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 $G = \{g_1, \dots, g_m\}$.
  - 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 happen 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 the 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

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

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

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

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

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

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

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

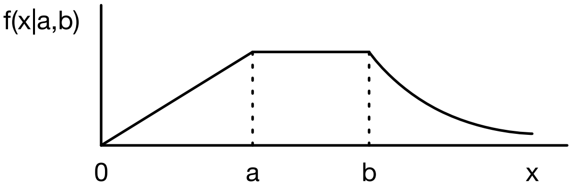


Fig. S1: **The three-piece distribution (Linear, Uniform, Exponential) for population size sampling.** Ideally, a new population size should be drawn from the prior $\Gamma(2, \psi)$. However, the inverse function Γ^{-1} cannot be solved directly. We designed the three-piece distribution as a replacement. 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 $\frac{p}{(1-p)(b-0.5a)} = \frac{6}{2b-a}$.

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

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 $u_2 \sim \mathrm{Uniform}(0,1)$. 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 $u_3 \sim \mathrm{Uniform}(0,1)$. 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 $u_4 \sim \mathrm{Uniform}(0,1)$. 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 $u_5 \sim \mathrm{Uniform}(0,1)$. We have $\theta_{e_{11}} = h(u_5)$. The partial derivative is $h'(u_5)$.
- the population size of e_{21}. Generate $u_6 \sim \mathrm{Uniform}(0,1)$. 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

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

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.

1.3 Testing the MCMC sampler 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.

1.3.1 Simulation settings

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

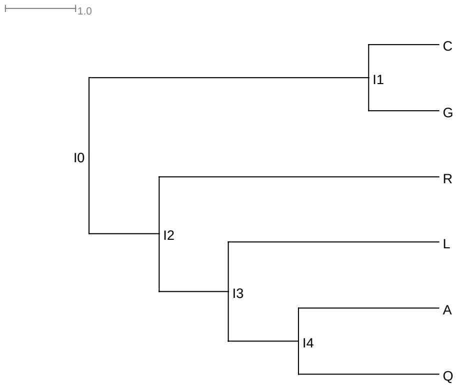


Fig. S2: **The true species tree used to generate a simulated data for testing the MCMC sampler.** The branch lengths of the species tree are measured in coalescent units.

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 (13) 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.

1.3.2 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.

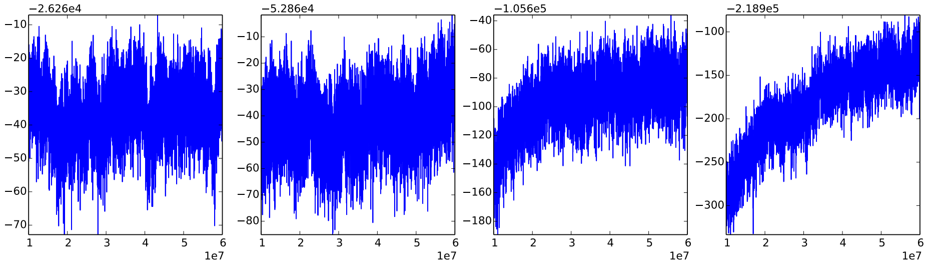


Fig. S3: **Trace plots of the MCMC chains using $*$BEAST given four data sets with 16, 32, 64, and128 loci, respectively (from left to right), simulated from the true species tree in Fig. S2.** The trace plots of the MCMC chains for 16, 32, 64-locus data sets indicate good mixing and convergence from 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 (14) 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

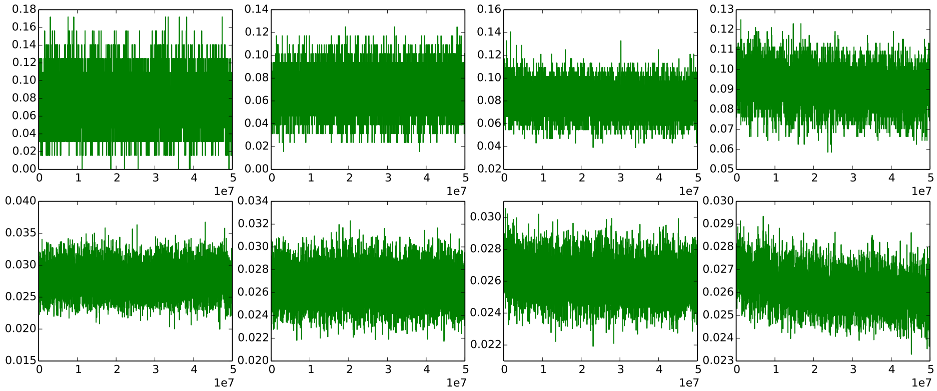


Fig. S4: **Plots of the RF distances (top row) and nrBSs (bottom row) between gene tree samples inferred by ∗BEAST and the true ones.** From left to right, the four plots correspond to the four data sets of 16, 32, 64, and 128 loci, respectively, simulated from the true species tree in Fig. S2. The RF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

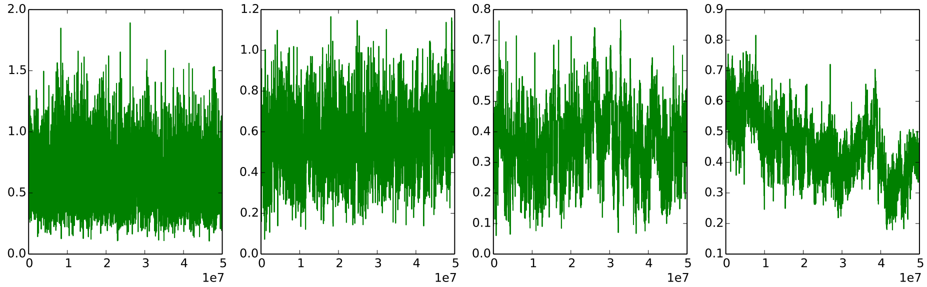


Fig. S5: **Plots of the nrBS values between the species tree sample inferred by ∗BEAST and the true species tree in Fig. S2.** The divergence times of the samples are converted to coalescent units. From left to right, the plots correspond to the four data sets (16, 32, 64, and 128 loci, respectively) 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.

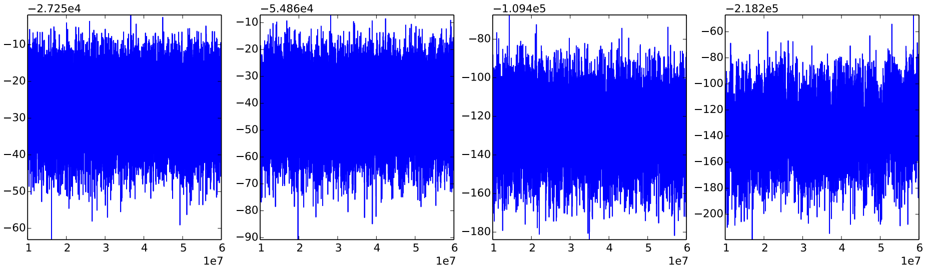


Fig. S6: **Trace plots of the MCMC chains using our method on the four data sets (16, 32, 64, and 128 loci, respectively, from left to right) simulated from the true species tree in Fig. S2.** The results indicate good mixing and convergence.

- 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.

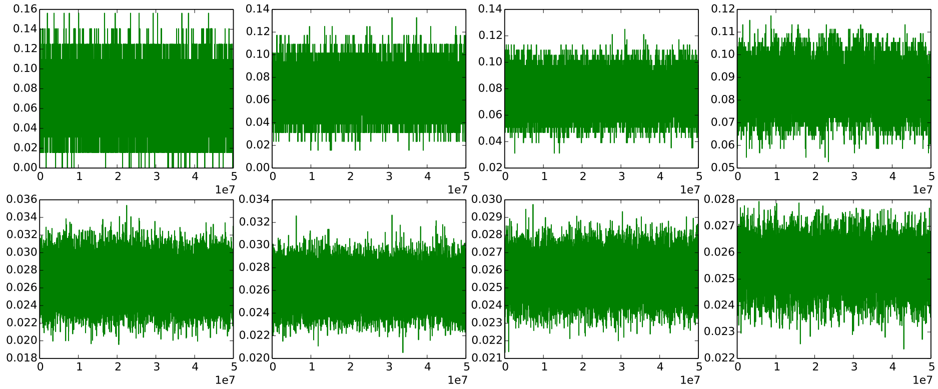


Fig. S7: **Plots of the RF distances (top row) and nrBS values (bottom row) between gene tree samples inferred by our method and the true ones.** From left to right, the plots correspond to the four data sets of 16, 32, 64, and 128 loci, respectively, simulated from the true species tree in Fig. S2. 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.

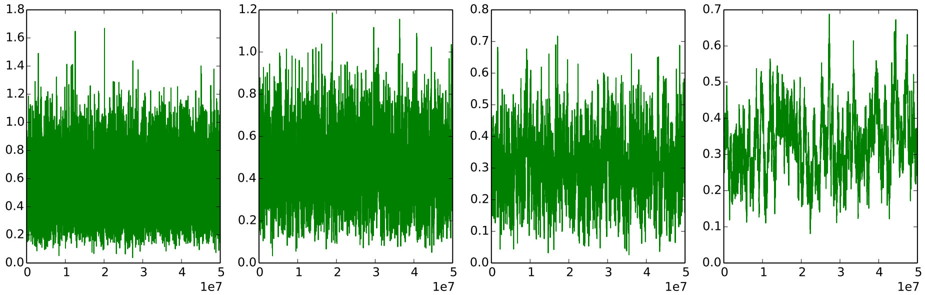
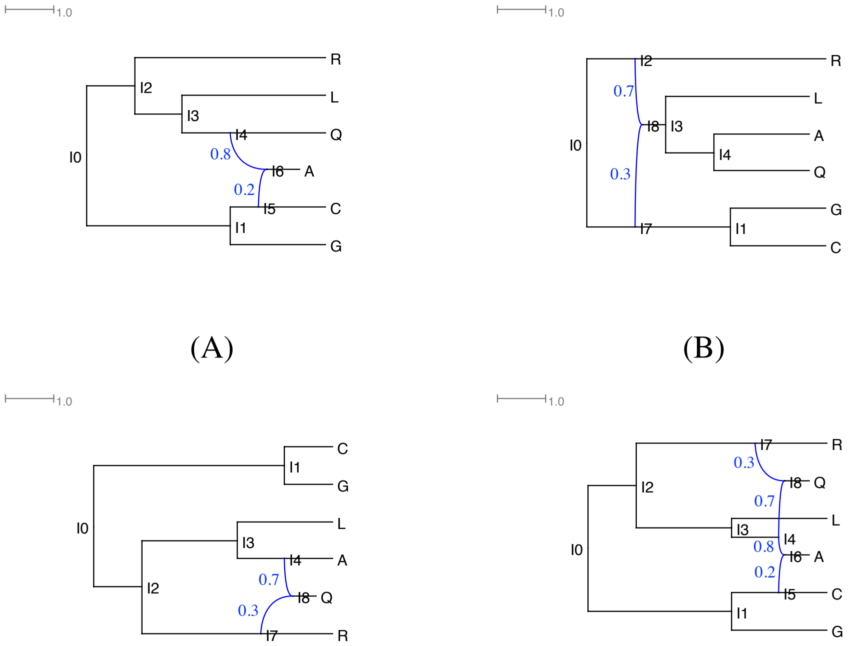


Fig. S8: **Plots of the nrBS values between the species tree sample inferred by our method and the true species tree in Fig. S2.** The divergence times of the samples are converted to coalescent units. From left to right, the plots correspond to the four data sets (16, 32, 64, 128 loci, respectively) 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.

2 Simulations

2.1 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).



(A)

(B)

(C)

(D)

