Syst. Biol. 66(2):283–298, 2017

DOI:10.1093/sysbio/syw097
Advance Access publication December 24, 2016

Displayed Trees Do Not Determine Distinguishability Under the Network Multispecies Coalescent

SHA ZHU[1] AND JAMES H. DEGNAN[2,*]

[1]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK;* [2]*Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87110, USA*
[*]*Correspondence to be sent to: Department of Mathematics and Statistics, 311 Terrace NE Albuquerque, NM 87110, USA; E-mail: jamdeg@unm.edu*

Received 10 March 2015; reviews returned 21 September 2015; accepted 8 March 2016
Associate Editor: Peter Foster

Abstract.—Recent work in estimating species relationships from gene trees has included inferring networks assuming that past hybridization has occurred between species. Probabilistic models using the multispecies coalescent can be used in this framework for likelihood-based inference of both network topologies and parameters, including branch lengths and hybridization parameters. A difficulty for such methods is that it is not always clear whether, or to what extent, networks are identifiable—that is whether there could be two distinct networks that lead to the same distribution of gene trees. For cases in which incomplete lineage sorting occurs in addition to hybridization, we demonstrate a new representation of the species network likelihood that expresses the probability distribution of the gene tree topologies as a linear combination of gene tree distributions given a set of species trees. This representation makes it clear that in some cases in which two distinct networks give the same distribution of gene trees when sampling one allele per species, the two networks can be distinguished theoretically when multiple individuals are sampled per species. This result means that network identifiability is not only a function of the trees displayed by the networks but also depends on allele sampling within species. We additionally give an example in which two networks that display exactly the same trees can be distinguished from their gene trees even when there is only one lineage sampled per species. [gene tree, hybridization, identifiability, maximum likelihood, species tree, phylogeny.]

Hybridization between distinct species or populations is often represented using a rooted phylogenetic network rather than a tree (Huson et al. 2010; Bapteste et al. 2013; Nakhleh 2013). In much of the literature on networks representing hybridization, there has been interest in which trees are *displayed* by a network, where a network displays a particular tree if removing some subset of hybridization edges results in the given tree (Huson and Scornavacca 2011; Morrison 2011). For example, several papers investigate finding a network with the minimum number of hybridization events that displays two conflicting input trees (Albrecht et al. 2012; Baroni et al. 2006; Bordewich and Semple 2007; Chen and Wang 2010; van Iersel et al. 2014). These input trees are often described as *gene trees*, and could arise, for example, from estimating trees from sequences from two different loci (e.g., one mitochondrial and one nuclear gene). However, it is not always clear in the literature if a displayed tree in a network refers to a gene tree or a species tree (representing species history rather than ancestry for a specific locus).

A number of methods have recently been developed to infer species networks that explicitly represent species relationships using a network while relationships at the gene level are modeled as gene trees within the network (Jones et al. 2013; Kubatko 2009; Meng and Kubatko 2009; Yu et al. 2011, 2012, 2014; Yu and Nakhleh 2015; Solís-Lemus and Ané 2016). These models are motivated by cases in which hybridization and incomplete lineage sorting are likely to occur simultaneously. In probabilistic versions of these models, gene trees are assumed to be strictly tree-like, and although they are embedded within the network, they do not have to be displayed by the network. In particular, by modeling species networks under the multispecies coalescent, all gene trees have positive probability whether or not they are displayed by the network. We refer to the multispecies coalescent model applied to networks as the Network Multispecies Coalescent (NMSC), and this model is the focus of this article.

The NMSC is intended to represent the case of two populations merging so that the hybrid population is expected to have many individuals with ancestry from both parental populations. Hybrid speciation due to changes in ploidy can result in all descendants of the hybrid having one ancestor, in which case incomplete lineage sorting would not occur. Our model is therefore restricted to homoploid hybridizations; see (Jones et al. 2013) for models applicable to polyploid hybridization. The NMSC is also not intended to model horizontal gene transfer, which causes much of the reticulation in bacterial networks, or recombination, two other processes that motivate network representations, including in likelihood frameworks (Strimmer and Moulton 2000; Jin et al. 2006; Abbott et al. 2010; Nguyen and Roos 2015).

An early study concerning the multispecies coalescent approach to networks assumed that a hybrid species occurred sometime in the past, and that a single allele is sampled from a population descended from this hybrid species (Meng and Kubatko 2009). Under this assumption, an allele from the hybrid species could have descended from one of two possible ancestral populations. This results in the probability of a gene tree being a linear combination of the gene tree probabilities from two parent species trees, where the parent species trees are obtained by removing one of the two hybridization edges. This reduces the network into a set of two species trees, and takes advantage of the fact that probabilities of gene trees given species trees under the multispecies coalescent can be computed. This approach is useful in cases where only one allele is sampled per locus from any species that is the result of hybridization. However, this approach does not generalize easily to cases where an ancient hybrid subsequently speciates or in which more than one allele is sampled from a hybrid species.

To compute more general likelihoods than the approach of Meng and Kubatko (2009), Yu et al. (2012) developed an algorithm that represented a species network as a multilabeled tree (MUL-tree) where species descended from hybrids are represented more than once in the tips of the MUL-tree. The likelihood is computed by summing over possible assignments of alleles to these nonuniquely labeled tips. This approach allows multiple alleles to be sampled within populations as well as hybrids to occur anciently in the network so that populations descended from hybrids can subsequently speciate.

A problem for inferring phylogenetic networks, however, is that they are not always identifiable. That is, examples can be found where two networks that correspond to distinct biological hypotheses about speciation and hybridization events can give rise to the same distribution of gene trees. Two networks that give the same probabilities on all gene tree topologies can be said to be mathematically indistinguishable. We might also differentiate between mathematical distinguishability, by which we mean that two models lead to distinct probability distributions, and practical distinguishability, which would mean that one can perform reasonably accurate model selection from finite data. In this article, we are primarily interested in mathematical distinguishability; however, we also do simulations to address the more practical sense of distinguishability.

Mathematical indistinguishability means that there are some sets of networks for which no amount of data could determine which of the networks gave rise to the data. Although several positive results have been found for identifying species trees from gene trees and sequences evolving on gene trees (DeGiorgio and Degnan 2010; Allman et al. 2011a,b; Chifman and Kubatko 2015), the identifiability of networks is a more challenging problem theoretically. One reason for this is that the space of phylogenetic networks is much larger than that of phylogenetic trees, and is infinite if the number of hybridization events is not bounded. Networks can also have "ghost" lineages (lineages that once existed but that went extinct) that can also make identifiability more difficult for networks than for trees (Marcussen et al. 2015).

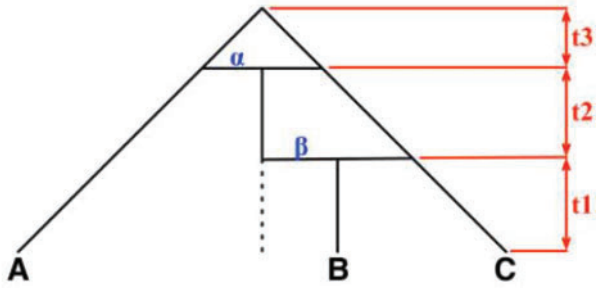

FIGURE 1. Example of three-taxon network in which parameters are not identifiable from gene tree topologies. The example is taken from Yu et al. (2012), Figure 4, doi:10.1371/journal.pgen.1002660.g004.

One factor that affects network identifiability is whether or not gene trees have branch lengths (Pardi and Scornavacca 2015). If only gene tree topologies are used, then many distinct networks will give equivalent gene tree topology probabilities if speciation times and hybridization parameters are allowed to vary. In some cases, the distribution of the coalescence times in the gene trees (which are functions of the gene tree branch lengths) will depend on hybridization events, thus allowing it to be possible to distinguish two networks that could not be distinguished using only topologies.

If only gene tree topologies are used, then the number of hybridization events that it is possible to infer may also be limited. An example is given in Yu et al. (2012) in which a network has three species and two hybridization events (Fig. 1). In that example, there are three times corresponding to either speciation or hybridization events, and there are two hybridization parameters. With only three observed gene tree topologies and five parameters, even if the network topology is known, this results in a system of three (estimated) equations and five unknowns (one equation for each gene tree topology). It is therefore not surprising that it is not possible to determine the five parameters using the gene tree topologies alone.

Yu et al. (2012) show that for the three-taxon example, identifiability is improved by allele sampling. If two alleles are sampled from species B, then there are 15 possible gene tree topologies (since we now have gene trees with four leaves). The 15 gene tree probabilities can then be used to estimate the five parameters.

