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

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

[1] *Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA;* [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.*

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 three biological data sets. Our results demonstrate the significance of not only co-estimating 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. [Multispecies network coalescent; reticulation; incomplete lineage sorting; phylogenetic network; Bayesian inference; 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 (Bouckaert *et al.* 2014; Edwards *et al.* 2007; Heled and Drummond 2010; Rannala and Yang 2003). 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; Barton 2001; Gogarten *et al.* 2002; Koonin *et al.* 2001; Mallet 2005, 2007; Rieseberg 1997), 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, co-estimating the species phylogeny and gene trees results in better estimates of the gene trees themselves (DeGiorgio and Degnan 2014; Zimmermann *et al.* 2014).

We report here on a Bayesian method for co-estimating 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 of this generative process. In particular, our method co-estimates, in the form of posterior 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 three biological data sets, 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.

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

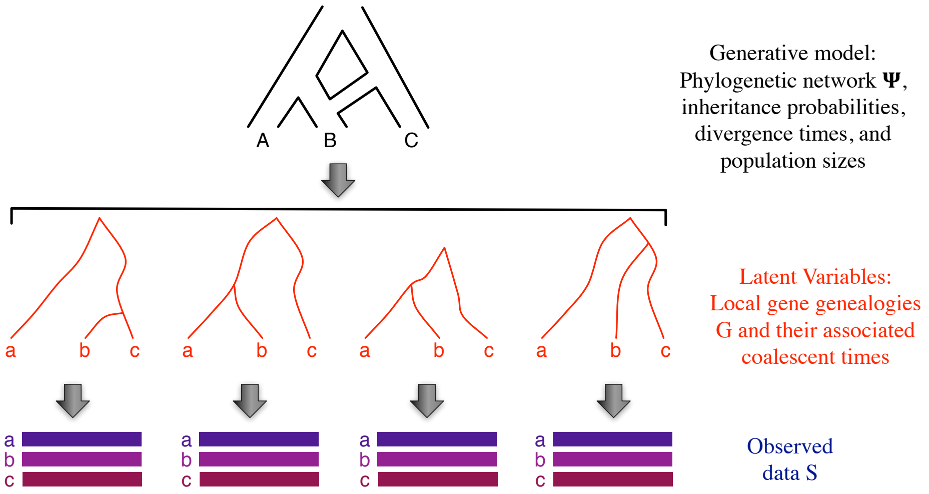

FIGURE **1.** From a phylogenetic network to multi-locus sequences via latent gene genealogies. The multispecies network coalescent (Yu *et al.* 2014) is a stochastic process that defines a probability distribution on gene genealogies along with their coalescent times. The parameters of the process consist 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.

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 by in (Long 1991), where two models of admixture were presented: the intermixture model and the gene flow model (Figure 2). The MSNC employs the intermixture model, whereas the population genetics community mostly uses the gene flow model (Gronau *et al.* 2011; Hey and Nielsen 2004, 2007; Leaché *et al.* 2013; Slatkin and Maddison 1989; Strasburg and Rieseberg 2010; Whitlock and Mccauley 1999). Note that the intermixture model also underlies the admixture graph model of (Pickrell and Pritchard 2012; Reich *et al.* 2009) 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.

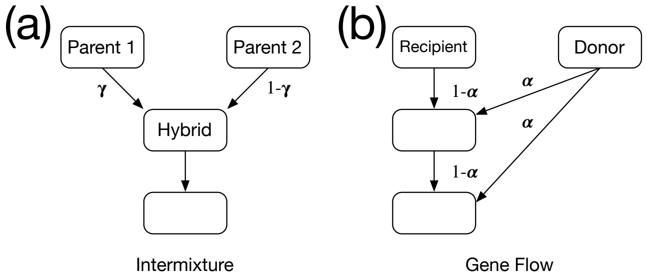

FIGURE **2.** Two admixture models for a hybrid population (Long 1991). (a) The hybrid population is formed by a single intermixture event between two parental populations, where γ is the inheritance probability measuring the proportion of the parental populations. (b) The hybrid population (recipient)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

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

Finally, as the model underlying out method extends the multispecies coalescent to cases that include admixture, our method is applicable to data from different sub-populations, not only different species, and to data where more than one individual per species or sub-population is sampled. The method is implemented and publicly available in the PhyloNet software package (Than *et al.* 2008).

METHODS

0.1 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, $\ell: V_L\to\mathscr{X}$ 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 edge 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. $\Psi_\theta\in(\mathbb{R}^+)^{|E(\Psi)|}$

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 topology of a phylogenetic network is a rooted, directed, acyclic graph, whose leaves are labeled by the taxa under study. Every node in the network has at most two parents, and nodes with two parents are called reticulation nodes. Associated with every internal node of the phylogenetic network is a divergence time parameter (the leaves are all assumed to be at time 0). Associated with every branch of the network, including one incident into the root, is a population size parameter. Furthermore, associated with the branches coming into reticulation nodes are the inheritance probabilities given by Γ.

