Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section: Phylogenetics

Assessing the fit of the multi-species network coalescent to multi-locus data

Ruoyi Cai [1] and Cécile Ané [1,2]*

[1]Department of Statistics, University of Wisconsin - Madison, Madison, 53706, USA and
[2]Department of Botany, University of Wisconsin - Madison, Madison, 53706, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: With growing genome-wide molecular data sets from next-generation sequencing, phylogenetic networks can be estimated using a variety of approaches. These phylogenetic networks include events like hybridization, gene flow, or horizontal gene transfer explicitly. However, the most accurate network inference methods are computationally heavy. Methods that scale to larger data sets do not calculate a full likelihood, such that traditional likelihood-based tools for model selection are not applicable to decide how many past hybridization events best fit the data. We propose here a goodness-of-fit test to quantify the fit between data observed from genome-wide multi-locus data, and patterns expected under the multi-species coalescent model on a candidate phylogenetic network.
Results: We identified weaknesses in the previously proposed TICR test, and proposed corrections. The performance of our new test was validated by simulations on real-world phylogenetic networks. Our test provides one of the first rigorous tools for model selection, to select the adequate network complexity for the data at hand. The test can also work for identifying poorly-inferred areas on a network.
Availability: Software for the goodness-of-fit test is available as a Julia package at https://github.com/cecileane/QuartetNetworkGoodnessFit.jl.
Contact: cecile.ane@wisc.edu
Supplementary information: Supplementary material is available at *Bioinformatics* online, and scripts are available at https://osf.io/eg6ju/.

1 Introduction

The importance of reticulate evolution is now widely recognized across all areas in the "tree" of life (Folk *et al.*, 2018). Consequently, many methods have been developed to estimate phylogenetic networks (reviewed in Degnan, 2018; Elworth *et al.*, 2019). These network phylogenies are trees with added reticulation branches to explicitly represent events such as hybrid speciation, introgression, gene flow, or horizontal gene transfer. Methods to estimate phylogenetic networks are either based on parsimony or based on likelihood calculations, and take as input either genome sequence data or gene trees inferred from multiple sequence alignments. Parsimony approaches identify a network that displays (or "almost" displays) all of the input gene trees, using the minimum number of reticulations (Yu *et al.*, 2013b; Wu, 2013; Markin *et al.*, 2019). One major disadvantage of parsimony-based methods is that there is no criterion to select the appropriate number of reticulations. Likelihood methods have the advantage to consider a model of evolution for how gene trees evolve within a given species network, including an inheritance parameter γ at each reticulation event to quantify the proportion of genes inherited from each parental population. Discordance between a gene tree and the species network due to incomplete lineage sorting within ancestral populations may be modelled by the multi-species coalescent (Meng and Kubatko, 2009; Yu *et al.*, 2012). Likelihood-based methods include maximum likelihood methods (Kubatko, 2009; Yu *et al.*, 2014), pseudo-likelihood methods to better scale with the number of species (Solís-Lemus and Ané, 2016; Yu and Nakhleh, 2015; Zhu and Nakhleh, 2018), and Bayesian approaches (Wen *et al.*, 2016; Wen and Nakhleh, 2017; Zhang *et al.*, 2018).

Likelihood-based methods are more accurate than parsimony-based methods but are dramatically slower, and even more so as the number of reticulation events increases. In addition, the network likelihood and pseudolikelihood scores, like the parsimony score, necessarily improve

as the user allows for more reticulations in the network. Therefore, the complexity of the network needs to be penalized to estimate the appropriate number of reticulations. Information criteria like AIC have been used, but are unfortunately not appropriate for the purpose of estimating a network because they do not account for the exploding number of network models when the number of reticulations is increased, so they run the risk of falsely detecting extra reticulations (Blair and Ané, 2020). The model selection problem is even harder with pseudolikelihood methods: a full likelihood is needed to perform a likelihood ratio test or to use information criteria like AIC or BIC. An empirical solution consists in finding a number of reticulations after which the rate of improvement in the pseudolikelihood score stabilizes. Bayesian methods circumvent the model selection problem because the number of reticulations is implicitly penalized by the prior distribution. Unfortunately, the scalability of these methods is very limited (Elworth *et al.*, 2019), and default priors typically use a very small expected number of reticulations to mitigate the computational burden.

In short, researchers currently face a dilemma: curtail their data set to a few taxa and use a Bayesian approach to estimate the number of reticulations, or use pseudolikelihood or parsimony approaches on more taxa, but with no rigorous way to estimate the number of reticulations. Moreover, model selection does not assess model adequacy (Brown and Thomson, 2018). We provide here a method to assess the adequacy of a candidate phylogenetic network. Our goal is to assess if a network of a given complexity is adequate to explain the data at hand, and if not, which areas in the network do not fit the data adequately. Our work builds on TICR (Stenz *et al.*, 2015), which was developed to test the adequacy of a candidate species tree (without reticulation) with polytomies to represent episodes of current or ancestral panmixia. Input data are quartet concordance factors (CFs), that is, the proportion of genes that support each four-taxon tree, considering every subset of four taxa from the entire taxon set. For each four-taxon set, the quartet CFs expected from the candidate species phylogeny under the multispecies coalescent are calculated, and compared to the quartet CFs observed in the data. If the observed quartet CFs are too far from the CFs expected from the species phylogeny, the four-taxon set is labelled as an "outlier", via an outlier test that returns a p-value. TICR then compares the distribution of these outlier p-values (across all four-taxon sets) to a uniform distribution, as expected if the species phylogeny provides a good fit to the quartet CF data, to provide an overall goodness-of-fit measure of the phylogeny.

In this work, we modify TICR to extend it to phylogenetic networks, and we eliminate fundamental flaws that we discovered in this test. More specifically, our test uses a different model for the distribution of the quartet CFs and new ways to conduct the outlier test for each four-taxon subnetwork. Finally, when assessing the overall goodness-of-fit of a candidate network, we propose a simulation-based method to account for the dependency between the outlier test results across four-taxon sets. This dependency is ignored by TICR's overall test, which is shown here to have an unacceptably inflated type I error rate.

In what follows, we provide background on quartet CFs expected under the multispecies coalescent, then describe our new method. We present simulations to study the effect of dependence across outlier tests, and a theoretical bound for a conservative correction to control for dependence. The new test is used on empirical data from prior studies. In particular, the conclusions of Stenz *et al.* (2015) are unchanged qualitatively.

2 The coalescent on phylogenetic networks

Phylogenetic networks use reticulation edges to depict events where a population received genetic material from two distinct parental lineages, such as via hybrid speciation, introgression, gene flow, or horizontal gene transfer (Fig. 1). Such an event is represented by a node that has two parent branches, called hybrid edges. Each hybrid edge e has an inheritance value γ_e that quantifies the proportion of genes the node inherited through e. In Fig. 1, one reticulation node is highlighted as a red dot. Its two parent edges are annotated by their inheritance probabilities, to illustrate a scenario with 10% introgression. Branch lengths and inheritance probabilities are used to calculate the distribution of gene trees evolving along the phylogenetic network, based on the multi-species coalescent process (Kubatko, 2009; Yu *et al.*, 2012). This model accounts for both incomplete lineage sorting (via the coalescent) and reticulation (via the network topology). To measure the fit of a candidate network to multi-locus data, one would ideally measure the fit between the expected distribution of gene trees from the network under the coalescent model, and the distribution of gene trees observed from the data. However, the space of trees is immense when there are more than a handful of taxa (Semple and Steel, 2003), which makes it difficult to compare distributions of full gene trees.

Following (Stenz *et al.*, 2015), our approach consists of pruning gene trees to just four taxa at a time, and then combining results across all possible sets of four taxa. For a given subset of four taxa, it is easy to compare the distribution of gene trees expected from the network with the distribution of gene trees observed in the data, because there are only 3 possible unrooted gene tree topologies, or quartets, on 4 taxa. For instance, if the 4 chosen taxa are a, b, c and d, then the 3 unrooted topologies correspond to the 3 ways that we can split the 4 taxa into pairs: $ab|cd$, $ad|bd$ and $ad|bc$. The concordance factor (CF) of each quartet is the proportion of genes that evolved under that quartet topology, either inferred from data, or expected from a network under the coalescent (Baum, 2007). If the species network contains no reticulation, then it is a species tree, and calculating the quartet CFs expected under the coalescent is straightforward (e.g. Allman *et al.*, 2011). If the species network has topology $ab|cd$, then the expected CFs are

$$\mathrm{CF}_{ab|cd} = 1 - \frac{2}{3}e^{-t} \quad \text{and} \quad \mathrm{CF}_{ac|bd} = \mathrm{CF}_{ad|bc} = \frac{1}{3}e^{-t},$$

where t is the length of the internal branch in the species quartet, that is, the total length of the path that connects the two separate groups of taxa a, b and c, d in the species tree. This branch length is in coalescent units, that is, number of generations scaled by the effective population size. If the species phylogenetic network contains a reticulation, the subnetwork pruned to taxa a, b, c and d may or may not be a tree, and the calculation of the expected CFs depends on the topology of the four-taxon subnetwork. Solís-Lemus and Ané (2016) derived the formulas for expected quartet CFs in case the network is of level 1, that is, if different reticulations create cycles that do not share any edges. More formally, a phylogenetic network is of level 1 if each biconnected component contains at most one reticulation node. For instance, the quartet CFs expected from the four-taxon subnetwork in Fig. 1 are

$$\begin{aligned}
\mathrm{CF}_{ab|cd} &= 0.9\left(1 - \frac{2}{3}e^{-t_1}\right) + 0.1\frac{1}{3}e^{-t_2} \\
\mathrm{CF}_{ac|bd} &= 0.9\frac{1}{3}e^{-t_1} + 0.1\frac{1}{3}e^{-t_2} \\
\mathrm{CF}_{ad|bc} &= 0.9\frac{1}{3}e^{-t_1} + 0.1\left(1 - \frac{2}{3}e^{-t_2}\right)
\end{aligned}$$

The formulas for all possible cases in level-1 networks are found in the supplementary material of Solís-Lemus and Ané (2016), and are straightforward to apply. When the network is of level 2 or higher, the calculation of expected quartet CFs needs to employ the algorithm from Yu *et al.* (2012) or Yu *et al.* (2013a).