A more difficult case of identifiability might appear to be that given by Pardi and Scornavacca (2015) (Fig. 2). In this example, when there is one allele sampled per species, the distribution of the gene trees, including their branch lengths, is identical for two different networks. The authors point out that there are three species trees displayed by the networks and that the three species trees

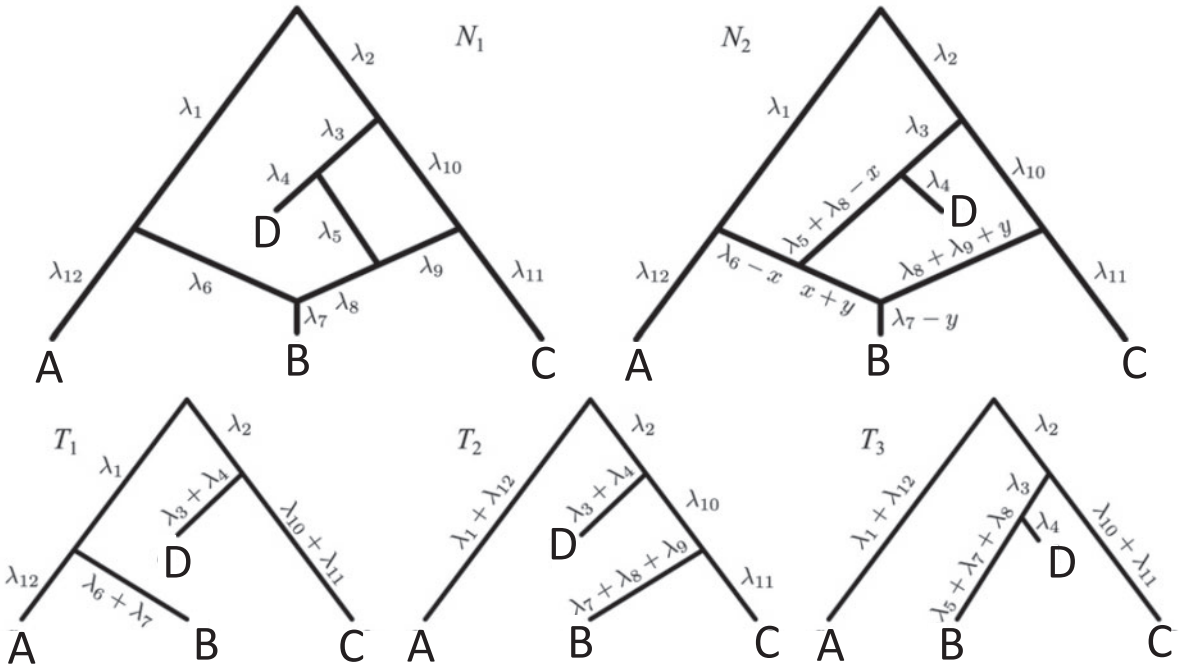

FIGURE 2. Networks N_1 and N_2 from Pardi and Scornavacca (2015), doi:10.1371/journal.pcbi.1004135.g003. The two networks both display exactly the same three trees, T_1, T_2, and T_3.

can have identical branch lengths given certain choices of parameters in the two networks. The authors claim that "no method based on this definition of likelihood will be able to discriminate between" the two networks.

While we agree that the likelihood used in Yu et al. (2012) cannot distinguish the two networks if one allele per species is used (whether or not branch lengths are used for this case), we disagree if the data can have multiple alleles per species and if incomplete lineage sorting is possible. Some likelihood approaches assume that sequence alignments evolve on gene trees displayed by the network (Jin et al. 2006; Park and Nakhleh 2012; Pardi and Scornavacca 2015), leading to the likelihood:

$$\prod_{i=1}^{m} \sum_{T \in \mathcal{T}(N_k)} P(A_i|T)P(T|N_k) \qquad (1)$$

where $\mathcal{T}(N_k)$ is the set of trees displayed by network N_k and A_i is the sequence alignment for the ith locus. This likelihood sums over the trees displayed by the network, and is motivated by cases such as horizontal gene transfer in bacteria and hybrid speciation, in which gene trees are expected to be trees displayed by the network.

The likelihood used in Meng and Kubatko (2009) and Yu et al. (2012) treats gene trees as data, and can be written instead as

$$\prod_{i=1}^{m} \sum_{W_j} \omega_j P(g_i|W_j) \qquad (2)$$

where W_j are species trees called *parental trees* in Meng and Kubatko (2009), and the ω_j are weights based on the probability that lineages take certain paths through the network. In Meng and Kubatko (2009), in which there is only one descendant from any hybrid node, the trees W_j are indeed displayed by the network, whereas in Yu et al. (2012), the trees W_j are generally MUL-trees which are not always displayed the network. The approach in this article can also be written using equation (2), where the W_j terms are uniquely labeled trees (not MUL-trees), and can be interpreted as parental trees, similarly to Meng and Kubatko (2009), but are not necessarily displayed by the network.

The description of the likelihood in Pardi and Scornavacca applies to cases where gene trees are considered known and can only arise as displayed trees within the network. This assumption might be reasonable for a number of biological processes such as hybrid speciation, in which an individual hybrid can be ancestral to a new species (Abbott et al. 2010), recombination among viruses, and horizontal gene transfer. Under the NMSC, gene trees are not necessarily displayed by the network. In the parental species tree approach of Meng and Kubatko (2009), parental species trees are displayed by the network if there is only one individual descended from each hybrid, or if all lineages are constrained to coalesce more recently than a hybridization event (such as for hybrid speciation). However, for cases where there are several lineages in a hybrid population, such as is allowed in Yu et al.

(2012), parental species trees under the NMSC are also not necessarily displayed by the network.

The likelihood used in Yu et al. (2012) is calculated over a sum of probabilities based on MUL trees, with the summation being over allele assignments. Some allele assignments will correspond to displayed trees, but some may not, particularly when two lineages (whether or not they are from the same species) follow different paths up the network at a hybridization node. In these cases, the probability cannot be written in terms of a displayed tree obtained by dropping one of the hybridization edges. Consequently, equation (1) is not in general an accurate representation of the likelihood used in Yu et al. (2012).

In the next section, we describe an alternative method for representing gene tree probabilities that does not use MUL trees, and describes the probabilities of gene trees as linear combinations as in Equation (1), except that the sum is not necessarily over trees displayed by the network. This helps to explain why equivalence of displayed trees is not sufficient for determining that two networks are indistinguishable.

Gene Tree Probabilities as Linear Combinations Under Different Species Trees

An alternative way of deriving the likelihood of the network given the gene trees can be obtained by conditioning on events at hybridization nodes and branches descended from them and using recursion. This results in an alternative algorithm to that of Yu et al. (2012) for computing the likelihood of a gene tree and results in an expression more similar to the strategy of Meng and Kubatko (2009) of reducing the probability given a network to a linear combination of probabilities given species trees. Following Meng and Kubatko, we refer to these species trees as *parental trees* or *parental species trees*. For networks with more than one lineage descended from a hybridization node, the recursion results in a linear combination including some species trees that are not displayed by the network. An example is shown in Figure 3, which gives an intuitive picture of the procedure. The example generalizes the three-taxon example in Figure 1 by splitting taxon B into two species and making hybridization edges to not be horizontal.

At each step in the recursive approach, we condition on whether lineages either coalesce or do not coalesce, or we condition on whether lineages go left versus right at a hybridization node. Each step reduces the network into a larger number of smaller networks until the process ends with a collection of species trees. Details of the algorithm are given in the Appendix.

Distinguishability of Networks with the Same Displayed Trees

Decomposition of Networks N_1 and N_2

We use the networks N_1 and N_2 described as indistinguishable (Pardi and Scornavacca 2015) (Fig. 2).

These networks are slightly modified from Pardi and Scornavacca (2015) with species written with capital letters. We then consider a modified version in which the population descended from both hybrid nodes undergoes speciation, resulting in species B and E (networks N_1' and N_2' in Fig. 4). The networks N_1' and N_2' are similar to N_1 and N_2, respectively, when there are two lineages sampled from B (Fig. 2). The number of lineages sampled per species affects the decomposition of the networks into parental species trees.

When there are two lineages sampled from species B in N_1, we denote the lineages by b_1 and b_2. They fail to coalesce in this branch with probability $g_{22}(\lambda_7)=e^{-\lambda_7}$. Assuming no coalescence, the lineages from species B either both go to the left, one goes leftward and one rightward, or both go to the right at the lower hybridization node. The cases are listed in Supplementary Table 1 (available on Dryad). An example corresponding to case W_2 in Supplementary Table 1 (available on Dryad) is shown in Figure 5. We use γ_1 and γ_2 for the probability that a lineage goes left at the more recent and less recent hybridization nodes, respectively, in network N_1. Similarly, γ_3 and γ_4 are the probabilities that a lineage goes left at the more recent and less recent hybridization nodes, respectively, in N_2. These hybridization parameters are also called *inheritance probabilities* (Pardi and Scornavacca 2015). The parental species trees W_1–W_{28} referred to in Supplementary Table 1 (available on Dryad) are given in newick format in Supplementary Tables 2 and 3 (available on Dryad).

We wish to show that there is at least one gene tree topology with different probabilities under the two networks. The calculations are simplest for a gene tree that is very unlikely, in which case calculations can be done "by hand." For example, consider the gene tree $g=((((a,d),c),b_1),b_2)$. The probability of this gene tree topology conditional on the above parental species trees is given in Supplementary Table 1 (available on Dryad).

The probabilty of the gene tree g under the two networks can be written as

$$P_{N_1}(g)=\sum_{i=1}^{14}\omega_i P(g|W_i),\quad P_{N_2}(g)=\sum_{i=15}^{28}\omega_i P(g|W_i)$$

where $\omega_i=P_{N_j}(W_i|N_j)$. To illustrate using Supplementary Table 1 (available on Dryad), the probability of g under N_1 is

$$\begin{aligned}P_{N_1}(g)=&\,g_{22}(\lambda_7)\gamma_1^2 g_{22}(\lambda_6)g_{33}(\lambda_1)g_{22}(\lambda_2)/180\\&+g_{22}(\lambda_7)\gamma_1(1-\gamma_1)\gamma_2 g_{22}(\lambda_1)g_{22}(\lambda_3)g_{33}(\lambda_2)/180\\&+\cdots+g_{21}(\lambda_7)(1-\gamma_1)(1-\gamma_2)\cdot 0\end{aligned}$$

Here $g_{ii}(t)=e^{-\binom{i}{2}t}$ from equation (A.2).

