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

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

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

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

[*]nakhleh@rice.edu

Contents

1 Sampling from the Posterior using RJMCMC

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

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

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

We describe the proposal workflow as follows:

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

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

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

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

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

1.1 Moves for the phylogenetic network and inheritance probabilities

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

Note that these moves might

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

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

1.1.1 Change-Parameters

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

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

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

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

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

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

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

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

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

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

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

1.1.2 Change-Topology

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

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

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

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

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

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

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

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

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

and similarly,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.1.3 Change-Dimension

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

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

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

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

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

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

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

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

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

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

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

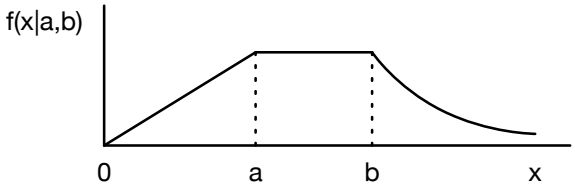

Fig. S1: **The three-piece distribution (Linear, Uniform, Exponential) for population size sampling.** Ideally, a new population size should be drawn from the prior $\Gamma(2, \psi)$. However, the inverse function Γ^{-1} cannot be solved directly. We designed the three-piece distribution as a replacement.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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The Hastings ratio of Add-Reticulation is

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

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

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

The Hastings ratio of Delete-Reticulation is

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

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

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

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

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

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

1.2 Convergence diagnostics

We make use of three commonly used diagnostics:

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

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

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

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

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

2.1 Simulation settings

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

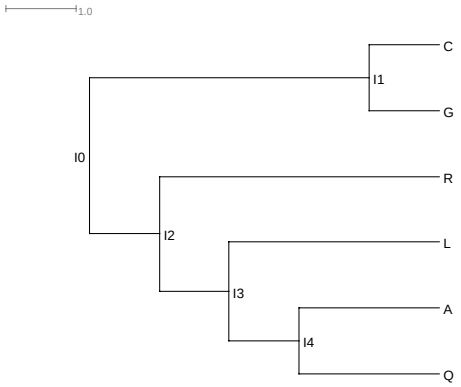

Fig. S2: **The true species tree used to generate a simulated data for testing the MCMC sampler.** The branch lengths of the species tree are measured in coalescent units.

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

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

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

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

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

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

2.2 Results

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

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

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

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

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

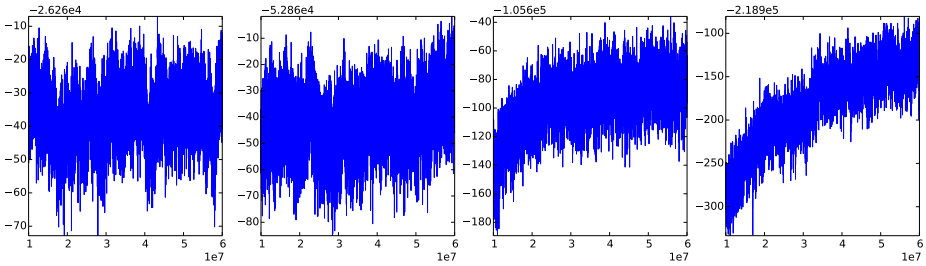

Fig. S3: **Trace plots of the MCMC chains using ∗BEAST given four data sets with 16, 32, 64, and128 loci, respectively (from left to right), simulated from the true species tree in Fig. S2.** The trace plots of the MCMC chains for 16, 32, 64-locus data sets indicate good mixing and convergence from 1×10^7,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

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

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

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

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

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

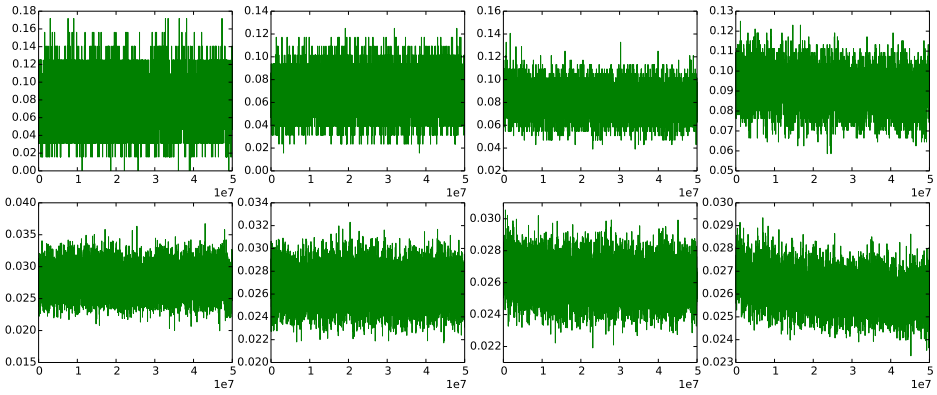

Fig. S4: **Plots of the RF distances (top row) and nrBSs (bottom row) between gene tree samples inferred by ∗BEAST and the true ones.** From left to right, the four plots correspond to the four data sets of 16, 32, 64, and 128 loci, respectively, simulated from the true species tree in Fig. S2. The RF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

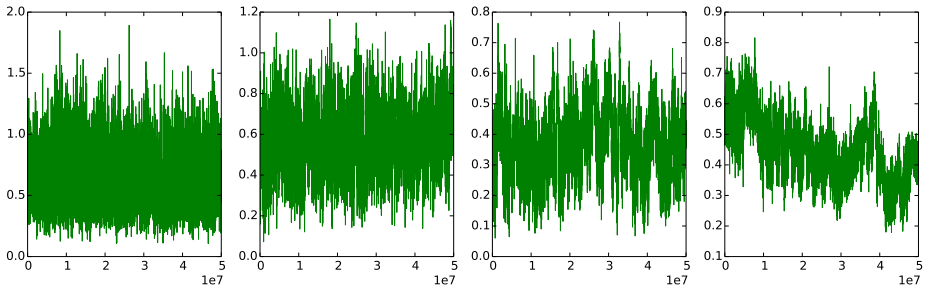

Fig. S5: **Plots of the nrBS values between the species tree sample inferred by ∗BEAST and the true species tree in Fig. S2.** The divergence times of the samples are converted to coalescent units. From left to right, the plots correspond to the four data sets (16, 32, 64, and 128 loci, respectively)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

in the quality of samples.

