Bjorner, M. B., Molloy, E. K., Dewey, C. N., & Solís-Lemus, C. (2024). Detectability of varied hybridization scenarios using genome-scale hybrid detection methods. *Bulletin of the Society of Systematic Biologists*. https://doi.org/10.18061/bssb.v3i1.9284

Descriptions

Detectability of Varied Hybridization Scenarios Using Genome-Scale Hybrid Detection Methods

Marianne B Bjorner[1], Erin K Molloy[2], Colin N Dewey[3], Claudia Solís-Lemus[4]

[1] Department of Computer Sciences & Wisconsin Institute for Discovery, University of Wisconsin - Madison, [2] Department of Computer Science, University of Maryland, College Park, [3] Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, [4] Department of Plant Pathology & Wisconsin Institute for Discovery, University of Wisconsin - Madison

https://doi.org/10.18061/bssb.v3i1.9284

Bulletin of the Society of Systematic Biologists

Abstract

Hybridization events complicate the accurate reconstruction of phylogenies, as they lead to patterns of genetic heritability that are unexpected under traditional, bifurcating models of species trees. This phenomenon has led to the development of methods to infer these varied hybridization events, both methods that reconstruct networks directly, as well as summary methods that predict individual hybridization events from a subset of taxa. However, a lack of empirical comparisons between methods – especially those pertaining to large networks with varied hybridization scenarios – hinders their practical use. Here, we provide a comprehensive review of popular summary methods: TICR, MSCquartets, HyDe, Patterson's D-Statistic (ABBA-BABA), D3, and Dp. TICR and MSCquartets are based on quartet concordance factors gathered from gene tree topologies and HyDe, Patterson's D-Statistic, D3, and Dp use site pattern frequencies to identify hybridization events between sets of three taxa. We then use simulated data to address questions of method accuracy and ideal use scenarios by testing methods against complex networks which depict gene flow events that differ in depth (timing), quantity (single vs. multiple, overlapping hybridizations), and rate of gene flow (γ). We find that deeper or multiple hybridization events may introduce noise and weaken the signal of hybridization, leading to higher relative false negative rates across all methods. Despite some forms of hybridization eluding quartet-based detection methods, MSCquartets displays high precision in most scenarios. While HyDe results in high false negative rates when tested on hybridizations involving extinct or unsampled ghost lineages, HyDe is the only method able to identify the direction of hybridization, distinguishing the source parental lineages from recipient hybrid lineages. Lastly, we test the methods on a dataset of ultraconserved elements from the bee subfamily Nomiinae, finding possible hybridization events between clades which correspond to regions of poor support in the species tree estimated in a previous study.

1 Introduction

Phylogenetics studies the evolutionary history between organisms. In many popular phylogenetic inference models, these relationships are assumed to be best represented as a binary tree, where each child node arises from only one direct parent (Bouckaert, 2019; Nguyen et al., 2015; Stamatakis, 2014). However, a binary tree model ignores the possibility of a reticulation or gene flow event. Gene flow occurs when members of one population reproduce or otherwise exchange genetic information with another population, which leads to the formation of admixed populations or new hybrid species lineages (Barton & Hewitt, 1985). These reticulation events transform bifurcating phylogenetic trees into network structures, wherein the taxa affected have more than one parental lineage (Moret et al., 2004). Non-tree-like evolution is common across the tree of life, found in groups such as insects (Suvorov et al., 2022), plants (Hibbins & Hahn, 2021) and mammals (Racimo et al., 2015).

The study of gene flow events in the tree of life have been aided by recent advances in sequencing technology, granting evolutionary researchers access to genome-scale information. This abundance of information can be used to infer reticulation events such as introgression, hybrid speciation, and horizontal gene transfer. Each mechanism for gene flow leaves behind various traces in a population's genetic information, and may be identified through hybridization detection methods that leverage gene trees or

sequence information (Hibbins & Hahn, 2021). Though the exact biological processes originating gene flow might differ, in this paper, we broadly refer to descendants of reticulation events as hybrids, and the methods that detect them as hybridization methods.

Many existing methods to infer phylogenies are based on a binary tree, where they do not account for reticulations between taxa. While this assumption limits the search space to only trees, it might be an unreasonable assumption, especially for populations where gene flow is common or expected. Methods to infer phylogenetic networks, such as those that use maximum likelihood, Bayesian inference, and combinatorial techniques (Allman et al., 2019; Solís-Lemus & Ané, 2016; Wen et al., 2018; Wen & Nakhleh, 2017; Zhang et al., 2017) are becoming increasingly popular for their ability to overcome the strictly bifurcating assumption. While valuable for studies of few taxa, these methods become very computationally expensive with increasingly large datasets. Furthermore, most network methods require specification of the number of expected hybrid events, as any increase in number of reticulations artificially increases the likelihood of the network (Markin et al., 2022).

Alternatively, to find evidence for individual hybridization events, summary methods analyze subsets of triples or quartets of taxa – which is an intrinsically more scalable endeavor than the search in network space – without any predetermination of the total of number of hybrids in the phylogeny (Hibbins & Hahn, 2021). Despite these advantages, summary methods still require comparisons to each other in order to address questions of method accuracy. The hybrid detection methods compared in this simulation study (Table in Supplementary Material) are MSCquartets (Mitchell et al., 2019), TICR (Stenz et al., 2015), HyDe (Kubatko & Chifman, 2019), Patterson's D-Statistic (Patterson et al., 2012), also known as the ABBA-BABA test, as well as methods derived from the D-Statistic such as D_p (Hamlin et al., 2020) and D_3 (Hahn & Hibbins, 2019). These methods (excluding TICR) identify specific hybrid relationships within either subsets of four or three taxa. TICR, in contrast, tests for how well a binary population tree fits the data (with failure to reject the population tree suggesting no hybridization). However, a significant result from any of these methods is not isolated to hybridization alone, and could also be due to assumptions these methods make related to substitution models, ultrametric trees, differences in population sizes, and so forth.

Here, we address questions of method accuracy and ideal use scenarios by testing these five summary methods against complex networks which depict gene flow events that differ in depth (timing), quantity (number of hybridization events: single vs. multiple), and proportion of genes transferred through the hybridization event (inheritance probability γ). We note that we are not treating hybridization events as continuous flow of genes over a period of time; instead, we treat each hybridization event as being instantaneous, representing it by a single arrow (see Figure 2). The hybridization scenarios also differ in terms of time consistency, a characteristic of reconstructed networks that, when violated, arises in the presence of incomplete sampling or ghost lineages (Moret et al., 2004; Pang & Zhang, 2022; Tricou et al., 2022). Finally, we use hybridization detection methods to analyze published empirical data from the bee subfamily Nomiinae, a dataset complete with sequences, and estimated gene and species trees.

2 Materials and Methods

We begin this section by reviewing the five methods evaluated in our study, breaking them into two classes: methods that take gene trees as input and methods that take molecular sequences as input. We then describe the generation of synthetic data and the metrics for evaluating methods. Lastly, we outline our re-analysis of a dataset of ultraconserved elements (UCEs) for the subfamily Nomiinae.

2.1 Methods based on gene trees

MSCquartets (Allman et al., 2021, 2021) and TICR (Cai & Ané, 2020; Stenz et al., 2015) both rely on frequencies of quartet gene tree topologies to conduct tests into how well a tree-like evolution fits the data and whether there is evidence for hybridizations. These tests are based on the multispecies coalescent model (MSC) which defines expected distributions of gene tree topologies under incomplete lineage sorting (ILS) (Allman et al., 2011; Rannala & Yang, 2003). ILS, also called deep coalescence, occurs when individual gene histories fail to coalesce at the same time as their given species history (Maddison, 1997). An example is shown in Figure 1. Note that this figure displays gene trees that are rooted and ultrametric; however, only the unrooted gene tree topologies are used by MSCquartets and TICR (i.e., the input to these methods is not required to be ultrametric).

A resolved tree on four taxa (called a quartet) can take on one of three unrooted topologies: one concordant and two discordant with the species tree (Figure 1). Under the MSC model, the probability of the concordant gene tree is strictly greater than that of the two discordant gene trees, which have equal probability (Allman & Rhodes, 2007) (note that for five or more taxa, the most probable unrooted gene tree may not be concordant with the unrooted species tree (Degnan, 2013)). Given a model species tree and a number of gene trees, we can compute the expected frequencies of each quartet, referred to as concordance factors (CFs), where the CFs of concordant topologies are called major CFs, as they align with the major tree, or species tree, and the CFs of discordant topologies are called minor CFs (Allman et al., 2021). The observed quartet counts CFs (qcCFs) are gathered by counting how often each of three possible resolved quartet trees appears (Figure 1) across the input set of gene trees (Allman et al., 2021), typically normalizing by the number of gene trees displaying any one of the three possible quartets. After normalization, we can denote them estimated CFs, instead of qcCFs. We will continue to use the terminology "qcCF" to make it clear we refer to empirical quantities, as opposed to expected probabilities (simply denoted CFs). Major qcCFs are expected