The probabilities do not depend on λ_4, λ_{11}, or λ_{12} due to there only being one lineage on each of these pendant edges and therefore no probability of coalescence on these edges. The terms $P(g|W_i)$ are equal to 0 for five

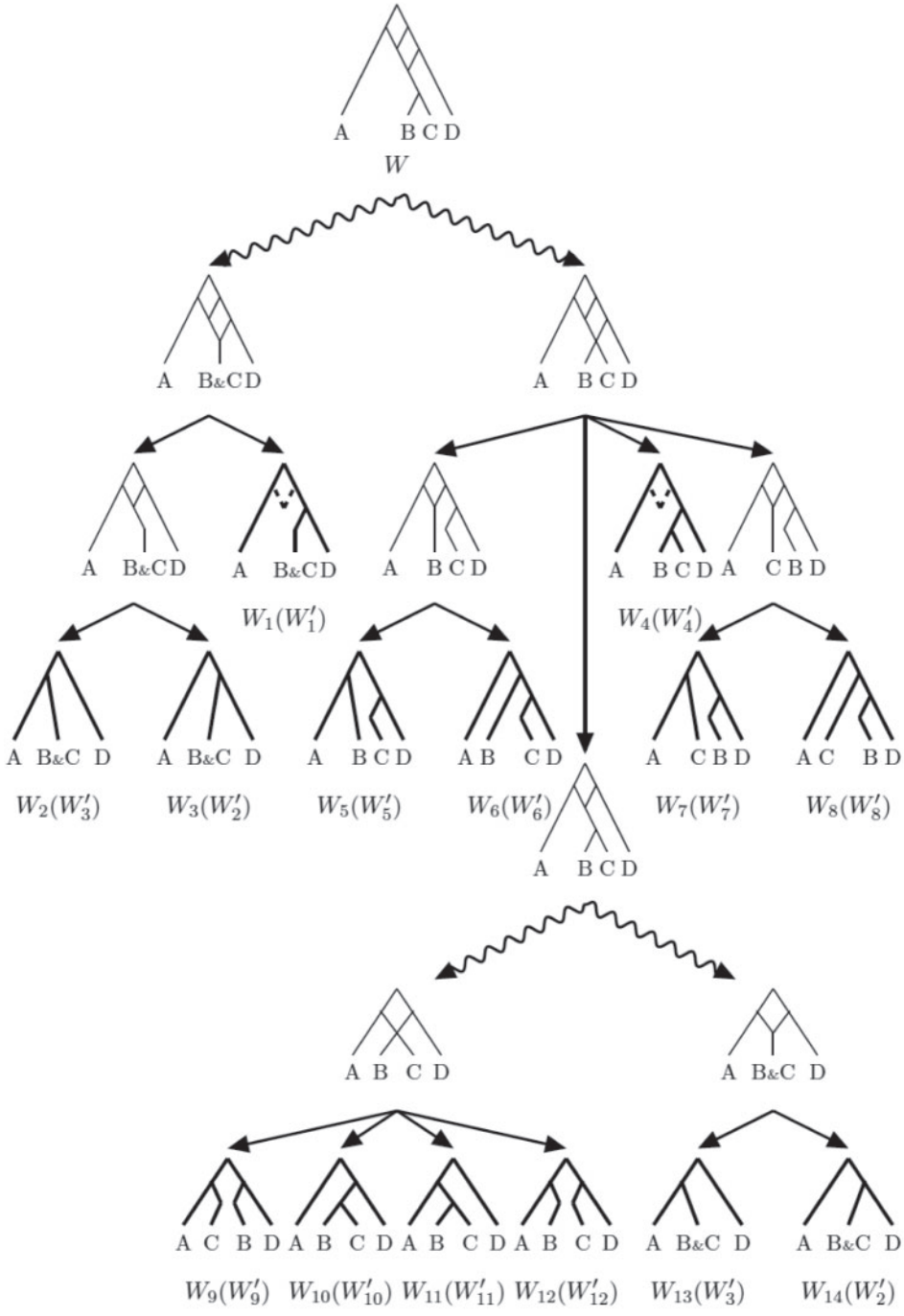

FIGURE 3. Decomposition of network into parental species trees. Wiggly arrows indicate conditioning on a coalescence event. Solid arrows indicate conditioning on paths taken at a hybridization node. When a taxon is labeled *B*&*C*, this can be interpreted as a leaf where *B* and *C* have been merged, or as a two-taxon tree where there is an infinite branch for the ancestor of *B* and *C*, guaranteeing that lineages sampled from *B* and *C* coalesce with each other more recently than with any other taxa.

choices of i under both networks. These cases correspond to parental species trees that have conditioned on the event that lineages b_1 and b_2 have coalesced more recently than one of the hybridization nodes, which is impossible for gene tree g. This reduces the number of parental species trees needed in the sums from 14 to 9.

In addition, the gene tree forces all coalescences to occur more anciently than the root of the network for

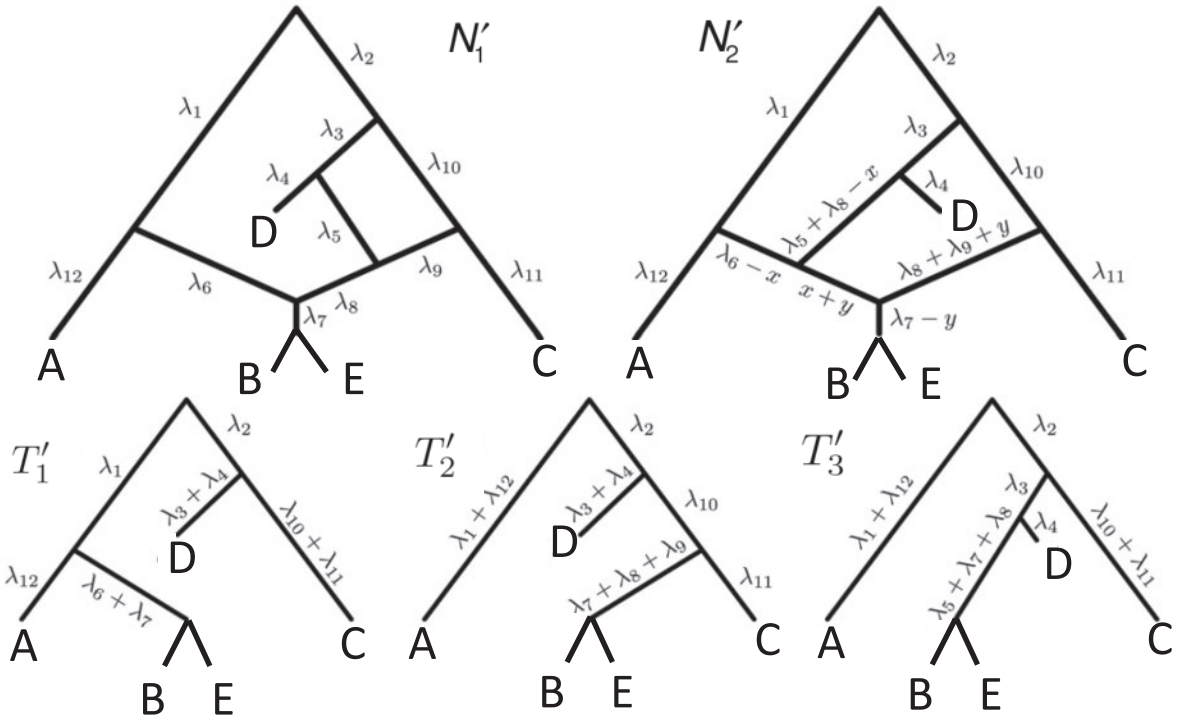

FIGURE 4. Extension of networks N_1 and N_2 to allow two species descended from the most recent hybridization node. The figure is modified from doi:10.1371/journal.pcbi.1004135.g003.

both networks. This means that only one coalescent history needs to be computed (instead of enumerating over several coalescent histories for each parental species tree). Because of the asymmetry in the gene tree, only one sequence of coalescences, out of $\binom{5}{2}\binom{4}{2}\binom{3}{2}\binom{2}{2}=180$, produces the gene tree, which leads to the denominators in the probabilities.

When there is only one lineage sampled per species, N_1 and N_2 are indistinguishable under the following conditions (Pardi and Scornavacca 2015):

1. $\gamma_3=1-(1-\gamma_1)(1-\gamma_2)$
2. $\gamma_4=\gamma_1/\gamma_3$
3. $0<x<\min\{\lambda_6,\lambda_5+\lambda_8\}$
4. $0<y<\lambda_7$

We pick a particular set of parameters to show that the networks are distinguishable when two lineages are sampled in species B. For the choice of parameters

$$\gamma_1=1/3,\ \gamma_2=2/3,\ \gamma_3=7/9,\ \gamma_4=3/7,$$
$$x=y=1/2,\ \lambda_i=1,\ i\in\{1,\ldots,12\},$$

the conditions for indistinguishability specified by Pardi and Scornavacca (2015) are met, and the probability of gene tree topology $g=((((a,d),c),b_1),b_2)$ under the two networks is

$$P_{N_1}(g)\approx 7.7\times 10^{-6},\ \ P_{N_2}(g)\approx 7.6\times 10^{-6},$$

Thus, the gene tree probability is approximately 1.4% higher under N_1 than N_2. Both probabilities are small because this gene tree is quite unlikely for both networks, requiring no coalescences to occur except more anciently than the root. Nevertheless, it shows that the two networks have different gene tree distributions.

