Bayesian Inference of Phylogenetic Networks from Bi-allelic Genetic Markers

Jiafan Zhu [1], Dingqiao Wen [1], Yun Yu [1], Heidi M. Meudt [2], and Luay Nakhleh [1,3,*]
[1]Computer Science, Rice University, Houston, TX, USA,
[2]Museum of New Zealand Te Papa Tongarewa, Wellington, New Zealand, and
[3]BioSciences, Rice University, Houston, TX, USA.
[*]**Corresponding author:** nakhleh@rice.edu.

Abstract

Phylogenetic networks are rooted, directed, acyclic graphs that model reticulate evolutionary histories. Recently, statistical methods were devised for inferring such networks from either gene tree estimates or the sequence alignments of multiple unlinked loci. Bi-allelic markers, most notably single nucleotide polymorphisms (SNPs) and amplified fragment length polymorphisms (AFLPs), provide a powerful source of genome-wide data. In a recent paper, a method called SNAPP was introduced for statistical inference of species trees from unlinked bi-allelic markers. The generative process assumed by the method combined both a model of evolution for the bi-allelic markers, as well as the multispecies coalescent. A novel component of the method was a polynomial-time algorithm for exact computation of the likelihood of a fixed species tree via integration over all possible gene trees for a given marker. Here we report on a method for Bayesian inference of phylogenetic networks from bi-allelic markers. Our method significantly extends the algorithm for exact computation of phylogenetic network likelihood via integration over all possible gene trees. Unlike the case of species trees, the algorithm is no longer polynomial-time on all instances of phylogenetic networks. Furthermore, the method utilizes a reversible-jump MCMC technique to sample the posterior of phylogenetic networks given bi-allelic marker data. Our method has a very good performance in terms of accuracy as we demonstrate on simulated data, as well as a data set of multiple New Zealand species of the plant genus *Ourisia* (Plantaginaceae). We implemented the method in the publicly available, open-source PhyloNet software package.

Key words: multispecies network coalescent; phylogenetic networks; bi-allelic markers; reticulation; incomplete lineage sorting.

Introduction

The availability of genome-wide data from many species and, in some cases, many individuals per species, has transformed the study of evolutionary histories, and given rise to phylogenomics—the inference of gene and species evolutionary histories from genome-wide data. Consider a data set $S=\{S_1,\ldots,S_m\}$ consisting of the molecular sequences of m loci under the assumptions of free recombination between loci and no recombination within a locus. The likelihood of a species phylogeny Ψ (topology and parameters) is given by

$$\mathcal{L}(\Psi|S) = \prod_{i=1}^{m} \mathcal{L}(\Psi|S_i) = \prod_{i=1}^{m} \int_G p(S_i|g)p(g|\Psi)\mathrm{d}g \quad (1)$$

where the integration is taken over all possible gene trees. The term $p(S_i|g)$ is the likelihood of gene tree g given the sequence data of locus i (Felsenstein, 1981). The term $p(g|\Psi)$ is the density function (pdf) of gene trees given the species phylogeny and its parameters. For example, (Rannala and Yang, 2003) derived this pdf under the multispecies coalescent (MSC) (Degnan and Rosenberg, 2009). This formulation underlies the Bayesian inference methods of (Heled and Drummond, 2010; Liu and Pearl, 2007; Rannala and Yang, 2003).

Debate has recently ensued regarding the size of genomic regions that would be recombination-free (or almost recombination-free) and could truly have a single underlying evolutionary tree (Edwards *et al.*, 2016; Springer and Gatesy, 2016). One way to overcome this issue is to use unlinked single nucleotide polymorphisms (SNPs) or amplified fragment length polymorphisms (AFLPs). Such data provide a powerful signal for inferring species phylogenies and the issue of recombination within a locus becomes irrelevant. Furthermore, as long as those markers are sampled far enough from each other the assumption of free recombination within loci holds. Indeed, this is the basis of the SNAPP method that was recently introduced in the seminal paper of (Bryant *et al.*, 2012). Since a bi-allelic SNP or AFLP marker has no signal by itself to resolve much of the branching patterns of a gene genealogy, a major contribution of Bryant *et al.* was an algorithm for analytically computing the integration in Eq. (1) for bi-allelic markers.

While trees constitute an appropriate model of the evolutionary histories of many groups of species, it is well known that other groups of species have evolutionary histories that are reticulate (Mallet *et al.*, 2016). Horizontal gene transfer is ubiquitous in prokaryotes (Gogarten *et al.*, 2002; Koonin *et al.*, 2001), and several bodies of work are pointing to much larger extent and role of hybridization in eukaryotic evolution than once thought (Arnold, 1997; Barton, 2001; Mallet, 2005, 2007; Mallet *et al.*, 2016; Rieseberg, 1997). Not only does hybridization play an important role in the genomic diversification of several eukaryotic groups, but increasing evidence is pointing to the adaptive role it has played, for example, in wild sunflowers (Rieseberg *et al.*, 2003), humans (Racimo *et al.*, 2015), macaques (Stevison and Kohn, 2009), mice (Liu *et al.*, 2015), butterflies (Zhang *et al.*, 2016), and mosquitoes (Fontaine *et al.*, 2015; Wen *et al.*, 2016b).

Reticulate evolutionary histories are best modeled by *phylogenetic networks*. Two statistical methods were recently introduced for inference under the formulation given by Eq. (1), when Ψ is a phylogenetic network (Wen and Nakhleh, 2016; Zhang *et al.*, 2017), and other methods were also introduced for statistical inference of phylogenetic networks using gene tree estimates as the input data (Solís-Lemus and Ané, 2016; Wen *et al.*, 2016a; Yu and Nakhleh, 2015; Yu *et al.*, 2012, 2014).

The methods of (Wen and Nakhleh, 2016; Zhang *et al.*, 2017) assume that the data for each locus consists of a sequence alignment that has no recombination. In this paper, we devise an algorithm that builds on the algorithm of (Bryant *et al.*, 2012) for analytically computing the integral in Eq. (1) when Ψ is a phylogenetic network. In other words, our algorithm allows for computing the likelihood of a phylogenetic network from unlinked bi-allelic markers while analytically integrating out the gene trees for the individual markers. We couple this likelihood function with priors on the phylogenetic network and its parameters to obtain a Bayesian formulation, and then employ the reversible-jump MCMC (RJMCMC) kernel from (Wen and Nakhleh, 2016) to sample the posterior of the phylogenetic networks and their associated parameters given the bi-allelic data.

We implemented our algorithm and the RJMCMC sampler in PhyloNet (Than *et al.*, 2008), which is a publicly available open-source software package for inferring and analyzing reticulate evolutionary histories. We studied the performance of our method on simulated and biological data. For simulations, we extended the framework of (Bryant *et al.*, 2012) so that the evolution of bi-allelic markers could be simulated within the branches of a phylogenetic network. For the biological data, we analyzed two data sets of multiple New Zealand species of the plant genus *Ourisia* (Plantaginaceae). The results on the simulated data show very good accuracy as reflected by the method's ability to recover the true phylogenetic networks and their associated parameters. For the biological data, the method recovers two established hybrids and their putative parents correctly.

The proposed method and Bayesian sampler provide a new tool for biologists to infer reticulate evolutionary histories, while also account for the complexity arising from incomplete lineage sorting, from bi-allelic markers, thus complementing existing tools that use gene tree estimates or sequence alignments of the individual loci as the input data. The use of such bi-allelic markers, particularly when they are sampled far enough across the genome, completely sidesteps potential problems that could arise due to the presence of recombination within loci.

Methods

Phylogenetic networks and gene trees

A *phylogenetic* $\mathscr{X}$*-network*, or $\mathscr{X}$-network for short, Ψ is a rooted, directed, acyclic graph (DAG) whose leaves are bijectively labeled by set $\mathscr{X}$ of taxa. We denote by $V(\Psi)$ and $E(\Psi)$ the sets of nodes and edges, respectively, of the phylogenetic network Ψ. Every node, except for the root, of the network has in-degree 1, which we call *tree node*, or in-degree 2, which we call reticulation node. The edges whose head is a reticulation node are the *reticulation edges* of the network; all other edges constitute the *tree edges* of the network. We assume all phylogenies considered here (trees and networks) are binary—no node has out-degree higher than 2.

Each node in the network has a species divergence time and each edge b has an associated population mutation rate $\theta_b = 4N_b\mu$. The network has a special edge $er(\Psi) = (s,r)$, where r is the root of the network. This special edge is infinite in length so that all lineages that enter it coalesce on it eventually. 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. We use Ψ to refer to the topology, species divergence times and population mutation rates of the phylogenetic network.