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

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

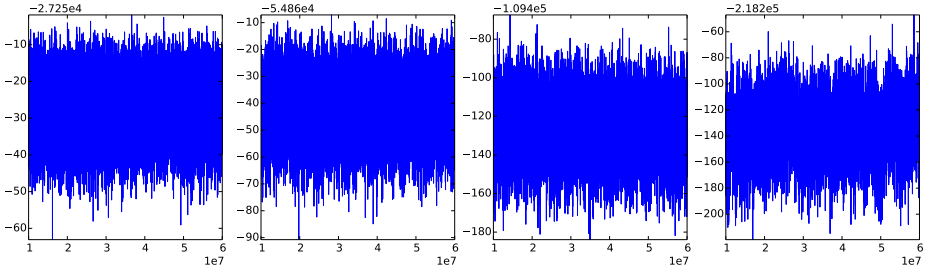

Fig. S6: **Trace plots of the MCMC chains using our method on the four data sets (16, 32, 64, and 128 loci, respectively, from left to right) simulated from the true species tree in Fig. S2.** The results indicate good mixing and convergence.

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

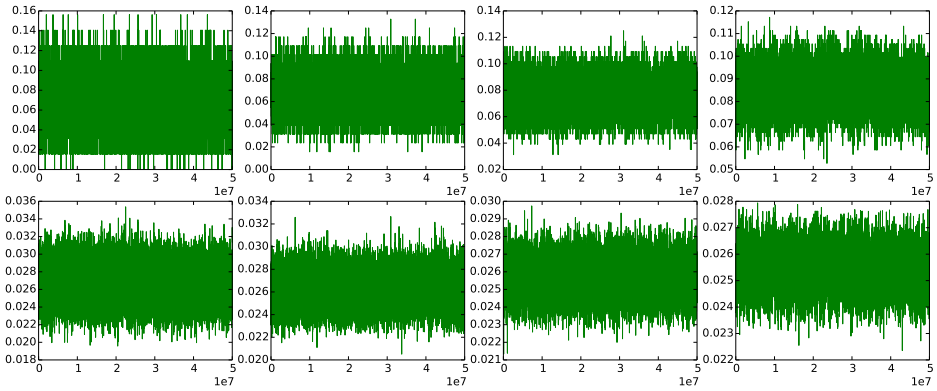

Fig. S7: **Plots of the RF distances (top row) and nrBS values (bottom row) between gene tree samples inferred by our method and the true ones.** From left to right, the plots correspond to the four data sets of 16, 32, 64, and 128 loci, respectively, simulated from the true species tree in Fig. S2.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

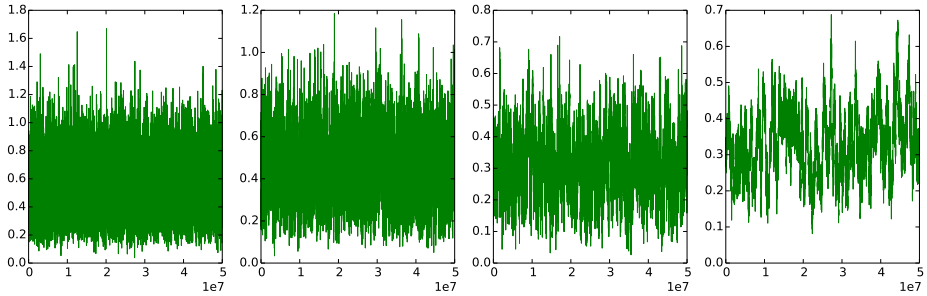

Fig. S8: **Plots of the nrBS values between the species tree sample inferred by our method and the true species tree in Fig. S2.** The divergence times of the samples are converted to coalescent units. From left to right, the plots correspond to the four data sets (16, 32, 64, 128 loci, respectively)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

3 Our Method vs. ∗BEAST on Data with Reticulations

3.1 Simulations settings

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

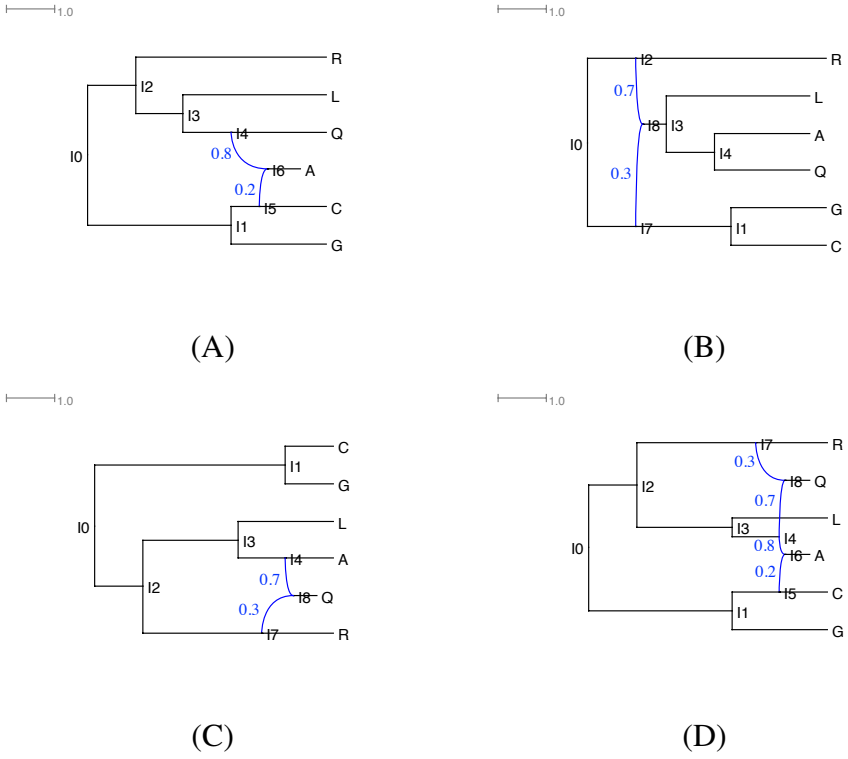

Fig. S9: **The four model phylogenetic networks used to generate the simulated data sets.** The branch lengths of the phylogenetic networks are measured in coalescent units. The inheritance probabilities are marked in blue.