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 (13) under the GTR model. We used $\theta = 0.036$ for the population mutation rate and 500 bps for the sequence length. The command is:

```
seq-gen -mgtr -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, 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.

2.2 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 $0.17 \sim 0.18$.

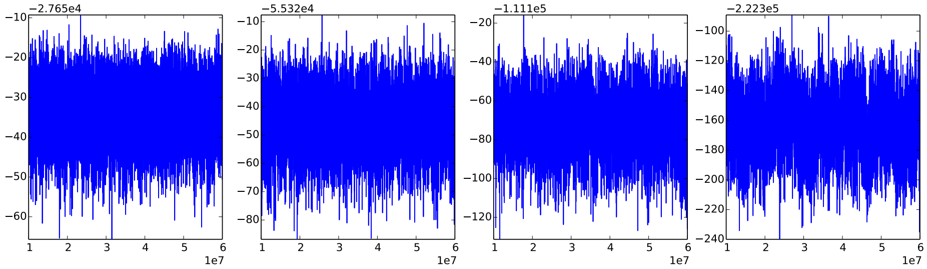


Fig. S10: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(A).** From left to right: 16, 32, 64, and 128 loci, respectively.

- 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).

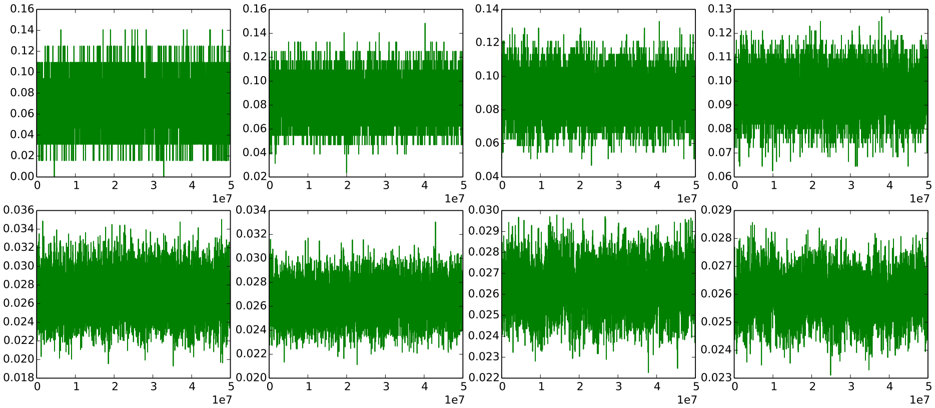


Fig. S11: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(A).** From left to right: 16, 32, 64, and 128 loci, respectively.

  - 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-

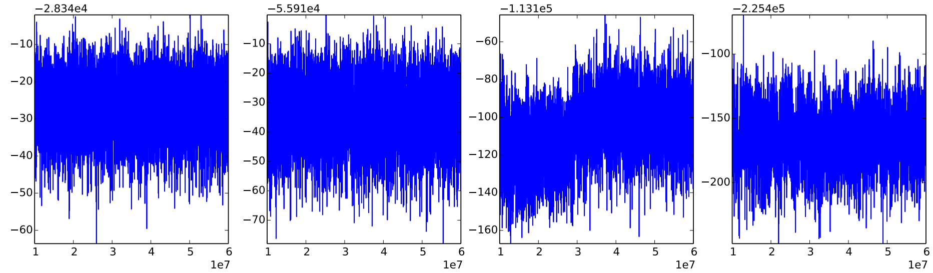


Fig. S12: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(B).** From left to right: 16, 32, 64, and 128 loci, respectively.

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

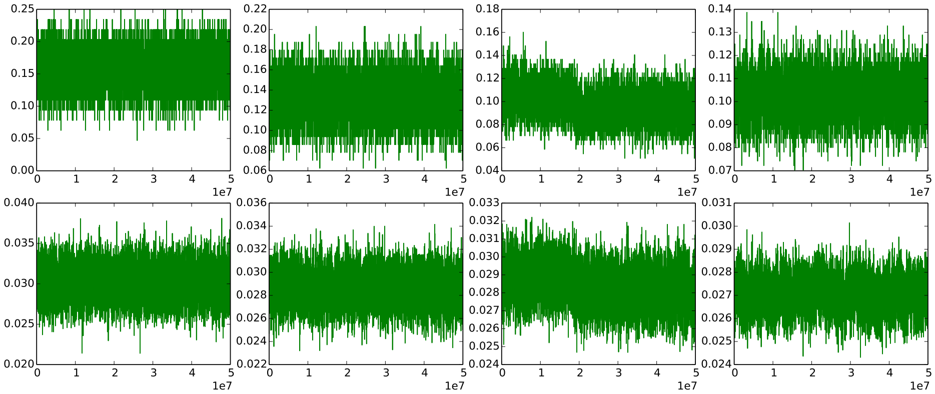


Fig. S13: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(B).** From left to right: 16, 32, 64, and 128 loci, respectively.

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 $0.16 \sim 0.17$.

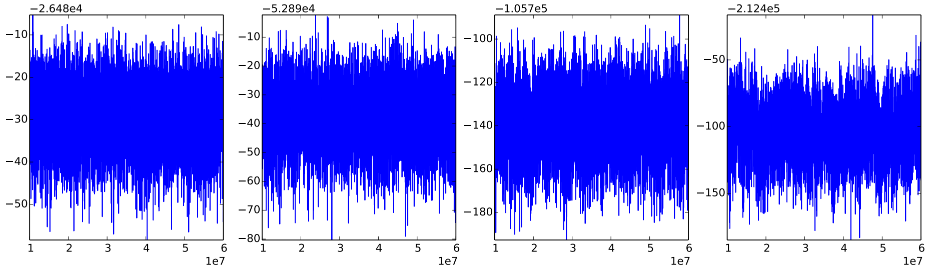


Fig. S14: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(C).** From left to right: 16, 32, 64, and 128 loci, respectively.

- 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.

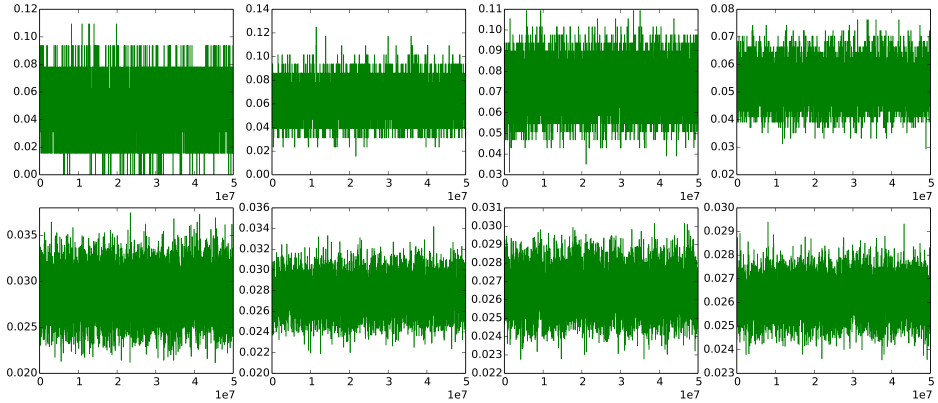


Fig. S15: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(C).** From left to right: 16, 32, 64, and 128 loci, respectively.

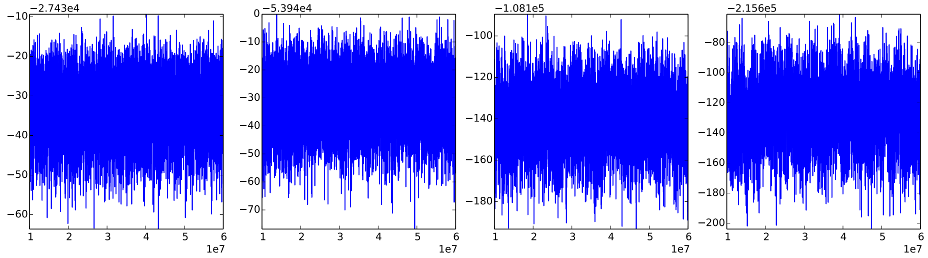


Fig. S16: **Trace plots of the MCMC chains using our method on the data sets simulated on the phylogenetic network of Fig. S9(D).** From left to right: 16, 32, 64, and 128 loci, respectively.

- 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.

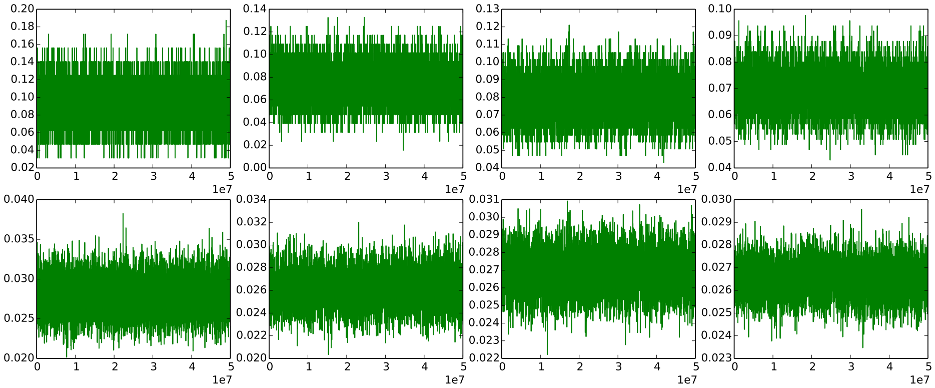


Fig. S17: **Plots of the RF distances (upper) and nrBS values (lower) between gene tree samples inferred by our method and the true ones on the data sets simulated on the phylogenetic network of Fig. S9(D).** From left to right: 16, 32, 64, and 128 loci, respectively.

2.3 ∗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.

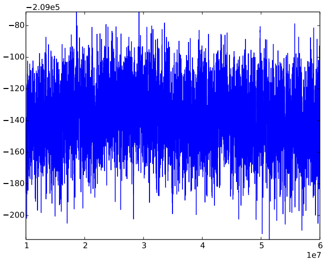


Fig. S18: **Trace plot of the MCMC chain using ∗BEAST on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).**

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 (maximum 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

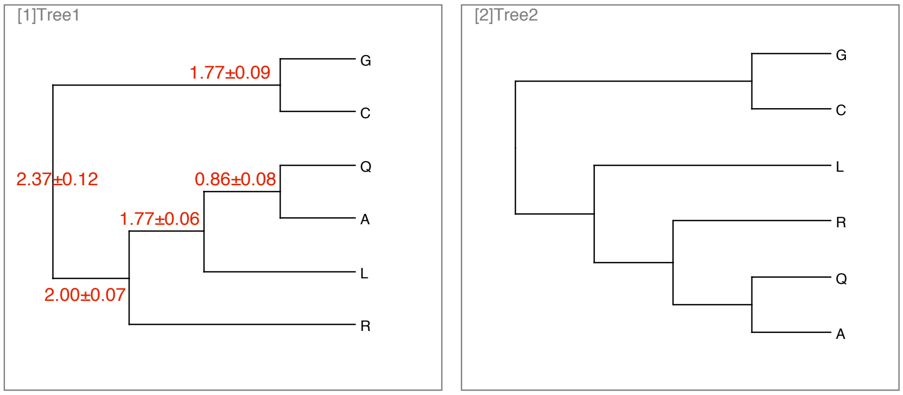


Fig. S19: **The two trees in the 95% credible set obtained by $*$BEAST on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).** The proportions of the two sampled species tree topologies are 94% (for the topology of Tree 1) and 4% (for the topology of Tree 2). The MAP topology (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.

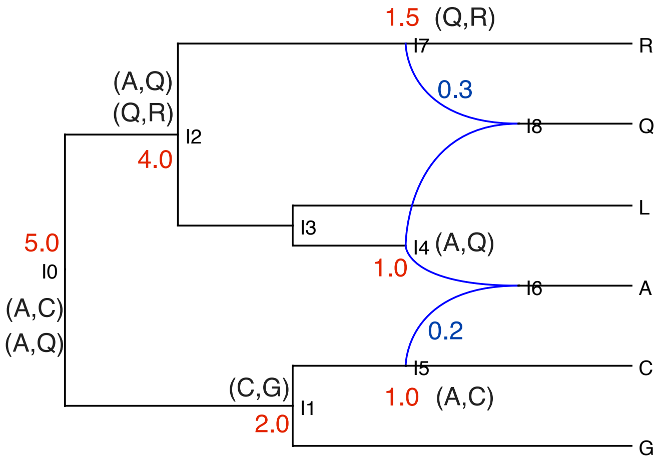


Fig. S20: **The true phylogenetic network used to simulate the 128-locus data set.** The divergence times in coalescent units are shown in red. The inheritance probabilities associated with the two reticulation edges are shown in blue. Node I1 is the MRCA of (C,G). The MRCA of (A,Q) could be one of the 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.

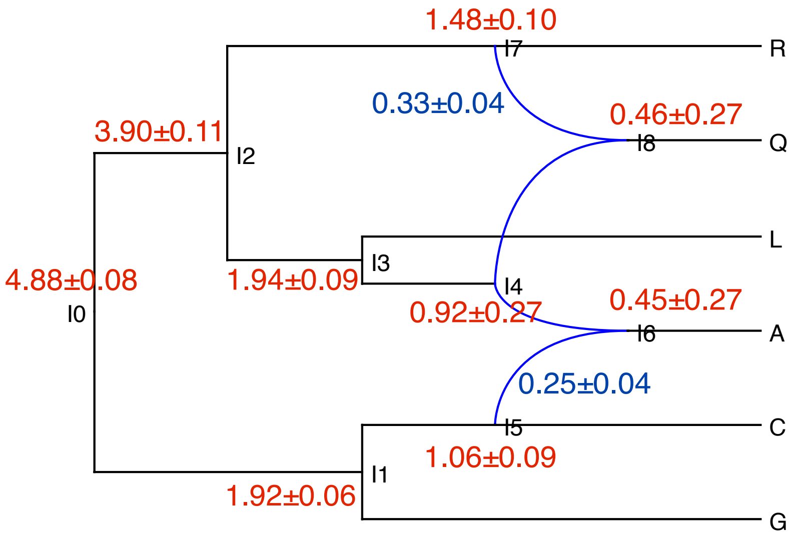


Fig. S21: **The 95% credible set obtained by our method on the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).** The single topology in the 95% credible set is the true network. The divergence times in coalescent units are shown in red and the inheritance probabilities are shown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

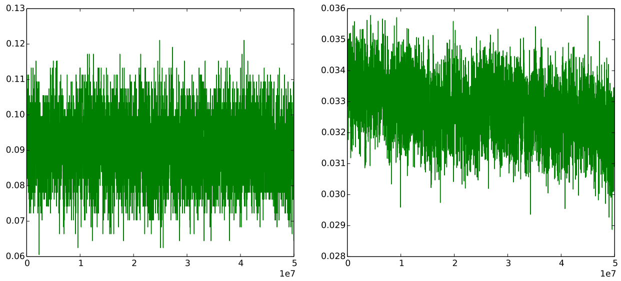


Fig. S22: **Plots of RF distances (left) and nrBS values (right) between gene trees sampled by $*$BEAST and the true ones.** The input is the 128-locus data set simulated on the phylogenetic network of Fig. S9(D).

- 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-

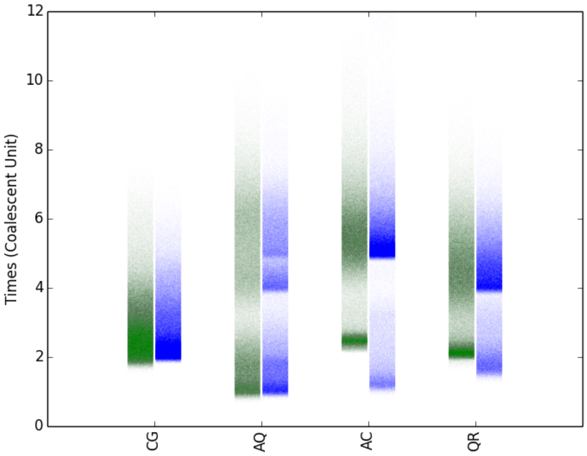


Fig. S23: **The divergence times in coalescent units of the MRCA of (C,G), (A,Q), (A,C), (Q,R) from co-estimated gene trees inferred by $*$BEAST (green) and our method (blue).** The input is the 128-locus data set simulated from phylogenetic network of Fig. S9(D).

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.

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

We used RAxML (16) 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.

2.5 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 (19), which infers phylogenetic networks and inheritance probabilities given gene tree topologies (command `MCMC_GT` in PhyloNet (18)). 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.

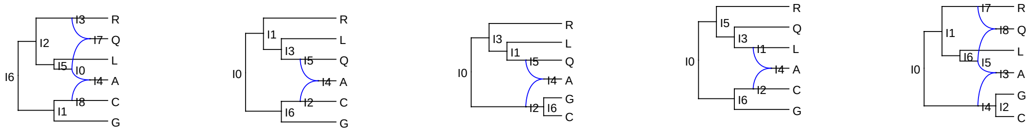


Fig. S24: **The five network topologies sampled using the method (19) on the true gene trees from the four data sets (16, 32, 64, and 128 loci) simulated on the phylogenetic network of Fig. S9(D).** Left to right: the true network, the 1-reticulation network missing the reticulation edge $R \rightarrow Q$, 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.

3 Simulation with Replicates

3.1 Simulations settings

Model phylogenetic network. We used the phylogenetic nework shown in Fig. S25 as the model species phylogeny. The scale parameter of the divergence times s was varied to take on values in the set $\{0.1, 0.25, 0.5, 1.0\}$. Setting $s = 0.1$ results in very short branches and, consequently, the hardest data sets on which to estimate parameters. Setting $s = 1.0$ results in longer branches and higher signal for a more accurate estimate of the parameter values. It is important to note that the topology, reticulation event, divergence times (with $s = 1.0$) and population size are inspired by the species phylogeny recovered from the Anopheles mosquitoes data set (Fontaine *et al.* 5, Wen *et al.* 20).

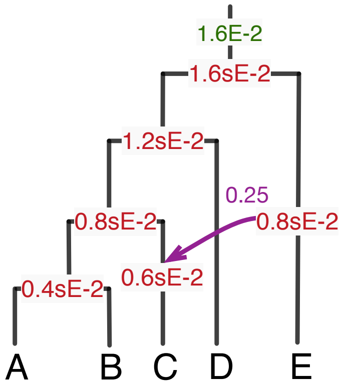


Fig. S25: A model phylogenetic network used to generate simulated data. The divergence times in units of expected number of mutations per site, the population size parameter in units of population mutation rate per site, and the inheritance probability are marked in red, green, and purple, respectively. Parameter s is used to scale the divergence times.

Model gene trees. For each setting of the four settings of s values, we simulated 20 data sets with 128 independent loci. For each of those 20 data sets, the program ms (Hudson 8) was used to simulate 128 gene trees on each dataset. The commands are listed as follows.

$s = 0.1$ ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.025 4 3 -es 0.0375 1 0.3 -ej 0.05 6 3 -ej 0.05 2 1 -ej 0.075 5 3 -ej 0.1 3 1

$s = 0.25$
```
ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.0625 4 3 -es 0.09375 1 0.3 -ej 0.125 6 3 -ej 0.125 2 1 -ej 0.1875 5 3 -ej 0.25 3 1
```

$s = 0.5$
```
ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.125 4 3 -es 0.1875 1 0.3 -ej 0.25 6 3 -ej 0.25 2 1 -ej 0.375 5 3 -ej 0.5 3 1
```

$s = 1.0$
```
ms 5 128 -T -I 5 1 1 1 1 1 -ej 0.25 4 3 -es 0.375 1 0.3 -ej 0.5 6 3 -ej 0.5 2 1 -ej 0.75 5 3 -ej 1.0 3 1
```

Sequences. The program Seq-gen (Rambaut and Grassly 13) was used to generate sequence alignments down the gene trees under the Jukes Cantor model. Sequence alignments were generated with lengths of 250, 500, and 1000 sites. The command is:

```
seq-gen -mHKY -lseqLen -s0.008
```

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

3.2 MCMC settings

For each data set, we ran an MCMC chain of 8×10^6 iterations with 3×10^6 burn-in. One sample was collected from every 5,000 iterations, resulting in a total of 1,000 collected samples.

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

3.3 Performance on data simulated under the intermixture model

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

Divergence times. Fig. S26 shows the estimates obtained for the divergence time at the root of the network. Three observations are in order. First, for any combination of sequence length and scaling parameter value, the divergence time estimate converges to the true value as the number of loci increases. Second, for any combination of number of loci and scaling parameter value, the divergence time estimate converges to the true value. Third, the estimates are relatively poor only under the extreme settings of scaling parameter value 0.1 and sequence length 250. In this case, the signal in the sequence data is too weak to obtain good estimates. However, it is worth noting that even under this setting, using 128 loci produces a very accurate estimate of the divergence time.

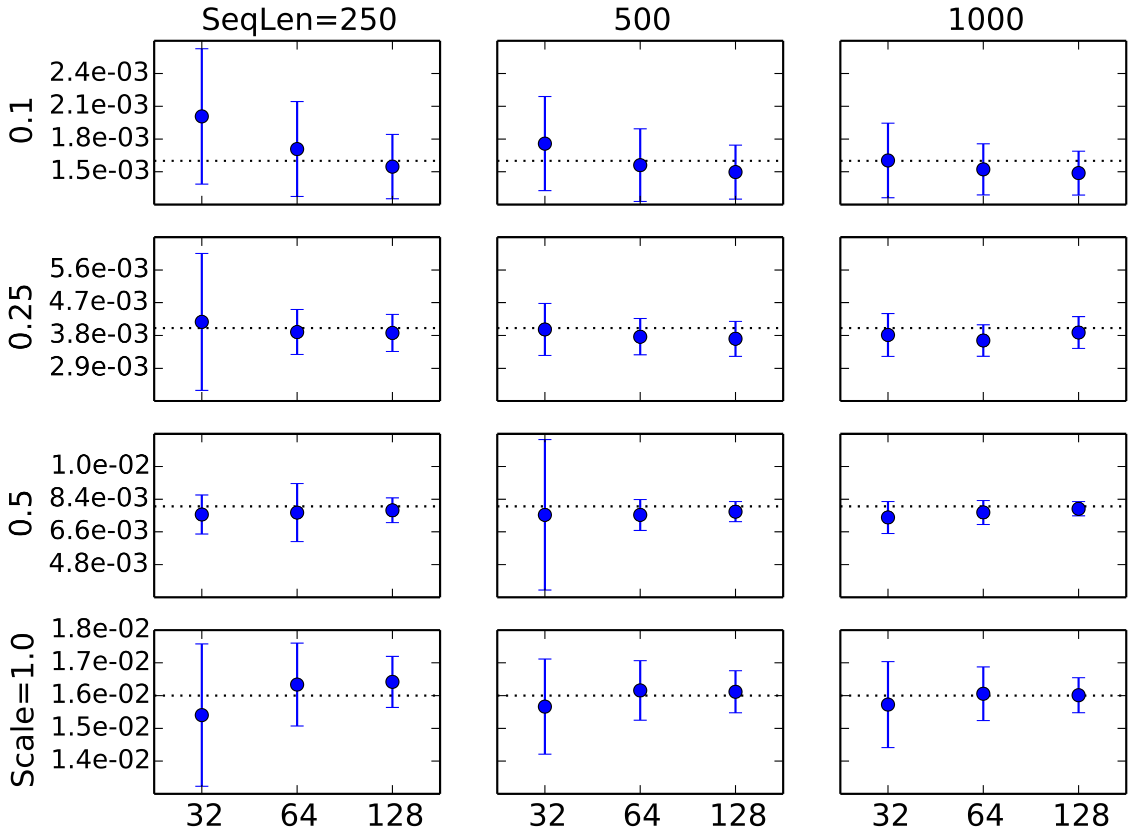


Fig. S26: Divergence time estimates at the root under different values of the scaling parameter s (different rows), sequence lengths (different columns), and numbers of loci (three values within each panel). The dashed line indicates the true value in the model network.

Population size. Fig. S27 shows the estimates obtained for the population mutation rate parameter (one value across all branches of the species network was assumed). The results show very similar trends to those obtained for the divergence time estimates, with the main difference being that the estimates now are very accurate even for the hardest of cases: $s = 0.1$ and sequence length 250, regardless of the number of loci used.

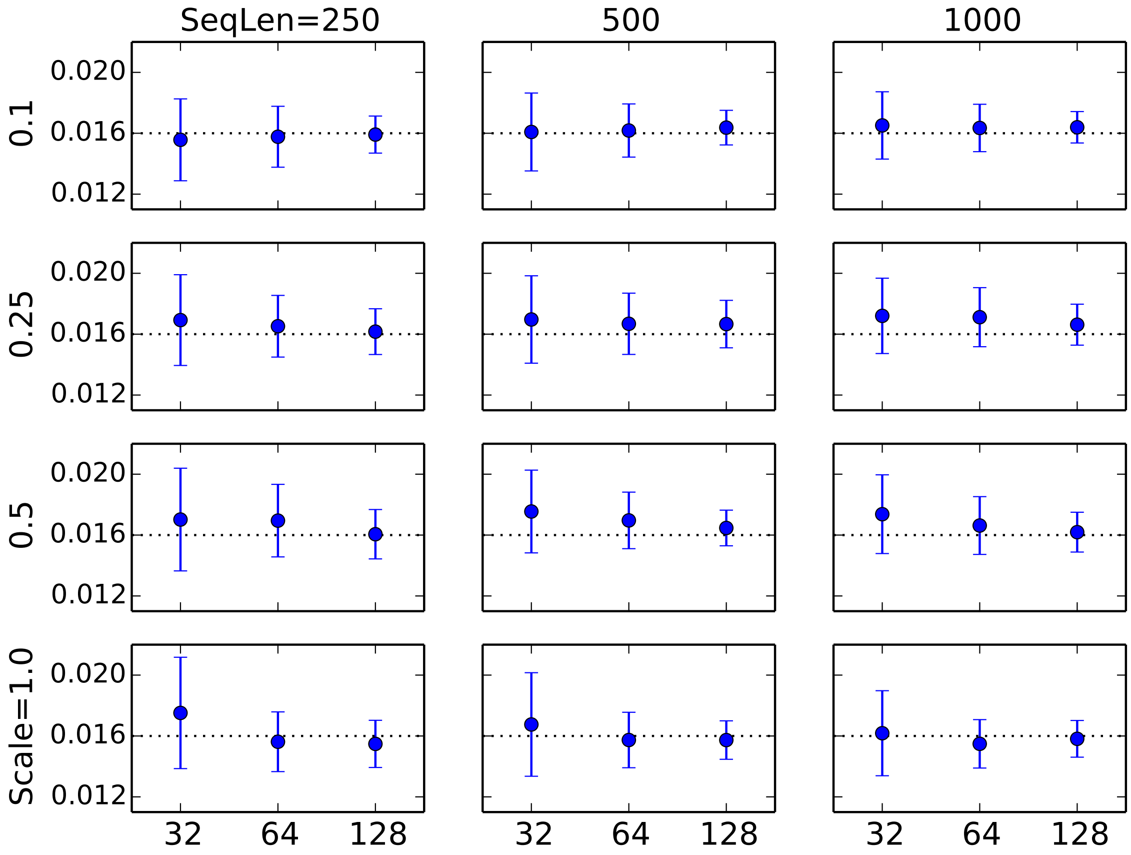


Fig. S27: Population mutation rate estimates under different values of the scaling parameter s (different rows), sequence lengths (different columns), and numbers of loci (three values within each panel). The dashed line indicates the true value in the model network.

The number of reticulations. The results are quite different when it comes to estimating the number of reticulations and the topology of the phylogenetic network itself. Fig. S28 shows the estimates of the number of reticulations under different settings. As the figure clearly shows, under the case of extremely short branches ($s = 0.1$), the method recovers

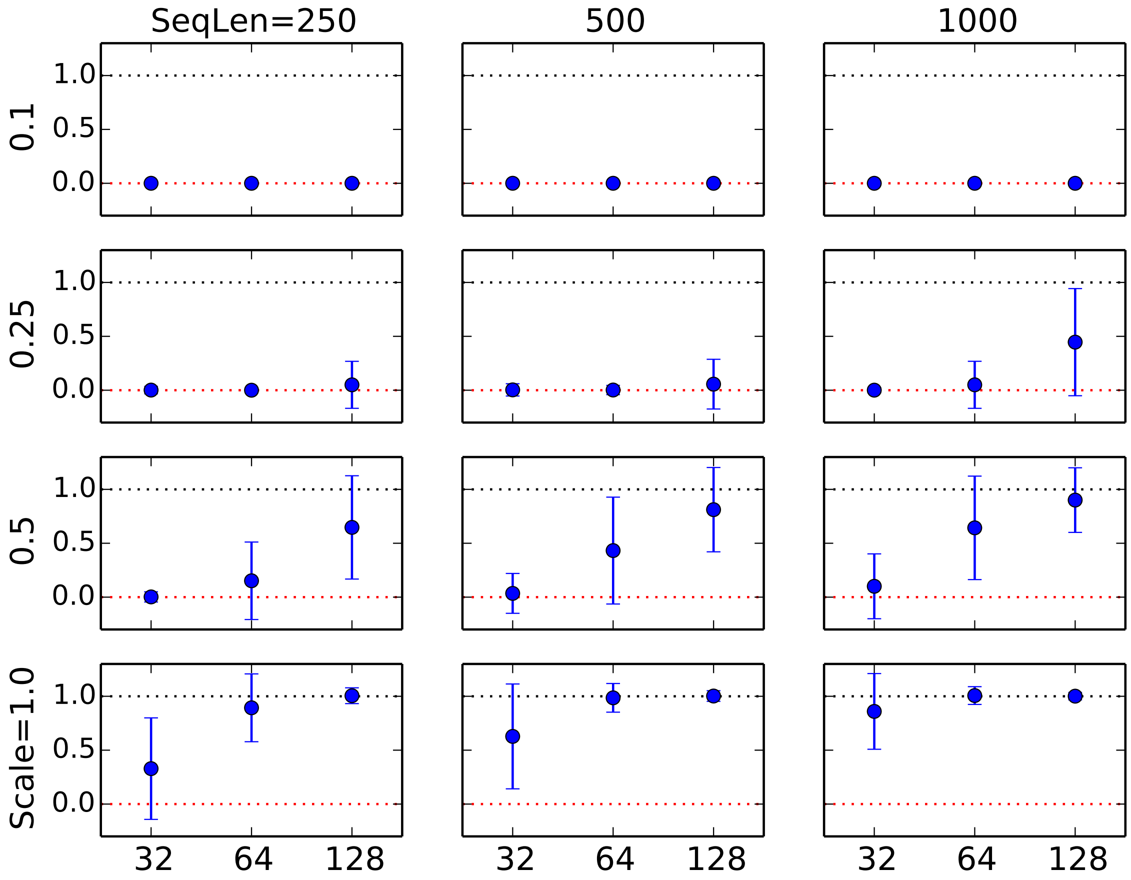


Fig. S28: The number of reticulations inferred under different simulation conditions. The model network has a single reticulation.

a tree; that is, it estimates the number of reticulations to be 0, regardless of the number of loci or sequence length used. Here, the signal is too weak to recover any reticulation. In the case of slightly longer branches ($s = 0.25$), the estimate of the number of reticulations becomes slightly more accurate when the sequences are long and 128 loci are used. Given the observed trend, the method could recover the true number of reticulations if a thousand or so loci are used. In the case of $s = 0.5$, a fast convergence towards the true number is observed as the number of loci increases. It is worth pointing out that, in the case of $s = 0.5$, increasing the number of loci, even when the sequences are very short, is much more advantageous than increasing the sequence lengths of the individual loci. It is also important to note here that in analyzing biological data sets, one cannot use longer sequences without risking violating the recombination-free loci assumption. In the case of $s = 1.0$, the method does very well at estimating the number of reticulations. Finally, observe that the method almost never overestimates the number of reticulations on these data sets.

The topology of the inferred species phylogeny. In assessing the quality of the estimated network topology itself, we analyzed the recovered networks in two ways. First, we compared the inferred network to the true network using a topological dissimilarity measure (Nakhleh 12). Second, when the method infers a tree, rather than a network, we compared the tree to the "backbone tree" of the true network (the tree resulting from removing the arrow in Fig. S25) using the Robinson-Foulds metric (Robinson and Foulds 14). The latter comparison allows us to answer the question: When the method estimates the species phylogeny to be a tree, how does this tree compare to the backbone tree of the true network? Fig. S29 shows the results. The results in terms of the topological difference between the inferred and true networks parallel those that we discussed above in terms of the estimates of the number of reticulations: Poor accuracy and no sign of convergence to the true network in cases of very small scaling parameter values, and very good accuracy and fast convergence to accurate estimates in cases of larger scaling parameter values. However, the topological difference between the inferred trees (in the cases where trees were inferred) and the backbone tree reveal an important insight: When the method fails to recover the true network, it does a very good job at recovering the backbone tree of the true network.

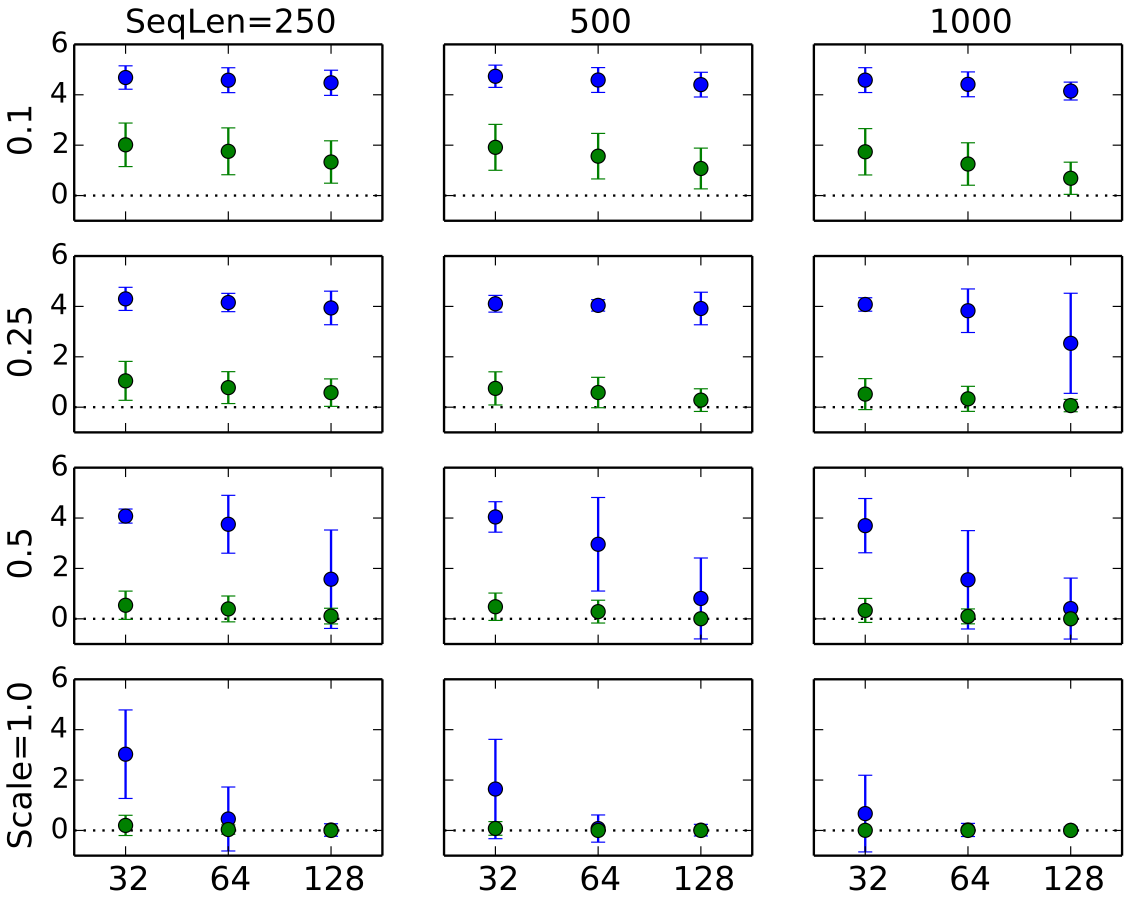


Fig. S29: The topological difference between the true and inferred networks in blue and the Robinson-Foulds distance between the inferred tree (if a network is inferred, this case is not included) and the backbone tree of the true network.

4 Simulation with Intermixture/Gene flow patterns

Intermixture and gene flow provide two different abstract models of reticulation. Furthermore, the program ms (Hudson 8) allows for generating data under models. While the MSNC is based on an intermixture model, we study here how it performs on data simulated under a gene flow model. We set up the experiment so that data are generated under the same phylogenetic networks and their parameters, yet under the scenarios of intermixture and gene flow separately. Furthermore, in this part, we assess the performance when multiple reticulation events occur between the same pair of species—a very realistic scenario in practice.

4.1 Simulations settings

Model species phylogenies. Fig. S30 shows the six phylogenetic networks we used to generate data.

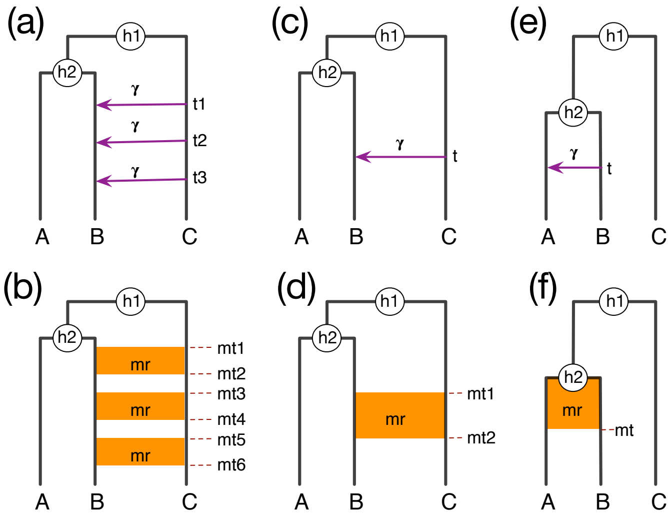


Fig. S30: True phylogenetic histories with intermixture and gene flow models. Recurrent reticulations between non-sister taxa (a,b), a single reticulation between non-sister taxa (c,d), and a single reticulation between sister taxa (e,f) is captured under both the intermixture model (top) and gene flow 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.

We set $h_1 = 9$, $h_2 = 6$. For the intermixture model (Fig. S30(a)), we set $t_2 = 3$, and varied (t_1, t_3) to take on the values $(4, 2)$, $(5, 1)$, and $(6, 0)$ so that the elapsed time, denoted by Δt, between subsequent reticulation events is 1, 2, or 3. For the gene flow model (Fig. S30(b)), we set $(mt_1, \ldots, mt_6)$ to $(6, 4, 4, 2, 2, 0)$ and $(6, 5, 3.5, 2.5, 1, 0)$, so that the duration of each gene flow epoch, denoted by Δmt, is either 1 or 2. Notice that, under our setting, the time elapsed between two consecutive gene flow epochs is smaller for $\Delta mt = 2$ than for $\Delta mt = 1$.

We set the inheritance probability γ and the migration rate mr each to 0.20.

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.

S30**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

S30**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

S30**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

S30**A** $|mt| = 1$: 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

S30**A** $|mt| = 2$: 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

S30**B** $t = 1$: 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

S30**B** $t = 0$: 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

S30**B** $mt_2 = 1$: 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

S30**B** $mt_2 = 0$: 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

S30**C** $t = 1$: 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

S30**C** $t = 0$: 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

S30**C** $mt = 1$: 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

S30**C** $mt = 0$: 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 13) 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 -mHKY -l1000 -s0.01
```