An $\mathscr{X}$-phylogenetic tree, or $\mathscr{X}$-tree, is an $\mathscr{X}$-network with no reticulation nodes. A gene tree is an $\mathscr{X}$-tree. Each node in the gene tree has an associated coalescence time. In the algorithm below, we make use of a coloring function c: $(E(g),t) \rightarrow \{0,1\}$, similar to that used in (Bryant *et al.*, 2012), where $c(e,t)$ indicates the color, or allele, at time t along the branch e of gene tree g.

Notations

Bryant *et al.* devised an algorithm for exact computation of the likelihood of a species tree given bi-allelic markers. We extend the algorithm to compute the likelihood of a phylogenetic network given bi-allelic markers. To make connections to the SNAPP method as clear as possible, we use the notations from (Bryant *et al.*, 2012) and extend them for our purposes.

Looking forward in time (from the root toward the leaves), let u and v be the mutation rate from red allele to green allele and the mutation rate from green allele to red allele, respectively. The stationary distribution of the red and green alleles at the root is given by $v/(u+v)$ and $u/(u+v)$, respectively. Observed alleles are indicated by values of the coloring function c at gene tree leaves.

Given a gene history embedded within the branches of the network, the numbers and types of lineages at both ends of each branch of the network are needed to compute the likelihood. Let x be a branch in the phylogenetic network. We denote by n_x^T and n_x^B the total numbers of lineages at the top and bottom of x, respectively, and by r_x^T and r_x^B the numbers of red lineages at the top and bottom of x, respectively. See Fig. 1 for an illustration.

Labeled partial likelihoods

Let x be an arbitrary branch in the phylogenetic network and let $\mathcal{R}_x$ be the event that for every

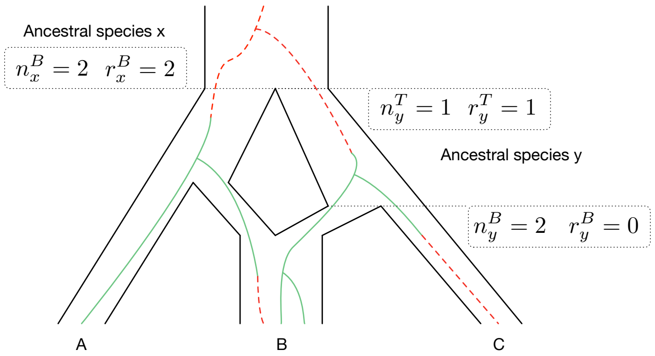

FIG. 1. Illustrating the "growth" of lineages of a gene tree in a phylogenetic network. The histories of green and red alleles are shown as solid (green) lines and dashed (red) lines, respectively.

external branch z that is a descendant of x, the actual number of red alleles in z equals to r_z^B.

Bryant *et al.* defined two partial likelihoods: $\mathbf{F}_x^B$ is the product of the likelihood of a subtree rooted at the bottom of x and the probability $Pr[n_x^B = n]$, and $\mathbf{F}_x^T$ is the product of the likelihood at the top of branch x and the probability $Pr[n_x^T = n]$. In the case of a species tree (i.e., no reticulation nodes in the species phylogeny), the partial likelihood vectors $\mathbf{F}_x^B$ and $\mathbf{F}_x^T$ are given by

$$\mathbf{F}_x^B(n,r) = Pr[\mathcal{R}_x | n_x^B = n, r_x^B = r] Pr[n_x^B = n] \quad (2)$$

and

$$\mathbf{F}_x^T(n,r) = Pr[\mathcal{R}_x | n_x^T = n, r_x^T = r] Pr[n_x^T = n]. \quad (3)$$

Here $\mathbf{F}_x^B$ and $\mathbf{F}_x^T$ are indexed by nonnegative integers n and r, where $r \leq n$. Let m be the maximum possible value of n_x^B and n_x^T over all branches. Then, each of $\mathbf{F}_x^B$ and $\mathbf{F}_x^T$ has at most $l = (1+(m+1))(m+1)/2$ entries.

In the case of a species tree, the path from a leaf to the root is unique. However, this might not be the case for phylogenetic networks: If there is a reticulation node on a path from a leaf to the root, then multiple paths exist between that leaf and the root. This is the issue that necessitates modifying the algorithm of (Bryant *et al.*, 2012) significantly, and that leads to much larger computational requirements in the case of phylogenetic networks. The key idea behind the modification is as follows. As the algorithm proceeds to compute the likelihood in a bottom-up fashion from the leaves to the root, whenever a reticulation node is encountered, the current set of lineages is bipartitioned in every possible way so that one side of the bipartition tracks one parent of the reticulation node and the other side tracks the other parent. As the network has a unique root, the two sides of each bipartition eventually come back together at an ancestral node. At that point, these two sides are merged properly.

To achieve this proper merger, we introduce "labeled partial likelihoods," or LPL. Given a phylogenetic network Ψ with k reticulation nodes numbered $0,1,\cdots,k-1$, an LPL P is an element of $\mathbb{R}^l \times \mathbb{Z}^k$, where the first element of the pair is a partial likelihood as in (Bryant *et al.*, 2012). The second element is the label to keep track of partial likelihoods that originated from a split of the same partial likelihood at a reticulation node so that these two could be merged. More formally, we say two LPLs $P_1 = (\mathbf{F}_1, \mathbf{s}_1)$ and $P_2 = (\mathbf{F}_2, \mathbf{s}_2)$, where $|\mathbf{s}_1| = |\mathbf{s}_2|$, are compatible if and only if for every $0 \leq i < |\mathbf{s}_1|$, either $\mathbf{s}_1(i) = \mathbf{s}_2(i)$ or $\mathbf{s}_1(i) \cdot \mathbf{s}_2(i) = 0$.

We denote by $\mathscr{P}_x^T$ and $\mathscr{P}_x^B$ the sets of LPLs that are associated with the top and bottom of branch x, respectively. These two quantities are computed in a bottom-up fashion, proceeding from the leaves of the network towards its root. Once the LPLs at the root are computed, the overall likelihood of a give site is computed. As the algorithm proceeds from the leaves towards the root, it needs to compute LPLs at the leaves, the top of a branch, the bottom of reticulation edges, and the bottom of tree edges. We now describe each of those computations; the overall algorithm is simply a bottom-up traversal of the network while applying the appropriate computation as a node is encountered.

Computing LPLs for leaf nodes

Consider an external branch x that is connected to a leaf node. Let n_x denote the number of individuals sampled from the species associated with that leaf, and let r_x be the number of red lineages among those individuals. We create LPL $P_x^B = (\mathbf{F}_x^B, \mathbf{s}_x^B)$, where

$$\mathbf{F}_x^B(n,r) = \begin{cases} 1, & \text{if } n = n_x \text{ and } r = r_x \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$\mathbf{s}_x^B = \mathbf{0}$. Finally, we associate $\mathscr{P}_x^B = \{P_x^B\}$ with the bottom of branch x.

As pointed out in (Bryant *et al.*, 2012), the input data may contain dominant markers like AFLPs, which means heterozygotes and homozygotes are not distinguishable for the dominant band. If there are dominant markers in the data, and the red allele is dominant, $\mathbf{F}_x^B$ is computed by

$$\mathbf{F}_x^B(n,r) = \begin{cases} \frac{n!}{(r-r_x)!(2r_x-r)!(n-r_x)!} 2^{2r_x-r} \binom{2n}{r}^{-1}, & \\ \qquad \text{if } n = 2n_x \text{ and } r_x \le r \le 2r_x & \\ 0, & \\ \qquad \text{otherwise} & \end{cases} . \tag{5}$$

instead of using Eq. (4).

Computing LPLs at the top of a branch

Bryant *et al.* computed partial likelihoods using a continuous-time Markov chain whose transition rate matrix $\mathbb{Q}$ is indexed by $((n,r);(n',r'))$ for transitioning from n lineages r of which are red alleles to n' individuals r' of which are red alleles, and its entries are given by

$$\begin{cases} \mathbb{Q}_{(n,r);(n,r-1)} = (n-r+1)v \\ \mathbb{Q}_{(n,r);(n,r+1)} = (r+1)u \\ \mathbb{Q}_{(n,r);(n-1,r)} = (n-1-r)n/\theta \\ \mathbb{Q}_{(n,r);(n-1,r-1)} = (r-1)n/\theta \\ \mathbb{Q}_{(n,r);(n,r)} = -n(n-1)/\theta - (n-r)v - ru \end{cases} . \tag{6}$$

Let x be any branch in the phylogenetic network, with θ and t being the population mutation rate and branch length of x, respectively, and assume $\mathscr{P}_x^T$ has already been computed. Then,

$$\mathscr{P}_x^T = \{(exp(\mathbb{Q}t)\mathbf{F},\mathbf{s}) : (\mathbf{F},\mathbf{s}) \in \mathscr{P}_x^B\}. \tag{7}$$

Computing LPLs at the bottom of reticulation edges

Consider a reticulation node given by two reticulation edges y and z, with inheritance probabilities γ and $1-\gamma$, respectively, and branch x emanating from the reticulation node, as illustrated by Fig. 2. The main idea in this part is as follows. Given a set of lineages at the top of branch x, a subset of those lineages is inherited along branch y and the remaining lineages is inherited along branch z. Since there are multiple ways of bipartitioning the set of lineages, the labels in an LPL allow the algorithm to keep track of the subsets of lineages that originated from the same split. We now describe this formally.

