The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection

Yun Yu[1], **James H. Degnan**[2,3], **Luay Nakhleh**[1]*

[1] Department of Computer Science, Rice University, Houston, Texas, United States of America, [2] Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, [3] National Institute of Mathematical and Biological Synthesis, Knoxville, Tennessee, United States of America

Abstract

Gene tree topologies have proven a powerful data source for various tasks, including species tree inference and species delimitation. Consequently, methods for computing probabilities of gene trees within species trees have been developed and widely used in probabilistic inference frameworks. All these methods assume an underlying multispecies coalescent model. However, when reticulate evolutionary events such as hybridization occur, these methods are inadequate, as they do not account for such events. Methods that account for both hybridization and deep coalescence in computing the probability of a gene tree topology currently exist for very limited cases. However, no such methods exist for general cases, owing primarily to the fact that it is currently unknown how to compute the probability of a gene tree topology within the branches of a phylogenetic network. Here we present a novel method for computing the probability of gene tree topologies on phylogenetic networks and demonstrate its application to the inference of hybridization in the presence of incomplete lineage sorting. We reanalyze a *Saccharomyces* species data set for which multiple analyses had converged on a species tree candidate. Using our method, though, we show that an evolutionary hypothesis involving hybridization in this group has better support than one of strict divergence. A similar reanalysis on a group of three *Drosophila* species shows that the data is consistent with hybridization. Further, using extensive simulation studies, we demonstrate the power of gene tree topologies at obtaining accurate estimates of branch lengths and hybridization probabilities of a given phylogenetic network. Finally, we discuss identifiability issues with detecting hybridization, particularly in cases that involve extinction or incomplete sampling of taxa.

Citation: Yu Y, Degnan JH, Nakhleh L (2012) The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection. PLoS Genet 8(4): e1002660. doi:10.1371/journal.pgen.1002660

Editor: Joseph Felsenstein, University of Washington, United States of America

Received October 10, 2011; **Accepted** March 5, 2012; **Published** April 19, 2012

Funding: This work was supported in part by NSF grant CCF-0622037, NSF grant DBI-1062463, grant R01LM009494 from the National Library of Medicine, and an Alfred P. Sloan Research Fellowship to LN. JHD was funded by the New Zealand Marsden Fund and the National Institute for Mathematical and Biological Synthesis, an institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award EF-0832858, with additional support from the University of Tennessee, Knoxville. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nakhleh@cs.rice.edu

Introduction

A molecular systematics paradigm that views molecular sequences as the characters of gene trees, and gene trees as characters of the species tree [1] is being increasingly adopted in the post-genomic era [2,3]. Several models of evolution for the former type of characters have been devised [4], while the coalescent has been the main model of the latter type of characters [5,6]. However, hybridization, a process that is believed to play an important role in the speciation and evolutionary innovations of several groups of plant and animal species [7,8], results in reticulate (species) evolutionary histories that are best modeled using a *phylogenetic network* [9,10]. Further, as hybridization may occur between closely related species, incongruence among gene trees may also be partly due to deep coalescence, and distinguishing between the two factors is hard under these conditions [11]. Therefore, to enable a more general application of the new paradigm, a phylogenetic network model that allows simultaneously for deep coalescence events as well as hybridization is needed [12]. This model can be devised by extending the coalescent model to allow for computing gene tree probabilities in the presence of hybridization. In this paper we focus on gene tree topologies and analyze the signal they contain for detecting hybridization in the presence of deep coalescence.

Applications of probabilities of gene tree topologies given species trees include determining statistical consistency (or inconsistency) of topology-based methods for inferring species trees [13–15], testing the multispecies coalescent model [13,16], determining identifiability of species trees using linear invariants of functions of gene tree topology probabilities [17,18], delimiting species [19], designing simulation studies for species tree inference methods [20–22], and inferring species trees [23,24]. We expect that similar applications may be useful for probabilities of gene tree topologies given species networks. In particular, it will be useful to be able to evaluate the performance of methods that infer species trees in the presence of hybridization as well as the performance of methods for inferring species networks. Knowing the distribution of gene tree topologies could also be useful for

Author Summary

Species trees depict how species split and diverge. Within the branches of a species tree, gene trees, which depict the evolutionary histories of different genomic regions in the species, grow. Evolutionary analyses of the genomes of closely related organisms have highlighted the phenomenon that gene trees may disagree with each other as well as with the species tree that contains them due to deep coalescence. Furthermore, for several groups of organisms, hybridization plays an important role in their evolution and diversification. This evolutionary event also results in gene tree incongruence and gives rise to a species phylogeny that is a *network*. Thus, inferring the evolutionary histories of groups of organisms where hybridization is known, or suspected, to play an evolutionary role requires dealing simultaneously with hybridization and other sources of gene tree incongruence. Currently, no methods exist for doing this with general scenarios of hybridization. In this paper, we propose the first method for this task and demonstrate its performance. We revisit the analysis of a set of yeast species and another of *Drosophila* species, and show that evolutionary histories involving hybridization have higher support than the strictly diverging evolutionary histories estimated when not incorporating hybridization in the analysis.

estimating the probability that two gene trees have the same topology, a quantity that is used in constructing the prior which models gene tree discordance in BUCKy [25], a program that is often used to estimate species trees or concordance trees.

A method for computing the probability mass function of gene tree topologies in the absence of hybridization (i.e., under the multispecies coalescent model is assumed) is given by Degnan and Salter [26]. However, to handle hybridization and deep coalescence simultaneously, this method has to be extended to allow for reticulate species evolutionary histories.

Indeed, attempts have been made recently for this very task [27–30], all of which have focused on very limited special cases where the phylogenetic network topology is known and contains one or two hybridization events, and a single allele sampled per species. However, a general formula for the probability of a gene tree topology given a general (any number of taxa, hybridizations, gene trees, and/or alleles) phylogenetic network has remained elusive.

A binary phylogenetic network topology W contains two types of nodes: *tree nodes*, each of which has exactly one parent (except for the root, which has zero parents), and *reticulation nodes*, each of which has exactly two parents. The edge incident into a tree node is called *tree edge*, and the edges incident into a reticulation node are called *reticulation edges*. In our context, we associate with a phylogenetic network W a vector of branch lengths $\boldsymbol{\lambda}$ (in units of $2N$ generations, where N is the effective population size in that branch) and a vector of hybridization probabilities $\boldsymbol{\gamma}$ (which indicates for each allele in a hybrid population its probability of inheritance from each of the two parent populations); see Text S1 for formal definition. The gene tree topology G can be viewed as a random variable with probability mass function $P_{W,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G=g)$. In this paper, we solve the aforementioned open problem by reporting on a novel method for computing the probability of a gene tree topology given a phylogenetic network, $P_{W,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G=g)$.

We illustrate the use of gene tree topology probabilities to estimate the values of species network parameters using the likelihood of the gene tree topologies. This application allows for disentangling hybridization and deep coalescence when analyzing a set of incongruent gene trees, as both events can give rise to similar incongruence patterns. Given a collection $\mathcal{G}$ of gene tree topologies, one per locus, in a set of sampled loci, the likelihood function is given by

$$L(W,\boldsymbol{\lambda},\boldsymbol{\gamma}|\mathcal{G}) = \prod_{g\in\mathcal{G}} P_{W,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G=g). \tag{1}$$

This formulation provides a framework for estimating the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ of an evolutionary history hypothesis W, given a collection of gene trees $\mathcal{G}$. Estimates of 0 or 1 for the entries in the $\boldsymbol{\gamma}$ vector reflect the absence of evidence for hybridization based on the gene tree topology distribution.

