Maximum likelihood inference of reticulate evolutionary histories

Yun Yu, Jianrong Dong, Kevin J. Liu, and Luay Nakhleh

Departments of Computer Science and Ecology and Evolutionary Biology, Rice University, Houston, TX 77005

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved October 7, 2014 (received for review April 30, 2014)

Hybridization plays an important role in the evolution of certain groups of organisms, adaptation to their environments, and diversification of their genomes. The evolutionary histories of such groups are reticulate, and methods for reconstructing them are still in their infancy and have limited applicability. We present a maximum likelihood method for inferring reticulate evolutionary histories while accounting simultaneously for incomplete lineage sorting. Additionally, we propose methods for assessing confidence in the amount of reticulation and the topology of the inferred evolutionary history. Our method obtains accurate estimates of reticulate evolutionary histories on simulated datasets. Furthermore, our method provides support for a hypothesis of a reticulate evolutionary history inferred from a set of house mouse (Mus musculus) genomes. As evidence of hybridization in eukaryotic groups accumulates, it is essential to have methods that infer reticulate evolutionary histories. The work we present here allows for such inference and provides a significant step toward putting phylogenetic networks on par with phylogenetic trees as a model of capturing evolutionary relationships.

Phylogenetic trees have long been a mainstay of biology, providing an interpretive model of the evolution of molecules and characters and a backdrop against which comparative genomics and phenomics are conducted. Nevertheless, some evolutionary events, most notably horizontal gene transfer in prokaryotes and hybridization in eukaryotes, necessitate going beyond trees (1). These events result in reticulate evolutionary histories, which are best modeled by phylogenetic networks (2). The topology of a phylogenetic network is given by a rooted, directed, acyclic graph (rDAG) that is leaf-labeled by a set of taxa (Fig. 1; more details are provided in Model and SI Appendix). Reticulation events result in genomic regions with local genealogies that are incongruent with the speciation pattern. Several methods and heuristics use this incongruence as a signal for inferring reticulation events and reconstructing phylogenetic networks from local genealogies. These methods, which are surveyed elsewhere (2–4), assume that reticulation events are the sole cause of all incongruence among the gene trees and seek phylogenetic networks to explain all of the incongruence. A serious limitation of these methods is that they would grossly overestimate the amount of reticulation in a dataset when other causes of incongruence are at play. Indeed, several recent studies (5–9) have shown that detecting hybridization in practice can be complicated by the presence of incomplete lineage sorting (ILS) (Fig. 1).

Recently, a set of methods was devised to analyze data where reticulation and ILS might both be simultaneously at play (10–15). However, these methods are all applicable to simple scenarios of species evolution and mostly assume a known hypothesis about the topology of the phylogenetic network. As reported (16, 17), we devised methods for computing the likelihood of a phylogenetic network, given a set of gene tree topologies. Still, these methods did not allow for inference of phylogenetic networks (they assume a given phylogenetic network topology and compute its likelihood).

To the best of our knowledge, the first method to conduct a search of the phylogenetic network space in search of optimal phylogenies is described in a study by our group (18). However, this method is based on the maximum parsimony criterion: It seeks a phylogenetic network that minimizes the number of “extra lineages” resulting from embedding the set of gene tree topologies within its branches.

Progress with phylogenetic network inference notwithstanding, methods of inferring reticulate evolutionary histories while accounting for ILS are still considered to be in their infancy and inapplicable broadly (9). This inapplicability stems mainly from two major issues: the lack of a phylogenetic network inference method and the lack of a method to assess the confidence in the inference. Here, we develop methods that resolve both issues and carry phylogenetic networks into the realm of practical phylogenomic applications. For the inference, we propose operations for traversing the phylogenetic network space, as well as methods for assessing the complexity of a network. For measuring branch support of inferred networks, we use the bootstrap method. Furthermore, we derive, for the first time to our knowledge, the distribution (density) of gene trees with branch lengths, given a phylogenetic network, and use it in inference. Our methods provided very good results on simulated datasets. We also applied our methods to a dataset of thousands of loci from five house mouse (Mus musculus) genomes. The analysis yielded a well-supported evolutionary history with two hybridization events.

Model

We seek to infer a phylogenetic network Ψ that models the (potentially reticulate) evolutionary history of a set X of species, where multiple individuals might be sampled per species. We use the phylogenetic network model given by Nakhleh (2). A

Significance

Phylogenetic trees play a central role in biology, modeling evolutionary histories of taxa ranging from genes, to genomes, and to species. Although trees will continue to be an essential modeling tool in evolution, phenomena such as hybridization, or gene flow more generally, result in evolutionary histories that are best modeled by phylogenetic networks. Inference of such networks is complicated by the presence of other evolutionary events, such as incomplete lineage sorting (ILS). Here, we provide a maximum likelihood method for inferring reticulate evolutionary histories while accounting for ILS. The method enables new evolutionary analyses under more complex evolutionary scenarios than existing methods can handle.

Author contributions: Y.Y. and L.N. designed research; Y.Y., J.D., and L.N. performed research; Y.Y., J.D., and K.J.L. contributed new reagents/analytic tools; Y.Y., J.D., K.J.L., and L.N. analyzed data; and Y.Y., J.D., K.J.L., and L.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence may be addressed. Email: nakhleh@cs.rice.edu or yy9@cs.rice.edu.

16448-16453 | PNAS | November 18, 2014 | vol. 111 | no. 46

www.pnas.org/cgi/doi/10.1073/pnas.1407950111

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1407950111/-/DCSupplemental.
Fig. 1. Phylogenetic networks. Here, the MRCA of A and B split from its MRCA with C, and some time after A and B split, hybridization occurred between B and C. Four independent loci, Δ, ●, ■, and ♦, are illustrated, for which a single individual is sampled from each of A and C and six individuals are sampled from B. Two gene trees are depicted for the △ and ♦ loci, and both trees agree in terms of their shapes. However, the disagreement of the species splitting pattern with the gene tree in red is due to ILS, whereas the disagreement with the gene tree in blue is due to hybridization. Furthermore, the Δ locus exhibits no evidence of hybridization in B, the ● locus has lost all signal of vertical inheritance from the MRCA of B with A, and the other two loci exhibit varying degrees of hybridization signal in the population. Locus-specific inheritance probabilities are needed to capture such scenarios.

phylogenetic Φ-network, or Φ-network for short, Ψ is an rDAG whose leaves are bijectively labeled by the set Φ of taxa and whose every internal node (except the root) has in-degree 1 and out-degree greater than 1 (tree nodes) or in-degree 2 and out-degree 1 (reticulation nodes). We use V(Ψ) and E(Ψ) to denote the set of nodes and edges, respectively, of phylogenetic network Ψ. Every edge (or branch) b of Ψ has a length \(\lambda_b = t_b/N_b\) in coalescent units, where \(t_b\) is the duration of edge \(b\) in generations and \(N_b\) is the population size corresponding to branch \(b\). A consequence of this setting is that the phylogenetic network does not have to be ultrametric. Furthermore, whereas the model does not require or necessitate a constant population size across all branches of the network, the population size and number of generations of each branch are dependent, given the branch’s length. In other words, the values of neither of these two parameters can be uniquely determined, given the length of a branch in our model (e.g., doubling both keeps the branch length unchanged). As is common in the literature in this area, we use a single composite parameter Ψ to denote the phylogenetic network topology and its branch lengths.

Tracing the evolution of a lineage from a leaf of the network back toward the root follows the multispecies coalescent model on trees, yet with one major difference: As a lineage encounters a reticulation node, it tracks one of the two parents of that node according to an inheritance probability. Because the probabilities of inheritance vary from one hybridization event to another in the network, and because different loci may provide different hybridization signals in the population (Fig. 1), the inheritance probabilities are given by a \(E(Ψ) \times m\) matrix Γ, where \(m\) is the number of independent loci (given the species phylogeny) in the dataset being analyzed and the entries of Γ satisfy three conditions for every 1 ≤ j ≤ m: (i) \(Γ(b,j) \in [0,1]\) for every \(b \in E(Ψ)\), (ii) \(Γ(b,j) = 1\) for every edge \(b\) incident into a tree node, and (iii) \(Γ(b,j) + Γ(b',j) = 1\) for every distinct pair \(b, b' \in E(Ψ)\) such that \(b\) and \(b'\) are incident into the same reticulation node. For an edge \(b\) incident into node \(v\) in Ψ, the entry \(Γ[b,j])\) denotes the probability that a sample from locus \(i\) tracks branch \(b\) when “entering” the population represented by node \(v\). It is important to note here that the topology and branch lengths of Ψ, as well as the matrix Γ, are to be inferred from the data; details are given below and in SI Appendix.

Likelihood Formulation Based on Sequence Data. Consider \(m\) independent loci along with a set \(S = \{S_1, \ldots, S_m\}\) of sequence alignments, where \(S_i\) corresponds to locus \(i\). The number of sequences in each \(S_i\) equals the total number of individuals from which a sequence is available for locus \(i\), and this number can vary from one locus to another. Under the independence assumption, the likelihood of an evolutionary history Ψ and inheritance probabilities Γ is given by

\[
L(Ψ, Γ | S) = \prod_{i=1}^{m} \int p(S_i | g) p(g | Ψ, Γ) dg, \tag{1}
\]

where \(p(S_i | g)\) is the probability of the (sequence) data, given a particular gene genealogy \(g\), and \(p(g | Ψ, Γ)\) is the distribution (density) of gene genealogies (topologies and branch lengths), given the model parameters. The integral in the equation is taken over all possible values of \(g\), where \(g\) represents a gene genealogy (topology and branch lengths). It is important to note here that for computing the probability \(p(S_i | g)\), the genealogy’s branch lengths need to be converted to coalescent units. Given the population mutation rate \(θ = 4Nμ\), where \(Nμ\) is the effective population size and \(μ\) is the per-site mutation rate, the conversion from units of the number of nucleotide substitutions per site to coalescent units can be done by multiplying every gene tree branch length by \(2/θ\).

Likelihood Formulation Based on Estimated Genealogies. Although the likelihood formulation given by Eq. 1 uses all of the information in the data, inference of the species phylogeny from estimated genealogies can significantly speed up the inference. In this case, the likelihood formulation becomes

\[
L(Ψ, Γ | S) = \prod_{i=1}^{m} p(G_i | Ψ, Γ), \tag{2}
\]

where \(G_i\) is the genealogy estimated for locus \(i\) and \(S = \{G_1, \ldots, G_m\}\). Here, \(p(G_i | Ψ, Γ)\) is the probability mass function (pmf) or probability density function (pdf), depending on whether the \(G_i\) s are given by their topologies alone or by topologies and branch lengths, respectively. Indeed, for the case when the topology of Ψ is a tree, the Species Tree Estimating using Maximum Likelihood (STEM) method (19) and the Species Tree Inference with Likelihood for Lineage Sorting (STELLS) method (20) use this formulation for inference of Ψ, where the former makes use of the gene genealogies’ topologies and branch lengths and the latter makes use of only the genealogies’ topologies.