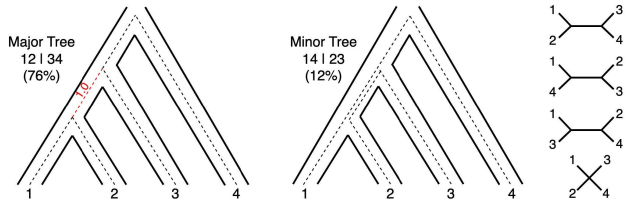

Figure 1. Concordant (left) and discordant (middle) gene and species histories. Solid lines denote species history, and dotted lines denote an individual gene history. The middle tree shows an example of incomplete lineage sorting (ILS) when the gene history fails to coalesce with the ancestral speciation event. The tree on the left (after unrooting) corresponds to the bipartition 12|34 and agrees with the unrooted species tree; thus it is referred to as the "major tree". Under the MSC model, this tree, with an internal branch length of 1:0 coalescent units, is expected to have a frequency of 76% in a given sample of gene trees. The tree in the center (after unrooting) corresponds to the bipartition 14j23 and disagrees with the unrooted species tree; thus, it is referred to as a "minor tree". Under the MSC model, the minor trees have an expected frequency of 12%. The far right depicts four possible quartet relationships, used by MSCquartets (Allman et al., 2021) and TICR (Stenz et al., 2015) depicted from top to bottom: the major tree (12|34), the minor trees (14|23 and 13|24), and the star tree (1234), which neither agrees nor disagrees with the species tree.

to be greater than the two minor qcCFs, and the two minor qcCFs are expected to be equal under MSC. Deviation from this expectation violates an ILS-only model of genetic inheritance, and MSCquartets leverages this invariant, looking at singularities in the space of possible topologies for each subset of four species. TICR, on the other hand, uses normalized qcCFs to conduct a χ^2 goodness-of-fit test against a specific model species tree.

2.1.1 MSCquartets

MSCquartets (Allman et al., 2021) is an R package (Rhodes et al., 2021) that takes as input previously inferred gene trees, from which it computes the observed qcCFs. If the input gene trees are not fully resolved, it is possible to have a star tree when looking at four taxa (Figure 1). Unresolved (star) trees can optionally be removed or redistributed among the resolved topologies. The resulting qcCFs are then compared to the expected invariant defined in (Mitchell et al., 2019) derived from the MSC model to test the hypothesis of whether a specific four-taxon subset follows a tree-like pattern (in agreement with the MSC model) or not. Each hypothesis test produces a p-value. We note that MSCquartets does not require a specific tree topology to test the CF expectations against, instead utilizing the information on all three qcCFs per four-taxon subset to compute the test statistic. When expectations are violated, it leads to low support for a tree-like species history between the four taxa. In other words, significant results fail to support an ILS-only model of evolution, and hybridization becomes a possible explanation for the imbalanced relationship between quartets. However, certain hybrid relationships elude detection using these quartet based methods. These correspond to hybridizations between sister taxa, as such hybrid relationships will not lead to an imbalance in the frequency of minor quartets.

In terms of computational efficiency, MSCquartets is used in a pipeline that requires costly preprocessing steps to produce the input, including aligning sequences and estimating gene trees. With these precalculated, the two primary factors that influence the speed of this approach are the number n of taxa and the number g of gene trees. Just consider that computing the observed qcCFs can be done by identifying the quartet displayed by a gene tree for each of the $\frac{n!}{4!(n-4)!}$ possible subsets of four species, repeating across all gene trees. This procedure alone would give the time complexity of MSCquartets a lower bound of $O(n^4 g)$. Thus, MSCquartets may be time consuming for large numbers of taxa.

2.1.2 TICR

In its original implementation, TICR – Tree Incongruence Checking in R – (Stenz et al., 2015) was used as part of a pipeline that begins with a set of alignments (one per gene), estimates gene trees, calculates qcCFs from the (estimated) gene trees, and finally calculates a population tree based on the qcCFs using the software `Quartet-Max-Cut` (Snir & Rao, 2012; Stenz et al., 2015). However, the only inputs required by TICR are the observed qcCFs computed from gene trees and the expected CFs calculated from a hypothesized population tree. Recently, this method was extended to test goodness-of-fit on a given population network, rather than population tree (Cai & Ané, 2020). We note that in our experiments we use the TICR version implemented in the Julia package called `QuartetNetworkGoodnessFit.jl` (Cai & Ané, 2020) though we restrict our tests to the case of population trees, not networks.

Given a fully resolved population tree with branch lengths in coalescent time, the expected probabilities of observing quartet relationships can be directly computed. For example, in Figure 1 assuming the internal branch in red has length $t = 1.0$ coalescent units, the probability of the major gene tree is given by $1 - \frac{2}{3}e^{-t} = 0.76$ (Allman et al., 2011). TICR computes a χ^2 goodness-of-fit test statistics that evaluates the fit of the observed qcCFs to the expected CFs under the ILS-only model. TICR can also be used to test for the likelihood of panmixia, or a star tree which occurs when all taxa arise from the same common ancestor and diverge at the same time, though any occurrences of star trees in the input gene trees are ignored when calculating qcCFs. TICR uses the p-values of the individual tests to form an overall test that inspects whether the distribution of observed qcCFs falls within the expected CFs of the input tree or network. This overall test indicates whether the proposed population tree fits the observed qcCFs. Although TICR does not directly test for the presence or absence of

specific hybridizations, by failing to reject a specific tree model, it provides lack of evidence for hybridization.

In terms of computational efficiency, the remarks made above for MSCquartets apply to TICR, although it is worth noting that the TICR pipeline additionally needs to estimate a species tree and compute the expected CFs based on it.

2.2 Methods based on (aligned) sequences

HyDe (Kubatko & Chifman, 2019), Patterson's D-Statistic (Patterson et al., 2012), D_3 (Hahn & Hibbins, 2019), and D_p (Hamlin et al., 2020) are all methods that use site pattern frequencies or pairwise differences to test the null hypothesis of tree-like evolutionary patterns (ILS-only). This eliminates the need for the estimation of gene trees, such that sequences can be used directly as input. However, it is important to use the input of many gene sequences, as opposed to few long sequences, as these tests are based on a model of coalescent independent sites, or designed for use at the allele-level (Kubatko & Chifman, 2019; Patterson et al., 2012). These sequences must also come from equidistant gene trees, where each taxon is equidistant from the root. While HyDe, Patterson's D-Statistic, and D_p use rooted triples plus an outgroup, D_3 uses a rooted triple without an outgroup. This is advantageous if a root is known as a poorly chosen, or distant outgroup can result in inclusion of ghost hybridizations and lead to false interpretations (Hahn & Hibbins, 2019; Tricou et al., 2022). Ghost hybridizations are defined as hybridization events when one (or both) of the parent lineages that provide genetic material to the hybrid node are either extinct or unsampled. Note that the Patterson's D-statistic, D_3, and D_p are intended as tests for introgression between two species, but do not indicate directionality, while HyDe is intended to test for a hybridization event between two taxa, that results in a third taxon. However, none of these tests are designed to detect reticulations between sister taxa, as they all rely on the disruption of symmetry between sister taxa to indicate the presence of hybridization; hybridization between sister taxa produces no such signal (Allman et al., 2011; Hahn & Hibbins, 2019; Hibbins & Hahn, 2021; Kubatko & Chifman, 2019). Additionally, they require that the species relationship between the triple and its outgroup (if any) is known.

2.2.1 HyDe

Distinct from other methods which evaluate for the overall presence of hybridization, HyDe (Blischak et al., 2018) identifies a singular parent-hybrid relationship between a triple, given its outgroup. It can also be used to estimate a mixing parameter γ, depicting the proportion of genetic material contributed by each parental lineage.

HyDe is based on phylogenetic invariants, or a function of site pattern probabilities, which evaluate to zero when consistent with given displayed tree models (Allman et al., 2011; Kubatko & Chifman, 2019). The linear invariants (f_1 and f_2) depend on mixing parameters γ and $1-\gamma$, respectively,

$$f_1 = p_{iijj} - p_{ijij}$$
$$f_2 = p_{ijji} - p_{ijij}$$

where p_{iijj} is the probability for the site pattern $iijj$ with mixing parameter γ. As a result, HyDe can also be used to estimate the mixing parameter γ between two taxa that are putative parent lineages of a proposed hybrid as $\frac{f_1}{f_2} = \frac{\gamma}{1-\gamma}$. When there is no hybridization, γ is 0, so the ratio is expected to be zero.

Site pattern probabilities observed in the sample are used to form estimates of f_1 and f_2, along with means and variance. These are then rearranged with the Geary-Hinkley transformation to form the Hils statistic:

$$H := \frac{\hat{f}_2(\frac{\hat{f}_1}{\hat{f}_2} - \frac{\mu_{f_1}}{\mu_{f_2}})}{\sqrt{\hat{\sigma}^2{}_{f_2}(\frac{\hat{f}_1}{\hat{f}_2})^2 - 2\hat{\sigma}_{f1,f2}\frac{\hat{f}_1}{\hat{f}_2} + \hat{\sigma}^2_{f1}}}$$

When the number of sampled sites is large, this follows the normal distribution $N(0,1)$, under the null hypothesis of ILS-only, or no hybrid speciation (Kubatko & Chifman, 2019). This allows direct interpretation of HyDe test results without the need for resampling by bootstrapping. However, significance levels should be adjusted with a Bonferroni correction due to multiple hypothesis testing, as HyDe considers all possible combinations of sets of three taxa, where one hybrid taxon is tested for every two distinct parent taxa.