Rather than using gene tree probabilities, clade probabilities could also be used to distinguish the two networks for many parameter values. Let λ_6 and λ_8 both be very large and let x be very small, so that any two lineages on branches with these lengths will almost certainly coalesce. Similarly, let λ_7 be very small, so that b_1 and b_2 are very unlikely to coalesce. For these parameters, with high probability, $\{b_1,b_2\}$ is a clade on the gene tree when, and only when, both lineages both go to the left or both go to the right at the more recent hybridization node. Then using the above values of γ_1 and γ_3, the probability that a gene tree has clade $\{b_1,b_2\}$ is approximately $\gamma_1^2+(1-\gamma_1)^2=45/81$ under N_1 and is approximately $\gamma_3^2+(1-\gamma_3)^2=53/81$ under N_2. The clade probability will therefore distinguish the two networks.

We emphasize that if there is only one lineage sampled per species, then there is at most one lineage present at each hybrid node for N_1 and N_2. In this case, the methods of Meng and Kubatko (2009) can be applied to calculate gene tree probabilities, but we agree with Pardi and Scornavacca (2015) that N_1 and N_2 are indistinguishable in this situation.

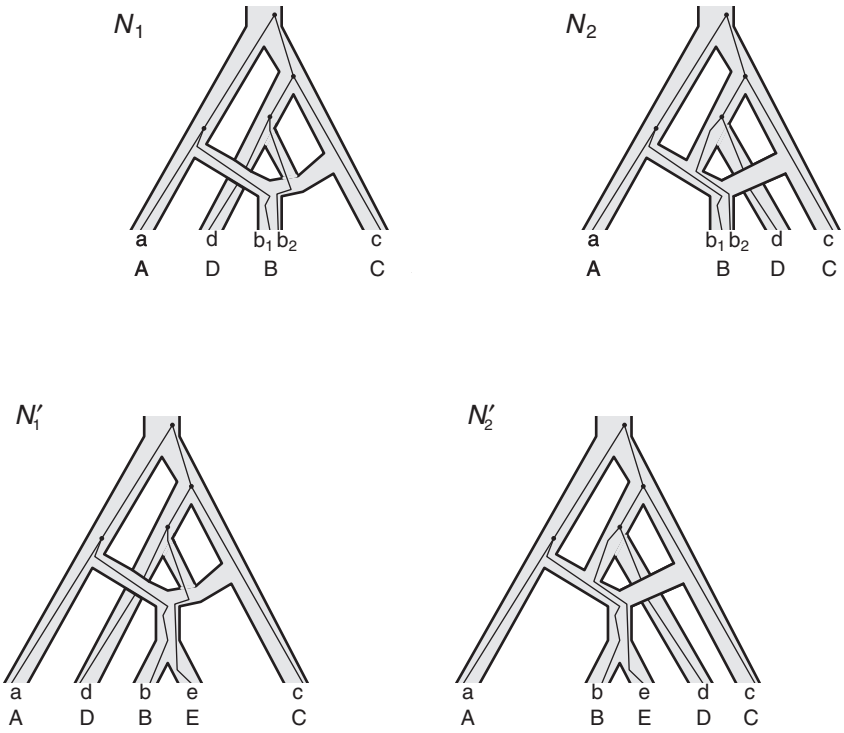

FIGURE 5. Gene trees in networks that display the same trees. Here N_1 and N_2 display the same trees with the same branch lengths given suitable choices of parameters. Similarly, N_1' and N_2' display the same trees. Two gene trees are shown with coalescence times that are compatible with both N_1 and N_2, and another two gene trees are shown with coalescence times compatible with both N_1' and N_2' so that knowing the coalescence times in the gene tree does not determine which network it evolved in. The gene trees in this figure cannot be represented as having evolved in a species tree displayed by the networks. The gene tree in N_1 corresponds to case W_2 where b_1 goes left, b_2 goes right, then left. The gene tree in N_2 corresponds to case W_{16}, where both go left, then b_1 goes left, b_2 goes right.

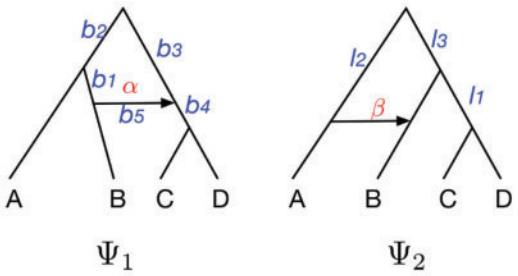

FIGURE 6. Two networks taken from Figure 2 of Yu and Nakhleh (2015) that display the same trees and triplets. The networks are distinguishable using gene tree probabilities but not using rooted triple probabilities.

Rooted triples and quartets have also been used to reconstruct or infer networks under the NMSC (Yu and Nakhleh 2015) and Solís-Lemus and Ané (2016). Yu and Nakhleh (2015) give an example of networks that are not distinguishable using probabilities of triples in the gene trees that have evolved in the network (Fig. 6). They give an explanation that the networks display the same sets of triples. We agree that for this particular example, triples cannot be used to distinguish their networks Ψ_1 and Ψ_2,. However, for the case of N_1 and N_2, triples can be used to distinguish the networks when there are two lineages sampled from B, even though N_1 and N_2 display the same set of rooted triples. For example, using the previous parameters, the probability of triple $((a,b_1),b_2)$ is approximately 0.081 under N_1 and approximately 0.089 under N_2.

As an alternative explanation for why triples cannot be used to distinguish Ψ_1 and Ψ_2, it is noticed that equating triple probabilities for the two networks, such as $P_{\Psi_1}[((a,b),c)] = P_{\Psi_2}[((a,b),c)]$ results in a system of 12 equations. Removing linear dependencies, such as $P_{\Psi_i}[((a,b),c)] = P_{\Psi_i}[((a,b),d)]$ and that for any three taxa, the sum of the three rooted triple probabilities sums to 1. Removing such linearly dependencies from the system results in five linearly independent equations for a system with nine parameters, making the system underdetermined. This makes it possible to find parameters for Ψ_2 that will make the rooted triple probabilities match those for Ψ_1.

Probabilities of unrooted quartets can also be calculated, and again, these distinguish the networks N_1 and N_2 when there are two lineages sampled from species B. For the same parameters as above, with $\lambda_i = 1$, $x = y = 1/2$, $\gamma_1 = 1/3$, and $\gamma_2 = 2/3$, the probability that a rooted gene tree displays the quartet $((a,b_1),(d,b_2))$ is approximately 0.10 and 0.14 for networks N_1 and N_2, respectively. We note this example in particular because the recently introduced method for inferring networks from quartets (Solís-Lemus and Ané 2016) cited the

results in Pardi and Scornavacca (2015) as a reason to not apply the method to level-k networks for $k > 1$ (networks in which an edge can appear in more than one cycle of the graph, such as N_1 and N_2).

The example from Yu and Nakhleh (2015) suggests that caution is indeed needed, since there are cases where trees but not summary statistics such as rooted triples can distinguish the networks. The example of N_1 and N_2 with multiple lineages per species, however, suggests that even more complicated networks are potentially distinguishable from rooted triples or quartets. In the Yu and Nakhleh (2015) example, we suspect that distinguishability would be achieved by sampling additional lineages. Their example is somewhat different from that of N_1 and N_2 in that the different networks do not have the same species descended from the hybrid, and the number of descendants of the hybrid is not the same for the two networks. The networks are also level 1, with only one hybridization event, and distinguishability is a problem not because of the complexity of the networks but rather because of the small number of taxa and resulting small set of linearly independent rooted triple probabilities for the number of parameters.

It is also possible to have distinguishability between two networks that each display the same set of trees when there is only one lineage sampled per species. In particular, the networks N_1' and N_2' are essentially identical to N_1 and N_2, respectively, when the pendant branches leading to species B and E have length 0. In this case, the most recent ancestral population to B and E is a single population, and the lineages b and e are two lineages from the same population. As a result, the distributions of gene tree topologies under N_1' and N_2' are identical with those of N_1 and N_2 when lineage b is replaced by b_1 and lineage e is replaced by b_2. Similarly, the networks N_1' and N_2' both display the trees T_1', T_2', and T_3' (Fig. 4), which are equivalent to T_1, T_2, and T_3 (Fig. 2), respectively, when B is replaced by (B, E). The length of the pendant edges does not affect gene tree probabilities when one lineage is sampled per species. Consequently, gene tree $g' = ((((a,d),c),b),e)$ has the same probabilities under N_1' and N_2' ($((((a,d),c),b_1),b_2)$ has under N_1 and N_2, respectively: $P_{N_i'}(g') = P_{N_i}(g)$ for $i = 1, 2$.