As gene tree topologies are estimated from sequence data, there is often uncertainty about them. In our method, we account for that in two ways: (1) by considering a set of gene tree topology candidates, along with their associated probabilities (produced, for example, by a Bayesian analysis), and (2) by considering for each locus the strict consensus of all optimal tree topologies computed for that locus (produced, for example, by a maximum parsimony analysis).

Finally, to account for model complexity, we employ a simple technique based on three information criteria, AIC [31], AICc [32] and BIC [33]. While these criteria have their shortcomings for model selection, the question of how to account for phylogenetic network complexity is still wide open and no methods exist for addressing it systematically [10].

We have implemented our method in the publicly available software package PhyloNet [34] and demonstrated its broad utilities in three domains. First, we reanalyze a *Saccharomyces* data set and a *Drosophila* data set, and find support for hybridization in both data sets. Second, we show the identifiability of the parameter values of certain reticulate evolutionary histories. Third, we highlight and discuss the lack of identifiability of the parameters in other scenarios that involve extinctions.

Materials and Methods

We begin by reviewing Degnan and Salter's method for computing the probability gene tree topologies on species trees, and then describe our novel extension to the case of species networks.

The probability of a gene tree topology within a species tree

Degnan and Salter [26] gave the mass probability function of a gene tree topology g for a given species tree with topology ψ and vector of branch lengths λ as

$$P_{\psi,\lambda}(G=g) = \sum_{h\in H_{\psi}(g)} \frac{\omega(h)}{d(h)} \prod_{b=1}^{n-2} \frac{\omega_b(h)}{d_b(h)} p_{u_b(h)v_b(h)}(\lambda_b), \tag{2}$$

which is taken over coalescent histories h from the set of all coalescent histories $H_{\psi}(g)$. The product is taken over all internal branches b of the species tree. The term $p_{u_b(h)v_b(h)}(\lambda_b)$ is the probability that $u_b(h)$ lineages coalesce into $v_b(h)$ lineages on branch b whose length is λ_b. And the terms $\omega_b(h)/d_b(h)$ and $\omega(h)/d(h)$ represents the probability that the coalescent events agree with the gene tree topology. In particular, $\omega_b(h)$ is the number of ways that coalescent events can occur consistently with the gene tree and $d_b(h)$ is the number of sequences of coalescences that give the number of coalescent events specified by h. However, this equation assumes that ψ is a tree and as such is inapplicable to reticulate evolutionary histories. Recently, this equation was adapted to very special cases of species phylogenies with hybridization [28–30]. However, none of these adaptations is general enough to allow for multiple hybridizations, multiple alleles per species, or arbitrary divergence patterns following hybridization. We present a novel approach for generalizing this equation to handle hybridization. Our approach is general enough in that it allows for computing gene tree probabilities on any binary phylogenetic network topology, thus overcoming limitations of recent works.

The probability of a gene tree topology within a species network

Our approach for computing the probability of a gene tree g given a species network W has three steps. First, W is converted into a multilabeled (MUL) tree T (a tree whose leaves are not uniquely labeled by a set of taxa; see Text S1); second, the alleles at the tips of g are mapped in every valid way to the tips of T; and, finally, the probability of g is computed as the sum, over all valid allele mappings, of probabilities of g given T (see Figure 1).

Step 1: Converting the phylogenetic network W to MUL tree T. Let W be a phylogenetic network on set $\mathcal{X}$ of species, and with branch lengths vector λ and hybridization probabilities vector γ. The conversion of W into a MUL tree is done as follows. Traversing the network W from the leaves towards the root, every time a reticulation node u is encountered, the two reticulation edges incident into it are removed, an additional copy of the subtree rooted at u's child is created, one copy is attached as child of one of u's original parents, and the other is attached as a child of u's other original parent. For example, in Figure 1, traversing the phylogenetic network from the leaves towards the root, the reticulation node u is encountered, two copies of the subtree rooted at its child (i.e., the most recent common ancestor of B and C) are created, and one is attached as a child of u's parent x, and the other is attached as a child of u's parent y, resulting in the MUL tree shown in the figure. In order to keep track of which branches in the MUL tree originated from the same branch in the phylogenetic network, we build during the conversion a mapping ϕ from the set of the MUL tree branches to the set of the phylogenetic network branches, such that $\phi(e)=e'$ if branch e in the MUL tree corresponds to branch e' in the phylogenetic network. We make use of ϕ in two ways. The first is in transferring the branch lengths and hybridization probabilities from N to the resulting MUL tree T, as illustrated briefly in Figure 1 and in more details in Text S1, and the second use is for computing the probabilities of gene trees, as becomes clearer below. Upon completion of this step of converting the phylogenetic network W, its branch lengths $\boldsymbol{\lambda}$ and hybridization probabilities $\boldsymbol{\gamma}$, the result is a MUL tree T along with its branch lengths $\boldsymbol{\lambda}'$, hybridization probabilities $\boldsymbol{\gamma}'$, and the branch mapping ϕ. The full description of the procedure NetworkToMULTree for achieving this conversion is given in Text S1.

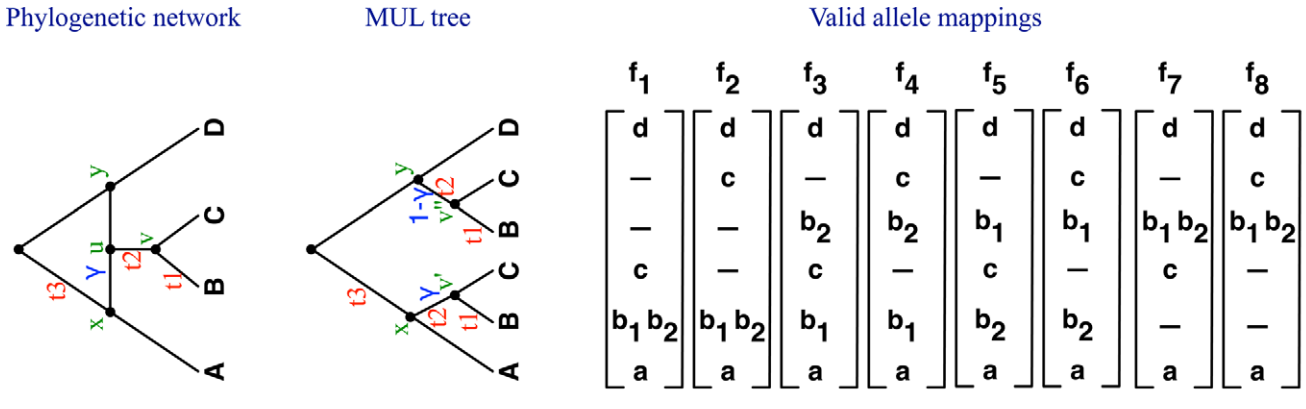
Step 2: Mapping the alleles to the leaves of the MUL tree. In computing the probability of a gene tree given a species phylogeny (tree or network), all the alleles sampled from species x are mapped to the single leaf labeled x in the species phylogeny. However, unless the species phylogeny W does not have any reticulation nodes, the resulting MUL tree T contains leaf sets that are labeled by the same species x. For example, in Figure 1, the MUL tree has two leaves labeled B and two leaves labeled C. In this case, it is important to map the alleles systematically to the leaves of the MUL tree so as to cover *exactly* all the coalescence patterns that would arise had the alleles been mapped to the phylogenetic network.