As a C-backed python package, HyDe is designed for use with multiple individuals per taxon, from a phylogeny where the outgroup is specified, and aligned sequence information is provided in PHYLIP format. Power increases with increasing sequence length, with a recommendation that sequence length is at minimum 50 kbp (Blischak et al., 2018; Kong & Kubatko, 2021). In addition, HyDe outputs counts of site pattern observations, namely AABB, ABBA, AABC,... that can be used for calculations of other statistics based on site pattern frequencies such as the Patterson's D-Statistics (ABBA-BABA) (Patterson et al., 2012), D_p (Hamlin et al., 2020), and can be rearranged for use in pairwise distance metrics, such as D_3 (Hahn & Hibbins, 2019).

In terms of computational efficiency, HyDe calculates site pattern frequencies from subsets of sequences in an alignment; thus, its speed is impacted by the number of taxa and alignment length. HyDe is capable of processing alignments of 20 taxa and 100,000 sites in under a minute (Kubatko & Chifman, 2019). For each set of three taxa (T_1, T_2, T_3), HyDe labels two as parental populations, and one as a hybrid population. HyDe evaluates each of the three possible hybrid relationships between the taxa.

Sites from each taxon T_1, T_2, and T_3 are then compared to sites from an outgroup to calculate the Hils statistic from $f1$ and $f2$. For a set of $n+1$ taxa, (where n is number of taxa excluding the outgroup), each with an aligned sequence of length L, HyDe performs $3\frac{n!}{3!(n-3)!}$ tests, where each test involves comparison of L sites, across each of the three taxa for hybrid testing, plus the outgroup resulting in a time complexity of $O(n^3L)$.

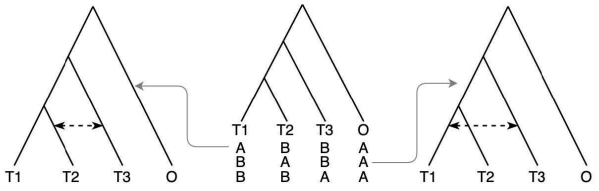

Figure 2. **Center:** Population tree of the form (((T1,T2),T3),O). Under each taxa, ABBA, BABA, and BBAA are three possible site patterns, with positions in the site pattern sequence corresponding to particular taxa. Under ILS, the proportion of sites where T3 and T1 match should be equal to the proportion of sites where T2 and T3 match, given that T3 is equidistant on the species tree from T1 and T2. **Left:** An excess of ABBA, corresponds to a gene flow event between T2 and T3. **Right:** An excess of BABA, corresponds to a gene flow event between T1 and T3.

2.2.2 Patterson's D-Statistic (ABBA-BABA)

Patterson's D-Statistic (Patterson et al., 2012), much like HyDe, involves the calculation of site pattern frequencies from SNPs and scales in a similar manner with respect to sequence length and number of taxa to compare. Patterson's D-statistic is defined as $D = \frac{ABBA-BABA}{ABBA+BABA}$ where A and B are distinct alleles. Each of the four positions in these nucleotide sequences belong to one of four taxa: three of these taxa are compared to each other for hybrid identification, and the fourth is assumed to be an outgroup. The D-statistic uses site patterns frequencies, where the frequencies of site patterns between taxa T_3, T_2 and between taxa T_3, T_1 are assumed to be equal, given a known species tree relationship, where T_1 and T_2 are sisters (such as in $(((T_1, T_2), T_3), O)$, further illustrated in Figure 2). Under the null hypothesis of no hybridization, the D-statistic is expected to be zero, as the frequency of "ABBA" and "BABA" patterns should be equal. Significant deviation from 0 are attributed to gene flow events. We note that the Patterson's D-statistic does not indicate the directionality of gene flow.

In terms of computational efficiency, testing on all permutations of three taxa, with a specified outgroup, computing time of Patterson's D-Statistic on an alignment of length L, containing n taxa scales with respect to $4L \times \frac{n!}{(n-3)!}$, giving a time complexity is $O(n^3L)$. However, it is worth noting that Patterson's D-statistic is symmetric, such that the resulting D-statistic of the topology (T_1,T_2),T_3) is equal to the negative D-statistic given the topology ((T_2,T_1),T_3). Further reduction of the number of tests occurs when provided the correct topology of the underlying species tree, as is necessary, as the D-statistic relies upon an assumed structure of the major tree. As a result, only one test is necessary per subset of three taxa.

2.2.3 D_3

Motivated by the original Patterson's D-Statistic, D_3 was created as an alternative method that does not require an outgroup, relying only on three taxa (Hahn & Hibbins, 2019). It uses pairwise distances instead of site patterns frequencies, where the pairwise distances between taxa T_3, T_2 and between taxa T_3, T_1 are assumed to be equal, given a known species tree relationship, where T_1 and T_2 are sisters (such as in $((T_1, T_2), T_3)$).

D_3 can be calculated as a ratio of pairwise distances between three sequences $D_3 = \frac{d_{T_2T_3} - d_{T_1T_3}}{d_{T_2T_3} + d_{T_1T_3}}$ where d_{ij} corresponds to the distance between taxa i and j. Here, significant deviation of D_3 from 0 may imply gene flow between taxa T_3 and T_2, in the case of a negative result, and between T_1 and T_3, in the case of a positive result. The distance used in this calculation can either be the uncorrected distance, i.e. Hamming, or a measure of distance corrected for multiple hits. This operates much like the original D-statistic to test for the presence, but not the directionality, of gene flow events.

In terms of computational efficiency, unlike the Patterson's D-Statistic and D_p, D_3 does not include comparison to an outgroup. As a result, the method is slightly faster as there is 25% less sequence information to analyze and compare to, as there are now three sequences instead of four. As with the Patterson's D-Statistic, the species topology must be known, and due to symmetry, only one test per combination of three taxa is necessary; these methods have the same time complexity: $O(n^3L)$.

2.2.4 D_p

D_p adds the site pattern frequency BBAA to the denominator of the original Patterson's D-Statistic in order to estimate the net proportion of the genome resulting from introgression (Hamlin et al., 2020). This feature provides comparability with HyDe's computation of γ. The denominator in D_p accounts for the total number of variable sites: $D_p = \frac{|ABBA-BABA|}{BBAA+ABBA+BABA}$.

In terms of computational efficiency, as with other forms of the D-test, D_p tests each combination of n taxa given an alignment of length L and a specified topology. Its time complexity is equivalent to that of Patterson's D-Statistic: $O(n^3L)$.

2.3 Simulations

All methods were tested on the same proposed networks and compared in their ability to test for the presence of hybridization events in relation to either the three or four taxa used as input. Networks used for simulation are illustrated in Figures in the Supplementary Material. These vary in number of reticulations, number of taxa, depth of reticulations and their mixing parameter γ, which denotes how much ancestral DNA is passed from the minor hybrid edge to the hybrid node. Size ranges from 4–25 taxa, where the number of reticulation events for networks with 10, 15, and 25 taxa are 20% of the number of taxa. Each reticulation event can have singular or multiple affected taxa downstream of the hybridization event. We name the networks based on the number of taxa (n) and number of hybridizations (h). For example, the network denoted n4h1 has four taxa ($n = 4$) and one hybridization ($h = 1$). Six

out of the twelve networks are replicated from earlier studies; specifically, four networks (n4h1, $\text{n4h1}_{\text{introgression}}$, n8h3 and n5h2) were used in (Kong & Kubatko, 2021), and two networks (n10h2 and n15h3) were used in (Solís-Lemus & Ané, 2016). The mixing parameters γ were also kept consistent between prior studies, where for $\text{n4h1}_{\text{introgression}}$, n8h3, and n5h2 $\gamma = 0.5$, and n4h1 was tested with multiple values ($\gamma \in 0, 0.1, 0.2, 0.3, 0.4, 0.5$). In n10h2, $\gamma \in 0.2, 0.3$, and in n15h3, $\gamma \in 0.2, 0.2, 0.3$, where the deepest hybridization had the highest mixing parameter. We consider variants of networks n10h2 (with two hybridization events) and n15h3 (with three hybridization events), where some hybrid edges are removed, leaving a single reticulation with its original mixing parameter γ. This results in derived networks $\text{n10h1}_{\text{deep}}$ (γ = 0.3), $\text{n10h1}_{\text{shallow}}$ (γ = 0.2), $\text{n15h1}_{\text{deep}}$ (γ = 0.3), $\text{n15h1}_{\text{intermediate}}$ (γ = 0.2), and $\text{n15h1}_{\text{shallow}}$ (γ = 0.2), viewable in Figures in the Supplementary Material. These single-hybridization networks allow us to measure the ability of methods to detect single hybridization events without the possible influence of overlapping hybridizations, and to compare the performance of the methods on shallow vs deep hybridizations, as it has been reported that deep hybridizations are more difficult to detect (Hibbins & Hahn, 2021). Note that here we use the term "overlapping hybridizations" not to refer to hybridizations that share edges (e.g. level-2 networks), but to hybridizations that affect the same set of taxa downstream. Finally, to represent how well methods perform at a larger scale, both in terms of computational efficiency and accuracy, we evaluate their performance on a larger network labeled n25h5 ($\gamma \in$ 0.024, 0.334, 0.396, 0.449, 0.395, in order top to bottom) .