The important point is that there exist pairs of networks that display the same trees but can be distinguished under the NMSC model, even when there is only one lineage sampled per species. This example demonstrates that showing that two networks display the same trees (including branch lengths and inheritance probabilities) is not sufficient for showing that the networks are indistinguishable. In this particular case, the networks N_1' and N_2' are also distinguishable using rooted triplets, quartets, or clades, in spite of the two networks displaying exactly the same rooted triplets, quartets, and clades. A crucial reason for the ability to distinguish networks is the following: if there is more than one lineage descended from a hybrid node (either due to the hybrid population speciating or due to sampling more than one lineage from a population descended from a hybrid), there can exist gene trees that are not embedded in a tree displayed by the network.

Simulation

Distinguishability of N_1 and N_2 Using Model Selection

To illustrate the ability of network methods to distinguish two networks that display the same trees, we simulated gene trees from N_1 with two and three alleles sampled from species B and one allele sampled from each of the other species. We used phylonet (Than et al. 2008) to compute likelihood scores by optimizing branch lengths and hybridization parameters assuming the fixed network topologies N_1 and N_2. The network branch lengths were based on using the network N_1 (Fig. 5) with the height of the network being 10 coalescent units. The networks used for simulation in hybrid-Lambda (Zhu et al. 2015) are, in coalescent units:

$$(((A{:}5, (B{:}3)h_1\#.5{:}2)s_2{:}5, ((D{:}5.6, (h_1\#.5{:}1.3)h_2\#.6{:}1.3)s_3{:}2.3,$$
$$(h_2\#.6{:}.1, C{:}4.4)s_4{:}3.5)s_5{:}2.1)s_6{:}10, O{:}20)r;$$
$$(((A{:}5, (B{:}1)h_1\#.5{:}4)s_2{:}5, ((D{:}5.6, (h_1\#.5{:}3.3)h_2\#.6{:}1.3)s_3{:}2.3,$$
$$(h_2\#.6{:}.1, C{:}4.4)s_4{:}3.5)s_5{:}2.1)s_6{:}10, O{:}20)r;$$

where species O is an out-group. In this notation, all internal nodes (both hybridization and speciation nodes) are labeled. After the hybridization nodes, h_i, the first number represents the probability of going "left," and the second number represents the branch length from the hybridization node to the next node (either left or right). Thus, for both networks used in the simulation, the probability of going left for h_2 is $\gamma_2 = 0.6$. In the extended newick string, the branch length after the first (second) instance (reading from the left) of h_i is the length of the branch leading from h_i to the left (right) parent of h_i.

The networks have identical topologies, inheritance probabilities, and branch lengths, except that $\lambda_7 = 3.0$ for the first network and $\lambda_7 = 1.0$ for the second network. The second network has a higher probability that the lineages sampled from B will fail to coalesce more recently than the most recent hybridization node, and therefore has a higher level of incomplete lineage sorting. We therefore refer to this as the "high ILS" network. Gene trees on this network are much less likely to have monophyly of lineages sampled from B. The other network is referred to as the "low ILS" network.

For each set of gene trees, the likelihood under the estimated parameters was compared with the two networks and the proportion of times that N_1' had a higher likelihood than N_2' was reported. The simulation was performed with 50, 100, 200 and 400 independent loci. For each gene tree, alignments with 500 sites were simulated using seq-gen (Rambaut and Grassly 1997) under the $GTR+\Gamma+I$ model with base frequencies of 0.3, 0.2, 0.2, and 0.3 for A, C, G, and T, respectively, with

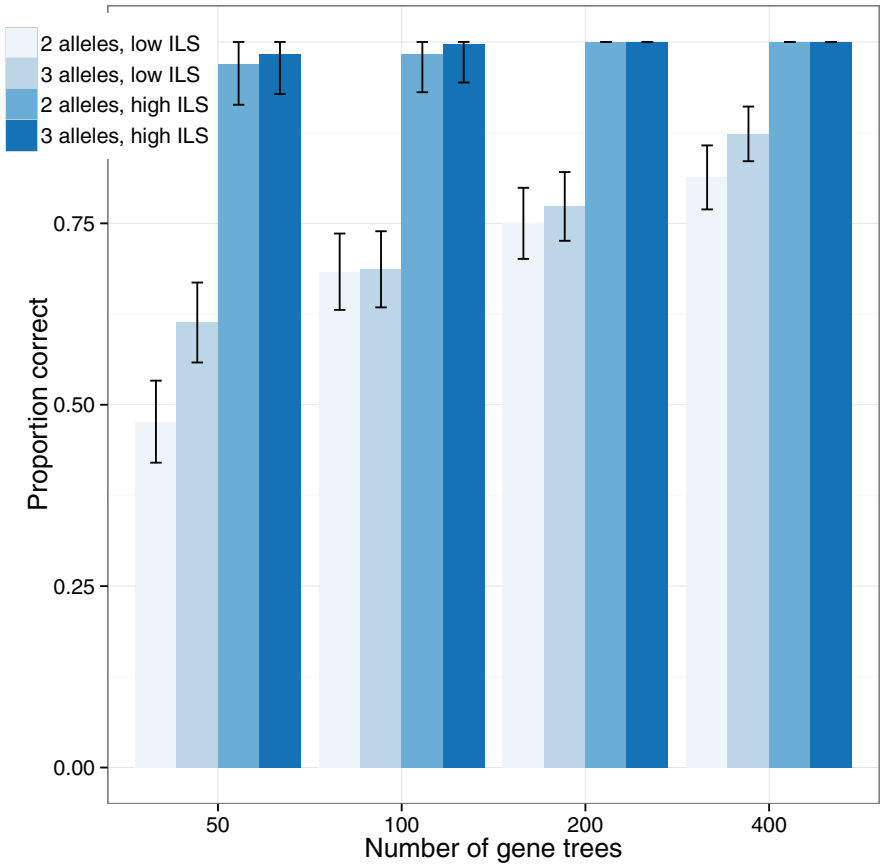

FIGURE 7. Performance of phylonet for distinguishing networks N_1 and N_2 when data was simulated from N_1, with the indicated number of alleles sampled from species B and all other species having one allele sampled. The fraction of times out of 300 iterations that the N_1 had a higher likelihood than N_2 is reported when both networks were fixed and phylonet optimized branch lengths and inheritance probabilities. "High ILS" refers to $\lambda_{=}1.0$, and "Low ILS" refers to $\lambda_7 = 3.0$, with all other parameters kept the same.

four rate categories and 10% invariable sites. As is typical with multilocus simulations, gene trees were independent with no recombination within loci. An out-group was added to the network with the MRCA of the out-group and in-group taxa being 10 coalescent units deeper than the root of the in-group taxa. This ensures extremely high probabilities that the in-group taxa are monophyletic in the gene trees. Gene trees were estimated using phyml (Guindon et al. 2010) under the $GTR+\Gamma+I$ model assuming four rate categories and estimating all other parameters, and using default tree searches. Unrooted gene trees estimated by phyml were rooted using the out-group, and the out-group was then removed before inputting the estimated rooted gene trees into phyml.

Not surprisingly, increasing the number of loci increased the ability to distinguish the two networks (Fig. 7). Increasing the number of alleles (from 2 to 3) increased the ability of maximum likelihood to distinguish the networks. An intuitive explanation is that with more alleles, it is more likely that lineages cannot be embedded in a tree displayed by the network. Having the higher level of ILS lineages (obtained by having a smaller value for λ_7) greatly increases the ability of phylonet to distinguish the two networks, and this also has the explanation that since B lineages are less likely to have coalesced more recently than the first hybridization node, gene trees in the high ILS case are less likely to be embedded in a tree displayed by the network than gene trees in the low ILS case. The fact that increasing allele sampling can improve inference of species relationships has been emphasized in the species tree literature as well (Maddison and Knowles 2006; DeGiorgio and Degnan 2014; Heled and Drummond 2010; Huang et al. 2010). In this case, however, sampling multiple alleles not only improves inference, but is also crucial for being able to distinguish the networks at all.

The "high ILS" case, with a branch length of 1.0 coalescent units, is typical for cases known to have significant amounts of ILS For example, the level of gene tree incongruence, for which approximately 60–80% of trees have humans and chimpanzees being the most closely related among humans, chimpanzees, and gorillas (Ebersberger et al. 2007) suggests an internal branch length of between 0.5 and 1.2 coalescent units (Ané 2010; Degnan 2010). The probability that two lineages coalesce within 1.0 coalescent units is $1-e^{-1.0}=0.63$, whereas the probability that two lineages coalesce within 3.0 units (the low ILS case in our simulations) is 0.95. Thus, for the low ILS case with

two alleles from B, more than 19 in 20 gene trees evolve on a tree displayed by the network, making it difficult to distinguish the two networks; for the high ILS case, using $\gamma_1 = 0.5$, the proportion of gene trees evolving on a tree not displayed by the network is close to $(1-0.36)(0.5) \approx 32\%$.

Comparison of Phylonet and Hybrid-Coal

Both phylonet and hybrid-coal compute probabilities of gene tree topologies given species networks, but the two programs use different algorithms. The program hybrid-coal uses a recursion that allows representing the gene tree topology probability as a linear combination of probabilities given species trees. In contrast, phylonet initially represented probabilities as a sum over probabilities of coalescent histories given MUL-tree representations of networks, and more recently also implemented the ancestral configuration approach (Wu 2012), which tends to run more quickly than the coalescent history approach for larger trees (roughly more than 10 taxa, depending on tree shape). There are also many features in phylonet not implemented in hybrid-coal, such as algorithms to infer the network from a set of gene trees, searching over network topologies, branch lengths, and inheritance probabilities, and using branch lengths in the gene trees.