We denote by c_x the set of leaf nodes in T that are labeled by species x. For example, c_B for the MUL tree in Figure 1 is the set of the two leaves labeled by B. Now, consider a locus ℓ. We denote by A_x (for $x\in\mathcal{X}$) the set of alleles sampled from species x for locus ℓ, and by a_x the size of this set (i.e., $a_x=|A_x|$). In the example of Figure 1, two alleles were sampled from species B; hence, $A_B=\{b_1,b_2\}$ and $a_B=2$. A *valid allele mapping* is a function $f:(\bigcup_{x\in\mathcal{X}}A_x)\rightarrow(\bigcup_{x\in\mathcal{X}}c_x)$ such that if $f(a)=d$, and $d\in c_x$, then $a\in A_x$. In other words, f maps an allele from species x to a leaf in the MUL tree labeled by x. Let $\mathcal{F}$ denote the set of all such valid allele mappings f; in Figure 1, $\mathcal{F}=\{f_1,f_2,\ldots,f_8\}$.

Step 3: Computing the probability of a gene tree on the MUL tree. Once the phylogenetic network W is converted into MUL tree T and the set of all valid allele mappings is produced (a straightforward computational task, yet results in a number of valid allele mappings that is exponential in a combination of the number of alleles sampled and the number of reticulation nodes), the probability of observing gene tree topology g is found by summing the probability of g given the MUL tree over all possible allele mappings. Then, the probability of observing gene tree topology g is found by summing over all possible allele mappings:

$$P_{W,\lambda,\gamma}(G=g)=\sum_{f\in\mathcal{F}}P_{T,\lambda',\gamma',f}(G=g). \qquad (3)$$

In this equation, the $P_{T,\lambda',\gamma',f}$ term accounts for all coalescent histories of a given mapping, which, when combined with the



f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
d	d	d	d	d	d	d	d
—	c	—	c	—	c	—	c
—	—	b_2	b_2	b_1	b_1	b_1 b_2	b_1 b_2
c	—	c	—	c	—	c	—
b_1 b_2	b_1 b_2	b_1	b_1	b_2	b_2	—	—
a	a	a	a	a	a	a	a

Figure 1. Phylogenetic networks, MUL trees, and valid allele mappings. In this example, single alleles a, c, and d were sampled from each of the three species A, C, and D, respectively, whereas two alleles (b_1 and b_2) were sampled from species B. See text and Text S1 for details.
doi:10.1371/journal.pgen.1002660.g001

summation over all valid allele mappings, accounts for all coalescent histories within the branches of a phylogenetic network. Finally, the likelihood for a collection of gene trees is the product of the individual gene tree probabilities. This formulation naturally gives rise to a likelihood setup for estimating the parameters of a reticulate evolutionary history from a collection of gene trees described by their topologies.

To complete our framework, we now provide a formula for $P_{T,\lambda',\gamma',f}(G=g)$, which is the probability of a gene tree given a MUL tree and a valid allele mapping. Special attention needs to be paid to sets of branches in the MUL tree that correspond to single branches in the phylogenetic network, since coalescence events within these branches are not independent. Let us illustrate this issue using valid allele mapping f_3 and the MUL tree T in Figure 1. Under this mapping, each of the two alleles sampled from species B is mapped to a different B leaf in T. Tracing these two alleles independently from the two B leaves implicitly indicates that tracing the evolution of these two alleles in the phylogenetic network, no coalescence event should occur within time t_1 on the branch incident into leaf B in the network. Additionally, each branch in the MUL tree may have a hybridization probability associated with it that is neither 0 nor 1, and must be accounted for in computing the probabilities. Accounting for these two cases gives rise to

$$P_{T,\lambda',\gamma',f}(G=g)=\sum_{h\in H_{T,f}(g)}\frac{\omega(h)}{d(h)}\prod_{b=1}^{n-2}{\gamma'_b}^{v_b(h)}P'_b(h), \qquad (4)$$

where the $P'_b(h)$ terms are symbolic quantities, that do not individually evaluate to any value. Instead, they play a role in simultaneously computing the probability along pairs of branches in the MUL tree that share a single source branch in the phylogenetic network. More formally, let $b'=(u,v)$ be a branch in W such that u is a reticulation node. Given the mapping ϕ from the branches of T to the branches of W, the pre-image (or, inverse image) $\phi^{-1}(b')$ is the set of all branches in T that map to b' under ϕ. That is, $\phi^{-1}(b')=\{e\in E(T):\phi(e)=b'\}$, where $E(T)$ is the set of T's branches. Then, we define

$$u_{b'}(h)=\sum_{b\in\phi^{-1}(b')}u_b(h) \quad \text{and} \quad v_{b'}(h)=\sum_{b\in\phi^{-1}(b')}v_b(h). \qquad (5)$$

This equation states that the number of lineages $u_{b'}(h)$ that enters (working backward in time) branch b' in the phylogenetic network equals the sum of the numbers of lineages that enter all branches of the MUL tree that map to branch b'. The number of lineages $v_{b'}(h)$ that exists branch b' is defined similarly. In Figure 1, the number of lineages that enters branch $b'=(u,v)$ in the phylogenetic network equals the sum of the number of lineages that enter branch $b_1=(x,v')$ and the number of lineages that enter branch $b_2=(y,v'')$ in the MUL tree.

Then, we use the following equation to evaluate the probability in Equation (4):

$$\prod_{b\in\phi^{-1}(b')}P'_b(h)=\frac{1}{d_{b'}(h)}p_{u_{b'}(h)v_{b'}(h)}(\lambda_{b'})(u_{b'}(h)-v_{b'}(h))! \prod_{b\in\phi^{-1}(b')}\frac{\omega_b(h)}{(u_b(h)-v_b(h))!}, \qquad (6)$$

where $d_{b'}(h)$ is computed using the formula in [26], with $u_{b'}(h)$ and $v_{b'}(h)$ as parameters. In the example of branches b', b_1 and b_2 that we just illustrated, Equation (6) states that $P'_{b_1}(h)P'_{b_2}(h)$ evaluates to

$$\frac{1}{d_{b'}(h)}p_{u_{b'}(h)v_{b'}(h)}(\lambda_{b'})(u_{b'}(h)-v_{b'}(h))! \frac{\omega_{b_1}(h)}{(u_{b_1}(h)-v_{b_1}(h))!}\frac{\omega_{b_2}(h)}{(u_{b_2}(h)-v_{b_2}(h))!}.$$

The term $p_{u_{b'}(h)v_{b'}(h)}(\lambda_{b'})$ gives the probability that $u_{b'}(h)$ lineages coalesce into $v_{b'}(h)$ lineages within time $\lambda(b')$. The term

$$[(u_{b'}(h)-v_{b'}(h))!\prod_{b\in\phi^{-1}(b')}(\omega_b(h)/(u_b(h)-v_b(h))!)]$$

corresponds to the quantity $\omega_{b'}(h)$ in [26]. Finally, the term

$$\prod_{b\in\phi^{-1}(b')}(\omega_b(h)/(u_b(h)-v_b(h))!)$$

is the number of restrictions for the ordering of coalescent events within branch b'.