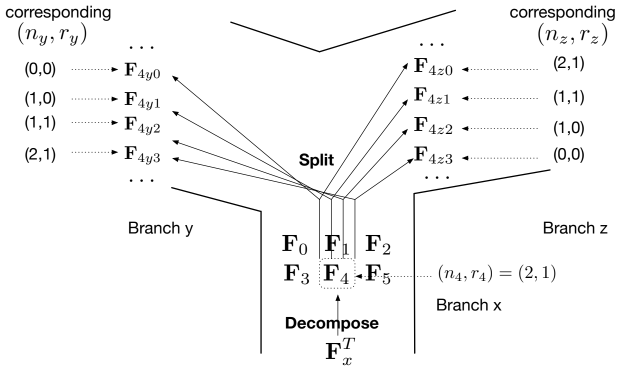

FIG. 2. Illustration of the decompose-and-split operation. In this example, partial likelihood $\mathbf{F}_x^T$ is decomposed into six vectors $\mathbf{F}_0$ to $\mathbf{F}_5$. An illustrating of how $\mathbf{F}_4$ is split in the four possible ways to trace branches y and z is shown, and every split is assigned a unique label.

Decomposing: Let $(\mathbf{F},\mathbf{s})$ be an LPL in $\mathscr{P}_x^T$. Given that $\mathbf{F}$ has l entries, we decompose $\mathbf{F}$ into l vectors, each with l entries: $\mathbf{F}_0$, $\mathbf{F}_1$, $\cdots$, $\mathbf{F}_{l-1}$. Let $\phi:\{(n',r'):n',r'\in\mathbb{N},r'\leq n'\leq m\}\to\mathbb{N}$ be given by $\phi(n',r')=n'(m+1)+r'$. The entries of $\mathbf{F}_i$ are set according to

$$\mathbf{F}_i(n',r')=\begin{cases}\mathbf{F}(n',r') & \text{if } i=\phi(n',r')\\ 0 & \text{otherwise}\end{cases}. \quad (8)$$

Splitting: Consider vector $\mathbf{F}_i$ and assume $i=\phi(n_i,r_i)$. The existence of n_i lineages out of which r_i are red at the top of branch x means that any $0\leq n_y\leq n_i$ lineages of those could be inherited along branch y, and out of those $0\leq r_y\leq n_y$ could be red; the remaining $n_z=n_i-n_y$ lineages, out of which $r_z=r_i-r_y$ are red, are inherited along branch z. Such a split gives rise to two LPLs: $P_y=(\mathbf{F}_y,\mathbf{s}_y)$ and $P_z=(\mathbf{F}_z,\mathbf{s}_z)$ with $\mathbf{s}_y$ and $\mathbf{s}_z$ assigned the same value that is unique to the specific split. For this specific split we define

$$\delta_i=\mathbf{F}_i(n_i,r_i)\binom{n_i}{n_y}\gamma^{n_y}(1-\gamma)^{n_i-n_y}, \quad (9)$$

and compute $\mathbf{F}_y$ and $\mathbf{F}_z$ by

$$\mathbf{F}_y(n,r)=\begin{cases}\delta_i, & \text{if } n=n_y \text{ and } r=r_y\\ 0, & \text{otherwise}\end{cases}, \quad (10)$$

and

$$\mathbf{F}_z(n,r)=\begin{cases}1, & \text{if } n=n_z \text{ and } r=r_z\\ 0, & \text{otherwise}\end{cases}. \quad (11)$$

The resulting P_y and P_z from all possible splits constitute the elements of the sets $\mathscr{P}_y^B$ and $\mathscr{P}_z^B$, respectively. The full procedure for executing the decompose-and-split operations is given in Algorithm 1.

Computing LPLs at the bottom of a tree edge

Consider an internal tree node j with its three associated edges $x=(u,j)$, $y=(j,v)$, and $z=(j,w)$. We are interested in computing the set $\mathscr{P}_x^B$ in terms of the two sets $\mathscr{P}_y^T$ and $\mathscr{P}_z^T$. The labels in LPLs allow the algorithm to determine whether two LPLs originated from a split at a descendant reticulation node or not (including the case of no descendant reticulation nodes of node j). Let $P_y=(\mathbf{F}_y,\mathbf{s}_y)$ and $P_z=(\mathbf{F}_z,\mathbf{s}_z)$ be two elements of $\mathscr{P}_y^T$

```
Input: Reticulation node $j$ and the three distinct
       branches $y=(u,j)$, $z=(v,j)$ and $x=(j,w)$
       associated with it; set $\mathscr{P}_x^T$.
Output: Sets $\mathscr{P}_y^B$, $\mathscr{P}_z^B$.
$o_j \leftarrow 0$;
$\mathscr{P}_y^B \leftarrow \emptyset$;
$\mathscr{P}_z^B \leftarrow \emptyset$;
foreach $(\mathbf{F},\mathbf{s}) \in \mathscr{P}_x^T$ do
    $l \leftarrow$ number of entries in $\mathbf{F}$;
    if $l=0$ then
        $\mathscr{P}_y^B \leftarrow \mathscr{P}_y^B \cup \{(\mathbf{F},\mathbf{s})\}$;
        $\mathscr{P}_z^B \leftarrow \mathscr{P}_z^B \cup \{(\mathbf{F},\mathbf{s})\}$;
    // decompose
    for $i=1 \rightarrow l-1$ do
        $(n_i,r_i) \leftarrow \phi^{-1}(i)$;
        Compute $\mathbf{F}_i$ using Eq. (8);
        // split
        foreach $0 \le n_y \le n_i$ and $0 \le r_y \le \min(r_i,n_y)$
        do
            $(n_z,r_z) \leftarrow (n_i-n_y, r_i-r_y)$;
            $\mathbf{s}_y \leftarrow \mathbf{s}$; $\mathbf{s}_z \leftarrow \mathbf{s}$;
            $\mathbf{s}_y(j) \leftarrow o_j+1$; $\mathbf{s}_z(j) \leftarrow o_j+1$;
            Compute $\mathbf{F}_y$ and $\mathbf{F}_z$ Eqs. (10) and (11);
            $\mathscr{P}_y^B \leftarrow \mathscr{P}_y^B \cup \{(\mathbf{F}_y,\mathbf{s}_y)\}$;
            $\mathscr{P}_z^B \leftarrow \mathscr{P}_z^B \cup \{(\mathbf{F}_z,\mathbf{s}_z)\}$;
            $o_j \leftarrow o_j+1$;
return $\mathscr{P}_y^B, \mathscr{P}_z^B$;
```

Algorithm 1: **Compute LPLs at Bottom of Reticulation Edges.**

and $\mathscr{P}_z^T$, respectively, that are compatible. A label $\mathbf{s}_x$ is computed by

$$\mathbf{s}_x(i)=\begin{cases}\mathbf{s}_y(i) & \text{if } \mathbf{s}_y(i)=\mathbf{s}_z(i) \text{ or } \mathbf{s}_z(i)=0\\ \mathbf{s}_z(i) & \text{if } \mathbf{s}_y(i)=0\end{cases} \tag{12}$$

for $0 \le i \le |\mathbf{s}_y|$. Furthermore, $\mathbf{F}_x$ is computed by

$$\mathbf{F}_x(n,r)=\sum_{n_y=1}^{n-1}\sum_{r_y=0}^{r}\mathbf{F}_y^T(n_y,r_y)\mathbf{F}_z^T(n-n_y,r-r_y) \times \frac{\binom{r}{r_y}\binom{n-r}{n_y-r_y}}{\binom{n}{n_y}}. \tag{13}$$