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

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

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

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

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

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

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

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

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

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

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

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

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

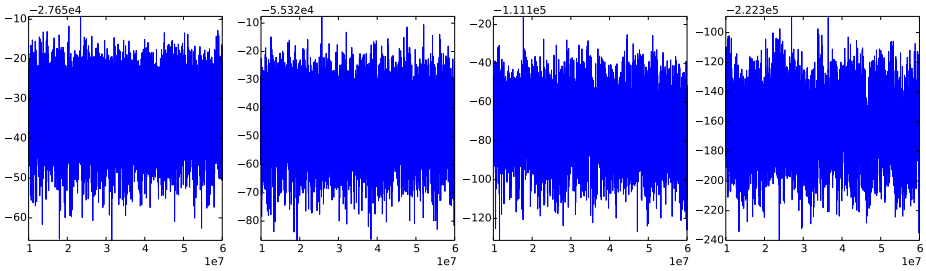

Fig. S10: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(A).** From left to right: 16, 32, 64, and 128 loci, respectively.

- Evaluation of gene tree and species tree samples.

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

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

- 95% credible sets.

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

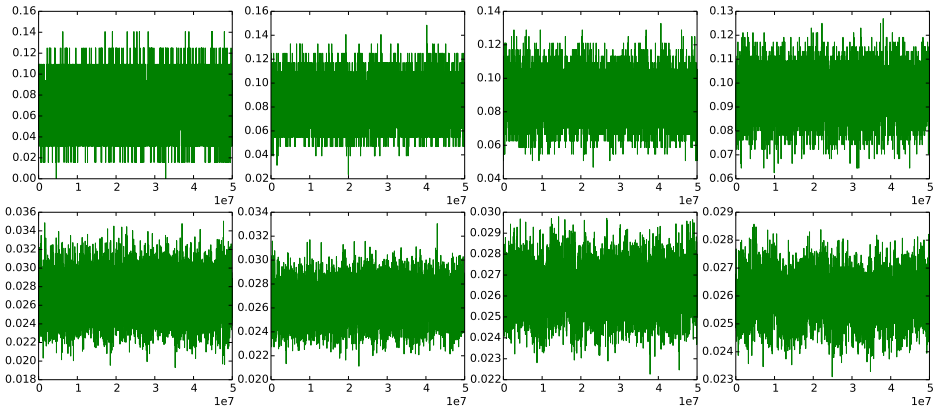

Fig. S11: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(A).** From left to right: 16, 32, 64, and 128 loci, respectively.

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

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

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

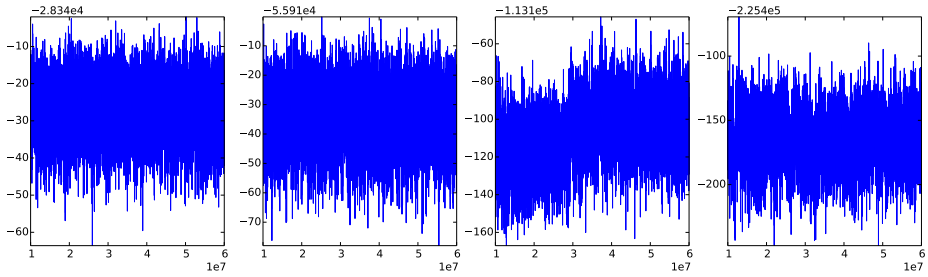

Fig. S12: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(B).** From left to right: 16, 32, 64, and 128 loci, respectively.

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

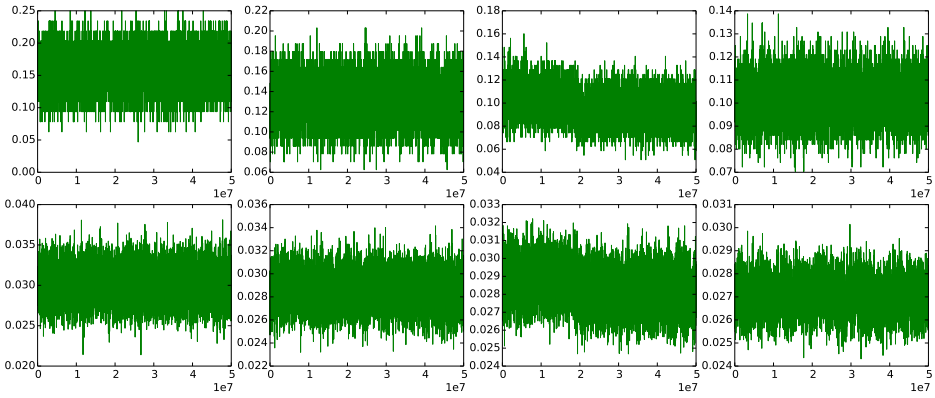

Fig. S13: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(B).** From left to right: 16, 32, 64, and 128 loci, respectively.

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

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

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

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

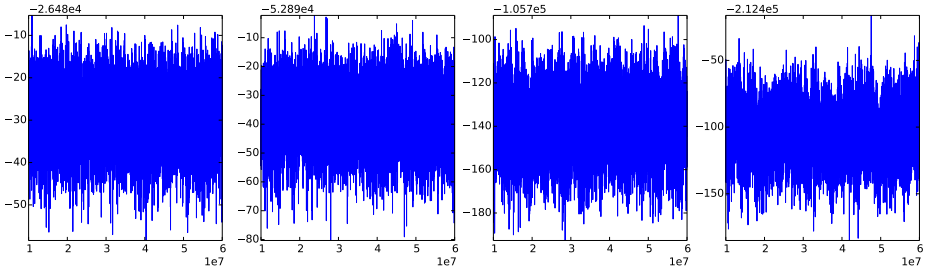

Fig. S14: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(C).** From left to right: 16, 32, 64, and 128 loci, respectively.

- Evaluation of gene tree and species tree samples.