Accounting for uncertainty in gene tree topologies

Thus far, we have assumed that we have an accurate, fully resolved gene tree for each locus. However, in practice, gene tree topologies are inferred from sequence data and, as such, there is uncertainty about them. In Bayesian inference, this uncertainty is reflected by a posterior distribution of gene tree topologies. In a parsimony analysis, several equally optimal trees are computed. We propose here a way for incorporating this uncertainty into the framework above. Assume we have k loci under analysis, and for each locus i, a Bayesian analysis of the sequence alignment returns a set of gene trees $g_1^i,\ldots,g_q^i$, along with their associated posterior probabilities $p_1^i,\ldots,p_q^i$ ($p_1^i+\cdots+p_q^i=1$). Now, let $\mathcal{G}$ be the set of all distinct tree topologies computed on all k loci, and for each $g\in\mathcal{G}$ let p_g be the sum of posterior probabilities associated with all gene trees computed over all loci whose topology is g. Thus, $p_g=\sum_{i=1}^k p_g^i$ and $\sum_{g\in\mathcal{G}}p_g=k$. Then, we replace Eq. (1) by

$$L(W,\boldsymbol{\lambda},\boldsymbol{\gamma}|\mathcal{G})=\prod_{g\in\mathcal{G}}\left[P_{W,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G=g)\right]^{p_g}. \qquad (7)$$

We note that if $p_j^i=1$ or 0 for each i and j, then Eq. (7) is equivalent to Eq. (1), and both are multinomial likelihoods. This multinomial approach has also been used elsewhere for both species networks under simple hybridization scenarios [28] and species trees [24]. We additionally allow the p_j^i terms to be between 0 and 1 (and therefore p_g to be non-integer values) in order to reflect uncertainty in the estimated gene trees.

In the case where a maximum parsimony analysis is conducted to infer gene trees on the individual loci, a different treatment is necessary, since for each locus, all inferred trees are equally optimal. For locus i, let g be the strict consensus of all optimal gene tree topologies found. Then, Eq. (1) becomes

$$L(W,\boldsymbol{\lambda},\boldsymbol{\gamma}|\mathcal{G})=\prod_{g\in\mathcal{G}}\max_{g'\in b(g)}\{P_{W,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G=g')\}, \qquad (8)$$

where $b(g)$ is the set of all binary refinements of gene tree topology g.

Results

Support for hybridization in yeast

Using our method to compute the likelihood function given by Eq. (1), we reanalyzed the yeast data set of [35], which consists of 106 loci, each with a single allele sampled from seven Saccharomyces species *S. cerevisiae* (*Scer*), *S. paradoxus* (*Spar*), *S. mikatae* (*Smik*), *S. kudriavzevii* (*Skud*), *S. bayanus* (*Sbay*), *S. castellii* (*Scas*), *S. kluyveri* (*Sklu*), and the outgroup fungus *Candida albicans* (*Calb*). Given that there is no indication of coalescences deeper than the MRCA of *Scer*, *Spar*, *Smik*, *Skud*, and *Sbay* [36], we focused only on the evolutionary history of these five species (see Text S1). We inferred gene trees using Bayesian inference in MrBayes [37] and using maximum parsimony in PAUP* [38] (see Text S1 for settings).

The species tree that has been reported for these five species, based on the 106 loci, is shown in Figure 2A [35]. Further, additional studies inferred the tree in Figure 2B as a very close candidate for giving rise to the 106 gene trees, under the coalescent model [36,39]. Notice that the difference between the two trees is the placement of *Skud*, which flags hybridization as a possibility. Indeed, the phylogenetic network topologies in Figure 2C and 2D have been proposed as an alternative evolutionary history, under the stochastic framework of [40], as well as the parsimony framework of [30].

Using the 106 gene trees, we estimated the times t_1, t_2, t_3, t_4 and γ for the six phylogenies in Figure 2 that maximize the likelihood function (we used a grid search of values between 0.05 and 4, with step length of 0.05 for branch lengths, and values between 0 and 1 with step length of 0.01 for γ). Table 1 lists the values of the parameters computed using Eq. (7) on the gene trees inferred by MrBayes and Table 2 lists the values of the parameters computed using Eq. (8) on the gene trees inferred by PAUP*, as well as the values of three information criteria, AIC [31], AICc [32] and BIC [33], in order to account for the number of parameters and allow for model selection.

Out of the 106 gene trees (using either of the two inference methods), roughly 100 trees placed *Scer* and *Spar* as sister taxa, which potentially reflects the lack of deep coalescence involving this clade (and is reflected by the relatively large t_3 values estimated). Roughly 25% of the gene trees did not show monophyly of the group *Scer*, *Spar*, and *Smik*, thus indicating a mild level of deep coalescence involving these three species (and reflected by the relatively small t_2 values estimated). However, a large proportion of the 106 gene trees indicated incongruence involving *Skud*; see Text S1. This pattern is reflected by the very low estimates of the time t_1 on the two phylogenetic trees in Figure 2. On the other hand, analysis under the phylogenetic network models of Figure 2C and 2D indicates a larger divergence time, with substantial extent of hybridization. These latter hypotheses naturally result in a better likelihood score. When accounting for model complexity, all three information criteria indicated that these two phylogenetic network models with extensive hybridization and larger divergence time between *Sbay* and the (*Smik*,(*Scer*,*Spar*)) clade provide better fit for the data. Further, while both networks produced identical hybridization probabilities, the network in Figure 2D had much lower values of the information criteria than those of the network in Figure 2E. The networks in Figure 2E and 2F have lower support (under all measures) than the other four phylogenies. In summary, our analysis gives higher support for the hypothesis of extensive hybridization, a low degree of deep coalescence, and long branch lengths than to the hypothesis of a species tree with short branches and extensive deep coalescence. It is worth mentioning that while the three networks in Figure 2C–2E were reported as equally optimal under a parsimonious reconciliation [36], our new framework can distinguish among the three, and identifies the network in Figure 2D as best, followed by the one in Figure 2C (the network of Figure 2E is found to be a worse fit than either of the two species tree candidates).

Support for hybridization in *Drosophila*

We reanalyzed the three-species *Drosophila* data set of [41], which includes *D. melanogaster* (*Dmel*), *D. yakuba* (*Dyak*), and *D. erecta* (*Dere*).

The data set consisted of 9,315 loci supporting the three possible gene tree topologies as follows:

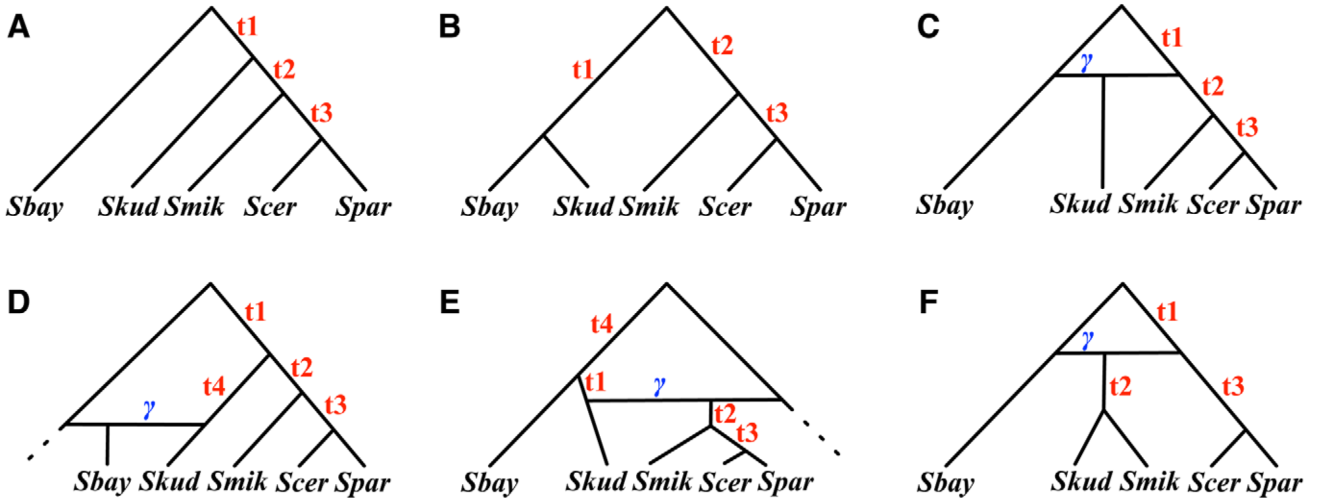


Figure 2. Various hypotheses for the evolutionary history of a yeast data set. (A) The species tree for the five species *Sbay*, *Skud*, *Smik*, *Scer*, and *Spar*, as proposed in [35], and inferred using a Bayesian approach [39] and a parsimony approach [36]. (B) A slightly suboptimal tree for the five species, as identified in [36,39]. (C–E) The three phylogenetic networks that reconcile both trees in (A) and (B), and which we reported as equally optimal evolutionary histories under a parsimony criterion in [30]. (F) A phylogenetic network that postulates *Smik* and *Skud* as two sister taxa whose divergence followed a hybridization event.
doi:10.1371/journal.pgen.1002660.g002