To quantify the fit of a candidate species network to a given genomic dataset, we compare the quartet CFs expected from the network with

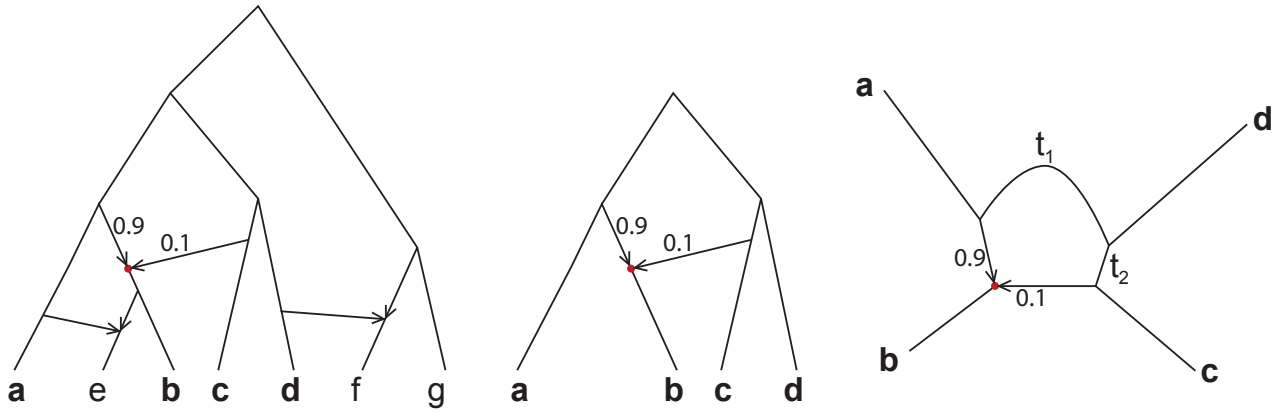

Fig. 1. Left: example of a phylogenetic network on 7 species, with 3 reticulations. For the reticulation node denoted by the red dot, the two parent hybrid edges are annotated by their inheritance probabilities, γ and $1-\gamma$. Middle: subnetwork extracted from the network on the left, on species a, b, c and d, after removing all edges that do not contribute any genetic material to a, b, c or d. Right: semi-directed version of the four-taxon subnetwork. The quartet concordance factors (proportion of genes having each quartet tree) only depend on this semi-directed topology, its branch lengths t_1 and t_2 in coalescent units, and its inheritance parameters γ (here 0.9 and 0.1).

quartet CFs obtained from the data. The simplest way to obtain these CFs is to estimate a gene tree from each multisequence alignment, and then calculate the number of genes that support each particular quartet. With this approach, one may collapse branches with low support in each gene tree. A gene will contribute information to quartets on subset a, b, c, d if this gene has sequences for all 4 of these taxa, and if its support for the quartet is sufficiently high. Therefore, the quartet CFs for different four-taxon sets may be estimated from different numbers of genes. Alternatively, a Bayesian approach can be used to estimate quartet CFs for each subset of 4 taxa, to integrate out gene tree uncertainty in a principled way (Ané *et al.*, 2006), as implemented in BUCKy (Larget *et al.*, 2010). With this approach, the estimated CF for a given quartet might not be an exact fraction of the number of genes that have sequences for the 4 taxon in this quartet. Like TICR, our test (below) accommodates this possibility. Unlike TICR, however, our tests requires knowledge of the number of genes that were used to estimate the observed quartet CFs, for each four-taxon set.

3 Methods

Our goodness-of-fit test takes as input a candidate network, and for each four-taxon set, the observed quartet CFs on this four-taxon set and the number of genes used to estimated these observed CFs. First, an outlier test is conducted on each four-taxon set to quantify the fit between observed and expected CFs. Second, the outlier test results are combined across all four-taxon sets into a single overall goodness-of-fit test, and a simulation-based strategy is employed to correct for the dependency of the test results between taxon subsets.

3.1 Goodness-of-fit on four taxa

3.1.1 Test statistics

Discrepancy between observed CFs and CFs expected from the network is due to sampling a limited number of genes, and to factors such as undetected paralogy or systematic biases during gene tree estimation. TICR models all these errors with a Dirichlet distribution, whose concentration parameter α is estimated from the data. Unfortunately, the Dirichlet distribution provides a poor fit when there are just a handful of genes that disagree (or agree) with a particular quartet (CFs close to 0 or to 1). This is because the Dirichlet distribution predicts continuous CF values, not a discrete fraction of genes having a particular quartet. In our work, we use instead the multinomial distribution for the 3 quartet resolutions given a known number of genes, and with probabilities set to the CFs expected from the candidate network. This multinomial distribution provides a much better fit at CFs close to 0 or 1 (see Fig. S1), and does not have any parameter to be estimated.

When the species phylogeny is a tree, the "major" quartet can be defined as the quartet present in the species tree. TICR focuses on this major quartet to quantify how the expected CFs fit the observed CFs. A species network may display several of the quartets, however, for a given four-taxon set. For example, the species network in Fig. 1 displays both $ab|cd$ and $ad|bc$. To generalize TICR to species networks, we modified the outlier test on four taxa to use all three quartet CFs symmetrically. We propose three outlier tests, depending on the choice of the test statistic. The Pearson's statistic X^2 is the most common choice for a goodness-of-fit test with the multinomial distribution:

$$X^2 = n \sum_{i=1}^{3} \frac{(o_i - e_i)^2}{e_i}\,, \tag{1}$$

where the sum is taken over the three quartets; e_i is the expected CF for quartet i from the network; and o_i is the observed CF for quartet i. Since CFs are defined as proportions, our formula has a factor n, the number of genes available for the given four-taxon set. The Pearson X^2 is known to be unreliable when one of the expected counts (e_i) is close to 0. This situation is expected to be very frequent, whenever a four-taxon relationship is supported across most genes with no discordance. Therefore, we propose alternative test statistics, with broader reliability: the $Q_{\log}$ statistic (Lorenzen, 1995)

$$Q_{\log} = 2n \sum_{i} \frac{(o_i - e_i)^2}{e_i(\log o_i - \log e_i)} \tag{2}$$

and the likelihood ratio test statistic G:

$$G = 2n \sum_{i=1}^{3} o_i * \log\left(\frac{o_i}{e_i}\right)\,. \tag{3}$$

All three test statistics measure the discrepancy between the expected CFs e_i and the observed CFs o_i. They take large values when there is large discrepancy, and $X^2 = Q_{\log} = G = 0$ when $e_i = o_i$, that is, when the four-taxon network fits the data perfectly. Under the null hypothesis that the four-taxon network and the multi-species coalescent model are correct, X^2, $Q_{\log}$ and G follow a chi-squared distribution χ^2_2 with 2 degrees of

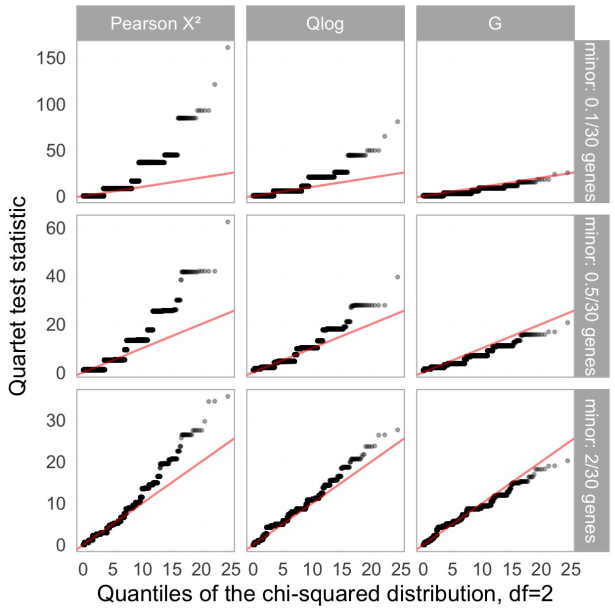

Fig. 2. Distribution of X^2, $Q_{\log}$ and G for measuring the fit of quartet CFs, when the data consist of 30 gene trees drawn from a multinomial distribution on 3 quartets with given expected CFs. The expected number of genes supporting each quartet was 0.1, 0.1, 29.8 (top), 0.5, 0.5, 29 (middle) and 2, 2, 26 (bottom). These values would arise from a tree-like four-taxon phylogeny with an internal branch length of 4.6 (top), 3.0 (middle) and 1.6 (bottom) coalescent units. 100,000 data sets were simulated in each case. The sorted values of the 100,000 test statistics are plotted on the vertical axis, versus the theoretical quantiles expected under a χ^2_2 distribution on the horizontal axis. The test statistic is χ^2_2-distributed if the points fall on the diagonal line (in red). When the test statistic is above the diagonal, the p-value obtained by comparison with the χ^2_2 distribution is too small.

freedom approximately. Therefore, we test the goodness of fit, for a given set of 4 taxa, by comparing the chosen test statistic to the χ^2_2 distribution. We call the resulting p-value an *outlier* quartet p-value, to avoid confusion with the p-value in the next section, for the fit of the overall network.

3.1.2 Comparison of quartet test statistics

We studied the behavior of X^2, $Q_{\log}$ and G using simulations, to know which one to recommended for our phylogenetic problem, in which we expect many four-taxon relationships to be well resolved. For instance, two taxa could be from one clade and the other two taxa from another well-separated clade. One might easily expect very few genes to have a quartet that disagrees with these clades. Yet, the approximation of X^2 by the χ^2_2 distribution deteriorates if any of the 3 quartets is expected to be supported by 5 or fewer genes.

We simulated observed CFs for four-taxon sets that challenge the reliability of the χ^2_2 approximation: with 30 genes and expected CFs such that two quartets are expected to be supported by 0.1, 0.5, or 2 genes only. Fig. 2 shows the quality (or lack thereof) of the χ^2_2 approximation, for each choice of test statistic. Results are similar (Fig. S2) when only one quartet is supported by very few genes, the other two having equally high CFs (which requires reticulation). As expected, the Pearson statistic X^2 performs poorly, leading to p-values that tend to be much smaller than they should be. $Q_{\log}$ performs much better, although its p-values also tend to be too small. The likelihood ratio statistics G performs best, with p-values behaving adequately. When all 3 quartets are supported by 2 or more genes, the χ^2_2 distribution approximation performs adequately for both the $Q_{\log}$ and G statistics. Consequently, we recommend the use of G for future studies.

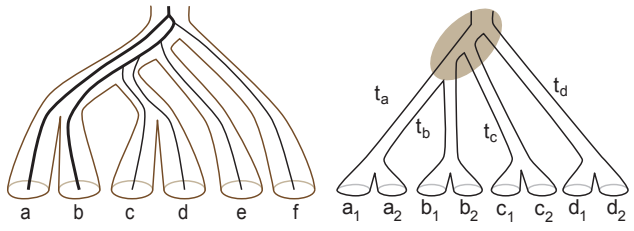

Fig. 3. Edge overlap between four-taxon sets causes dependence between their quartet CFs. Left: the four-taxon sets a, b, c, d and a, b, e, f share edges ancestral to a, b. If a and b coalesce deeply in a given gene tree (thick black lines), then this gene is less likely to have quartets where a and b are sister, for both a, b, c, d or a, b, e, f. Right: the four-taxon sets a_1, b_1, c_1, d_1 and a_2, b_2, c_2, d_2 have no taxon overlap, but share many edges. If the alleles from each clade coalesce early for a given gene (that is, a_1 and a_2 coalesce along the branch of length t_a, etc.) then this gene has the same topology when restricted to a_1, b_1, c_1, d_1 or to a_2, b_2, c_2, d_2. This shared topology depends on coalescences along ancestral branches (highlighted area). Therefore, the dependence between these two four-taxon sets' CFs increases with branch lengths t_a, t_b, t_c and t_d. More generally, a higher overall degree of dependence is expected with increased branch lengths.