The posterior of the model is given by

$$\begin{aligned} p(\mathscr{M}|\mathscr{S}) &\propto p(\mathscr{S}|\mathscr{M})p(\mathscr{M}) \\ &= p(\mathscr{M})\prod_{i=1}^{m}\int_G p(S_i|g)p(g|\mathscr{M})dg, \end{aligned} \tag{0.1}$$

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. (0.1) 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 as given by

$$\begin{aligned} p(\Psi,\Gamma,G|S) &\propto p(\mathscr{M})\prod_{i=1}^{m}p(S_i|g_i)p(g_i|\mathscr{M}) \\ &= p(\Psi)p(\Gamma)\prod_{i=1}^{m}p(S_i|g_i)p(g_i|\Psi,\Gamma), \end{aligned} \tag{0.2}$$

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

The Likelihood Function

Felsenstein (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, and

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

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.* (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

$$p(g|\Psi,\Gamma) = \sum_{h \in H_\Psi(g)} p(h|\Psi,\Gamma), \qquad (0.3)$$

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

$$\begin{aligned} &p(h|\Psi,\Gamma) = \\ &\prod_{b \in E(\Psi)} \left[\prod_{i=1}^{|T_b(h)|-1} \frac{2}{\theta_b} e^{-(\frac{2}{\theta_b})\binom{u_b(h)-i+1}{2}(T_b(h)_{i+1} - T_b(h)_i)} \right] \\ &\times e^{-(\frac{2}{\theta_b})\binom{v_b(h)}{2}(\tau_\Psi(b) - T_b(h)_{|T_b(h)|})} \times \Gamma_b^{u_b(h)}, \end{aligned} \qquad (0.4)$$

where $\theta_b = 4N_b\mu$ 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. (0.5),

$$\begin{aligned} p(\Psi|\nu,\delta,\eta,\psi) = &p(\Psi_{ret}|\nu) \times p(\Psi_{top}|\Psi_{ret},\Psi_\tau,\eta) \\ &\times p(\Psi_\tau|\delta) \times p(\Psi_\theta|\psi) \end{aligned} \qquad (0.5)$$

where $p(\Psi_{ret}|\nu)$, the prior on the number of reticulation nodes, and $p(\Psi_{top}|\Psi_{ret},\Psi_\tau,\eta)$, 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 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 $\Gamma_b \sim \mathrm{Beta}(\alpha,\beta)$. 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 hyper-parameters α and β can be appropriately set to bias the inheritance probabilities to values close to 0 and 1 (a U-shaped distribution).

The RJMCMC Sampler

As computing the posterior distribution given by Eq. (0.2) 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, one for sampling the inheritance probabilities, and one for sampling the gene trees. To propose a new state of the Markov chain, one element from $(\Psi,\gamma_1,...,\gamma_{\Psi_{ret}},g_1,...,g_m)$ 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 Materials.

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

RESULTS

Performance on Data Simulated Under the Intermixture Model

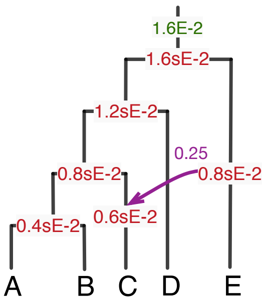

FIGURE 3. A model phylogenetic network used to generate simulated data. The divergence times in units of expected number of mutations per site, the population size parameter in units of population mutation rate per site, and the inheritance probability are marked in red, green, and purple, respectively. Parameter s is used to scale the divergence times.

We used the phylogenetic nework shown in Fig. 3 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 each setting of the four settings of s values, we simulated 20 data sets with 128 independent loci. For each of those 20 data sets, the program ms (Hudson 2002) was used to simulate the gene trees and the program Seq-gen (Rambaut and Grassly 1997) was used to generate sequence alignments down the gene trees under the Jukes Cantor model. Sequence alignments were generated with lengths of 250, 500, and 1000 sites. 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. For each data set, we ran an MCMC chain of 8×10^6 iterations with 3×10^6 burn-in. One sample was collected from every 5,000 iterations, resulting in a total of 1,000 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).

In assessing the performance of our method, we evaluated the estimates obtained for the various parameters of interest: divergence times, population size, the number of reticulations, and the topology of the inferred species phylogeny. Fig. 4 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. 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.