Table 1. Analysis results for the six phylogenies in Figure 1 using gene tree topologies inferred by a Bayesian analysis (using MrBayes).

Species phylogeny	t_1	t_2	t_3	t_4	γ	$-lnL$	AIC	AICc	BIC
Figure 1A	0.05	0.85	2.05	N/A	N/A	284	575	576	583
Figure 1B	0.2	0.85	2.05	N/A	N/A	276	559	560	567
Figure 1C	0.4	0.65	2.05	N/A	0.59	274	556	556	567
Figure 1D	2.95	0.7	2.1	0.85	0.5	247	504	504	517
Figure 1E	0.6	0.05	2.05	0.2	0.0	276	563	564	577
Figure 1F	0.9	0.05	2.15	N/A	0.27	325	659	659	669

doi:10.1371/journal.pgen.1002660.t001

- gene tree (*Dmel*,(*Dere*,*Dyak*)) is supported by 5,381 (57.8%) loci;
- gene tree ((*Dmel*,*Dere*),*Dyak*) is supported by 2,188 (23.5%) loci; and,
- gene tree ((*Dmel*,*Dyak*),*Dere*) is supported by 1,746 (18.7%) loci.

For a species tree with three species and one individual sampled per species, the multispecies coalescent predicts that the two gene trees with topologies different from that of the species tree each occur with probability $(1/3)\exp(-t)$, where t is the length of the one internal branch in coalescent units [42]. Two important predictions under the coalescent are therefore that the two nonmatching gene trees are expected to be tied in frequency and that both occur less than $1/3$ of the time, with the matching gene tree topology occurring more than $1/3$ of the time. This tie in the expected frequency of nonmatching gene trees is observed in some three-taxon data sets, but not in others, including the *Drosophila* data set.

Although this deviation from symmetry can be explained by a model of population subdivision, where the subdivision must occur in the internal branch as well as the population ancestral to all three species [43], the asymmetry can also be explained by the simplest hybridization network on three species with just one hybridization parameter (Figure 3).

We considered six candidates for the species phylogeny: three with no hybridization, and three with hybridizations involving different pairs of species (see Figure 3). For the three phylogenetic trees, we estimated the time t that maximizes the probability of observing all 9,315 gene trees, and for the three phylogenetic networks, we additionally estimated the hybridization probability γ.

The results in Table 3 show that of the three phylogenetic trees, the one in Figure 3A provides the best fit of the data, which is in agreement with the analysis in [41]. In fact, the value of t we estimated on the other two trees was the lowest value we used in the estimation procedure. Clearly, this value can be arbitrarily small for these two trees, since the unresolved phylogeny (*Dmel*, *Dere*, *Dyak*) fits the data better.

Among the three network candidates, the one in Figure 3D has the best fit of the data. This network, with a value of $\gamma = 0.11$, indicates that 89% of the alleles sampled from *Dere* shared a common ancestor first with alleles from *Dyak* (reflecting the tree in Figure 3A), while 11% of the alleles from *Dere* shared a common ancestor first with alleles from *Dmel* (reflecting the tree in Figure 3B). Indeed, this network is the smallest network (in terms of the number of reticulation nodes) that reconciles both trees. Further, the change in AIC for this network is $18143 - 18095 = 48$, indicating a much better fit than the best tree (Figure 3A). As noted previously [43], a χ-square test will also strongly reject the hypothesis that the species relationships are tree-like with random mating.

This three-taxon example can be analyzed analytically. Fitting a hybridization parameter allows a perfect fit to any observed frequencies of gene tree topologies for three species for one of the three networks in Figure 3. We let p_1, p_2, and p_3 represent the probabilities of topologies (*Dmel*,(*Dere*, *Dyak*)), ((*Dmel*, *Dere*), *Dyak*), and ((*Dmel*, *Dyak*), *Dere*) under the network in Figure 3D. Then

$$p_1 = (1-\gamma)(1-e^{-t}) + e^{-t}/3$$

$$p_2 = \gamma(1-e^{-t}) + e^{-t}/3$$

Table 2. Analysis results for the six phylogenies in Figure 1 using gene tree topologies inferred by maximum parsimony (using PAUP*).

Species phylogeny	t_1	t_2	t_3	t_4	γ	$-lnL$	AIC	AICc	BIC
Figure 1A	0.3	1.25	3.6	N/A	N/A	205	416	417	424
Figure 1B	0.2	1.35	3.6	N/A	N/A	208	423	423	431
Figure 1C	1.1	1.05	3.6	N/A	0.34	188	384	385	395
Figure 1D	3.45	1.15	3.6	3.05	0.34	157	325	326	338
Figure 1E	0.3	1.25	3.6	N/A	1.0	205	420	421	434
Figure 1F	1.55	0.05	3.7	N/A	0.18	252	512	512	523

