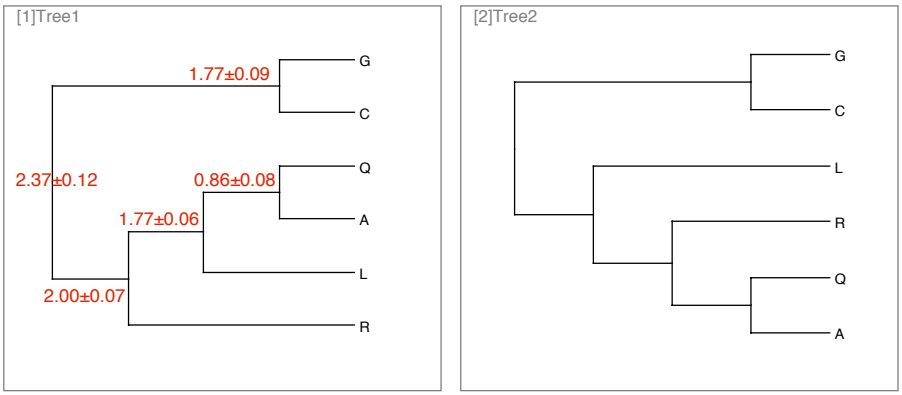

Fig. S19: **The two trees in the 95% credible set obtained by ∗BEAST on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).** The proportions of the two sampled species tree topologies are 94% (for the topology of Tree 1) and 4% (for the topology of Tree 2). The MAP topology (Tree 1) can be embedded into the true phylogenetic network (that is, the true phylogenetic network could be obtained by adding horizontal edges to Tree 1). Divergence times, in coalescent units, of the MAP topology are marked in red.

  mum a posteriori) topology can be embedded in the network inferred by our method (which is the true network). The divergence times in coalescent units of the MAP topology are marked in red. Comparing to the divergence times in the true network (Fig. S20) and the times estimated by our method (Fig. S21), the times from ∗BEAST are significantly underestimated. For instance, the true divergence times of

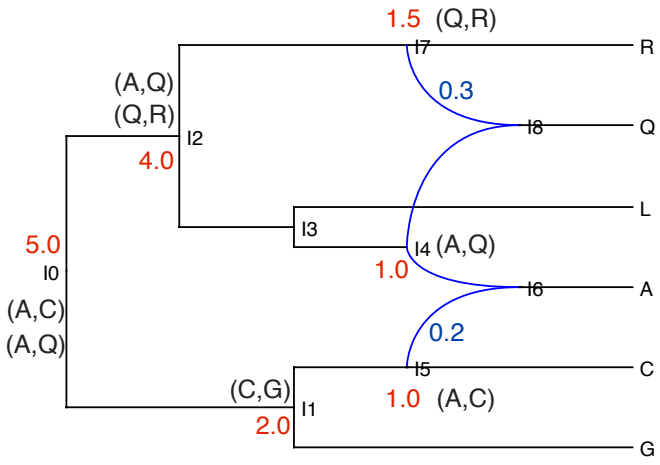

Fig. S20: **The true phylogenetic network used to simulate the 128-locus data set.** The divergence times in coalescent units are shown in red. The inheritance probabilities associated with the two reticulation edges are shown in blue. Node I1 is the MRCA of (C,G). The MRCA of (A,Q) could be one of the three nodes $I4$, $I2$, and the root $I0$, depending on which two of the four reticulation edges are "used" by the coalescent history of a given locus. The MRCA of (A,C) could be node $I5$ or the root, depending on whether reticulation edge $(I5, I6)$ is used or not, respectively. The MRCA of (R,Q) could be node $I7$ or node $I2$, depending on whether reticulation edge $(I7, I8)$ is used or not, respectively.

the root is 5.0, and the estimated value is around 4.88 from our method; however, the average value from $*$BEAST is only 2.37.

- Plots of the RF distances and nrBS values between the sampled gene trees and the original true gene trees. The range of RF distances, $[0.07, 0.11]$ from $*$BEAST (Fig. S22) is larger than $[0.05, 0.09]$ from our method (Fig. S17); and the range of nrBS values, $[0.030, 0.035]$ from $*$BEAST is larger than $[0.025, 0.028]$ from our method. These numbers indicate the gene trees inferred by our method are more accurate in both topology and divergence times.

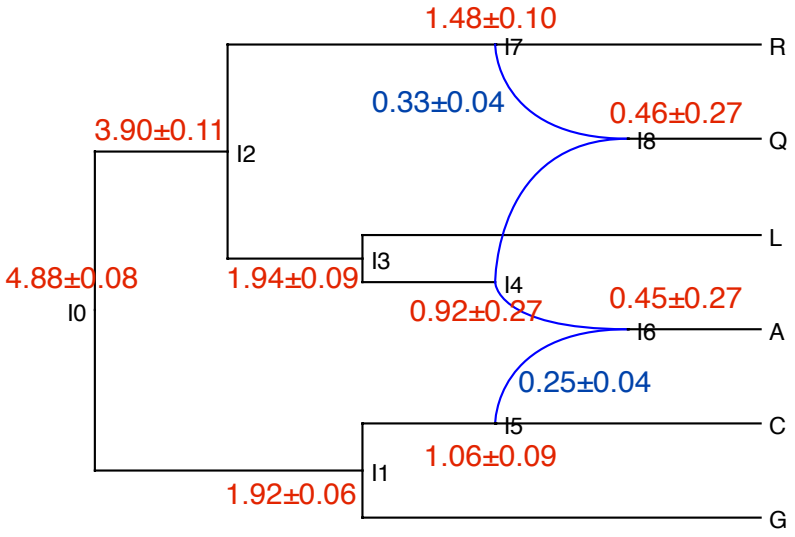

Fig. S21: **The 95% credible set obtained by our method on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).** The single topology in the 95% credible set is the true network. The divergence times in coalescent units are shown in red and the inheritance probabilities are shown in blue.

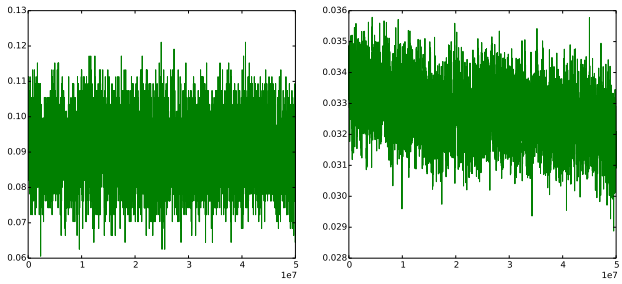

Fig. S22: **Plots of RF distances (left) and nrBS values (right) between gene trees sampled by ∗BEAST and the true ones.** The input is the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).

