Bayesian Inference of Species Networks from Multilocus Sequence Data

Chi Zhang[1,2,*], Huw A. Ogilvie[3,4], Alexei J. Drummond[4,5], Tanja Stadler[1,2,*]

April 6, 2017

[1] *Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zürich, 4058 Basel, Switzerland*
[2] *Swiss Institute of Bioinformatics (SIB), Switzerland*
[3] *Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia*
[4] *Centre for Computational Evolution, University of Auckland, Auckland, New Zealand*
[5] *Department of Computer Science, University of Auckland, Auckland, New Zealand*
[*] *Corresponding author: E-mail: tanja.stadler@bsse.ethz.ch or chi.zhang@bsse.ethz.ch*

Abstract

Reticulate species evolution, such as hybridization or introgression, is relatively common in nature. In the presence of reticulation, species relationships can be captured by a rooted phylogenetic network, and orthologous gene evolution can be modeled as bifurcating gene trees embedded in the species network. We present a Bayesian approach to jointly infer species networks and gene trees from multilocus sequence data. A novel birth-hybridization process is used as the prior for the species network. We assume a multispecies network coalescent (MSNC) prior for the embedded gene trees. We verify the ability of our method to correctly sample from the posterior distribution, and thus to infer a species network, through simulations. We reanalyze a large dataset of genes from closely related spruces, and verify the previously suggested homoploid hybridization event in this clade. Our method is available within the BEAST 2 add-on `SpeciesNetwork`, and thus provides a general framework for Bayesian inference of reticulate evolution.

Keywords: reticulate evolution, hybridization, multispecies coalescent, incomplete lineage sorting

1 Introduction

Hybridization during speciation is relatively common in animals and plants (Mallet, 2005, 2007). However, when reconstructing the evolutionary history of species, typically non-reticulating species trees are inferred (Guindon et al., 2010; Stamatakis, 2014; Drummond and Bouckaert, 2015; Ronquist et al., 2012), and the potential for hybridization events is ignored.

To account for the distribution of evolutionary histories of genes inherited from multiple ancestral species, the multispecies coalescent model (Rannala and Yang, 2003) was extended to allow reticulations among species, named multispecies network coalescent (MSNC) model (Yu et al., 2014). Orthologous genes are modeled as gene trees embedded in the species network. The MSNC model accounts for gene tree discordance due to incomplete lineage sorting and reticulate species evolution events, such as hybridization or introgression. There have been computational methods developed based on the MSNC to infer species networks using maximum likelihood (Yu et al., 2014; Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2016) and Bayesian inference (Wen et al., 2016). These methods use gene trees inferred from other resources as input. Due to the model complexity, applying the MSNC model in a full Bayesian framework, i.e., to infer the posterior distribution of

species network and gene trees directly from the multilocus sequence data, is challenging. Recently Wen and Nakhleh (2016) have developed a Bayesian method that can co-estimate species networks and gene trees from multilocus sequence data, but a process-based prior for the species network is still lacking. Their method also integrates over all possible gene tree embeddings at each MCMC step, which means that the estimated histories of individual gene trees within the species network are not available for subsequent analysis, and is limited to the JC69 (Jukes and Cantor, 1969) and GTR (Tavaré, 1986) substitution models.

In this paper, we present a Bayesian method to infer ultrametric species networks jointly with gene trees and their embeddings from multilocus sequence data. Our method assumes a birth-hybridization model for the species network, the MSNC model for the embedded gene trees with analytical integration of population sizes, and employs novel MCMC operators to sample the species network and gene trees along with associated parameters. It is able to use the full range of substitution models implemented in BEAST2, including models with gamma rate variation across sites (Yang, 1994).

2 New Approaches

In this section, we specify our approach to sample from the posterior distribution of species networks and gene trees, given a multilocus sequence alignment. First we derive the unnormalized posterior distribution. Then we introduce operators to move through the space of species networks, the space of gene trees, and finally to update the gene tree embeddings within species networks.

2.1 The posterior distribution of species networks and gene trees

2.1.1 The probability density of a species network

The birth-hybridization process provides a prior probability for a given species network Ψ (Fig. 1). The process starts from t_0 (time of origin) in the past with a single species. A species gives birth to a new species with a constant rate λ (speciation rate), and two species merge into one with a constant rate ν (hybridization rate). That is, at the moment of k species, the speciation rate is $k\lambda$, the hybridization rate is $\binom{k}{2}\nu$, and the waiting time to the next event is an exponential distribution. The process ends at time

0 (the present). For the network shown in Figure 1, the probability density of the species network Ψ given λ, ν, and t_0 is

$$\begin{aligned} f(\Psi \mid \lambda, \nu, t_0) = & \lambda e^{-\lambda(t_0 - t_1)} \\ & \lambda e^{-(2\lambda+\nu)(t_1 - t_2)} \\ & \lambda e^{-(3\lambda+3\nu)(t_2 - t_3)} \\ & \nu e^{-(4\lambda+6\nu)(t_3 - t_4)} \\ & e^{-(3\lambda+3\nu)t_4}. \end{aligned}$$

In general, the probability density of a species network with n species descending from $n-1+m$ speciation events and m hybridization events, and these events happening at time $t_1 > t_2 > \ldots > t_{n+2m-1}$, is,

$$f(\Psi \mid \lambda, \nu, t_0) = \lambda^{n+m-1} \nu^m \prod_{i=0}^{n+2m-1} e^{-\left(\lambda k_i + \nu \binom{k_i}{2}\right)(t_i - t_{i+1})}, \qquad (1)$$

where k_i is the number of lineages within time interval (t_i, t_{i+1}) and $t_{n+2m} = 0$ is the present time. In our Bayesian analysis, the parameters λ, ν, and t_0 can be assigned hyperpriors.

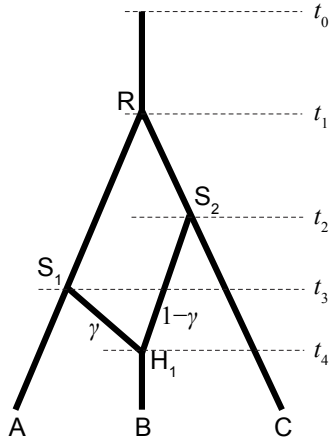

Figure 1: A species network with 3 tips, 3 bifurcations, and 1 reticulation. The inheritance probability at branch S_1H_1 is γ, and that at S_2H_1 is $1-\gamma$. In the simulations, $t_i = \tau_i$ ($\mu = 1.0$), $\tau_1 = 0.05, \tau_2 = 0.03, \tau_3 = 0.02, \tau_4 = 0.01$, and $\gamma = 0.3$. The population sizes are all $\theta = 0.01$.

2.1.2 The probability of the sequence data given the gene trees

Assuming complete linkage within each locus and free recombination among loci, the probability of the data $D = \{D_1, D_2, \ldots, D_{\text{L}}\}$ given gene trees $G = \{G_1, G_2, \ldots, G_{\text{L}}\}$ is the product of phylogenetic likelihoods (Felsenstein, 1981) at individual loci:

$$\Pr(D \mid G, \boldsymbol{\mu}, \boldsymbol{\varphi}) = \prod_{i=1}^{\text{L}} \Pr(D_i \mid G_i, \mu_i, \varphi_i), \tag{2}$$

where G_i is the gene tree with coalescent times, μ_i is the substitution rate per site per time unit, and φ_i represents the parameters in the substitution model (e.g., the transition-transversion rate ratio κ in the HKY85 model (Hasegawa et al., 1985)), at locus i ($i = 1, \ldots, \text{L}$).

There are two sources of evolutionary rate variation: across gene tree lineages at the same locus and across different gene loci. In the strict molecular clock model (Zuckerkandl and Pauling, 1965), μ is the global clock rate, i.e., no rate variation across gene lineages at each locus. To extend to a relaxed molecular clock model (e.g., Thorne and Kishino, 2002; Drummond et al., 2006; Lepage et al., 2007; Rannala and Yang, 2007), the molecular clock rate is variable across gene lineages following certain distributions with μ as the mean. To account for rate variation across genes, gene-rate multipliers $\{m_1, m_2, \ldots, m_{\text{L}}\}$ are constrained to average to 1.0 ($\sum_{i=1}^{\text{L}} m_i x_i = 1$, where x_i is the proportion of sites in locus i to the total number of sites). Then the substitution rate at locus i is $\mu_i = \mu m_i$. Thus, when multiplying the gene tree lineages in G_i by μ_i, all the branch lengths are then measured by genetic distance (substitutions per site).

The gene-rate multipliers are assigned a flat Dirichlet prior. The average substitution rate (clock rate) μ can be either fixed to 1.0 such that branch lengths are measured by genetic distance, or assigned an informative prior to infer branch lengths measured in absolute time.