Inference of high-quality species phylogenies based on Eq. 2 requires accurate estimates of the individual gene genealogies. Because the methods are aimed at data from closely related species and potentially multiple individuals from populations, the signal in the sequence data might be too low for estimating accurate gene genealogies. Although inference from sequences (Eq. 1) accounts naturally for this issue, it is important to account for it explicitly when conducting inference from estimates of gene genealogies. Assume that for each locus \(i\), the uncertainty in estimation is accounted for by having a collection of gene genealogies \(G_i = \{G_{i1}, \ldots, G_{im}\}\); for example, these gene
genealogies could be the trees inferred for locus $i$ based on $p$ bootstrap replicates. In this case, we have

$$p(G_i | \Psi, \Gamma) = \left( \prod_{j \in G_i} p(g_j | \Psi, \Gamma) \right)^{1/|G_i|}, \quad [3]$$

where, once again, $p$ is given by the pmf or pdf, depending on whether the individual genealogy estimates are given by their topologies alone or by their topologies and branch lengths, respectively. The likelihood model is now given by Eq. 2, with $p$ from Eq. 3 being used instead of the pmf or pdf for individual genealogies. We demonstrate the performance of this formulation in Results.

**Maximum Likelihood Inference.** Under maximum likelihood (ML), the inference problem amounts to computing the pair $(\Psi, \Gamma)$ that maximizes the likelihood function based on sequence data using Eq. 1 or based on estimated gene genealogies using Eq. 2. Inference based on Eq. 1 requires computing the integral over all possible gene genealogies. Bryant et al. (21) provided an efficient algorithm for computing this integral when each independent locus is given by a biallelic marker. To enable ML inference based on Eq. 1, the algorithm of Bryant et al. (21) needs to be extended along three axes: allowing for sites with more than two states, allowing for the species history to have reticulations, and allowing for each marker to consist of more than a single site. Although extensions along all three axes are technically achievable, inference of even three-taxon networks with a single reticulation from a few sites is computationally prohibitive (Discussion). We therefore focus on inference based on Eq. 2 in this work. Using this formulation, the pmf $p(G_i | \Psi, \Gamma)$, when $G_i$ is the gene genealogy’s topology alone, is computable using the algorithms of Yu et al. (16, 17). In Results, we derive the pdf of gene genealogies (with branch lengths), given a phylogenetic network.

Given all of these tools, the inference problem is still very hard computationally, because the optimal $\Psi$ and $\Gamma$ need to be computed. It is standard in the case of species tree inference to use heuristics that walk the tree space in search of optimal solution candidates. It makes sense, therefore, to devise techniques for walking the phylogenetic network space in search of optimal phylogenetic networks while optimizing branch lengths and the $\Gamma$ matrix. However, extra caution must be taken when searching the network space. In the case of trees, all rooted, binary trees on a given number of taxa are essentially different models with the same number of parameters. In the case of networks, on the other hand, an arbitrarily large number of reticulation nodes can be added during the search, resulting in more complex models that, by definition, could fit the data at least as well as simpler models. Because the goal is to estimate the true amount of reticulation, rather than only fitting the data, we address this challenge in two ways. First, we devise a search heuristic that searches the phylogenetic network space in layers. Second, we explore the use of cross-validation as a method to ameliorate overfitting the data, which adds to the array of other methods (e.g., information criteria) that have already been used (12, 16). Finally, to assess the fit of the inferred phylogenetic network to the data, we devise a parametric bootstrap approach that allows us to quantify branch support for the phylogenetic network. We give details for all of these methods below and in SI Appendix.

**Results**

**Probability Density of a Gene Tree.** Given a phylogenetic network $\Psi$ and a gene genealogy $G_j$ for locus $j$ (topology and branch lengths in both cases), we denote by $H(G_j)$ the set of all coalescent histories of $G_j$ within the branches of $\Psi$. Then, the distribution (density) of gene trees is given by

$$p(G_j | \Psi, \Gamma) = \sum_{h \in H(G_j)} p(h | \Psi, \Gamma), \quad [4]$$

where $\Gamma$ is the inheritance probabilities matrix, as described above. For an edge $b = (x, y) \in E(\Psi)$, we define $T_b(h)$ to be the vector of times (in increasing order) of coalescence events that occur on branch $b$ under the coalescent history $h$ and the time of node $y$ (a formal definition is provided in SI Appendix). We denote by $u_i(h)$ the $i$th element of the vector. Furthermore, we denote by $v_i(h)$ the number of gene lineages entering edge $b$ and by $v_0(h)$ the number of gene lineages leaving edge $b$ under $h$. Then, we have

$$p(h | \Psi, \Gamma) = \prod_{b \in E(\Psi)} \prod_{i=1}^{u_i(b)-1} e^{-\left(\frac{u_i(b)-i+1}{2}\right)} \left(T_b(h), \ldots, T_b(h)\right) \times e^{-\left(\frac{v_i(h)}{2}\right)} \left(v_0(b) - T_b(h) + \frac{\omega_c}{n}\right) \times \Gamma(b, j)^{u_i(h)}, \quad [5]$$

where $\omega_c(b)$ for edge $b = (x, y)$ is the time of node $x$ in the phylogenetic network $\Psi$. A full derivation of the formula and a more efficient algorithm for computing it along the lines of Yu et al. (17), which avoid explicit summations over the possible coalescent histories, are given in SI Appendix.

**Searching the Space of Phylogenetic Networks.** Letting $\Omega(n)$ denote the space of all phylogenetic networks on $n$ taxa, we denote by $\Omega(n, k)$ the subspace of $\Omega(n)$ that contains all phylogenetic networks (rDAGs) with $n$ leaves and $k$ reticulation nodes. In particular, $\Omega(n, 0)$ is the subspace that contains all phylogenetic trees. To search the phylogenetic network space in a layered fashion, we define two operations that allow for searching within $\Omega(n, k)$ for a given $k$: one operation that allows the search to ascend a layer from $\Omega(n, k)$ to $\Omega(n, k + 1)$ and one operation that allows the search to descend a layer from $\Omega(n, k)$ to $\Omega(n, k - 1)$. For searching within a layer, the operations either relocate the destination of a reticulation edge or relocate the source of an edge (reticulation or not). For ascending a layer, the operation consists of adding a reticulation edge between two existing edges in the network, and for descending a layer, the operation removes a reticulation edge (more details are provided in SI Appendix). It is worth mentioning that although the space of all phylogenetic tree topologies on $n$ taxa is finite, the space of all phylogenetic network topologies on $n$ taxa is, in theory, infinite, because $\Omega(n) = \cup_{k \geq 0} \Omega(n, k)$ and $k$ are unbounded. For example, consider the case of only two taxa. There is a unique, rooted tree in this case. However, because multiple hybridization events could happen between the same two sister taxa at different times, any number of horizontal edges can be added between these two taxa. Nevertheless, whether such repetitive hybridization scenarios are identifiable from typical genomic datasets is a different question.

A heuristic for estimating the optimal branch lengths for a fixed species tree topology, given gene tree topologies, that is based on repeated application of Brent’s method (22) was introduced by Wu (20). We use a similar heuristic for estimating the phylogenetic network branch lengths and inheritance probabilities (full details are given in SI Appendix). Coupling topological transformations and parameter estimation heuristics with the likelihood formulation above enables searching the space in a hill-climbing manner to infer an ML phylogenetic network. Given the existence of local optima within each layer, multiple, independent runs can be made.

**Controlling for Model Complexity.** Because networks in $\Omega(n, k + 1)$ provide more complex models than networks in $\Omega(n, k)$, the approaches described above must handle the model selection problem. Information criteria have already been used in the context of phylogenetic networks (12, 16), and we use them here
(instead of searching based on the likelihood score, the search proceeds based on the values of these criteria, which incorporate the likelihood scores). Another approach that we propose here, for the first time to our knowledge, is the use of K-fold cross-validation, whereby the input set of gene trees is partitioned into K subsets of equal sizes, the parameters of the model are inferred from K - 1 subsets, and the model’s fit of the remaining subset is computed. This fit is computed by comparing the frequencies of the gene trees in the validation subset with the distribution of the gene trees produced by the inferred network. If the fit of the best network Ψ* found in \(\Omega(n, k + 1)\) is not much better (we use a cutoff of 3% improvement, chosen based on empirical observations) than the fit of the best network Ψ found in \(\Omega(n, k)\), we declare k to be the correct estimate of the number of reticulation nodes and Ψ* to be the optimal phylogenetic network. It is important to note here that this cross-validation idea works only for fully resolved gene tree topologies, because in the case of gene trees with branch lengths, the frequencies of the gene trees in the validation subset are not informative.

Finally, to assess the support of the phylogenetic networks we infer, we propose using parametric bootstrapping. Having inferred a network Ψ from the data G, we use Ψ to generate l datasets, from which we infer l phylogenetic networks Ψ1, . . . , Ψl. We then estimate the support of each branch b in Ψ as the number of networks (out of the l) that have an equivalent branch. We say that two edges in two phylogenetic networks are equivalent if (i) either or both are reticulation edges or both are not and (ii) both define the same clusters (the cluster defined by a branch is the set of all taxa under that branch in the network).

**Performance on Simulated Data.** We implemented all of the methods described above in the publicly available, open-source software package PhyloNet (23) and studied the performance of the methods on several simulated datasets. In the simulation study whose results are reported in Fig. 2, we used phylogenetic network Ψ as the model network, and for various numbers of loci, we evolved gene trees under the coalescent within the branches of the network and then simulated sequence evolution on these gene trees with varying sequence lengths. We then estimated for each sequence alignment 100 gene trees using ML with bootstrapping. Finally, we inferred networks using our ML method from (i) true gene tree topologies, (ii) estimated gene tree topologies, (iii) true gene tree topologies and branch lengths, and (iv) estimated gene tree topologies and branch lengths. The results of (i) and (ii) are shown in Fig. 2B, whereas the results of (iii) and (iv) are shown in Fig. 2C. For each setting of the number of loci and sequence length, we generated 30 datasets and conducted inferences on all of them.

Whereas the hybridization in the model network involves B and the most recent common ancestor (MRCA) of C and D, the length of the branch between the hybridization event and the divergence of C and D from their MRCA can have a big effect on distinguishing between the true hybridization scenarios and the two given by Ψ2 and Ψ3 in Fig. 2A. Therefore, for every dataset, we recorded whether the method inferred one of the three networks shown in Fig. 2A, as opposed to any other network with a single reticulation.

Several trends can be observed in Fig. 2A. First, using the true gene tree topologies with branch lengths results in more accurate inferences than using gene tree topologies alone. This finding is not surprising, because the former type of data contains more information than the latter. In particular, when using 80 or 160 loci, the inferred network from the true gene trees with branch lengths is always the true network. On the other hand, when using only the gene tree topologies for 160 loci, in five of the 30 cases, the inference returned one of the two alternative networks Ψ1 and Ψ2. Second, the accuracy of the inferences improved as the number of loci increases and as the sequence length increases, although the increase in the number of loci had much more of a positive effect on the inference accuracy. Third, a very surprising result is that when using gene tree topologies alone, using the true gene trees almost never resulted in better accuracy than when using estimated gene tree topologies for a given number of loci. This result attests to the fact that when accounting carefully for uncertainty in the gene tree estimates, the method can obtain very good results. Even when using gene tree topologies and branch lengths, the gain in accuracy when using the true gene trees is very small compared with using the gene tree estimates with uncertainty taken into account. Fourth, the combination of a low value of inheritance probability (0.1 in this simulation) and a relatively short time between hybridization and subsequent speciation results in uncertainty in identifying the donor and recipient of the hybridization event. For example, when using gene tree topologies alone for 160 loci, the inferred network is always one of the three networks Ψ1, Ψ2, and Ψ3, even thought it is mostly Ψ1. We found that increasing the branch lengths or the inheritance probabilities would result in higher accuracies. Furthermore, in our simulations, we found that increasing the number of individuals sampled per taxon would result in improved accuracy, albeit rather slightly (SI Appendix). However, we expect that sampling more individuals would result in more significant improvements on larger or more complex datasets. In terms of the inferred inheritance probabilities, the true gene trees resulted in very accurate estimates, whereas estimated gene trees with branch lengths resulted, in general, in more accurate estimates of the probabilities. Finally, we found that cross-validation generally does better than information criteria at determining the number of hybridization events (including on the biological dataset, as discussed below). More extensive simulation results under scenarios that are easier for inference than the ones we discussed here are contained in SI Appendix.