We simulated gene trees under each network with the software `ms` (Hudson, 2002), with a single individual per taxon. Note that this approach reduces the power of HyDe, when compared to simulations with multiple individuals per population (Kong & Kubatko, 2021; Kubatko & Chifman, 2019). The software `ms` allows hybridization events to be modeled with -es t i p and -ej t i j events which correspond to population admixture and population splitting, respectively (Hudson, 2002). With this approach, we circumvent the alternative of decomposing each network into 2^r trees as in (Kong & Kubatko, 2021), where r is the number of reticulations, and sampling a proportion of each tree to represent the mixing parameter γ. We note that the gene tree distribution under a network is not equivalent as the gene tree distribution under 2^r displayed trees, unless there is only one taxon sampled beneath the hybrid node, and thus, directly modeling reticulation events from a network ensures that simulated gene trees follow their underlying network structure, which has a different probability density than the combination of individual trees, especially under complex reticulation events (Y. Yu et al., 2012). The decomposition to individual trees may produce reticulate gene tree patterns that are artificially clearer, and may not be as accurately representative of the timing of natural gene flow events.

For TICR and MSCquartets, we simulate gene trees for $g \in \{30, 100, 300, 1000, 3000\}$ to be used directly as input. For methods which require sequences, `ms` generates g unlinked gene trees where $g \in \{30, 100, 1000, 3000, 10000\}$ and from each gene tree, `seqgen` (Rambaut & Grass, 1997) is used to generate sequences with 100 base pairs generated per gene tree, similar to the approach used in (Kong & Kubatko, 2021) and then concatenated to form sequences of total length L. The `seqgen` parameters used to generate these base pairs are -mHKY -s0.036 -f0.300414,0.191363,0.196748,0.311475 -n1 -l100. Additionally, `IQ-TREE` was used to estimate gene trees from sequences with 100bp lengths using parameters -m HKY85 -s . The estimated gene trees are also used as input for TICR and MSCquartets. Full simulation details including `ms` commands and newick structures of these networks can be found on the GitHub repository https://github.com/mbjorner/hybrid-detection-comparison. Thirty trials were simulated for each combination of network and gene tree number or sequence length. A pipeline of this simulation is shown in Figure 3.

Note that the D-derived tests rely on pre-specification of topology and thus, we can expect increased false positives when testing on inputs of $(((t_1, t_3), t_2), O)$ when $(((t_1, t_2), t_3), O)$ is the true topology. For D_3, we use the uncorrected genetic distance, as all simulations have stationary mutation rates. In addition, HyDe relies on the existence of concurrent parental lineages to test for hybrid speciation. Where only one parental lineage is sampled (see network n10h2 for an example), we investigate the influence of introgression events from these "ghost" lineages on HyDe's output. To run MSCquartets, we chose to remove unresolved star trees, and use the T3 model, which represents an unspecified tree topology. Last, TICR requires an input estimated population tree to be used for the expected CFs. The estimated population tree that we use is the major tree from the input network. This is equivalent to the network with any minor hybrid edges removed. Deviations from the expected CFs could indicate deviations from the ILS-only model, but also, it could indicate that the wrong population tree, either topologically or metrically, was used for comparison. In our simulation studies, we use the known major tree as the input population tree for TICR so that any significant TICR results are interpreted with the possibility of hybridization.

Each method is also evaluated in terms of computing time, as measured in CPU time in seconds, given their gene tree or sequence data inputs, for the purpose of predicting how well each summary method accommodates the addition of sampled taxa. We note that we do not include the time to estimate gene trees in the running time, but for the number of taxa here tested, IQ-Tree is very fast. As genetic sequence information has become more widely available, so too have the datasets that biologists use to construct phylogenies and infer these reticulation events. In practice, often tens or hundreds of taxa are compared (Bossert et al., 2020; Suvorov et al., 2022). Since the Patterson's D-Statistic, D_3, and D_p were computed from the output of HyDe, which describes all possible site pattern frequencies, timing was omitted for D-statistic related tests.

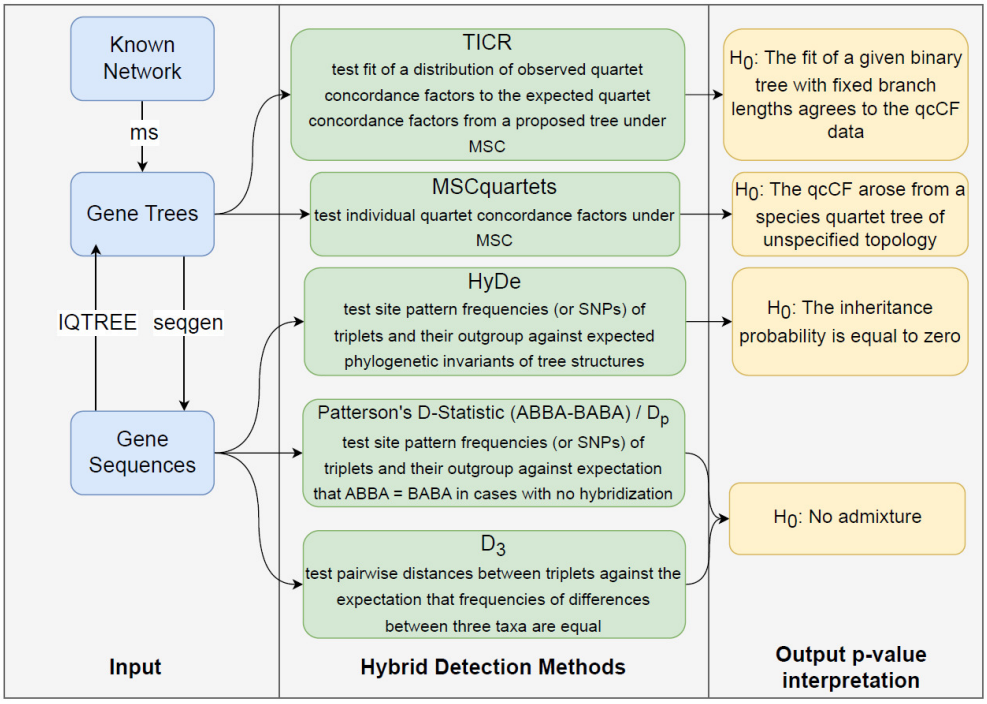

Figure 3. Simulation Pipeline. The software ms is used to simulate g gene trees from a known network where $g \in \{30, 100, 300, 1000, 3000\}$. These gene trees are used as input for the hybrid detection methods TICR and MSCquartets. Additionally, sequences of length L are generated using seqgen from g gene trees where $g \in \{30, 100, 300, 1000, 3000, 10000\}$. Each gene tree is used to generate sequences with 100 base pairs generated per gene tree as in (Kong & Kubatko, 2021). These sequences are used as input for HyDe, D_3, Patterson's D-Statistic and D_p. Additionally, these sequences were used as input for IQTREE, in order to transform them into estimated gene trees. The estimated gene trees were again used as input for TICR and MSCquartets. This process was repeated for each network structure in Figures in the Supplementary Material with 30 replicates each.

2.3.1 Evaluation of Accuracy on Simulated Datasets

We now describe the computation of the false positive/negative rates and precision (see also Figure 4). Every triple (or quartet for MSCquartets and TICR) could have a hybrid or not. For example, in the n4h1 network, the triple $\{1, 2, 3\}$ contains a hybrid (taxon 2), but the triple $\{1, 3, 4\}$ does not contain any hybrid. If the triple (quartet) contains a hybrid, and the method detects it (pvalue$< \alpha$ for significance level α), we consider this a true positive (TP). If the triple (quartet) contains a hybrid, but the method does not detect it (pvalue$> \alpha$), we consider this a false negative (FN). If the triple (quartet) does not contain a hybrid, and the method finds no hybrid (pvalue$> \alpha$), we consider this a true negative (TN). If the triple (quartet) does not contain a hybrid, but the method detects a hybrid (pvalue$< \alpha$), we consider this a false positive (FP). The False Positive Rate (FPR) is computed as $FP/(FP + TN)$. The recall, or sensitivity, is computed as $TP/(TP + FN)$. The precision is computed as $TP/(TP + FP)$. For HyDe, an additional metric, Wrong Hybrid Rate (WHR) describes the rate at which hybridization is detected but is falsely attributed to the incorrect hybrid taxon. That is, if the triple contains a hybrid, and HyDe detects it, but identifies the wrong taxon as the hybrid taxon, we consider this a wrong hybrid (WH). For example, in the n4h1 network, the triple $\{1, 2, 3\}$ contains a hybrid (taxon 2). HyDe could test whether 1 and 3 are the parents of hybrid taxon 2 (correct hybrid), or whether 2 and 3 are parents of hybrid taxon 1 (wrong hybrid). If the latter test is significant, then HyDe correctly identified that there is a hybrid relationship among these taxa, but wrongly identified the hybrid taxon. We define the Wrong Hybrid Rate (WHR) as $WHR = WH/(WH + TP + FP)$. We use `PhyloNetworks` (Solís-Lemus et al., 2017), a Julia package that allows for efficient manipulation of phylogenetic networks to easily identify triples or quartets with hybrid relationships in all networks under study. `PhyloNetworks` also allowed us to filter for only identifiable hybrids between non-sister lineages, as these patterns can be simplified to a tree structure. Any gene flow between sister taxa in the triplet or quartet is excluded from consideration as a hybrid in the results, due to their lack of detectability.