2.1.3 The probability density of the gene trees given a species network

The gene trees $G = \{G_1, G_2, \ldots, G_{\text{L}}\}$ are embedded in the species network Ψ under the multispecies network coalescent (MSNC) model (Yu et al., 2014) (Fig. 2). Hybridizations or horizontal gene transfers are modelled by reticulations in the species network. The effective population sizes $N = \{N_1, N_2, \ldots, N_{\text{B}}\}$ are assumed to be identically and independently distributed

(i.i.d.) for each of the B branches in Ψ, while each locus has the same effective population size N_i at branch i ($i = 1, \ldots, \text{B}$). For each locus j, the number of coalescences of gene tree G_j within branch b of Ψ is denoted by k_{jb}, and the number of lineages at the tipward end of b is denoted by n_{jb}, thus the number of lineages at the rootward end of b is $n_{jb} - k_{jb}$. The $k_{jb}+1$ coalescent time intervals between the tipward and rootward of branch b are denoted by c_{jbi} ($0 \leq i \leq k_{jb}$). p_j is the gene ploidy of locus j (e.g., 2 for autosomal nuclear genes and 0.5 for mitochondrial genes in diploid species). $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_{\text{H}}\}$ are the inheritance probabilities, one per hybridization node in Ψ. For each lineage of G_j traversing the hybridization node H_h backward in time, with probability γ_h it goes to the parent branch associated with that inheritance probability, and to the alternate parent branch with probability $1 - \gamma_h$. The corresponding number of traversing lineages are denoted by u_{jh} and v_{jh} respectively.

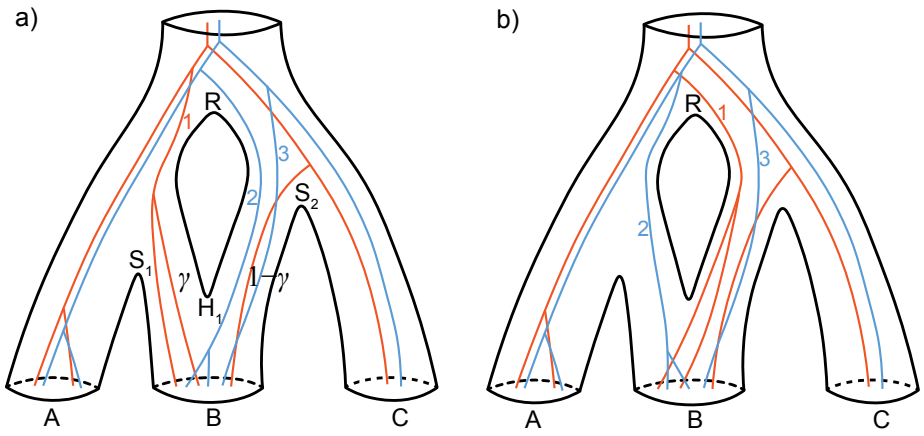

Figure 2: Two gene trees embedded in the species network. There are 2 samples from species A, 3 samples from B, and 1 sample from C. But note that we do not restrict the number of samples to be the same for each gene in each species. For each gene tree lineage traversing the hybridization node H_1 backward in time, it goes to the left population with probability γ, and to the right with probability $1 - \gamma$. a) and b) show two possibilities of gene-tree embeddings. The three lineages that can traverse either side through root node R are labeled as 1, 2, and 3.

The coalescent probability of the gene trees G in species network Ψ with

time being measured in calendar units is thus:

$$
\begin{aligned}
&f(G \mid \Psi, \boldsymbol{\gamma}, N) \\
&= \prod_{j=1}^{\mathrm{L}} \left[\prod_{b=1}^{\mathrm{B}} (p_j N_b)^{-k_{jb}} \exp\left(-(p_j N_b)^{-1} \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb}-i}{2} \right) \prod_{h=1}^{\mathrm{H}} \gamma_h^{u_{jh}} (1-\gamma_h)^{v_{jh}} \right] \\
&= \Gamma \prod_{b=1}^{\mathrm{B}} r_b N_b^{-q_b} \exp(-\sigma_b N_b^{-1}), \qquad (3)
\end{aligned}
$$

where $q_b = \sum_j k_{jb}$, $r_b = \prod_j p_j^{-k_{jb}}$, $\sigma_b = \sum_j p_j^{-1} \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb}-i}{2}$, and $\Gamma = \prod_j \prod_h \gamma_h^{u_{jh}} (1-\gamma_h)^{v_{jh}}$. When there is no reticulation in the species network (i.e., species tree), then $\Gamma = 1$ and Equation 3 is equivalent to Equation 2 in Jones (2017).

Note here, when time is measured by genetic distance, we use $\theta_b = N_b \mu$ as the population size parameter of branch b, and $\tau_i = t_i \mu$ as the height of node i. The prior for $\boldsymbol{\gamma}$ can be any distribution on $[0, 1]$, we use throughout $f(\gamma_h) \sim \text{beta}(1, 1)$. In the next section, we discuss how to integrate out the population sizes, which will improve computational speed.

2.1.4 Integrating out the population sizes analytically

Equation 3 has the form of unnormalized inverse gamma densities. The population sizes N can be integrated out through the use of i.i.d. inverse-gamma(α, β) conjugate prior distributions (Jones, 2017; Hey and Nielsen, 2007), that is,

$$
\begin{aligned}
f(G \mid \Psi, \boldsymbol{\gamma}) &= \int f(G \mid \Psi, \boldsymbol{\gamma}, N) f(N \mid \alpha, \beta) dN \\
&= \Gamma \prod_{b=1}^{\mathrm{B}} \int_0^{\infty} r_b N_b^{-q_b} \exp(-\sigma_b N_b^{-1}) \frac{\beta^{\alpha}}{\Gamma(\alpha)} N_b^{-\alpha-1} \exp(-\beta N_b^{-1}) dN_b \\
&= \Gamma \prod_{b=1}^{\mathrm{B}} \frac{r_b \beta^{\alpha}}{(\beta + \sigma_b)^{\alpha+q_b}} \frac{\Gamma(\alpha + q_b)}{\Gamma(\alpha)}. \qquad (4)
\end{aligned}
$$

The symbolic notations follow Equation 3.

2.1.5 The joint posterior distribution

The joint posterior distribution of the parameters is

$$f(\Psi, G, \Theta \mid D) \propto \Pr(D \mid G, \boldsymbol{\mu}, \boldsymbol{\varphi}) f(G \mid \Psi, \boldsymbol{\gamma}) f(\Psi \mid \lambda, \nu, t_0) \\ f(\boldsymbol{\mu}) f(\boldsymbol{\varphi}) f(\boldsymbol{\gamma}) f(\lambda, \nu) f(t_0). \quad (5)$$

Here Θ represents $(\boldsymbol{\mu}, \boldsymbol{\varphi}, \boldsymbol{\gamma}, \lambda, \nu, t_0)$.

2.2 MCMC operators for the species network

2.2.1 Node slider

The node-slider operator only changes the node heights of the species network, not the topology. It selects an internal node or the origin randomly, then proposes a new height centered at the current height according to a normal distribution: $t' \mid t \sim N(t, \sigma^2)$, where σ is a tuning parameter controlling the step size. The lower bound is the oldest child-node height, the upper bound is the youngest parent-node height (except for the origin, Fig. 3). If the proposed value is outside this range, the excess is reflected back into the interval. Note that for the origin, if the proposed height is outside the range of its prior, this move is aborted. A variation of this operator can use a uniform proposal instead of the normal proposal: $t' \mid t \sim U(t - w/2, t + w/2)$, where w is the window size. The proposal ratio is 1.0 in both cases.

2.2.2 Node uniform

The node-uniform operator also changes the internal-node heights of the species network while keeping the topology. It selects an internal node randomly, then proposes a new height uniformly between the lower and upper bounds (Fig. 3ab). The lower bound is the oldest child-node height, the upper bound is the youngest parent-node height. The proposal ratio is 1.0. Unlike node slider, this operator does not change the time of origin. A separate operator for the origin, such as multiplier or scaler, can be coupled to update all the node heights.