**Analysis of a Multilocus House Mouse Dataset.** We also used our method to analyze a multilocus dataset of house mouse (M. musculus) genomes, obtained from the studies of Staubach...
et al. (7), Didion et al. (24), and Yang et al. (25) (more details are provided in SI Appendix). Staubach et al. (7) found substantial genome-wide evidence of subspecific introgression in all four populations, amounting to 3% of the genome in the two *M. m. domesticus* populations (one from France and the other from Germany), 4% in an *M. m. musculus* population from Kazakhstan, and 18% in an *M. m. musculus* population from the Czech Republic. However, it is important to note that the HAPMIX method (26), which was used by Staubach et al. (7), does not explicitly account for ILS.

Our study included all of the samples in the study of Staubach et al. (7). Furthermore, our study included additional samples from an *M. m. musculus* population from China (25) that were not used in the study of Staubach et al. (7). In this analysis, we used estimated gene tree topologies alone. The reason for doing so is that the genomic sequences are obtained from very closely related individuals (these individuals are five individuals from the same species), and very little variation exists in the data to estimate branch lengths with any accuracy. Furthermore, this low variation does not allow for proper bootstrap analysis of gene trees for the individual loci. The powerful signal in this dataset comes from the very large number of loci. In our analysis, we found a significant improvement in a phylogenetic network with a single reticulation over no reticulations and a significant improvement in a phylogenetic network with two reticulations over a single reticulation. However, when we continued the search for the optimal network with three reticulations, we found that the improvement gained by considering a third reticulation event was insignificant based on the information criteria, and that there was no improvement at all based on cross-validation. We thus called the optimal phylogenetic network with two reticulations as our hypothesis for the evolutionary history of this set of genomes. This evolutionary history is shown in Fig. 3 (more details of the results and analyses are provided in SI Appendix). The phylogenetic network is not ultrametric, and it is worth emphasizing that the branch lengths are given in coalescent units. Thus, the lack of ultrametricity could be due to different population sizes or, to a lesser degree, different generation times.

Our analysis of house mouse genomes produces an evolutionary history that differs from that reported by Staubach et al. (7) not only in terms of the number of populations involved but also by accounting for the evolutionary history of the populations involved. We consider the percentages of the genome with introgressed origin reported by Staubach et al. (7) to be overestimates, because introgression involving an ancestral population that later split into more than one extant population would be multiply reported for each extant population in the case of the study by Staubach et al. (7). On the other hand, the same percentages would be underestimated in the case where admixed populations were used in place of the nonadmixed reference populations required by HAPMIX, as Staubach et al. (7) did by using putatively introgressed mouse samples to construct the reference populations. Notably, our methodology does not require the use of nonadmixed reference populations.

We hypothesize that the more recent introgression event in Fig. 3 is due to gene flow from secondary contact, where the ranges of the two *M. musculus* subspecies overlapped, roughly at the border between Germany and the Czech Republic. The biological interpretation of the more ancient introgression event is less clear. We conjecture that the event is related to gene flow during and after subspecific divergence. Further study may provide important clues to the mechanistic basis of the evolution of subspecies in *M. musculus* and the process of speciation itself.

It is important to note that although we used a very large number of loci, there was still uncertainty in the inferred origins of the two hybridization events (as shown in Fig. 3), a similar pattern to the one observed in the simulation results and discussed above. This uncertainty is a reflection of the weak signal in these data, coupled with the low inheritance probability and short branch length between the hybridization and the MRCA of *M. m. musculus* from China and *M. m. musculus* from Kazakhstan and *M. m. domesticus* from France and *M. m. domesticus* from Germany, which is an issue that we discussed above in the context of the simulated data. The samples used are very closely related, resulting in genomes with a very small number of segregating sites, and hence a weaker signal for inference. Nonetheless, the uncertainty is localized in the sense that the potential donors of the genetic material of each hybridization event revolve around a single ancestral node. Because all five populations under analysis are closely related, most of the reconstructed gene trees were not binary, due to identical sequences of multiple alleles. Because bootstrapping is not useful in these scenarios (every locus has a handful of sites, most of which are monomorphic), we used the nonbinary gene tree topologies for the loci and considered the set of all resolutions as the set of gene tree estimates to use in Eq. 3.

Discussion

We have devised methods that enable revisiting existing evolutionary analyses and conducting new ones when both hybridization and ILS are either suspected or observed. Programs implementing all of these methods are publicly available in the open-source software package PhyloNet (23). We illustrated the power of our method in extensive simulations and demonstrated its utility on a dataset of mouse genomes. In our model, we abstract the notion of hybridization such that each reticulation edge can be viewed as a “tunnel” through which genetic material can flow repetitively and at different, yet close, times. In other words, the interpretation of a reticulation edge is not that it is a single event of mating between two individuals from two populations or species; rather, it encompasses an ongoing gene flow within a time interval that can be abstracted with one edge and one inheritance probability. This abstraction is a major difference between our model and the more detailed population
genetic models that account explicitly for rates of gene flow, such as the isolation-with-migration model. A major direction for future research is scaling up our methods to larger datasets. Currently, it takes a few seconds to a few minutes to evaluate the likelihood of a phylogenetic network with 10–20 taxa (17). This running time can vary significantly even among networks with the same numbers of taxa and reticulation events, because the shape of the gene tree and the configuration of the reticulation nodes in the network (their locations and interdependencies) are the crucial factors (27). However, optimizing the branch lengths and inheritance probabilities, coupled with the phylogenetic network search, is the bottleneck for computation. Furthermore, as our analyses, both on simulated and biological data, demonstrated, it might often be the case that several evolutionary histories have similar likelihoods. This observation calls for Bayesian approaches to inference of phylogenetic networks, whereby a distribution of networks, rather than a point estimate, is computed. In this case, modern Markov chain Monte Carlo techniques can replace the traditional hill-climbing technique we used here.

Although we discussed the model above with respect to a single population mutation rate (θ), it is generalizable in a straightforward manner to allow for different rates across the branches of the phylogenetic network if the branch-specific population size is known. Furthermore, a rate r_i can be specific for locus i or vary the recombination rates across loci (all gene tree branch lengths for locus i are multiplied by 1/r_i). Similarly, the model can naturally incorporate a single set of inheritance probabilities for the various hybridization events and allow for rate parameters, one per locus, to vary the inheritance probabilities across loci.

A major assumption underlying our models and methods is free recombination between loci and no recombination within. This assumption is common to the majority of methods and tools that focus on hybridization. Relaxing this assumption requires introducing spatial dependence in the data, similar to a method we recently introduced (28). However, this extension only makes the model more complex and significantly increases the computational requirements of the inference methods. Currently, to use such inference methods, it is assumed that independent loci are sampled and that each locus is recombination-free. If a locus contains recombination, it can be partitioned into recombination-free regions, potentially at the expense of creating regions that are too short for reliable estimation of gene trees, further emphasizing the need to account carefully for uncertainty in gene tree estimates.

Although we focused on using gene trees, the ultimate goal is to enable inference directly from sequences (Eq. 1), because such an approach uses the full signal in the data and bypasses the issue of uncertainty in gene tree estimates and the need to deal with it carefully. As discussed above, the SNP (single nucleotide polymorphism) and AFLP (amplified fragment length polymorphism) package for Phylogenetic analysis (SNAPP) method of Bryant et al. (21) enables such an inference from biallelic data in the case of phylogenetic trees (when no hybridization is allowed), even though the authors presented a Bayesian approach based on the likelihood function, rather than an ML approach. Extending the algorithms of SNAPP to allow for an ML inference based on Eq. 1 is doable, yet the application of such an extension is computationally prohibitive even for the smallest phylogenetic network (three taxa and a single reticulation), as we have observed from preliminary work.

Finally, although we varied the number of individuals sampled per species in our simulations, more thorough investigations need be conducted of the data requirements (more taxa, more loci, or more alleles) to tease apart introgression signals from those signals arising from population effects. These investigations would inform the data collection and help focus the efforts aimed at ameliorating the computational requirements. For example, in the mouse dataset we considered here, the five genomes are very closely related, giving a very weak signal for estimating gene tree branch lengths with any reasonable accuracy. In this case, the large number of loci provided a powerful signal for the network inference. The simulations, on the other hand, show that with stronger signal within the individual markers, fewer loci would be needed for accurate inferences.

ACKNOWLEDGMENTS. L.N. was supported, in part, by National Science Foundation Grants DBI-1062463 and CCF-1302179, as well as by Grant R01LM009494 from the National Library of Medicine (NLM). K.J.L. was partially supported by a training fellowship from the Keck Center of the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics (Grant T15LM007093).

Supplementary material

Maximum Likelihood Inference of Reticulate Evolutionary Histories

Y. Yu\textsuperscript{1,*}, J. Dong\textsuperscript{1}, K. Liu\textsuperscript{1,2}, and L. Nakhleh\textsuperscript{1,2,*}

\textsuperscript{1}Department of Computer Science, Rice University
\textsuperscript{2}Department of Ecology and Evolutionary Biology, Rice University
*Contact authors: \{yy9,nakhleh\}@rice.edu

Contents

1 Phylogenetic networks \hspace{1cm} 3
2 Distribution of gene tree topologies \hspace{1cm} 4
3 Distribution of gene trees with their branch lengths \hspace{1cm} 6
4 Inference of networks and inheritance probabilities \hspace{1cm} 14
  4.1 Optimizing branch lengths and inheritance probabilities of a phylogenetic network \hspace{1cm} 14
  4.1.1 Using gene tree topologies alone \hspace{1cm} 14
  4.1.2 Using both topologies and branch lengths of gene trees \hspace{1cm} 15
  4.2 Inferring an ML phylogenetic network \hspace{1cm} 16
    4.2.1 Neighborhood of a phylogenetic network \hspace{1cm} 16
  4.3 Search heuristic \hspace{1cm} 18
5 Assessing phylogenetic networks
  5.1 Information criteria ............................................. 18
  5.2 Cross-validation .................................................. 19
  5.3 Parametric bootstrap ............................................. 20

6 Simulations
  6.1 Settings for the simulations in the main text (Figure 2) ............... 22
  6.2 Other simulations .................................................. 25
    6.2.1 Results ..................................................... 27
    6.2.2 A complete simulation involving network, gene trees and sequences 28

7 Analysis of a house mouse (*Mus musculus*) data set ..................... 36

8 PhyloNet implementation and use ........................................... 41
1 Phylogenetic networks

In order to account for both hybridization and incomplete lineage sorting, we use the phylogenetic network model given in [17], which is described briefly below.

**Definition 1** A phylogenetic $\mathcal{X}$-network, or $\mathcal{X}$-network for short, $\Psi$ is a directed, acyclic graph (DAG) with $V = \{r\} \cup V_L \cup V_T \cup V_N$, where

- $\text{indeg}(r) = 0$ ($r$ is the root of $\Psi$);
- $\forall v \in V_L, \text{indeg}(v) = 1$ and $\text{outdeg}(v) = 0$ ($V_L$ are the external tree nodes, or leaves, of $\Psi$);
- $\forall v \in V_T, \text{indeg}(v) = 1$ and $\text{outdeg}(v) \geq 2$ ($V_T$ are the internal tree nodes of $\Psi$);
- $\forall v \in V_N, \text{indeg}(v) = 2$ and $\text{outdeg}(v) = 1$ ($V_N$ are the reticulation nodes of $\Psi$),

$E \subseteq V \times V$ are the network’s edges, including reticulation edges whose heads are reticulation nodes, and tree edges whose heads are tree nodes, and $\ell : V_L \rightarrow \mathcal{X}$ is the leaf-labeling function, which is a bijection from $V_L$ to $\mathcal{X}$.