The main idea of hybrid-coal is compatible with both the coalescent history and ancestral configuration approaches, since once the parental species trees have been enumerated, either coalescent histories or ancestral configurations could be used to compute the probability of the gene tree given the parental species tree. Currently, only the coalescent history approach is implemented in hybrid-coal, but the ancestral configuration method could be added in the future. In comparison with phylonet, hybrid-coal breaks down the network into a larger number of smaller problems, with the parental species trees tending to be smaller trees than the MUL-tree representation of the network, which can have more leaves than there are taxa. This could potentially be an advantage in future parallel programming implementations of the algorithm.

The main advantage for having the new algorithm in hybrid-coal is perhaps the theoretical insight it gives in terms of representing gene tree probabilities in terms of parental species trees. This appears to be a fairly intuitive way to think about the relationship between gene trees and species networks (Holland et al. 2008; Meng and Kubatko 2009), although we have shown that perhaps counterintuitively, the parental species trees are often not displayed by the network. We hope that future theoretical work will make use of the representation of gene tree probabilities as mixtures arising from different species trees. In other contexts, mixtures of trees have proved identifiable (Allman et al. 2012; Rhodes and Sullivant 2012), and this might be a useful approach for thinking about identifiability of networks.

Discussion

The Effect of Branch Lengths

Consideration of branch lengths in the gene trees can lead to the ability to distinguish networks which could not be distinguished using only topologies. For example, if a species history has multi-edges—cases where two nodes are directly connected by two distinct edges—it can still be possible to estimate parameters of this model and to distinguish it from a model in which multi-edges are collapsed. An example of networks with multi-edges is shown in Figure 8. Biologically, a multi-edge could depict a population temporarily splitting into two populations with no gene flow followed by the populations merging at a later time before either population splits. This type of history would be desirable to be able to estimate since it could occur, for example, due to glaciation or other episodic events that temporarily divide populations (Comes and Kadereit 1998; Marshall et al. 2009).

Coalescence times are potentially useful in these cases because multi-edges affect the distribution of coalescence times. Theoretically, a multi-edge will result in a bimodal distribution of coalescence times. In practice, estimated coalescence times are highly variable and subject to estimation error, so that a bimodal signature might be difficult to detect. However, this does not affect the point that multi-edges are potentially inferable given ideal data.

The example from Figure 8 essentially arises from the three-taxon network in Figure 1 when taxon A is dropped and $\alpha = 0$. The example illustrates that this network has potentially identifiable parameters when using branch lengths in the gene trees even when topologies cannot identify the parameters in the network.

A deeper difficulty with identifiability is that it is not clear that hybridization can be distinguished from other population genetic processes that can result in gene tree incongruence and complicated distributions of coalescence times. For example, alternating bottlenecks and population expansions can result in similar multimodal distributions of coalescence times as that found in Fig. 8 (DeGiorgio et al. 2011). Bottlenecks can also be a problem in practice for distinguishing two networks since a smaller population size makes incomplete lineage sorting less likely, thereby making gene trees more likely to be displayed by the network. The extreme case is a bottleneck of size one, which can occur in hybrid speciation, and guarantees that all gene trees are displayed by the network.

As another example, it is well known that the multispecies coalescent on a three-taxon tree with no ancestral population structure predicts that one triplet is most frequent while the other two triplets are tied in probability, and that these tied probabilities are less frequent (Nei 1987). Consequently, a test of equality of proportions is sometimes used for the less frequent triplets as a goodness of fit test for the multispecies coalescent (Degnan and Rosenberg 2009; White et al.

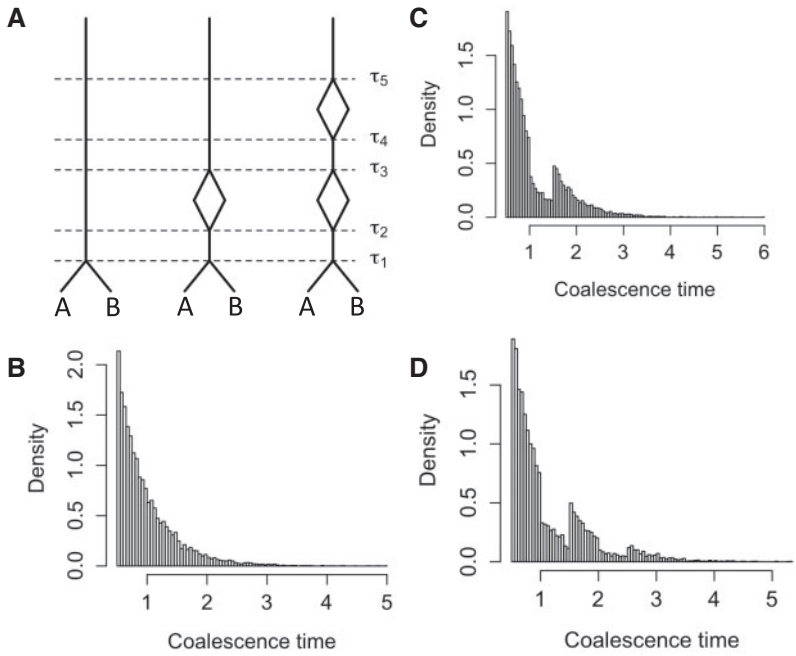

FIGURE 8. Coalescence times for networks with multi-edges. a) The leftmost species history is a two-taxon species tree. The middle and right species histories have one or two sets of multi-edges, reflecting species diverging and subsequently hybridizing without any other speciation. Histograms of 100,000 coalescence times for single genes sampled from species A and B are depicted in (b)–(d). (c) and (d) correspond to the middle and rightmost species histories, respectively. Simulations are based on $\tau_i = i/2$ coalescent units and were done in the program ms (Hudson 2002).

2009; Ané 2010; Cranston 2010; Song et al. 2012). Asymmetry in the less frequent triplet can be explained by hybridization, but could also be explained by ancient population structure (Slatkin and Pollack 2008). Distinguishing hybridization from processes such as ancient population structure and changing population sizes might be at least as challenging as distinguishing one hybridization network from another assuming that population structure and population sizes that do not fluctuate.

Summary

To summarize our results, we find that

- Two networks that display the same trees, including branch lengths and inheritance probabilities, might or might not be distinguishable under the NMSC in the sense of leading to the same probability distribution of gene tree topologies. In particular, there are examples where two networks display exactly the same trees, clades, triples, and quartets, yet are distinguishable from the probabilities of trees, clades, triples, and quartets.

- Network distinguishability can be improved in some cases by using branch length information and/or by sampling more than one individual per species descended from a hybrid population.

- Higher levels of incomplete lineage sorting can make inference of hybridization events easier in some cases.

- A desirable property of a network inference method is to be able to distinguish networks that are in fact distinguishable, even when they display the same trees. We have shown that maximum likelihood can do this in at least some cases.

We agree with Pardi and Scornavacca (2015) that identifiability is an important topic when trying to infer networks. Much of the effort in the literature on hybridization networks has focused on constructing networks that display a set of input trees, which are treated as data (Bordewich and Semple 2007; Holland et al. 2008; van Iersel and Linz 2013). From this point of view, it is crucial to understand when two networks might display the same set of trees.

Much less work has been done on what we are calling the NMSC, which has only recently become an active area of research. We have shown that identifiability results from the combinatorial point of view do not necessarily immediately transfer to the NMSC framework, and that many cases thought to be indistinguishable turn out to be distinguishable using this probabilistic modeling approach. An analogy is that in the case of trees (rather than networks), unrooted trees might not be expected to have any information about the root of the trees from which they came. However, under a probabilistic model, unrooted trees can have

information about the root (Steel 2012), and in particular, under the multispecies coalescent, the distribution of unrooted trees determines the rooted species tree when there are five or more taxa (but not for four taxa) (Allman et al. 2011b).

We hope that there will be more of an intersection in future phylogenetic network research between combinatorial approaches and the NMSC framework. A particular problem in need of more theoretical work is that of distances between networks. In particular, standard definitions of distance between networks, such as cluster-based definitions which extend Robinson–Foulds distances (Robinson and Foulds 1981) to networks (Cardona et al. 2009), return a distance of 0 between N_1 and N_2 and between N_1' and N_2'. This makes it difficult to determine whether an inferred network is closer to N_1 versus N_2 (or N_1' vs. N_2'), even for methods capable of distinguishing these networks.

The increased ability to distinguish networks using probabilistic models is good news for biologists interested in being able to infer biologically meaningful networks. However, much is still not understood about the space of networks in which we are interested in making inferences, and more theory is needed to determine what is and what is not distinguishable or identifiable under the NMSC. We showed that the particular example given by Pardi and Scornavacca (2015) turned out to be distinguishable if there is more than one lineage sampled from species B, and that generally there are cases of two networks that display exactly the same species trees (including branch lengths) that are nevertheless distinguishable under the NMSC, even with one lineage sampled per species.

However, we did not establish that N_1 and N_2 are distinguishable from all networks on four taxa, even if the number of hybridization nodes is capped. Nor did we establish that if, say, the topology of N_1 is known, then the parameters of the network would be identifiable. Here lack of identifiability would mean that two distinct sets of branch lengths and/or hybridization parameters (γ_i), lead to the same distribution of gene trees. Since networks of any complexity can be conceived, we can construct networks on n taxa with more parameters than there are gene tree topologies (assuming a fixed number of alleles sampled per species), and this will certainly result in lack of identifiability of the parameters from gene tree topologies even if the network topology is known. The challenge remains to determine what is and what is not identifiable for networks under the NMSC.