2.2.3 Branch relocator

The branch-relocator operator can change the topology, but keeps the number of reticulations in the species network constant. It first selects an internal node at random. If the selected node is a bifurcation node, the rootward end of either its child branches is relocated (Fig. 4a); if the selected node is a

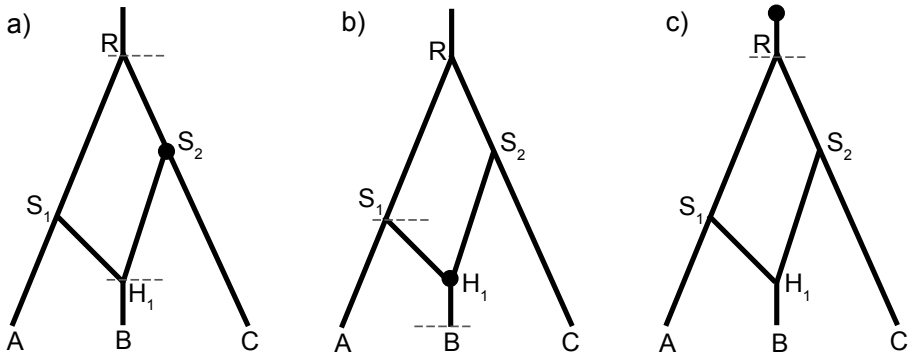

Figure 3: Three cases when the node-slider operator is applied: a) a bifurcation node S_2 is selected; b) the reticulation node H_1 is selected; c) the origin is selected. The dashed lines are the lower and upper bounds for changing its height (only the lower bound is applicable in c)). For the node-uniform operator, a) and b) apply but c) does not.

reticulation node, the tipward end of either its parent branches is relocated (Fig. 4b). Then the selected branch is detached at the side of the selected node, and a destination branch chosen randomly from all possible branches to attach to is proposed (excluding the original position).

There are two variants of this operator. The narrow move doesn't change any node heights (Fig. 4). Since there are equal numbers of possible attachment points for the forward and backward moves, the proposal ratio is 1.0. The wide move proposes a new height of the selected node uniformly between a lower (v') and an upper (u') bound. If the selected node is a bifurcation node, the lower bound v' is the maximum of the tipward ends of the selected branch and the destination branch, and the upper bound u' is the height of the rootward end of the destination branch (Fig. 4a). If the selected node is a reticulation node, the lower bound v' is the height of the tipward end of the destination branch, and the upper bound u' is the minimum of the rootward ends of the selected branch and the destination branch (Fig. 4b). We denote with v and u the lower and upper bounds of the backward move. The proposal ratio is $(u' - v')/(u - v)$.

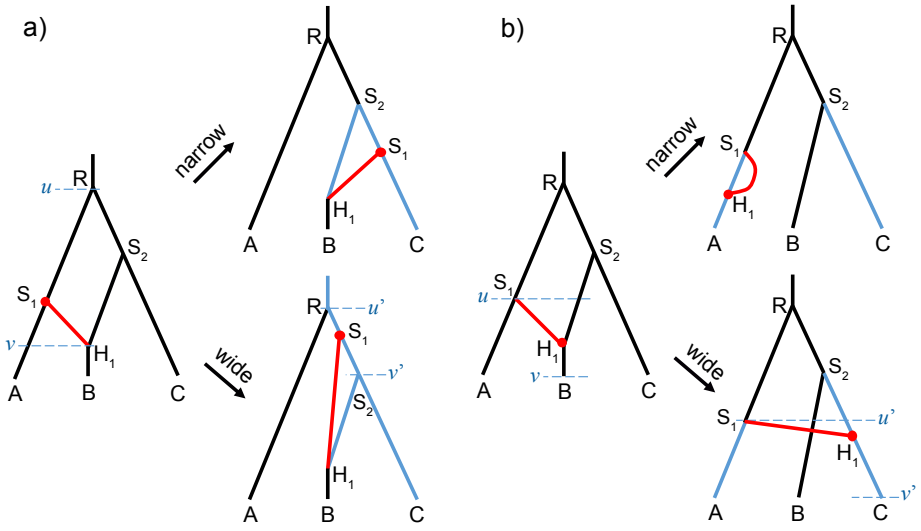

Figure 4: Two cases when the branch-relocator operator is applied: a) a bifurcation node S_1 is selected; b) a reticulation node H_1 is selected. Branch S_1H_1 is relocated (the selected node and branch are in red, the candidate destination branches are in blue). In the narrow move, the height of S_1 or H_1 is kept unchanged. In the wide move, the height of S_1 or H_1 is proposed uniformly between v' and u'. The lower and upper bounds of the backward move are v and u.

2.2.4 Add- and delete-reticulation

The add-reticulation and delete-reticulation operators are reversible-jump MCMC (rjMCMC) proposals that can add and delete a reticulation event respectively.

In the add-reticulation operator, a new branch is added by connecting two randomly selected branches with length l_1 and l_2 (Fig. 5). The same branch can be selected twice so that $l_1 = l_2$ (Fig. 5b). Then three values ω_1, ω_2 and ω_3 are drawn from $U(0, 1)$. One attaching point cuts the branch length l_1 to $l_{11} = l_1\omega_1$ (and thus $l_{12} = l_1(1 - \omega_1)$); the other attaching point cuts the branch length l_2 to $l_{21} = l_2\omega_2$ (and thus $l_{22} = l_2(1 - \omega_2)$). Analogously, if we select the same branch twice, the attachment times of the new branch are $l_1\omega_1$ and $l_1\omega_2$. An inheritance probability $\gamma = \omega_3$ is associated to the new branch. We will operate on the inheritance probability γ of this added branch, while the inheritance probability of the second reticulation branch (i.e., $1 - \gamma$) changes accordingly. We denote k as

the number of branches in the current network, and m as the number of reticulation branches in the proposed network. The Hastings ratio is then $(1/m)/[(1/k)(1/k) \times 1 \times 1 \times 1] = k^2/m$. The Jacobian is $|\frac{\partial(l_{11},l_{21},\gamma)}{\partial(\omega_1,\omega_2,\omega_3)}| = l_1 l_2$. Thus the proposal ratio of add-reticulation is $l_1 l_2 k^2/m$.

In the delete-reticulation operator, a random reticulation branch together with the inheritance probability γ is deleted (Fig. 5). Joining the singleton branches at each end of the deleted branch, resulting in two branches with length l_1 and l_2 completes the operator ($l_1 = l_2$ when forming a single branch, Fig. 5b). If there is no reticulation, or the selected branch is connecting two reticulation nodes, the move is aborted. For example in Figure 5a, deleting reticulation branch H_1H_2 will result in an invalid network. We denote k as the number of branches in the proposed network, and m as the number of reticulation branches in the current network. The proposal ratio of delete-reticulation is $m/(k^2 l_1 l_2)$.

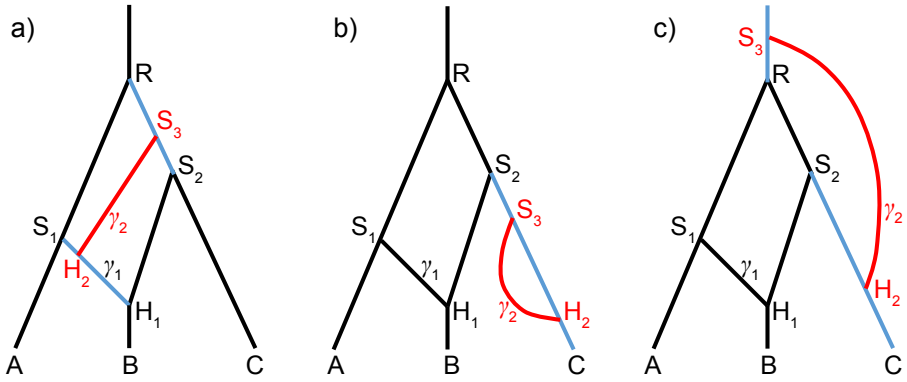

Figure 5: Three cases when the add-reticulation operator is applied. The number of branches in the current network (i.e., the network without the red branch) is $k = 8$. The probability of selecting the illustrated branches (in blue) is $1/k^2$. The number of reticulation branches in the proposed network is $m = 4$. In the reverse move, delete-reticulation, the probability of selecting the added branch (in red) is $1/m$. a) Branches S_1H_1 and RS_2 are selected and a new branch S_3H_2 is added together with γ_2. The length of S_1H_1 is $l_1 = l_{S_1H_1}$, and that of RS_2 is $l_2 = l_{RS_2}$. In the delete-reticulation move, if H_1H_2 is selected, the operator is aborted. b) The same branch S_2C is selected twice. $l_1 = l_2 = l_{S_2C}, l_{11} = l_{S_2S_3}, l_{21} = l_{S_2H_2}$. c) The root branch and S_2C are selected. S_3 becomes the new root.

2.2.5 Gamma uniform

The gamma-uniform operator selects a reticulation node randomly, and propose a new value of the inheritance probability $\gamma' \sim U(0,1)$. The proposal ratio is 1.0.