The LPL $(\mathbf{F}_x,\mathbf{s}_x)$ is added to $\mathscr{P}_x^B$. The full procedure for computing set $\mathscr{P}_x^B$ is give in Algorithm 2.

```
Input: Internal tree node $j$ and the three distinct
       branches $y=(j,v)$, $z=(j,w)$ and $x=(u,j)$
       associated with it; sets $\mathscr{P}_y^T$, $\mathscr{P}_z^T$.
Output: Set $\mathscr{P}_x^B$.
$\mathscr{P}_x^B \leftarrow \emptyset$;
foreach $(\mathbf{F}_y,\mathbf{s}_y) \in \mathscr{P}_y^T$ do
    foreach $(\mathbf{F}_z,\mathbf{s}_z) \in \mathscr{P}_z^T$ do
        if $\mathbf{s}_y$ and $\mathbf{s}_z$ are compatible then
            Compute $\mathbf{s}_x$ using Eq. (12);
            Compute $\mathbf{F}_x$ using Eq. (13);
            $\mathscr{P}_x^B \leftarrow \mathscr{P}_x^B \cup \{(\mathbf{F}_x,\mathbf{s}_x)\}$;
return $\mathscr{P}_x^B$;
```

Algorithm 2: **Compute LPLs at Bottom of Tree Edge.**

Termination: Computation above root node

Let the infinite-length branch associated with root be ρ. Then, we let $\mathbf{F}_\rho^B$ be the sum of all vectors $\mathbf{F}$ in elements $(\mathbf{F},\mathbf{s})$ of set $\mathscr{P}_\rho^B$.

To obtain the overall likelihood $\mathcal{L}(\Psi|S_i)$ given the data S_i for site i, vector $\mathbf{x}$ is obtained as a solution of $\mathbb{Q}\mathbf{x}=\mathbf{0}$, and the likelihood is computed by

$$\mathcal{L}(\Psi|S_i)=\mathbf{F}_\rho^B\cdot\mathbf{x}. \tag{14}$$

Optimizing the computation

As described above, the partial likelihood vectors are split to follow every possible way of bipartitioning a set of lineages at a reticulation node. It is this operation that leads to a significant increase in the running time and memory requirement of the likelihood computation as compared to the case of species trees. Here we describe an optimization step that we employ to improve performance in terms of computational requirements, without affecting the correctness of the likelihood computation.

An *articulation node* in a graph is a node whose removal disconnects the graph into two or more components. In a directed graph, a *lowest articulation node* is an articulation node that has at least one child that is not an articulation node. For example, in a tree, every node is an articulation node. However, in a phylogenetic network that is not necessarily the case. For example, in the phylogenetic network of Fig. 1, the reticulation node is an articulation node. However, the root of the network is the only lowest articulation node.

The main idea of the optimization is that all LPLs are a lowest articulation node could be merged into a single LPL, thus avoiding carrying forth all that information. More formally, given a set of LPLs at the bottom of a lowest articulation node, a new LPL is produced by summing all the partial likelihood vectors in the LPLs, and assigning it an empty label. This new LPL is the only one assigned to the bottom of the articulation node; all other LPLs are deleted.

Time complexity

Our algorithm computes the likelihood of a phylogenetic network given a set of biallelic markers. This algorithm computes matrix exponential along every branch, and processes the network's nodes in a post-order traversal. Computation at a leaf takes $O(1)$ time. At a tree node, computation is mostly spent on evaluating Eq. (13). Let n be the number of individuals present under an internal tree node. Then, this evaluation takes $O(n^4)$ time for a pair of compatible LPLs. The total time consumption of processing tree nodes also depends on the number of LPLs. Assuming k reticulation nodes in the phylogenetic network, there are at most $O(n^{4k})$ pairs of compatible LPLs. Therefore the time complexity of processing a tree node is $O(n^{4k+4})$.

At a reticulation node, the time consumption increases after each reticulation node is processed, due to the accumulation of (split) LPLs. In the last processed reticulation node, the number of LPLs in its descendant is at most $O(n^{4(k-1)})$. There are at most $O(n^4)$ new LPLs generated due to decompose-and-split operation for each original LPL. Therefore the time complexity of processing a reticulation node is at most $O(n^{4k})$. We adopted the same approximation of matrix exponential as in (Bryant *et al.*, 2012), so the time complexity of computing matrix exponentiation is $O(n^2)$, and computation along every branch is at most $O(n^{4k+2})$.

In total, the time complexity of the algorithm is $O(mn^{4k+4})$, where m is the number of species, n is the total number of lineages sampled from the species, and k is the number of reticulation nodes. Notice that when $k=0$, which means the species phylogeny is a tree, the time complexity is

$O(mn^4)$, which is the running time of the SNAPP algorithm without fast Fourier transform.

Bayesian inference

The prior on the phylogenetic network is the same as that employed in (Wen and Nakhleh, 2016). It is composed of the prior on the number of reticulation nodes (Poisson distribution), the prior on the diameters of reticulation nodes (Exponential distribution), the prior on the species divergence times (Exponential distribution), and the prior on the population mutation rate (Gamma distribution). For the prior on the population mutation rate, we use the Gamma distribution $\Gamma(2,\psi)$ with mean value 2ψ and shape parameter 2. We also used the non-informative prior $P_\psi(x)=1/x$ for hyper-parameter ψ. For the prior on the inheritance probabilities, we use Beta(α,β). Unless there is some specific knowledge on the inheritance probabilities, a uniform prior on [0,1] is adopted by setting $\alpha=\beta=1$.

We employed the reversible-jump MCMC, or RJMCMC (Green, 1995) algorithm implemented in PhyloNet (Than *et al.*, 2008) to sample from the posterior distribution given by

$$p(\Psi|S)\propto\mathcal{L}(\Psi|S)p(\Psi)$$

where Ψ here denotes the topology of the network and all its parameters, and $p(\Psi)$ is the prior on the network and its parameters. We only make use of the 12 proposals designed for sampling phylogenetic networks and inheritance probabilities described in (Wen and Nakhleh, 2016), but not the proposals aimed at sampling gene trees, as gene trees are integrated out.

Synthetic data generation

We implemented a program to simulate bi-allelic markers on a given phylogenetic network. Bryant *et al.* simulated bi-allelic markers by first generating gene trees inside a species tree (under the multispecies coalescent model), and then simulating the markers down the gene trees. In our case, we replaced the first step by generating gene trees inside a phylogenetic network under the multispecies network coalescent (Yu *et al.*, 2014); the second step of simulating bi-allelic markers down gene trees remains the same as that employed in (Bryant *et al.*, 2012) . When requiring the data set to contain only polymorphic sites, if the generated site is not polymorphic, we discard both gene tree and markers, and repeat until a polymorphic site is generated.

We used following commands in PhyloNet to generate four data sets to exam the ability to recover topology of our Bayesian inference. Each of these commands was also repeated with “-num 10000”, “-num 100000”, “-num 1000000”, instead of “-num 1000”, for different numbers of sites.

```
SimBiMarkersinNetwork -pi0 0.5 -sd 12345678 -num
    1000 -cu 0.036 -truenet "(((((A:0.7)I6#H1
    :1.3::0.8,Q:2.0)I4:1.0,L:3.0)I3:1.0,R:4.0)I2
    :1.0,(G:2.0,(I6#H1:0.7::0.2,C:1.4)I5:0.6)I1
    :3.0)I0;"
```

```
SimBiMarkersinNetwork -pi0 0.5 -sd 12345678 -num
    1000 -cu 0.036 -truenet "(((((Q:2.0,A:2.0)I4
    :1.0,L:3.0)I3:0.5)I8#H1:0.5::0.7,R:4.0)I2
    :1.0,(I8#H1:0.5::0.3,(G:2.0,C:2.0)I1:2.0)I7
    :1.0)I0;"
```

```
SimBiMarkersinNetwork -pi0 0.5 -sd 12345678 -num
    1000 -cu 0.036 -truenet "(((((Q:0.5)I8#H1
    :0.5::0.7,A:1.0)I4:1.0,L:2.0)I3:2.0,(I8#H1
    :1.0::0.3,R:1.5)I7:2.5)I2:1.0,(G:1.0,C:1.0)I1
    :4.0)I0;"
```

```
SimBiMarkersinNetwork -pi0 0.5 -sd 12345678 -num
    1000 -cu 0.036 -truenet "(((((Q:0.5)I8#H1
    :0.5::0.7,(A:0.5)I6#H2:0.5::0.8)I4:1.0,L:2.0)
    I3:2.0,(I8#H1:1.0::0.3,R:1.5)I7:2.5)I2:1.0,((
    I6#H2:0.5::0.2,C:1.0)I5:1.0,G:2.0)I1:3.0)I0;"
```

The true networks of those commands correspond to four models, given by the four phylogenetic networks, their branch lengths, and inheritance probabilities, shown in Fig. 3. These networks and parameters were inspired by the phylogenetic networks inferred from six mosquito genomes in (Fontaine *et al.*, 2015; Wen *et al.*, 2016b). For each of the four models, we simulated data sets consisting of 1000, 10000, 100000, and 1000000 bi-allelic sites. In the simulations, we set $u=1$ and $v=1$ as the mutation rate. Furthermore, we used $\theta=0.036$ as the population mutation rate in the unit of population mutation rate per site. Under these settings, we observed that each of the 16 data sets contained between 34% and 37.5% polymorphic sites; the remaining sites were all monomorphic.