doi:10.1371/journal.pgen.1002660.t002

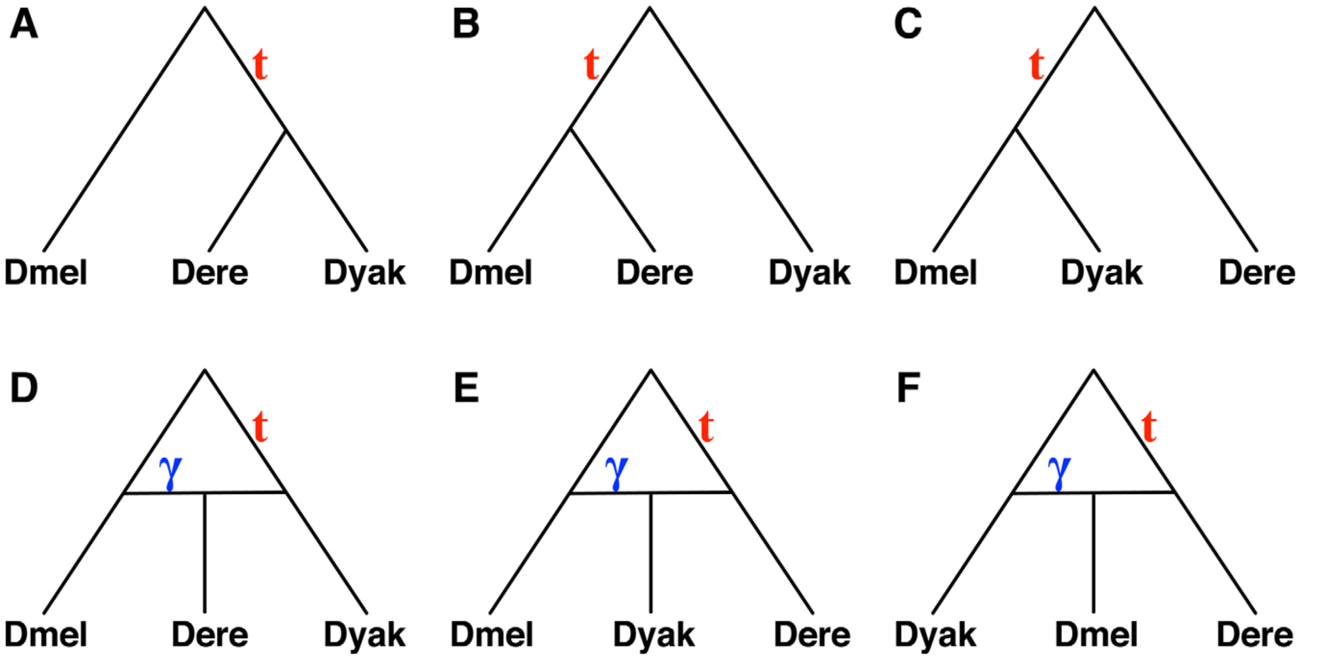


Figure 3. Six hypotheses for the evolutionary history of a *Drosophila* data set. (A–C) The three possible species tree topologies. (D–E) The three possible single-hybridization species network topologies (excluding extinction events).
doi:10.1371/journal.pgen.1002660.g003

$$p_3 = e^{-t}/3$$

This system has the unique solution

$$t = -\log(3p_3), \quad \gamma = \frac{p_2 - p_3}{1 - 3p_3} \tag{9}$$

for $p_3 < 1/3$ and $0 < p_3 < p_2, p_1$ (either at least one of the gene tree probabilities is less than $1/3$ if since they sum to 1.0; or if they are all exactly $1/3$, then a star tree with $t_3 = 0$ and any α exactly fits the data). Thus we can estimate t and γ using the observed $\hat{p}_2$ and $\hat{p}_3$ in equation (9), and this also maximizes the likelihood.

Identifiability of hybridization using gene tree topologies: A simulation study

For the simulated data, we evolved gene trees within the branches of phylogenetic networks, while varying branch lengths and hybridization probabilities, and investigated two questions: (1) how much data (gene trees) is needed to obtain accurate inference of the parameters (branch lengths and/or hybridization probabilities)? (2) are the parameters always identifiable? To answer these two questions, we investigated six different phylogenetic network topologies that involved single reticulation scenario, two reticulation scenarios (dependent and independent), and cases with extinctions involving the species that hybridize (see Text S1).

Our results show that both hybridization probabilities and branch lengths can be estimated with very high accuracy provided that no extinction events were involved in the parents of hybrid populations (see Text S1). Further, this accuracy can be achieved even when using the smallest number of gene trees we used in our study, which is 10. Under these settings, estimates using our framework seemed to converge quickly to the true values.

We also investigated the performance of the method, as well as identifiability issues when phylogenetic signal from at least one of the species involved in the hybridization is completely lost. Figure 4 shows the results for one such scenario (see Text S1 for another scenario that involves the loss of phylogenetic signal from both species involved in the hybridization).

Table 3. Estimates of time t and hybridization probability γ (when applicable) on the six candidate species phylogenies shown in Figure 2 for the three *Drosophila* species *Dmel*, *Dere*, and *Dyak*.

Species phylogeny	$-\ln L$	t	γ	AIC	AICc	BIC
Figure 2A	9070	0.46	N/A	18143	18143	18150
Figure 2B	10233	$1E-10$	N/A	20469	20469	20476
Figure 2C	10233	$1E-10$	N/A	20469	20469	20476
Figure 2D	9045	0.58	0.11	18095	18095	18109
Figure 2E	9070	0.46	0.0	18145	18145	18159
Figure 2F	10233	$1E-10$	0.0	20471	20471	20485

doi:10.1371/journal.pgen.1002660.t003

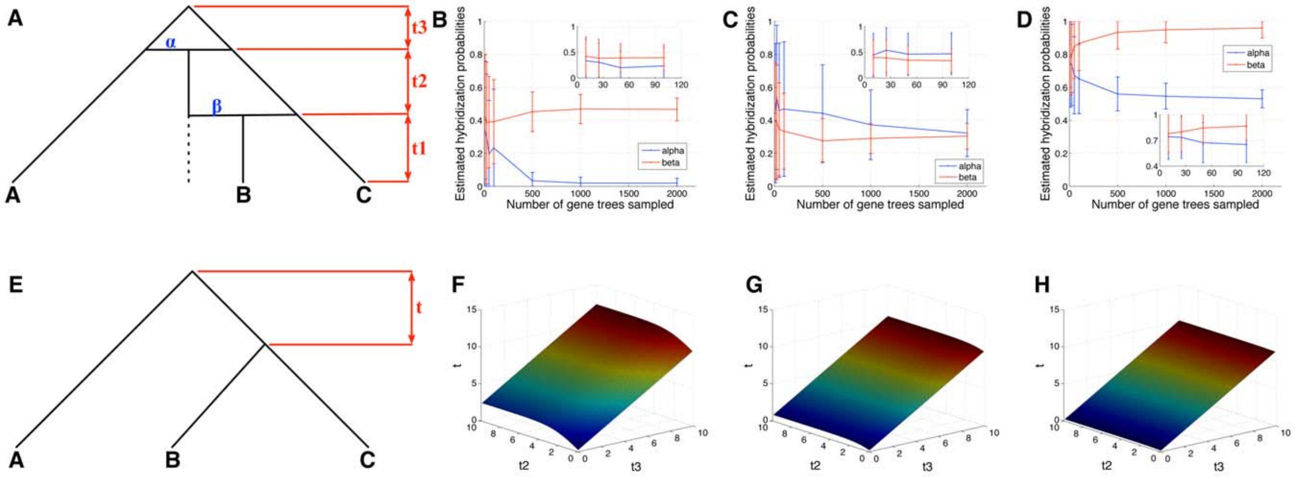


Figure 4. Identifiability in detecting hybridization. (A) A phylogenetic network with two hybridization probabilities, where the second hybridization involves the first hybrid population, and extinction is involved. (B–D) Estimates of α and β, as a function of the number of gene trees used, when the true values of $t_1 = t_2 = t_3 = 1.0$ are assumed in the inference, and for true (α,β) values of (0.0,0.5), (0.3,0.3), and (0.5,1.0), respectively (insets zoom in on the left parts of the figure). (E) A phylogenetic tree with three taxa, and with divergence time t between the two speciation events. (F–H) The value of t for the tree in (E) that yields the same probability of the data under the scenario depicted in (A) when $\alpha = 0.0$, as a function of t_2 and t_3, and for β value of 0.1, 0.5, and 0.9, respectively. Since a single allele was sampled per species, the data is uninformative for estimating the value of t_1 here.
doi:10.1371/journal.pgen.1002660.g004

Panels Figure 4B–4D show that when the true values of t_2 and t_3 are assumed to be known in the estimation procedure (the value of t_1 is irrelevant in the case when a single allele is sampled per species), the estimates of the hybridization probabilities converge to the true values. However, unlike the cases that did not involved extinctions, a larger number of gene trees is now required to obtain an accurate estimate (while there are only three possible gene tree topologies, a large number of gene trees need be sampled in order for the three topologies' frequencies to be informative). The time intervals of $t_2 = t_3 = 1.0$ coalescent units amount to a large extent of deep coalescence events, which blurs the phylogenetic signal, and results in slight over- or under-estimation of the hybridization probabilities (Text S1 shows the results for the time interval with $t_2 = t_3 = 2.0$).