2.2.6 Gamma random-walk

The gamma random-walk operator selects a reticulation node randomly, and applies a uniform sliding window to the logit of the inheritance probability γ, that is $y' \mid y \sim U(y-w/2, y+w/2)$, where $y = \text{logit}(\gamma) = \log(\gamma) - \log(1-\gamma)$. Since the proposal ratio for the transformed variable y is 1.0, and $\frac{d\gamma}{dy} = \frac{d}{dy}\left[e^y/(1+e^y)\right] = e^y/(1+e^y)^2$, the proposal ratio for the original variable γ is $\frac{d\gamma'}{dy'}/\frac{d\gamma}{dy} = e^{(y'-y)}(1+e^y)^2/(1+e^{y'})^2$.

2.3 MCMC operators for gene trees

The standard tree operators in BEAST 2 (Bouckaert et al., 2014) are applied to update the gene trees, including the scale, uniform, subtree-slide, narrow- and wide-exchange, and Wilson-Balding (Wilson and Balding, 1998). The scale and uniform operators only update the node heights without changing the tree topology, while the other operators can change the topology (Drummond and Bouckaert, 2015). The species network is kept unchanged when operating on the gene trees, and vice versa.

2.4 MCMC operator for the gene tree embedding

The gene trees must be compatibly embedded in the species network (Fig. 2). When a new gene tree is proposed using one of the tree operators, the rebuild-embedding operator proposes a new embedding for that gene tree. When a new species network is proposed, the rebuild-embedding operator proposes a new embedding for each gene tree in the species network. If there is no valid embedding for any gene tree, the gene-tree or species-network operator is aborted.

The rebuild-embedding operator goes through all the gene tree lineages from the root to the tips recursively. For each lineage,

1. if it can traverse either side through a network node forward in time, increase the counter n' by one, then embed the lineage in either side of the descendent species with equal probability 0.5;

2. if it can traverse through a network node in only one way, or stays within the same species, embed it that way;

3. otherwise, abort the move, as there is no valid embedding.

The proposal ratio is $2^{n'}/2^{n} = 2^{n'-n}$, where n is the traversing counter of the backward move.

For example, if the current embedding is Figure 2a and the proposed embedding is Figure 2b, $n' = n = 3$ so that the proposal ratio is 1.0, because lineages 1, 2 and 3 can traverse either side through root node R, and the other lineages either traverse in only one way or stay within the same species. Note though that in contrast to Figure 2, typically the gene trees or the species networks are different for the forward and backward moves, such that n' can be different from n.

3 Results

The components from the last section, i.e., the unnormalized posterior density and the operators, allow us to implement a Markov chain Monte Carlo (MCMC) procedure to sample species networks and gene trees from the posterior distribution, given a multilocus sequence alignment. The implementation is available within BEAST 2 as an add-on `SpeciesNetwork`. A convenient format for the species networks, and a link to our source code, is presented in Materials and Methods. We investigate the performance of the implementation in this section, first based on simulations and then based on empirical data.

3.1 Simulations

We performed several simulation studies to verify the implementation of our Bayesian MCMC method.

We first compared networks simulated forward-in-time under the birth-hybridization process with those sampled under MCMC using the network operators (see Material and Methods for details of simulations and MCMC sampling). Theoretically, we expect both forward-in-time simulated networks and the MCMC sampled networks to be identical. Indeed, the networks obtained from the simulator and the MCMC match when comparing the network length, root height, number of reticulations, and time of the youngest reticulation (Fig. 6).

We then tested the rebuild-embedding operator by traversing a bipartite gene tree in the species network with 1 tip and 4 reticulations (Fig. 7a,

details in Materials and Methods). The result shown in Figure 7b confirms equal probability of the 36 embeddings.

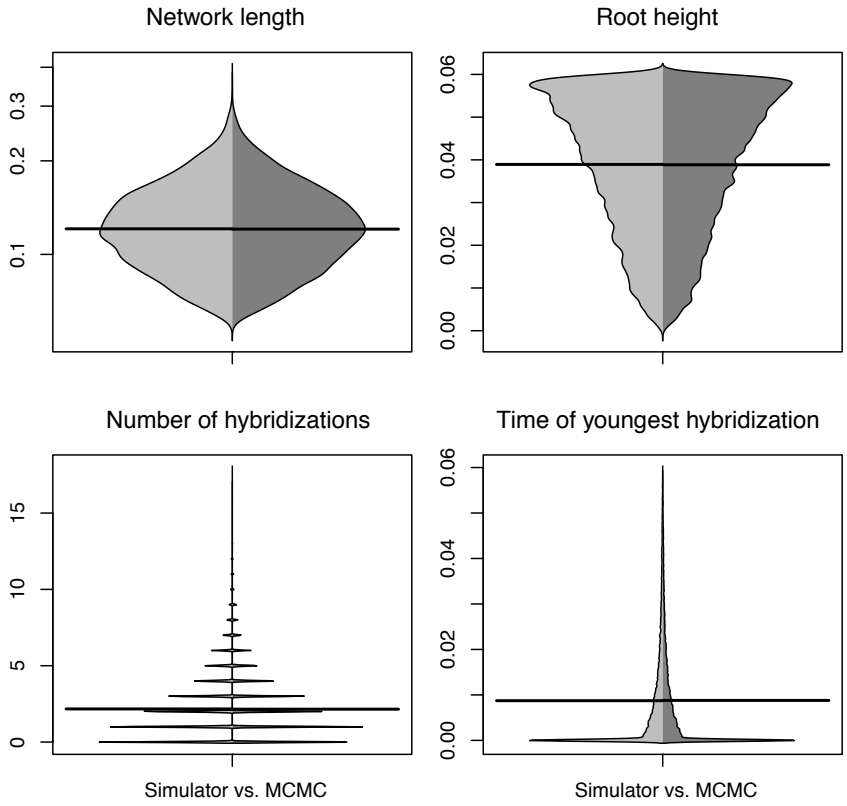

Figure 6: Beanplot of network summary statistics comparing 3-tips networks simulated under the birth-hybridization process (left, light gray) with those sampled using the network operators (right, dark gray). The horizontal bar is the mean.

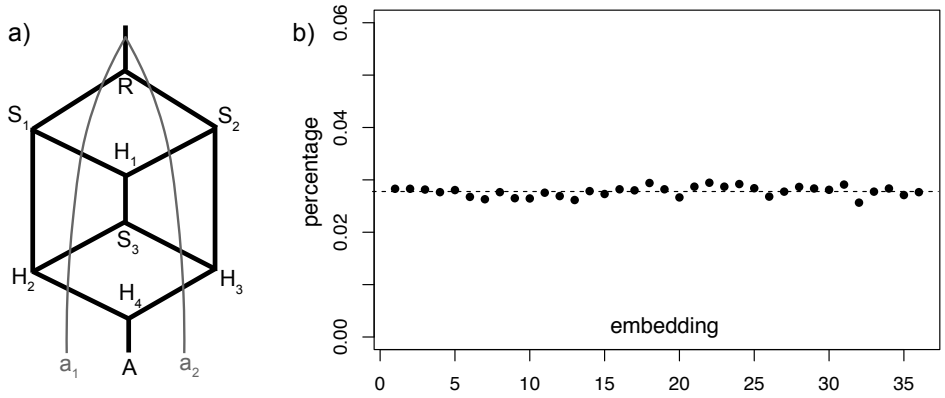

Figure 7: a) The embeddings of gene tree (a_1, a_2) within the species network with 1 tip and 4 reticulations are sampled using MCMC. For each lineage, there are 6 ways traversing the network from root R to tip A, resulting in $6 \times 6 = 36$ possible embeddings in total. b) The percentage of 36 embeddings sampled using the rebuild-embedding operator. The dashed line is the true value 1/36.

We also compared gene trees simulated under MSNC backward-in-time with those sampled under MCMC using the gene-tree operators, given the species network in Figure 8. When the population sizes were fixed to the truth (0.01), the tree sets from MSNC and MCMC give rise to the same distribution of tree length, gamma-statistic (Pybus and Harvey, 2000), and Colless' index (Blum et al., 2006) as expected (Fig. 9). When the population sizes were integrated out analytically using inverse-gamma(10, 0.1) in the MCMC (Eq. 4), there is a slight mismatch between the two tree sets (Fig. 10), as the integration averages over all the possible values of population sizes under MCMC, while the population sizes were fixed as 0.01 in the simulation. But the inverse-gamma prior used is very informative, centered around the true value (0.01), thus the difference is minor.