3.2 Goodness-of-fit of a candidate network

If the candidate network provides a good fit, one would expect outlier quartet p-values of four-taxon sets to have a uniform distribution, with 5% of four-taxon sets having outlier p-values below 0.05, for instance. TICR compares the proportion of outlier p-values in the intervals 0-0.01, 0.01-0.05, 0.05-0.1, 0.1-1, to the expected proportions of 0.01, 0.04, 0.05, 0.90. Here, we use a one-sided test instead, to detect if there are *more* outlier four-taxon sets than expected. Each four-taxon set is classified as an "outlier" if its outlier quartet p-value is below 0.05. If the network provides a good fit, we expect 5% of four-taxon sets to be outliers. If there is less than 5% outliers, then the network can be said to provide a good fit to the data. If there is more than 5% outliers, then a rigorous test is needed. A chi-square goodness-of-fit test for proportions is used by TICR, but four-taxon sets violate a fundamental assumption of this test: they are not all independent.

For example, consider the four-taxon sets a, b, c, d and a, b, e, f (Fig. 3). The quartets obtained by pruning a gene tree to a, b, c, d and a, b, e, f are not independent, because a, b and their adjacent edges are retained in both. So the outlier p-values of these two four-taxon sets could be correlated. Consider now the four-taxon sets a_1, b_1, c_1, d_1 and a_2, b_2, c_2, d_2, that represent one taxon from each of 4 well-defined clades, such that a_1 and a_2 are close sisters, etc. (Fig. 3). Rapid coalescence between sister taxa (a_1 with a_2, b_1 with b_2 etc.) would cause the quartet tree on a_1, b_1, c_1, d_1 to be equal to the quartet tree on a_2, b_2, c_2, d_2, for a given gene. Therefore, non-overlapping four-taxon sets can also have correlated outlier p-values. The main reason for dependence is gene sharing (not taxon sharing), in the data used to calculate outlier p-values: the same gene trees are used across different four-taxon sets.

To determine if the proportion of outlier four-taxon sets deviates from 5% significantly, we propose a simulation-based strategy to account for dependence. We first consider the traditional z statistic for a proportion test:

$$Z = \frac{\hat{p}_{\text{out}} - p_{\text{out}}}{\sqrt{p_{\text{out}}(1 - p_{\text{out}})/N}} = \frac{\hat{p}_{\text{out}} - 0.05}{\sqrt{0.0475/N}} \qquad (4)$$

where $p_{\text{out}} = 0.05$ is the expected proportion of outliers, $\hat{p}_{\text{out}}$ is the observed proportion of outliers, and N is the total number of four-taxon sets. If four-taxon sets were independent, then $Z \sim \mathcal{N}(0, 1)$ approximately.

To account for dependence, we simulate a large number B of data sets under the coalescent on the candidate network, each with the same number of gene trees as in the original data. From each simulation i we calculate

Z_i using (4). These values are then used to estimate $\sigma^2 = \mathbb{E}Z^2$ as

$$\hat{\sigma}^2 = \frac{1}{B}\sum_{i=1}^{B} Z_i^2 \,.$$

Finally, we quantify the overall goodness-of-fit of the network by comparing Z from the original data to the normal distribution $\mathcal{N}(0, \hat{\sigma}^2)$ with variance $\hat{\sigma}^2$.

This test assumes that the right tail of the distribution of Z remains that of a normal distribution approximately, even when four-taxon sets are dependent. We assess the reliability of this assumption next. In any case, the simulations used to estimate σ^2 would naturally reveal if this assumption is not satisfied for the network at hand. In this case, an empirical p-value can be estimated using the simulated Z_i values directly, as the proportion of Z_i values greater than or equal to the original Z. This alternative estimation of the p-value requires a larger number of simulations B, because it amounts to estimating the tail, instead of the variance, of the distribution of Z.

4 Results

4.1 Simulations

We simulated the distribution of Z on three different networks, from low dependence to high dependence across four-taxon sets. As our baseline network ("net1", moderate dependence), we used the network phylogeny of New World kingsnakes *Lampropeltis*. This network has 23 taxa, one hybridization, was inferred using 304 loci by Burbrink and Gehara (2018), and was rooted at the outgroup *Cemophora coccinea*. We modified the branch lengths around the reticulation to make the network time-consistent, as required by `HybridLambda` for simulation: such that all paths from the root to any given hybrid node are of equal length. External branch lengths were set to make all paths from the root to the tips of equal length. Internal branch lengths, which determine the level of ILS, had an average of 0.66 coalescent units. For a network with minimal dependence, we used the star topology on 23 taxa, that is, a network whose internal branch lengths are all 0 to represent ancestral panmixia. This topology is treated more theoretically below. For a network with high dependence, we used our baseline network and multiplied all its branch lengths by a factor or 3 ("net3"). The maximum of the three quartet CFs, averaged across all four-taxon sets, was one third under the star network, 77% under net1 and 91% under net3. `HybridLambda` (Zhu *et al.*, 2015) was used to simulate 304 gene trees under the coalescent along the network. We then calculated the observed quartet CFs from the simulated gene trees, the outlier quartet p-values (using either X^2, $Q_{\log}$ or G), and the Z statistics in (4) using `QuartetNetworkGoodnessFit`, which builds on `PhyloNetworks` (Solís-Lemus *et al.*, 2017). These simulations were repeated 1000 times on each network.

Fig. 4 shows the results using the G statistic. Results are similar when using X^2 or $Q_{\log}$ instead of G for calculating outlier quartet p-values (Fig. S3 and S4). The distributions of outlier quartet p-values are relatively uniform with almost 5% of four-taxon sets labelled as outliers, except on the network with long branches (net3). This is expected, because long branches in coalescent units means that deep coalescence is very rare, and genes are expected to be highly concordant with each other (except where reticulation is involved). In this case, many four-taxon sets will have one or two quartets with almost no genes supporting them, a situation that degrades the χ^2_2 approximation, as seen in section 3.1.2. Still, the proportion of outliers is conservative and remains close to 5%: 0.036.

Failing to account for dependence using a traditional proportion test that compares Z to a standard normal distribution ($\sigma^2 = 1$), like TICR does, leads to greatly inflated rate of type 1 errors: from 20% to 36% depending on the network (Fig. 4, second column). P-values from this uncorrected test cluster near zero and one. To correct for dependence, we approximated the distribution of Z using $\mathcal{N}(0, \hat{\sigma}^2)$. As expected, the star topology showed to the least amount of dependence, as quantified by the smallest $\hat{\sigma}^2 = 19.1$. The baseline network (net1) had a larger $\hat{\sigma}^2 = 82.7$, that is, required more correction for dependence. The network with increased branch lengths (net3) had the largest $\hat{\sigma}^2 = 150.1$, indicating the highest level of dependence across four-taxon sets. Regardless of the network, these $\hat{\sigma}^2$ values are very far from 1, confirming that the dependence between four-taxon sets is a major factor. On all three networks, Z shows some degree of right skew, due in part to its lower bound at $-0.05/\sqrt{.0475/N}$ when the proportion of outliers is zero. However, the right tail of Z is fairly well approximated by the corrected normal distribution, which is what matters for our one-sided test. After correction, the p-value for testing the overall goodness-of-fit has an acceptable rate of type I error close to the desired 5% level, or conservatively below 5% for the network with long branches (Fig. 4, right). Therefore, our test accounts for dependence appropriately.

4.2 Star phylogeny

In this section, we assume a start phylogeny, that is, a phylogeny where all internal branches have length 0 in coalescent units. In this case, we derive the value of σ^2 exactly, asymptotically when the number of genes is large:

$$\sigma^2 = 1 + \frac{n_{\text{tax}} - 4}{0.0475}\big(3(n_{\text{tax}} - 5)v_2 + 4v_3\big) \qquad (5)$$

where n_{tax} is the number of tips, $v_2 \simeq 0.000373186$ and $v_3 \simeq 0.002275837$.

We conjecture that the star phylogeny has the smallest σ^2 among all phylogenies with the same number of taxa, such that (5) provides a lower bound for any phylogeny, which is very fast to calculate. Even if this conjecture is true, this lower bound can be far from the true σ^2, as exemplified by the range of values from the simulations above. As seen in the proof below (and in the Supplementary Material), this bound assumes no bias in Z (that is, the outlier test statistic has type I error rate of 0.05 exactly and $\mathbb{E}Z = 0$), and only quantifies dependence due to taxon-sharing between four-taxon sets. It does not account for dependence due to shared coalescent events on internal branches (Fig. 3), since there are no such events on a star phylogeny.

In the rest of this section, we sketch the proof of (5), and refer to the Supplementary Material for details. We can rewrite (4) as

$$Z = \frac{1}{\sqrt{N}}\sum_{q=1}^{N}\frac{Y_q - 0.05}{\sqrt{0.0475}}$$

where $Y_q = 1$ if the four-taxon set q is detected as an outlier, $Y_q = 0$ otherwise. Y_q was designed to satisfy $\mathbb{P}\{Y_q = 1\} = 0.05$ under the null hypothesis, so it has mean 0.05 and variance 0.0475 asymptotically (with a large number of genes). In the Supplementary Material, we prove that the covariance of Y_q and $Y_{q'}$ for distinct q and q' is 0 if q and q' share one taxon at most; v_2 if they share two taxa, and v_3 if they share three taxa. Given these pairwise covariances, (5) follows from counting the number of pairs of four-taxon sets that share exactly 2 taxa: $3(n_{\text{tax}}-4)(n_{\text{tax}}-5)N$; and the number of pairs that share exactly 3 taxa: $4(n_{\text{tax}} - 4)N$.

4.3 Case studies

We re-analyzed the *Arabidopsis thaliana* data from Stenz *et al.* (2015) on 23 taxa and over 3000 genes (Table 1). The outlier p-values from our new test are much smaller than the TICR p-values. All population histories considered by Stenz *et al.* (2015) are found inadequate to describe the pattern in the data, with much stronger evidence than in the original study.