- Plots of divergence times. We plotted the divergence times of the most recent common ancestors (MRCAs) of (C,G), (A,Q), (A,C), (Q,R) from gene trees inferred by ∗BEAST (green) and our method (blue) in Fig. S23. We scaled the divergence times into coalescent units by dividing $\theta/2 = 0.018$ for comparison purposes. The diver-

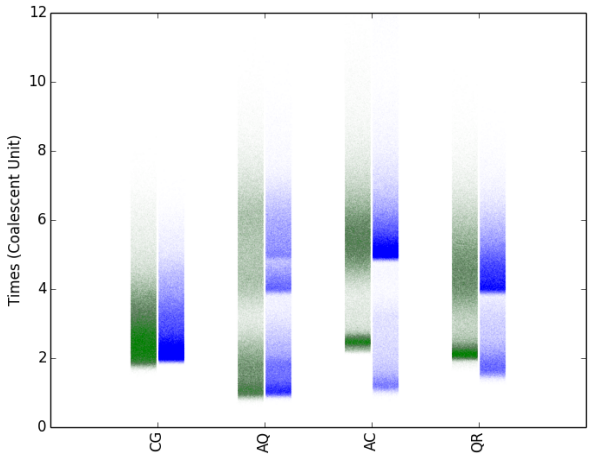

Fig. S23: **The divergence times in coalescent units of the MRCA of (C,G), (A,Q), (A,C), (Q,R) from co-estimated gene trees inferred by ∗BEAST (green) and our method (blue).** The input is the 128-locus data set simulated from phylogenetic network of Fig. S9(D).

gence times provided by the true network, 2.0, 1.0, 1.0, and 1.5 for (C,G), (A,Q), (A,C), and (Q,R), respectively (marked in red in Fig. S20), serve as the temporal constraints, or low bounds for the time estimates of gene trees. We compare the lower bound from the true network with the time estimates of the gene tree samples for (C,G), (A,Q), (A,C), and (Q,R), respectively.

- (C,G): the minimum times inferred by ∗BEAST and our method are both close to the lower bound of 2.0. The variation of samples from our method is smaller.
- (A,Q): the minimum times inferred by ∗BEAST and our method are both close to the lower bound of 1.0. Note that if the edge $(I4, I8)$ in Fig. S20 is removed, the time of MRCA of (A,Q)—node I2—is 4.0; if $(I4, I6)$ is removed, the time of MRCA of (A,Q)—node I0—is 5.0. We can see three groups of divergence times grouped around the values of 1.0, 4.0, and 5.0 obtained by our method. The time samples obtained by ∗BEAST are almost evenly distributed.
- (A,C) and (Q,R): the minimum times inferred from our method are lower and more accurate. Similar to results for (A,Q), we can see two groups of diver-

gence times obtained by our method, depending on which reticulation edge was "used" by the coalescent history of the individual loci.

3.4 Simultaneous inference of phylogenetic networks and gene trees provides more accurate gene trees than gene trees estimated from individual loci

We used RAxML (14) to construct 100 bootstrap trees for each locus in the 128-locus data set simulated on the phylogenetic network of Fig. S9(D). We computed the average RF-distance between bootstrap trees and true gene trees for all loci. The value is 0.099, which is greater than 0.09 and 0.07 calculated from samples inferred by $*$BEAST (Fig. S22) and our method (Fig. S17), respectively.

3.5 Inference from gene tree estimates requires more data than inference from sequences directly

We fed the true gene trees of the four data sets (16, 32, 64, and 128 loci) generated from the phylogenetic network of Fig. S9(D) to the Bayesian inference method of (16), which infers phylogenetic networks and inheritance probabilities given gene tree topologies (command `MCMC_GT` in PhyloNet (15)). We ran $5,050,000$ iterations with $50,000$ burn-in and sampled every 1,000 iterations. The five network topologies sampled are shown in Fig. S24.

- For the 16-locus data set, the 95% credible set contains 0% true network, 75.8% 1-reticulation network, and 20.1% other networks.
- For the 32-locus data set, the 95% credible set contains 39.1% true network, 51.2% 1-reticulation network, and 5.6% other networks.
- For the 64-locus data set, the 95% credible set contains 44.9% true network, 50.2% 1-reticulation network, and 3.0% other networks.

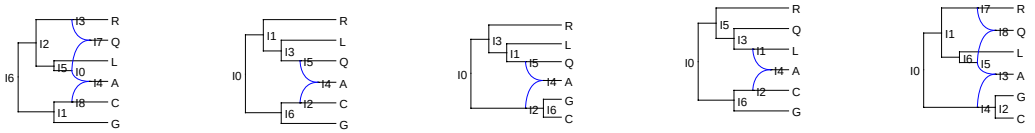

Fig. S24: **The five network topologies sampled using the method (16) on the true gene trees from the four data sets (16, 32, 64, and 128 loci) simulated on the phylogenetic network of Fig. S9(D).** Left to right: the true network, the 1-reticulation network missing the reticulation edge $R \to Q$, and the other three networks similar to the true network or the 1-reticulation network.

- For the 128-locus data set, the 95% credible set contains 60.2% true network, 34.6% 1-reticulation network, and 2.8% other networks.

The proportions of the true network in the samples are 0, 39.1%, 44.9%, and 60.2% for data sets with 16, 32, 64 and 128 gene tree topologies, respectively. Besides the true network, the 95% credible set contains several topologies that are similar to the true one. Inference using the sequence data, obtained by our new method that is reported on here, requires fewer loci to obtain comparable or more accurate results.

4 Simulations under Intermixture/Gene Flow Models

Fig. S25 shows the six phylogenetic networks we used to generate data.