The analyses up to here show that we can correctly sample species networks, gene tree embeddings for a given network, and gene trees for a given network. Next, we simulated sequence alignments of multiple loci to reveal the ability of our method to recover the true species network from multilocus sequence data. When the species network topology was fixed to Figure 1 in the inference, the results are shown in Figure 11. With small sample size (2, 4, 2) (meaning species A has 2, B has 4, and C has 2 sampled sequences) and only 5 loci, the posterior estimates are sensitive to the priors. When

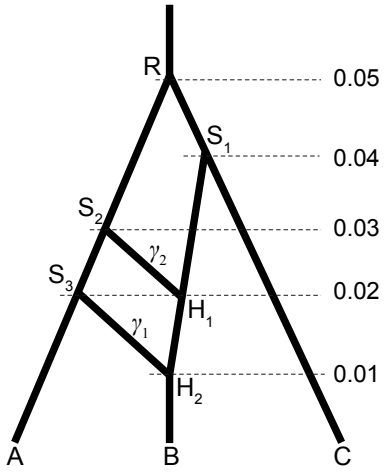

Figure 8: The species network under which to simulate gene trees without data. The inheritance probabilities are $\gamma_1 = 0.3$ and $\gamma_2 = 0.7$. The population sizes are all $\theta = 0.01$.

sample size increases, the posterior estimates become increasingly accurate. When we also inferred species network topology, with all network operators enabled, the results are shown in Figure 12.

From Figure 12 (and also Fig. 11), we observe that adding more loci increases the accuracy of inference much more than adding more individuals. For example, by comparing (5, 10, 5) 5 loci with (2, 4, 2) 10 loci, the latter has much higher posterior probability recovering the true species network (Fig. 12a). Conditional on the true species network topology (i.e., Fig. 1), the estimate of inheritance probability γ becomes increasingly accurate as the number of loci increases (Fig. 12b).

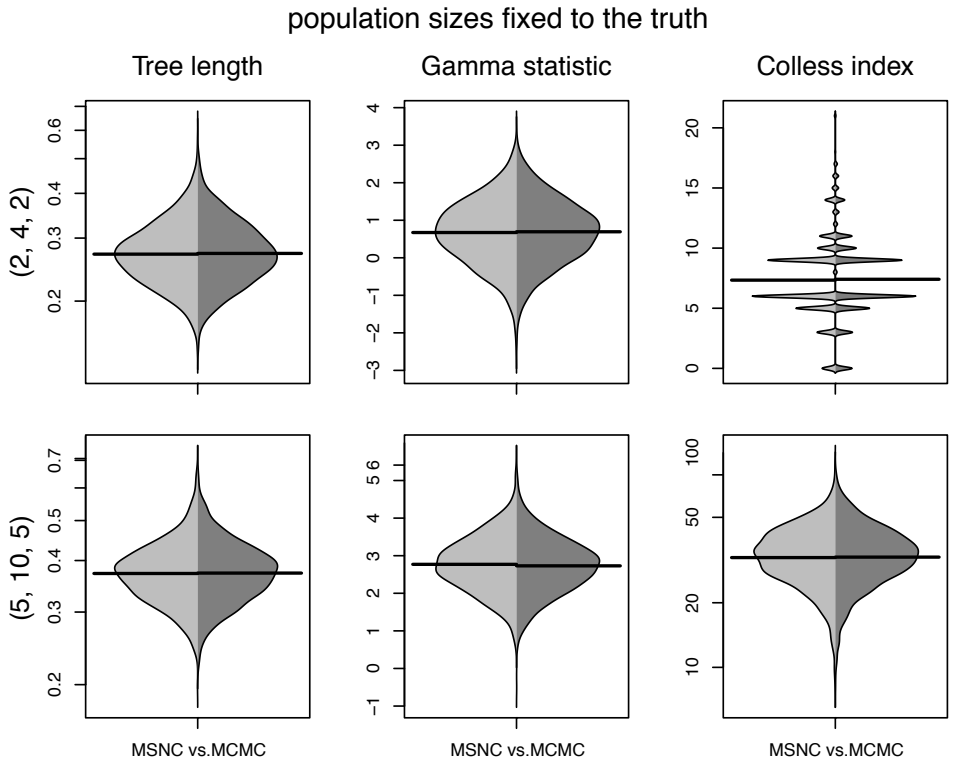

Figure 9: Beanplot of three tree summary statistics comparing gene trees simulated under MSNC (left, light gray) with those sampled using the gene-tree operators (right, dark gray), given the species network in Figure 8. The sample configuration were (2, 4, 2) (2 samples from A, 4 from B, and 2 from C) or (5, 10, 5) (5 samples from A, 10 from B, and 5 from C). The population sizes were fixed to the truth (0.01) in the MCMC.

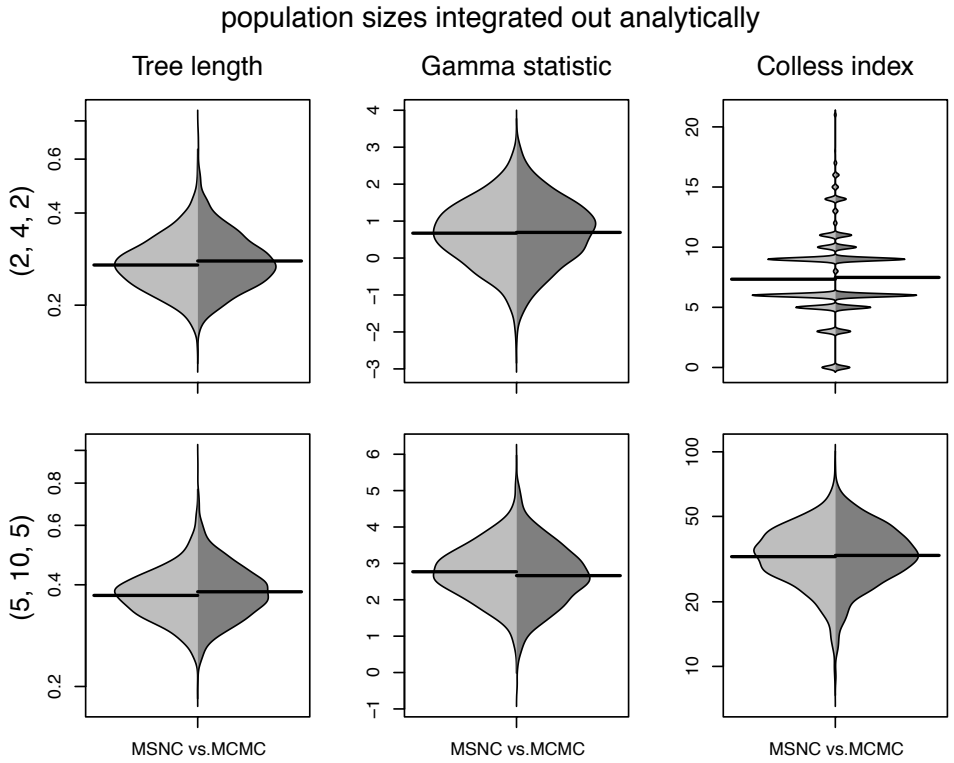

Figure 10: Beanplot of three tree summary statistics comparing gene trees simulated under MSNC (left, light gray) with those sampled using the gene-tree operators (right, dark gray), given the species network in Figure 8. The sample configuration were (2, 4, 2) or (5, 10, 5). The population sizes were integrated out analytically using inverse-gamma(10, 0.1) in the MCMC.