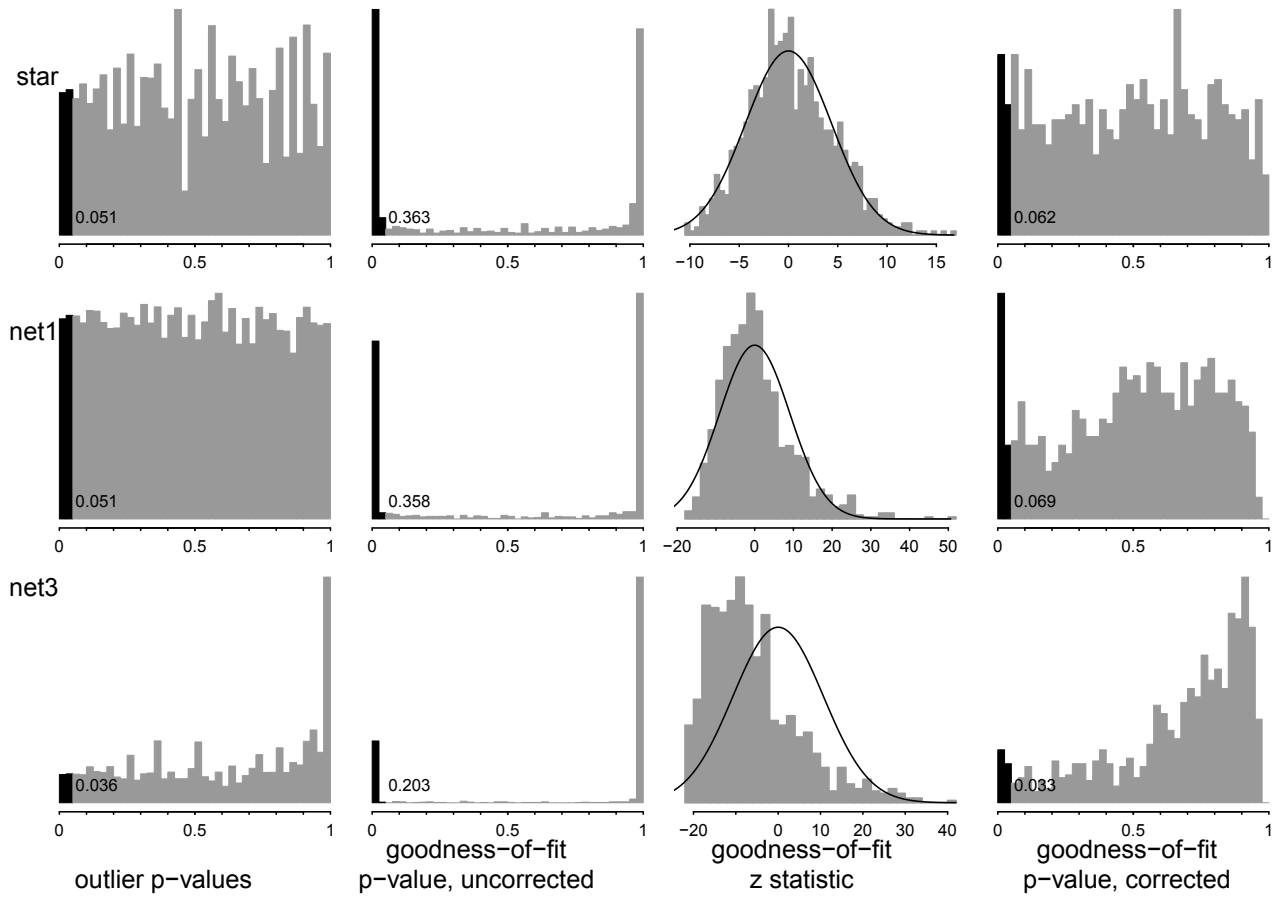

Fig. 4. Simulations of the goodness-of-fit test using the likelihood ratio statistic G for the outlier test on each four-taxon set, on three different 23-taxon networks (top: star phylogeny; middle: baseline network from Burbrink and Gehara (2018); bottom: baseline network where all branch lengths have been multiplied by 3). For each network, 1000 data sets were simulated from that network, so the null hypothesis of adequacy was true. Left column: distribution of outlier p-values, across all simulation replicates and all four-taxon sets. About 5% of four-taxon set appear as outliers with a p-value below 5% (shown as black histogram bars, with actual proportion value indicated in black). Second column: distribution of the p-value obtained by comparing z to $\mathcal{N}(0, 1)$, without any correction for dependence. The black histogram bars represent the simulation replicates with a p-value below 5%, with the proportions indicated in black (all above 20%). Third column: distribution of z from (4) (grey histogram) and density of the centered normal $\mathcal{N}(0, \hat{\sigma}^2)$ used to approximate the right tail of z (black curve). Right column: distribution of the goodness-of-fit p-value with correction for dependence. The black histogram bars represent the simulation replicates with a p-value below 5%, with the proportions indicated in black (all around 5%).

Their qualitative conclusions remain: full panmixia is strongly rejected, and a partially resolved population tree with episodes of panmixia provides a better fit to the data than either panmixia or a fully resolved tree. However, we find that this fit is still inadequate when the test properly accounts for dependence.

The outlier taxa detected by Stenz *et al.* (2015) are still detected as outliers in our new test. In the stratified subsample, we ranked outlier p-values under the tree with optimized branch lengths and retained the 519 four-taxon sets with an outlier p-value below 10^{-50}. This list is of comparable size as the list of 483 outlier four-taxon sets found to have a TICR p-value below 0.01, of which 212 contained two accessions from the United Kingdom hypothesized to have undergone recent gene flow, Cnt_1 and Vind_1 (Stenz *et al.*, 2015). In our 519 most outlier four-taxon sets from our new test, 322 contain both Cnt_1 and Vind_1, confirming the original conclusion by Stenz *et al.* (2015).

Next, we analyzed data from Karimi *et al.* (2020) on baobabs (*Adansonia*), from 372 loci across 17 accessions (8 ingroup and 2 outgroup species). Karimi *et al.* (2020) found evidence for gene flow, but uncertainty about the placement and number of gene flow events, so we tested the two network topologies estimated from the 372 loci: one estimated using all taxa, and one estimated with the outgroups excluded. Both networks were found to provide an inadequate fit to the data (Table 2). When considering the most outlying four-taxon sets (with outlier p-values below 0.01), we

Data set	Topology	z	$\hat{\sigma}$	$z/\hat{\sigma}$
PD subsample	panmixia	641.66	6.24	103.8
	full tree	557.54	6.28	88.7
	BPP partial tree	636.81	6.25	101.9
	TICR partial tree	556.21	6.49	85.7
Stratified subsample	panmixia	676.36	6.13	110.4
	full tree	584.51	7.04	83.0
	BPP partial tree	602.05	6.69	90.0
	TICR partial tree	580.40	6.90	84.2

Table 1. Analysis of the *Arabidopsis thaliana* data sets from Stenz et al. (2015), each with 30 taxa, using the G statistics. The test statistic $z/\hat{\sigma}$ corrects for dependence, with $\hat{\sigma}$ obtained using simulations of 1000 data sets, with the same number of gene trees as in the original data (PD: 3064; stratified: 3191). All trees were inadequate with overwhelming evidence (p-values $\ll 10^{-100}$).

Outgroups	Network	z	$\hat{\sigma}$	$z/\hat{\sigma}$	p-value
included	3(a)	25.8	6.32	4.07	0.00002
excluded	3(a)	22.2	4.92	4.51	0.000003
excluded	3(c)	20.3	5.17	3.93	0.00004

Table 2. Analysis of the 372-locus baobab data from Karimi et al. (2020) using the G statistic. The network in their Figure 3(a) was estimated using outgroups, with gene flow ($\gamma = 12\%$) from *A. digitata* into the stem of *Brevitubae*. Network 3(c) was estimated using ingroup taxa only, with gene flow ($\gamma = 17\%$) from *A. rubrostipa* into the stem of the core Malagasy Longitubae clade. Branch lengths were optimized for each network. Simulations used 10,000 replicates to estimate $\hat{\sigma}$. All networks fit poorly.

found an over-representation (81% or more) of sets containing either accession from *A. za* or *A. perrieri*. In fact, a majority of the most outlying four-taxon sets had either both accessions from *A. za*, or both accessions from *A. perrieri*. Note that in both candidate networks, *A. za* were not monophyletic, and *A. perrieri* formed a clade sister to one of the *A. za* accession. When both *A. za* accessions or both *A. perrieri* accessions were excluded, the network adequacy increased substantially, with p-values of 0.01 (*A. za* excluded) or above 0.04 (*A. perrieri* excluded). When *A. za* and *A. perrieri* accessions were both excluded, the networks fitted adequately, with p-values above 0.05. This pattern suggests that both candidate networks are missing possible gene flow involving *A. za* and *A. perrieri*.

Finally, we analyzed data on *Stachyurus*, a genus of shrubs and small trees in which species delimitation is difficult. We analyzed data on 17 taxa and 1362 genes from Feng *et al.* (2020), who reported evidence for reticulation in this genus. Their networks did not fit the data adequately, even with up to 3 reticulations. Using outlier quartets, we identified a subset of taxa for which the subnetwork with two reticulations fits adequately, whereas one or no reticulations did not fit adequately (see Supplementary Material).

4.4 Implementation

Our new tests and TICR are implemented in `Julia` (Bezanson *et al.*, 2012), in package `QuartetNetworkGoodnessFit`, available at `github.com/cecileane/QuartetNetworkGoodnessFit.jl`. Since it depends on `PhyloNetworks` (Solís-Lemus *et al.*, 2017), the current implementation assumes that the species network is of level 1. By default, the test uses the G statistic to obtain outlier quartet p-values, and 1000 replicates for simulations. These options can be adjusted by the user, as demonstrated in the package documentation. The scripts for our analyses are available at `https://osf.io/eg6ju/`.

5 Discussion

We developed a method to decide if the multi-species coalescent model on a given (possibly reticulate) phylogeny adequately describes the concordance pattern in a genome-wide data set. The strategy extends that from TICR (Stenz *et al.*, 2015), with proper handling of dependence among four-taxon sets and extension to phylogenetic networks. Like TICR, this method provides information beyond network fit by detecting outlier species that drive the lack of fit. Species over-represented in outlier four-taxon sets could be misplaced in the network, or might be of undetected reticulate origin. Even if most of the phylogeny fits the data well, the outlier tests could help identify local areas that fit the concordance pattern poorly.

A limitation of our method is its reliance on simulations to quantify the dependence across four-taxon sets. Future theoretical work could extend our results on the star phylogeny to general network topologies, to avoid reliance on simulations.

Another limitation is the use of quartet CFs as input data, as estimated from multi-locus data. Estimating CFs from concatenated alignments, such as concatenated bi-allelic or SNP data, has only recently been attempted (Olave and Meyer, 2020). Other measures of goodness-of-fit could be based on SNP patterns instead of CF patterns, but there are many more possible site patterns on 4 taxa for DNA ($4^4 = 256$) or biallelic markers ($2^4 = 16$), than possible quartets (3), making the use of site patterns more complicated. Using site patterns could bring more power, but could also be more prone to inconsistencies due to long branch attraction or rate variation across sites and/or lineages, something that can be mitigated by using best-fitting substitution models when estimating gene trees and CFs. Ruffley *et al.* (2018) used SNP data from multiple alleles in two populations. They used site patterns as summarized by the joint allele frequency spectrum, then quantified the goodness-of-fit of various migration models with a global likelihood ratio statistic and a simulation approach. The limitation of this SNP-based approach is that it requires few populations (or taxa), multiple individuals per population and within-taxon allele variation that is shared between taxa.