4.2 MCMC settings

For each data set, we ran an MCMC chain of 8×10^6 iterations with 3×10^6 burn-in. One sample was collected from every 5,000 iterations, resulting in a total of 1,000 collected samples.

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

4.3 Recurrent reticulations

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. S30(a) and Fig. S30(b). As the results show, the method performs very well in terms of estimating

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 in the model of Fig. S30(a) and Δmt in the model of Fig. S30(b). The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2 = 0.01$.

Case	θ	h_1	h_2	#reti
$\Delta t = 1$	$2.2 \pm 0.2e^{-2}$	8.9 ± 0.1	5.9 ± 0.1	1.2 ± 0.4
$\Delta t = 2$	$2.2 \pm 0.2e^{-2}$	8.9 ± 0.1	5.9 ± 0.1	2.0 ± 0.0
$\Delta t = 3$	$2.1 \pm 0.3e^{-2}$	9.0 ± 0.1	6.0 ± 0.1	2.6 ± 0.5
$\Delta mt = 1$	$2.3 \pm 0.3e^{-2}$	8.9 ± 0.1	6.0 ± 0.1	2.1 ± 0.3
$\Delta mt = 2$	$2.3 \pm 0.3e^{-2}$	8.9 ± 0.1	6.9 ± 0.1	2.0 ± 0.1

the divergence times and population mutation rates, regardless of whether the data was generated under an intermixture model or a gene flow model. Furthermore, for these two parameters, the estimates are stable while varying the elapsed times between consecutive reticulation events.

As for the estimated number of reticulations, it becomes more accurate as the elapsed times between consecutive reticulations is larger. To better understand the factors that affect the detectability of reticulations, we plotted histograms of the true and estimated coalescent times of the most recent common ancestor (MRCA) of alleles from B and C in Fig. S31. As Fig. S30(a) and Fig. S30(b) show, the coalescent times of alleles from B and C would form a mixture of four distributions: three due to the three reticulation events, and one above the root of the phylogenetic network.

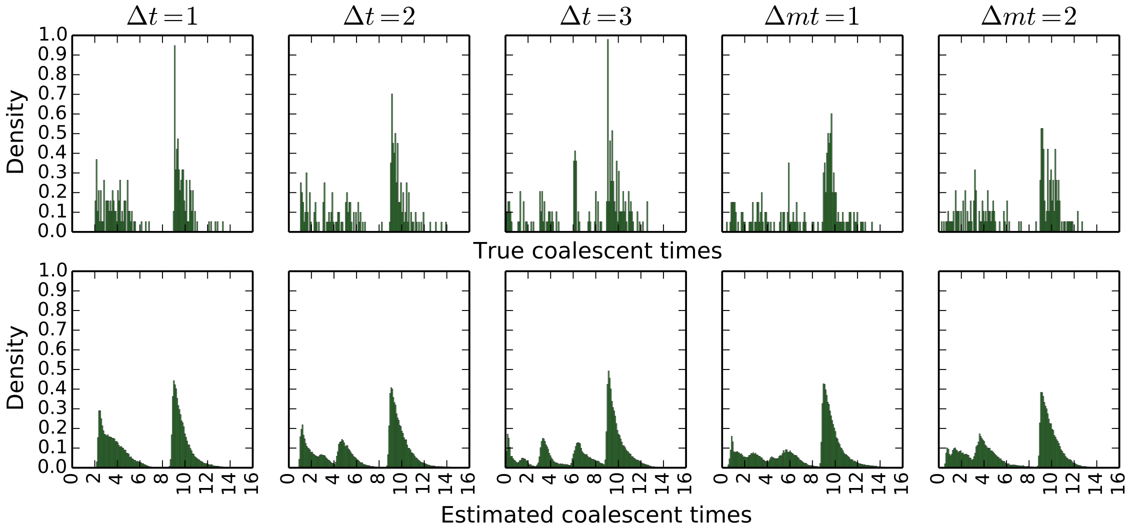


Fig. S31: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on data generated under the models of Fig. S30(a) and Fig. S30(b).

As the left three columns of panels in the figure show, under an intermixture model, as Δt increases, the signal for a mixture of four distributions of (A, B) coalescent times becomes much stronger, thus pointing to three reticulations in addition to the coalescent events above the root of the phylogeny. This is why, under the intermixture model, the method's performance in terms of the estimated number of reticulations improves as Δt

increases. However, this is not the case under the gene flow model (the right two columns of panels in the figure). It is important to note that for $\Delta mt = 2$, the three gene flow epochs actually form one continuous epoch of gene flow from time mt_1 to mt_6.

Fig. S32 shows results similar to those reported in Fig. S31, with the only difference being that these are the coalescent times from all $4,000$ loci generated from the 20 data sets of 200 loci each. Effectively, this is the signal in a data set of $4,000$ independent loci.

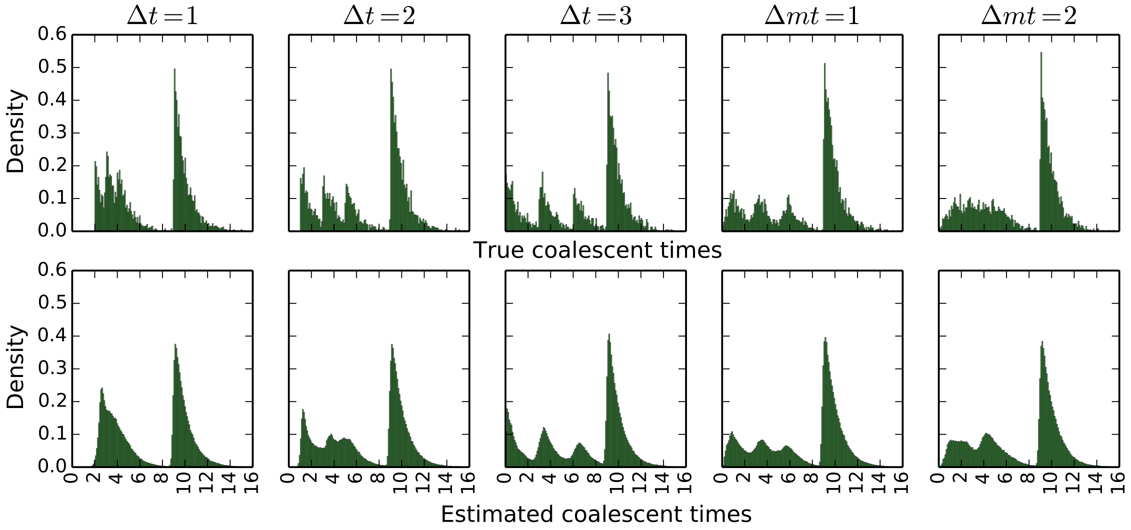


Fig. S32: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on $4,000$ loci generated under the models of Fig. S30(a) and Fig. S30(b).

Clearly, the signal is much stronger than in data sets of 200 loci, and all reticualtions would be recoverable under the intermixture model for $\Delta t = 2, 3$ and for the gene flow model for $\Delta mt = 1$.

4.3.1 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. S30(a).

Table S4 shows the population mutation rates, divergence times, and numbers of reticulations estimated by our method on data generated under the models of Fig. S30(a) with varying number of individuals sampled from species B. As the results show, the method

Table S4: 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. S30(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

performs very well in terms of estimating the divergence times and population mutation rates.

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 in Table S3 to 1.8 ± 0.4). However, increase the number of individuals from 3 to 5 doesn't 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.

4.4 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. S30(c) and Fig. S30(d), set $h_1 = 2.5$, $h_2 = 1.5$, and $mt_1 = h_2$, and varied $t, mt_2 \in \{1, 0\}$.

As the results in Table S5 demonstrate, our method estimated the population mutation rate θ, the divergence times h_1 and h_2, and the inheritance probability/migration rate very accurately under all cases. A single reticulation was detected for all cases of intermixture and gene flow. We plotted the histograms of the true and estimated coalescent times of

Table S5: Estimated population mutation rates (θ), divergence times (h_1 and h_2), inheritance/migration rates, and numbers of reticulations (#reti) as a function of varying t in the model of Fig. S30(c) and mt_2 in the model of Fig. S30(d). The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2 = 0.01$.

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

the MRCA of alleles from B and C in Fig. S33. As the figure shows, the distributions of estimated coalescent times match the distributions of true coalescent times very well.

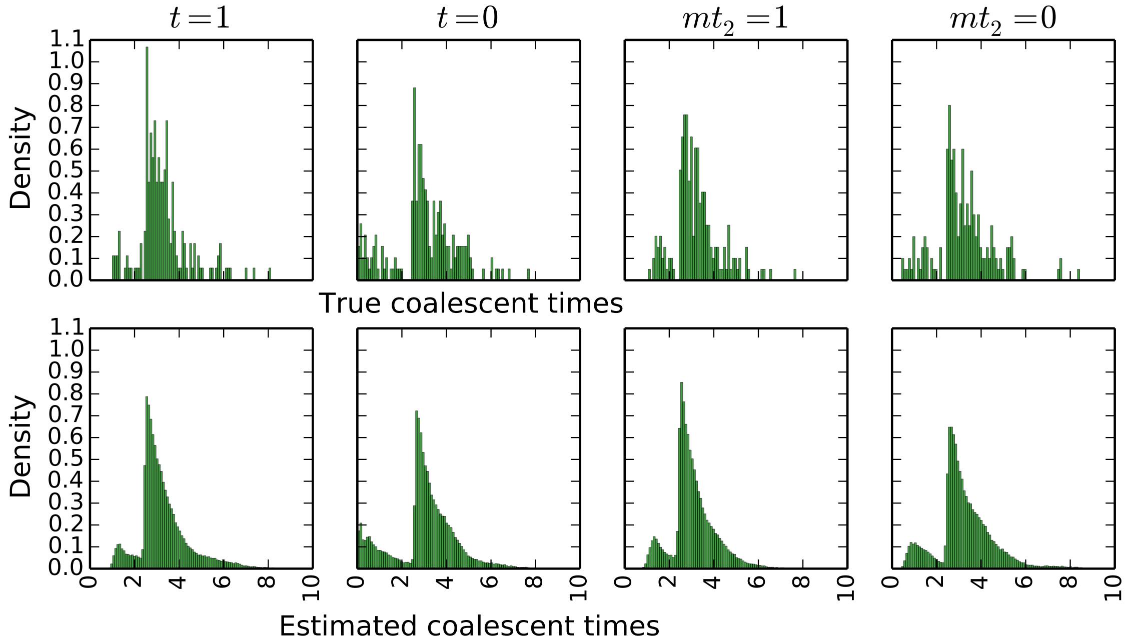


Fig. S33: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on data generated under the models of Fig. S30(c) and Fig. S30(d).