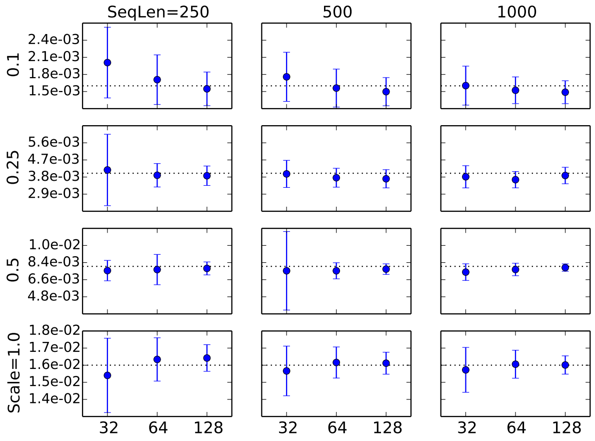

FIGURE 4. Divergence time estimates at the root under different values of the scaling parameter s (different rows), sequence lengths (different columns), and numbers of loci (three values within each panel). The dashed line indicates the true value in the model network.

Fig. 5 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.

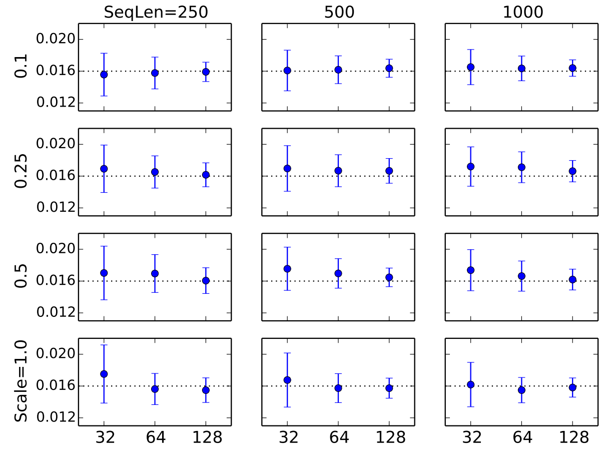

FIGURE 5. Population mutation rate estimates under different values of the scaling parameter s (different rows), sequence lengths (different columns), and numbers of loci (three values within each panel). The dashed line indicates the true value in the model network.

The results are quite different when it comes to estimating the number of reticulations and the topology of the phylogenetic network itself. Fig. 6 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; that is, it estimates the number of reticulations to be 0, 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. 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.

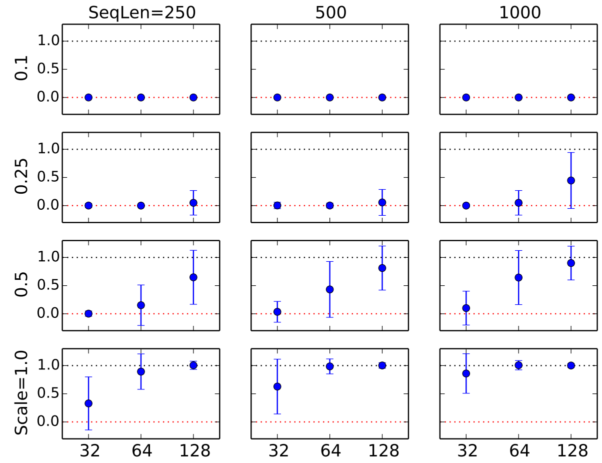

FIGURE 6. The number of reticulations inferred under different simulation conditions. The model network has a single reticulation.

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. 3) 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? Fig. 7 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 scaling parameter values, and very good accuracy and fast convergence to accurate estimates in cases of larger scaling parameter values. 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.

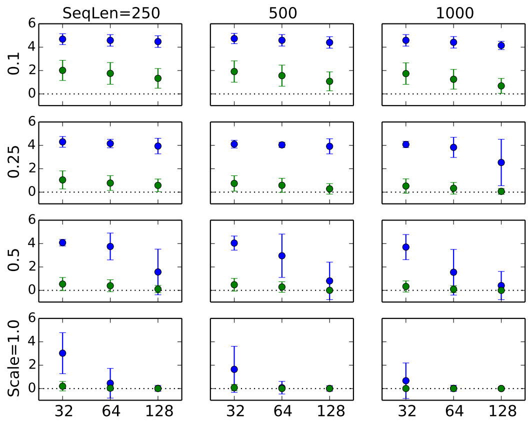

FIGURE 7. The topological difference between the true and inferred networks in blue and the Robinson-Foulds distance between the inferred tree (if a network is inferred, this case is not included) and the backbone tree of the true network.

Contrasting the Performance on Data Simulated Under the Intermixture and Gene Flow 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 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. Fig. 8 shows the six phylogenetic networks we used to generate data.