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

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

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

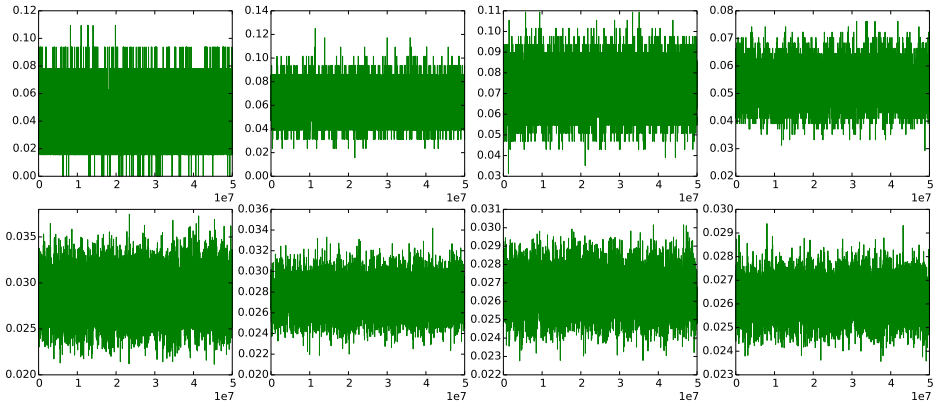

Fig. S15: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(C).** From left to right: 16, 32, 64, and 128 loci, respectively.

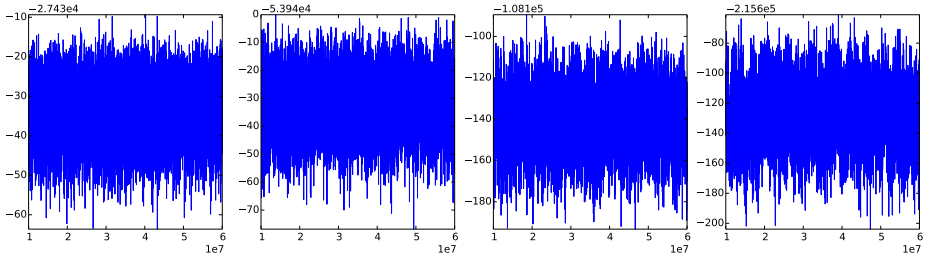

Fig. S16: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(D).** From left to right: 16, 32, 64, and 128 loci, respectively.

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

- Evaluation of gene tree and species tree samples.

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

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

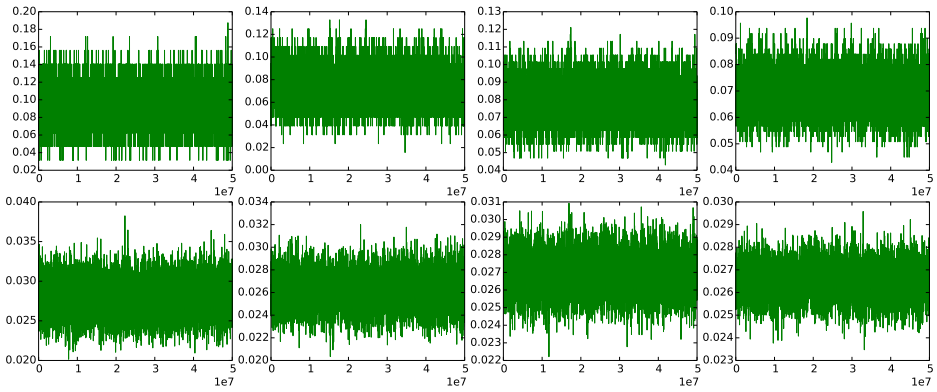

Fig. S17: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(D).** From left to right: 16, 32, 64, and 128 loci, respectively.

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

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

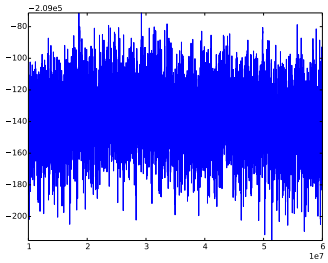

Fig. S18: **Trace plot of the MCMC chain using ∗BEAST on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).**

We considered two hypotheses:

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

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

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

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

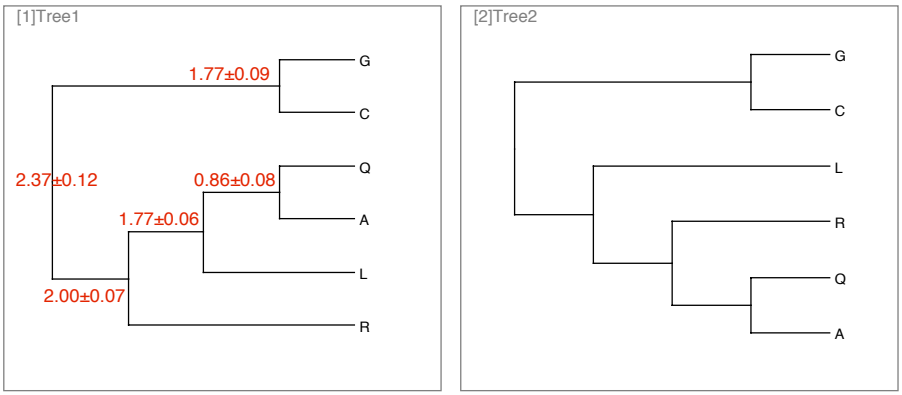

Fig. S19: **The two trees in the 95% credible set obtained by ∗BEAST on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).** The proportions of the two sampled species tree topologies are 94% (for the topology of Tree 1) and 4% (for the topology of Tree 2). The MAP topology (Tree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

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

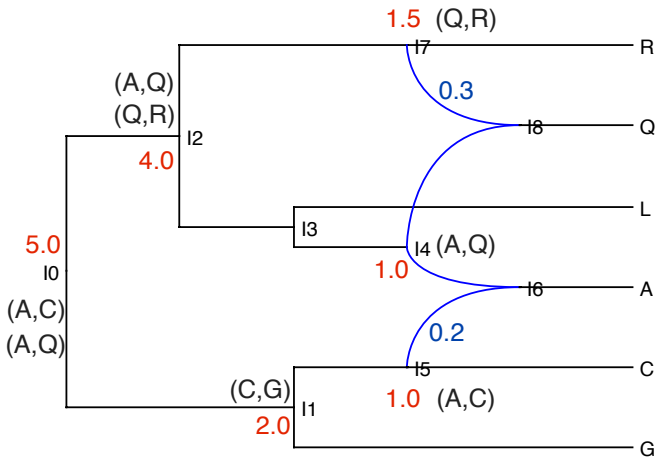

Fig. S20: **The true phylogenetic network used to simulate the 128-locus data set.** The divergence times in coalescent units are shown in red. The inheritance probabilities associated with the two reticulation edges are shown in blue. Node I1 is the MRCA of (C,G). The MRCA of (A,Q) could be one of the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