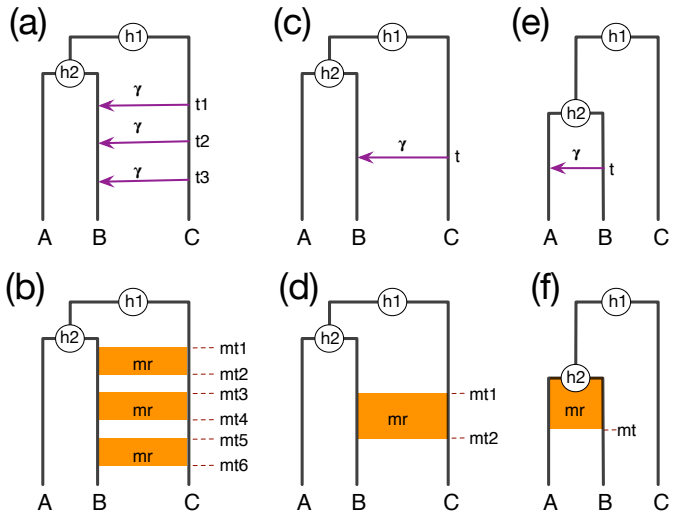

Fig. S25: True phylogenetic histories with intermixture and gene flow models. Recurrent reticulations between non-sister taxa (a,b), a single reticulation between non-sister taxa (c,d), and a single reticulation between sister taxa (e,f) is captured under both the intermixture model (top) and gene flow model (bottom). Parameters h_1 and h_2 denote divergence times (in coalescent units), t_i parameters denote intermixture times, mt_i parameters denote start/end of migration epochs, γ is the inheritance probability, and mr is the migration rate.

Model gene trees. For each simulation setting, we simulated 20 data sets with 200 1-kb loci. The program ms (8) was used to simulate 200 gene trees on each dataset. The commands used are listed as follows.

S25(a) $\Delta t = 1$: ms 3 200 -T -I 3 1 1 1 -es 1.0 2 0.2 -ej 1.0 2 1 -es 1.5 4 0.2 -ej 1.5 4 1 -es 2.0 5 0.8 -ej 2.0 6 1 -ej 3.0 3 5 -ej 4.5 5 1

S25(a) $\Delta t = 2$: ms 3 200 -T -I 3 1 1 1 -es 0.5 2 0.2 -ej 0.5 2 1 -es 1.5 4 0.2 -ej 1.5 4 1 -es 2.5 5 0.8 -ej 2.5 6 1 -ej 3.0 3 5 -ej 4.5 5 1

S25(a) $\Delta t = 3$: ms 3 200 -T -I 3 1 1 1 -es 0.0 2 0.2 -ej 0.0 2 1 -es 1.5 4 0.2 -ej 1.5 4 1 -es 3.0 5 0.8 -ej 3.0 3 5 -ej 3.0 6 1 -ej 4.5 5 1

S25(b) $\Delta mt = 1$: ms 3 200 -T -I 3 1 1 1 -em 0.0 2 1 0.4 -em 0.5 2 1 0.0 -em 1.25 2 1 0.4 -em 1.75 2 1 0.0 -em 2.5 2 1 0.4 -em 3 2 1 0.0 -ej 3 3 2 -ej 4.5 2 1

S25(b) $\Delta mt = 2$: ms 3 200 -T -I 3 1 1 1 -em 0.0 2 1 0.2 -em 3 2 1 0.0 -ej 3 3 2 -ej 4.5 2 1

S25(c) $t = 1$: ms 3 200 -T -I 3 1 1 1 -es 0.5 2 0.8 -ej 0.5 4 1 -ej 0.75 3 2 -ej 1.25 2 1

S25(c) $t = 0$: ms 3 200 -T -I 3 1 1 1 -es 0.0 2 0.8 -ej 0.0 4 1 -ej 0.75 3 2 -ej 1.25 2 1

S25(d) $mt_2 = 1.5$: ms 3 200 -T -I 3 1 1 1 -em 0.0 2 1 0.0 -em 0.5 2 1 0.8 -em 0.75 2 1 0.0 -ej 0.75 3 2 -ej 1.25 2 1

S25(e) $t = 1$: ms 3 200 -T -I 3 1 1 1 -es 0.5 1 0.8 -ej 0.5 4 2 -ej 0.75 2 1 -ej 1.25 3 1

S25(e) $t = 0$: ms 3 200 -T -I 3 1 1 1 -es 0.0 1 0.8 -ej 0.0 4 2 -ej 0.75 2 1 -ej 1.25 3 1

S25(f) $mt = 1$: ms 3 200 -T -I 3 1 1 1 -em 0.0 3 2 0.0 -em 0.5 3 2 0.8 -em 0.75 3 2 0.0 -ej 0.75 3 2 -ej 1.25 2 1

S25(f) $mt = 0$: ms 3 200 -T -I 3 1 1 1 -em 0.0 3 2 0.2666667 -em 0.75 3 2 0.0 -ej 0.75 3 2 -ej 1.25 2 1

Sequences. The program Seq-gen (Rambaut and Grassly 11) was used to generate sequence alignments down the gene trees under the Jukes-Cantor model. Sequence alignments were generated with length of 1000 sites. The command is:

seq-gen -m HKY -l 1000 -s 0.01

4.1 MCMC settings

For each data set, we ran an MCMC chain of 8×10^6 iterations with 1×10^6 burn-in. One sample was collected from every 5,000 iterations, resulting in a total of 1,400 collected samples.

We summarized the results based on 20,000 samples from 20 replicates for each of the 36 simulation settings (four values of s, three sequence lengths, and three numbers of loci).

4.1.1 The effect of the number of individuals

To study the effect of the number of individuals in the inference, we varied the number of individuals sampled from species B (we sampled 1, 3, and 5 individuals) given the true species phylogeny in Fig. S25(a).

Table S3 shows the population mutation rates, divergence times, and numbers of reticulations estimated by our method on data generated under the models of Fig. S25(a) with varying number of individuals sampled from species B. As the results show, the method performs very well in terms of estimating the divergence times and population mutation rates.

Table S3: Estimated population mutation rates (θ), divergence times (h_1 and h_2), and numbers of reticulations (#reti) as a function of varying Δt and varying number of individuals sampled from species B in the model of Fig. S25(a). The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2 = 0.01$.