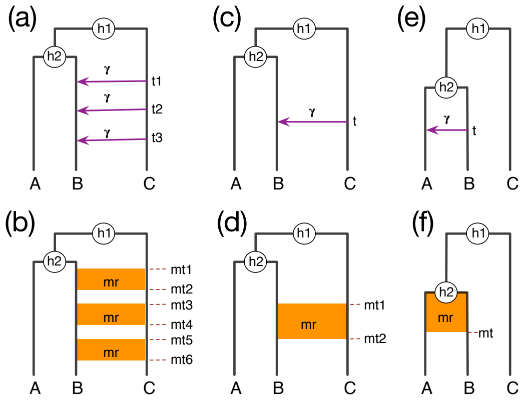

FIGURE 8. True phylogenetic histories with intermixture and gene flow models. Recurrent reticulations between non-sister taxa (a,b), a single reticulation between non-sister taxa (c,d), and a single reticulation between sister taxa (e,f) is captured under both the intermixture model (top) and gene flow 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.

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. The MCMC settings were set as discussed above.

We set $h_1=9$, $h_2=6$. For the intermixture model (Fig. 8(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. 8(b)), we set $(mt_1,...,mt_6)$ to (6,4,4,2,2,0) and (6,5,3.5,2.5,1,0), so that the duration of each gene flow epoch, denoted by Δmt, is either 1 or 2. Notice that, under our setting, the time elapsed between two consecutive gene flow epochs is smaller for $\Delta mt=2$ than for $\Delta mt=1$.

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. 8(a) and Fig. 8(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 was generated under an intermixture model or a gene flow model. Furthermore, for these two parameters, the estimates are stable while varying the elapsed times between consecutive reticulation events.

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. 8(a) and Δmt in the model of Fig. 8(b). 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
$\Delta t=1$	$2.2\pm0.2e^{-2}$	8.9±0.1	5.9±0.1	1.2±0.4
$\Delta t=2$	$2.2\pm0.2e^{-2}$	8.9±0.1	5.9±0.1	2.0±0.0
$\Delta t=3$	$2.1\pm0.3e^{-2}$	9.0±0.1	6.0±0.1	2.6±0.5
$\Delta mt=1$	$2.3\pm0.3e^{-2}$	8.9±0.1	6.0±0.1	2.1±0.3
$\Delta mt=2$	$2.3\pm0.3e^{-2}$	8.9±0.1	6.9±0.1	2.0±0.1

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 coalescent times of the most recent common ancestor (MRCA) of alleles from B and C in Fig. 9. As Fig. 8(a) and Fig. 8(b) show, the coalescent 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.

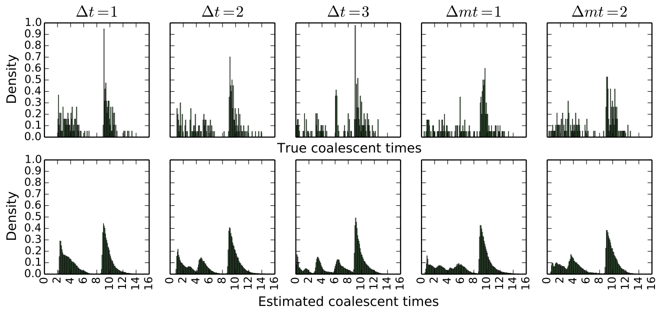

FIGURE 9. Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on data generated under the models of Fig. 8(a) and Fig. 8(b).

As the left three columns of panels in the figure show, under an intermixture model, as Δt increases, the signal for a mixture of four distributions of (A,B) coalescent times becomes much stronger, thus pointing to three reticulations in addition to the coalescent 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, this is not the case under the gene flow model (the right two columns of panels in the figure). It is important to note that for $\Delta mt = 2$, the three gene flow epochs actually form one continuous epoch of gene flow from time mt_1 to mt_6.

Fig. 10 shows results similar to those reported in Fig. 9, with the only difference being that these are the coalescent times from all 4,000 loci generated from the 20 data sets of 200 loci each. Effectively, this is the signal in a data set of 4,000 independent loci. Clearly, the signal is much stronger than in data sets of 200 loci, and all reticualtions would be recoverable under the intermixture model for $\Delta t = 2,3$ and for the gene flow model for $\Delta mt = 1$.

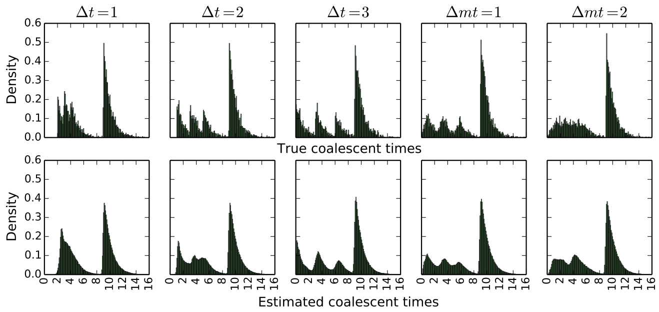

FIGURE 10. Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on 4,000 loci generated under the models of Fig. 8(a) and Fig. 8(b).