Figure 4. Visual description of false positives, false negatives, wrong hybrids (for HyDe only), true positives and true negatives.

2.4 Hybridizations in the bee subfamily Nomiinae

To demonstrate the use of these hybrid detection methods on real data, we compare method performance on a dataset of ultraconserved elements (UCEs) from the bee subfamily Nomiinae. This data originates from a paper investigating

the impacts of gene tree estimation error on species tree reconstruction (Bossert et al., 2020), and was used to demonstrate improved tree reconstruction with weighted ASTRAL (Zhang & Mirarab, 2022), a new version of ASTRAL that weights quartets based on their uncertainty (branch support) and terminal branch lengths in input gene trees. The dataset is available for download on https://datadryad.org/stash/dataset/doi:10.5061/dryad.z08kprrb6.

This dataset contains sequences and gene trees of up to 852 UCEs, for a total concatenated sequence length of 576,041 base pairs for each of 32 taxa. In the original study (Bossert et al., 2020), gene trees were estimated using six different methods, (1) IQ-Tree2 with the GTR-G substitution model, (2) IQ-Tree2 with the substitution model chosen by ModelFinder, (3) MrBayes with the GTR-G substitution model, (4) MrBayes with reversible jump MCMC, (5) PhyloBayes, and (6) RAxML. The original investigation found a consensus tree using PhyloBayes on concatenated UCEs.

Here, we use each of the proposed sets of gene trees created using the six different methods, as input for MSCquartets, and apply a Bonferroni correction to evaluate significant quartets which may contain hybridization. We use the proposed species tree and gene trees in combination for TICR, for which we interpret a poor fit of the observed qcCFs to either possibility of incorrect species tree, presence of hybridization, or a combination of the two. Next, we use the original UCE sequences and concatenate them to run HyDe, using *Lasioglossum albipes* as the outgroup, as indicated by the consensus tree constructed with PhyloBayes (Bossert et al., 2020), and a Bonferroni correction for significance. As the original study included two outgroups, *Lasioglossum albipes* and *Dufourea novaeangilae*, we removed *Dufourea novaeangilae* from all gene trees and sequences prior to running hybrid detection methods because these methods require only one outgroup.

3 Results

3.1 Simulations

Figure 5 shows the proportion of times that TICR correctly rejects the major tree from true and estimated gene trees, and thus, detects the presence of hybridizations under the different networks under study. We highlight that TICR accurately detects hybridizations for the case of single shallow hybridizations, $\text{n10h1}_{\text{shallow}}$ and $\text{n15h1}_{\text{shallow}}$. However, TICR does not detect deeper hybridizations as in $\text{n10h1}_{\text{deep}}$, $\text{n15h1}_{\text{intermediate}}$, and $\text{n15h1}_{\text{deep}}$ or multiple hybridizations in the same network as in n5h2, n8h3, n10h2, n15h3, and n25h5. TICR also does not detect hybridizations on networks with four taxa (as n4h1 or $\text{n4h1}_{\text{introgression}}$) and those results are not included in the figure. We highlight the decreased accuracy in performance when using estimated gene trees across all tested networks.

Figure 6 shows the false positive rate (yellow), precision (pink) and recall (gray) for MSCquartets (from true and estimated gene trees), HyDe, Patterson's D-Statistic, D_p, and D_3 on the networks: $\text{n4h1}_{\text{introgression}}$ (network with single shallow introgression event), n5h2 (network with two overlapping hybridization events), n8h3 (network with three overlapping hybridization events), and n25h5 (network with five overlapping hybridization events). As in (Kong & Kubatko, 2021), an overlapping hybridization event is defined as a hybridization where one taxon is the parent of multiple hybridization events. For HyDe, an additional metric, wrong hybrid rate (blue) describes the rate at which hybridization is detected but is falsely attributed to the incorrect hybrid taxon. The network $\text{n4h1}_{\text{introgression}}$ displays an introgression event which is easily detected by all methods (high precision and high recall). All methods also display no false positive rates on this network, as all triples or quartets tested contain a hybrid relationship. For the case of two hybridizations (n5h2), all methods display a high precision and high recall, except for HyDe which has a lower recall than others. False positive rate is low and comparable for all methods in this network. For three hybridizations (n8h3), all methods have high precision and lower recall. As more taxa become part of the network, certain combinations contain hybrids that arise from ghost lineages, which may not have strong signal to detect the hybridization events. In this figure, all test are Bonferroni-corrected at a level of significance $\alpha = 0.05/\text{number of tests}$, but we also show the uncorrected version ($\alpha = 0.05$) in Figures in the Supplementary Material.

Figure 7 also shows the results on the largest network under study (n25h5). Again, all methods show a low recall and low false positive rate both of which could be explain by a weakening of the hybridization signal when multiple hybridizations are affecting the same taxa. All methods have a high precision which means that when a hybrid is detected, it is very likely a true hybrid. HyDe has slightly lower precision compared to other methods, but this is due to the fact that HyDe (unlike other methods) test for a very specific parent-hybrid relationship. When HyDe is tested in the setup of clear parent-hybrid relationships (Figures in the Supplementary Material), HyDe indeed displays high precision. It is notable that HyDe's precision is better for n25h5 compared to n15h3 or n10h2. This is due to the fact that the hybridizations in n10h2 and n15h3 involve ghost lineages which is not accounted for in HyDe. In this figure, all test are Bonferroni-corrected at a level of significance $\alpha = 0.05/\text{number of tests}$, but we also show the uncorrected version ($\alpha = 0.05$) in Figure 8 in the Supplementary Material.

Figure 7 shows the results for the networks: n10h2 (network with two hybridization events), $\text{n10h1}_{\text{shallow}}$ (network with single shallow hybridization event) and $\text{n10h1}_{\text{deep}}$ (network with a single deep hybridization event). All methods report a lower recall compared to the simpler networks (n4h1 and n5h2), although precision continues to be high for all methods, except for HyDe. HyDe's lower precision is due to the fact that some hybridizations involve ghost lineages (hybridizations when one or both parental lineages contributing to the hybrid node are extinct or unsampled) and HyDe cannot account for this scenario. False positive rate is controlled in all methods. This combined with the lower recall allows us to conclude that multiple overlapping

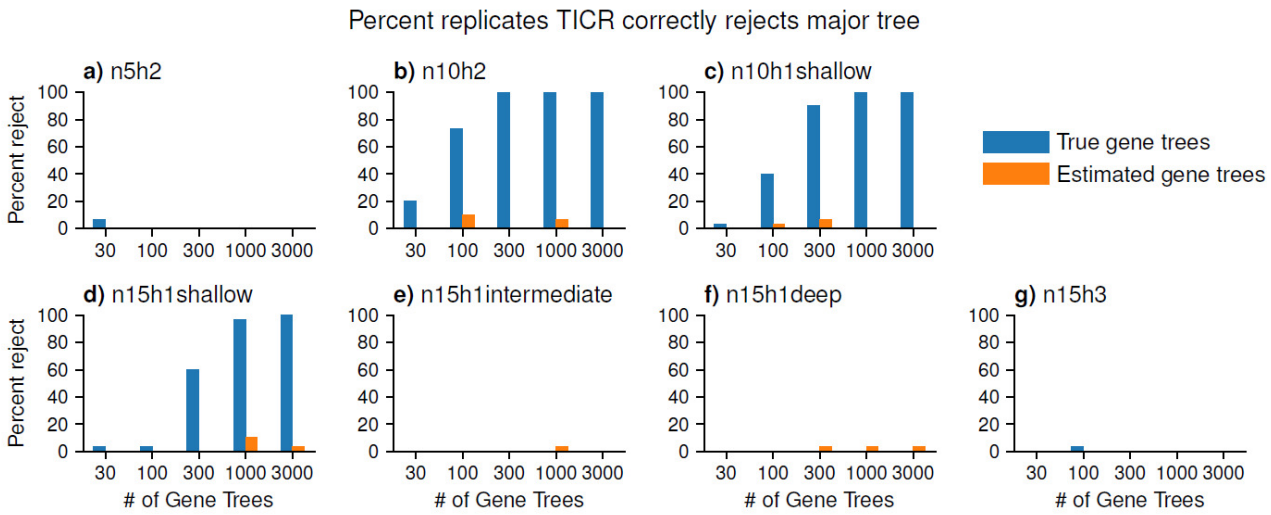

Figure 5. Y axis corresponds to the proportion of replicates in which TICR correctly rejects the true major tree at a level of 0.05 and X axis corresponds to the number of gene trees. With true gene trees as input (blue bars), TICR rejects the null hypothesis of the major tree (equivalent to the input network sans hybridization events) for networks with single, shallow hybridizations. TICR fails to detect deeper hybridizations such as $\text{n10h1}_{\text{deep}}$, $\text{n15h1}_{\text{intermediate}}$, and $\text{n15h1}_{\text{deep}}$. For networks with multiple hybridization events (n5h2, n8h3, n10h2, n15h3, and n25h5), TICR is also unable to reliably detect non-treelike patterns. With estimated gene trees as input (orange bars), TICR is unable to reject the major tree across tested networks.

hybridizations result in loss of signal for hybridization, rather than contradicting signal pointing at wrong hybrids. In addition, the recall is low across all methods for the single deep hybridization case ($\text{n10h1}_{\text{deep}}$) which means that it is not only multiple hybridizations that result in loss of signal, hybridizations occurring in deeper parts of the tree have also lost signal to be detected. We also note that unlike previous cases (e.g., Figure 6) where HyDe's wrong hybrid rate (blue) and false positive rate (yellow) were overlapped, for these cases, the wrong hybrid rate is much higher than the false positive rate. This implies that HyDe is better able to identify hybrid relationships for these networks, but not the correct hybrid taxon. In this figure, all test are Bonferroni-corrected at a level of significance $\alpha = 0.05/\text{number of tests}$, but we also show the uncorrected version ($\alpha = 0.05$) in Figure 9 in the Supplementary Material. Results for network n4h1 are also in the Supplementary Material. This network is among the simplest cases with a single shallow hybridization, and thus, all methods have a high precision, low false positive rate, and high recall for as little as 100,000 sites or 300 gene trees.