We use $V(\Psi)$ and $E(\Psi)$ to denote the set of nodes and edges of phylogenetic network $\Psi$ respectively. In addition to the topology of a phylogenetic network $\Psi$, each edge $b = (u, v)$ in $E(\Psi)$ has a length $\lambda_b$ measured in coalescent units, which is the number of generations divided by effective population size on that branch. We use $\Psi$ to refer to both the topology and branch lengths of the phylogenetic network.
2 Distribution of gene tree topologies

Given a phylogenetic network $\Psi$, the gene tree topology is a random variable whose probability mass function (pmf) was given in [3] for the case where the topology of $\Psi$ is a tree, and in [36] for the case where the topology of $\Psi$ is a network. We now briefly review the pmf given in [36].

We denote by $\Psi_u$ the set of nodes that are reachable from the root of $\Psi$ via at least one path that goes through node $u \in V(\Psi)$. Then given a phylogenetic network $\Psi$ and a gene tree $G$ for some locus $j$, a coalescent history is a function $h : V(G) \rightarrow E(\Psi)$ such that the following two conditions hold:

- if $v$ is a leaf in $G$, then $h(v) = (x, y)$ where $y$ is the leaf in $\Psi$ with the label of the species from which the allele labeling leaf $v$ in $G$ is sampled;
- if $v$ is a node in $G_u$, and $h(u) = (p, q)$, then $h(v) = (x, y)$ where $y \in \Psi_q$.

In Fig. 1, we show an example of all the possible coalescent histories for a given gene tree and phylogenetic network.

Given a phylogenetic network $\Psi$ and a gene tree $G$ for locus $j$, we denote by $H_{\Psi}(G)$ the set of all coalescent histories of $G$ within the branches of $\Psi$. Then the pmf of the gene tree is given by

$$P(G|\Psi, \Gamma) = \sum_{h \in H_{\Psi}(G)} P(h|\Psi, \Gamma), \quad (1)$$

where $\Gamma$ is the inheritance probabilities matrix (see the main text) and $P(h|\Psi, \Gamma)$ gives the pmf of the coalescent history random variable, which can be computed as

$$P(h|\Psi, \Gamma) = \frac{w(h)}{d(h)} \prod_{b \in E(\Psi)} \frac{u_b(h)}{d_b(h)} \Gamma[b, j]^{u_b(h)} p_{u_b(h)v_b(h)}(\lambda_b). \quad (2)$$

In this equation, $u_b(h)$ and $v_b(h)$ denote the number of lineages enter and exit edge $b$ of $\Psi$ under coalescent history $h$. The term $p_{u_b(h)v_b(h)}(\lambda_b)$ is the probability of $u_b(h)$ gene lineages coalescing into $v_b(h)$ during time $\lambda_b$ [28]. And $w_b(h)/d_b(h)$ is the proportion of all coalescent scenarios resulting from $u_b(h) - v_b(h)$ coalescent events that agree with the topology of the gene tree [3]. This quantity without the $b$ subscript corresponds to the root.
of $\Psi$. In Table 1, we gave an example of how Eq. 2 is computed given the phylogenetic network $\Psi$ and $G$ in Fig. 1.

Recently, we proposed the first method for computing $P(G|\Psi, \Gamma)$ based on the concept of MUL-tree [36]. Basically, the phylogenetic network is first converted to a MUL-tree, and then the probability is calculated as the sum of the probabilities of observing the gene tree within the branches of the MUL-tree under all allele mappings. Later, we proposed another more efficient way of computing $P(G|\Psi, \Gamma)$ based on the concept of weighted ancestral configuration [37]. It is a bottom-up algorithm working on the network $\Psi$ directly without explicitly enumerating any coalescent history.
Table 1: The probabilities of all coalescent histories in Fig. 1. For every coalescent history $h$, columns from 2 to 7 list the probability of having $h$ on every branch of the species network $\Psi$, where $t_i$ is the branch length of branch $i$ and $g_{uv}(t_i)$ is the probability of $u$ gene lineages coalescing into $v$ within time $t_i$ [3]. Branch 6 corresponds to the branch incident into the root of the species network $\Psi$. A dash means no gene lineages enter that branch. Therefore, the total probability of a coalescent history is the product taken over all branches of the species network. In Fig. 1, coalescent events $y$ and $z$ can only happen above the root of $\Psi$. For every coalescent history, the highlighted cell shows where coalescent event $x$ happens.

<table>
<thead>
<tr>
<th>$h$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>$g_{21}(t_1)$</td>
<td>$\gamma$</td>
<td>$-$</td>
<td>$g_{22}(t_4)$</td>
<td>1</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_2$</td>
<td>$g_{21}(t_1)$</td>
<td>$-$</td>
<td>$1 - \gamma$</td>
<td>$1$</td>
<td>$g_{22}(t_5)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_3$</td>
<td>$g_{22}(t_1)$</td>
<td>$\gamma^2 g_{21}(t_2)$</td>
<td>$-$</td>
<td>$g_{22}(t_4)$</td>
<td>1</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_4$</td>
<td>$g_{22}(t_1)$</td>
<td>$-$</td>
<td>$(1 - \gamma)^2 g_{21}(t_3)$</td>
<td>$1$</td>
<td>$g_{22}(t_5)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_5$</td>
<td>$g_{22}(t_1)$</td>
<td>$\gamma^2 g_{22}(t_2)$</td>
<td>$-$</td>
<td>$\frac{1}{3} g_{32}(t_4)$</td>
<td>1</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_6$</td>
<td>$g_{22}(t_1)$</td>
<td>$-$</td>
<td>$(1 - \gamma)^2 g_{22}(t_3)$</td>
<td>$1$</td>
<td>$\frac{1}{3} g_{32}(t_5)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_7$</td>
<td>$g_{22}(t_1)$</td>
<td>$\gamma^2 g_{22}(t_2)$</td>
<td>$-$</td>
<td>$g_{33}(t_4)$</td>
<td>1</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_8$</td>
<td>$g_{22}(t_1)$</td>
<td>$-$</td>
<td>$(1 - \gamma)^2 g_{22}(t_3)$</td>
<td>$1$</td>
<td>$g_{33}(t_5)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_9$</td>
<td>$g_{22}(t_1)$</td>
<td>$\gamma$</td>
<td>$1 - \gamma$</td>
<td>$g_{22}(t_4)$</td>
<td>$g_{22}(t_5)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$h_{10}$</td>
<td>$g_{22}(t_1)$</td>
<td>$\gamma$</td>
<td>$1 - \gamma$</td>
<td>$g_{22}(t_4)$</td>
<td>$g_{22}(t_5)$</td>
<td>$\frac{1}{3}$</td>
</tr>
</tbody>
</table>

3 Distribution of gene trees with their branch lengths

Given a species tree, the probability density function (pdf) of a gene tree with branch lengths was given in [20]. Now we propose the first method for computing this pdf when the given species phylogeny is a network. We discuss in the main text the conversion between branch lengths of gene trees in units of expected numbers of mutations and branch lengths of phylogenetic networks in coalescent units. Therefore, neither the phylogenetic network nor the gene trees have to be ultrametric in our model, unless time (in both cases)
is measured in standard units (calendar time).

Given a gene tree $G$ and a species tree $\Psi$ (both given by their topologies and branch lengths), there is only one way of reconciling $G$ within the branches of $\Psi$. However, when the species phylogeny is a network, there might be more than one reconciliation due to different paths that the gene lineages can take at reticulation nodes of $\Psi$ when tracing them backwards in time.

We use $\tau_\Psi(v)$ to denote the height of node $v$ in phylogeny $\Psi$ with branch lengths $\lambda$. Given a gene tree $G$ whose branch lengths are given by $\lambda'$ and a phylogenetic network $\Psi$ whose branch lengths are given by $\lambda$, we define a coalescent history with respect to coalescence times to be a function $h : V(G) \to E(\Psi)$, such that the following condition holds:

- for $h \in H_\Psi(G)$, if $h(v) = (x, y)$ and $\tau_\Psi(x) > \tau_G(v) \geq \tau_\Psi(y)$, then $h(v) = (x, y)$.

The quantity $\tau_G(v)$ indicates at which point of branch $(x, y)$ coalescent event $v$ happens. We denote the set of coalescent histories with respect to coalescence times for gene tree $G$ and phylogenetic network $\Psi$ by $H_\Psi(G)$. Clearly, in this case, the set $H$ depends on $\lambda$ and $\lambda'$. To illustrate this, an example is shown in Fig. 2, where the same phylogenetic network and gene tree are used as the ones in Fig. 1, but with branch lengths. We can see that there are only two coalescent histories with respect to coalescence times, $h_1$ and $h_2$, resulting from different paths $b_1$ and $b_2$ took at the reticulation node. And their corresponding coalescent histories in Fig. 1 are $h_5$ and $h_6$, respectively. It is important to note that some $\lambda$ and $\lambda'$ may result in $H_\Psi(G) = \emptyset$, which means $G$ cannot be reconciled within the branches of $\Psi$ with respect to their coalescence times.

Given a phylogenetic network $\Psi$, the pdf of the gene tree (topology and branch lengths) random variable is given by

$$p(G|\Psi, \Gamma) = \sum_{h \in H_\Psi(G)} P(h|\Psi, \Gamma),$$

where $p(h|\Psi, \Gamma)$ gives the pdf of the coalescent history (with respect to coalescence times) random variable.

Let us now consider a locus $j$, whose gene tree is $G$ and an arbitrary $h \in H_\Psi(G)$. For an edge $b = (x, y) \in E(\Psi)$, we define $T_b(h)$ to be a vector of the elements in the
Figure 2: A phylogenetic network $\Psi$, a gene tree $G$, and the two possible coalescent histories with respect to coalescence times of $G$ within the branches of $\Psi$. One allele is sampled from taxa $A$ and $C$, and two alleles from taxon $B$. As shown in the figure, $\tau_1$, $\tau_2$ and $\tau_3$ are the heights of the three internal nodes of $G$, and $\eta_1$, $\eta_2$, $\eta_3$ and $\eta_4$ are the heights of four internal nodes of $\Psi$.

Set $\{\tau_G(w) : w \in h^{-1}(b)\} \cup \{\tau_\Psi(y)\}$ in increasing order. We denote by $T_b(h)[i]$ the $i$-th element of the vector. Furthermore, we denote by $u_b(h)$ the number of gene lineages entering edge $b$ and $v_b(h)$ the number of gene lineages leaving edge $b$ under $h$. Then we have

$$p(h|\Psi, \Gamma) = \prod_{b=(x,y) \in E(\Psi)} \left[ \prod_{i=1}^{\left|T_b(h)\right|-1} f_c(u_b(h) - i + 1, T_b(h)_{i+1} - T_b(h)_i) \right] \times \frac{1}{\left(u_b(h) - i + 1\right)^{2}} \times f_n(v_b(h), \tau_\Psi(x) - T_b(ht)|T_b(h)|) \times \Gamma[b, j]^{u_b(h)},$$

where $f_c(j, t)$ is the pdf of the waiting time $t$ ($t \geq 0$) for $j$ lineages to coalesce into $j - 1$ [11, 12]

$$f_c(j, t) = \binom{j}{2} e^{-\binom{j}{2}t}.$$  

Furthermore, $1/\left(u_b(h) - i + 1\right)^{2}$ is the probability of a particular pair of gene lineages among $u_b(h) - i + 1$ lineages coalescing in a manner that is consistent with the topology of $G$.  