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 Materials.

To assess the performance of our method on the simpler case of a single reticulation event, we considered the networks in Fig. 8(c) and Fig. 8(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. A single reticulation was detected for all cases of intermixture and gene flow. We plotted the histograms of the true and estimated coalescent times of the MRCA of alleles from B and C in Fig. 11. As the figure shows, the distributions of estimated coalescent times match the distributions of true coalescent times very well.

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. 8(c) and mt_2 in the model of Fig. 8(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	γ (mr)	#reti
$t = 1$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	0.20 ± 0.05	1.0 ± 0.0
$t = 0$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	0.21 ± 0.04	1.0 ± 0.0
$mt_2 = 1$	$2.0 \pm 0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	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	0.17 ± 0.04	1.0 ± 0.0

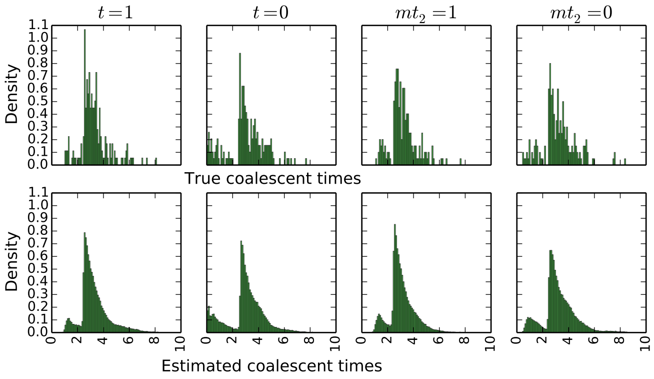

FIGURE 11. Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on data generated under the models of Fig. 8(c) and Fig. 8(d).

Finally, we assessed the performance of our method on cases where the reticulation event involves sister taxa. Fig. 8(e) and Fig. 8(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.

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. 8(e) and mt in the model of Fig. 8(f). 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	γ	#reti
$t=1$	$2.0\pm0.2e^{-2}$	2.5 ± 0.1	1.3 ± 0.1	NA	0.0 ± 0.0
$t=0$	$2.0\pm0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.0	0.21 ± 0.06	1.0 ± 0.0
$mt=1$	$2.0\pm0.2e^{-2}$	2.5 ± 0.1	1.4 ± 0.1	NA	0.0 ± 0.0
$mt=0$	$2.2\pm0.2e^{-2}$	2.5 ± 0.1	1.5 ± 0.1	0.11 ± 0.06	1.0 ± 0.0

We plotted the histograms of the true and estimated coalescent times of the MRCA of alleles from A and B in Fig. 12. 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 coalescent times match those of true coalescent times very well.

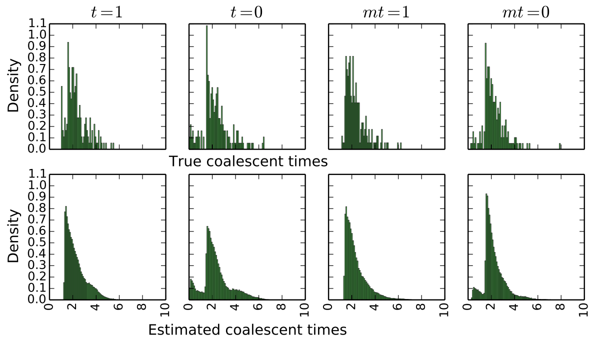

FIGURE 12. Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from A and B on data generated under the models of Fig. 8(e) and Fig. 8(f).

Performance on Biological Data Sets

Analysis of a bread wheat data set The bread wheat data set consists of three subgenomes of *Triticum aestivum*, TaA (A subgenome), TaB (B subgenome) and TaD (D subgenome), and five diploid relatives Tm (*T. monococcum*), Tu (*T. urartu*), Ash (*Ae. sharonensis*), Asp (*Ae. speltoides*) and At (*Ae. tauschii*). Marcussen *et al.* found that each of the A and B lineages is more closely related to D than to each other, as represented by the phylogenetic network in Fig. 13(a) inferred using the parsimony approach of (Yu *et al.*, 2011) given gene tree topologies of TaA, TaB, and TaD. Based on this network, they proposed an evolutionary history of *Triticum aestivum*, where about 7 million years ago the A and B genomes diverged from a common ancestor and 1∼2 million years later these genomes gave rise to the D genome through homoploid hybrid speciation (Marcussen *et al.*, 2014). We fed 68 loci of the eight genomes into our method. The only network in the 95% credible set is identical to the one in (Marcussen *et al.*, 2014), shown in Fig. 13(b) where A, B, and D represent ((TaA,Tu),Tm), (TaB,Asp), and ((TaD,At),Ash), respectively. Assuming a mutation rate of 1×10^{-9} per-site per-generation and 1 year per generation, a plausible evolutionary history posits that a common ancestor of A, B, and D started differentiation ∼8.5 Ma into (A,D) and B genome lineages. Subsequently, (A,D) speciated at ∼7 Ma into A and D lineages. The hybridization occurred around 4-4.5 Ma from B to D genome lineages. Although both proposed evolutionary histories contain one hybridization, phylogenetic networks with two or more reticulations were inferred on larger data sets by our method and by the authors of the original study (Marcussen *et al.*, 2014). See Supplementary Materials for full details.