We also used following command in PhyloNet to generate one data set to test the robustness to the misspecification of the value of the hyperparameter ψ of our algorithm to recover continuous parameters. The true value for ψ is 0.018.

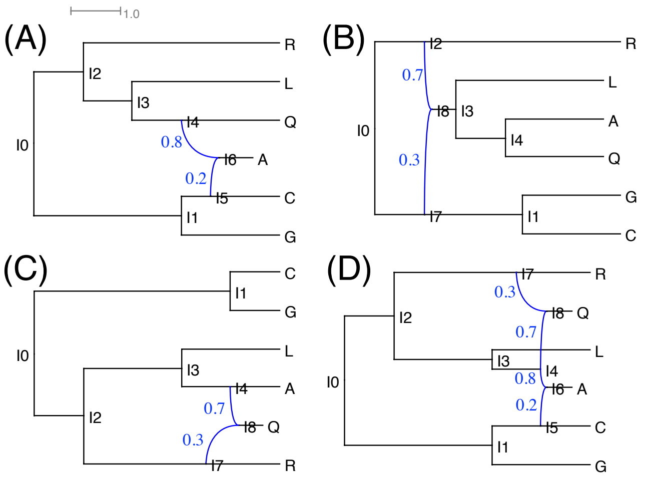

FIG. 3. The four model phylogenetic networks used to generate the simulated data sets. The branch lengths of the phylogenetic networks are measured in coalescent units (scale is shown). The inheritance probabilities are marked in blue. All four networks are based on the same "backbone" tree: If for every pair of reticulation edges sharing the same reticulation node the edge with the smaller inheritance probability is removed, all four networks give rise to the tree ((C,G),(R,(L,(A,Q)))). The hybridization events in each of the four panels can be viewed as involving pairs of branches of this tree. (A) The hybridization is from C to A. (B) The hybridization is from (C,G) to (L,(A,Q)). (C) The hybridization is from R to Q. (D) One hybridization is from R to Q and another from C to A.

```
SimBiMarkersinNetwork -diploid -pi0 0.5 -sd
    12345678 -num 10000 -cu 0.036 -truenet "(((((
    A:0.7)I6#H1:1.3::0.8,Q:2.0)I4:1.0,L:3.0)I3
    :1.0,R:4.0)I2:1.0,(G:2.0,(I6#H1:0.7::0.2,C
    :1.4)I5:0.6)I1:3.0)I0;"
```

We considered the network in Fig. 3(A) to show the ability of our algorithm to estimate the continuous parameters (branch lengths, inheritance probabilities, and population mutation rates) given different values of the hyperparameter ψ. In this case, we assumed two individuals for each taxon and generated 10000 bi-allelic sites using our simulator.

Monomorphic sites help estimate parameter values, but sometimes they are removed because they are uninformative for estimating the topology and to reduce the computation time for the phylogenetic analyses. If there are only polymorphic sites in the data set, sampling multiple individuals could improve parameter estimation. To investigate this aspect, we set up a simulation with the phylogenetic network in Fig. 4. In the simulation, we set $u=1$ and $v=1$ as the mutation rates. Furthermore, we used $\theta=0.005$. We sampled one diploid individual for each of the three species A, B, and D, and four diploid individuals for species C. We generated 10000 polymorphic sites with dominant markers for each of those individuals.

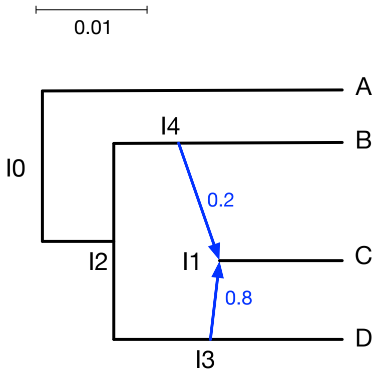

FIG. 4. The phylogenetic network used to investigate effect of multiple individuals. The branch lengths of the phylogenetic networks are measured in units of expected number of mutations per site. The inheritance probabilities are marked in blue.

We used following command in PhyloNet to generate the data set:

```
SimBiMarkersinNetwork -diploid -dominant -op -pi0
    0.5 -sd 123456 -num 10000 -tm <A:A_0; B:B_0;
    C:C_0,C_1,C_2,C_3; D:D_0> -truenet
    "[0.005]((((C:0.005:0.005)I1#H1
    :0.006:0.005:0.8,D:0.011:0.005):0.009:0.005,(
    B:0.014:0.005,I1#H1:0.009:0.005:0.2)
    :0.006:0.005):0.005:0.005,A:0.025:0.005);"
```

We ran the method on the entire data set (7 diploid individuals, amounting to 14 haploid individuals), and on a subset that consists of a single diploid individual from each of the four species (8 haploids in total).

Results and Discussion

Simulations

The method’s ability to recover the phylogenetic network topology

To test the ability of our algorithm to recover the topology of the true phylogenetic network, we ran the RJMCMC sampler on simulated data sets consisting of 1000, 10000, 100000, and 1000000 bi-allelic sites of the four phylogenetic networks in Fig. 3. We ran an MCMC chain for 1.5×10^6 iterations, and one sample was collected from every 500 iterations in the last 5×10^5 iterations. While sampling topologies, inheritance probabilities and branch lengths of the phylogenetic network, we assume a correct population mutation rate along every branches.

Before we discuss the quality of the sampled networks, we introduce the notion of a “backbone tree.” Given a phylogenetic network with inheritance probabilities on its reticulation edges, removing for each reticulation node the incoming edge with the smaller inheritance probability results in a tree, which we call the backbone tree. For example, for the network in Fig. 3(D), the reticulation edges with inheritance probabilities 0.2 and 0.3 would be removed, resulting in the backbone tree ((G,C),((L,(A,Q)),R)).

For each data set and collected samples from the RJMCMC results, we computed the 95% credible set of phylogenetic networks and their parameters. The results were as follows:

- Data corresponding to the phylogenetic network of Fig. 3(A):
  - For the 1000-site data set, 85.0% in the 95% credible set consist of the backbone tree of the true phylogenetic network; the remaining topologies were all trees that differed from the backbone tree. In other words, using 1000 sites, the true network was not recovered.
  - For all other three data sets, the 95% credible sets contain only the true phylogenetic network topology.
- Data corresponding to the phylogenetic network of Fig. 3(B):
  - For the 1000-site data set, 85.1% in the 95% credible set consist of the backbone tree of the true phylogenetic network; the remaining topologies were all trees that differed from the backbone tree. In other words, using 1000 sites, the true network was not recovered.
  - For the 10000-site data set, the 95% credible set contains only the backbone tree of the true phylogenetic network. In other words, using 10000 sites, the true network was not recovered.
  - For the other two data sets, the 95% credible sets contain only the true phylogenetic network topology.
- Data corresponding to the phylogenetic network of Fig. 3(C):
  - For the 1000-site data set, 81.0% in the 95% credible set consist of the backbone tree of the true phylogenetic network; the remaining topologies were all trees that differed from the backbone tree. In other words, using 1000 sites, the true network was not recovered.
  - For all other three data sets, the 95% credible sets contain only the true phylogenetic network topology.
- Data corresponding to the phylogenetic network of Fig. 3(D):
  - For the 1000-site data set, 29.1% of the 95% credible set consist of the backbone tree of the true phylogenetic network; the remaining topologies were all trees that differed from the backbone tree. In other words, using 1000 sites, the true network was not recovered.
  - For all other three data sets, the 95% credible sets contain only the true phylogenetic network topology.