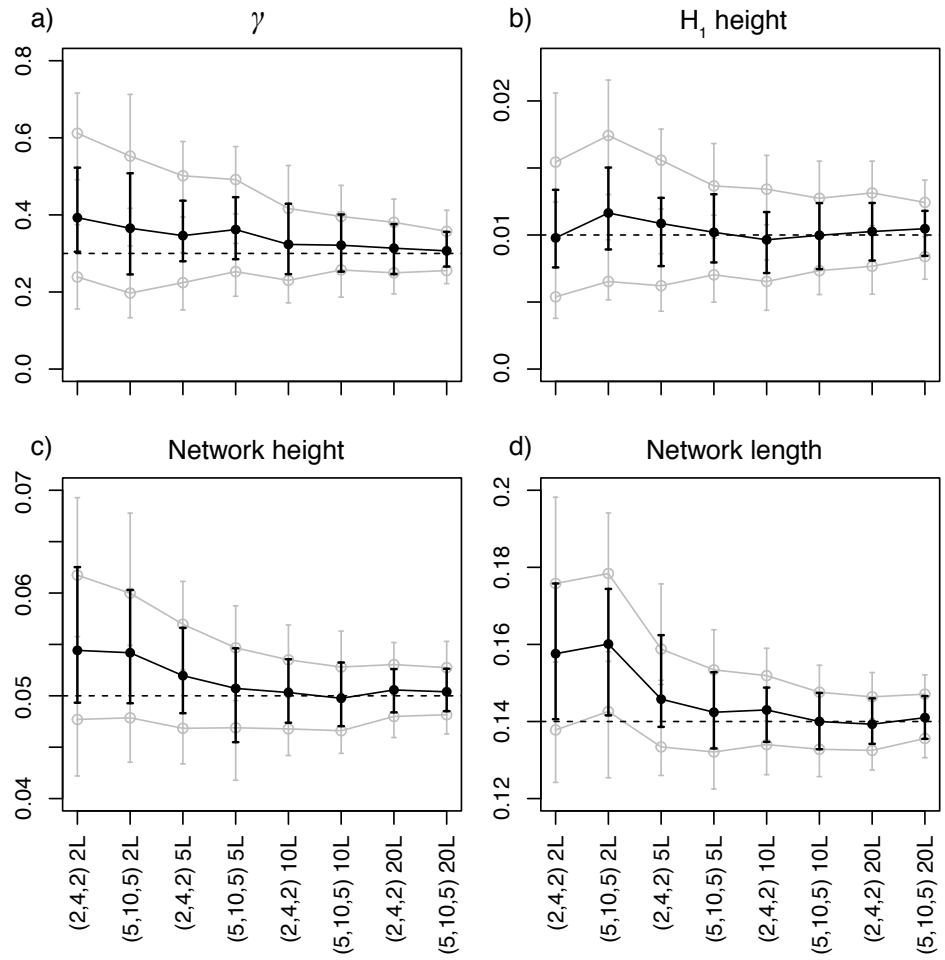

Figure 11: Posterior estimates of γ, H_1 height, network height and network length, when the data were simulated under the network in Figure 1 with sample configuration (2, 4, 2) (species A has 2, B has 4, and C has 2 sampled sequences) or (5, 10, 5), and 2, 5, 10 or 20 loci, respectively. The species network topology was fixed to the truth when performing inference. For each setting, the black dot with error bars are the median and the 1^{st} and 3^{rd} quantiles of the posterior medians of 100 replicates, the gray circle with error bars are the same for the 1^{st} and 3^{rd} quantiles of the posterior samples. The dashed lines indicate the true values.

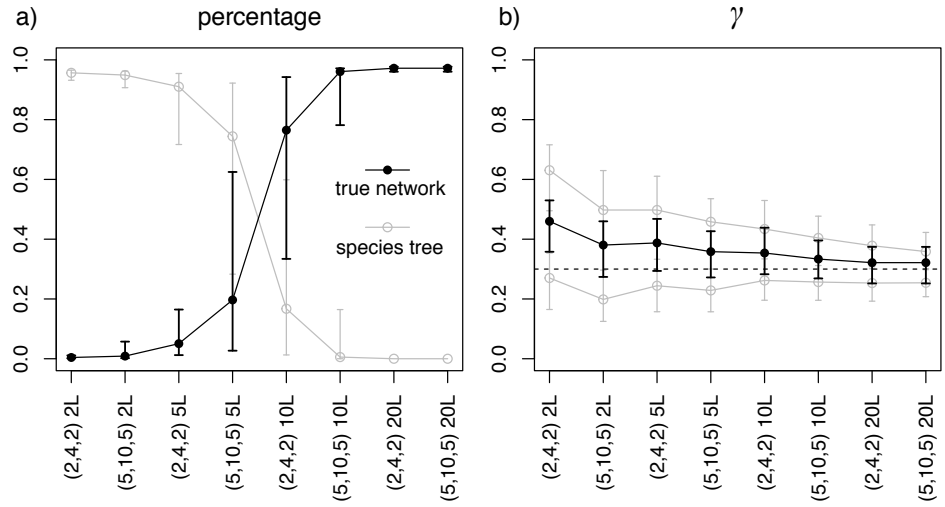

Figure 12: Simulation settings are the same as the ones described in Figure 11. Now, the species network topology was additionally inferred using MCMC. a) Posterior probabilities of the true network (black) and of species tree (gray). For each setting, the dot/circle with error bars are the median and the 1^{st} and 3^{rd} quantiles of the percentages of 100 replicates. b) Posterior estimates of γ when the species network is true. The dot/circle with error bars have the same meaning as in Figure 11a. The dashed line indicates the true value of γ (0.3).

3.2 Analysis of biological data

We analyzed a dataset of three spruce species (*Picea purpurea*, *P. likiangensis* and *P. wilsonii*) in the Qinghai-Tibet Plateau. *P. purpurea* was inferred to be a homoploid hybrid of *P. likiangensis* and *P. wilsonii* (Sun et al., 2014). The data consists of 11 loci per individual, and ≥ 50 individuals per species. To achieve proper mixing and convergence in a reasonable time, the data was truncated (w.r.t. the number of individuals) into two non-overlapping datasets, each having 10 individuals per species. Each individual has two phased haplotype sequences per locus.

Analyses were performed as described in the Materials and Methods section. The species network shown in Figure 13 has the highest posterior probability for both datasets, which are 0.65 and 0.62 respectively. The parameter estimates are similar for both datasets, so that we combined the MCMC samples to summarize the posteriors. The 95% credible set contains networks of 1, 2, and 3 reticulations, with probability 0.64, 0.28, 0.07, respectively. The species networks with one reticulation all have the same topology as Figure 13. However, the estimate of γ is 0.49 (95% CI = [0.23, 0.79]) for the networks with one reticulation, meaning the prior mean and posterior mean are not very different. It appears that these datasets do not have enough information to infer γ precisely.

To improve the inference of the parameters, we added back more individuals to the previously truncated data while fixing the network topology to the one shown in Figure 13. Each of the two non-overlapping datasets now has 20 individuals from *P. purpurea*, 15 from *P. likiangensis*, and 15 from *P. wilsonii* (100 sequences per locus). We were not able to obtain good mixing if co-estimating the network topology for this size of data. The MCMC samples from the two datasets were combined. The posterior estimates of node heights, γ, population sizes, and gene-rate multipliers (means and 95% HPD intervals) are shown in Table 1. The parameters estimates are similar, regardless of whether the population sizes are inferred or integrated out.

The results above confirm that *P. purpurea* is a hybrid species of *P. likiangensis* and *P. wilsonii*. About 35% of the nuclear genome of *P. purpurea* was derived from *P. wilsonii* (and 65% from *P. likiangensis*). This estimate is close to the original estimate of 31% made using approximate Bayesian computation (ABC) (Sun et al., 2014). Assuming an average substitution rate $\mu = 2 \times 10^{-4}$ per site per million years (Sun et al., 2014), and dividing the node heights (τ's in Table 1) by μ, we get the times measured by million years, as shown in Figure 13. The time of hybridization is inferred to be around 1 Ma. The estimate was 1.3 (95% CI = [0.73, 2.2]) Ma in the

original analysis assuming the same height for nodes D, E, and H. Moreover, we get an older and narrower estimate for the root age of 6.0 (95% CI = [5.1, 7.5]) Ma, compared to 2.7 (95% CI= [1.4, 6.5]) Ma in the original analysis. Similarly, dividing estimates of θ (Table 1) by $\mu = 1 \times 10^{-8}$ per site per generation, we get the mean effective population sizes of *P. purpurea*, *P. wilsonii*, and *P. likiangensis* as 2.1×10^5, 0.5×10^5, and 1.4×10^5, respectively, which are smaller than estimated using ABC (cf. Table 4 in Sun et al., 2014).

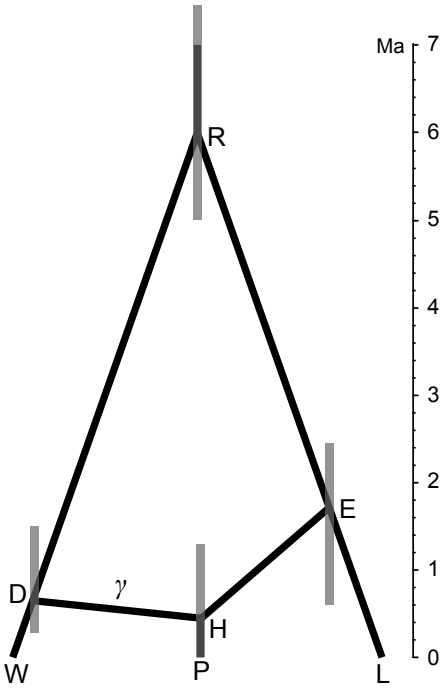

Figure 13: The species network with highest posterior probability (0.64) inferred from the spruce data. The node heights (means and 95% HPD intervals) are in unit of million years. The inheritance probability γ is about 0.35. See also Table 1.