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

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

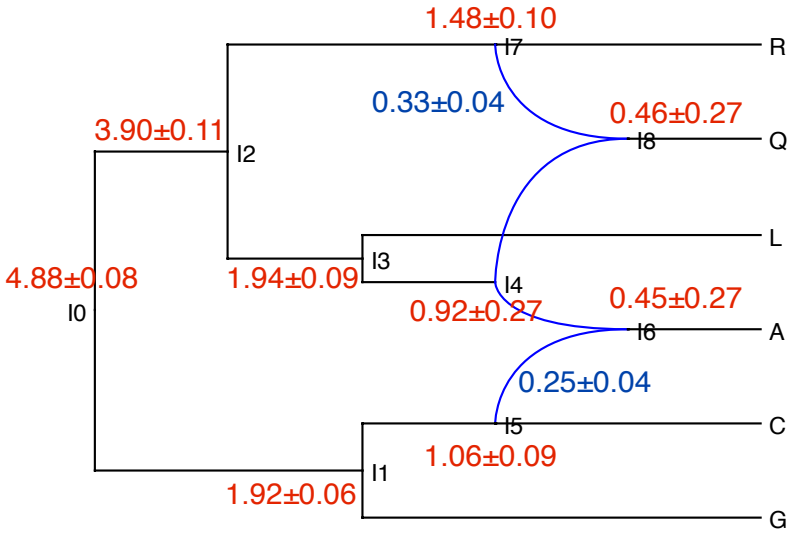

Fig. S21: **The 95% credible set obtained by our method on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).** The single topology in the 95% credible set is the true network. The divergence times in coalescent units are shown in red and the inheritance probabilities are shown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

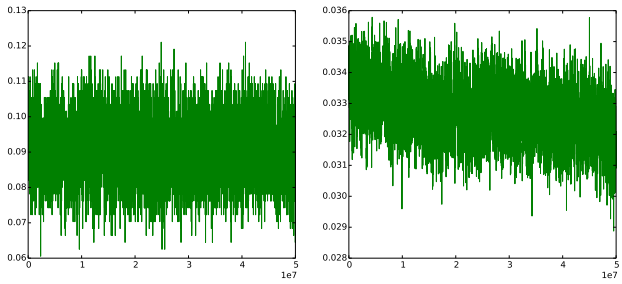

Fig. S22: **Plots of RF distances (left) and nrBS values (right) between gene trees sampled by ∗BEAST and the true ones.** The input is the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).

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

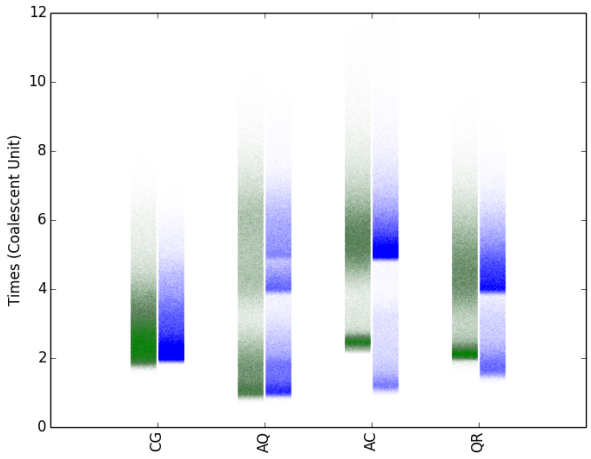

Fig. S23: **The divergence times in coalescent units of the MRCA of (C,G), (A,Q), (A,C), (Q,R) from co-estimated gene trees inferred by ∗BEAST (green) and our method (blue).** The input is the 128-locus data set simulated from phylogenetic network of Fig. S9(D).

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

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

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

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

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

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

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

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

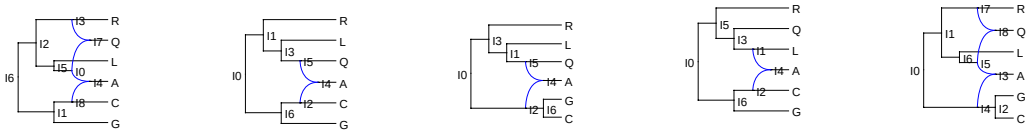

Fig. S24: **The five network topologies sampled using the method (16) on the true gene trees from the four data sets (16, 32, 64, and 128 loci) simulated on the phylogenetic network of Fig. S9(D).** Left to right: the true network, the 1-reticulation network missing the reticulation edge $R \to Q$, and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

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

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

4 Simulations under Intermixture/Gene Flow Models

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

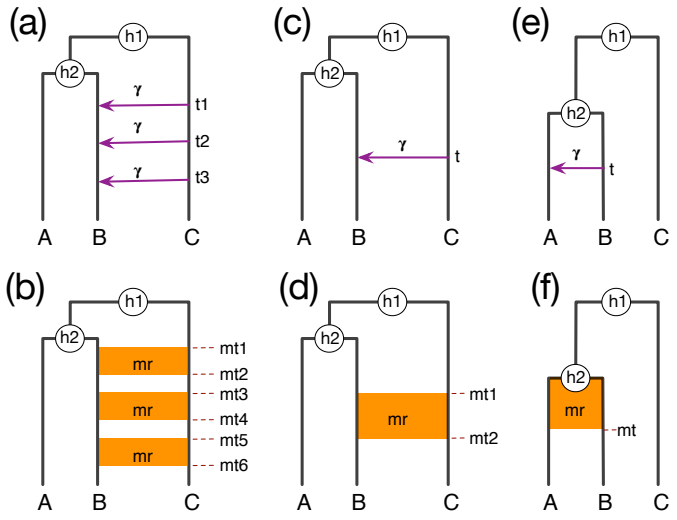

Fig. S25: True phylogenetic histories with intermixture and gene flow models. Recurrent reticulations between non-sister taxa (a,b), a single reticulation between non-sister taxa (c,d), and a single reticulation between sister taxa (e,f) is captured under both the intermixture model (top) and gene flow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

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

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

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

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

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

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

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

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

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

S25(d) $mt_2 = 0$: ms 3 200 -T -I 3 1 1 1 -em 0.0 2 1 0.2666667 -em 0.75 2 1 0.0 -ej 0.75 3 2 -ej 1.25 2 1