Supplemental Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.t2d38.

Software Availability

The software hybrid-coal, which computes gene tree probabilities recursively, and hybrid-Lambda, which simulates gene trees in phylogenetic networks, are both available at Github at https://github.com/hybridLambda/ and can be freely used under GNU GPL Version 3 or later. Both have been tested in Mac OS X 10.9.5 and Ubuntu.

Funding

Much of this work was completed while SZ was a PhD student at University of Canterbury, supervised by JD and Mike Steel, [SZ and JD were funded by the New Zealand Marsden fund during 2010–2013]. JD was additionally supported by National Institutes of Health [grant R01 GM117590].

Acknowledgments

We are grateful to David Bryant and Mike Steel for comments on the description of the algorithm. We thank Bengt Oxelman and two anonymous reviewers for additional helpful comments.

Appendix

Recursive Method to Compute Gene Tree Probabilities Given Species Networks

In this section, we introduce a novel method to compute gene tree probabilities of a given species network. Nodes of the network are visited in a modified post-order traversal so that the algorithm works on the deepest nodes descended from hybrid nodes first and works from this node toward the root until all hybrid nodes are eliminated. We introduce two operations to decompose a network into reduced networks that have a smaller number of edges or nodes. The post-order traversal ensures that we perform the simplification operations in a correct order—a node is never operated on before removing its descendant internal nodes.

Decomposition Operations

In this section, we propose two operations to simplify a complex phylogeny structure into simpler structures with fewer hybridization events. To demonstrate this procedure, we first consider simple cases where one individual is sampled from each population at the present. We first make several restrictions and assumptions for the gene tree T and the network W in this section:

- The gene tree T and the network W are rooted.
- Gene tree T and network W have the same number of external edges.
- Gene tree T is binary.
- An interior node of W can only have at most two parent nodes; a hybrid node refers to an internal node of W which has two parent nodes.

- We do not consider the case that a hybrid node is also a leaf node.

The network W is initially reduced to a set of simpler networks ($SG(W)$) in a single step in the reduction process. Let $P(T|W)$ be the probability of gene tree T given a species network W, by the law of total probability, we have the following:

$$\begin{aligned} P(T|W) &= \sum_{w^* \in SG(W)} P(T|W^*=w^*, W)P(W^*=w^*|W) \\ &= \sum_{w^* \in SG(W)} P(T|W^*=w^*)P(W^*=w^*|W), \end{aligned} \tag{A.1}$$

where W^* is a random variable that depends on W.

For any $w^* \in SG(W)$, w^* implies either a particular parental branch that some lineages have followed at a hybrid node or some specific coalescences that have occurred beneath a hybridization node.

Prior to decomposing a network, nodes are ranked from the leaves of the network to the top: tip nodes have rank one; an interior node's rank is one plus the highest rank of its child nodes.

The key to simplifying a network is to remove the interior nodes of the network in a specific order, along with the branches that are connected to the node. Here we define several functions to assist us identifying which nodes should be removed first. Let V be the set of nodes in the network; for $v \in V$, let $r(v)$ be the rank of v (the number of edges from v to the root), and $p(v)$ be the number parent node of v. We use indicator function $h(v)$ to identify if a node v is a hybrid node:

$$h(v) = \begin{cases} 1, & \text{if } p(v)=2; \\ 0, & \text{otherwise.} \end{cases}$$

Let $hd(v)$ and $t(v)$ be the indicator functions that take values

$$hd(v) = \begin{cases} 1, & \text{if } v \text{ is a descendant node of a hybrid node;} \\ 0, & \text{otherwise;} \end{cases}$$

and

$$t(v) = \begin{cases} 1, & \text{if } v \text{ is a leaf node;} \\ 0, & \text{otherwise} \end{cases}$$

respectively.

Thus, we can apply Algorithm 1 to find which node should be removed from the network: If the algorithm returns value -1, it means that W is already tree-like, and does not need to be simplified; otherwise, it returns the index of the node that we need to perform the following operations.

Decomposition operation 1.—If the chosen node is an interior descendant node s of a hybrid node, then this implies that s has a single parent node (otherwise s is a hybrid node), and child nodes of s are the leaf nodes of W (since s has the lowest rank beside the tips). The first step of operation 1 is to remove s from W, along with all of the edges that are connected to s.

Algorithm 1 Algorithm to choose the *index* of the node to be removed in order to simplify a network.

```
index = |V| − 1; I = 1;
for I < |V| do
    if (h(v_I) + hd(v_I)) ∗ (1 − t(v_I)) ≥ 1 and r(v_I) < r(v_index) then
        index = I;
    end if
    I = I + 1;
end for
if I = |V| − 1 then
    return index = −1
else
    return index
end if
```

Let D denote all of the leaf nodes descended from s. We now enumerate all possible ways to partition D. For example, if $D=\{\alpha_1,\alpha_2,\alpha_3\}$, let D' be one of the possible partitions of D. D' could be $\{\{\alpha_1\},\{\alpha_2\},\{\alpha_3\}\}$, $\{\{\alpha_1,\alpha_2\},\{\alpha_3\}\}$, $\{\{\alpha_1\},\{\alpha_2,\alpha_3\}\}$, $\{\{\alpha_1,\alpha_3\},\{\alpha_2\}\}$ or $\{\{\alpha_1,\alpha_2,\alpha_3\}\}$. We treat every element of any D' as a new leaf node. In the second part of operation 1, we create a new graph w^*, by connecting the elements of D' to the parent node of s. Notice, if the element of D' contains more than one leaf node, this implies that by changing from graph W to w^*, we need to coalesce these leaves on the branch that connects s and its parent node.

To calculate the probability of these events, we let $u=|D|$, and $v=|D'|$ and t be the branch length from s to its parent node. Then the probability of u lineages coalesce into v lineages within time t is (Tajima 1983; Saunders et al. 1984; Takahata and Nei 1985; Rosenberg 2002; Degnan and Salter 2005):

$$g_{ij}(t) = \sum_{k=j}^{i} e^{-\binom{k}{2}t} \frac{(2k-1)(-1)^{k-j}}{j!(k-j)!(j+k-1)} \times \prod_{y=0}^{k-1} \frac{(j+y)(i-y)}{(i+y)}. \tag{A.2}$$

Therefore, we have:

$$P(W^*=w^*|W) = \frac{w}{c} g_{ij}(t) \mathcal{I}_{w^*}(T), \text{ for } w^* \in SG(W), \tag{A.3}$$

where c is the number of ways for i lineages to coalesce into j lineages, which is equal to $\prod_{k=j}^{i}\binom{k}{2}$, and w is the number of sequences of coalescences resulting in the same topology with i lineages coalescing into j lineages. This is equal to $w=(i-j)!\prod_{k=1}^{i-j}\frac{1}{1+a_k}$, where a_j is the number of interior nodes that are descended from the coalesced nodes (Degnan and Salter 2005), and c is the number of ways for i lineages to coalesce into j lineages, which is equal to $\prod_{i=v}^{u}\binom{i}{2}$. The indicator function is

defined as

$$\mathcal{I}_{w^*}(T)=\begin{cases}1, & \text{if the lineages in } w^* \text{ can lead to topology } T;\\ 0, & \text{otherwise.}\end{cases}$$

For instance, if the gene tree is $(((a,d),c),b)$ and $w^*=W_1$ in Fig. 3, then $\mathcal{I}_{w^*}(T)=0$.

Operation 1 removes an internal node of network W. Therefore, any reduced network w^*, $w^*\in SG(W)$, has fewer interior nodes than network W.

Decomposition operation 2.—Before applying operation 2 on a hybrid node h of W, we need to make sure that operation 1 has been applied to all of the interior nodes descended from h. This implies that all of the child nodes of h are the leaf nodes of W. Let p_L and p_R be the two parent nodes of h. We use H to denote the set of child nodes of h and $\mathbb{C}_H$ to denote the collection of all of the subsets of H. The first step of operation 2, is to remove h from W, and all of the edges connected to h.

We then introduce two new nodes, h_L and h_R. For any $L\in\mathbb{C}_H$, we have a new graph w^*, connect $l\in L$ to h_L, then connect h_L to p_L, and connect $r\in H\setminus L$ to h_R, then connect h_R to p_R. Let $m_L=|L|$, $m_R=|H\setminus L|$, and $m=|H|$. The parameter γ is the probability that one lineage is attached to p_L. Thus, we obtain the set of simpler networks $SG(W)$ and the probabilities $P(W^*=w^*|W)$ for any $w^*\in SG(W)$:

$$P(W^*=w^*|W)=\gamma^{m_L}(1-\gamma)^{m_R},\ \text{ where } m_L+m_R=m. \quad (A.4)$$

Operation 2 removes an internal node of network W. The newly added nodes h_L and h_R are effectively external nodes: as all of the nodes descended from h_L and h_R are leaf nodes, we can treat h_L and h_R as leaf nodes, but sampling multiple lineages from each of them. Therefore, any reduced network w^*, $w^*\in SG(W)$, has fewer interior nodes than network W.