8
Table 2: The individual terms of the pdf of all coalescent histories with respect to coalescence times in Fig. 2. For every \( h \), the six columns labeled 1—6 give the term for having the coalescence events given by \( h \) on every branch of the species network \( \Psi \). Since there is only one reticulation node and we are illustrating an arbitrary locus, we replace \( \Gamma \) by a single \( \gamma \) value for edge 2 in the network and \( 1 - \gamma \) for edge 3 (the \( \Gamma \) values for every tree edge in the network is 1). Branch 6 is the one incident into the root of the species network. A dash means no gene lineages enter that branch. The relative likelihood for the coalescent history random variable to take on the value of a specific history is the product of all the six terms in the corresponding row of that coalescent history. In Fig. 2, coalescent events \( y \) and \( z \) can only happen above the root of \( \Psi \). For every \( h \), the highlighted cell shows where coalescent event \( x \) happens.

<table>
<thead>
<tr>
<th>( h )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h_1 )</td>
<td>( e^{-\eta_4} )</td>
<td>( \gamma^2 e^{-(\eta_5 - \eta_4)} )</td>
<td>( - )</td>
<td>( 3e^{-(\gamma_3 - \gamma_5)}e^{-(\eta_1 - \gamma_3)} )</td>
<td>( 1 )</td>
<td>( 3e^{-(\gamma_2 - \eta_1)}e^{-(\eta_1 - \gamma_2)} )</td>
</tr>
<tr>
<td>( h_2 )</td>
<td>( e^{-\eta_4} )</td>
<td>( - )</td>
<td>( (1 - \gamma)^2 e^{-(\eta_2 - \eta_4)} )</td>
<td>( 1 )</td>
<td>( 3e^{-(\gamma_2 - \eta_1)}e^{-(\eta_1 - \gamma_2)} )</td>
<td>( 3e^{-(\gamma_2 - \eta_1)}e^{-(\eta_1 - \gamma_2)} )</td>
</tr>
</tbody>
</table>

And \( f_n(j, t) \) is the probability of no coalescent events happening among \( j \) gene lineages for time \( t \) which can be computed as [11, 12]

\[
f_n(j, t) = e^{-\frac{j}{2}t}
\]

(6)

After substituting Eq. 5 and Eq. 6 into Eq. 4, we have

\[
p(h|\Psi, \Gamma) = \prod_{b \in E(\Psi)} \left[ \prod_{i=1}^{[T_b(h)]-1} e^{-\left(\frac{u_b(h)(i+1)}{2}\right)(T_b(h)_{i+1} - T_b(h)_i)} \right] \times e^{-\left(\frac{\tau_{\Psi}(b) - T_b(h)}{2}\right)} \times \Gamma[b, j]^{u_b(h)}. \]

(7)

In Table 2, we give an example of how Eq. 7 is computed given the phylogenetic network \( \Psi \) and gene tree \( G \) in Fig. 2.

We can use the same technique in [36] but with Eq. 3 to calculate \( p(G|\Psi, \Gamma) \). Basically, we first convert the phylogenetic network to a MUL-tree. Then under every allele mapping, we compute the set of coalescent histories with respect to coalescence times and use Eq. 7 to compute the probability of every coalescent history. Note that as in [36] special attention needs to be paid to the sets of edges in the MUL-tree that come from the same edge in the original network.
Additionally, we can also compute \( p(G|\Psi, \Gamma) \) based on weighted ancestral configurations, which is faster than computing it based on MUL-trees. The main idea is similar to the method we proposed in [37], which was built on the work of [32] for the case of phylogenetic trees. We now describe this method in detail.

We first describe briefly the concept of weighted ancestral configuration (AC, or simply configuration) we introduced in [37]. An ancestral configuration can be associated with a node \( v \) of \( \Psi \), denoted by \( AC_v \), or an edge \( b \) of \( \Psi \), denoted by \( AC_b \). It is a triplet \((B, a, w)\) interpreted as follows:

- **B**: a set of lineages that exist at the point with which the AC is associated.
- **a\([i]\)**, for \( 1 \leq i \leq |V_N| \): an index for the AC split that occurred at reticulation node \( i \) and gave rise to \( B \).
- **w**: weight of the AC which is the probability of observing \( B \) at the point with which the AC is associated.

Given two ACs, \( AC_1 = (B_1, a_1, w_1) \) and \( AC_2 = (B_2, a_2, w_2) \), if \( B_1 = B_2 \) and \( a_1 = a_2 \), we say that \( AC_1 \) and \( AC_2 \) are identical. Furthermore, we define two ACs to be compatible if for each \( i, 1 \leq i \leq q \), either \( a_1[i] = a_2[i] \) or \( a_1[i] \cdot a_2[i] = 0 \).

Basically, we traverse the nodes of \( \Psi \) in post-order and build a set of ancestral configurations for every node we visit. We denote a set of ACs of node \( v \) by \( AC_v \), and a set of ACs of edge \( b \) which are about to leave \( b \) by \( AC_b \). After \( AC \) is built for the root of the network, we could obtain \( p(G|\Psi, \Gamma) \). There are three important operations during this process:

- **Merging ACs** whenever an internal tree node is encountered. Let \( u \) be an internal tree node with two child nodes \( x_1 \) and \( x_2 \). To construct \( AC_u \), every compatible AC pair \((AC_1, AC_2)\) where \( AC_1 = (B_1, a_1, w_1) \in AC_{(u,x_1)} \) and \( AC_2 = (B_2, a_2, w_2) \in AC_{(u,x_2)} \) are merged into one AC \((B, a, w)\) where \( B = B_1 \cup B_2 \), \( a[i] = \max\{a_1[i], a_2[i]\} \) for all \( 1 \leq i \leq q \) and \( w = w_1 \cdot w_2 \).

- **Splitting ACs** whenever an reticulation node is encountered. Let \( u \) be the \( k_{th} \) reticulation node whose two parent nodes are \( y_1 \) and \( y_2 \). For each reticulation node \( i \) of
\(N\), we have a counter \(o\) initialized to 0. Let \((B, a, w)\) be an AC in \(\mathcal{AC}_u\). Then we split \(B\) into all possible ordered pairs \((B_1, B_2)\), such that \(B_1 \cup B_2 = B\) and \(B_1 \cap B_2 = \emptyset\). For each pair, we make \(AC_1 = (B_1, a_1, w)\) and \(AC_2 = (B_2, a_2, 1)\), where \(a_1 = a_2 = a\) except for \(a_1[k] = a_2[k] = o_k + 1\) and \(o_k\) is incremented by 1. \(AC_1\) then is a configuration about to go along edge \((y_1, u)\), and \(AC_2\) is a configuration about to go along edge \((y_2, u)\).

- Coalescing ACs along an edge. We define a function called \textbf{CoalACs} which takes a gene tree \(G\), an edge \((x, y)\) of \(\Psi\) and a set of configurations \(\mathcal{AC}_y\) that enter edge \((x, y)\), and returns a set of configurations \(\mathcal{AC}_{(x, y)}\) that leave edge \((x, y)\). See Alg. 1 for details.

The algorithm for computing \(p(G|\Psi, \Gamma)\) is shown in Alg. 2. Further, we can use the same technique we introduced in [37] to reduce the number of configurations at articulation nodes of the network.
Algorithm 1: CoalACs.

**Input:** a gene tree $G$, an edge $(x, y) \in E(\Psi)$, a set of ACs $\mathcal{AC}_y$

**Output:** a set of ACs $\mathcal{AC}_{(x, y)}$

Let $V(G)$ be the set of internal nodes of $G$ ordered by their heights in increasing order;

$\mathcal{AC}_{(x, y)} \leftarrow \emptyset$;

foreach $(B, a, w) \in AC_y$ do

$t \leftarrow \tau_{\Psi}(y)$;

$B^+ \leftarrow B$;

$p \leftarrow \lambda_{(x, y)}^{[B]}$;

foreach $v \in V(G)$ do

if $\tau_{\Psi}(y) \leq \tau_G(v) < \tau_{\Psi}(x)$ then

Let $L_v$ be the set of taxa under node $v$ in $G$;

Let $L_B$ be the set of taxa that coalesce into $B$;

if $L_v \subseteq L_B$ then

$p \leftarrow p \cdot e^{-(\frac{|B^+|}{2})(\tau_G(v) - t)}$;

$t \leftarrow \tau_G(v)$;

Apply the coalescent event represented by $v$ to $B^+$ and the resulting $B^+$ contains one less lineages;

else if $L_v \cap L_B \neq \emptyset$ then

$p \leftarrow 0$;

Break;

if $p \neq 0$ then

if $|B^+| \neq 1$ then

$p \leftarrow p \cdot e^{-(\frac{|B^+|}{2})(\tau_{\Psi}(x) - t)}$;

$\mathcal{AC}_{(x, y)} \leftarrow \mathcal{AC}_{(x, y)} \cup (B^+, a, w, p)$;

return $\mathcal{AC}_{(x, y)}$;
Algorithm 2: CalProb.

**Input:** Phylogenetic network $\Psi$, gene tree $G$, and inheritance probabilities matrix $\Gamma$

**Output:** $p(G|\Psi, \Gamma)$

while traversing the nodes of $\Psi$ in post-order do

if node $v$ is a leaf, whose parent is $u$ then

$\mathcal{A}_v \leftarrow \{(B, a, 1)\}$ where $B$ is the set of leaves in $G$ sampled from the species associated with $v$ and $a$ is a vector of $q$ 0's;

$\mathcal{A}_{(u,v)} \leftarrow \text{CoalACs}(G, (u, v), \mathcal{A}_v)$;

else if node $v$ is a reticulation node, who has child $w$, and two parents $u_1$ and $u_2$ then

$\mathcal{A}_v \leftarrow \mathcal{A}_{(v,w)}$;

$S_1 \leftarrow \emptyset$;

$S_2 \leftarrow \emptyset$;

foreach $AC \in \mathcal{A}_v$ do

Split $AC$ in every possible way into pairs of ACs, and for each pair, add one to $S_1$ and the other to $S_2$;

$\mathcal{A}_{(u_1,v)} \leftarrow \text{CoalACs}(G, (u_1, v), S_1)$;

$\mathcal{A}_{(u_2,v)} \leftarrow \text{CoalACs}(G, (u_2, v), S_2)$;

else if node $v$ is an internal tree node, who has two children $w_1$ and $w_2$ then

foreach pair $(AC_1, AC_2)$ of compatible ACs in $\mathcal{A}_{(v,w_1)} \times \mathcal{A}_{(v,w_2)}$ do

Merge $AC_1$ and $AC_2$ and add the resulting AC to $\mathcal{A}_v$;

if node $v$ is an internal tree node, and its parent is $u$ then

$\mathcal{A}_{(u,v)} \leftarrow \text{CoalACs}(G, (u, v), \mathcal{A}_v)$;

else

Create a dummy node $r'$ with height $+\infty$;

$\mathcal{A}_{(r',v)} \leftarrow \text{CoalACs}(G, (r', v), \mathcal{A}_v)$;

return $\sum_{(B,a,w) \in \mathcal{A}_{(r',v)}} w$;
4 Inference of networks and inheritance probabilities

The search consists of (1) optimizing a candidate network’s branch lengths and inheritance probabilities, and (2) searching the network topologies space. We assume here that all loci share the same inheritance probability (denoted by $\gamma$) for a given branch in the phylogenetic network. Extending this to allow for varying these probabilities across loci is straightforward (of course, while increasing the running time).