Figure 8 shows the results for the networks: n15h3 (network with three hybridization events), $\text{n15h1}_{\text{shallow}}$ (network with a single shallow hybridization event), $\text{n15h1}_{\text{intermediate}}$ (network with a single intermediate hybridization event), and $\text{n15h1}_{\text{deep}}$ (network with a single deep hybridization event). As already shown in the case of n10h2 (Figure 7), all methods have a higher false negative rate, but controlled false positive rate except for Patterson's D-statistic with a high false positive rate for the case of three hybridizations (n15h3). HyDe shows lower precision compared to other methods which is due to the fact that the hybridizations in n15h3 involve ghost lineages which HyDe cannot account for. The fact that there is low recall even for the single shallow hybridization ($\text{n15h1}_{\text{shallow}}$) across of methods could provide some evidence that ghost lineages create challenges, not just for HyDe. In this figure, all test are Bonferroni-corrected at a level of significance $\alpha = 0.05/\text{number of tests}$, but we also show the uncorrected version ($\alpha = 0.05$) in Figure 10 in the Supplementary Material.

3.2 Empirical running time

Each method is predicted to increase linearly with respect to the number of gene trees or sequence length used as input. Due to the nature of summary methods' triple- or quartet-wise analysis, an increase in network size corresponds to a cubic or quartic increase in time, respectively, and indeed, the running time of methods (HyDe, MSCquartets, and TICR) dramatically increase with the number of taxa and the number of gene trees (Figure 9). It is worth noting that time complexity does not account for many practical issues, like memory locality and cache performance, that greatly impact runtime in practice.

3.3 Nomiinae bee subfamily

Figure 10 displays the heatmap of the proportion of times that each taxon is involved in a significant hybridization event as identified by MSCquartets, HyDe, Patterson's D-statistics and D_3. This figure was created using `ggtree` (G. Yu et al., 2017).

We selected the estimated gene trees from the IQ-Tree2-GTRG model to use as input in MSCquartets and dis-

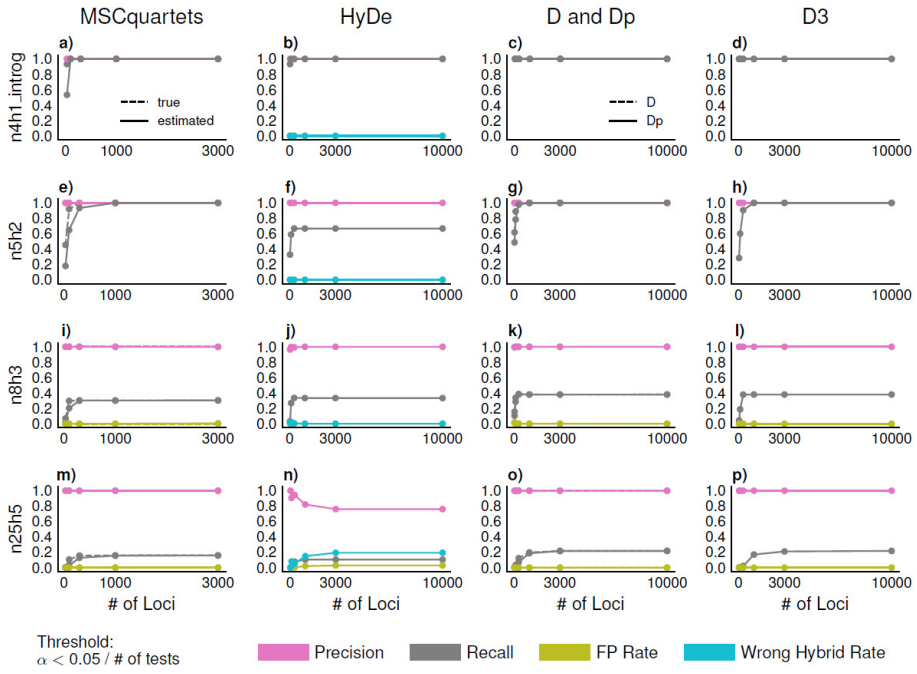

Figure 6. Precision, recall, and false positive rate for MSCquartets (from true and estimated gene trees), HyDe, Patterson's D-Statistic, D_p, and D_3 on the networks: n4h1introgression, n5h2, n8h3, and n25h5. All tests are Bonferroni-corrected at a level of significance $\alpha = 0.05 \div$ number of tests (see the Supplementary Material for results on α = 0.05). For HyDe, an additional metric, Wrong Hybrid Rate (blue) describes the rate at which hybridization is detected but is falsely attributed to the incorrect hybrid taxon. X axis corresponds to the number of loci. For the introgression event ($\mathrm{n4h1}_{\mathrm{introgression}}$), MSCquartets, HyDe, Patterson's D-Statistic, and D3 all behave similarly to the hybrid speciation scenario in n4h1, with near-perfect recovery and identification of hybrid scenarios. High recall is noted for n5h2 across MSCquartets and D-related methods. For n5h2, HyDe correctly recovers parent-hybrid relationships, but has a slightly decreased recall. Network n8h3 displays low recall across all methods. However, precision is high, and false positive rate across all methods are comparable. For n25h5, high precision is noted for all tests.

play the proportion of times that each taxa is involved in a significant hybridization (Figure 10) for the Bonferroni-corrected significance level of $0.05/14950$, or 3.3×10^6 and Figure 11 in the Supplementary Material for significance level of $\alpha = 0.05$) *Stictonomia spp.* is implicated in all significant quartets, with *Stictonomia schubotzi* appearing with the highest frequency. In addition, HyDe detects *Stictonomia spp.* implicated 923 times over 2851 significant hybrid speciation events. We also show the proportion that each taxon is identified as a parent (ancestral lineage contributing genetic material to the hybrid) in the HyDe tests which provides a broader picture of hybridization than any of the other methods. The results of Patterson's D-statistics or D_3 align with those of HyDe and MSCquartets in the identification of *Stictonomia spp.* (especially *Stictonomia schubotzi*) as involved in hybridization events. The D-related tests, however, cannot separate the hybrid taxon from the parents as HyDe. Though MSCquartets primarily implicates *Stictonomia spp.* in hybridization, other methods find widespread hybridization across the tree. These differences may be due to methods' sensitivity to the depth of hybridization events.

These results suggest that these closely related species may not be reproductively isolated, which can lead to gene tree estimation error, and difficulty in reconstructing the phylogenetic tree. In the original study (Bossert et al., 2020), gene tree estimation error was identified as a source of the discordance and conflict. However, here we identify hybridization events as a plausible explanation for the gene tree discordance.

4 Discussion

Here, we present a deep investigation of the performance of genome-wide hybrid detection methods. We found that all five methods compared (TICR (Stenz et al., 2015), MSCquartets (Mitchell et al., 2019), HyDe (Kubatko & Chifman, 2019), Patterson's D-Statistic (Patterson et al., 2012), D_p (Hamlin et al., 2020) and D_3 (Hahn & Hibbins, 2019)) have similar good performance (i.e., high precision and low false positive/negative rates) on single shallow hybridizations involving few taxa (n4h1 or n5h2). Our investigation confirms previous findings (Kong & Kubatko, 2021), and extends the conclusions to previously untested scenarios.