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

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

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

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

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

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

4.1 MCMC settings

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

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

4.1.1 The effect of the number of individuals

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

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

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

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

As for the estimated number of reticulations, we found when $\Delta t = 1$, increasing the number of individuals from 1 to 3 leads to a increase in the number of reticulations (from

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

4.2 Paraphyletic intermixture/gene flow

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

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

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

4.3 Isolation-migration between sister species

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

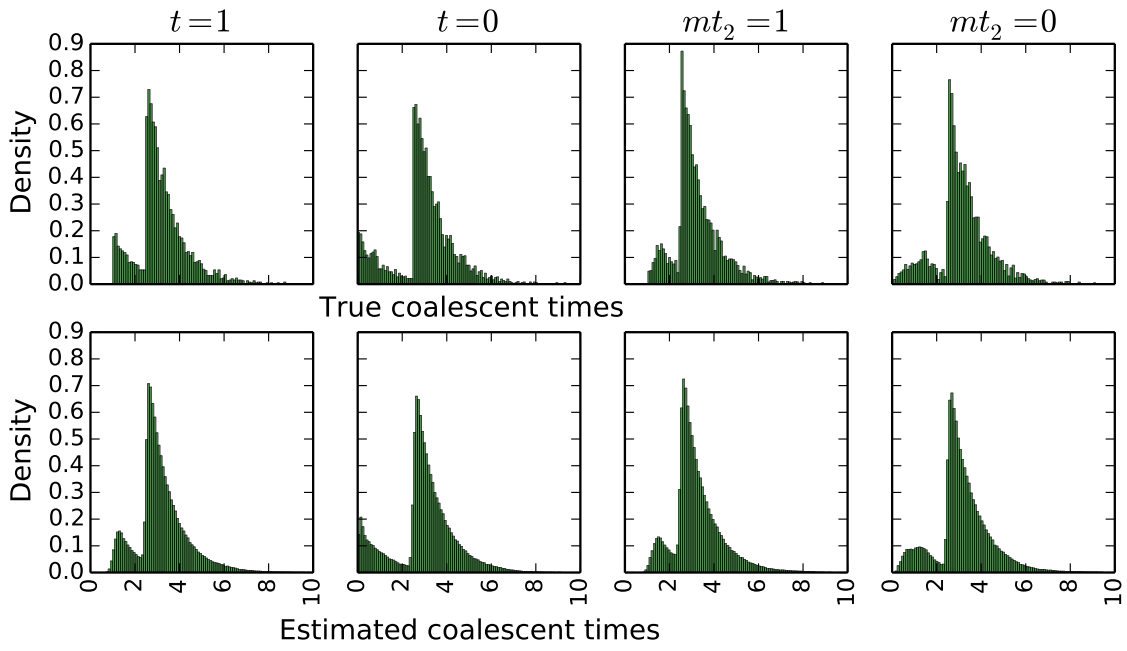

Fig. S26: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on $4,000$ loci generated under the models of Fig. S25(c) and Fig. S25(d).

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

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

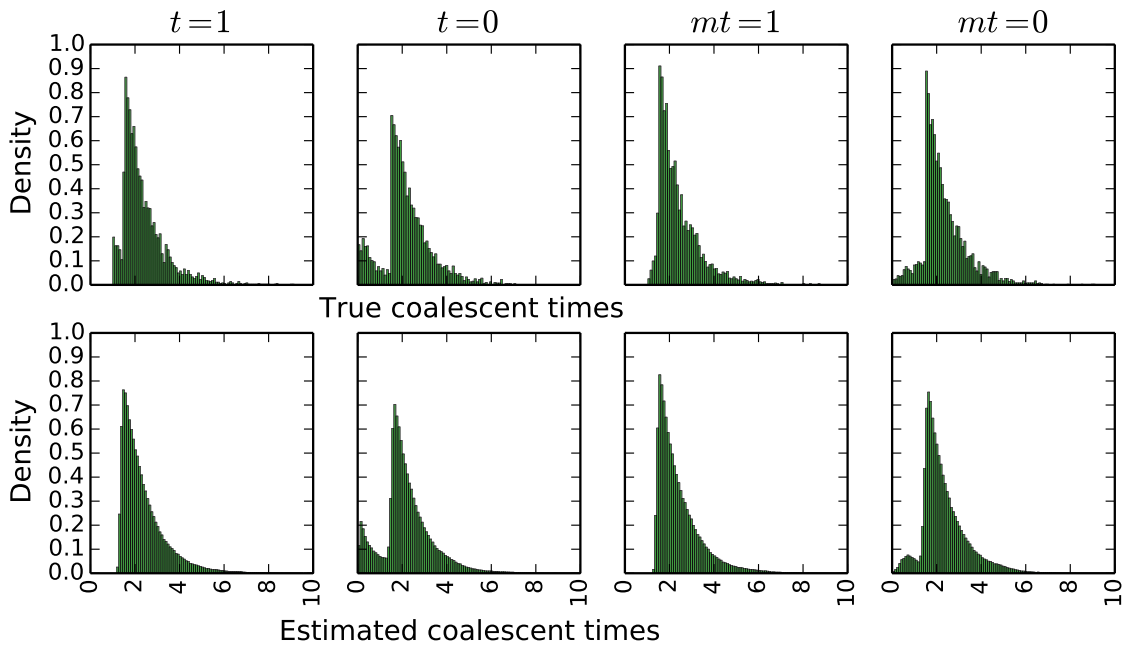

Fig. S27: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on $4,000$ loci generated under the models of Fig. S25(e) and Fig. S25(f).

5 Analysis of a Yeast Data Set

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

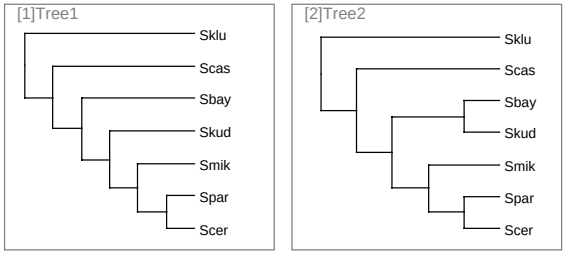

Fig. S28: **The species trees of seven Saccharomyces species.** (Left) The topology inferred from concatenation method (13) and the main topology sampled by BEST (4). (Right) The topology sampled by BEST with the second highest proportion (4).