Case	θ	h_1	h_2	#reti
$\Delta t = 1$, #3	$2.0 \pm 0.1e^{-2}$	9.0 ± 0.1	6.0 ± 0.1	1.8 ± 0.4
$\Delta t = 1$, #5	$2.0 \pm 0.1e^{-2}$	9.0 ± 0.1	6.0 ± 0.1	1.8 ± 0.4
$\Delta t = 2$, #3	$2.0 \pm 0.1e^{-2}$	9.0 ± 0.1	6.0 ± 0.1	2.1 ± 0.3
$\Delta t = 2$, #5	$2.1 \pm 0.1e^{-2}$	9.0 ± 0.1	6.0 ± 0.1	2.2 ± 0.4

As for the estimated number of reticulations, we found when $\Delta t = 1$, increasing the number of individuals from 1 to 3 leads to a increase in the number of reticulations (from 1.2 ± 0.4 to 1.8 ± 0.4). However, increase the number of individuals from 3 to 5 does not change the inference significantly. When $\Delta t = 2$ and the number of individuals in species B is 1, the estimated number of reticulations is 2.0 ± 0.0, while increase the number of individuals to 3 or 5, the number of reticulations only increased slightly.

4.2 Paraphyletic intermixture/gene flow

To assess the performance of our method on the simpler case of a single reticulation event, we considered the networks in Fig. S25(c) and Fig. S25(d), set $h_1 = 2.5$, $h_2 = 1.5$, and $mt_1 = h_2$, and varied $t, mt_2 \in \{1, 0\}$. Results are in Table S4 and Fig. S26.

Table S4: Estimated population mutation rates (θ), divergence times (h_1 and h_2), intermixture/migration time (t/mt), inheritance/migration rates, and numbers of reticulations (#reti) as a function of varying t in the model of Fig. S25(c) and mt_2 in the model of Fig. S25(d). The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2 = 0.01$.

Case	θ	h_1	h_2	t/mt	γ (mr)	#reti
$t = 1$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	1.02 ± 0.15	0.20 ± 0.05	1.0 ± 0.0
$t = 0$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	0.06 ± 0.04	0.21 ± 0.04	1.0 ± 0.0
$mt_2 = 1$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	1.14 ± 0.16	0.18 ± 0.05	1.0 ± 0.0
$mt_2 = 0$	$2.2 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	0.41 ± 0.16	0.17 ± 0.04	1.0 ± 0.0

4.3 Isolation-migration between sister species

We assessed the performance of our method on cases where the reticulation event involves sister taxa. Fig. S25(e) and Fig. S25(f) show the cases we considered, with setting $h_1 = 2.5$ and $h_2 = 1.5$, and varying $t, mt \in \{1, 0\}$. Results are in Table S5 and Fig. S27.

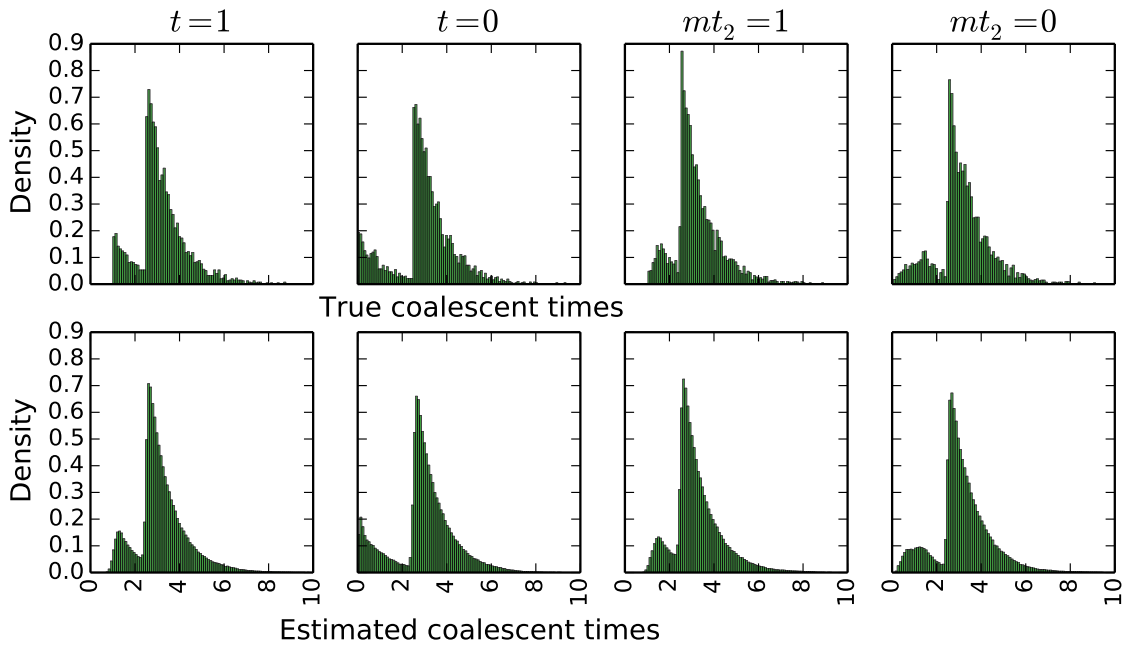

Fig. S26: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on $4,000$ loci generated under the models of Fig. S25(c) and Fig. S25(d).

Table S5: Estimated population mutation rates (θ), divergence times (h_1 and h_2), intermixture/migration time (t/mt), inheritance/migration rates, and numbers of reticulations (#reti) as a function of varying t in the model of Fig. S25(e) and mt in the model of Fig. S25(f). The times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2 = 0.01$.

Case	θ	h_1	h_2	t/mt	γ	#reti
$t=1$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.3 ± 0.1	NA	NA	0.0 ± 0.0
$t=0$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.0	0.07 ± 0.04	0.21 ± 0.06	1.0 ± 0.0
$mt=1$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.4 ± 0.1	NA	NA	0.0 ± 0.0
$mt=0$	$2.2 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	0.30 ± 0.18	0.11 ± 0.06	1.0 ± 0.0