These results indicate a very good performance of the method. First, as the number of sites increases, the ability of the method to recover the true network improves. In particular, in all cases, the method was able to recover the true network topology when using more than 10,000 sites.[1] Second, even for small data sets (in terms of the number of sites), when the method fails to recover the true network, it recovers the backbone tree of the network. That is, the method misses the reticulation signal. This is not unexpected. Given that percentage of polymorphic sites in the data (an average of around 36%), and the low inheritance probability on the reticulation edges not present in the backbone trees, this implies that very few, if any at all, polymorphic sites in a data set of 1000 sites support reticulation edges with the low inheritance probability. In other words, it could very well be the case that there is no signal at all for recovering those reticulation edges. Third, we observe that only in the case of the phylogenetic network of Fig. 3(B) that the method does not recover the network from 10,000 sites. In this network, the reticulation is much deeper in the phylogeny (immediately after the split from the root) than in the other three model networks. Ancient reticulation events are in general much harder to detect than newer ones. In the case of bi-allelic markers in particular, most of the mutations could have happened after this reticulation and there is hardly any signal for recovering it.

---

[1] It is worth noting that while not many empirical AFLP- or SNP-based studies currently include as many as 10,000 loci, such large data sets may become commonplace as genomic technologies continue to advance.

This simulation was performed on NOTS (Night Owls Time-Sharing Service), which is a batch scheduled High-Throughput Computing (HTC) cluster. We used 4 cores, with two threads per core running at 2.6GHz, and 4G RAM per thread. The runtimes, in hours, for analyzing the 1000-, 10000-, 100000-, and 1000000-site data sets, respectively, on each of the four networks in Fig. 3 were as follows. The network of Fig. 3(A): 0.9, 2.0, 2.0, 2.1; The network of Fig. 3(B): 0.9, 1.1, 5.3, 5.1; The network of Fig. 3(C): 0.9, 1.8, 1.8, 2.2; The network of Fig. 3(D): 1.0, 21.0, 5.3, 6.6.

The method's ability to recover the continuous parameters

The analysis was run twice: the first time it was fed the correct starting value for ψ (0.018), and the second time it was fed an incorrect starting value for ψ (0.0018). Each time we let the sampler sample the value of ψ, and we ran an MCMC chain for 1.5×10^6 iterations, with 5×10^5 burn-in iterations, one sample was collected from every 500 iterations.

The posterior distribution of branch lengths for the data is shown in Fig. 5. The posterior distribution of population mutation rate is shown in Fig. 6. The posterior distribution of inheritance probability is shown in Fig. 7.

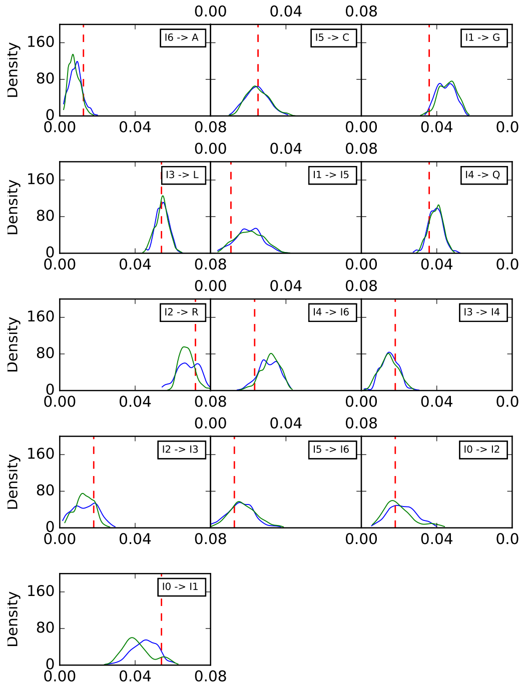

FIG. 5. The posterior distribution of branch lengths using our method on the simulated data set of the phylogenetic network of Fig. 3(A). Blue: Correct starting prior. Green: Incorrect starting prior. The red dashed lines correspond to the true values.

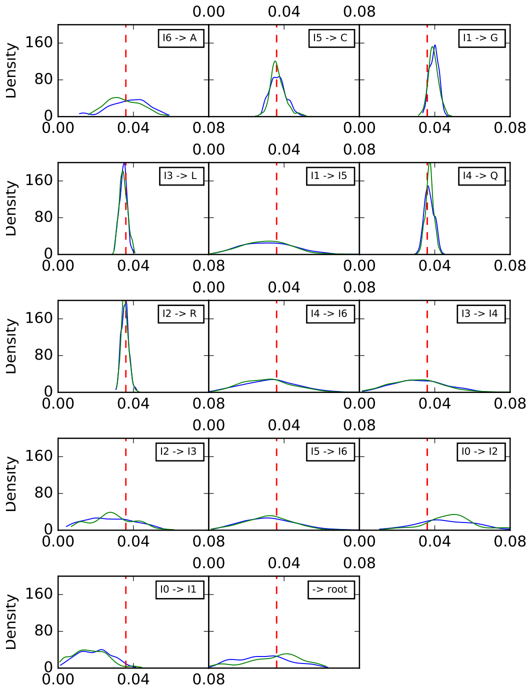

FIG. 6. The posterior distribution of population mutation rates using our method on the simulated data set of the phylogenetic network of Fig. 3(A). Blue: Correct starting prior. Green: Incorrect starting prior. The red dashed lines correspond to the true values.

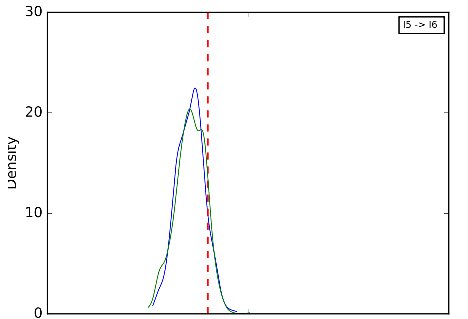

FIG. 7. The posterior distribution of inheritance probability using our method on the simulated data set of the phylogenetic network of Fig. 3(A). Blue: Correct starting prior. Green: Incorrect starting prior. The red dashed lines correspond to the true value.

These results indicate a very good performance of the method in terms of the robustness to misspecification of the hyperparameter to estimate parameter values. The blue and green curves are in very good agreement, indicating that our method recovered very similar posterior distributions of parameters given two different starting values of ψ. In other words, our method is robust to the (mis)specification of the hyperparameter ψ. Second, the posterior distributions of parameters fit well with the true value, which is marked by red dashed lines. This further demonstrates the robustness of our method, because the true parameters are correctly recovered under both correct and incorrect specifications of the hyperparameter ψ. However, the posterior distributions in Fig. 6 are widespread for the branches near the root. The reason is that those deep branches are where the mutation signal is very weak, if at all existent.

This simulation was performed on NOTS. We used 16 cores, with two threads per core running at 2.6GHz, and 4G RAM per thread. The runtime for analyzing the data set is about 3.7 hours.

The effect of the number of sampled individuals on parameter estimates

We ran each test using an MCMC chain for 1.0×10^6 iterations, with 5×10^5 burn-in iterations, and one sample was collected from every 500 iterations.

The posterior distribution of branch lengths for the data is shown in Fig. 8. The posterior distribution of population mutation rates is shown in Fig. 9. The posterior distribution of inheritance probability is shown in Fig. 10.

These results show that the method's performance improves as more individuals are sampled from the hybrid species. The biggest improvement is achieved for the branch length and population mutation rate estimates of branch

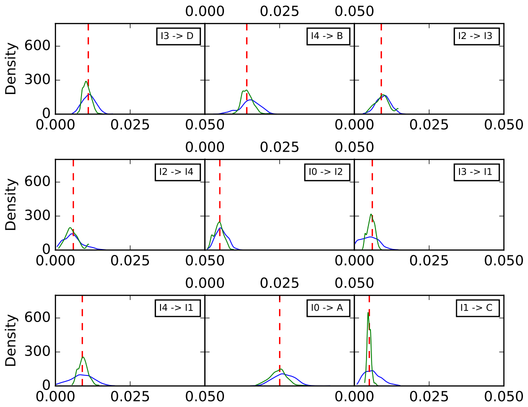

FIG. 8. The posterior distribution of branch lengths using our method on the simulated data set of the phylogenetic network of Fig. 4. In all cases, a single individual was sampled from A, B, and D. Blue: A single individual is sampled from C. Green: Four individuals are sampled from C. The red dashed lines correspond to the true values.

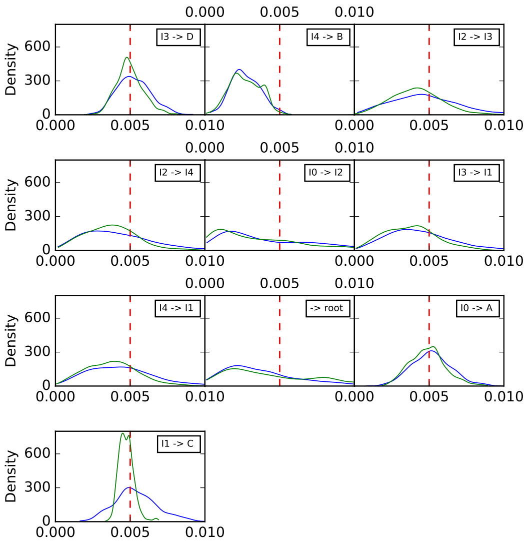

FIG. 9. The posterior distribution of population mutation rates using our method on the simulated data set of the phylogenetic network of Fig. 4. In all cases, a single individual was sampled from A, B, and D. Blue: A single individual is sampled from C. Green: Four individuals are sampled from C. The red dashed lines correspond to the true values.