By design, both MSCQuartets and TICR should also be able to detect complex hybridization of more than one instance of gene flow among four taxa by relying on the rejection of a tree hypothesis. However, as more hybridizations are added involving similar groups of taxa (n8h3, n10h2 and n15h3), all methods have a higher false negative

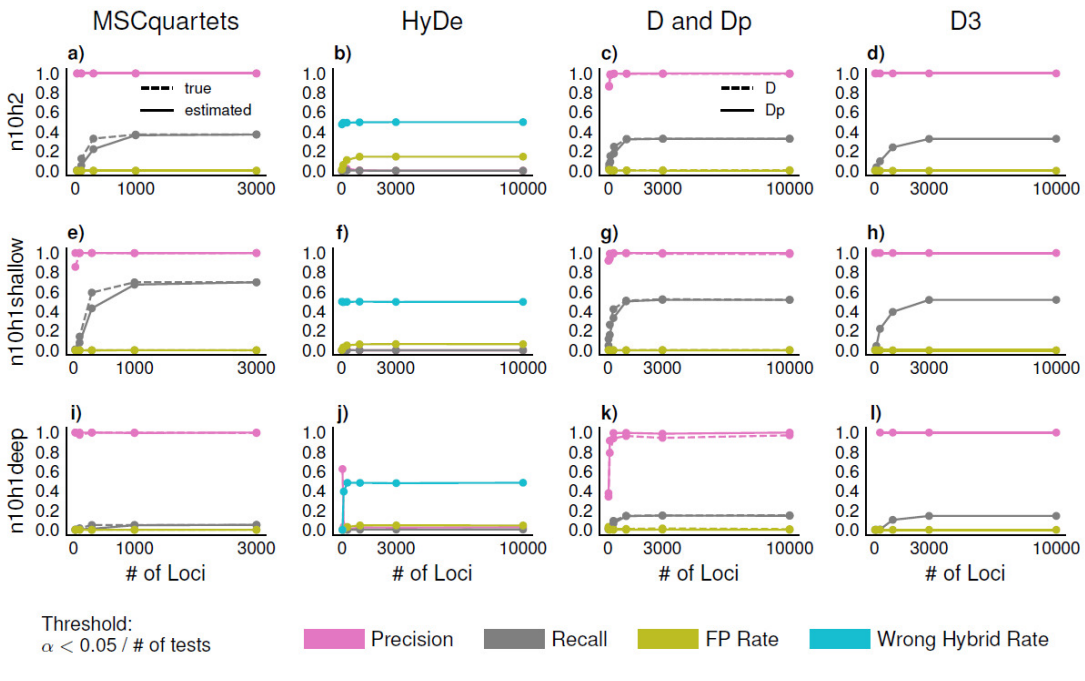

Figure 7. False positive rate (yellow), precision (pink) and recall (gray) for MSCquartets (from true and estimated gene trees), HyDe, Patterson's D-Statistic, and D_3 on the networks: n10h2, $\text{n10h1}_{\text{shallow}}$ (n10h2 with only the shallow hybridization event), $\text{n10h11}_{\text{deep}}$ (n10h2 with only the deeper hybridization event). All tests are Bonferroni-corrected at a level of significance $\alpha = 0.05$ number of tests (see the Supplementary Material for results on $\alpha = 0.05$). X axis corresponds to the number of loci. D_3, Patterson's D-Statistic, and MSCquartets behave comparably across network structures, with high precision, low false positive rate, but lower recall, which is seen in the deepest hybridization ($\text{n10h1}_{\text{deep}}$). A shallower hybridization that creates a single hybrid taxon ($\text{n10h1}_{\text{shallow}}$) is readily detected by these tests, but still with low recall. HyDe has the lowest recall with the $\text{n10h1}_{\text{deep}}$ hybridization event. While HyDe correctly detects hybridization in triples, it often misattributes the hybridization to the wrong parent or child. When both hybridization events are present, as in n10h2, all methods result in low recall with HyDe also having a higher FP rate.

rate which suggests that combinations of gene flow events weaken the signal to detect such hybridizations as opposed to creating discordant signal to identify wrong hybridizations (which would have been evidenced by an increased false positive rate). This is also confirmed by the results of TICR which is *unable* to reject the major tree in most cases, except for those involving single shallow hybridizations even if they involve ghost lineages. This finding for true gene trees did not hold for estimated gene trees , as TICR was rarely able to reject the major tree when given estimated gene trees, even for this easier model condition.

HyDe had a lower precision that other methods when ghost lineages were involved (n10h2 and n15h3) which aligns with previous studies on the subject (Pang & Zhang, 2022; Tricou et al., 2022). It also showed a higher rate of wrong hybrid identified within the hybrid triple. However, HyDe is the only method able to detect the hybrid taxon and the parent taxa involved in the hybridization as shown in the bee dataset. The results of methods using site pattern frequencies or pairwise differences is highly influenced by the topology of the underlying species tree from which taxa arise. Longer coalescent times introduce noise to sequence data such that comparison to a distant outgroup or comparison between distant species is no longer advantageous, as the infinite-sites mutation model on which Patterson's D-Statistic is based expects a single mutation per site (Hibbins & Hahn, 2021). With increased branch lengths, or increased distance between taxa, convergent substitutions can cause ABBA and BABA (and other) site patterns to accumulate (Hibbins & Hahn, 2021). Similarly, HyDe was derived under Jukes-Cantor model, and while previous simulations (Kubatko & Chifman, 2019) showed some level of robustness to model misspecification, the fact that we do not simulate sequences under Jukes-Cantor could also explain HyDe's poor performance.

Finally, we re-analyzed the dataset of the bee subfamily Nomiinae. While the original study (Bossert et al., 2020) concludes that gene tree estimation error could be the source of discordance in the clade, here we show that hybridization is another plausible explanation for the discordant patterns with all methods identifying *Stictonomia spp.* (especially *Stictonomia schubotzi*) as involved in hybridization events.

Practical advice for evolutionary biologists. From our investigation, we can conclude that MSCquartets (Allman & Rhodes, 2007) is an accurate method to detect hybridization events under a variety of different scenarios. HyDe is the only method that can identify which taxon is the hybrid taxon among the taxa involved in the hybridization event. However, HyDe cannot perform well when the parents of

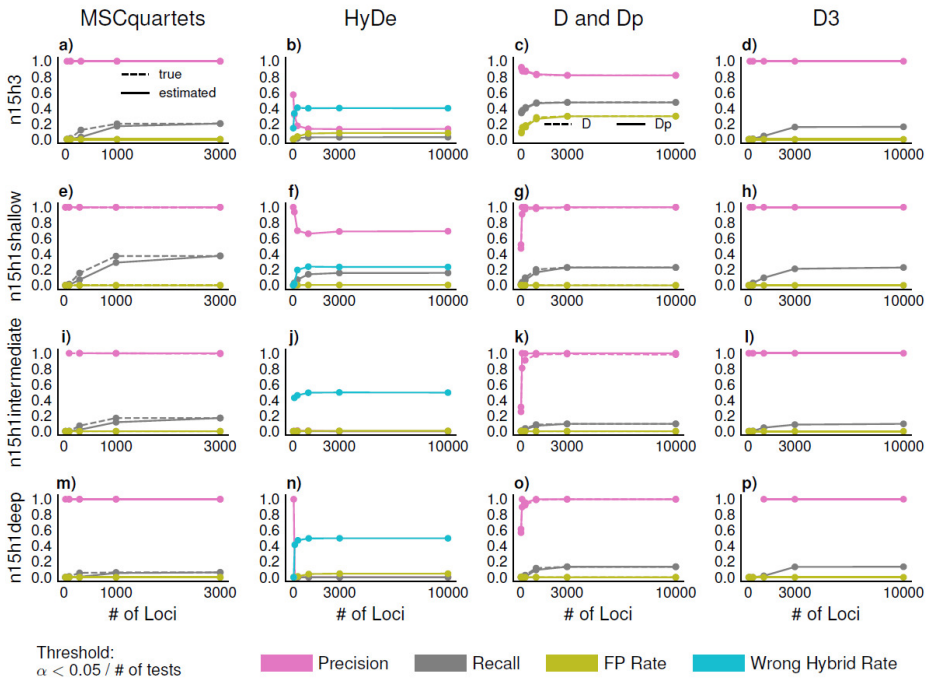

Figure 8. False positive rate (yellow), precision (pink) and recall (gray) for MSCquartets (from true and estimated gene trees), HyDe, Patterson's D-Statistic, and D_3 on the networks: $\mathrm{n15H3}, \mathrm{n15h1}_{\mathrm{deep}}$ ($n15h3$ with only the deepest hybridization event), $\mathrm{n15h}_{\mathrm{shallow}}$ ($n15h3$ with only the shallowest hybridization event), and $\mathrm{n15h1}_{\mathrm{intermediate}}$ ($n15h3$ with a hybridization event of intermediate depth). All tests are Bonferroni-corrected at a level of significance $\alpha = 0.05/$ number of tests (see Figure 10 in the Supplementary Material for results on $\alpha = 0.05$). X axis corresponds to number of loci. MSCquartets results in lower recall given deeper hybridization events, with a stable, high precision rate given more than 30 gene trees as input. All methods have a controlled false positive rate, except Patterson's D-Statistic with n15h3. HyDe has a similar or lower recall when compared to other tests, and again misidentifies the exact hybrid-parent relationship in which there is a hybridization present. However, in the cases of single hybridizations, its false positive rate remains low. For all single hybridization events, Patterson's D-Statistic and D_3 perform comparably, with the same pattern in recall given hybridization depth. However, in the case of n15h3, D_3 outperforms Patterson's D-Statistic in terms of its higher precision and lower false positive rate. In this case, Patterson's D-Statistic has a very high false positive rate, of approximately 25%.