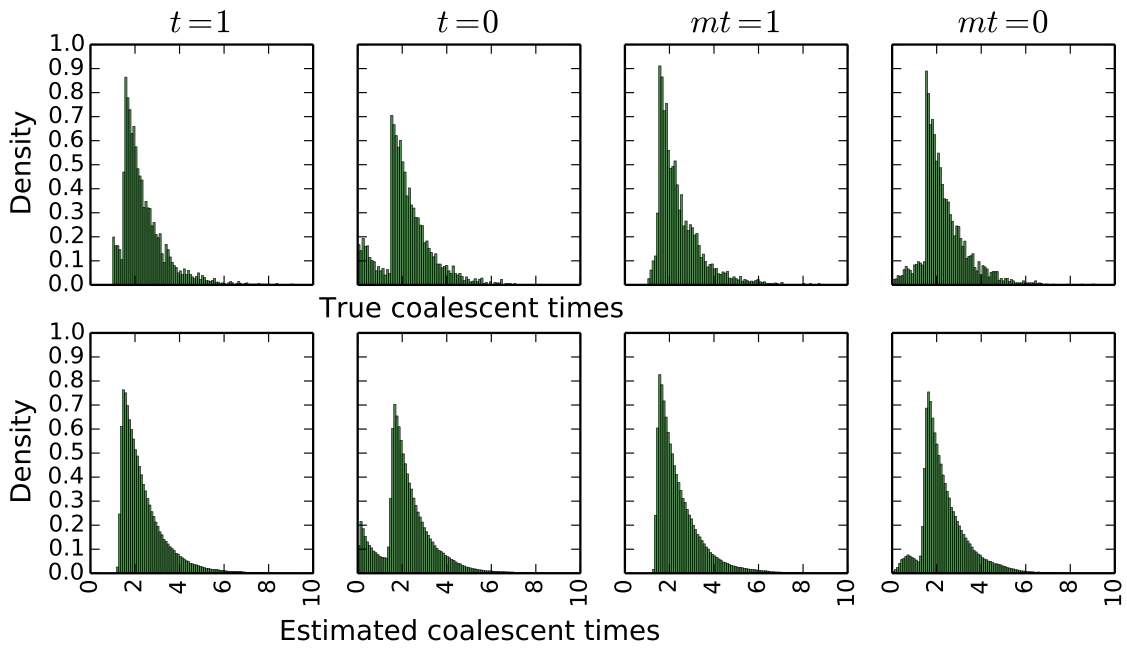

Fig. S27: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on $4,000$ loci generated under the models of Fig. S25(e) and Fig. S25(f).

5 Analysis of a Yeast Data Set

Rokas et al. (13) reported on extensive incongruence of single-gene phylogenies of seven Saccharomyces species, *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae* (Smik), *S. kudriavzevii* (Skud), *S. bayanus* (Sbay), *S. castellii* (Scas), *S. kluyveri* (Sklu). The data set consists of 106 loci of the seven species, and fungus *Candida albicans* (Calb) serves as the outgroup. They revealed the species tree from concatenation method shown in Fig. S28 (left). Edwards *et al.* (4) reported two main gene tree / species tree topologies sam-

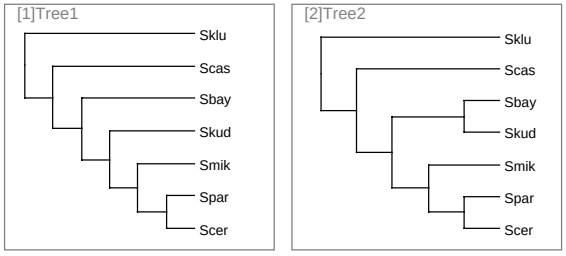

Fig. S28: **The species trees of seven Saccharomyces species.** (Left) The topology inferred from concatenation method (13) and the main topology sampled by BEST (4). (Right) The topology sampled by BEST with the second highest proportion (4).

pled from BEST, a multispecies coalescent Bayesian inference method, as shown in Fig. S28. Although the two species trees support $(Sklu, (Scas, ...))$, other gene tree topologies (Fig. S29) sampled from BEST indicate the weak phylogenetic signal for resolving the relationship of Sklu and Scas to the five other species.

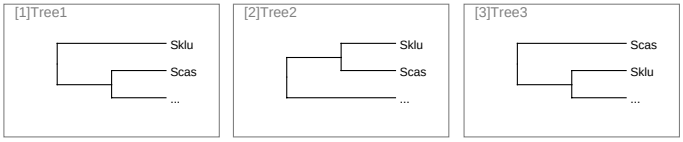

Fig. S29: **Relationships of Sklu and Scas in several gene tree topologies of seven Saccharomyces species.**

Bloomquist and Suchard (2) revisited the data set and studied the ancestral recombination graphs (ARGs) from the data set via Bayesian inference approach. They removed

Sklu from the data set as it presents a noisier signal with Scas. Their approach keeps adding non-vertical events (introgressions) between Scas and the rest species because the lineage specific rate variation in Scas are much stronger compared to the remaining species. They did not report the number of non-vertical events, the topologies, or the parameter values. In terms of gene trees, they stated that 31 and 75 genes support the trees in Fig. S28(left) and Fig. S28(right).

Yu *et al.* (17) focused on the five species Scer, Spar, Smik, Skud, and Sbay, and analyzed the data set using a parsimonious inference approach. The maximum parsimony phylogenetic network with 1 reticulation supports $Skud \rightarrow Sbay$ with inheritance probability of 0.38 (see Fig. 8 in (17)).

We reanalyzed the data set using our new method, as we now describe.

5.1 MCMC settings

We used the Jukes-Cantor substitution model (9). We assumed a constant population size θ across all branch of the species network ($\theta \sim \Gamma(2, \psi)$, ψ is a hyper-parameter sampled from non-informative prior $p_\psi(x) = 1/x$).

We employed Metropolis-coupled MCMC (MC3) (1) to help the sampler traverse the posterior landscape as follows:

- Number of MC3 chains: three (one cold chain, two heated chains);
- Temperature settings: 1 (cold chain), 2, 4 (heated chains);
- Swap frequency: considers swapping states of two random chains once every 100 iterations.

5.2 Data preprocessing

We downloaded the 106 gene sequence alignments of seven Saccharomyces species from the website of Rokas Lab. The sequence lengths of the individual loci varied between 390 and 2994 bps (in the sequence alignments).

5.3 Results for the full data set

For the yeast data set of 106 loci from seven Saccharomyces species, we ran three MC3 chains with 3.5×10^7 iterations with 1×10^7 burn-in. One sample was collected from every 5,000 iterations.

- 95% credible sets of the phylogenetic network topologies. The 95% credible sets contains 12 topologies, the main three topologies are shown in Fig. S30.