4.1 Optimizing branch lengths and inheritance probabilities of a phylogenetic network

In this section, we describe our approach for optimizing branch lengths $\lambda^*$ and inheritance probabilities $\gamma^*$ for a fixed network topology $\Psi$ ($\lambda^*$ is part of $\Psi$ here), given a set $G$ of gene trees, in order to maximize $p(G|\Psi, \gamma)$. We discuss separately the cases of using gene tree topologies alone and using gene tree topologies and branch lengths.

4.1.1 Using gene tree topologies alone

A heuristic for finding the optimal branch lengths for a fixed species tree topology was introduced in [32]. Here, we are using the same method but in our case of phylogenetic networks we are optimizing not only branch lengths but also inheritance probabilities. In particular, an initial value of likelihood is first calculated with every branch length initialized to be 1.0 and inheritance probability initialized to be 0.5. Then the elements in $[\lambda, \gamma]$ are optimized one by one separately using Brent’s method [2]. More specifically, while Brent’s method is varying the value of one element in $[\lambda, \gamma]$ in order to find a local optimum, the values of all other elements are fixed. After the local optimum is found, the element is replaced by this new value and then Brent’s method moves to the next element for optimization. Updating all elements in $[\lambda, \gamma]$ once is called a round. After each round of optimization, we compare the likelihood of the network with the newly updates branch lengths and inheritance probabilities to the likelihood from the previous round. If the improvement is smaller than some pre-specified threshold or some pre-specified maximum
number of rounds is reached, then the new branch lengths and inheritance probabilities are declared to be optimal and the optimization process terminates. This process is repeated multiple times to handle the issue of local optima. All parameters used in this optimization process, including those for Brent’s method, are listed in Section 8.

4.1.2 Using both topologies and branch lengths of gene trees

As we mentioned, the set of coalescent histories of a gene tree $G$ within the branches of a phylogenetic network $\Psi$ does not change with the branch lengths or inheritance probabilities if only the topology of gene tree $G$ is considered. However, this is not the case if both topologies and branch lengths of the gene trees need to be taken into account. The main reason for this is that the coalescence times in the gene tree provide constraints on where coalescence events could take place within the branches of the phylogenetic network. Further, it is important here to note that the time units in the gene tree and those in the species network must be matched for our method below to work. Consider a branch $b$ with length $\lambda_b$ (in coalescent units) in a phylogenetic network $\Psi$. Now, consider a branch $d$ with length $\tau_d$ (in units of expected number of mutations) in a gene tree $G$. The length of branch $d$ in coalescent units is

$$\tau_d \times \frac{2}{\theta}$$

where $\theta = 4N\mu$ is the population mutation rate.

In our discussion below, as well as in our implementation, we assume that $\theta$ is the same across all loci and all branches, and that the population size and generation time are the same across all branches. This implies that both the network and gene tree must be ultrametric. However, it is important to notice that removing these assumptions does not affect the model, but rather increase the running time of the method as more parameters require optimization.

In order to guarantee the ultrametricity requirement, instead of optimizing branch lengths of phylogenetic network $\Psi$ and inheritance probabilities, we optimize the height of every internal node of $\Psi$ and inheritance probabilities. Second, in order to ensure that the resulting phylogenetic network allows for embedding the gene trees in the input, we use the
coalescence times from the gene trees to compute upper bounds on the heights of nodes in \( \Psi \), and use these upper bounds to constrain the search for height values. Then the iterative process for optimization itself is similar to what we described in Section 4.1.1. The full details of the optimization procedure are available in open source in the software package PhyloNet [30].

4.2 Inferring an ML phylogenetic network

For inferring an ML phylogenetic network, we couple the optimization procedures of the previous section with a procedure for traversing the phylogenetic network space, which we now describe.

4.2.1 Neighborhood of a phylogenetic network

For a fixed number of taxa \( n \), the space of phylogenetic networks, denoted by \( \Omega(n) \), consists of an infinite set of non-overlapping subspaces, each of which contains phylogenetic networks that have the same number of reticulation nodes. We denote each subspace of \( \Omega(n) \) by \( \Omega(n, k) \), where \( k \) is the number of reticulation nodes. In particular, \( \Omega(n, 0) \) is the tree space.

Given a phylogenetic network \( \Psi \in \Omega(n, k) \), we define four types of operations for network rearrangement as follows.

- Adding a reticulation edge (\( \delta_1 \)):
  1. Let \((u_1, v_1)\) and \((u_2, v_2)\) be two distinct edges in \( \Psi \) such that \( v_2 \) is not a predecessor of \( u_1 \).
  2. Delete the two edges \((u_1, v_1)\) and \((u_2, v_2)\).
  3. Add two new nodes \( x_1 \) and \( x_2 \) and five new edges \((u_1, x_1)\), \((x_1, v_1)\), \((u_2, x_2)\), \((x_2, v_2)\), and \((x_1, x_2)\) to network \( \Psi \).

- Removing a reticulation edge (\( \delta_2 \)):
  1. Let \((u, v)\) be an edge in \( \Psi \) such that \( v \) is a reticulation node and \( u \) is a tree node.
2. Delete the two nodes $u$ and $v$ and the five edges $(w, u), (u, z), (u, v), (x, v)$ and $(v, y)$, where $w$ is the parent node of $u$, $z$ is the child node of $u$ other than $v$, $x$ is the parent node of $v$ other than $u$, and $y$ is the child node of $v$.

3. Add two new edges $(w, z)$ and $(x, y)$ to network $\Psi$.

- Relocating the destination of a reticulation edge ($\delta_3$):

  1. Let $(u_1, v_1)$ and $(u_2, v_2)$ be two distinct edges in $\Psi$ such that $v_1$ is a reticulation node and $v_2$ is not a predecessor of $u_1$.
  2. Delete node $v_1$ and the four edges $(u_1, v_1), (u_2, v_1), (w, v_1),$ and $(v_1, z)$, where $w$ is the parent node of $v_1$ other than $u_1$ and $z$ is the child node of $v_1$.
  3. Add a new nodes $x$ and four new edges $(u_2, x), (x, v_2), (u_1, x),$ and $(w, z)$ to network $\Psi$.

- Relocating the source of an edge ($\delta_4$):

  1. Let $(u_1, v_1)$ and $(u_2, v_2)$ be two distinct edges in $\Psi$ such that $u_1$ is neither a reticulation node nor a predecessor of $v_2$.
  2. Delete node $u_1$ and the four edges $(u_1, v_1), (u_2, v_2), (w, u_1),$ and $(u_1, z)$, where $w$ is the parent node of $u_1$ and $z$ is a child node of $u_1$ other than $v_1$.
  3. Add a new nodes $x$ and four new edges $(u_2, x), (x, v_2), (x, v_1),$ and $(w, z)$ to network $\Psi$.

We denote the set of phylogenetic networks that can be obtained by applying operation $\delta_i$ to $\Psi$ by $\delta_i(\Psi)$, where $1 \leq i \leq 4$. Clearly, $\Psi' \in \Omega(n, k + 1)$ if $\Psi' \in \delta_1(\Psi)$, $\Psi' \in \Omega(n, k)$ if $\Psi' \in \delta_3(\Psi)$ or $\Psi' \in \delta_4(\Psi)$ and $\Psi' \in \Omega(n, k - 1)$ if $\Psi' \in \delta_2(\Psi)$. Finally, we define the neighborhood of a phylogenetic network $\Psi$, denoted by $\Delta(\Psi)$, to be $\bigcup_{1 \leq i \leq 4} \delta_i(\Psi)$. So a phylogenetic network $\Psi'$ is a neighbor of $\Psi$ if $\Psi'$ can be obtained by applying any operation defined above to $\Psi$. 
4.3 Search heuristic

We employ a hill-climbing heuristic to search the network space in order to find an optimal phylogenetic network $\Psi$ from a set $G$ of gene trees. Starting from some network, the search proceeds by sampling networks from the neighborhood of the current network, optimizing its branch lengths as well as the inheritance probabilities, and accepting the proposed network if its likelihood improves upon the current one. The process terminates if no neighboring network improves upon the current one (our implementation allows for pre-specifying a number of failed neighbor proposals, since the number of neighbors can be very large for large numbers of leaves).

For the starting network, it is reasonable to start the search from some species trees, e.g., the set of all binary resolutions of majority consensus of the input gene trees, or the optimal species tree under the MDC criterion [13, 29, 38, 39]. For moving from a current network, a neighbor can be generated by applying one of the four types of operations of network rearrangement we defined in the Section 4.2.1. We associate each of these four operations a weight. When we propose a random neighbor of a network, the type of operation to be applied to generate the neighbor is first randomly selected according to their weights and the edges involved in that operation are then randomly chosen. The entire search process is repeated multiple times to handle the problem of local optima.

An illustration of the search is given in Fig. 3.

5 Assessing phylogenetic networks

5.1 Information criteria

The Akaike Information Criterion [1] (AIC) is defined as follows for a phylogenetic network:

$$AIC = 2k - 2 \ln L$$

where $k$ is the number of free parameters which includes both branch lengths and inheritance probabilities, and $L$ is the likelihood of the network.
Layer $\Omega(n,k)$ contains all $n$-taxon phylogenetic networks with $k$ reticulation nodes. Four simple transformations enable searching this entire space and guarantee reachability of any point from any other point (see main text and SI). These operations allow for the search to proceed within a given layer, ascend a layer, or descend a layer. Weighting, or assigning rates to the transformations, allows for controlling the behavior of the search (e.g., never descend a layer, or ascend a layer with very low probability). When searching for optimal networks, multiple searches can be initiated from different starting points; these searches may or may not converge onto a single optimal point estimate. Further, the number and features of local optima vary from one layer to another.

The Bayes Information Criterion [24] (BIC) is defined as follows for a phylogenetic network:

$$BIC = -2 \ln L + k \ln n$$

where $k$ and $L$ are defined as in AIC, and $n$ is the number of gene trees in the set.

5.2 Cross-validation

Cross-validation is another model validation technique that assesses how well the model fits a data set. $K$-fold cross-validation partitions a data set into $K$ equal-size subsets. It uses one set, the training set, which consists of $K - 1$ subsets, to infer the model parameters and use the remaining subset, the validation set, to assess prediction. The difference between predictions and the real data in the validation set can be computed. To reduce variability, $K$
rounds of cross-validations are performed using different partitions, and the differences are averaged over the number of rounds. If there are multiple models, the one with the lowest average difference is the most appropriate model. In our case, for each distinct gene tree topology in the validation subset, we compute its frequency, as well as its probably under the learned network. The difference between these two values is taken to reflect the quality of the model.

5.3 Parametric bootstrap

With the increasing interest in reconstruction of phylogenetic trees, in order to evaluate how confident one should be in a reconstructed phylogeny, bootstrapping has been widely used for decades since it was first proposed as a method for obtaining confidence limits on phylogenies [5]. Here, we employ parametric bootstrap assess support for the branches in an inferred phylogenetic network (illustrated in Fig. 4).

The idea is as follows. An inferred phylogenetic network $\Psi$ is used to generate $k$ sets of gene trees independently, each of which has the same size as the number of loci in the input (in PhyloNet, the default value of $k$ is 100). Then from each simulated set of gene trees, a phylogenetic network is inferred using the same method and settings as the one used to obtain the original phylogenetic network $\Psi$ from gene trees $\mathcal{G}$. Finally, by comparing the $k$ inferred phylogenetic networks with $\Psi$, we obtain the support of every edge in $\Psi$.