Relying on quartets also means that the computing time increases polynomially with the number of taxa, as $N = O(n_{\text{tax}}^4)$ if all four-taxon sets are used. Extensions should consider reducing the number of input four-taxon sets while minimizing the loss of power to detect a lack of fit, such as using efficient quartet representations of input gene trees (Davidson *et al.*, 2018).

Performing tests on subsets of 3 or 4 taxa is a common strategy on large genome-wide data sets. Most methods using quartets or rooted triplets focus on detecting introgression, such as the ABBA-BABA test (Green *et al.*, 2010; Durand *et al.*, 2011), HyDe (Blischak *et al.*, 2018), Quartet Sampling (Pease *et al.*, 2018) or D_3 (Hahn and Hibbins, 2019). Our method is similar as it performs a very large number of tests, one on each four-taxon subset. It is unique in that it combines all these test results into an overall test, accounting for dependence across them. Some quartet / triplet methods aim to estimate an overall phylogenetic network (Solís-Lemus and Ané, 2016; Yu and Nakhleh, 2015; Allman *et al.*, 2019; Zhu *et al.*, 2019), although without assessing the adequacy of the estimated network. The NANUQ method (Allman *et al.*, 2019) is closest to our work, as it involves a test based on the quartet CFs on each four-taxon set. They also use the G statistic, but perform different tests then we do: one test to decide if a star tree is an adequate explanation for the CF data (to avoid using four-taxon sets with too little information), and one test to decide if there is evidence for a four-taxon reticulate network over a four-taxon species tree. In our work, our null hypothesis corresponds to the candidate network, because our method aims to measure the goodness-of-fit instead of inferring a network.

Our test offers the first rigorous method to measure and compare the fit of various candidate phylogenetic networks with different number of reticulations, when the data include more than a handful of taxa. The test could be used to select the appropriate number of past hybridizations, to avoid over-parametrization with unnecessary reticulations. Yu *et al.* (2014) used cross-validation, an adequate way to select the number of reticulations. However, it is intensive (10-fold cross-validation requires about 10 times the computing time of a single network). More importantly, it does not scale to many taxa because the quality of a given network is calculated by comparing the frequency of every gene tree topology in the validation subset, and its probability under the estimated network. With more taxa, the number of possible gene topologies grows exponentially, rendering this quality measure difficult to calculate, and possibly unstable.

Some empirical studies have used BIC, AIC or AICc to select the optimal number of reticulations (e.g. Mason *et al.*, 2019; Kleinkopf *et al.*,

2019; Zhang *et al.*, 2019). However, AIC and BIC approaches have two major drawbacks. First, they should not be used with pseudo-likelihood (also called composite likelihood, Varin *et al.*, 2011; Excoffier *et al.*, 2013). Pseudo-likelihood is valid for parameter estimation, but not valid for hypothesis testing or model selection, unless special care is taken (e.g. Godambe information, Varin *et al.*, 2011). Second, AIC and BIC are not meant to deal with a number of models that grows exponentially with the number of reticulations (Blair and Ané, 2020). A similar problem is encountered for phylogenetic adaptive niche models, based on Ornstein-Uhlenbeck processes with variable niches represented by variable drift parameters. AIC and BIC are inappropriate criteria to select the number k of adaptive shifts, because the number of ways to place k shifts on a phylogeny grows very fast with k (Ho and Ané, 2014; Khabbazian *et al.*, 2016; Bastide *et al.*, 2018). Similarly, the number of ways to place k reticulations on a given phylogenetic network grows very fast with k, and this growth in model space is not accounted for by AIC or BIC.

Many studies used a slope (or broken stick) heuristic approach, looking at the number of reticulations beyond which the network score has little improvement (e.g. Burbrink and Gehara, 2018; Hejase *et al.*, 2018). This slope heuristic has proven guarantees in some situations (Baudry *et al.*, 2012, implemented in R package capusche), but its rigorous application still requires to estimate a network for a large range of reticulation numbers, which might be computationally prohibitive.

Alternative rigorous methods for the selection of network complexity include machine learning approaches, after obtaining a short list of candidate networks. Data sets are simulated under each network in the list, and used to train a model that predicts from which network (in the list) the data came from, based on summary statistics. For example, Burbrink and Gehara (2018) used a neural network approach to formally compare one tree-like and two reticulate phylogenies, after estimation with a pseudo-likelihood method SNaQ (Solís-Lemus and Ané, 2016). Approximate Bayesian computation (ABC) methods can then be used to estimate posterior probabilities for each network in the list (Nater *et al.*, 2015; Pudlo *et al.*, 2016; Smith *et al.*, 2018). The downside of these methods is that a suitable set of summary statistics needs to be chosen.

Our method provides a rigorous tool that can be used after network estimation, to measure the adequacy of networks with various number of reticulations and select the optimal network complexity. It scales to much larger phylogenies than other rigorous alternatives, and goes beyond model comparison by measuring the absolute fit of the coalescent on a candidate network.

Acknowledgements

We are very grateful to David Baum and Bret Larget for insightful feedback on this work, to Claudia Solís-Lemus for discussions about outlier tests for quartets, and to Marcelo Gehara for sharing kingsnakes phylogenetic networks with various numbers of reticulations. We warmly thank Yu Feng for sharing data on *Stachyurus*.

Funding

This work was supported in part by a University of Wisconsin-Madison book store academic excellence award (to RC), a H. I. Romnes faculty fellowship (to CA) provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation, and by National Science Foundation grant DMS 1902892.

References

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, **62**(6), 833–862.

Allman, E. S., Baños, H., and Rhodes, J. A. (2019). NANUQ: a method for inferring species networks from gene trees under the coalescent model. *Algorithms for Molecular Biology*, **14**(1), 24.

Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. (2006). Bayesian estimation of concordance among gene trees. *Molecular biology and evolution*, **24**(2), 412–426.

Bastide, P., Ané, C., Robin, S., and Mariadassou, M. (2018). Inference of adaptive shifts for multivariate correlated traits. *Systematic Biology*, **67**(4), 662–680.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, **22**(2), 455–470.

Baum, D. A. (2007). Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*, **56**(May), 417–426.

Bezanson, J., Karpinski, S., Shah, V. B., and Edelman, A. (2012). Julia: A Fast Dynamic Language for Technical Computing. *arXiv:1209.5145*, pages 1–27.

Blair, C. and Ané, C. (2020). Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. *Systematic Biology*, **69**(3), 593–601.

Blischak, P., Chifman, J., Wolfe, A. D., and Kubatko, L. S. (2018). HyDe: a Python package for genome-scale hybridization detection. *Systematic Biology*, **67**, 821–829.

Brown, J. M. and Thomson, R. C. (2018). Evaluating model performance in evolutionary biology. *Annual Review of Ecology, Evolution, and Systematics*, **49**(1), 95–114.

Burbrink, F. T. and Gehara, M. (2018). The biogeography of deep time phylogenetic reticulation. *Systematic biology*, **67**(5), 743–755.

Davidson, R., Lawhorn, M., Rusinko, J., and Weber, N. (2018). Efficient quartet representations of trees and applications to supertree and summary methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **15**(3), 1010–1015.

Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Systematic Biology*, **67**(5), 786–799.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**(8), 2239–2252.

Elworth, R. A. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. (2019). Advances in computational methods for phylogenetic networks in the presence of hybridization. In T. Warnow, editor, *Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret*, pages 317–360. Springer International Publishing, Cham.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and snp data. *PLOS Genetics*, **9**(10), 1–17.

Feng, Y., Comes, H. P., and Qiu, Y.-X. (2020). Phylogenomic insights into the temporal-spatial divergence history, evolution of leaf habit and hybridization in *Stachyurus* (Stachyuraceae). *Molecular Phylogenetics and Evolution*, **150**, 106878.

Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. (2018). New prospects in the detection and comparative analysis of hybridization in the tree of life. *American Journal of Botany*, **105**(3), 364–375.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, **328**(5979), 710–722.

Hahn, M. W. and Hibbins, M. S. (2019). A three-sample test for introgression. *Molecular Biology and Evolution*, **36**(12), 2878–2882.

Hejase, H. A., VandePol, N., Bonito, G. M., and J., L. K. (2018). FastNet: Fast and accurate statistical inference of phylogenetic networks using large-scale genomic sequence data. In M. Blanchette and A. Ouangraoua, editors, *Comparative Genomics. RECOMB-CG 2018*, volume 11183 of *Lecture Notes in Computer Science*, pages 242–259. Springer, Cham.

Ho, L. and Ané, C. (2014). Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*, **5**(11), 1133–1146.

Karimi, N., Grover, C. E., Gallagher, J. P., Wendel, J. F., Ané, C., and Baum, D. A. (2020). Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (*Adansonia*; Bombacoideae; Malvaceae). *Systematic Biology*, **69**(3), 462–478.

Khabbazian, M., Kriebel, R., Rohe, K., and Ané, C. (2016). Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*, **7**(7), 811–824.

Kleinkopf, J. A., Roberts, W. R., Wagner, W. L., and Roalson, E. H. (2019). Diversification of Hawaiian *Cyrtandra* (Gesneriaceae) under the influence of incomplete lineage sorting and hybridization. *Journal of Systematics and Evolution*, **57**(6), 561–578.

Kubatko, L. S. (2009). Identifying Hybridization Events in the Presence of Coalescence via Model Selection. *Systematic Biology*, **58**(5), 478–488.

Larget, B., Kotha, S., Dewey, C., and Ané, C. (2010). BUCKy: Gene tree / species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, **26**(22), 2910–2911.

Lorenzen, G. (1995). A new family of goodness-of-fit statistics for discrete multivariate data. *Statistics & probability letters*, **25**(4), 301–307.

Markin, A., Anderson, T. K., Vadali, V. S., and Eulenstein, O. (2019). Robinson-foulds reticulation networks. *bioRxiv*.

Mason, A. J., Grazziotin, F. G., Zaher, H., Lemmon, A. R., Moriarty Lemmon, E., and Parkinson, C. L. (2019). Reticulate evolution in nuclear middle america causes discordance in the phylogeny of palm-pitvipers (viperidae: Bothriechis). *Journal of Biogeography*, **46**(5), 833–844.

Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, **75**(1), 35–45.

Nater, A., Burri, R., Kawakami, T., Smeds, L., and Ellegren, H. (2015). Resolving evolutionary relationships in closely related species with whole-genome sequencing data. *Systematic Biology*, **64**(6), 1000–1017.

Olave, M. and Meyer, A. (2020). Implementing large genomic single nucleotide polymorphism data sets in phylogenetic network reconstructions: A case study of particularly rapid radiations of cichlid fish. *Systematic Biology*. Advance Access.

Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E., and Smith, S. A. (2018). Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany*, **105**(3), 385–403.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, **32**(6), 859–866.