Fig. S34 shows results similar to those reported in Fig. S33, with the only difference being that these are the coalescent times from all $4,000$ loci generated from the 20 data sets of 200 loci each. Effectively, this is the signal in a data set of $4,000$ independent loci.

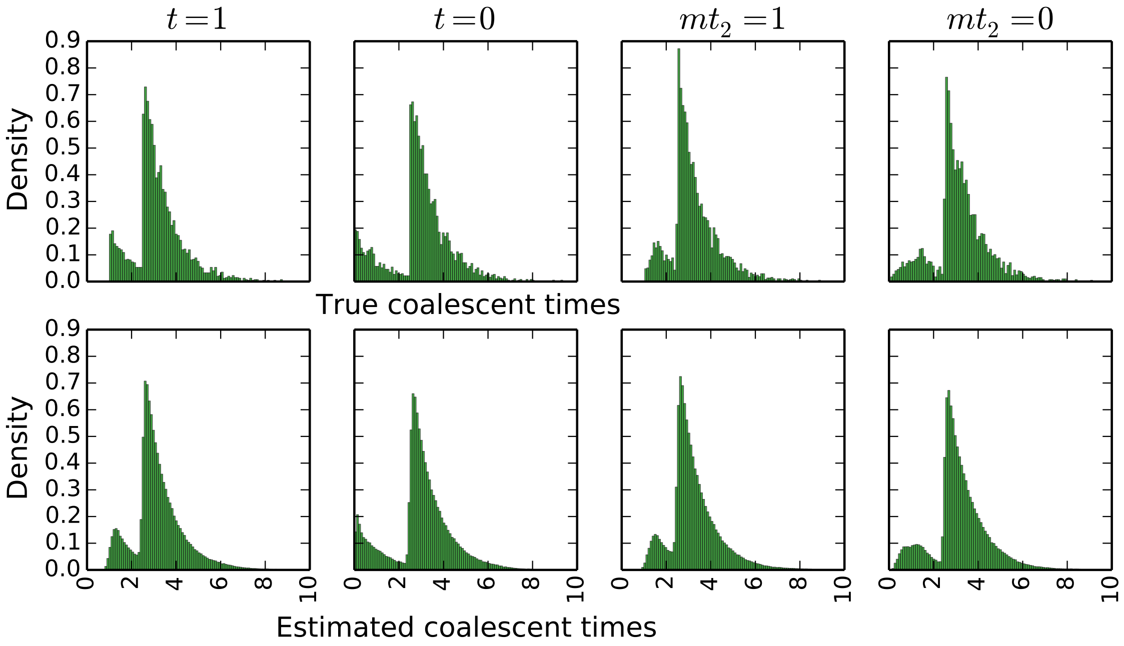


Fig. S34: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on $4,000$ loci generated under the models of Fig. S30(c) and Fig. S30(d).

Clearly, the signal is much stronger than in data sets of 200 loci.

4.5 Isolation-migration between sister species

we assessed the performance of our method on cases where the reticulation event involves sister taxa. Fig. S30(e) and Fig. S30(f) show the cases we considered, with setting $h_1 = 2.5$ and $h_2 = 1.5$, and varying $t, mt \in \{1, 0\}$.

As the results in Table S6 demonstrate, our method obtained very accurate estimates of the various parameters under $t = 0$ and $mt = 0$. Under the cases of intermixture with $t = 1$ and gene flow with $mt = 1$, our method did not detect the reticulation, which resulted in an underestimation of h_2. In the case of $mt = 0$, the migration rate was severely under-

Table S6: Estimated population mutation rates (θ), divergence times (h_1 and h_2), inheritance/migration rates, and numbers of reticulations (#reti) as a function of varying t in the model of Fig. S30(e) and mt in the model of Fig. S30(f). The divergence times were estimated in units of expected number of mutations per site and are reported in coalescent units by dividing by $\theta/2 = 0.01$.

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

estimated, most likely due to the short time interval between the migration and divergence events between A and B.

We plotted the histograms of the true and estimated coalescent times of the MRCA of alleles from A and B in Fig. S35. When $t = 1$ and $mt = 1$, the signal of reticulation is very low, which explains the failure of our method to detect it. In the cases of $t = 0$ and $mt = 0$, the distributions of estimated coalescent times match those of true coalescent times very well.

Fig. S36 shows results similar to those reported in Fig. S35, with the only difference being that these are the coalescent times from all $4,000$ loci generated from the 20 data sets of 200 loci each. Effectively, this is the signal in a data set of $4,000$ independent loci. Clearly, the signal is much stronger than in data sets of 200 loci.

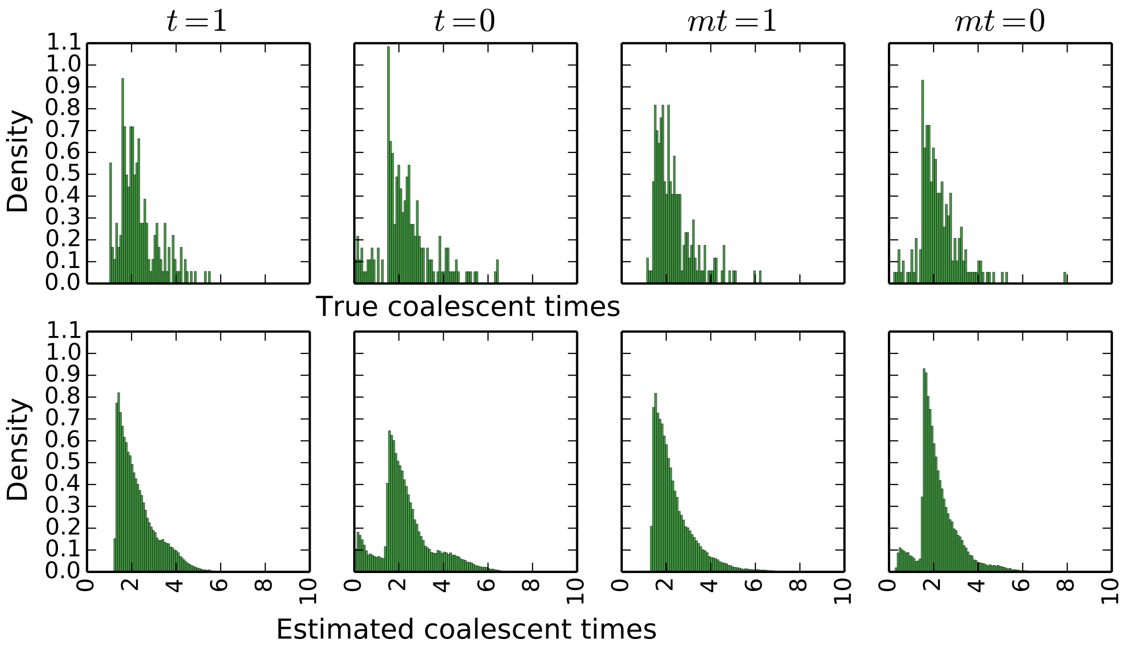


Fig. S35: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from A and B on data generated under the models of Fig. S30(e) and Fig. S30(f).

5 Analysis of a bread wheat (*Triticum aestivum*) data set

Marcussen *et al.* (11) investigated ancient hybridization among the ancestral genomes of bread wheat by performing

1. parsimonious inference of hybridizations in the presence of ILS (21) using PhyloNet (18). 2269 gene trees were constructed from three subgenomes of *T. aestivum* TaA (A subgenome), TaB (B subgenome) and TaD (D subgenome). Using these gene trees, several 1-reticulation and 2-reticulation species phylogenies were inferred, shown in Table 4S in (11).

2. gene tree analyses using BEAST. Gene trees and coalescent times were inferred from 275 sequence alignments of hexaploid bread wheat subgenomes TaA, TaB and TaD, and five diploid relatives Tm (*T. monococcum*), Tu (*T. urartu*), Ash (*Ae. sharonensis*), Asp (*Ae. speltoides*) and At (*Ae. tauschii*).

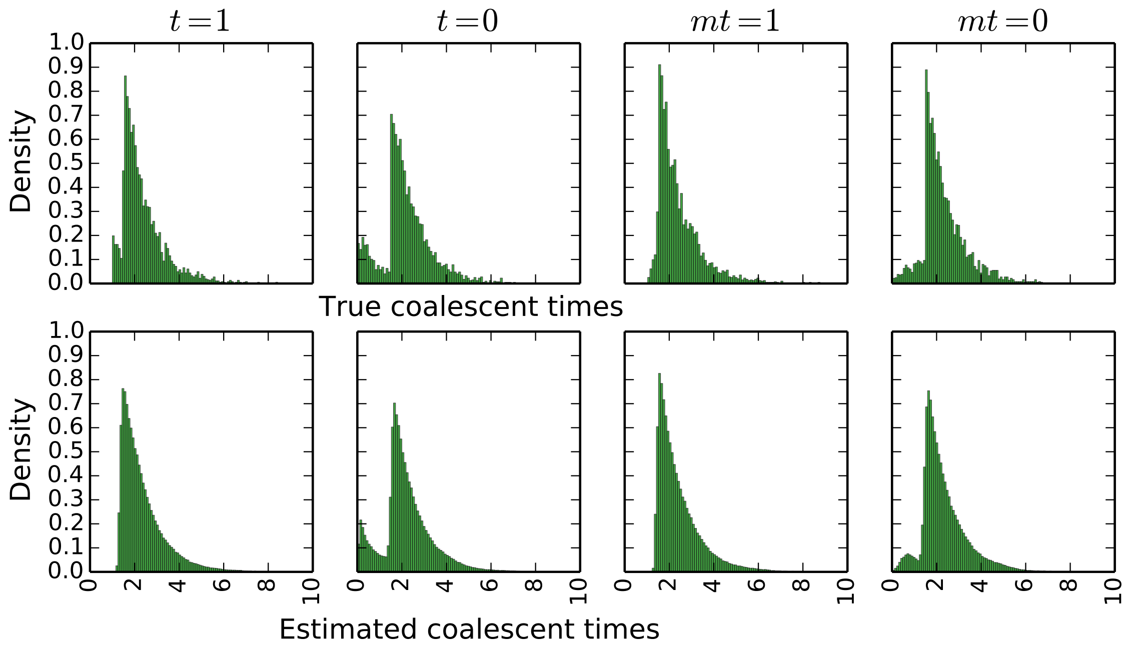


Fig. S36: Histograms of the true (top) and estimated (bottom) coalescent times (in coalescent units) of the MRCA of alleles from B and C on $4,000$ loci generated under the models of Fig. S30(e) and Fig. S30(f).

Given a 1-reticulation phylogenetic network topology from (1) and gene tree coalescent times from (2), the genome divergence times were further estimated from a Bayesian hierarchical model (17). This Bayesian hierarchical model only takes a network with at most one introgression event; thus, 2-reticulation networks were not analyzed. Marcussen *et al.* proposed a plausible evolutionary history of bread wheat based on the estimated divergence times of the selected 1-reticulation network.

Here we reanalyzed the 275 gene data set using our newly developed method, which provides a systematic way to infer the species network, gene trees and the genome divergence times from genome sequences simultaneously.

5.1 Data preprocessing

We downloaded the sequence alignments of 275 genes from Dryad Digital Repository (doi:10.5061/dryad.f6c34). Each alignment is composed of genes from hexaploid bread

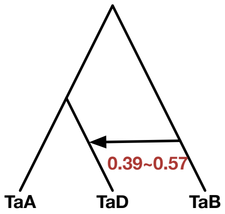


Fig. S37: **One of the species phylogenies inferred from a data set of 2269 gene trees using parsimonious inference in PhyloNet (18).** Marcussen *et al.* performed divergence time analyses and finally proposed a plausible evolutionary history of bread wheat based on this network (11).

wheat subgenomes TaA, TaB and TaD, and diploid relatives Tm, Tu, Ash, Asp and At. We created two data sets of 68 loci (25% of the full data set) and 137 loci (50% of the full data set) and analyzed them.

5.2 MCMC settings

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

5.3 Results for the 68-locus data set

For the 68-locus data set, we ran an MCMC chain with 6×10^7 iterations with 2.5×10^7 burn-in. One sample was collected from every 5,000 iterations.

- 95% credible set of the phylogenetic network topologies. The 95% credible set contains only one topology, see Fig. S38. The topology is identical to the one reported in (11) (inferred from the data set of 2269 gene trees using PhyloNet).

- Population size. The population size estimates are consistent among samples, with a mean values and the standard deviation of 51.48 ± 3.83 E-4.

- Convergence and mixing. The trace plot, shown in Fig. S39, indicates good convergence and mixing.

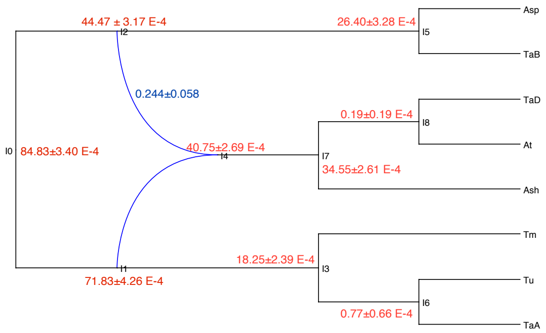


Fig. S38: **The phylogenetic network in the 95% credible set from the 68-locus wheat data set using our method.** The divergence times are labeled in red, and the inheritance probability is marked in blue.

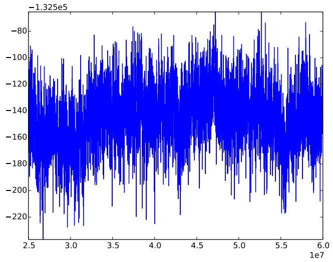


Fig. S39: **Trace plot of the MCMC chain using our method on the 68-locus wheat data set.**

We also ran an MCMC chain with 2.5×10^7 iterations with 5×10^6 burn-in using $*$BEAST. One sample was collected from every 5,000 iterations. The 95% credible set contains only one topology (Fig. S40) that can be embedded by the network inferred from our method in Fig. S38. We can see that

- the divergence times of Asp-TaB, At-TaD, TaA-Tu, Tm-(TaA, Tu), and Ash-(At, TaD) are similar from both methods.
- the divergence times of **A-B**, **A-D**, and **B-D** from $*$BEAST are much smaller than the ones inferred by our method. The reason is that $*$BEAST assumes a species tree, forcing the times to satisfy the temporal constraints of gene trees, resulting in divergence time underestimation.

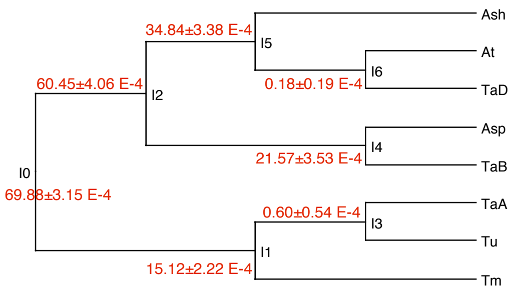


Fig. S40: **The species tree in the 95% credible set from the 68-locus wheat data set using $*$BEAST.** The divergence times are marked in red.

Assuming a mutation rate of 1×10^{-9} per-site per-generation and 1 generation per year, a common ancestor of **A**, **B**, **D** started differentiation around 8.5 Mya into the **A-D** and **B** genome lineages. **A-D** began speciation at $\sim$7 Mya into **A** and **D** lineages. The first hybridization occurred $\sim$4.5 Mya between the **B** and **D** genome lineages and led to the origin of the hybrid **D** genome lineage.