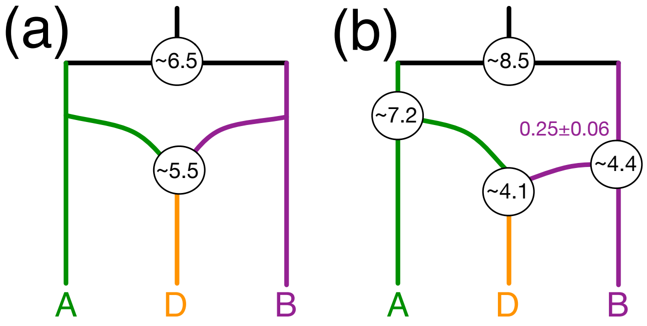

FIGURE 13. Phylogenetic history of the bread wheat data of (Marcussen *et al.*, 2014). (a) The phylogenetic network inferred using the parsimony approach of (Yu *et al.*, 2011) given gene tree topologies of TaA, TaB, and TaD. The times are estimated by gene tree analyses. (b) The phylogenetic network inferred by our method given 68 loci of eight genomes. The times at the internal nodes are in millions of years.

Analysis of a 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), *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. 14(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. 14(a-b). The other gene tree topologies (Fig. 14(c)) exhibited weak phylogenetic signals among Sklu, Scas and the other species. Bloomquist and Suchard (Bloomquist and Suchard, 2010) reanalyzed the data set without Sklu since it added too much noise to their analysis. Their

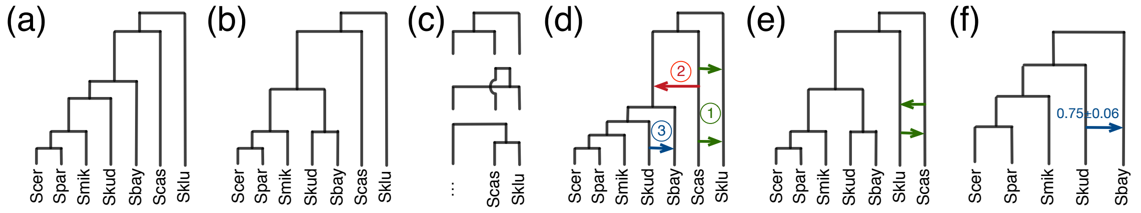

FIGURE 14. Results on the yeast data set of (Rokas *et al.*, 2003). (a) The species tree inferred using the concatenation method (Rokas *et al.*, 2003) and the main species tree and gene tree topology sampled using BEST (Edwards *et al.*, 2007). (b) The second most frequently sampled species and gene 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) A representative phylogenetic network inferred by our method on all 106 loci. (e) A representative phylogenetic network inferred by our method on the 28 loci with strong phylogenetic signal (see Supplementary Materials). (f) The single phylogenetic network inferred using all 106 loci from the five species Scer, Spar, Smik, Skud, Sbay.

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.* (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 representative network is shown in Fig. 14(d). All the previous findings are encompassed by the networks inferred by our method. The two hybridizations between Sklu and Scas (green edges in 14(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 edge in 14(d)) captures the stronger lineage-specific rate variation in Scas. Finally, the hybridization from Skud to Sbay (blue edge in 14(d)) resolves the incongruence between the two main species tree topologies in 14(a-b).

We further investigated the phylogenetic signal in each locus by counting the number of internal branches in the 70% majority-rule consensus of 100 maximum likelihood bootstrap trees. We found that only 28 out of the 106 loci contain four internal branches, and no locus had a consensus tree with all five internal branches. A representative phylogenetic network in the 95% credible set given these 28 loci, shown in Fig. 14(e), indicates the weak phylogenetic signals among Sklu, Scas and the rest of the species. 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 MPP phylogenetic network in Fig. 14(f) contains the hybridization from Skud to Sbay, which is identical to the sub-network in 14(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.