Table 1: Posterior estimates

	posterior (mean & 95%HPD)	
parameters	population sizes integrated	population sizes inferred
τ_R	0.00119 (0.00103, 0.00149)	0.00121 (0.00101, 0.00152)
τ_D	0.000139 (0.000057, 0.000318)	0.000126 (0.000054, 0.000287)
τ_E	0.000338 (0.000119, 0.000425)	0.000349 (0.000128, 0.000545)
τ_H	0.000092 (0.000000, 0.000268)	0.000086 (0.000000, 0.000251)
γ	0.352 (0.127, 0.568)	0.346 (0.121, 0.542)
θ_W	–	0.000495 (0.000187, 0.000879)
θ_P	–	0.00212 (0.00145, 0.00429)
θ_L	–	0.00135 (0.000778, 0.00195)
θ_{H_D}	–	0.00346 (0.00181, 0.00527)
θ_{H_E}	–	0.00151 (0.000254, 0.00309)
θ_D	–	0.00128 (0.000774, 0.00182)
θ_E	–	0.00328 (0.00149, 0.00523)
θ_R	–	0.00264 (0.000923, 0.00439)
m_{4cl}	1.513 (1.110, 1.923)	1.516 (1.122, 1.938)
m_{ebs}	0.380 (0.152, 0.626)	0.381 (0.160, 0.636)
m_{gi}	0.443 (0.221, 0.689)	0.445 (0.223, 0.688)
m_{moo2}	1.603 (0.986, 2.240)	1.592 (0.995, 2.209)
m_{m007d1}	1.069 (0.682, 1.467)	1.063 (0.685, 1.457)
m_{sb16}	1.393 (1.018, 1.789)	1.393 (1.022, 1.785)
m_{sb29}	0.929 (0.543, 1.357)	0.929 (0.544, 1.351)
m_{sb62}	0.994 (0.405, 1.655)	1.002 (0.416, 1.641)
m_{se1364}	0.431 (0.130, 0.783)	0.436 (0.133, 0.787)
m_{se1390}	1.379 (0.955, 1.817)	1.379 (0.953, 1.824)
m_{xy1420}	0.353 (0.075, 0.703)	0.351 (0.077, 0.700)

Posterior estimates (means and 95% HPD intervals) of node heights (τ's), inheritance probability (γ), population sizes (θ's), and gene-rate multipliers (m's for each of the 11 loci in the spruce data). The population sizes were either integrated out analytically using inverse-gamma(3, 0.003), or inferred under gamma(2, 2000) priors. Dividing τ by $\mu = 2 \times 10^{-4}$ per site per million years results in the time being measured in million years (Fig. 13).

4 Discussion

Methods to build a species network (e.g., Wu, 2010; Park et al., 2010; Albrecht et al., 2012) traditionally use inferred gene trees from each locus without accounting for their uncertainties, and employ nonparametric criterions such as parsimony. For population level data, the sequences are similar and the signal for gene tree topologies is typically low, using fixed gene trees is assigning too much certainty to the data. These methods typically assume that gene tree discordance is solely due to reticulation, thus may suffer in the presence of incomplete lineage sorting (Yu et al., 2011). The MSNC model (Yu et al., 2014) provides a statistical framework to account for both incomplete lineage sorting and reticulate evolution. But properly analyzing genetic data to infer species networks under the MSNC model is a challenging task. There have been methods using only the gene tree topologies from multiple loci under MSNC (Yu et al., 2012, 2014; Wen et al., 2016). However, gene trees with branch lengths are more informative for inferring species tree or network topology than gene tree topologies alone. Accounting for branch lengths can improve distinguishability of species networks (Pardi and Scornavacca, 2015; Zhu and Degnan, 2016). Although methods using estimated gene trees (with branch lengths) from bootstrapping or posterior sample as input take account gene tree uncertainty (Yu et al., 2014; Wen et al., 2016), directly using sequence data to co-estimate species networks and gene trees in a Bayesian framework showed improved accuracy (Wen and Nakhleh, 2016), where such uncertainties are averaged over using MCMC. Pseudo-likelihood approaches (Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2016) compute faster than full likelihood or Bayesian approaches, but have severe distinguishability issues and require more data to achieve good accuracy.

At the time of writing, another Bayesian method inferring species networks and gene trees simultaneously from multilocus sequence data has been published (Wen and Nakhleh, 2016). The general framework here is similar, but we highlight four major differences. We use a birth-hybridization prior for species network which naturally models the process of speciation and hybridization. The prior is extendable to account for extinction and incomplete sampling, and rates variation over time, as we outline below. Wen and Nakhleh (2016) used a descriptive prior combining Poisson distribution for the number of reticulations and exponential distributions for branch lengths. Secondly, we allow parallel branches (e.g., S_3H_2 in Fig. 5b) in the network. This is biologically possible. Even if the true species history has no parallel branches, the observed species network can still have

such features due to incomplete sampling. Note though that a very large number of individuals and loci are required to detect such parallel branches. To prevent the species network from growing arbitrarily big, such that it becomes indistinguishable by the gene trees (Pardi and Scornavacca, 2015; Zhu and Degnan, 2016), we typically assign informative prior on the birth rate to be larger than the hybridization rate. A similar strategy was used in Wen et al. (2016); Wen and Nakhleh (2016) by restricting the rate of the Poisson distribution. Third, we account for the uncertainty in the embedding of a gene tree within a species network by estimating the MSNC probability conditional on a proposed embedding at each MCMC step. This provides a posterior distribution of gene trees and their embeddings within a species network, enabling analysis of which alleles are derived from which ancestral species. Last but not least, we applied analytically integration for population sizes in the species network (Eq. 4). This reduces the number of parameters for the rjMCMC operators to deal with, and should improve convergence and mixing. Besides, our implementation in `SpeciesNetwork` is an extension to BEAST 2 (Bouckaert et al., 2014), to take advantage of many standard phylogenetic models, such as different substitution models, relaxed molecular clock models, and the BEAUTi graphical interface.

In our approach, we employ a simple prior for the species network based on a birth-hybridization model. Analogous to priors for species trees (e.g., Stadler, 2010; Heath et al., 2014), the prior for species network could be extended to account for speciation, extinction, hybridization, and incomplete sampling, each with a different rate, leading networks with present-day samples and potential past samples corresponding to fossils. The rates could also be allowed to vary over time, to model the diversification patterns during speciation (the skyline model for trees, Stadler et al., 2013). When considering networks instead of trees, techniques to derive the probability density of trees cannot be directly applied as the hybridization rate depends on pairs of lineages rather than individual lineages. This non-linearity necessitates solving differential equations to derive the species network probability densities, a task which we defer to a later study.

Our approach is limited in computational speed. The empirical analysis was done, e.g., on only 3 species with 10 individuals (20 sequences) each, and 11 loci. The main bottleneck is the MCMC operators. Due to hard constraints between the species network and embedded gene trees (Fig. 2), MCMC operators changing them separately limit the ability to analyze genomic scale data from many individuals. More specifically, updating the species network will likely violate a gene tree embedding, resulting in very low acceptance rate of the operator. Thus it will be essential to design more

efficient MCMC operators. There have been coordinated operators that can change species tree and gene trees simultaneously (Rannala and Yang, 2003, 2017; Jones, 2017). Such operators are possible to be extended to species networks, and will potentially improve efficiency of the MCMC algorithm. Moreover, under the assumption of independence among loci, parallelizing the calculation of phylogenetic likelihoods (Eq. 2) will further improve the speed.

Another important step for the Bayesian method is summarizing the posterior sample of species networks. For phylogenetic trees, the most common summaries include majority rule consensus tree and maximum clade credibility tree (Drummond and Bouckaert, 2015). But there is still a lack of good summaries for phylogenetic networks.

In summary, we developed a Bayesian method and computational tool for inferring species networks together with gene trees and evolutionary parameters from multilocus sequence data. The method provides a general Bayesian framework, with potential extensions in both theoretical and computational aspects.

5 Materials and Methods

5.1 Simulations

Time is measured by genetic distance (substitutions per site) throughout the simulations, so that $\theta = N\mu$ is used for all population sizes and $\tau_i = t_i\mu$ for the time of node i. The substitution rate μ is fixed to 1.0 across all gene lineages (strict molecular clock) and all loci (no rate variation).

5.1.1 Forward-in-time simulation and MCMC sampling of species networks

We first generated networks under the birth-hybridization process. The simulator starts from the time of origin (t_0) with one species. A species split into two (speciation) with rate λ, and two species merge into one (hybridization) with rate ν. At the moment of k branches, the total rate of change is $r_{tot} = k\lambda + \binom{k}{2}\nu$. We generate a waiting time $\sim \exp(r_{tot})$ and a random variable $u \sim U(0,1)$. If $u < k\lambda/r_{tot}$, we randomly select a branch to split; otherwise, we randomly select two branches to join. The simulator stops at time 0 (cf. Fig. 1). In this simulation, $\tau_0 = 0.06$, $\lambda = 30$, and $\nu = 20$, and we kept 20,000 networks with exactly 3 tips.

To compare with networks simulated above, we used our five network operators (the operators for γ are irrelevant in this case) to operate on networks with 3 tips. The gene trees were not considered (i.e., no effect from MSNC). τ_0, λ, and ν were fixed to the true values in the simulation. The MCMC chain was run for 30 million steps and sampled every 1000 steps. The last 20,000 sampled networks were kept (i.e., the burn-in was 33%).