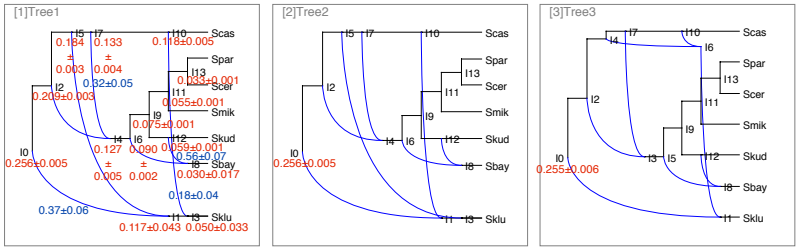

Fig. S30: **The three main phylogenetic networks in the 95% credible sets from the yeast data set using our method.** The divergence times are labeled in red, and the inheritance probabilities are marked in blue.

- Convergence and mixing. The trace plots shown in Fig. S31 indicate good convergence and mixing. The states across MC3 chains are slightly inconsistent in terms

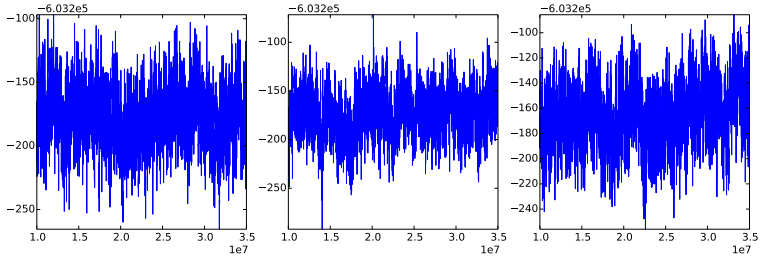

Fig. S31: **Trace plots of MC3 chains using our method on the yeast data set.**

of the range of the posterior values, while the average values are similar. It is difficult for the MCMC sampler to explore the spaces of phylogenetic networks with four or more reticulations, since there are many topologies with similar hybridization patterns but different orders, as shown in Fig. S30.

We fed the data set into ∗BEAST for comparison. We ran an MCMC chain of 3.5×10^7 iterations with 1×10^7 burn-in. One sample was collected from every 5,000 iterations.

From the densiTree plot of the species trees sampled from ∗BEAST in Fig. S32, we can see the phylogenetic signals among Scas, Sklu and the other 5 species are low.

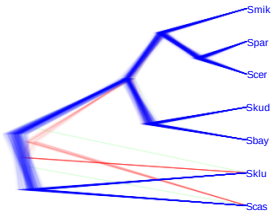

Fig. S32: **The densiTree plot of the species trees sampled from ∗BEAST given the yeast data set.**

The 95% credible set (Fig. S33) contains two topologies that can be embedded into the networks inferred by our program. The divergence time of the root 0.126 ± 0.003 obtained by ∗BEAST is much lower compared to 0.256 ± 0.005 inferred by our method.

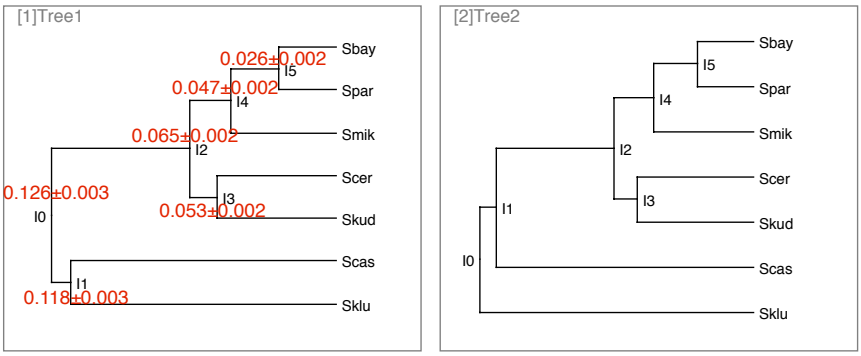

Fig. S33: **The two main species trees in the 95% credible set from the yeast data set using ∗BEAST.** The proportions for Tree 1 and Tree 2 are 78.4% and 16.9%, respectively. The divergence times are marked in red.

We plotted the divergence times of the MRCAs of (Sbay,Skub), (Scas,Sklu), (Scer,Spar), and (Scas,Spar) from gene tree samples inferred by ∗BEAST (green) and our method (blue) in Fig. S34. The ranges of the divergence times obtained by ∗BEAST and our method are similar.

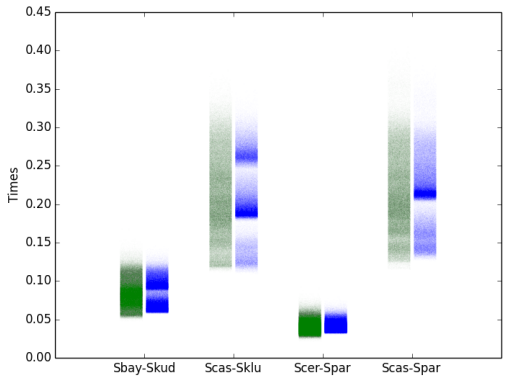

Fig. S34: **The divergence times of the MRCAs of (Sbay,Skub), (Scas,Sklu), (Scer,Spar), and (Scas,Spar) from estimated gene tree samples inferred by ∗BEAST (green), and our method (blue).** The input is the full yeast data set.

Analyzing the dataset using gene tree topologies as input We revisited the dataset via the Bayesian inference method (16) taking gene tree topologies as input. We ran two MCMC chains with 1.1×10^6 iterations with 1×10^5 burn-in. One sample was collected from every 1,000 iterations.

The 95% credible sets contain only one topology, as shown in Fig. S35. The inheritance

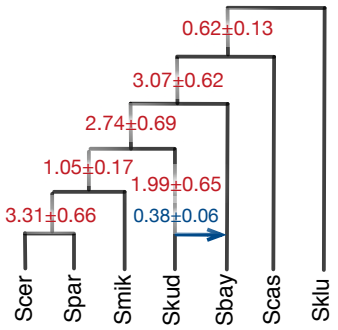

Fig. S35: **The phylogenetic network in the 95% credible set from the data set of 106 loci from seven Saccharomyces species using the Bayesian inference method taking gene tree topologies as input.** The divergence times are marked in red, and the inheritance probability is marked in blue.

probability of the horizontal reticulate edge is 0.38 ± 0.06, which is close to the value of 0.36 reported in (17).