Analysis of a mosquito data set The Anopheles mosquitoes (*An. gambiae* complex) data set of (Fontaine *et al.*, 2015) consists of genome alignment of *An. gambiae* (G), *An. coluzzii* (C), *An. arabiensis* (A), *An. quadriannulatus* (Q), *An. merus* (R) and *An. melas* (L). Fontaine *et al.* reported on extensive introgressions in the *An. gambiae* complex (Fontaine *et al.*, 2015). Gene tree analyses were performed to detect the donor, recipient and migration times of the reticulation edges. Three major introgressions were added to the species tree backbone recovered from the X chromosome, resulting in a plausible phylogenetic network. More recently, Wen *et al.* (Wen *et al.*, 2016b) reanalyzed the data set using the maximum likelihood method of (Yu *et al.*, 2014) and then using the Bayesian method of (Wen *et al.*, 2016a) and provided new insights into the evolutionary history of the *An. gambiae* complex.

Fontaine et al. inferred gene trees on 50-kb genomic windows using maximum likelihood and tabulated and analyzed the frequencies of distinct gene tree topologies across the chromosomes. However, such large genomic windows are very likely to include recombination. Indeed, a simple comparison of the 70% majority-rule consensus of 100 maximum likelihood bootstrap trees on the entire window against individual trees inferred from smaller regions of the same window highlight this issue (see Supplementary Materials for details).

To avoid using such large genomic windows, we randomly sampled 228 1-kb regions from the X chromosome. We fed the 228-locus data set into ∗BEAST and our method. Our method produces a phylogenetic network with many reticulations on this data set. We assessed the phylogenetic signal in each locus by computing the number of internal branches in the 70% majority-rule consensus of 100 maximum likelihood bootstrap trees. We found that only 59 out of the 228 loci contain three internal branches, and no locus had a consensus tree with all four internal branches. The MPP species tree inferred by ∗BEAST (Fig. 15(a)) groups (A,Q) with (C,G) to account for heterogeneity across loci by means of ILS alone. Analyzing those 59 loci data set using our method, the 95% credible set contains three topologies grouping (C,G) with R and positing hybridization from A, Q, or (A,Q) to (C,G) with inheritance probability 0.33 (Fig. 15(b)). The divergence times of the MRCAs of (C,G), (A,Q), (A,Q,C,G), and (R,C,G) inferred by ∗BEAST are similar to those inferred by our method. The minimum coalescent times of clades (C,G), (A,Q), (A,Q,C,G) and (R,C,G) co-estimated by ∗BEAST (green) and our method (blue) in Fig. 15(c) further confirm this statement. However, ∗BEAST obtains lower estimates of the divergence times of the MRCA of (R,L) (or the root) to reconcile the divergence times of (R,C,G). BEAST, which infers gene trees from sequences without regard to a species tree, significantly underestimates all the coalescent times (sandy brown bars in Fig. 15(c)). For the autosomes, we randomly sampled 382 1-kb regions with strong phylogenetic signal and fed the data set into our method. The MPP phylogenetic network shown in Fig. 15(e) groups Q with R, and reveals hybridization from (A,C,G) to Q. This network can be embedded in the phylogenetic network inferred by the Bayesian method of (Wen *et al.*, 2016a) given gene tree topologies from 2791 regions with varying lengths of 1∼20-kb from the autosomes (Fig. 15(d)). Using data with strong phylogenetic signal would significantly reduce the complexity of the model. See Supplementary Materials for details.

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 three biological data sets. Our results demonstrate several important aspects. First, ignoring hybridization when it had occurred results in underestimating the divergence times of species and overestimating the coalescent times of individual loci. Second, co-estimation 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 (Wen *et al.* 2016a; Yu *et al.* 2014) 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. Last but not the least, the phylogenetic signal in the individual loci on the inference must be taken into consideration when selecting loci of analysis of reticulate evolutionary histories. In particular, when there is low phylogenetic signal in the data, tree inference methods tend to result in unresolved trees. In the case of network methods, the counterpart to an unresolved tree is an overly complex network. In other words, while low signal is captured by a soft polytomy in trees, it is captured by multiple reticulations in networks. Therefore, it is very important that the signal in individual loci is carefully assessed in network inference, and indeed, in phylogenomics in general.

While the MSNC corresponds naturally to an intermixture model of admixture, we assesses 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 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. 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

Supplementary material, including data files and online-only appendices, can be found in the Dryad data repository at .

FUNDING

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

REFERENCES

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

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., *et al.* 2011. Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic biology*, 61: 170–173.

Barton, N. 2001. The role of hybridization in evolution. *Molecular Ecology*, 10(3): 551–568.

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

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.