5.4 Results for 137-locus data set

We ran 3 MCMC chains, each for 1×10^8 iterations and with 5×10^7 burn-in. One sample was collected from every 5,000 iterations.

- 95% credible sets. The 95% credible sets of topologies contain five species networks, shown in Fig. S41. The topologies are identical to the ones reported in Table S4 of (11) (inferred from the data set of 2269 gene trees using PhyloNet).

- Population size. The population size estimates are consistent across MCMC chains, with mean values and the standard deviations of 42.33 ± 2.54, 43.33 ± 2.23, and 42.92 ± 2.39 E-4, respectively.

- Convergence and mixing. Trace plots are shown in Fig. S42 indicating good convergence and mixing.

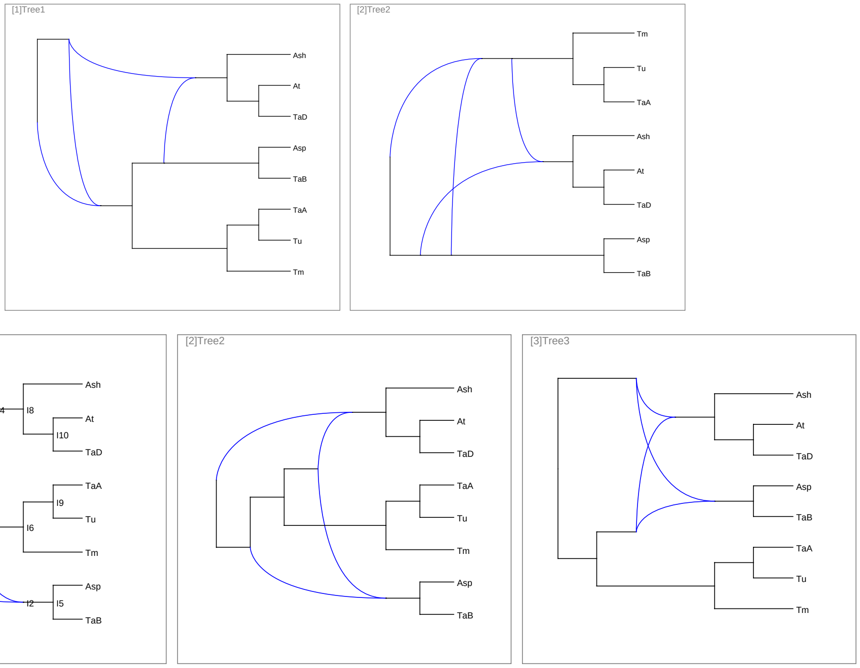


Fig. S41: **The phylogenetic network topologies in the 95% credible sets from the 137-locus wheat data set using our method.** The proportions are 33%, 33%, 20%, 7%, and 5% respectively. The posterior values are -256083.4 ± 31.1, -256082.8 ± 32.7, -256085.6 ± 32.7, -256104.0 ± 35.1, and -256106.7 ± 32.6, respectively.

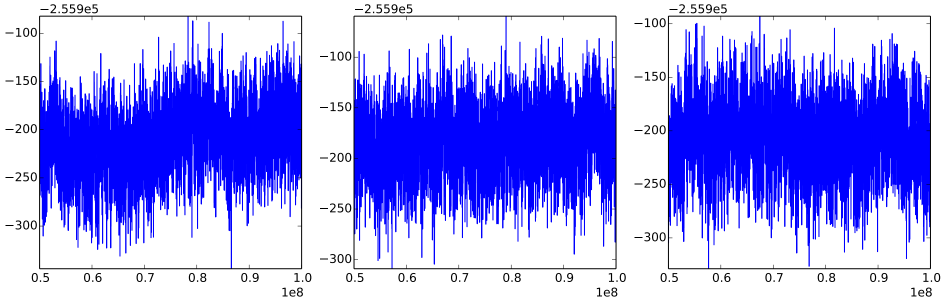


Fig. S42: **Trace plots of the MCMC chains given the 137-locus wheat data set using our method.**

6 Analysis of a yeast data set

Rokas et al. (15) 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. S43 (left). Edwards *et al.* (4) reported two main gene tree / species tree topologies sampled from BEST, a multispecies coalescent Bayesian inference method, as shown in Fig. S43. Although the two species trees support $(Sklu, (Scas, ...))$, other gene tree topologies (Fig. S44) sampled from BEST indicate the weak phylogenetic signal for resolving the relationship of Sklu and Scas to the five other species.

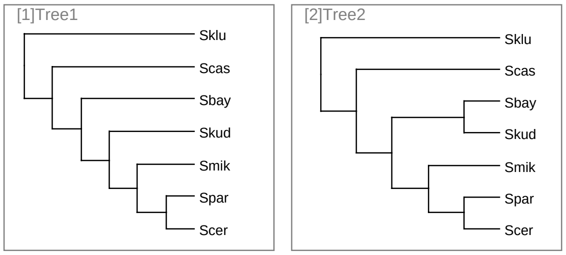


Fig. S43: **The species trees of seven Saccharomyces species.** (Left) The topology inferred from concatenation method (15) and the main topology sampled by BEST (4). (Right) The topology sampled by BEST with the second highest proportion (4).

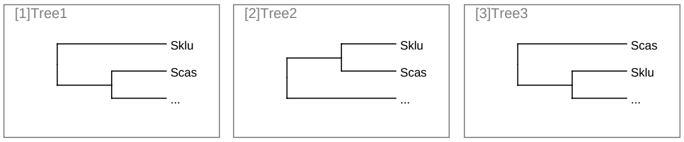


Fig. S44: **Relationships of Sklu and Scas in several gene tree topologies of seven Saccharomyces species.**

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. S43(left) and Fig. S43(right).

Yu *et al.* (22) 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 $Skud \to Sbay$ with inheritance probability of 0.38 (see Fig. 8 in (22)).

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

6.1 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.

6.2 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).

We quantified the phylogenetic signal in a locus based on the number of resolved branches in the majority-rule consensus of bootstrap trees. For each locus, we inferred 100 bootstrap trees using RAxML (16) and computed bootstrap values for the tree of that locus. Finally, we contracted every branch with bootstrap support smaller than 70. Table S7 shows the number of gene trees (total of 106) with each of the possible numbers of internal branches resulting from the bootstrap procedure.

Table S7: **The numbers of of trees with indicated numbers of internal branch lengths on the seven Saccharomyces species data set.** Gene trees with 0 internal branches are star phylogenies. Gene trees with 5 internal branches are fully resolved.

number of internal branches	0	1	2	3	4	5
number of trees	2	15	22	39	28	0

However, if we only focus on the five species (Scer, Spar, Smik, Skud, Sbay), the phylogenetic signal become stronger, as shown in Table S8.

Table S8: **The numbers of of trees with indicated numbers of internal branch lengths on the five Saccharomyces species data set (Scer, Spar, Smik, Skud, Sbay).** Gene trees with 0 internal branches are star phylogenies. Gene trees with 3 internal branches are fully resolved.

number of internal branches	0	1	2	3
number of trees	0	7	99	0

6.3 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. S45.

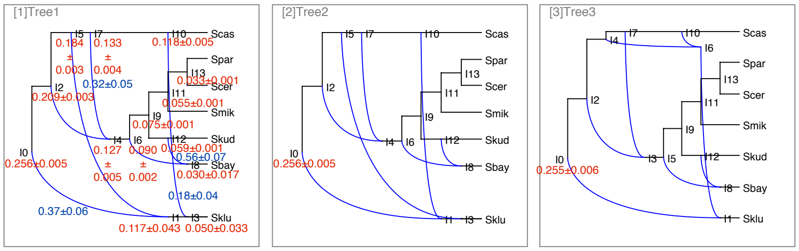


Fig. S45: **The three main phylogenetic networks in the 95% credible sets from the yeast data set using our method.** The divergence times are labeled in red, and the inheritance probabilities are marked in blue.

- Convergence and mixing. The trace plots shown in Fig. S46 indicate good convergence and mixing. The states across MC3 chains are slightly inconsistent in terms 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. S45.

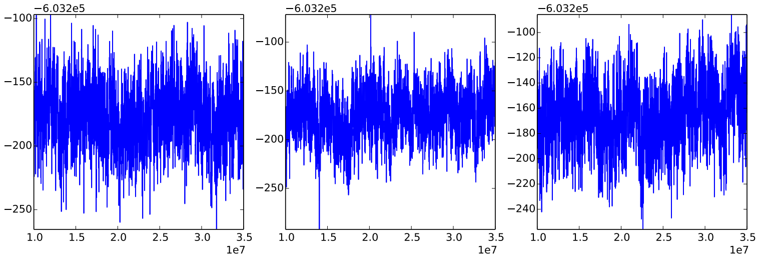


Fig. S46: **Trace plots of MC3 chains using our method on the yeast data set.**

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. S47, we can see the phylogenetic signals among Scas, Sklu and the other 5 species are low.

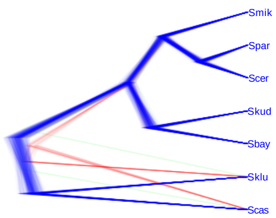


Fig. S47: **The densiTree plot of the species trees sampled from $*$BEAST given the yeast data set.**

The 95% credible set (Fig. S48) 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.

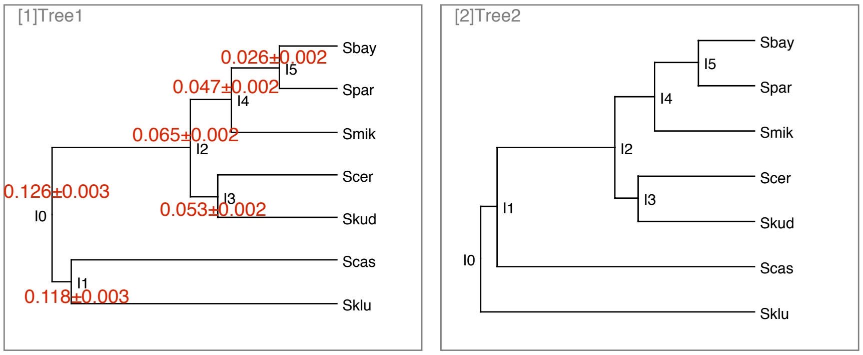


Fig. S48: **The two main species trees in the 95% credible set from the yeast data set using $*$BEAST.** The proportions for Tree 1 and Tree 2 are 78.4% and 16.9%, respectively. The divergence times are marked in red.

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. S49. The ranges of the divergence times obtained by $*$BEAST and our method are similar.

6.4 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

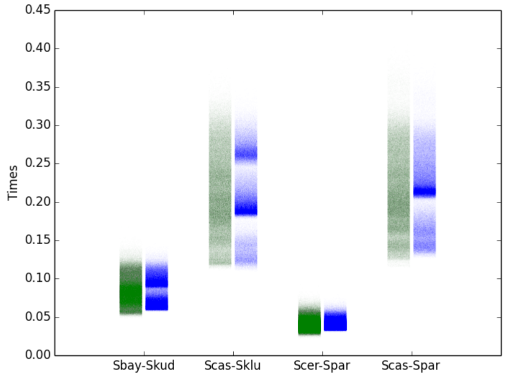


Fig. S49: **The divergence times of the MRCAs of (Sbay,Skub), (Scas,Sklu), (Scer,Spar), and (Scas,Spar) from estimated gene tree samples inferred by ∗BEAST (green), and our method (blue).** The input is the full yeast data set.

iterations.

- 95% credible sets of the phylogenetic network topologies. The 95% credible sets contain only one topology, as shown in Fig. S50. The topology is identical to the

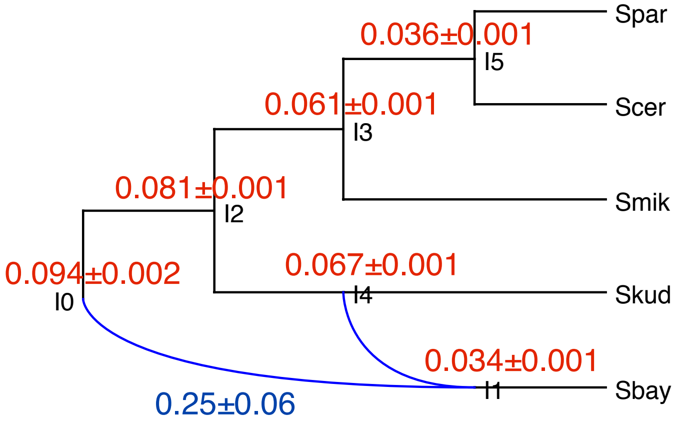


Fig. S50: **The phylogenetic network in the 95% credible sets from the data set of 106 loci from five Saccharomyces species using our method.** The divergence times are marked in red, and the inheritance probability is marked in blue.

one reported in (22), which reconciles the two main species tree topologies reported by (4) in Fig. S43). The inheritance probability of the horizontal reticulate edge is 0.75 ± 0.06, which differs from the value of 0.36 reported in (22). The divergence times are similar to the ones inferred from the full data set of seven species in Fig. S45.

- Convergence and mixing. The trace plots shown in Fig. S51 indicate good convergence and mixing.

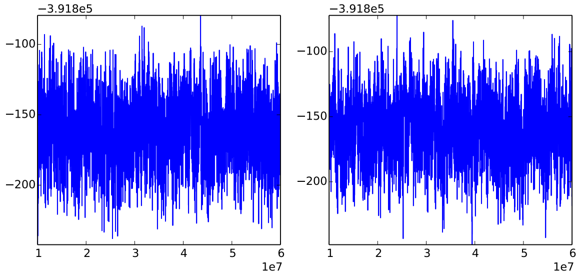


Fig. S51: **Trace plots of MC3 chains using our method given the data set of 106 loci from five Saccharomyces species.**

6.5 Results for the 28-locus data set with strong phylogenetic signals

For the data set of 28 loci where the number of internal branches in the bootstrapped tree of each locus is 3, 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 four topologies, as shown in Fig. S52. All the topologies demonstrate the weak phylogenetic signals involving Sklu, Scas and the other five species reported by (4) and represented in Fig. S44.

- Convergence and mixing. The trace plots shown in Fig. S53 indicate good convergence and mixing.

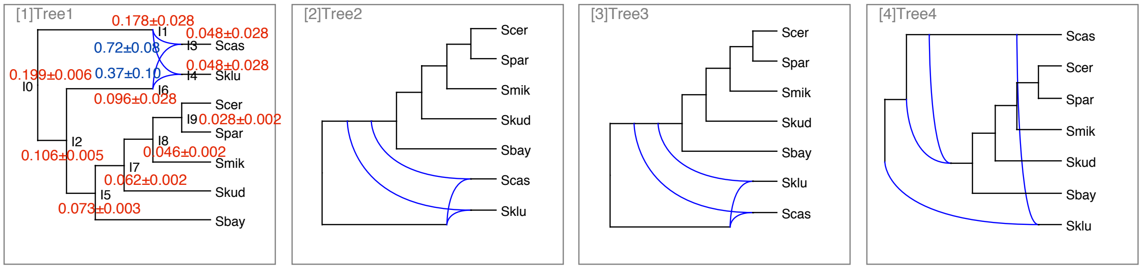


Fig. S52: **The phylogenetic networks in the 95% credible sets from the data set of 28 loci from seven Saccharomyces species using our method.** The proportions of the topologies are 77.9%, 8.4%, 5.4%, and 4.0%, respectively. The posterior values are -190654.0 ± 11.5, -190656.1 ± 11.9, -190656.9 ± 11.0, and -190656.3 ± 11.8, respectively. The divergence times are marked in red, and the inheritance probabilities are marked in blue.