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

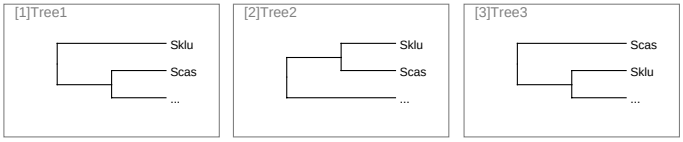

Fig. S29: **Relationships of Sklu and Scas in several gene tree topologies of seven Saccharomyces species.**

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

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

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

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

5.1 MCMC settings

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

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

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

5.2 Data preprocessing

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

5.3 Results for the full data set

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

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

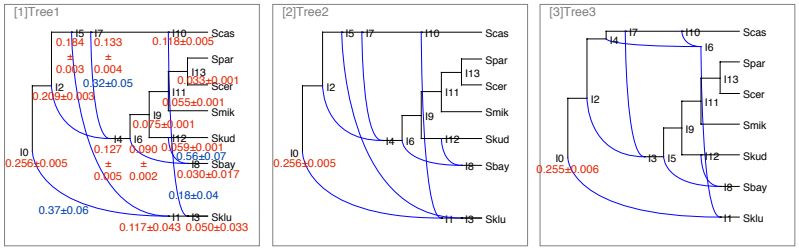

Fig. S30: **The three main phylogenetic networks in the 95% credible sets from the yeast data set using our method.** The divergence times are labeled in red, and the inheritance probabilities are marked in blue.

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

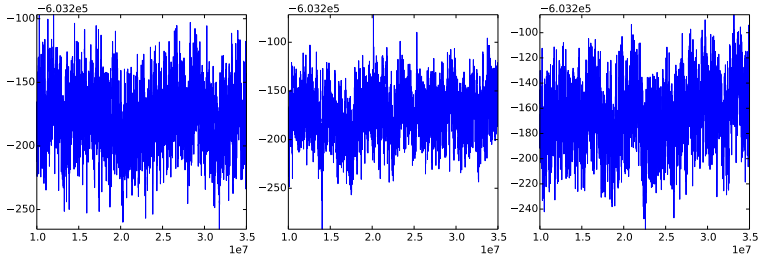

Fig. S31: **Trace plots of MC3 chains using our method on the yeast data set.**

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

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

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

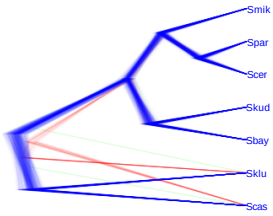

Fig. S32: **The densiTree plot of the species trees sampled from ∗BEAST given the yeast data set.**

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

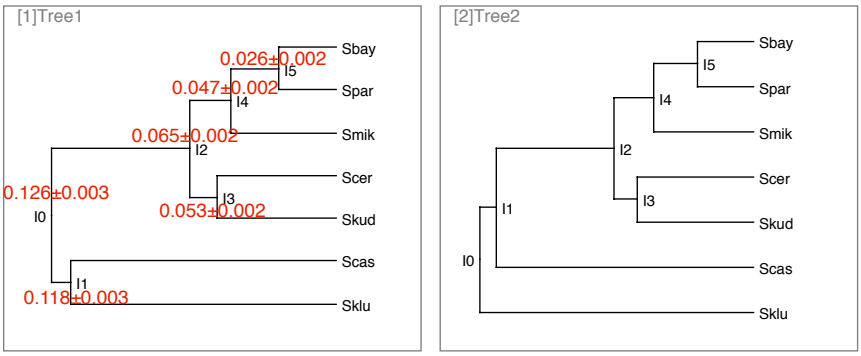

Fig. S33: **The two main species trees in the 95% credible set from the yeast data set using ∗BEAST.** The proportions for Tree 1 and Tree 2 are 78.4% and 16.9%, respectively. The divergence times are marked in red.

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

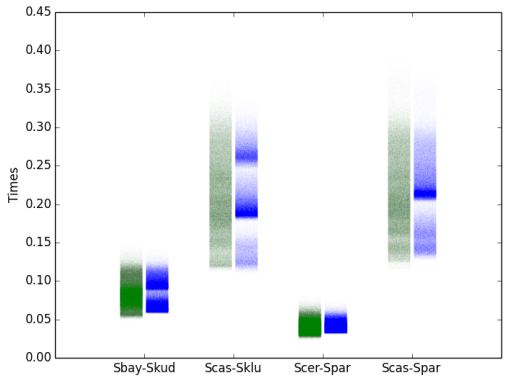

Fig. S34: **The divergence times of the MRCAs of (Sbay,Skub), (Scas,Sklu), (Scer,Spar), and (Scas,Spar) from estimated gene tree samples inferred by $*$BEAST (green), and our method (blue).** The input is the full yeast data set.

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

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

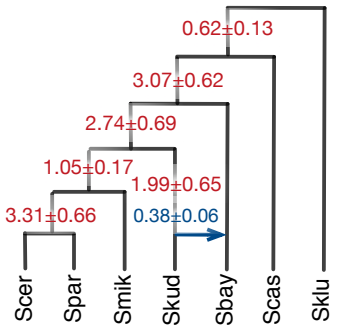

Fig. S35: **The phylogenetic network in the 95% credible set from the data set of 106 loci from seven Saccharomyces species using the Bayesian inference method taking gene tree topologies as input.** The divergence times are marked in red, and the inheritance probability is marked in blue.

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

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

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

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

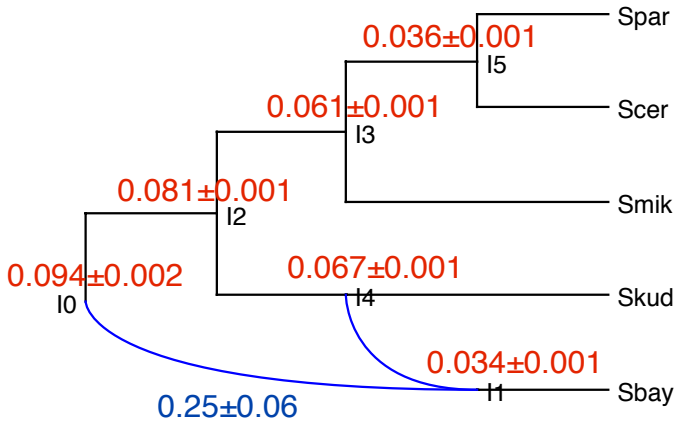

Fig. S36: **The phylogenetic network in the 95% credible set from the data set of 106 loci from five Saccharomyces species using our method.** The divergence times are marked in red, and the inheritance probability is marked in blue.