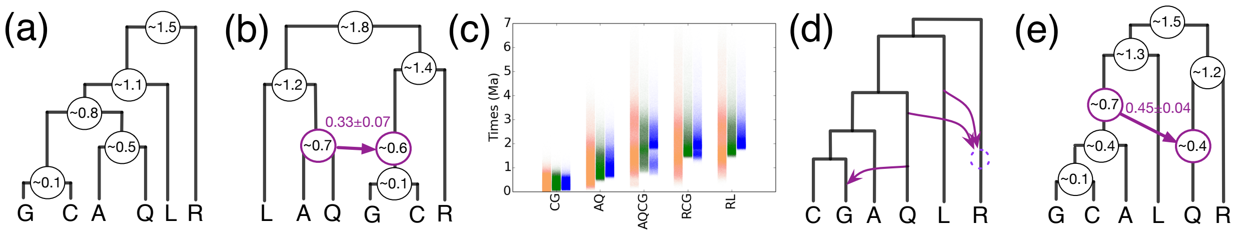

FIGURE 15. Results on the *An. gambiae* complex data of (Fontaine *et al.*, 2015). (a) The MPP species tree inferred by ∗BEAST on regions with strong phylogenetic signal from the X chromosome. (b) The phylogenetic network inferred by our method on the same regions as in **A**. (c) The coalescent times of the MRCAs of (C,G), (A,Q), (A,Q,C,G), (R,C,G) and (R,L) from gene trees inferred by BEAST (sandy brown), ∗BEAST (green), and our method (blue) on the same regions as in (a). (d) The phylogenetic network inferred by the Bayesian method of (Wen *et al.*, 2016a) given gene tree topologies from 2791 regions with varying lengths of 1∼20-kb from the autosomes. (e) The phylogenetic network inferred by our method given 382 1-kb regions with strong phylogenetic signals from the autosomes.

DeGiorgio, M. and Degnan, J. H. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Systematic biology*, 63(1): 66–82.

Degnan, J. H. and Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24(6): 332–340.

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.

Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*, 17(6): 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., *et al.* 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217): 1258524.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. 2002. Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution*, 19(12): 2226–2238.

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

Green, P. J. 2003. Trans-dimensional Markov chain Monte Carlo. In P. Green, N. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Processes*, pages 179–198. Oxford University Press, Oxford, UK.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10): 1031–1034.

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

Hey, J. and 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(2): 747–760.

Hey, J. and Nielsen, R. 2007. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8): 2785–2790.

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

Koonin, E. V., Makarova, K. S., and Aravind, L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification 1. *Annual Reviews in Microbiology*, 55(1): 709–742.

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. 2013. The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, page syt049.

Liu, L. 2008. Best: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21): 2542–2543.

Long, J. C. 1991. The genetic structure of admixed populations. *Genetics*, 127(2): 417–428.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends Ecol. Evol.*, 20(5): 229–237.

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

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B. B., Steuernagel, B., Mayer, K. F., Olsen, O.-A., *et al.* 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, 345(6194): 1250092.

Nakhleh, L. 2010a. Evolutionary phylogenetic networks: models and issues. In L. Heath and N. Ramakrishnan, editors, *The Problem Solving Handbook for Computational Biology and Bioinformatics*, pages 125–158. Springer, New York.

Nakhleh, L. 2010b. A metric on the space of reduced phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(2): 218–222.

Pickrell, J. K. and Pritchard, J. K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*, 8(11): e1002967.

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.

Rannala, B. and Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4): 1645–1656.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. 2009. Reconstructing indian population history. *Nature*, 461(7263): 489–494.

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

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

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.

Slatkin, M. and Maddison, W. P. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, 123(3): 603–613.

Strasburg, J. L. and Rieseberg, L. H. 2010. How robust are "isolation with migration" analyses to violations of the im model? a simulation study. *Molecular biology and evolution*, 27(2): 297–310.

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

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

Wen, D., Yu, Y., Hahn, M. W., and Nakhleh, L. 2016b. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25(11): 2361–2372.

Whitlock, M. C. and Mccauley, D. E. 1999. Indirect measures of gene flow and migration: Fst≠ 1/(4nm+ 1). *Heredity*, 82(2): 117–125.

Yu, Y., Warnow, T., and Nakhleh, L. 2011. Algorithms for mdc-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, 18(11): 1543–1559.

Yu, Y., Degnan, J. H., and Nakhleh, L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS genetics*, 8(4): e1002660.

Yu, Y., Ristic, N., and Nakhleh, L. 2013a. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics*, 14(Suppl 15): S6.

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

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46): 16448–16453.

Zimmermann, T., Mirarab, S., and Warnow, T. 2014. BBCA: Improving the scalability of *BEAST using random binning. *BMC genomics*, 15(Suppl 6): S11.