Ruffley, M., Smith, M. L., Espíndola, A., Carstens, B., Sullivan, J., and Tank, D. C. (2018). Combining allele frequency and tree-based approaches improves phylogeographic inference from natural history collections. *Molecular Ecology*, **27**(4), 1012–1024.

Semple, C. and Steel, M. (2003). *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press.

Smith, M. L., Ruffley, M., Rankin, A. M., Espíndola, A., Tank, D. C., Sullivan, J., and Carstens, B. C. (2018). Testing for the presence of cryptic diversity in tail-dropper slugs (*Prophysaon*) using molecular data. *Biological Journal of the Linnean Society*, **124**(3), 518–532.

Solís-Lemus, C. and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, **12**(3), e1005896.

Solís-Lemus, C., Bastide, P., and Ané, C. (2017). Phylonetworks: a package for phylogenetic networks. *Molecular Biology and Evolution*, **34**(12), 3292–3298.

Stenz, N., Larget, B., Baum, D. A., and Ane, C. (2015). Exploring tree-like and non-tree-like patterns using genome sequences: and example using the inbreeding plant species Arabidopsis thaliana (L.) Heynh. *Systematic Biology*, **64**(5), 809–823.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**(1), 5–42.

Wen, D. and Nakhleh, L. (2017). Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, **67**(3), 439–457.

Wen, D., Yu, Y., and Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet*, **12**(5), 1–17.

Wu, Y. (2013). An algorithm for constructing parsimonious hybridization networks with multiple phylogenetic trees. *Journal of Computational Biology*, **20**(10), 792–804.

Yu, Y. and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, **16(Suppl 10)**.

Yu, Y., Degnan, J. H., and Nakhleh, L. (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS genetics*, **8**(4), e1002660.

Yu, Y., Ristic, N., and Nakhleh, L. (2013a). Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics*, **14**(15).

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

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum Likelihood Inference of Reticulate Evolutionary Histories. *PNAS*, **111**(46), 16448–16453.

Zhang, B.-W., Xu, L.-L., Li, N., Yan, P.-C., Jiang, X.-H., Woeste, K. E., Lin, K., Renner, S. S., Zhang, D.-Y., and Bai, W.-N. (2019). Phylogenomics reveals an ancient hybrid origin of the Persian walnut. *Molecular Biology and Evolution*, **36**(11), 2451–2461.

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, **35**(2), 504–517.

Zhu, J. and Nakhleh, L. (2018). Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. *Bioinformatics*, **34**(13), i376–i385.

Zhu, J., Liu, X., Ogilvie, H. A., and Nakhleh, L. K. (2019). A divide-and-conquer method for scalable phylogenetic network inference from multilocus data. *Bioinformatics*, **35**(14), i370–i378.

Zhu, S., Degnan, J. H., Goldstien, S. J., and Eldon, B. (2015). Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. *BMC bioinformatics*, **16**(1), 292.

Supplementary Material for: Assessing the fit of the multi-species network coalescent to multi-locus data

Ruoyi Cai and Cécile Ané

1 Supplementary Figures

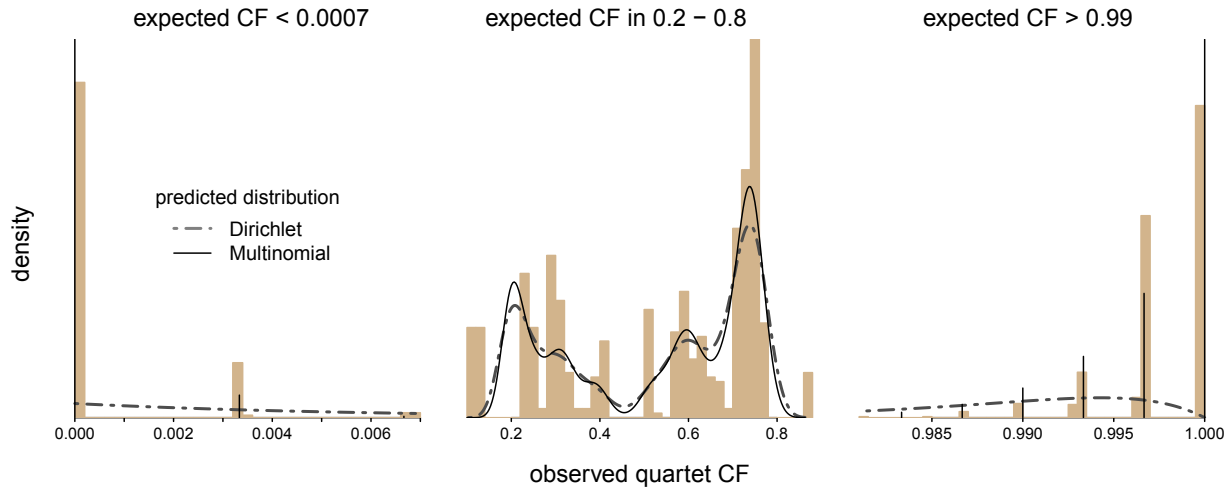

Figure S1: Distribution of quartet CFs from 300 loci simulated by Solís-Lemus and Ané (2016) under a network with 15 species and three hybridization events, from `https://github.com/crsl4/PhyloNetworks.jl/wiki/Example-Data`. Observed CFs are posterior means from a Bayesian inference with BUCKy (histogram). These CFs were used to fit a network with one reticulation using SNaQ (Solís-Lemus and Ané, 2016). CFs expected from this fitted network were then used to predict the distribution of observed CFs with a Dirichlet model whose precision parameter was fitted to the CF data (dashed line), and with a multinomial model (probability mass: vertical lines, or density: solid line). Each column shows observed and predicted distribution of CFs whose expectation falls in a particular interval: very close to 0 (left), intermediate (middle), or very close to 1 (right). For CFs with expectations close to 0 or close to 1, the multinomial distribution fits the data much better than the Dirichlet distribution.

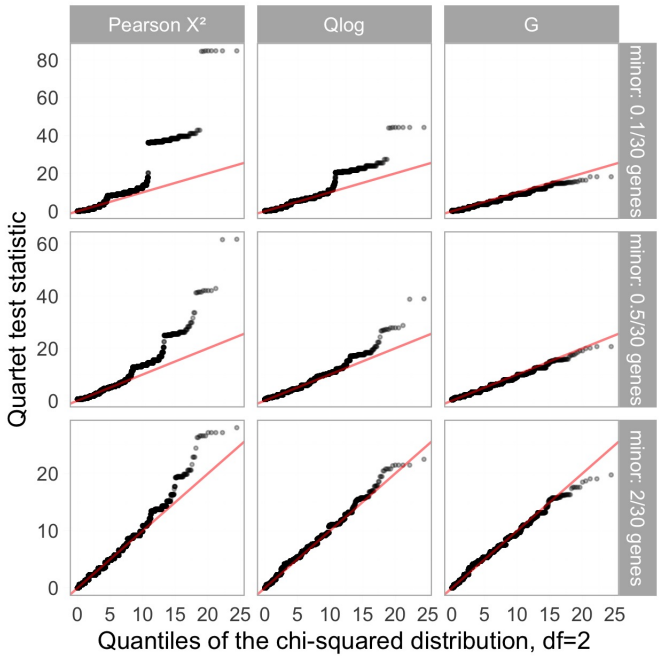

Figure S2: Distribution of X^2, $Q_{\log}$ and G for measuring the fit of quartet CFs, when the data consist of 30 gene trees drawn from a multinomial distribution on 3 quartets, with given expected CFs. The expected number of genes supporting each quartet was 0.1, 14.95, 14.95 (top), 0.5, 14.75, 14.75 (middle) and 2, 14, 14 (bottom). These values would arise, for example, from a reticulate four-taxon phylogeny as in Fig. 1 but with inheritance $\gamma = 0.5$ on both reticulation edges, and with branch lengths $t_1 = t_2 = 4.6$ (top), 3.0 (middle) and 1.6 (bottom) coalescent units. 100,000 data sets were simulated in each case. The sorted values of the 100,000 test statistics are plotted on the vertical axis, versus the theoretical quantiles expected under a χ^2_2 distribution on the horizontal axis. The test statistic is χ^2_2-distributed if the points fall on the diagonal line (in red). When the test statistic is above the diagonal, the p-value obtained by comparison with the χ^2_2 distribution is too small.

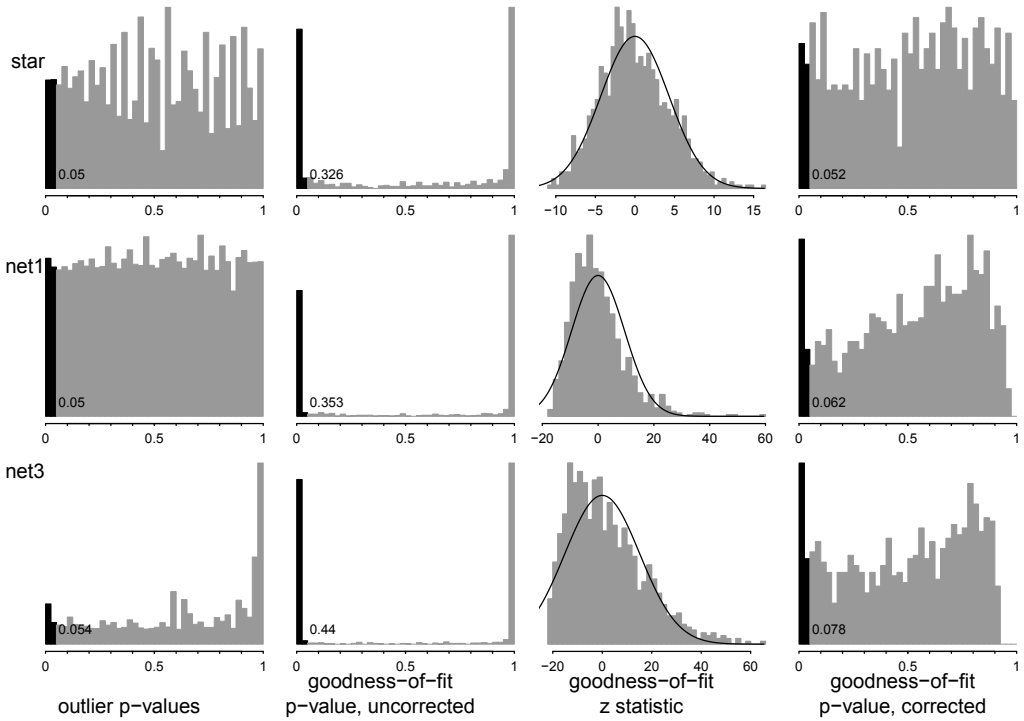

Figure S3: Simulations of the goodness-of-fit test as in Fig. 4, using the Pearson statistic X^2 for the outlier test on each four-taxon set,