This parametric bootstrap works for both inference from gene trees with and without branch lengths. In PhyloNet, for ML inference using only the topologies of gene trees, we implemented our own simulator to generate topologies of gene trees from a given phylogenetic network. And for ML inference using both the topologies and branch lengths of gene trees, we called an external software ms [6] to simulate gene trees with branch lengths.

The bootstrap value of a branch in the inferred phylogenetic network $\Psi$ is calculated as the proportion of networks in $\Psi_1, \ldots, \Psi_k$ that contain the same branch. We say that branch $b_1$ in network $\Psi_1$ and branch $b_2$ in network $\Psi_2$ are the same if they satisfy the following two conditions:

- $b_1$ and $b_2$ induce the same set of softwired clusters [9],
Figure 4: Illustration of parametric bootstrap to assess support for an inferred phylogenetic network’s branches.

- $b_1$ and $b_2$ are either both tree edges or both reticulation edges.
6 Simulations

6.1 Settings for the simulations in the main text (Figure 2)

**Phylogenetic network.** We used the phylogenetic network shown in Fig. 5. It captures a scenario where there is divergence followed by a hybridization with inheritance probability being 0.1 as shown in the figure.

![Phylogenetic Network](image)

**Figure 5:** A phylogenetic network used in the simulations reported in the main text. Branch lengths are in coalescent units and the inheritance probability is 0.1 for the reticulation edge \((i_3, i_2)\).

**True gene trees.** Gene trees were simulated using software ms [6]. See command below. We varied the number of loci by \(\text{loci} = \{10, 20, 40, 80, 160\}\), for each of which we generate 30 sets of gene trees.

\[
\text{ms} 4 \text{ loci} -T -I 4 1 1 1 1 -ej 0.5 4 3 -es 1.0 3 0.5 -ej 1.0 2 5 -ej 1.5 5 1 -ej 2.0 3 1
\]

**Sequences.** Sequences were generated using Seq-gen [19] under the GTR model. We used a population mutation rate of \(\theta = 0.036\). More specifically, for gene trees contained in file \(gtFile\), the following command was used:

\[
\text{seq-gen} -mGTR -s0.018 -fbaseFreq -rrates -lseqLen < gtFile
\]
where $\text{baseFreq} = 0.300414, 0.191363, 0.196748, 0.311475$ are the base frequencies of the nucleotides A, C, G and T, and $\text{rates} = 1.24284, 3.47484, 0.48667, 1.07118, 4.38510, 1.0$ are the relative rates of substitutions. We also varied the length of the sequences through $\text{seqLen} = \{250, 500, 1000\}$.

**Estimated gene trees.** Gene trees were estimated using PAUP* [27] under maximum likelihood. For each sequence alignment, we randomly generated 100 bootstrap replicates. And for each of them, we used the following commands in PAUP* to reconstruct an ultrametric gene tree:

```plaintext
execute seqFile;

nj;
lscore 1/tratio=estimate nst=6 rmatrix=estimate;
set criterion=likelihood;
lsset tratio=estimate nst=6 rmatrix=estimate clock=yes;
hsearch addseq=asis;
```

where `seqFile` is a NEXUS file that contains the sequence alignment. All branch lengths of the reconstructed gene tree were then multiplied by $2/\theta$ to convert them into coalescent units.

**Experiments.** We inferred the optimal networks from (i) true gene tree topologies, (ii) estimated gene tree topologies, (iii) true gene tree topologies and branch lengths, and (iv) estimated gene tree topologies and branch lengths. Default settings were used for inference (See Table. 4). See main text for results.

Furthermore, to study the effect of branch lengths of the phylogenetic network and the inheritance probability on the performance of the method, we investigated two cases. More specifically, assuming the network in Fig. 5 is $\Psi_1$, we considered two networks $\Psi_2$ and $\Psi_3$, where $\Psi_2$ is obtained by doubling the lengths of the internal branches $(i_5, i_4)$, $(i_5, i_2)$, $(i_4, i_3)$ and $(i_2, i_1)$ of $\Psi_1$ and not changing the inheritance probability, and $\Psi_3$ is obtained by changing the inheritance probability of $\Psi_1$ from 0.1 to 0.5 and keeping the
branch lengths unmodified. Then from $\Psi_2$ and $\Psi_3$ we generated gene trees using the same settings as above. Finally, we used the topologies of those gene trees to infer networks. The results are shown in Fig. 6. We can see that, as expected, increasing branch lengths and increasing inheritance probability result in improved accuracy of the method.

Figure 6: The number of correct inferences using true gene tree topologies simulated from species network $\Psi_1$, $\Psi_2$ and $\Psi_3$. The blue, green and red bars are the results for $\Psi_1$, $\Psi_2$ and $\Psi_3$, respectively.

Varying the number of individuals. We have also varied the number of individuals (lineages) sampled from each of taxa C and D and considered 1, 2, and 4 lineages (for each of the two taxa). For each case, we used the true gene tree topologies and true gene tree topologies and branch lengths to infer the networks. The accuracy results are shown in Fig. 7. The results show that increasing the number of individuals sampled from each of C and D improves the accuracy of the method, as expected (except for the case when using the gene tree topologies alone on 10 loci). Doubling the number of loci used in the inference results in a bigger improvement in accuracy in general than doubling the number of alleles (there are a few exceptions to this that can be seen in Fig 7). However, it is important to caution here that these results are obtained from relatively small networks (4 taxa) and using true gene trees. As the number of taxa increases, the number of reticulations increases, and the branch lengths get shorter, it would be expected that sampling more
individuals would show more significant gains in accuracy.

Figure 7: Accuracy of the method on simulated data with varying alleles. The data were generated down the phylogenetic network $\Psi_1$ (Fig. 5 and also shown in Fig. 2A in the main text). Results based on true gene tree topologies and true gene tree topologies and branch lengths are shown in the left and right panels, respectively. For every number of loci, the bars from left to right correspond to cases of 1, 2, and 4 individuals sampled from each of the two taxa C and D, respectively. The dark blue, cyan, and yellow regions correspond to the number of times each of the networks $\Psi_1$, $\Psi_2$, and $\Psi_3$, respectively (Fig.2A in the main text), were inferred. The maroon region corresponds to the number of times any other network with a single reticulation was inferred.

6.2 Other simulations

We also conducted other simulations under scenarios that are “easier” for the inference method (longer sequences, longer branch lengths, higher inheritance probabilities, and no speciation following hybridization).

The simulations make use of several tools and programs:

- PhyloNet [31], which has implementation of all our methods.
- Hybrid-Lambda [40] which simulates the evolution of gene trees within the branches of a phylogenetic network under the coalescent model.
- Seq-gen [19] which simulates the evolution of DNA sequences down a given (gene) tree.

- Fasttree [15, 16], which infers a maximum likelihood phylogenetic tree from a sequence alignment.

Using these tools, we conducted three types of simulations (Fig. 8).

Figure 8: Simulation Flow Chart. Hybrid-Lambda [40] is used to simulate gene trees within a network. Seq-gen [19] is used to simulate the evolution of DNA sequences down gene trees. Fasttree [15, 16] is used to estimated gene trees from sequence alignments. PhyloNet [31] is used to infer phylogenetic networks.

In each simulation run, we varied the number of gene trees: 10, 50, 100 and 500. We also consider multiple scenarios of phylogenetic network topologies, branch lengths, and inheritance probabilities, as shown in Fig. 9. For each setting, we conducted 30 runs and averaged the results.
Figure 9: Models used in the simulation. These models are ultrametric networks, required by Hybrid-Lambda.

6.2.1 Results

The plots are arranged as follows. First we show the six models we used for the short flow simulation. Figures 9a, 9b, 9c and 9d have one reticulation node with four, five, six, and seven taxa respectively. Figures 9e and 9f has two reticulation nodes with five and six taxa respectively.
respectively.

Although limited and empirical in nature, some observations are made from the plots.

- In the topology only case, cross-validation captures much better the correct reticulation nodes than both BIC and AIC. More trees help. But even with very few trees, cross validation is still superior to the other two. It shows that cross-validation is an effective way to determine the proper number of reticulation events.

- In the case of gene trees with branch lengths, AIC and BIC perform similarly when there is a single reticulation node. BIC does a better job than AIC when the number of reticulations is two.

- The cluster distance between the original network and the inferred network with correct network nodes is always smaller than that between the original network and the inferred network with incorrect reticulations when the number of gene trees grows larger.

- Using gene tree branch lengths results in better estimates of the inheritance probabilities. When the number of trees increases, the estimates improve. When using gene tree topologies alone, the inheritance probability estimates are slightly less accurate.

### 6.2.2 A complete simulation involving network, gene trees and sequences

We use a network as shown in Figure 16 to generate gene trees and nucleotide sequences. Then the process is inverted to use these sequences to estimate gene trees and network. Several public domain software are used as well as PhyloNet. Hybrid-Lambda [40] uses the network to generate gene trees which form the input of Seq-gen [19] to produce sequences. For each gene tree, only a sequence is generated. These sequences are fed into Fasttree [15, 16] to generate unrooted rooted gene trees. PhyloNet reroots these gene trees via the outgroup and later infers the network.

The gene trees are organized the same way as in the simulation section. The Hybrid-Lambda parameter settings used to generate gene trees from the network are the same as well. The Seq-gen settings used to generate sequences from gene trees is
Figure 10: Results for Model 1. Left column: Input consists of true gene tree topologies alone. Right column: Input consists of true gene tree topologies and branch lengths. Top row: inferred number of hybridizations. Middle row: Distance between the true network and inferred network. Bottom row: Inferred inheritance probabilities (averages over 30 runs with standard deviation bars shown).
Figure 11: Results for Model 2. Left column: Input consists of true gene tree topologies alone. Right column: Input consists of true gene tree topologies and branch lengths. Top row: inferred number of hybridizations. Middle row: Distance between the true network and inferred network. Bottom row: Inferred inheritance probabilities (averages over 30 runs with standard deviation bars shown).
Figure 12: Results for Model 3. Left column: Input consists of true gene tree topologies alone. Right column: Input consists of true gene tree topologies and branch lengths. Top row: inferred number of hybridizations. Middle row: Distance between the true network and inferred network. Bottom row: Inferred inheritance probabilities (averages over 30 runs with standard deviation bars shown).
Figure 13: Results for Model 4. Left column: Input consists of true gene tree topologies alone. Right column: Input consists of true gene tree topologies and branch lengths. Top row: inferred number of hybridizations. Middle row: Distance between the true network and inferred network. Bottom row: Inferred inheritance probabilities (averages over 30 runs with standard deviation bars shown).
Figure 14: Results for Model 5. Left column: Input consists of true gene tree topologies alone. Right column: Input consists of true gene tree topologies and branch lengths. Top row: inferred number of hybridizations. Middle row: Distance between the true network and inferred network. Bottom row: Inferred inheritance probabilities (averages over 30 runs with standard deviation bars shown).
Figure 15: Results for Model 6. Left column: Input consists of true gene tree topologies alone. Right column: Input consists of true gene tree topologies and branch lengths. Top row: inferred number of hybridizations. Middle row: Distance between the true network and inferred network. Bottom row: Inferred inheritance probabilities (averages over 30 runs with standard deviation bars shown).
For nucleotide substitution model, we used HKY. By setting the nucleotide frequencies equal (with the -fe option) and the transition transversion ratio to 0.5 (with the -t0.5 option), it becomes JC69 as a special case of HKY. The length of character sequence is 5000. The scale factor of 0.0005 equals the expected number of substitutions per site per coalescent unit for each branch is 0.0005. By multiplying with the branch length, we find the expected number of substitutions per site for each branch. The output uses the PHYLIP format.