5.1.2 Sampling the gene tree embeddings in a given species network

We sampled the embeddings of gene tree (a_1, a_2); in the species network shown in Figure 7a, using the rebuild-embedding operator alone. The gene tree and species network were both fixed. The priors, MSNC, and likelihood were set to be constant functions. The MCMC chain was run 2 million steps and sampled every 100 steps.

5.1.3 Sampling gene trees without data in a given species network

We compared gene trees simulated under the backward-in-time MSNC with those sampled using the gene-tree operators. 10,000 random gene trees were generated with the backward-in-time simulation given the network shown in Figure 8, with population sizes $\theta = 0.01$. The sample configurations were (2, 4, 2) (2 samples from A, 4 from B, and 2 from C) and (5, 10, 5) (5 samples from A, 10 from B, and 5 from C).

In the MCMC, the gene-tree operators include scale, uniform, subtree-slide, narrow- and wide-exchange, and Wilson-Balding (Drummond and Bouckaert, 2015). The species network topology, node heights and inheritance probabilities were fixed to the true values in the simulation (Fig. 8). The population sizes were either fixed to the truth (0.01), or integrated out analytically using inverse-gamma(10, 0.1) (Eq. 4). The probability of the sequence data was set to be constant (no data). The chain was run 15 million steps and sampled every 1000 steps. The last 10,000 sampled gene trees were kept (i.e., the burn-in was 33%).

5.1.4 Inference of species networks from sequences

We simulated sequence alignments of multiple loci under the true network shown in Figure 1. A random gene tree was generated for each locus under the MSNC. Then DNA sequences of length 200 bp were simulated under JC69 model (Jukes and Cantor, 1969) along each tree, with $\tau_1 = 0.05, \tau_2 =$

$0.03, \tau_3 = 0.02, \tau_4 = 0.01, \gamma = 0.3$, and population sizes $\theta = 0.01$. The sample configurations were (2, 4, 2) and (5, 10, 5), and the number of loci was 2, 5, 10, 20, respectively. Under each setting, the simulation was repeated 100 times.

We first fixed the network topology, to infer the node heights and inheritance probability from each simulated dataset. The priors were $\tau_0 \sim \exp(10)$, $d = \lambda - \nu \sim \exp(0.1)$, $r = \nu/\lambda \sim \text{beta}(1, 2)$, and $\gamma \sim \text{beta}(1, 1)$. The population sizes were integrated out analytically using inverse-gamma(10, 0.1) (Eq. 4). The substitution model was set to JC69 (the truth). We fixed $\mu = 1.0$ for all genes as in the simulation (strict molecular clock and no rate variation). The MCMC chain was run 20 million steps and sampled every 2000 steps. The first 25% samples were discarded as burn-in.

We then also inferred the species network topology from each simulated dataset, with all network operators enabled. The priors were kept unchanged. The chain was run 40 million steps (doubled chain length) and sampled every 2000 steps. The first 25% samples were discarded as burn-in.

5.2 Analysis of biological data

We analyzed a dataset of three spruce species (*Picea purpurea*, *P. likiangensis* and *P. wilsonii*) in the Qinghai-Tibet Plateau (Sun et al., 2014). The original data has 166 diploid individuals and 11 nuclear loci (50 from *P. wilsonii*, 56 from *P. purpurea*, 60 from *P. likiangensis*, and two phased haplotype sequences per individual per locus).

The original data is too large for this Bayesian method to achieve proper mixing and convergence in a reasonable time. Thus we truncated the data into two datasets by randomly selecting individuals, each having 10 individuals per species. Each diploid individual has 2 sequences, thus there are 60 sequences (from 30 individuals) and 11 loci for each dataset. The two truncated datasets have no overlapping sequences, to confirm that they can produce similar posterior estimates. The priors for the species network were $\tau_0 \sim \exp(500)$, $d = \lambda - \nu \sim \exp(0.02)$, $r = \nu/\lambda \sim \text{beta}(1, 4)$, and $\gamma \sim \text{beta}(1, 1)$. The population sizes were integrated out analytically using inverse-gamma(3, 0.003) (Eq. 4). The substitution model was HKY85 (Hasegawa et al., 1985), with independent κ (transition-transversion rate ratio) and state frequencies at each locus. The clock rate was fixed to 1.0 (strict molecular clock across branches) and gene-rate multipliers were used to account for rate variation across loci. The MCMC chain was run 1 billion steps and sampled every 20,000 steps. The first 30% samples were discarded as burn-in. For each dataset we obtained two independent runs, and the

two runs were checked for the effective sample sizes (ESS) and trace plots of parameters, to ensure consistency. The MCMC samples from the two truncated datasets were combined.

We noticed that the estimate of γ was close to 0.5 (prior mean) using the datasets above, with a large HPD interval. Thus we added back more individuals, and fixed the network topology to which having the highest posterior probability (Fig. 13), to infer the node heights and γ more accurately. Each of the two non-overlapping datasets had 20 individuals from *P. purpurea*, 15 from *P. likiangensis*, and 15 from *P. wilsonii* (100 sequences per locus). The population sizes were either inferred using MCMC under gamma(2, 2000) priors (prior mean is 0.001), or integrated out analytically using inverse-gamma(3, 0.003). The other priors and MCMC settings were unchanged. The MCMC samples from the two truncated datasets were combined.

5.3 Representation of phylogenetic networks

The species network can be represented using extended Newick format (Cardona et al., 2008), which was also used in the software PhyloNet (Than et al., 2008).

For example, the species network in Figure 1 is written as

```
((A:0.02,(B:0.01)#H1[&gamma=0.3]:0.01)S1:0.03,
       (#H1:0.02,C:0.03)S2:0.02)R:0.03;
```

where the hash sign indicates a reticulation node, and the inheritance probability is in the brackets as metadata. Such extended Newick string can be read into IcyTree (`icytree.org`) and be displayed nicely.

5.4 Software availability

The method is implemented in the add-on `SpeciesNetwork` for BEAST2 (Bouckaert et al., 2014), and is hosted publicly on GitHub (`https://github.com/zhangchicool/speciesnetwork`).

6 Acknowledgments

This research was supported by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant number 335529 to T.S.). C.Z. acknowledges his salary as well as a visit covered by this grant to the Centre for Computational Evolution, University of

Auckland, New Zealand in mid-2016. We sincerely thank Simone Linz for detailed discussion on modeling phylogenetic networks.

References

Albrecht B, Scornavacca C, Cenci A, Huson DH. 2012. Fast computation of minimum hybridization networks. *Bioinformatics*. 28:191–197.

Blum MGB, François O, Janson S. 2006. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability*. 16:2195–2214.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*. 10:e1003537.

Cardona G, Rosselló F, Valiente G. 2008. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics*. 9:532.

Drummond AJ, Bouckaert RR. 2015. Bayesian Evolutionary Analysis with BEAST. Cambridge University Press.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*. 4:e88.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*. 17:368–376.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*. 59:307–321.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*. 22:160–174.

Heath TA, Huelsenbeck JP, Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences of the United States of America*. 111:E2957–66.

Hey J, 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 of the United States of America*. 104:2785–2790.

Jones G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology*. 74:447–467.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. *Mammalian Protein Metabolism*. pp. 21–132.

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution*. 24:2669–2680.

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

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

Pardi F, Scornavacca C. 2015. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. *PLoS Computational Biology*. 11:e1004135.

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

Pybus OG, Harvey PH. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B: Biological Sciences*. 267:2267–2272.

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

Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Systematic Biology*. 56:453–466.

Rannala B, Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*. p. syw119.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*. 61:539–542.

Solís-Lemus C, Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS genetics*. 12:e1005896.

Stadler T. 2010. Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*. 267:396–404.

Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences of the United States of America*. 110:228–233.

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

Sun Y, Abbott RJ, Li L, Li L, Zou J, Liu J. 2014. Evolutionary history of Purple cone spruce (Picea purpurea) in the Qinghai-Tibet Plateau: homoploid hybrid origin and Pleistocene expansion. *Molecular Ecology*. 23:343–359.

Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on mathematics in the life sciences*. 17:57–86.

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

Thorne JL, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology*. 51:689–702.

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

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

Wilson IJ, Balding DJ. 1998. Genealogical inference from microsatellite data. *Genetics*. 150:499–510.

Wu Y. 2010. Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics*. 26:i140–i148.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*. 39:306–314.

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

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

Yu Y, Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*. 16:S10.

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

Zhu S, Degnan JH. 2016. Displayed trees do not determine distinguishability under the network multispecies coalescent. *Systematic Biology*. p. syw097.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*. 97:97–166.