If the topology of the network in Figure 4A is assumed to be known, but both the branch lengths and hybridization probabilities are to be estimated, then these parameters are unidentifiable; that is, two different pairs of vectors of branch lengths and hybridization probabilities can be found to explain the observed data with exactly the same probability (see Text S1). If at least two alleles are sampled from species B, then the parameter values become identifiable; however, an extremely large, and potentially infeasible, number of gene trees need to be sampled to uniquely identify the parameter values in practice (see Text S1).

Furthermore, in the special case where $\alpha = 0.0$, a phylogenetic tree, with appropriate branch lengths can be found, to fit the data exactly with the same probability that the phylogenetic network would. Let λ be the branch lengths vector with $\lambda_1 \equiv t_1$, $\lambda_2 \equiv t_2$, and $\lambda_3 \equiv t_3$, and let γ be the hybridization probabilities vector with $\gamma_1 \equiv \beta$. Now, consider the phylogenetic tree T in Figure 4E. Then, if we set t as a function of β, t_2, and t_3, using $t(\beta,t_2,t_3) = -\ln(\beta e^{t_2} + 1 - \beta) + t_2 + t_3$, then, $P_{W,\lambda,\gamma}(g) = P_{T,t}(g)$ for any gene tree g. The values of $t(\beta,t_2,t_3)$ are shown in Figure 4F–4H. These results show that as t_2 increases, the value of t becomes unaffected by t_2, and that increasing t proportionally to the increase in t_3 always maintains identical probabilities of gene trees under both species phylogenies (see Text S1).

Our method for computing the probability of gene trees under hybridization and deep coalescence allows for analyzing data sets with arbitrary complexity of evolutionary histories in terms of the hybridization scenarios. When parameters are identifiable, our method estimates their values with high accuracy from a relatively small number of loci. Further, our method can be used to show lack of identifiability of model parameters for other cases. Our method supports a hypothesis of larger divergence time coupled with hybridization over short divergence times (with extensive deep coalescence) in a yeast data set. Finally, for a large *Drosophila* data set, our method indicated no hybridization based on the sampled loci.

Discussion

Using coalescence times versus topologies to infer species networks

We have focused on calculating probabilities of gene tree topologies and using these probabilities to infer species networks. In addition, the joint density of the coalescence times and topology in the gene trees could be used to infer species networks. Indeed, this approach has been used for networks where reticulation nodes have one descendant which is an extant species [29], using the density for coalescence times derived by Rannala and Yang [44]. This approach is computationally faster than computing gene tree topology probabilities because it is not necessary to sum over a large number of coalescent histories. To compute this joint density, each gene sampled can potentially have to trace through up to $\Pi_{i=1}^{n} 2^{m_i}$ possible paths through the network, where m_i is the number of hybridization events ancestral to the sampled gene from species i, and the density will take the form of a sum over possible paths through the network. (In contrast, computing the probability of a topology will require $\Pi_{i=1}^{n} 2^{m_i}$ mappings of alleles to the MUL-tree, and each gene topology calculation will require summing over coalescent histories.) This joint density for the gene trees with coalescence times could then be used in either maximum likelihood or Bayesian frameworks to infer the species network.

An important advantage of using coalescence times is that certain networks might be identifiable using coalescence times when probabilities of topologies might not identify the network. In the example of Figure 3A, although the gene tree topology probabilities can be obtained by a tree, the distribution of the coalescence times between lineages sampled from B and C is a mixture of three shifted exponential distributions if $\alpha > 0$, but a mixture of two shifted exponential distributions if $\alpha > 0$. For example, if t_1, t_2, and t_3 are known but α and β are unkown, then the likelihood of observing a coalescence between a B and C lineage for times slightly greater $t_1 + t_2$ will be very low if $\alpha = 0$, and much higher for $\alpha > 0$, thus making it possible to test whether $\alpha = 0$ when coalescence times are used.

Another identifiability issue is that both population subdivision and hybridization can lead to the asymmetry in gene tree topology probabilities in the 3-taxon case such as observed in the *Drosophila* example discussed earlier, where the two least frequently observed topologies are not tied in frequency. Either population subdivision, with a parameter describing the probability that the two most closely related species fail to coalesce in the ancestral population due to population structure, or hybridization can fit the data for the gene tree topologies. However, the two models could imply different distributions on coalescence times, which might therefore be useful in distinguishing the models. We note that identifiability in the case of three species with one individual per species might be especially limited due to the small number of gene tree topology probabilities that can be used to estimate parameters. In the case of identifying rooted species trees from unrooted gene trees with one lineage per species, for example, identifiability is achieved only with 5 or more species [17].

We consider it desirable to develop many methods for inferring species trees and species networks so that their properties and performances can be compared. In the case of species tree inference, there are advantages and disadvantages to using topology-based methods versus methods that include branch lengths, and in using likelihood versus Bayesian methods. We expect that many of these strengths and weaknesses may carry over to the case of inferring networks. For moderately sized data sets, Bayesian methods that model branch lengths and uncertainty in the gene trees such as BEST [45] and *BEAST [46] often have the best performance [47]. However, these methods require estimating the joint posterior distribution of the species tree and gene trees and therefore are difficult to implement for large numbers of loci. Maximizing the likelihood of the gene trees and their coalescent times (but without accounting for uncertainty in the gene trees), as in STEM [48], is fast and has very good performance on known gene trees but seems to be very sensitive to the assumption that branch lengths are estimated correctly [24,49]. Maximizing the likelihood of the species tree using only gene tree topologies using the program STELLS, even while not accounting for uncertainty in the gene trees, tended to have better performance than STEM for a large simulated data set (>100 loci on 8 taxa) and worse performance on fewer loci [24]. Which method is optimal for inferring species trees or networks might depend on many factors such as the number of loci, the number of lineages sampled per species, the accuracy with which branch lengths can be estimated, the extent to which there are model violations, and the speciation history [49].

Recombination and population size assumptions

Two common assumptions in multispecies coalescent models are that there is no recombination within loci (and free recombination between loci) and that ancestral population sizes are constant.

Recombination can lead to different portions of a gene alignment effectively having distinct gene tree topologies. Ideally, alignments should be chosen so that recombination within genes is unlikely. This can be achieved by testing alignments beforehand for recombination using many available methods [50–52], or for whole genome data, choosing the cutoffs for loci such that they are unlikely to occur at recombination breakpoints [53]. In addition, recombination may lead to greater violations of the coalescent model for branch lengths than for topologies [53], so that topology-based methods might be less sensitive to the assumption that there is no recombination within loci. In addition, a recent simulation study found that recombination within loci did not have much impact on species tree inference methods for a wide range of recombination rates [54].

Coalescent models often assume that ancestral populations have constant size for the duration of the population (i.e., a constant size for a given branch of the species tree, but not necessarily the same on different branches). The program *BEAST [46] allows for ancestral population sizes to change linearly with time. Nonconstant population sizes will tend to result in branch lengths that make topologies more (or less) star-like for populations that are increasing (or decreasing) in size [55]. One approach to modelling a changing population size would be to break up a branch into intervals that are relatively constant in size. Suppose, for instance that a branch consists of an interval of τ_1 generations with population size N_1, and τ_2 generations with size N_2. The total time of the branch in coalescent units is $t = \tau_1/N_1 + \tau_2/N_2$. Although unequal values of N_i can affect the distribution of coalescence times (for example, if $\tau_1 = \tau_2$ but $N_1 > N_2$, then coalescence events might be more likely to occur in the interval with size N_2), the probabilities of topologies arising in this branch are not affected and can be calculated just using the total time t. In particular, for the functions $p_{u,v}(t)$, which are the terms that depend on time in the calculations for gene tree topology probabilities, we have

$$p_{u,v}(t) = p_{u,v}(\tau_1/N_1 + \tau_2/N_2) = \sum_{k=v}^{u} p_{u,k}(\tau_1/N_1) p_{k,v}(\tau_2/N_2),$$

which is an instance of the Chapman-Kolmogorov equations because the number of lineages is a continuous time Markov chain (a death chain) [56].