The Fasttree settings used to generate unrooted gene trees from sequences.

FastTree -nt < fasttreeInputFileName > fasttreeOutputFileName

Fasttree use the Jukes-Cantor + CAT model. The -nt option shows that it works with nucleotide sequences.

We compared the gene trees that are used to generate the sequences and those estimated from sequences. In this model, all the gene trees have the outgroup directly connected to the root. The average Robinson-Foulds distance is 0.32 while its weighted version is 0.081. The standard deviation for RF distance is 0.53 and its weighted version is 0.13. There are totally 19800 gene trees on each side. Among them there are 14060 pairs are exactly the same (71%). If we increase the scale factor or increase the sequence length, the same pair rate will be higher.

When PhyloNet infers the network, the default maximum length of a network branches is 6. Since there are some large branch length values in the models, we assigned the maximum branch length to be 20 for all the models tested in this paper.
7 Analysis of a house mouse (*Mus musculus*) data set

Using our method of inferring a phylogenetic network, we analyzed a house mouse (*Mus musculus*) data set.

*M. musculus* samples. Our *Mus musculus domesticus* samples were provided by the previous study of [26] and represent one population from France (in the Massif Central) and another population from Germany (in the vicinity surrounding Cologne and Bonn).
Mus musculus musculus samples in our study also came from previous studies [26, 4, 35] and represent a population in Czechoslovakia (Studenec) [26], another population in Kazakhstan (Almaty) [26], and a third population from China (Urumqi in Xinjiang Province) [4, 35]. For simplicity, these five populations will be referred to as DF, DG, MZ, MK and MC, respectively.

Sequence data. Genome-wide sequence data for our samples was produced using the Mouse Diversity Array [34]. We called genotypes from raw intensity values for the Chinese M. m. musculus samples using the procedure described in [4]; genotypes for all other M. musculus samples were provided by [26]. Since our computational pipeline was constructed to analyze substitution-based variation, we filtered loci exhibiting short indel or structural variation found in previously reported whole-genome sequencing of M. musculus strains (including wild-derived M. m. domesticus and M. m. musculus strains) [10, 33]. We used the most recent M. musculus reference genome coordinates as of this writing (version GRCm38.p2) throughout our study.

The reference Rattus norvegicus genome (version RGSC Rnor_5.0) was used as an outgroup in our analyses. Orthology between the R. norvegicus genome and M. musculus was determined using the BLASTZ-produced [23] pairwise genome alignment provided by the UCSC Genome Browser [14].

In total, 387,923 loci from the M. musculus genome were sampled in our data sets.

Local phylogeny estimation. Genotypes were phased into haplotypes and missing bases were imputed using fastPHASE [22]. A larger superset of 416 M. musculus samples from the studies of [35, 4, 26] were used for this purpose.

We then estimated local phylogenies along haplotype sequences using a custom analytical pipeline. To satisfy the assumption of no intralocus recombination required by our new method, breakpoints inducing recombination-free intervals were inferred on haplotypes using the Four-Gamete Test [7]. We inferred phylogenies between breakpoints using the maximum likelihood method implemented in [18]. The maximum likelihood phylogenetic analysis used the General Time Reversible substitution model [21] with the CAT model.
of rate variation across sites [25]. Local phylogenies were rooted using *R. norvegicus* as an outgroup. To satisfy the assumption of free recombination between loci, as required by our new method, local phylogenies were sampled at 100 kb intervals so that linkage disequilibrium was negligible [26]. In total, 20639 local phylogenies were reconstructed.

**Species phylogeny inference.** From the reconstructed gene trees, we inferred the optimal phylogenetic networks with 0, 1, 2 and 3 reticulation nodes, respectively, using our method described in Section 2, 4.1 and 4.2 (only topologies of gene trees were used). For each of them, the search was run 50 times and top 5 networks were saved. All other parameters were set to their default values as listed in Section 8.

Since all five populations under analysis are closely related, most of the reconstructed gene trees were not binary due to identical sequences of multiple alleles. As bootstrap is not doable in this case, given the very short sequences for each locus and the low signal, we treated uncertainty differently. Consider a non-binary gene tree *G* that is inferred for some locus. Then, we use the following equation for the probability of *G*:

\[
P(G|\Psi, \Gamma) = \sum_{g' \in b(G)} P(g'|\Psi, \Gamma),
\]

where \(b(G)\) is the set of all binary resolutions of *G*. We then used this term in the likelihood formulation based on gene tree topologies.

The results are shown in Fig. 17. Furthermore, to account for model complexity, we calculated the values of three information criteria, AIC, AICc and BIC, as well as the error of cross-validation, for the optimal inferred networks with the number of reticulation nodes from 0 to 3 respectively, as shown in Table 3. More specifically, we did 10-fold cross-validation and only binary gene trees in the validation sets were used to calculate the error. We can see in the table that the error keeps decreasing from optimal network with 0 reticulation node to the one with 2 reticulation nodes, and there is no improvement from optimal network with 2 reticulation nodes to the one with 3 reticulation nodes. Similar trend holds for all three information criteria, where the improvement from the optimal network with 2 reticulation nodes to the one with 3 is relatively small compared to that from the optimal network with 0 reticulation node to the one with 1, as well as the optimal
network with 1 reticulation node to the one with 2.

Figure 17: The inferred phylogenetic networks of the *M. musculus* dataset. The rows from top to bottom contain top 5 phylogenetic networks with 0, 1, 2 and 3 reticulation nodes, respectively. In each row, networks are listed from left to right with an decreasing value of log likelihood shown under each of them. Branch lengths and inheritance probabilities are shown for the networks with two reticulations.

Furthermore, in order to check how the search covered the space of phylogenetic networks, we exhaustively enumerated all networks with 1 reticulation node and calculated their likelihood scores. More specifically, we first listed all possible 105 binary species trees over 5 taxa (DF, DG, MZ, MK and MC). Then from each of them, say $st$, we cal-
<table>
<thead>
<tr>
<th>$N(k)$</th>
<th>lnL</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>Error of cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(0)$</td>
<td>-47329</td>
<td>94664</td>
<td>94664</td>
<td>94688</td>
<td>$7.69 \times 10^{-5}$</td>
</tr>
<tr>
<td>$N(1)$</td>
<td>-46756</td>
<td>93527</td>
<td>93527</td>
<td>93583</td>
<td>$5.36 \times 10^{-5}$</td>
</tr>
<tr>
<td>$N(2)$</td>
<td>-46392</td>
<td>92806</td>
<td>92806</td>
<td>92893</td>
<td>$4.03 \times 10^{-5}$</td>
</tr>
<tr>
<td>$N(3)$</td>
<td>-46300</td>
<td>92635</td>
<td>92635</td>
<td>92754</td>
<td>$4.13 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

Table 3: The results of information criteria and cross validation of the optimal inferred species networks of the *M. musculus* dataset. $N(k)$ refers to the optimal inferred species network with $k$ reticulation nodes.

culated the likelihood score of every network in $\delta_1(st)$. We ordered all of them by their likelihood scores, and found the top 5 were exactly what we obtained by our heuristic search in Fig. 17(b).
8 PhyloNet implementation and use

PhyloNet [30] is an open-source software our group developed for phylogenetic analysis. All methods we discussed in this paper are implemented in it. We illustrate the usage of the command `inferNetwork ML` which infers a phylogenetic network from a set of gene trees. It takes a set of gene trees and the maximum number of reticulations and returns optimal inferred phylogenetic networks along with branch lengths and inheritance probabilities. There are many parameters for the users to specify; See Table 4 for details.
InferNetwork _ML (gt1 [, gt2...]) numReticulations [-a taxaMap] [-bl] [-b threshold] [-s startingNetwork] [-n numNetReturned] [-h {s1 [, s2...]}] [-w (w1,w2,w3,w4)] [-f maxFailure] [-x numRuns] [-m maxNetExamined] [-d maxDiameter] [-p (rel,abs)] [-r maxRounds] [-t maxTryPerBr] [-i improveThreshold] [-l maxBL] [-pl numProcessors] [-di]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Illustration</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>(gt1 [, gt2 ...])</td>
<td>Comma delimited list of gene tree identifiers.</td>
<td>-</td>
</tr>
<tr>
<td>numReticulations</td>
<td>Maximum number of reticulations to add to the species network.</td>
<td>-</td>
</tr>
<tr>
<td>-a taxaMap</td>
<td>Gene tree / species network taxa association.</td>
<td>-</td>
</tr>
<tr>
<td>-bl</td>
<td>Use the branch lengths of the gene trees for the inference.</td>
<td>No</td>
</tr>
<tr>
<td>-b threshold</td>
<td>Gene trees bootstrap threshold. Edges of gene trees whose bootstrap values are under it will be contracted.</td>
<td>100</td>
</tr>
<tr>
<td>-s startingNetwork</td>
<td>The network to start search from.</td>
<td>MDC tree</td>
</tr>
<tr>
<td>-n numNetReturned</td>
<td>Number of top optimal networks to return.</td>
<td>1</td>
</tr>
<tr>
<td>-h {s1 [, s2 ...]}</td>
<td>A set of specified hybrid species. The size of this set equals the number of reticulation nodes in the inferred network.</td>
<td>-</td>
</tr>
<tr>
<td>-w (w1, w2, w3, w4)</td>
<td>The weights of operations ($\delta_1$, $\delta_2$, $\delta_3$, $\delta_4$) for network arrangement during the network search.</td>
<td>(0.15, 0.15, 0.2, 0.5)</td>
</tr>
<tr>
<td>-f maxFailure</td>
<td>The maximum number of consecutive failures before the search terminates.</td>
<td>100</td>
</tr>
<tr>
<td>-x numRuns</td>
<td>The number of runs of the search.</td>
<td>10</td>
</tr>
<tr>
<td>-m maxNetExamined</td>
<td>Maximum number of network topologies to examine during the search in each run.</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>-d maxDiameter</td>
<td>Maximum diameter to make an rearrangement during network search.</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>-p (rel, abs)</td>
<td>The original stopping criterion of Brents algorithm for optimizing branch lengths and inheritance probabilities of a network.</td>
<td>(0.01, 0.001)</td>
</tr>
<tr>
<td>-r maxRound</td>
<td>Maximum number of rounds to optimize branch lengths and inheritance probabilities for a network topology.</td>
<td>100</td>
</tr>
<tr>
<td>-t maxTryPerBr</td>
<td>Maximum number of trial per branch in one round to optimize branch lengths and inheritance probabilities for a network topology.</td>
<td>100</td>
</tr>
<tr>
<td>-i improveThreshold</td>
<td>Minimum threshold of improvement to continue the next round of optimization of branch lengths and inheritance probabilities.</td>
<td>0.001</td>
</tr>
<tr>
<td>-l maxBL</td>
<td>Maximum branch lengths considered during optimization.</td>
<td>6</td>
</tr>
<tr>
<td>-pl numProcessors</td>
<td>Number of processors if you want the computation to be done in parallel.</td>
<td>1</td>
</tr>
<tr>
<td>-di</td>
<td>Output the Rich Newick string of the inferred network that can be read by Dendroscope [8].</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 4: The usage of command inferNetwork _ML in PhyloNet. The first two parameters are mandatory and all others are optional.
References