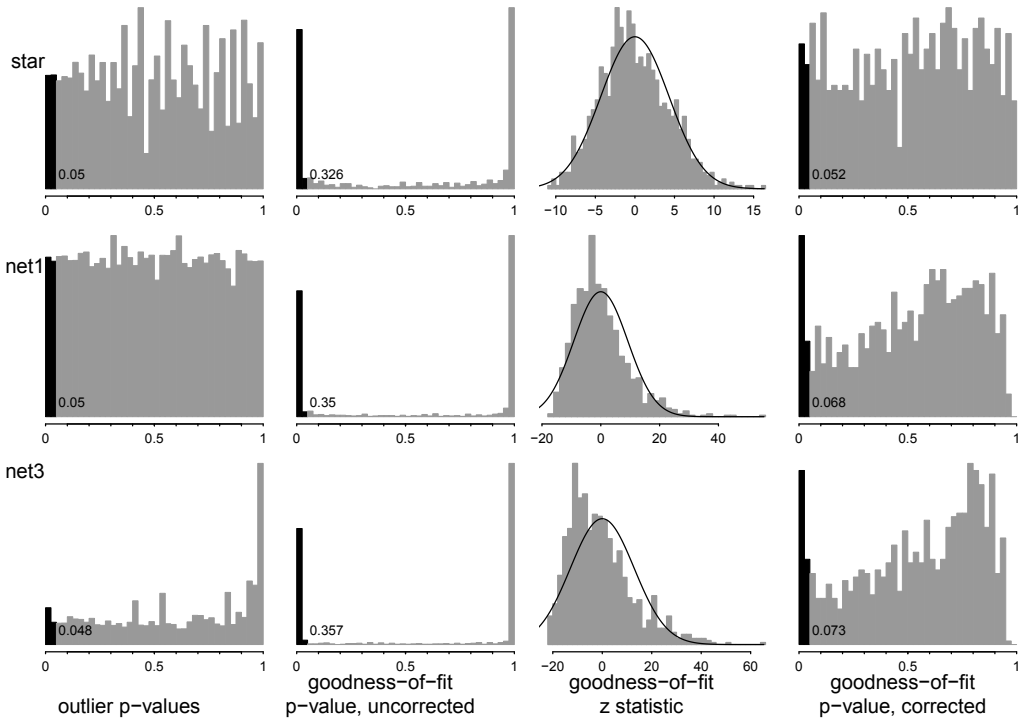

Figure S4: Simulations of the goodness-of-fit test as in Fig. 4, using the $Q_{\log}$ statistic for the outlier test on each four-taxon set,

2 Derivation of the null variance on a star phylogeny

In this section, we prove the claims that were used in the main text to derive (5) for the null variance σ^2 under a star phylogeny. Recall that for a four-taxon set q, the outcome of the outlier test was quantified with the Bernoulli variable $Y_q = 1$ if q is detected as an outlier, $Y_q = 0$ otherwise. We prove here the following claims about the covariance between Y_q and $Y_{q'}$ for two distinct four-taxon sets q and q', under a null phylogeny: $\text{cov}(Y_q, Y_{q'}) = 0$ if $|q \cap q'| = 0$ or 1, $\text{cov}(Y_q, Y_{q'}) = v_2 \simeq 0.000373186$ if $|q \cap q'| = 2$, and $\text{cov}(Y_q, Y_{q'}) = v_3 \simeq 0.002275837$ if $|q \cap q'| = 3$.

Without loss of generality, let $q = \{1, 2, 3, 4\}$. For a given gene, the quartet tree for this gene on q only depends on the coalescent events that occur above the root node, because the phylogeny is a star. Under the coalescent, this quartet can be generated by drawing six independent random variables from the exponential distribution $\mathcal{E}(1)$, that represent the potential coalescent times between each pair of two taxa: $T_{12}, T_{13}, T_{14}, T_{23}, T_{24}, T_{34}$. The smallest of these values determines which pair coalesces first, and hence which quartet the gene has on $\{1, 2, 3, 4\}$.

First consider the case when q and q' do not overlap, or overlap by one taxon. Without loss of generality, let $q' = \{5, 6, 7, 8\}$ (no overlap) or $q' = \{1, 5, 6, 7\}$ (1-overlap). For a given gene, this gene's quartet on q' is determined by random variables that are independent of those that determine the gene's quartet on q, namely: $T_{56}, T_{57}, T_{58}, T_{67}, T_{68}, T_{78}$ (no overlap), or $T_{15}, T_{16}, T_{17}, T_{56}, T_{57}, T_{67}$ (1-overlap). Therefore, the gene's quartet on q is independent of the gene's quartet on q', so Y_q is independent of $Y_{q'}$, and $\text{cov}(Y_q, Y_{q'}) = 0$.

Next, consider the case when q and q' share exactly three taxa. Without loss of generality, let $q' = \{1, 2, 3, 5\}$. In this case, T_{12}, T_{13}, T_{23} influence both the quartet on q and the quartet on q'. We first derive the covariance between the quartets on q (12|34, 13|24, 23|14) and the quartets on q' (12|35, 13|25, 23|15), then derive the covariance between the outlier test outcomes, Y_q and $Y_{q'}$. Let $E_3 = \min\{T_{12}, T_{13}, T_{23}\}$ and let $\{a, b\}$ be the pair that coalesce first among taxa $1, 2, 3$, that is, $T_{ab} = E_3$. The minimum coalescent time E_3 has an exponential distribution with rate 3, and $\{a, b\}$ equals each pair with probability 1/3, independently of E_3. Conditional of $E_3 = t$ and $\{a, b\} = \{1, 2\}$, say, the quartet on q is:

- 12|34 if $E_2 = \min\{T_{14}, T_{24}\} > t$, which occurs with probability e^{-2t} since $E_2 \sim \mathcal{E}(2)$,
- 13|24 if $E_2 < t$ and $T_{14} > T_{24}$, which occurs with probability $(1 - e^{-2t})/2$,
- 23|14 if $E_2 < t$ and $T_{14} < T_{24}$, which occurs with probability $(1 - e^{-2t})/2$.

For q', we get similar conditional probabilities for quartets 12|35, 13|25 and 23|15, depending on T_{15} and T_{25}. Since (T_{14}, T_{24}) and (T_{15}, T_{25}) are independent of each other, the quartets on q and on q' are independent and we get the following probabilities for each pair of quartet, conditional on $E_3 = t$ and $\{a, b\} = \{1, 2\}$.

		q		
		12|34	13|24	23|14
	12|35	e^{-4t}	$e^{-2t}(1-e^{-2t})/2$	$e^{-2t}(1-e^{-2t})/2$
q'	13|25	$e^{-2t}(1-e^{-2t})/2$	$(1-e^{-2t})^2/4$	$(1-e^{-2t})^2/4$
	23|15	$e^{-2t}(1-e^{-2t})/2$	$(1-e^{-2t})^2/4$	$(1-e^{-2t})^2/4$

Integrating over $t = E_3 \sim \mathcal{E}(3)$, we get the probabilities of quartet pairs conditional on $\{a, b\} = \{1, 2\}$:

		q		
		12|34	13|24	23|14
	12|35	15/35	3/35	3/35
q'	13|25	3/35	2/35	2/35
	23|15	3/35	2/35	2/35

Finally, integrating over $\{a, b\}$ gives the joint probabilities of quartet pairs, that is, the joint expected concordance factors:

$$\begin{pmatrix} 19 & 8 & 8 \\ 8 & 19 & 8 \\ 8 & 8 & 19 \end{pmatrix} /(3 \times 35)$$

using the same ordering of quartets for q and for q' as previously. Note that, as expected, the marginals are all 1/3, since each quartet of q (or of q') has equal CF, 1/3.

Asymptotically, all three outlier test statistics are equivalent to $S = 3(X_1^2 + X_2^2 + X_3^2)$ where $X_i = \sqrt{n}(\hat{p}_i - 1/3)$ is a centered and rescaled version of the observed CF for quartet i, such that, as the number of genes n gets large, $\mathbf{X}$ has a multivariate normal distribution with mean 0 and variance from the multinomial distribution:

$$\mathbf{\Sigma} = \frac{1}{9} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} .$$

The outlier test outcome for four-taxon sets q and q' are then: $Y_q = 1$ if $S_q > 5.99$ and 0 otherwise, where S_q is calculated from the quartets on q, and 5.99 is the 0.05 tail quantile for χ^2_2, the chi-square distribution with 2 degrees of freedom, since $S \sim \chi^2_2$ asymptotically. The covariance between Y_q and $Y_{q'}$ is therefore:

$$\mathrm{cov}(Y_q, Y_{q'}) = \mathbb{P}\{S_q > 5.99 \text{ and } S_{q'} > 5.99\} - 0.05 \times 0.05 \,.$$

To calculate this quantity, we consider the joint distribution of $\mathbf{X}_q$ and $\mathbf{X}_{q'}$, which is multivariate with mean 0 and covariance

$$\begin{pmatrix} \mathbf{\Sigma} & \mathbf{C} \\ \mathbf{C} & \mathbf{\Sigma} \end{pmatrix} \text{ where } \mathbf{C} = \frac{11}{9 \times 35} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

was derived from the joint probabilities of quartet pairs above and the general formula for multinomial covariances. Importantly, we can diagonalize $\mathbf{\Sigma}$ and $\mathbf{C}$ with the same eigenvectors:

$$\mathbf{\Sigma} = \mathbf{P}\begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 0 \end{pmatrix}\mathbf{P}^\mathsf{T} \text{ and } \mathbf{C} = \mathbf{P}\begin{pmatrix} \frac{11}{3\times 35} & 0 & 0 \\ 0 & \frac{11}{3\times 35} & 0 \\ 0 & 0 & 0 \end{pmatrix}\mathbf{P}^\mathsf{T}$$

with

$$\mathbf{P} = \begin{pmatrix} 2/\sqrt{6} & 0 & 1/\sqrt{3} \\ -1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \\ -1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \end{pmatrix} \text{ and } \mathbf{P}^\mathsf{T} = \mathbf{P}^{-1}.$$