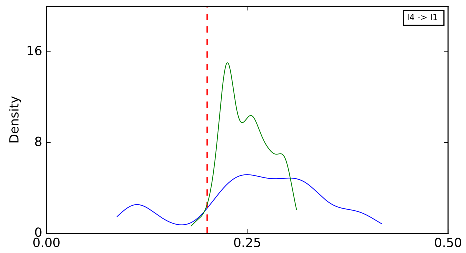

FIG. 10. The posterior distribution of inheritance probability using our method on the simulated data set of the phylogenetic network of Fig. 4. In all cases, a single individual was sampled from A, B, and D. Blue: A single individual is sampled from C. Green: Four individuals are sampled from C. The red dashed lines correspond to the true value.

"I1→C". In particular, when all four individuals are sampled from C, the posterior distributions become more concentrated around the true values. Furthermore, the branch length estimates of those branches adjacent to "I1→C" are also improved given more individuals of C. The population mutation rates are not improved much except for the branch "I1→C". The inheritance probability estimates also improve when four individuals are sampled, as the posterior samples become more concentrated and peak much closer to the true value.

This simulation was performed on NOTS. We used 8 cores, with two threads per core running at 2.6GHz, and 4G RAM per thread. The runtime for analyzing the full data set with four individuals sampled from C is 23.3 hours. The runtime for analyzing the subset with a single individual sampled from C is 0.5 hour. This shows the drastic effect of the number of individuals sampled on the running time of the method.

Analysis of an empirical data set

Two small subsets of a larger AFLP data set of multiple New Zealand species of the plant genus *Ourisia* (Plantaginaceae) (Meudt *et al.*, 2009) were analyzed, including previously unpublished AFLP profiles from two different hybrid individuals *O.* × *cockayneana* and *O.* × *prorepens* (herbarium codes follow Thiers [continuously updated]) . There is both morphological (Meudt, 2006) and molecular

(Meudt unpubl.) data supporting the hybrid nature of these two individuals. Although other *Ourisia* hybrid combinations have been reported in New Zealand (Meudt, 2006), *O.* × *cockayneana* and *O.* × *prorepens* are perhaps the most common, both involve *O. caespitosa* as a putative parent, and both have been formally named. Each data subset comprised five diploid individuals in total, which means ten haploid individuals were effectively analyzed due to the correction for dominant markers. A Poisson distribution with $\lambda=2$ as the prior of the number of reticulations was adopted. An MCMC chain was run on each data subset for 1.5×10^6 iterations, and sampled from every 500 iterations in the last 90% of iterations.

Data subset with hybrid O. × *cockayneana*

The first data subset comprises the following five individuals: *O. macrocarpa* (voucher: *Meudt 133a*, MPN 29546; herbarium codes follow Thiers [continuously updated]), *O. macrophylla* subsp. *lactea* (*Cameron 13392*, AK 294893), hybrid *O.* × *cockayneana* (*Meudt 175a*, MPN 29710), *O. caespitosa* (*Meudt 174a*, MPN 29705), and *O. calycina* (*Meudt 176a*, MPN 29713). The number of loci in this data set is 802.

The phylogenetic network with highest posterior probability is shown in Fig. 11. Other topologies in the 95% credible set have different ways of rooting the network, but all topologies successfully detected the hybrid and its putative parents. If the hybrid is removed, the topology in Fig. 11 also agrees with that of Fig. 3 in (Meudt *et al.*, 2009).

It should be noted that the posterior standard deviations reported in Fig. 11 is much larger than those in (Bryant *et al.*, 2012). This is perhaps not unexpected because we only used one individual per species in our analysis. Our simulation study shows that increased sampling of individuals helps the estimation of parameters, whereas when only one individual per species is sampled, the posterior distribution is much larger.

Data subset with hybrid O. × *prorepens*

The second data subset comprises *O. sessilifolia* subsp. *splendida* (*Heenan s.n.*, MPN 32149), *O. macrocarpa* (*Meudt 133a*, MPN 29713), hybrid *O.* × *prorepens* (*Meudt 203ª*, MPN 29774), *O. sessilifolia* subsp. sessilifolia (*Meudt 199a*, MPN 29771), and *O. caespitosa* (*Meudt 196a*, MPN 297695). The number of loci in this data set is 820.

The phylogenetic network with highest posterior probability is shown in Fig. 12. The result shows our method successfully detected the hybird and its putative parents. If the hybrid is removed, the topology in Fig. 12 also agrees with that of Fig. 3 in (Meudt *et al.*, 2009). As with the first data subset, the posterior standard deviations reported in Fig. 12 are large.

Nevertheless, the mean values of inferred parameters are very similar for the two species

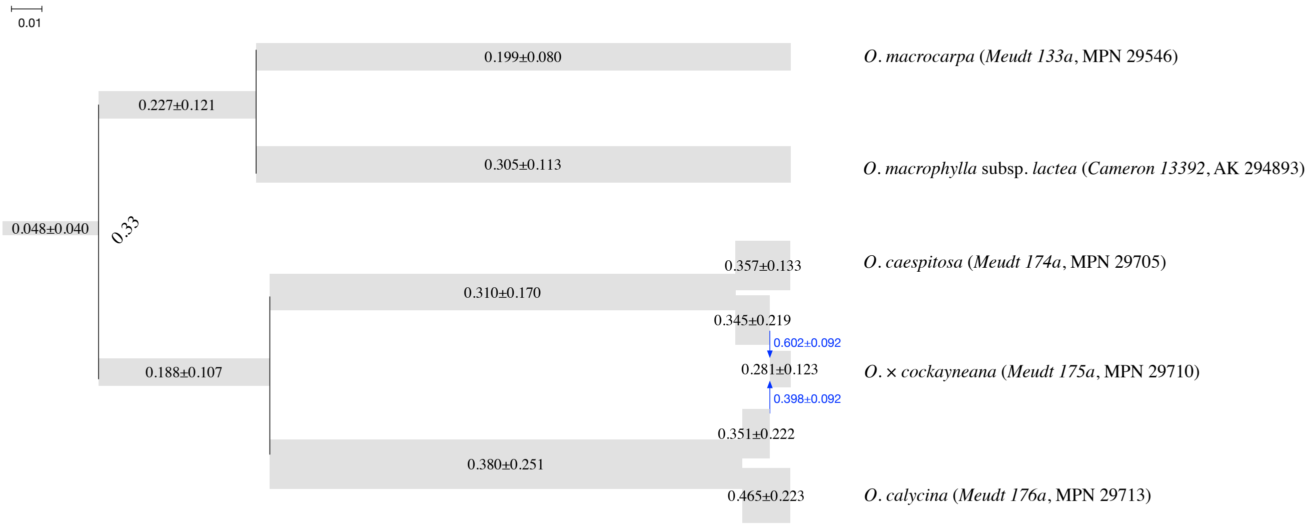
0.01
0.199±0.080
0.227±0.121
0.305±0.113
0.048±0.040
0.33
0.357±0.133
0.310±0.170
0.345±0.219
0.602±0.092
0.281±0.123
0.188±0.107
0.398±0.092
0.351±0.222
0.380±0.251
0.465±0.223
O. macrocarpa (*Meudt 133a*, MPN 29546)
O. macrophylla subsp. *lactea* (*Cameron 13392*, AK 294893)
O. caespitosa (*Meudt 174a*, MPN 29705)
O. × *cockayneana* (*Meudt 175a*, MPN 29710)
O. calycina (*Meudt 176a*, MPN 29713)

FIG. 11. Phylogenetic network with highest posterior distribution for the subset with the hybrid *O.* × *cockayneana* (*Meudt 175a*, MPN 29710) and putative parents. The width of each tube is proportional to the population mutation rate of each branch, which is printed on each tube. The length of each tube is proportional to the length of the corresponding branch in units of expected number of mutations per site (scale shown). Blue arrows indicate the reticulation edges and their inheritance probabilities are printed in blue.

that were common to the two data subsets, *O. caespitosa* and *O. macrocarpa*. The mean value of inferred population mutation rate of their corresponding leaves are similar. This shows that the method is both robust and consistent.