5.4 Results for the data set of 106 loci from five Saccharomyces species

For the data set of 106 loci from five Saccharomyces species, we ran two MC3 chains with 6×10^7 iterations with 1×10^7 burn-in. One sample was collected from every 5,000 iterations.

- 95% credible sets of the phylogenetic network topologies. The 95% credible sets contain only one topology, as shown in Fig. S36. The topology is identical to the

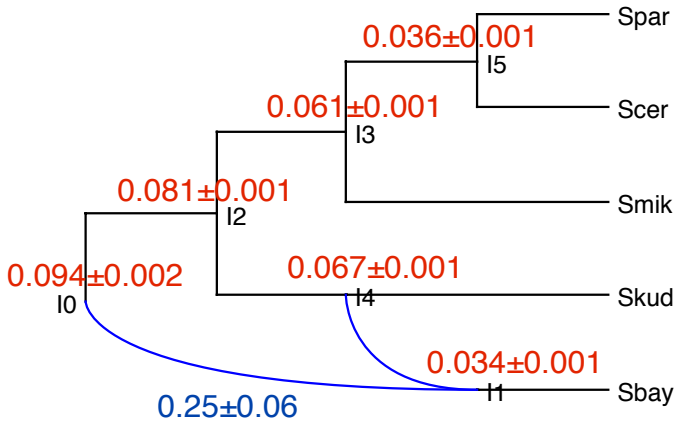

Fig. S36: **The phylogenetic network in the 95% credible set from the data set of 106 loci from five Saccharomyces species using our method.** The divergence times are marked in red, and the inheritance probability is marked in blue.

  one reported in (17), which reconciles the two main species tree topologies reported by (4) in Fig. S28). The inheritance probability of the horizontal reticulate edge is 0.75 ± 0.06, which differs from the value of 0.36 reported in (17). The population mutation rate is $1.7 \pm 0.2 \times 10^{-2}$. The divergence times are similar to the ones inferred from the full data set of seven species in Fig. S30.

- Convergence and mixing. The trace plots shown in Fig. S37 indicate good convergence and mixing.

Analyzing the dataset using gene tree topologies as input We revisited the dataset via the Bayesian inference method (16) taking gene tree topologies as input. We ran two

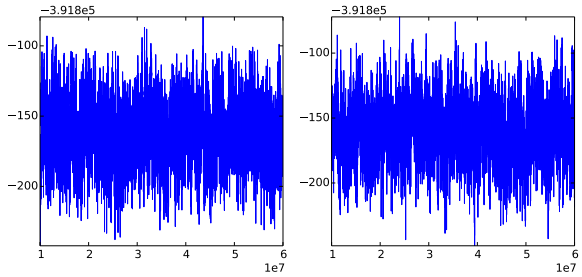

Fig. S37: **Trace plots of MC3 chains using our method given the data set of 106 loci from five Saccharomyces species.**

MCMC chains with 1.1×10^6 iterations with 1×10^5 burn-in. One sample was collected from every 1,000 iterations.

The 95% credible sets contain only one topology, as shown in Fig. S38. The inheritance

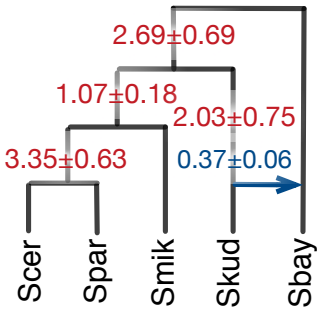

Fig. S38: **The phylogenetic network in the 95% credible sets from the data set of 106 loci from five Saccharomyces species using the Bayesian inference method taking gene tree topologies as input.** The divergence times are marked in red, and the inheritance probability is marked in blue.

probability of the horizontal reticulate edge is 0.37 ± 0.06, which is close to the value of 0.36 reported in (17).

6 Runtimes

All the results reported above were obtained by running the code on NOTS (Night Owls Time-Sharing Service), which is a batch scheduled High-Throughput Computing (HTC) cluster. We used 16 cores, with two threads per core running at 2.6GHz, and 1G RAM per thread.

6.1 Simulations

The runtimes, in hours, for analyzing the 16-, 32-, 64-, and 128-locus data sets, respectively, on each of the four networks in Fig. S9 were as follows:

- The network of Fig. S9(A): 6.1, 5.6, 5.9, 8.9
- The network of Fig. S9(B): 5.8, 6.0, 6.1, 9.1
- The network of Fig. S9(C): 6.3, 5.7, 6.0, 8.8
- The network of Fig. S9(D): 6.3, 6.8, 6.3, 9.3

The runtimes, in minutes, for analyzing the simulated data sets with 20 replicates for each of the 36 simulation settings ($s \in \{0.1, 0.25, 0.5, 1.0\}$, $seqLen \in \{250, 500, 1000\}$, $numLoci \in \{32, 64, 128\}$) under the true species phylogeny in Fig. 7 in the main text are given in in Fig. S39.

The runtimes, in minutes, for analyzing the simulated gene tree topologies with 20 replicates for each of the 3 simulation settings ($s = 1.0$, $seqLen = 250$, $numLoci \in \{32, 64, 128\}$) under the true species phylogeny in Fig. 7 in the main text were as follows:

- $numLoci = 32$: 15.8 ± 3.3
- $numLoci = 64$: 28.2 ± 7.0
- $numLoci = 128$: 49.2 ± 4.8

The runtimes, in minutes, for analyzing the simulated data sets with Intermixture/Gene flow patterns under the true species phylogenies in Fig. S25 were as follows:

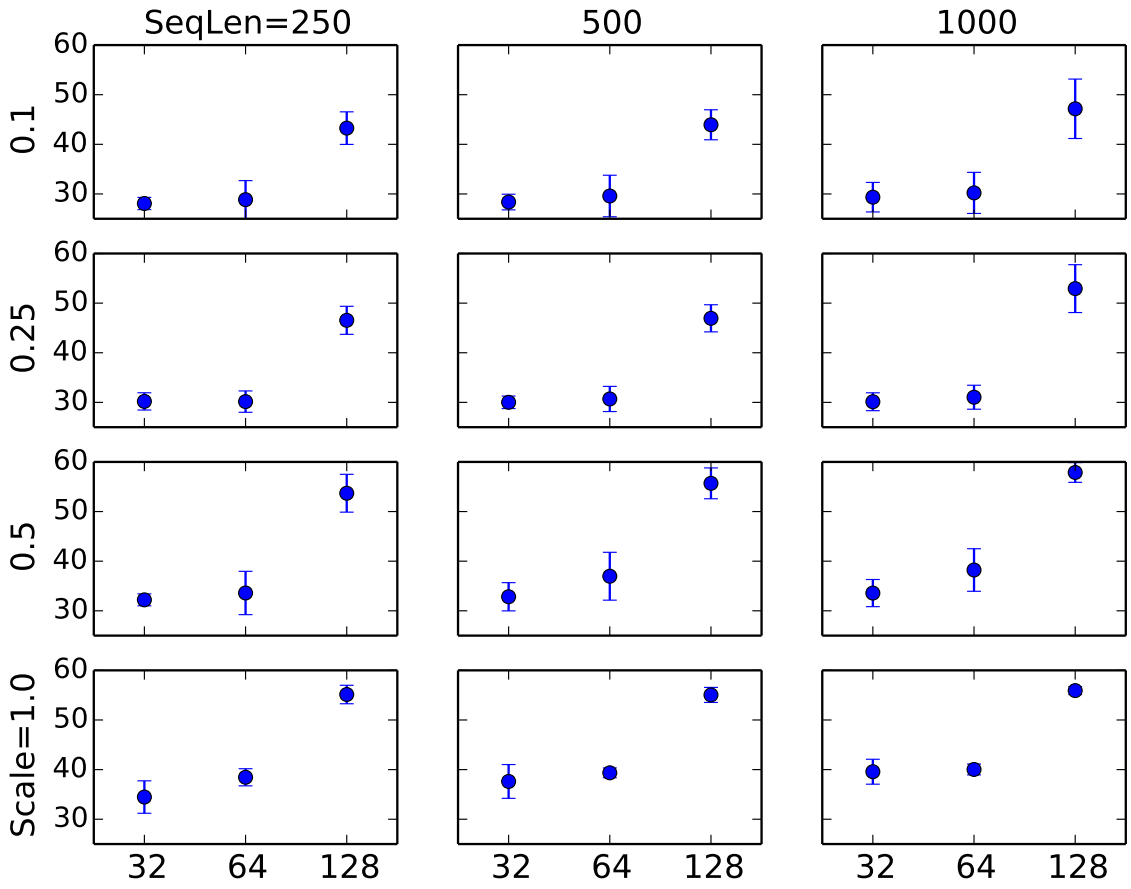

Fig. S39: **The runtimes in minutes under different simulation conditions.** From top to bottom: 0.1, 0.25, 0.5, 1.0 divergence time scale, respectively. From left to right: 250, 500, and 1000 bps sequence length, respectively. Within each plot: 32, 64, 128 loci, respectively.

- The recurrent intermixture in S25(a), $\Delta t = 1$: 44.9 ± 3.5
- The recurrent intermixture in S25(a), $\Delta t = 2$: 49.5 ± 6.0
- The recurrent intermixture in S25(a), $\Delta t = 3$: 54.1 ± 6.6
- The recurrent gene flow in S25(b), $\Delta mt = 1$: 50.6 ± 5.9
- The recurrent gene flow in S25(b), $\Delta mt = 2$: 49.7 ± 6.0
- The paraphyletic intermixture between non-sister species in S25(c), $t = 1$: 42.8 ± 3.4

- The paraphyletic intermixture between non-sister species in S25(c), $t = 0$: 42.1 ± 3.4
- The paraphyletic gene flow between non-sister species in S25(d), $mt_2 = 1$: 45.8 ± 4.6
- The paraphyletic gene flow between non-sister species in S25(d), $mt_2 = 0$: 46.5 ± 5.7
- The isolation-migration between sister species in S25(e), $t = 1$: 36.2 ± 3.8
- The isolation-migration between sister species in S25(e), $t = 0$: 48.8 ± 7.6
- The isolation-migration between sister species in S25(f), $mt = 1$: 36.5 ± 3.9
- The isolation-migration between sister species in S25(f), $mt = 0$: 47.7 ± 6.9

The runtimes, in minutes, for analyzing the simulated data sets with varying number of individuals under the true species phylogenies in Fig. S25(a) were as follows:

- The recurrent intermixture in S25(a), $\Delta t = 1, \#3$: 62.3 ± 4.7
- The recurrent intermixture in S25(a), $\Delta t = 1, \#5$: 82.0 ± 5.3
- The recurrent intermixture in S25(a), $\Delta t = 2, \#3$: 64.3 ± 2.8
- The recurrent intermixture in S25(a), $\Delta t = 2, \#5$: 85.3 ± 4.8

6.2 Biological data sets

For the yeast data set, the runtimes were as follows (when using three chains in Metropolis-Coupled MCMC):

- 7-taxon, 106-locus data set: $35 \sim 38$ hours
- 5-taxon, 106-locus data set: $16.6 \sim 18$ hours

7 PhyloNet Implementation and Usage

We implemented our method in PhyloNet (15), a publicly available, open-source software package for phylogenetic network inference and analysis. Description of the command options and the scripts used in the analyses described above are found under the *MCMC_SEQ* command of PhyloNet.

8 References

S1. Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3): 407–415.

S2. Bloomquist, E. and Suchard, M. 2010. Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. *Systematic Biology*, 59(1): 27–41.

S3. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. 2014. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4): e1003537.

S4. Edwards, S. V., Liu, L., and Pearl, D. K. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14): 5936–5941.

S5. Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., *et al.* 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217): 1258524.

S6. Green, P. J. 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4): 711–732.

S7. Heled, J. and Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, 27(3): 570–580.

S8. Hudson, R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18: 337–338.

S9. Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, NY.

S10. Kuhner, M. K. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3): 459–468.

S11. Rambaut, A. and Grassly, N. C. 1997. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosci.*, 13: 235–238.

S12. Robinson, D. and Foulds, L. 1981. Comparison of phylogenetic trees. *Math. Biosci.*, 53: 131–147.

S13. Rokas, A., Williams, B. L., King, N., and Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960): 798–804.

S14. Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

S15. Than, C., Ruths, D., and Nakhleh, L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC bioinformatics*, 9(1): 322.

S16. Wen, D., Yu, Y., and Nakhleh, L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genetics*, 12(5): e1006006.

S17. Yu, Y., Barnett, R. M., and Nakhleh, L. 2013. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic biology*, 62(5): 738–751.