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

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

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

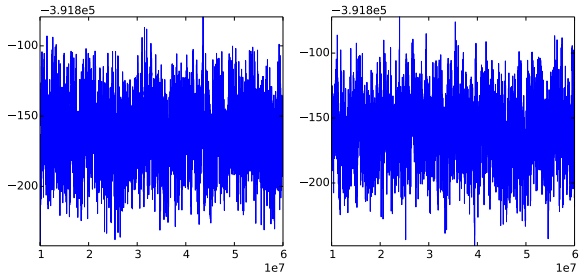

Fig. S37: **Trace plots of MC3 chains using our method given the data set of 106 loci from five Saccharomyces species.**

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

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

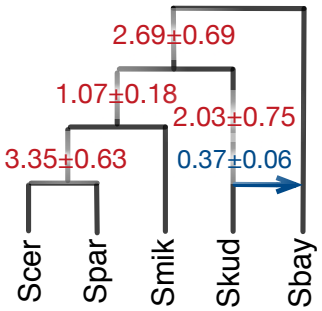

Fig. S38: **The phylogenetic network in the 95% credible sets from the data set of 106 loci from five Saccharomyces species using the Bayesian inference method taking gene tree topologies as input.** The divergence times are marked in red, and the inheritance probability is marked in blue.

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

6 Runtimes

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

6.1 Simulations

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

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

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

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

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

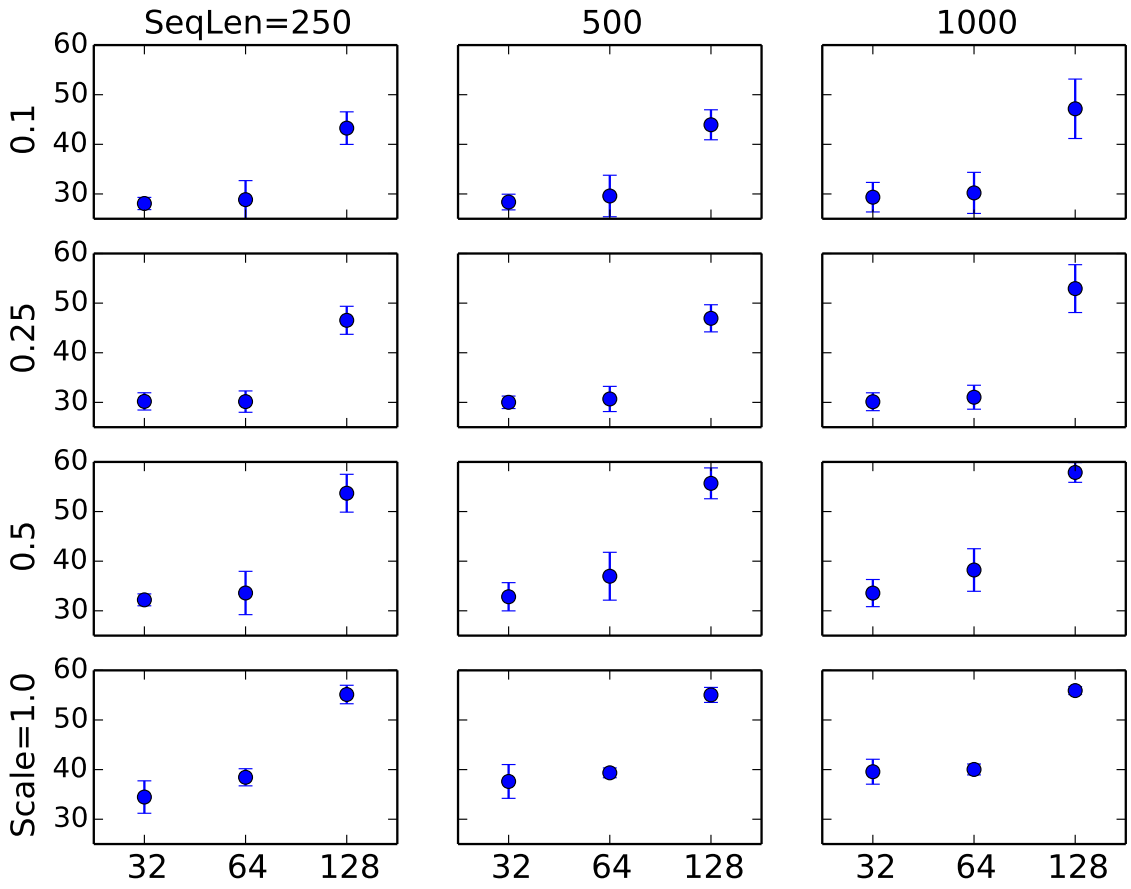

Fig. S39: **The runtimes in minutes under different simulation conditions.** From top to bottom: 0.1, 0.25, 0.5, 1.0 divergence time scale, respectively. From left to right: 250, 500, and 1000 bps sequence length, respectively. Within each plot: 32, 64, 128 loci, respectively.

- The paraphyletic intermixture between non-sister species in S25(c), $t = 1$: 42.8 ± 3.4
- The paraphyletic intermixture between non-sister species in S25(c), $t = 0$: 42.1 ± 3.4
- The paraphyletic gene flow between non-sister species in S25(d), $mt_2 = 1$: 45.8 ± 4.6
- The paraphyletic gene flow between non-sister species in S25(d), $mt_2 = 0$: 46.5 ± 5.7
- The isolation-migration between sister species in S25(e), $t = 1$: 36.2 ± 3.8
- The isolation-migration between sister species in S25(e), $t = 0$: 48.8 ± 7.6

- The isolation-migration between sister species in S25(f), $mt = 1$: 36.5 ± 3.9
- The isolation-migration between sister species in S25(f), $mt = 0$: 47.7 ± 6.9

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

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

6.2 Biological data sets

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

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

7 PhyloNet Implementation and Usage

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

8 References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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