We expect that topology-based methods may show more robustness to recombination and changing population sizes than approaches which explicitly model coalescence times. However, for estimating species trees and networks from gene trees, as in other areas of statistical inference, there is likely to be a tradeoff between power and robustness for methods that do and do not model branch lengths of the gene trees.

Searching for networks

A current limitation to the procedure we have outlined for estimating hybridization is that we require a set of candidate networks on which to perform model selection. In some cases, such a set of candidate networks can be obtained by considering specific hypotheses related to biogeographical information. Candidate networks can also be generated using supernetworks from gene trees [57] or other network methods [9]. Often these methods will generate very complicated networks if there are many conflicts in the data, so it might be useful to choose different random subsets of well-supported (or frequently occurring) gene tree topologies to generate candidate species networks. In the future it will be desirable to develop algorithms that directly search the space of species networks in order to automate searching for optimal species networks.

Supporting Information

Text S1 Supporting information file that contains formal definitions and additional results on synthetic data.
(PDF)

Acknowledgments

We are grateful to J. Felsenstein and two anonymous reviewers for comments which have improved the manuscript.

Author Contributions

Conceived and designed the experiments: YY JHD LN. Performed the experiments: YY. Analyzed the data: YY JHD LN. Contributed reagents/materials/analysis tools: YY. Wrote the paper: YY JHD LN.

References

1. Doyle JJ (1992) Gene trees and species trees: molecular systematics as one-character taxonomy. Syst Bot 17: 144–163.
2. Maddison W (1997) Gene trees in species trees. Syst Biol 46: 523–536.
3. Edwards SV (2009) Is a new and general theory of molecular systematic biology emerging? Evolution 63: 1–19.
4. Swofford D, Olsen G, Waddell P, Hillis D (1996) Phylogenetic inference. In: Hillis D, Mable B, Moritz C, eds. Molecular Syst Biol.s. Sunderland, Mass.: Sinauer Assoc. pp 407–514.
5. Rosenberg NA (2002) The probability of topological concordance of gene trees and species trees. Theor Pop Biol 61: 225–247.
6. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24: 332–340.
7. Arnold ML (1997) Natural Hybridization and Evolution. Oxford: Oxford University Press.
8. Mallet J (2007) Hybrid speciation. Nature 446: 279–283.
9. Huson D, Rupp R, Scornavacca C (2010) Phylogenetic Networks: Concepts, Algorithms and Applications. New York: Cambridge University Press.
10. Nakhleh L (2010) Evolutionary phylogenetic networks: models and issues. In: Heath L, Ramakrishnan N, eds. The Problem Solving Handbook for Computational Biology and Bioinformatics. New York: Springer. pp 125–158.
11. Mallet J (2005) Hybridization as an invasion of the genome. Trends Ecol Evol 20: 229–237.
12. Linder CR, Rieseberg LH (2004) Reconstructing patterns of reticulate evolution in plants. Am J Bot 91: 1700–1708.
13. Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA (2009) Properties of consensus methods for inferring species trees from gene trees. Syst Biol 58: 35–54.
14. Than CV, Rosenberg NA (2011) Consistency properties of species tree inference by minimizing deep coalescences. J Comput Biol 18: 1–15.
15. Wang Y, Degnan JH (2011) Performance of matrix representation with parsimony for inferring species from gene trees. Stat Appl Genet Mol 10: 21.
16. Ané C (2010) Reconstructing concordance trees and testing the coalescent model from genome- wide data sets. In: Knowles LL, Kubatko LS, eds. Estimating species trees: Theoretical and practical aspects. Hoboken, NJ: Wiley-Blackwell. pp 35–52.
17. Allman ES, Degnan JH, Rhodes JA (2011) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 62: 833–862.
18. Allman ES, Degnan JH, Rhodes JA (2011) Determining species tree topologies from clade probabilities under the coalescent. J Theor Biol 289: 96–106.
19. Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. Syst Biol 56: 887–895.
20. Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56: 17–24.
21. Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. Syst Biol 58: 468–477.
22. DeGiorgio M, Degnan JH (2010) Fast and consistent estimation of species trees using supermatrix rooted triples. Mol Biol Evol 27: 552–569.
23. Carstens B, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. Syst Biol 56: 400–411.
24. Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66: 763–775.
25. Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance factors. Mol Biol Evol 24: 412–426.
26. Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. Evolution 59: 24–37.
27. Than C, Ruths D, Innan H, Nakhleh L (2007) Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. J Comput Biol 14: 517–535.
28. Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theor Popul Biol 75: 35–45.
29. Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. Syst Biol 58: 478–488.
30. Yu Y, Than C, Degnan JH, Nakhleh L (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst Biol 60: 138–149.
31. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19: 716–723.
32. Burnham K, Anderson D (2002) Model selection and multi-model inference: a practical-theoretic approach. New York: Springer Verlag, 2nd edition.
33. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461–464.
34. Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9: 322.
35. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.
36. Than C, Nakhleh L (2009) Species tree inference by minimizing deep coalescences. PLoS Comput Biol 5: e1000501. doi:10.1371/journal.pcbi.1000501.
37. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.
38. Swofford DL (1996) PAUP*: Phylogenetic analysis using parsimony (and other methods). Sinauer Associates, Underland, Massachusetts, Version 4.0.
39. Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Natl Acad Sci U S A 104: 5936–5941.
40. Bloomquist EW, Suchard MA (2010) Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. Syst Biol 59: 27–41.
41. Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet 2: e173. doi:10.1371/journal.pgen.0020173.
42. Nei M (1987) Molecular Evolutionary Genetics. New York: Columbia University Press.
43. Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature Rev Genet 9: 477–485.
44. Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.
45. Liu L, Pearl DK (2007) Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol 56: 504–514.
46. Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27: 570–580.
47. Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: A com- parison of methods. Syst Biol 60: 126–137.
48. Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25: 971–973.
49. Huang H, He Q, Kubatko LS, Knowles LL (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.
50. Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from dna sequences: Computer simulations. P Natl Acad Sci USA 98: 13757–13762.
51. Bruen TC, Philippe H, Bryant D (2002) A simple and robust statistical test for detecting the presence of recombination. Genetics 172: 2665–2681.
52. Ruths D, Nakhleh L (2006) RECOMP: A parsimony-based method for detecting recombination. In: Proceedings of the 4th Asia Pacific Bioinformatics Conference. pp 59–68.
53. Ané C (2011) Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. Genome Biol Evol 3: 246–258.
54. Lanier H, Knowles L (2012) Is recombination a problem for species-tree analyses? Syst BiolIn press: DOI:10.1093/sysbio/syr128.
55. Wakeley J (2008) Coalescent Theory. Greenwood Village, CO: Roberts & Company.
56. Ross SM (2010) Introduction to Probability Models. New York: Academic Press, 10th edition.
57. Holland B, Benthin S, Lockhart P, Moulton V, Huber K (2008) Using supernetworks to distinguish hybridization from lineage-sorting. BMC Evol Biol 8: 202.