Therefore, we can write

$$\begin{aligned} \mathbf{X}_q &= \mathbf{P}\,\mathrm{diag}\left(\sqrt{\frac{11}{3\times 35}}, \sqrt{\frac{11}{3\times 35}}, 0\right)\mathbf{U} + \mathbf{P}\,\mathrm{diag}\left(\sqrt{\frac{24}{3\times 35}}, \sqrt{\frac{24}{3\times 35}}, 0\right)\mathbf{Z} \\ \text{and } \mathbf{X}_{q'} &= \mathbf{P}\,\mathrm{diag}\left(\sqrt{\frac{11}{3\times 35}}, \sqrt{\frac{11}{3\times 35}}, 0\right)\mathbf{U} + \mathbf{P}\,\mathrm{diag}\left(\sqrt{\frac{24}{3\times 35}}, \sqrt{\frac{24}{3\times 35}}, 0\right)\mathbf{Z}' \end{aligned}$$

where $\mathbf{U}$, $\mathbf{Z}$ and $\mathbf{Z}'$ are 3-dimensional independent $\mathcal{N}(0, \mathbf{I})$ random variables. The correlation between quartet CFs is mediated by $\mathbf{U}$, which affects both four-taxon sets q and q'. Finally, the outlier test statistics can be written as

$$\begin{aligned} S_q &= 3\mathbf{X}_q^\mathsf{T}\mathbf{X}_q = \left(\sqrt{\frac{11}{35}}U_1 + \sqrt{\frac{24}{35}}Z_1\right)^2 + \left(\sqrt{\frac{11}{35}}U_2 + \sqrt{\frac{24}{35}}Z_2\right)^2 \\ S_{q'} &= 3\mathbf{X}_{q'}^\mathsf{T}\mathbf{X}_{q'} = \left(\sqrt{\frac{11}{35}}U_1 + \sqrt{\frac{24}{35}}Z'_1\right)^2 + \left(\sqrt{\frac{11}{35}}U_2 + \sqrt{\frac{24}{35}}Z'_2\right)^2. \end{aligned}$$

This implies that, conditional on $\mathbf{U}$, $\frac{35}{24}S_q$ and $\frac{35}{24}S_{q'}$ are independent and both have a non-central chi-squared distribution with 2 degrees of freedom, with non-centrality parameter determined by $\lambda = \frac{11}{24}(U_1^2 + U_2^2)$. We can then use the density of the non-central chi-square distribution to calculate the probability of both $S_q > 5.99$ and $S_{q'} > 5.99$ conditional on $\mathbf{U}$, then integrate over $u = U_1^2 + U_2^2$, which has a (central) chi-squared distribution with 2 degrees of freedom:

$$\mathbb{P}\{S_q > 5.99 \text{ and } S_{q'} > 5.99\} = \int_{u=0}^{\infty} \mathbb{P}\left\{\chi^2_{2,\lambda=(11/24)u} > \frac{35}{24}\,5.99\right\}^2 \frac{1}{2}e^{-u/2}du \simeq 0.004775837.$$

This integral was calculated numerically, using the R package cubature v2.0.4. After subtracting 0.05×0.05, we get $v_3 \simeq 0.002275837$, as claimed in the main text.

Finally, consider the case when q and q' share exactly two taxa. This reasoning is similar to that in the previous case: we first derive the joint probability of each of the quartets on q

and on q', then the covariance matrix $\mathbf{C}$, which lets us write S_q and $S_{q'}$ as independent non-central chi-squared conditional on a shared $\mathbf{U}$. The last step is a slightly more complicated integral. Without loss of generality, let $q = \{1, 2, 3, 4\}$ and $q = \{1, 2, 5, 6\}$. They share a single coalescent time: T_{12}. Conditional of $T_{12} = t$, the quartet on q is:

- $12|34$ if $E_5 = \min\{T_{13}, T_{23}, T_{14}, T_{24}, T_{34}\} > t$, which occurs with probability e^{-5t} since $E_5 \sim \mathcal{E}(5)$; or if $E_5 \leq t$ and $T_{34} = E_5$, which occurs with probability $(1 - e^{-5t})/5$;
- $13|24$ if $E_5 \leq t$ and $E_5 = T_{13}$ or T_{24}, which occurs with probability $(1 - e^{-5t}) \times 2/5$;
- $23|14$ if $E_5 \leq t$ and $E_5 = T_{23}$ or T_{14}, which occurs with probability $(1 - e^{-5t}) \times 2/5$.

For q', the conditional probabilities are the same for quartets $12|56$, $15|26$ and $25|16$, depending on $T_{15}, T_{25}, T_{16}, T_{26}$ and T_{56}. Since these coalescent times are independent of those determining the quartet on q, the quartets on q and q' are independent conditional on $T_{12} = t$, so we can calculate their joint conditional probabilities. After integrating over T_{12} we get the joint (unconditional) probabilities:

		q		
		$12\|34$	$13\|24$	$23\|14$
q'	$12\|56$	$5/33$	$3/33$	$3/33$
	$15\|26$	$3/33$	$4/33$	$4/33$
	$25\|16$	$3/33$	$4/33$	$4/33$

It follows that the covariance between $\mathbf{X}_q$ and $\mathbf{X}_{q'}$ is now:

$$\mathbf{C} = \frac{1}{9 \times 11} \begin{pmatrix} 4 & -2 & -2 \\ -2 & 1 & 1 \\ -2 & 1 & 1 \end{pmatrix} = \mathbf{P} \begin{pmatrix} \frac{2}{3 \times 11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{P}^\mathsf{T}$$

such that we can write

$$\begin{aligned} \mathbf{X}_q &= \mathbf{P} \operatorname{diag}\left(\sqrt{\frac{2}{3 \times 11}}, 0, 0\right) \mathbf{U} + \mathbf{P} \operatorname{diag}\left(\sqrt{\frac{9}{3 \times 11}}, \sqrt{\frac{1}{3}}, 0\right) \mathbf{Z} \\ \text{and } \mathbf{X}_{q'} &= \mathbf{P} \operatorname{diag}\left(\sqrt{\frac{2}{3 \times 11}}, 0, 0\right) \mathbf{U} + \mathbf{P} \operatorname{diag}\left(\sqrt{\frac{9}{3 \times 11}}, \sqrt{\frac{1}{3}}, 0\right) \mathbf{Z}' \end{aligned}$$

where, as before, $\mathbf{U}$, $\mathbf{Z}$ and $\mathbf{Z}'$ are 3-dimensional independent $\mathcal{N}(0, \mathbf{I})$ random variables. The outlier test statistics can be written as

$$\begin{aligned} S_q &= 3\mathbf{X}_q^\mathsf{T}\mathbf{X}_q = \left(\sqrt{\frac{2}{11}}U_1 + \sqrt{\frac{9}{11}}Z_1\right)^2 + Z_2^2 \\ S_{q'} &= 3\mathbf{X}_{q'}^\mathsf{T}\mathbf{X}_{q'} = \left(\sqrt{\frac{2}{11}}U_1 + \sqrt{\frac{9}{11}}Z_1'\right)^2 + {Z_2'}^2 . \end{aligned}$$

Conditional on U_1, S_q and $S_{q'}$ are independent, but have a more complicated conditional distribution than before: the sum of a rescaled non-central chi-squared distribution and an independent chi-squared distribution, both with 1 degree of freedom: $\chi^2_1 + \frac{9}{11}\chi^2_{1,\lambda=(2/9)u^2}$. We can calculate the conditional probability that q is detected as an outlier numerically:

$$\mathbb{P}\{S_q > 5.99 | U_1 = u\} = \int_z \mathbb{P}\left\{\chi^2_{1,\lambda=(2/9)u^2} > \frac{11}{9}(5.99 - z^2)\right\} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \simeq 0.003030943$$

The numerical integration was performed using the R package cubature v2.0.4. After subtracting 0.05×0.05, we get $v_2 \simeq 0.000373186$, as claimed in the main text. As expected, v_2 is much smaller than v_3, because q and q' are less correlated (via a single pair only) when they share 2 taxa than when they share 3 taxa.

On 23 taxa, (5) gives $\sigma^2 = 12.7$. The discrepancy with our estimate $\hat{\sigma} = 19.1$ from simulations is likely due to the fact that 19.1 is an estimate of $\mathbb{E}(Z^2)$ rather than the variance $\mathbb{E}(Z^2) - (\mathbb{E}Z)^2$. For a conservative test, we recommend the normalization of Z by $\mathbb{E}(Z^2)$, as done in the main text.

3 Analysis of *Stachyurus* data from Feng *et al.* (2020)

Stachyurus is a genus of shrubs and small trees from East Asia, recently studied by Feng *et al.* (2020). We analyzed their data on 17 taxa and 1362 genes. For candidate networks, we used their phylogenetic networks estimated with 0 to 3 reticulations, because Feng *et al.* (2020) present a network with 2 reticulations, and because networks estimated with 4 or more reticulations conflicted with the rooting in which *S. praecox* (from Japan) is sister to the remainder of the genus (from the Asian mainland). All networks failed the goodness-of-fit test, but the lack of adequacy was very similar between networks with 1 to 3 reticulations (Table S1), in line with the conclusion by Feng *et al.* (2020) that there is support of one reticulation and uncertainty about other reticulations.

We sought to find a subnetwork that fits adequately, so we removed taxa from the network with $h = 2$ favored by Feng *et al.* (2020), until we obtained a subnetwork with a p-value above 0.05. The subsample was selected by removing taxa most represented in outlier quartets on the network with $h = 2$ reticulations, except that we kept taxa whose removal would have removed a reticulation in the network, such that the subnetwork retained $h = 2$ reticulations. The subsample had the following 9 taxa: *S. oblongifolius*, *S. yunnanensis* (EM), *S. obovatus*, *S. chinensis* (SX), *S. chinensis* (ZJ), *S. chinensis* (AH), *S. chinensis* (QCS), *S. retusus*, and *S. chinensis* var. *latus*. We then tested the goodness of fit of the networks with 1 or no reticulations on the same subsample. Both subnetworks were rejected (Table S1), showing that the 2 reticulations do contribute to the fit of the subnetwork with $h = 2$, not simply the pruning of taxa whose data were poorly fit by the larger network.

Taxon set	Topology: number of reticulations	z	$\hat{\sigma}$	$z/\hat{\sigma}$	p-value
full taxon set	$h = 0$	144.2	4.688	30.7	$p \ll 10^{-100}$
	$h = 1$	126.6	4.685	27.0	$p \ll 10^{-100}$
	$h = 2$	123.8	4.678	26.5	$p \ll 10^{-100}$
	$h = 3$	119.9	4.639	25.8	$p \ll 10^{-100}$
subsample	$h = 0$	23.18	2.234	10.4	2×10^{-25}
	$h = 1$	5.191	2.239	2.32	0.010
	$h = 2$	3.147	2.203	1.43	0.077

Table S1: Analysis of the *Stachyurus* data sets from Feng *et al.* (2020) with 1362 genes, using the G statistics. The test statistic $z/\hat{\sigma}$ corrects for dependence, with $\hat{\sigma}$ obtained using simulations of 10,000 data sets, each with 1362 genes. Branch lengths were optimized for each network. The full taxon set has the 17 *Stachyurus* taxa sampled by Feng *et al.* (2020). The subsample has 9 taxa. After pruning the network with $h = 3$ reticulations, the resulting subnetwork has only 2 visible reticulations, making it equivalent to the subnetwork with $h = 2$.

References

Feng, Y., Comes, H. P., and Qiu, Y.-X. (2020). Phylogenomic insights into the temporal-spatial divergence history, evolution of leaf habit and hybridization in *Stachyurus* (Stachyuraceae). *Molecular Phylogenetics and Evolution*, **150**, 106878.

Solís-Lemus, C. and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, **12**(3), e1005896.