In summary, our method was able to extract the signal of the hybrid and successfully recover its putative parents, as well as reconstruct network topologies which were consistent with a previous study of a larger dataset (Meudt *et al.*, 2009).

Conclusions

Phylogenetic networks allow for representing evolutionary relationships that involve both vertical and horizontal transmission of genetic material. Extensions of the multispecies coalescent process to include hybridization events have facilitated the development of statistical methods for inferring and analyzing phylogenetic networks from gene tree estimates and sequence data. A major challenge with using gene tree estimates as the input to species phylogeny inference methods is the error in these estimates. While using the sequence data directly overcomes this issue, the problem of recombinations within loci can confound inferences. Using bi-allelic markers from individual, independent loci could provide a way to avoid both the gene tree uncertainty and recombination problems (the two are not necessarily independent). Furthermore, it is important to note that many biological studies use data sets that consists of bi-allelic markers and no available sequence alignment data for individual loci.

Bryant *et al.* recently devised an algorithm for inferring species trees from bi-allelic genetic markers while analytically integrating out the

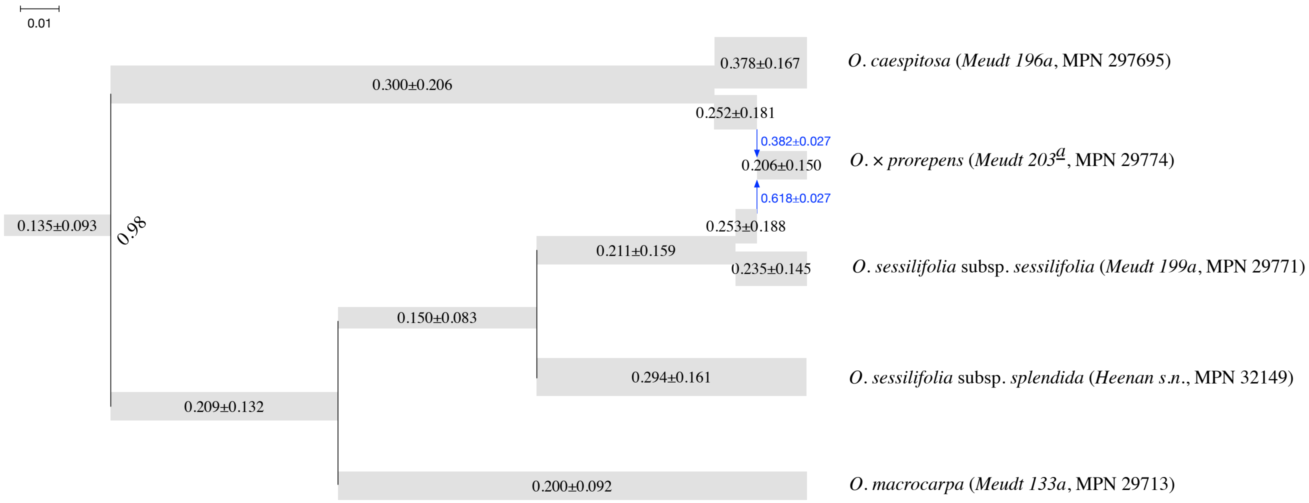
0.01
0.378±0.167
0.300±0.206
0.252±0.181
0.382±0.027
0.206±0.150
0.618±0.027
0.135±0.093
0.98
0.253±0.188
0.211±0.159
0.235±0.145
0.150±0.083
0.294±0.161
0.209±0.132
0.200±0.092
O. caespitosa (*Meudt 196a*, MPN 297695)
O. × *prorepens* (*Meudt 203$^{\underline{a}}$*, MPN 29774)
O. sessilifolia subsp. *sessilifolia* (*Meudt 199a*, MPN 29771)
O. sessilifolia subsp. *splendida* (*Heenan s.n.*, MPN 32149)
O. macrocarpa (*Meudt 133a*, MPN 29713)

FIG. 12. Phylogenetic network with highest posterior distribution for the subset with the hybrid *O.* × *prorepens* (*Meudt 203$^{\underline{a}}$*, MPN 29774) and putative parents. The width of each tube is proportional to the population mutation rate of each branch, which is printed on each tube. The length of each tube is proportional to the length of the corresponding branch in units of expected number of mutations per site (scale shown). Blue arrows indicate the reticulation edges and their inheritance probabilities are printed in blue.

gene trees for the individual loci (Bryant *et al.*, 2012). In this paper, we extended their algorithm significantly so as the likelihood of a phylogenetic network given bi-allelic markers could be computed while integrating out the gene trees. This method complements existing ones that use gene tree estimates or sequence alignments as input for statistical inference of phylogenetic networks.

We implemented a Bayesian method for sampling the posterior of phylogenetic networks and their associated parameters from bi-allelic data, and studied its performance on both simulated and empirical data. The results indicate a very good performance of the method. This work adds a powerful method to the biologist's toolbox that allows for estimating reticulate evolutionary histories.

A major bottleneck of the method is its computational requirements. While the SNAPP method is very time consuming on species trees, our method is much more time consuming given that reticulations in the phylogenetic network give rise to an explosion of the number of partial likelihoods that need to be computed and stored. More generally, the number of taxa in a data set has more of an effect on the running time of the method than the number of loci does. In particular, two aspects of the phylogenetic network under consideration affect the computational requirements of the method: The number of leaves under the reticulation nodes and the diameter of each of the reticulation nodes. As discussed above, the set of lineages entering a reticulation node must be bipartitioned in every possible way. This number of lineages is dependent on the number of leaves under that reticulation node. For example, if a single individual is sampled from a single species that exist under the reticulation node, then the number of bipartitions is very small (only two bipartitions exist). However, if n individuals are sampled from a single species that exist under the reticulation node or one individual is sampled per n species that exist under the reticulation node, then a number of bipartitions on the order of 2^n arises. This computation becomes much more demanding if there are more reticulation nodes on the path to a lowest articulation node. As for the diameter—which is the number of branches on the paths between the two parents of the reticulation node and a lowest articulation node above them, the larger its value, the more demanding the computation becomes. An important direction for future research is improving the computational requirements of the method to scale up to data sets with many taxa.

FUNDING

This work was supported by grants DBI-1355998 and CCF-1302179 from the National Science Foundation. This work was also supported in part by the Big-Data Private-Cloud Research Cyberinfrastructure MRI-award funded by NSF under grant CNS-1338099 and by Rice University.

References

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

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

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8): 1917–1932.

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., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leache, A. D., Liu, L., and David, C. C. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94: 447–462.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 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.

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

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.

Liu, K., Steinberg, E., Yozzo, A., Song, Y., Kohn, M., and Nakhleh, L. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences*, 112(1): 196–201.

Liu, L. and Pearl, D. K. 2007. Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3): 504–514.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5): 229–237.

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

Mallet, J., Besansky, N., and Hahn, M. 2016. How reticulated are species? *BioEssays*, 38(2): 140–149.

Meudt, H. M. 2006. Monograph of *Ourisia* (Plantaginaceae). *Systematic Botany Monographs*, 77: 1–188.

Meudt, H. M., Lockhart, P. J., and Bryant, D. 2009. Species delimitation and phylogeny of a New Zealand plant species radiation. *BMC Evolutionary Biology*, 9(1): 111.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. 2015. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6): 359–371.

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.

Rieseberg, L. 1997. Hybrid origins of plant species. *Annual Reviews of Ecology, Evolution and Systematics*, 28: 359–389.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J. L., Schwarzbach, A. E., Donovan, L. A., and Lexer, C. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, 301(5637): 1211–1216.

Solís-Lemus, C. and Ané, C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3): e1005896.

Springer, M. S. and Gatesy, J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94: 1–33.

Stevison, L. and Kohn, M. 2009. Divergence population genetic analysis of hybridization between rhesus and cynomolgus macaques. *Molecular Ecology*, 18(11): 2457–2475.

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

Thiers, B. [continuously updated]. Index herbariorum: A global directory of public herbaria and associated staff. New York Botanical Gardens Virtual Herbarium.

Wen, D. and Nakhleh, L. 2016. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. *bioRxiv*, page 095539.

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.

Yu, Y. and Nakhleh, L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*.

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., 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.

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. 2017. Bayesian inference of species networks from multilocus sequence data. *bioRxiv*, page 124982.

Zhang, W., Dasmahapatra, K. K., Mallet, J., Moreira, G. R., and Kronforst, M. R. 2016. Genome-wide introgression among distantly related heliconius butterfly species. *Genome Biology*, 17: 25.