Simplifying a Network Recursively

Operations 1 and 2 are applied recursively on any networks in $SG(W)$ until all of the simplified network structures are tree-like. Gene tree probabilities can be computed using either coalescent histories (Degnan and Salter 2005) or ancestral configurations (Wu 2012). The approach outlined in this article will therefore reduce the probability of a gene tree, given a species network, to a linear combination of gene tree probabilities of given species trees.

Let $AG_T(W)$ be an ordered list of directed graphs (trees or networks). Then $|AG_T(W)|$ is the number of elements in the list. Here we borrow the concepts of set operations "$\bigcup$" and "$\setminus$" for our use. Let $AG_T(W)\bigcup SG(W)$ denote gradually appending the elements of $SG(W)$ to the end of the list $AG_T(W)$, then indexing the new elements of $AG_T(W)$ from $|AG_T(W)|+1$ to $|AG_T(W)|+|SG(W)|$. For an element G of $|AG_T(W)|$, we define operation $AG_T(W)\setminus\{G\}$, as removing the element G from $AG_T(W)$, the index of any element behind G is now one less.

Algorithm 2 Recursive algorithm for simplifying a network.

```
Initialize AG_T(W)={W} and I=1;
while I≤|AG_T(W)| do
    Apply Algorithm 1 to G_I, G_I∈AG_T(W) to choose the index of the node that needs to be removed;
    if index is positive then
        if p(v_index) is 1 then
            Perform decomposition operation 1 on v_index, obtain SG(G_I).
        else
            Perform decomposition operation 2 on v_index, obtain SG(G_I).
        end if
        AG_T(W) ← AG_T(W)\{G_I}.
        AG_T(W) ← AG_T(W)⋃SG(G_I).
    else
        I=I+1;
    end if
end while
```

Then we apply Algorithm 2 to simplify a network W, and then compute the probability for gene tree T.

During the decomposition process, different sequences of removing the hybrid nodes may lead to the same subspecies trees W'. For $W'\in AG_T(W)$, we use $C(W,W')$ to denote the collection of ways to decompose W into W'. Each sequence of decomposition corresponds to a unique weight ω_c. Thus by simplifying Equation (A.1), we have:

$$P(T|W)=\sum_{W'\in AG_T(W)} P(T|W^*=W',W)\sum_{c\in C(W,W')}\omega_c. \quad (A.5)$$

Figure 3 illustrates the decomposition of a species network on four taxa with two hybridization nodes.

Notice that even though W_1 and W_{14} have the same topology, the branch lengths of these two trees differ. We consider them to be different species trees. For different gene trees, according to coalescent events, $AG_T(W^2_{\text{IV}})$ may differ. For example, if the gene tree is $(((a,d),c),b)$, $AG_T(W^2_{\text{IV}})=\{W'_4,W'_5,\ldots,W'_{12}\}$, but when the gene tree is $(((a,b),c),d)$, $AG_T(W^2_{\text{IV}})=\{W'_1,W'_2,\ldots,W'_{12}\}$.

References

Abbott R.J., Hegarty M.J., Hiscock S.J., Brennan A.C. 2010. Homoploid hybrid speciation in action. Taxon 59:1375–1386.

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

Allman E.S., Degnan J.H., Rhodes J.A. 2011a. Determining species tree topologies from clade probabilities under the coalescent. J. Theor. Biol. 289:96–106.

Allman E.S., Degnan J.H., Rhodes J.A. 2011b. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62:833–862.

Allman E.S., Rhodes J.A., Sullivant S. 2012. When do phylogenetic mixture models mimic other phylogenetic models? Syst. Biol. 61:1049–1059.

Ané C. 2010. Reconstructing concordance trees and testing the coalescent model from genome-wide data sets. In: Knowles L. L., Kubatko L. S., editors. Estimating species trees: theoretical and practical aspects. Hoboken, (NJ): Wiley-Blackwell. p. 35–52.
Bapteste E., van Iersel L., Janke A., Kelchner S., Kelk S., McInerney J.O., Morrison D.A., Nakhleh L., Steel M., Stougie L., Whitfield J. 2013. Networks: expanding evolutionary thinking. Trends Genet. 29:439–441.
Baroni M., Semple C., Steel M. 2006. Hybrids in real time. Syst. Biol. 55:46–56.
Bordewich M., Semple C. 2007. Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. IEEE/ACM Trans. Comp. Biol. Bioinform. 4:458–466.
Cardona G., Llabrés M., Rosselló F., Valiente G. 2009. Metrics for phylogenetic networks I: Generalizations of the Robinson-Foulds metric. IEEE/ACM Trans. Comp. Biol. Bioinform. 6:46–61.
Chen Z.-Z., Wang L. 2010. Hybridnet: a tool for constructing hybridization networks. Bioinformatics 26:2912–2913.
Chifman J., Kubatko L. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J. Theor. Biol. 374:35–47.
Comes H.P., Kadereit J.W. 1998. The effect of quaternary climatic changes on plant distribution and evolution. Trends Plant Sci. 3:432–438.
Cranston K.A. 2010. Summarizing gene tree incongruence at multiple phylogenetic depths. In: Knowles L.L., Kubatko L.S., editors. Estimating species trees: practical and theoretical aspects Hoboken (NJ): Wiley-Blackwell. p. 129–143.
DeGiorgio M., Degnan J.H., Rosenberg N.A. 2011. Coalescence-time distributions in a serial founder model of human evolutionary history. Genetics 189:579–593.
DeGiorgio M., Degnan J.H. 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. Mol. Biol. Evol. 27:552–569.
DeGiorgio M., Degnan J.H. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst. Biol. 63:66–82.
Degnan J.H., Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.
Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.
Degnan J.H. 2010. Probabilities of gene trees with intraspecific sampling given a species tree. In: Knowles L.L., Kubatko L.S., editors. Estimating Species Trees: Practical and Theoretical Aspects Wiley-Blackwell. p. 53–78.
Ebersberger I., Galgoczy P., Taudien S., Taenzer S., Platzer M., von Haeseler A. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266–2277.
Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. Syst. Biol. 59:307–321.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Holland B.R., Benthin S., Lockhart P.J., Moulton V., Huber K.T. 2008. Using supernetworks to distinguish hybridization from lineage-sorting. BMC Evol. Biol. 8:1.
Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.
Hudson R. 2002. Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics 18:337–338.
Huson D.H., Scornavacca C. 2011. A survey of combinatorial methods for phylogenetic networks. Genome Biol. Evol. 3:23–35.
Huson D., Rupp R., Scornavacca C. 2010. Phylogenetic networks: concepts, algorithms and applications. New York: Cambridge University Press.
Jin G., Nakhleh L., Snir S., Tuller T. 2006. Maximum likelihood of phylogenetic networks. Bioinformatics 22:2604–2611.
Jones G., Sagitov S., Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62:467–478.
Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.
Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.
Marcussen T., Heier L., Brysting A.K., Oxelman B., Jakobsen K.S. 2015. From gene trees to a dated allopolyploid network: Insights from the angiosperm genus *Viola* (Violaceae). Syst. Biol. 64: 84–101.
Marshall D.C., Hill K.B., Fontaine K.M., Buckley T.R., Simon C. 2009. Glacial refugia in a maritime temperate climate: Cicada (*kikihia subalpina*) MTDNA phylogeography in New Zealand. Mol. Ecol. 18:1995–2009.
Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor. Popul. Biol. 75:35–45.
Morrison D.A. 2011. Introduction to phylogenetic networks. Uppsala: RJR Productions.
Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. 28: 719–728.
Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.
Nguyen Q., Roos T. 2015. Likelihood-based inference of phylogenetic networks from sequence data by phylodag. In: Algorithms for computational biology. Springer. p. 126–140.
Pardi F., Scornavacca C. 2015. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. PLoS Comput. Biol. e1004135.
Park H.J., Nakhleh L. 2012. Inference of reticulate evolutionary histories by maximum likelihood: the performance of information criteria. BMC Bioinform. 13:S12.
Rambaut A., Grassly N.C. 1997. Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comp. Appl. Biosci. 13:235–238.
Rhodes J.A., Sullivant S. 2012. Identifiability of large phylogenetic mixture models. Bull. Math. Biol. 74:212–231.
Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Rosenberg N.A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Pop. Biol. 61:225–247.
Saunders I.W., Tavaré S., Watterson G.A. 1984. On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. 16:471–491.
Slatkin M., Pollack J.L. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. Mol. Biol. Evol. 25:2241–2246.
Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12:e1005896.
Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA 109: 14942–14947.
Steel M. 2012. Root location in random trees: a polarity property of all sampling consistent phylogenetic models except one. Mol. Phylogenet. Evol. 65:345–348.
Strimmer K., Moulton V. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. 17: 875–881.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.
Takahata N., Nei M. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9:322.
van Iersel L., Kelk S., Lekić N., Scornavacca C. 2014. A practical approximation algorithm for solving massive instances of hybridization number for binary and nonbinary trees. BMC Bioinformat. 15:1.

van Iersel L., Linz S. 2013. A quadratic kernel for computing the hybridization number of multiple trees. Inform. Process. Lett. 113:318–323.

White M.A., Ané C., Dewey C.N., Larget B.R., Payseur B.A. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5:e1000729.

Wu Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66:763–775.

Yu Y., Degnan J.H., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 8:e1002660–e1002660.

Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. USA 111:16448–16453.

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

Yu Y., Than C., Degnan J.H., Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst. Biol. 60:138–149.

Zhu S., Degnan J.H., Goldstien S.J., Eldon B. 2015. Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. BMC Bioinform. 16:292.