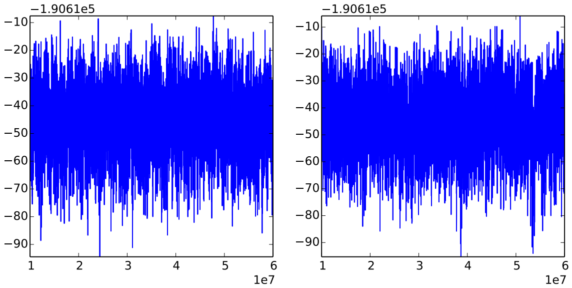


Fig. S53: **Trace plots of MC3 chains using our method given the data set of 28 loci from seven Saccharomyces species.**

7 Analysis of a mosquito (*An. gambiae* complex) data set

Here we revisited the data set of (5) using our newly developed method.

7.1 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 $p_\psi(x) = 1/x$). The divergence and migration times (in million years, Mya) reported below all assume a substitution rate of 1.1×10^{-9} per-site per-generation and 10 generations per year (based on (5)).

7.2 Data preprocessing

We downloaded the MAF genome alignment of high depth field samples from Dryad (doi:10.5061/dryad.f4114). The species we included in our analysis were *An. gambiae* (G), *An. coluzzii* (C), *An. arabiensis* (A), *An. quadriannulatus* (Q), *An. merus* (R) and *An. melas* (L). *An. christyi* served as the outgroup for gene tree reconstruction and rooting.

Fontaine et al. built one gene tree on every 50-kb genomic window and reported the topology frequencies of the maximum likelihood gene trees in Fig. 2 in (5). However, there are several potential problems:

- 50-kb is very long, which may violate the assumption that each locus is recombination-free.
- Using the optimal ML gene tree topology estimated to represent a 50-kb region does not account for gene tree uncertainty.

7.2.1 Regions with longer sequences exhibit stronger phylogenetic signals

We investigated the effect of sequence length on gene tree reconstruction. We randomly sampled 185 10-kb chunks from the 2R chromosome. For each chunk, we inferred 100 bootstrap trees, computed the majority-rule consensus tree (16), and finally analyzed the number of chunks that have gene trees with 0 to 4 internal branches. We also split the 10-kb chunks into ten 1-kb chunks (1850 chunks in total) and repeated the same steps. Table S9 shows that

Table S9: **The number of loci from the 2R chromosome that have majority-rule consensus trees with 0 to 4 internal branches.** Gene trees with 0 internal branches are star phylogenies. Gene trees with 4 internal branches are fully resolved.

data set	consensus threshold	0	1	2	3	4
185 regions	50	0	1	24	160 (86.5%)	0
10-kb	70	3	15	67	100 (54.1%)	0
1850 regions	50	136	302	587	825 (45.4%)	0
1-kb	70	530	618	501	201 (11.1%)	0

- There is no fully resolved tree by setting consensus threshold to 50 or 70. We found all the roots of consensus trees are non-binary.

- Using a higher consensus threshold decreases the number of internal branches.

- Comparing the 1850 1-kb regions data set to the 185 10-kb regions data set, the proportion of the regions that have three internal branches drops significantly, which means the bootstrap support, or the signal from data, becomes lower. We can also see from Fig. S54 and S55 that the proportion of the top 5 out of 286 topologies in 185 10-kb regions data set is around 47.9%, while in 1850 1-kb regions data set, the proportion of top 5 out of 796 topologies is around 9.1%.

We randomly chose one 10-kb region that has three internal branches, built the maximum likelihood trees of this region using RAxML (16), and compared the consensus trees (threshold 70) with the maximum likelihood trees (shown in Fig. S56).

- The consensus tree shows rooting issue among R, L and (Q,(C,(G,A))), but the ML tree groups (R,L) together.

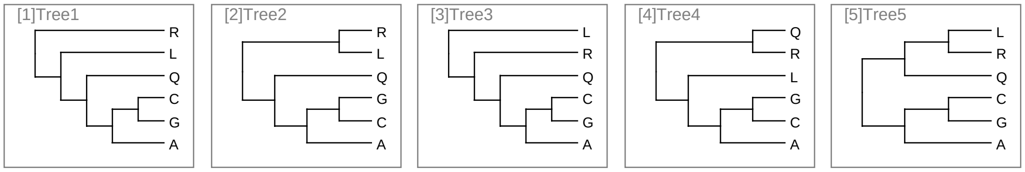


Fig. S54: **The top 5 out of 286 gene tree topologies from the 185 10-kb regions in the 2R chromosome.** The topology frequencies are 16.4%, 14.5%, 7.4%, 5.4%, and 3.9%, respectively.

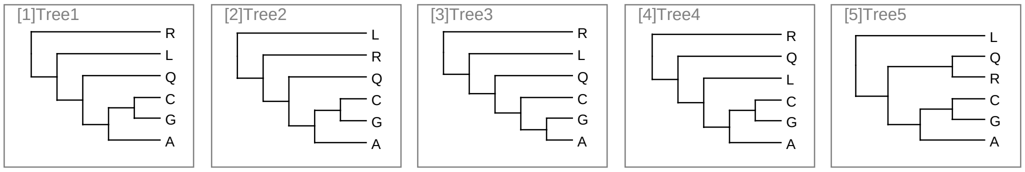


Fig. S55: **The top 5 out of 796 gene tree topologies from the 1850 1-kb regions in 2R chromosome.** The topology frequencies are 3.2%, 2.6%, 1.8%, 1.8%, and 1.5%, respectively.

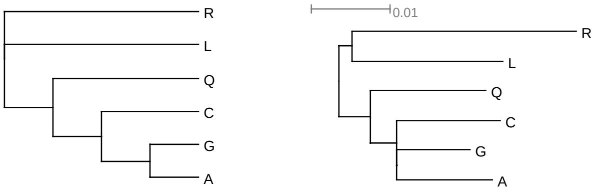


Fig. S56: **The majority-rule consensus tree** (left) and **the maximum likelihood tree** (right) of a 10-kb region from 2R chromosome. The majority-rule consensus tree of this region has three internal branches.

- Although the ML tree also supports (Q,(C,(G,A))), the branch length between C and (C,A) is 0, indicating C,G,A are unresolved.

We repeated the same analysis for the ten 1-kb regions within the 10-kb region; results in Fig. S57.

- Out of 10 consensus trees of the 1-kb regions (Fig. S57), 2 are fully unresolved (star tree), 5 have one internal branch, 2 have two internal branches and the rest one shows

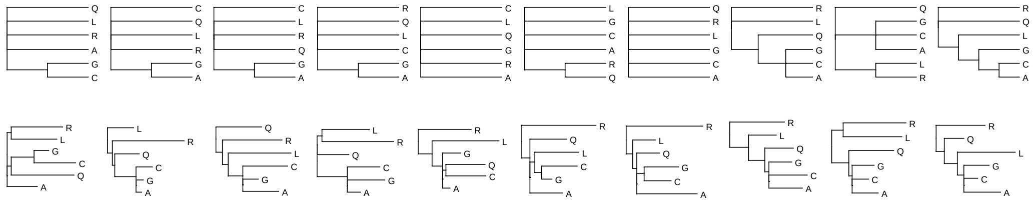


Fig. S57: **The majority-rule consensus trees** (top) and **the maximum likelihood trees** (bottom) of the ten 1-kb regions within the 10-kb region in Fig. S56.

three internal branches. All are different from the concatenation consensus tree in Fig. S56.

- Comparing the 10 maximum likelihood trees of the 1-kb regions (Fig. S57) with the concatenation maximum likelihood tree in Fig. S56, only 2 have the same topology as the concatenation one. 7 out of 10 trees show unresolved branches.

7.3 Comparison between $*$BEAST and our method given regions with weak phylogenetic signals from the X chromosome

We randomly sampled 228 1-kb regions from the X chromosome (at least 64-kb apart). For each region, we inferred 100 bootstrap trees, computed the majority-rule consensus tree, and finally analyzed the number of loci that have gene trees with 0-4 internal branches to assess the signal in the data (Table S10). We found there is no fully resolved tree, more specifically, all the roots are non-binary.

We fed the 228 regions as independent loci into $*$BEAST and our method.

- The co-estimated gene tree topologies and frequencies. The top 5 gene tree topologies in both $*$BEAST and our method are shown in Fig. S58.

- The species phylogenies in the 95% credible sets. There is only one species tree topology in the 95% credible set from $*$BEAST, shown in Fig. S59. The 95% credible set from our method contains three species networks with 4-reticulations (Fig. S60). The divergence times of the roots inferred by our method are twice higher than the ones inferred by $*$BEAST.

Table S10: **The number of loci from X chromosome that have majority-rule consensus trees with 0-4 internal branches.** Gene trees with 0 internal branches are star phylogenies with no signal at all. Gene trees with 4 internal branches are fully resolved. There is no fully resolved tree by setting consensus threshold to 50 or 70.

consensus threshold	0	1	2	3	4
50	13	18	57	140 (61.4%)	0
70	21	61	87	59 (25.9%)	0

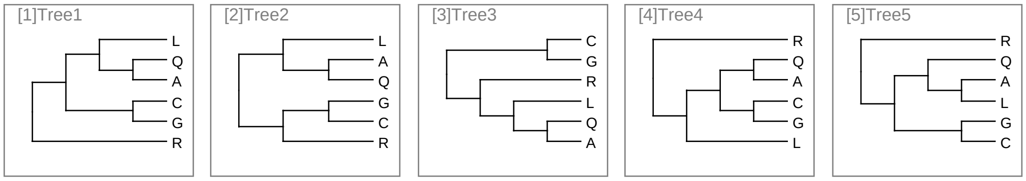


Fig. S58: **The top 5 gene tree topologies inferred from both ∗BEAST and our method on 228 1-kb regions from X chromosome.** The frequencies are 16.3%, 10.6%, 7.4%, 6.4%, and 4.2%, respectively from ∗BEAST, and 20.1%, 9.2%, 6.7%, 5.9%, and 5.1%, respectively from our method.

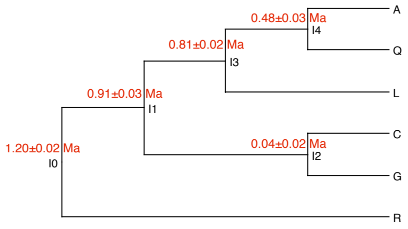


Fig. S59: **The species tree in the 95% credible set using ∗BEAST given 228 1-kb regions from X chromosome.** The divergence times are marked in red.

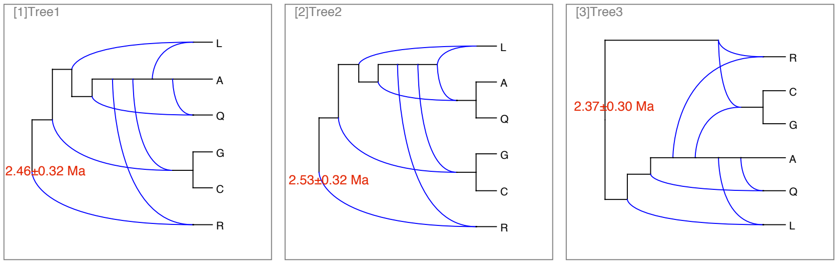


Fig. S60: **The phylogenetic networks in the 95% credible set using our method given 228 1-kb regions from X chromosome.** The proportions are 62%, 25%, and 10%, respectively. The posterior values are -337855.6 ± 78.1, -337849.5 ± 35.3, and -337887.8 ± 44.0, respectively. The population size parameters are 6.98 ± 0.43 E-3, 6.84 ± 0.32 E-3, and 6.66 ± 0.41 E-3, respectively. The divergence times of the root nodes are marked in red.

7.4 Comparison of BEAST, ∗BEAST and our method given regions with strong phylogenetic signals from the X chromosome

As we stated above, 59 out of 228 consensus trees have three internal branches using consensus threshold 70. We assume that using the sequence alignments from these 59 regions, the differences of species topologies and root divergence times between ∗BEAST and our method would be less significant, and our method would not add reticulations unnecessarily to account for the uncertainty of data.

We fed the 59 regions as independent loci to BEAST, ∗BEAST and our method.

- The co-estimated gene trees.
  - Topologies and frequencies. The top 6 gene tree topologies inferred by ∗BEAST and our method are shown in Fig. S61.
  - Divergence times. We plotted the divergence times of the most recent common ancestors (MRCAs) of (C,G), (A,Q), (A,C,G), (Q,C,G) and (A,Q,C,G) from gene tree samples inferred by BEAST (sandy brown), ∗BEAST (green) and our method (blue) in Fig. S62.
    - ∗ The divergence times estimated by BEAST vary widely, and the minimum

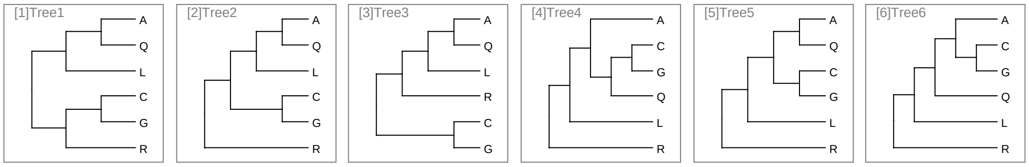


Fig. S61: **The top 6 gene tree topologies inferred by ∗BEAST and our method on the 59 1-kb regions from X chromosome.** The frequencies are 21.7%, 21.1%, 6.7%, 6.1%, 5.8%, and 4.5%, respectively, from ∗BEAST, and 29.7%, 16.1%, 4.9%, 8.1%, 5.4%, and 6.4%, respectively, from our method.

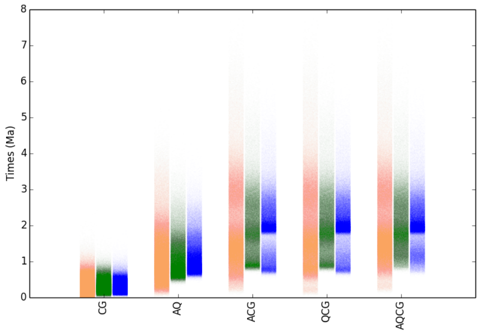


Fig. S62: **The divergence times of the MRCAs of (C,G), (A,Q), (A,C,G), (Q,C,G) and (A,Q,C,G) from estimated gene tree samples inferred by BEAST (sandy brown), ∗BEAST (green), and our method (blue).** The input is the 59 1-kb regions from the X chromosome.

values of (C,G), (A,Q), (A,C,G), (Q,C,G) and (A,Q,C,G) are close to zero. Using the times of gene trees estimated by BEAST to infer the speciation times of the species tree would result in gross underestimation of all the species divergence times. Comparing with BEAST, co-estimation methods provide more accurate results.

∗ The divergence times estimated by our method show smaller variation than BEAST and ∗BEAST. The minimum value of (C,G) is similar to the ones inferred by ∗BEAST. The minimum value of (A,Q) is closer to the minimum values of (A,C,G), (Q,C,G), (A,Q,C,G). Note that in the species networks inferred from our method (Fig. S63), the gene trees can either “use” the reticulation edges, or "grow" within the branches of the species tree backbone. Indeed, we can clearly see two groups of divergence times for (A,C,G), (Q,C,G), (A,Q,C,G). The gene trees using the reticulation edges could explain why the minimum values of (A,C,G), (Q,C,G), (A,Q,C,G) are smaller than $*$BEAST, while the minimum values of (A,Q) is slightly larger.

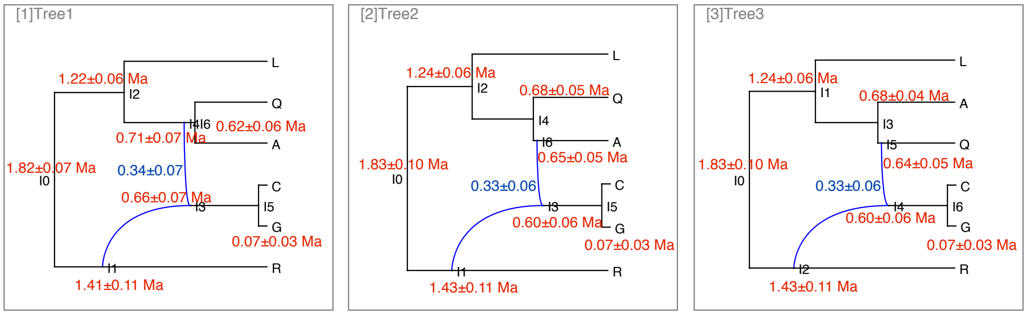


Fig. S63: **The species networks in the 95% credible set using our method given 59 1-kb regions from X chromosome.** The proportions are 77.2%, 12.6%, and 10.2%, respectively. The posterior values are -90491.4 ± 18.1, -90493.4 ± 18.1, and -90493.9 ± 18.0, respectively. The population size parameters are 1.26 ± 0.14, 1.31 ± 0.14, and 1.26 ± 0.13 E-2, respectively. The divergence times are marked in red and the inheritance probabilities are labeled in blue.

  - The densi-tree plots of 5 loci from BEAST, $*$BEAST and our method are shown in Fig. S64.

- The species phylogenies in the 95% credible sets.

  - There is only one species tree topology in the 95% credible set from $*$BEAST, shown in Fig. S65.
  - The 95% credible set from our method contains three 1-reticulation phylogenetic networks (Fig. S63). The divergence times of (C,G), (C,G,R), (L,A,Q) and the root are similar across topologies. The MAP topology supporting the hybridization (A,Q)→(C,G), shows lower divergence time (around 0.62) than

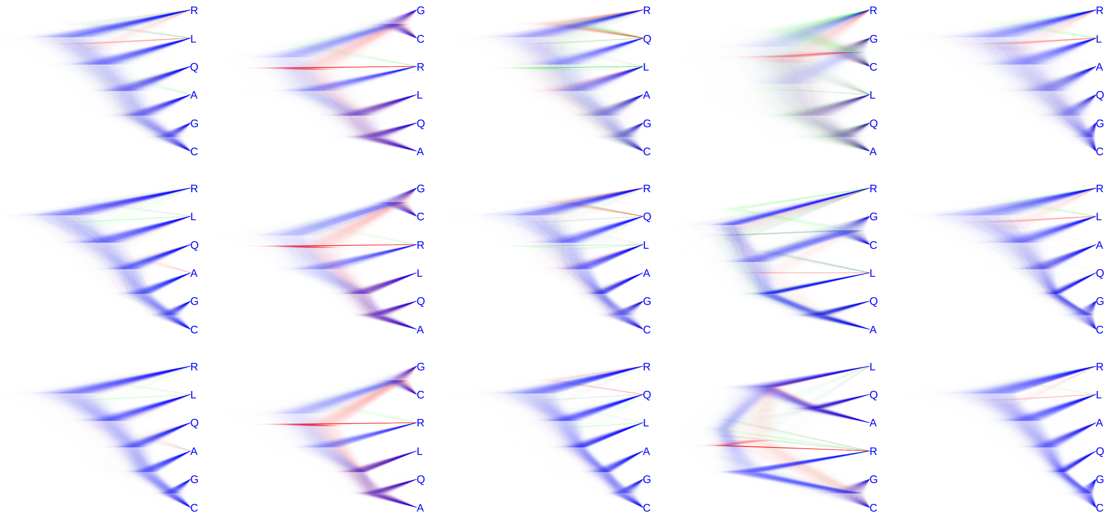


Fig. S64: **The DensiTree plots of gene tree samples of 5 loci from BEAST (top row), ∗BEAST (middle row) and our method (bottom row).**

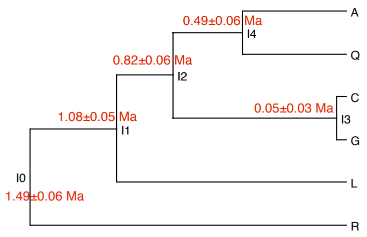


Fig. S65: **The species tree in the** 95% **credible set using ∗BEAST given 59 1-kb regions from X chromosome.** The divergence times are marked in red.

the other two topologies supporting the hybridization A/Q→(C,G) (around 0.68 Mya).

- The species tree inferred by ∗BEAST has the same topology as the MAP network inferred by our method by removing the reticulation edge R→(C,G).
- The divergence times of the root inferred by our method are around 1.83 Mya, which is higher than 1.49 Mya inferred by ∗BEAST (of course, the difference is not as significant as using data with weak phylogenetic signals).

- Species phylogenies and gene trees.

- The proportion of gene tree samples that have the same topology as the species tree inferred by $*$BEAST is only 5.8% from $*$BEAST and 5.4% from our method.
- The proportion of gene tree samples that have the same topology as the common species tree backbone by removing the reticulation edges from the networks inferred from our method is 21.7% from $*$BEAST and 29.7% from our method.

7.5 Bayesian inference given 382 regions with strong phylogenetic signals from autosomes

We randomly sampled 2791 1-kb regions from the autosomes (at least 64-kb apart). For each region, we inferred 100 bootstrap trees, computed the majority-rule consensus tree, and finally analyzed the number of regions that have gene trees with 0-4 internal branches (Table S11). We found there is no fully resolved tree, more specifically, all the roots are non-binary.

Table S11: **The number of loci from the X chromosome that have majority-rule consensus trees with 0-4 internal branches.** Gene trees with 0 internal branches are star phylogenies. Gene trees with 4 internal branches are fully resolved.

consensus threshold	0	1	2	3	4
70	624	985	800	382 (13.7%)	0

We fed the 382 regions whose bootstrapped trees had three internal branches as independent loci into our method. Out of 382 regions, 86, 93, 120, and 83 regions come from chromosomes 2L, 2R, 3L, and 3R respectively. We ran 4 MCMC chains of 6×10^7 iterations with 1×10^7 burn-in. One sample was collected from every 5,000 iterations.

- 95% credible sets. The 95% credible set of topologies contains only one species network, as shown in Fig. S66. The posterior value of this network is $-624378.9 \pm$

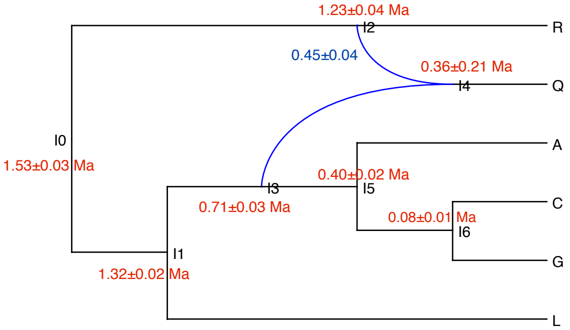


Fig. S66: **The phylogenetic network in the 95% credible set using our method on the 382 1-kb regions from the mosquito autosomes.** The divergence times are marked in red, and the inheritance probability is marked in blue. The posterior value of this network is -624378.9 ± 41.0. The population size is 1.56 ± 0.05 E-2.

41.0. The population size is 1.56 ± 0.05 E-2. The reticulation edge R→Q has an inheritance probability of 0.45 ± 0.04.

- Convergence and mixing. The trace plots are shown in Fig. S67. We can see the posterior values are similar across MCMC chains.

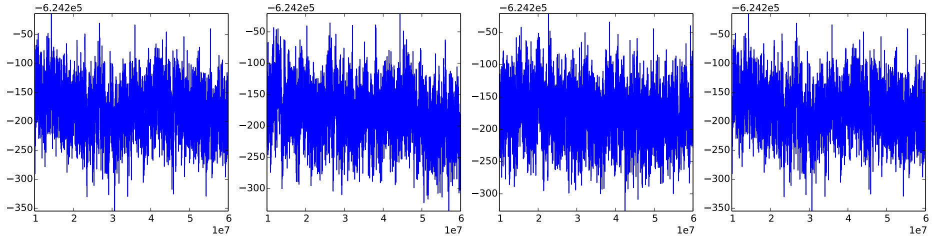


Fig. S67: **Trace plots of MCMC chains using our method given 382 1-kb regions from mosquito autosomes.**

8 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.

8.1 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. S25 is shown in in Fig. S68.

The runtimes, in minutes, for analyzing the simulated data sets with Intermixture/Gene flow patterns under the true species phylogenies in Fig. S30 were as follows:

- The recurrent intermixture in S30**A**, $\Delta t = 1$: 44.9 ± 3.5
- The recurrent intermixture in S30**A**, $\Delta t = 2$: 49.5 ± 6.0
- The recurrent intermixture in S30**A**, $\Delta t = 3$: 54.1 ± 6.6
- The recurrent gene flow in S30**A**, $|mt| = 1$: 50.6 ± 5.9
- The recurrent gene flow in S30**A**, $|mt| = 2$: 49.7 ± 6.0

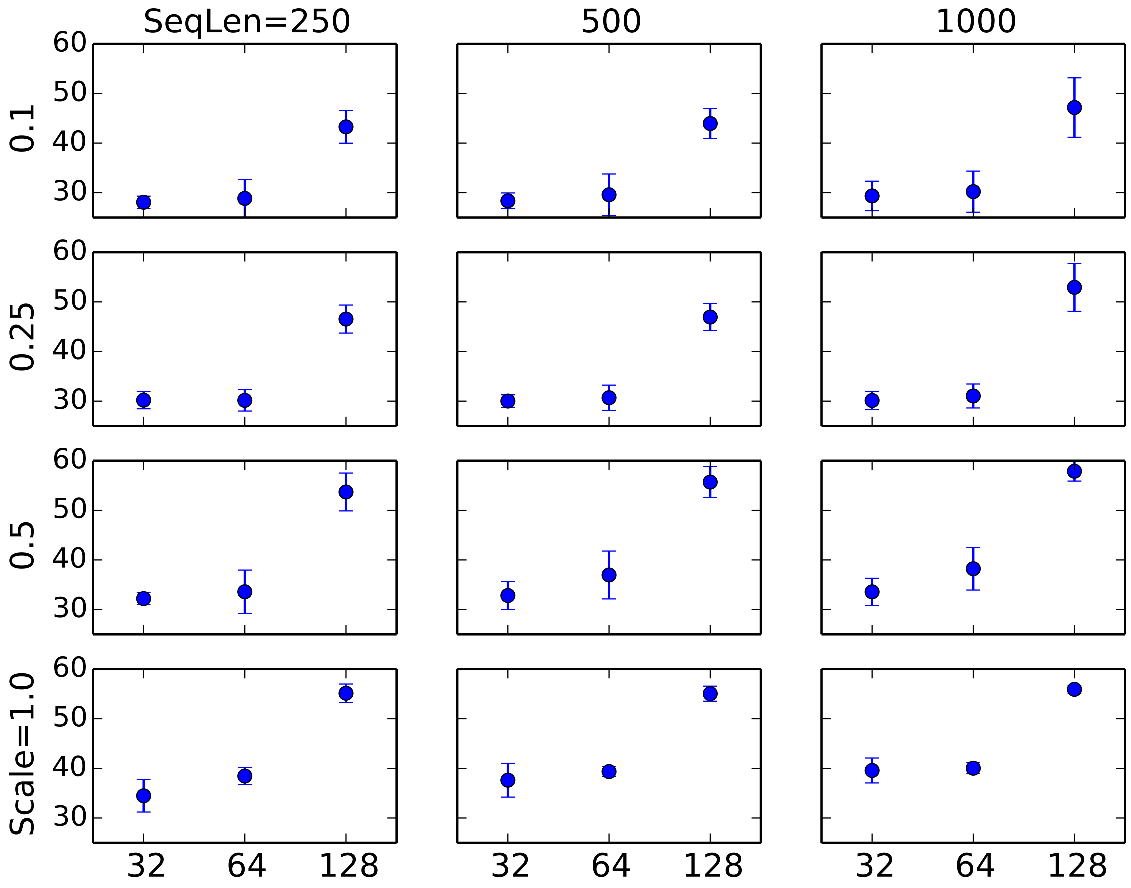


Fig. S68: **The runtimes in minutes under different simulation conditions.** From top to bottom: 0.1, 0.25, 0.5, 1.0 divergence time scale, respectively. From left to right: 250, 500, and 1000 bps sequence length, respectively. Within each plot: 32, 64, 128 loci, respectively.

- The paraphyletic intermixture between non-sister species in S30**B**, $t = 1$: 42.8 ± 3.4
- The paraphyletic intermixture between non-sister species in S30**B**, $t = 0$: 42.1 ± 3.4
- The paraphyletic gene flow between non-sister species in S30**B**, $mt_2 = 1$: 45.8 ± 4.6
- The paraphyletic gene flow between non-sister species in S30**B**, $mt_2 = 0$: 46.5 ± 5.7
- The isolation-migration between sister species in S30**C**, $t = 1$: 36.2 ± 3.8
- The isolation-migration between sister species in S30**C**, $t = 0$: 48.8 ± 7.6

- The isolation-migration between sister species in S30**C**, $mt = 1$: 36.5 ± 3.9
- The isolation-migration between sister species in S30**C**, $mt = 0$: 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. S30**A** were as follows:

- The recurrent intermixture in S30**A**, $\Delta t = 1, \#3$: 62.3 ± 4.7
- The recurrent intermixture in S30**A**, $\Delta t = 1, \#5$: 82.0 ± 5.3
- The recurrent intermixture in S30**A**, $\Delta t = 2, \#3$: 64.3 ± 2.8
- The recurrent intermixture in S30**A**, $\Delta t = 2, \#5$: 85.3 ± 4.8

8.2 Biological data sets

For the 8-taxon wheat data set, the runtimes were 13.5 hours for the 68-locus data set, and $13.7 \sim 14.5$ hours for the 137-locus data set.

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
- 7-taxon, 28-locus data set: $23.5 \sim 24$ hours
- 5-taxon, 106-locus data set: $16.6 \sim 18$ hours

For the mosquito data set, the runtimes were as follows:

- X chromosome, 6-taxon, 228-locus data set: $11.2 \sim 12.5$ hours
- X chromosome, 6-taxon, 59-locus data set: $5.2 \sim 5.7$ hours
- Autosome, 6-taxon, 382-locus data set: $19.2 \sim 24.5$ hours

9 PhyloNet implementation and usage

We implemented our method in PhyloNet (18), 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.

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

S12. Nakhleh, L. 2010. A metric on the space of reduced phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(2): 218–222.

S13. 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.

S14. Robinson, D. and Foulds, L. 1981. Comparison of phylogenetic trees. *Math. Biosci.*, 53: 131–147.

S15. 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.

S16. Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

S17. Sturtz, S., Ligges, U., Gelman, A., *et al.* 2005. R2winbugs: a package for running winbugs from r. *Journal of Statistical software*, 12(3): 1–16.

S18. Than, C., Ruths, D., and Nakhleh, L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC bioinformatics*, 9(1): 322.

S19. Wen, D., Yu, Y., and Nakhleh, L. 2016a. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genetics*, 12(5): e1006006.

S20. Wen, D., Yu, Y., Hahn, M. W., and Nakhleh, L. 2016b. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25(11): 2361–2372.

S21. Yu, Y., Warnow, T., and Nakhleh, L. 2011. Algorithms for mdc-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, 18(11): 1543–1559.

S22. 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.