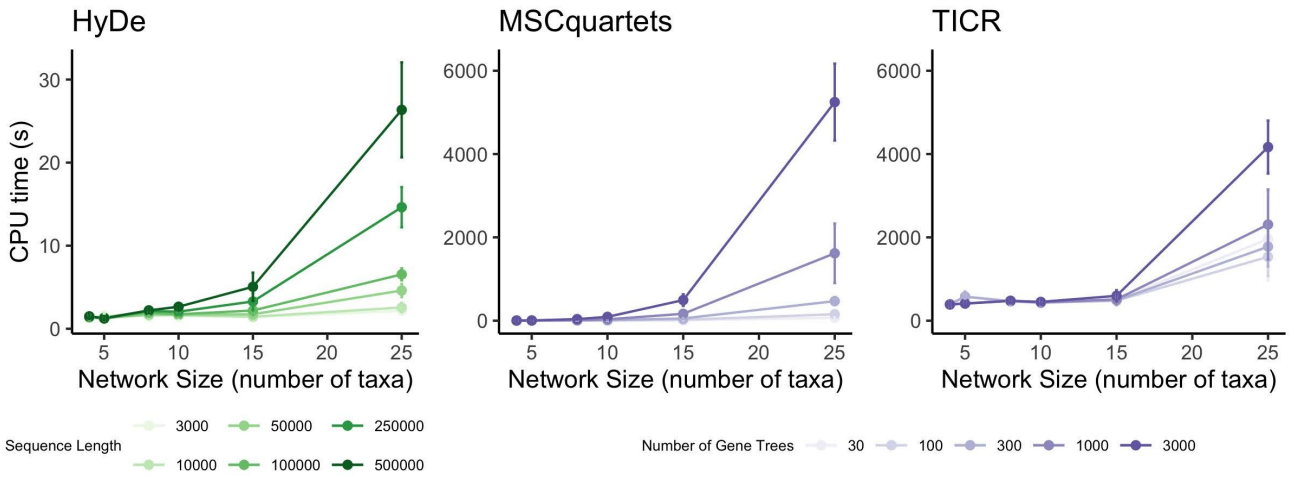

Figure 9. CPU time of HyDe (left), MSCquartets (center), and TICR (right) as network size and sequence length or number of gene trees change. Note the different limits on the Y axis for HyDe, which is the fastest of all three methods. For HyDe, a cubic increase in time is observed with respect to network size. For MSCquartets, a quartic increase in time is observed with respect to network size.

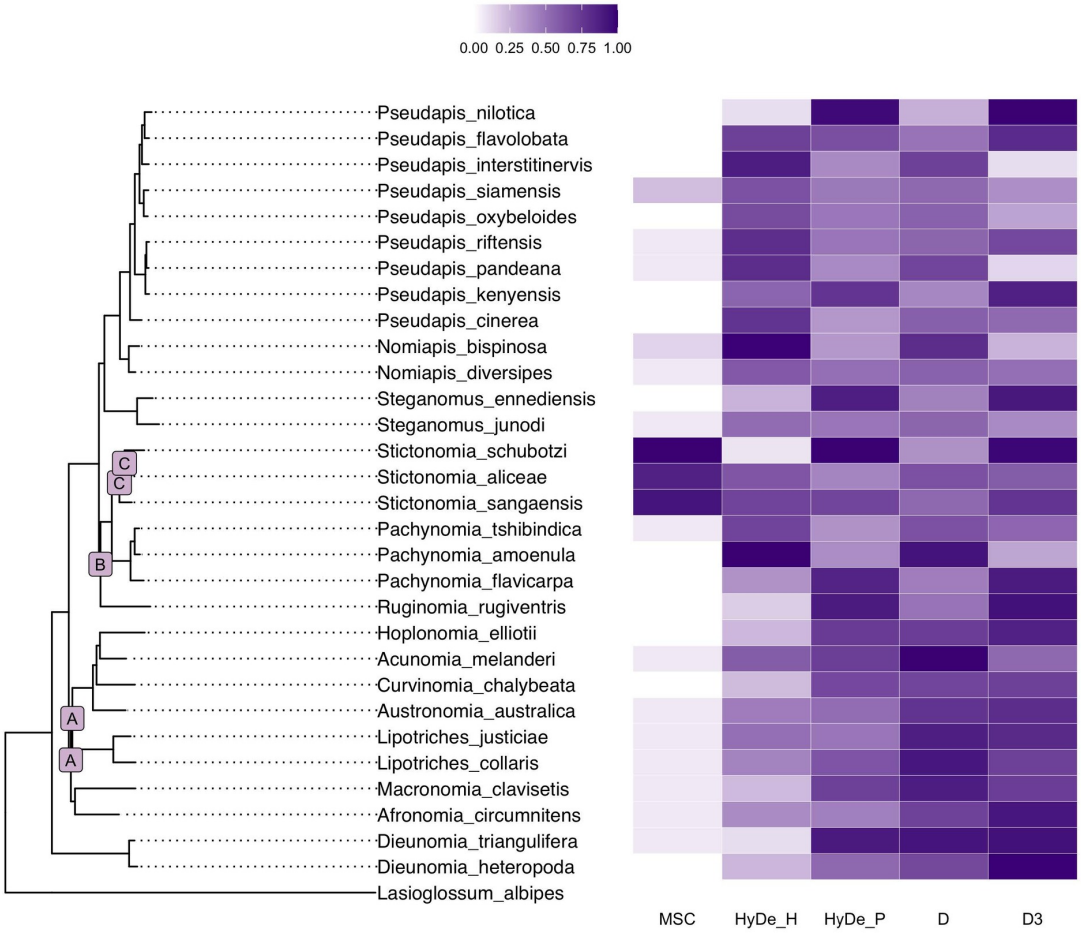

Figure 10. Left: Phylogenetic tree of selected species from the bee subfamily Nomiinae, as provided by (Bossert et al., 2020). Letters A-C identify nodes that were identified as conflicting relationships by the analyses conducted in (Bossert et al., 2020). Right: Heatmap of frequency of taxa identified as part of hybridization events (scaled between 0 and 1 for comparison across methods). MSCquartets was conducted with gene trees estimated using the IQTree2-GTRG model. Tests are Bonferroni-corrected (see Figure 11 in the Supplementary Material for significance level $\alpha = 0.05$). The clade with the highest proportion of proposed hybridizations contains *Stictonomia spp*.

hybridization are unsampled or extinct. Furthermore, all methods are unable to detect hybridizations when multiple events are affecting the same set of taxa or when hybridizations are deep. That is, when the hybridization event is close to the root. In this situation, we recommend testing a smaller sample of taxa suspected to be involved a single hybridization event, instead of testing all taxa at once. When a given parent-hybrid relationship is to be tested, HyDe outperforms D-Statistics-like tests by allowing the identification of hybrid taxa vs parent taxa. Finally, we conclude that TICR is a powerful method to detect single shallow hybridization events, even if they involve ghost lineages provided that gene trees can be accurately estimated.

Limitations and future work. All the networks in the simulation study are ultrametric in coalescent units, which implicitly assumes equal population sizes across lineages. While this assumption is unrealistic, it is convenient to disentangle the causes that create differences across methods. A more thorough investigation of the interaction between population structure and hybridization patterns is needed. Along these lines, multiple sequence alignment errors can occur in phylogenomics data sets (Zhang et al., 2021) and could impact the performance of all methods tested. We simulated data under a substitution-only model and thus all methods were given true alignments as input. Lastly, the substitution model is well behaved (homogeneous, stationary, and reversible), which is assumed by HyDe; however, these assumptions can be violated in practice (Naser-Khdour et al., 2019). All sequences generated from gene trees were the same length, which does not reflect real-life variation in gene length. Differences in performance based on substitution model were also not explored here. We point at a recent manuscript that explores the effect of rate variation on the performance of introgression tests (Frankel & Ané, 2023). Lastly, our evaluation of TICR allowed it to use the major tree derived from the true network, rather than an estimated species tree. It is not clear to what extent these practicalities will impact methods, and future work should explore them in simulations and in real data sets.

Data Availability

Scripts for generating simulated gene trees and conducting data analysis and plotting are available at https://github.com/mbjorner/hybrid-detection-comparison. Genetic data on the bee subfamily Nomiinae was collected from Bossert et. al.'s article (Bossert et al., 2020).

Supplementary materials are available at https://doi.org/10.5281/zenodo.13350781.

Acknowledgements

This work was supported by the Department of Energy [DE-SC0021016 to CSL] and by the National Science Foundation [DEB-2144367 to CSL]. We thank Laura Kubatko and Sungsik Kong for meaningful discussions about HyDe.

Submitted: October 31, 2023 EDT, Accepted: April 16, 2024 EDT

References

Allman, E. S., Baños, H., & 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. https://doi.org/10.1186/s13015-019-0159-2

Allman, E. S., Degnan, J. H., & 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. https://doi.org/10.1007/s00285-010-0355-7

Allman, E. S., Mitchell, J. D., & Rhodes, J. A. (2021). Gene tree discord, simplex plots, and statistical tests under the coalescent. *Systematic Biology*, *71*(4), 929–942. https://doi.org/10.1093/sysbio/syab008

Allman, E. S., & Rhodes, J. A. (2007). Phlogenetic invariants. In *Reconstructing evolution: New mathematical and computational advances* (pp. 108–146). Oxford University Press.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, *16*, 113–148. http://www.jstor.org/stable/2097045

Blischak, P. D., Chifman, J., Wolfe, A. D., & Kubatko, L. S. (2018). HyDe: A python package for genome-scale hybridization detection. *Systematic Biology*, *67*(5), 821–829. https://doi.org/10.1093/sysbio/syy023

Bossert, S., Murray, E. A., Pauly, A., Chernyshov, K., Brady, S. G., & Danforth, B. N. (2020). Gene tree estimation error with ultraconserved elements: An empirical study on pseudapis bees. *Systematic Biology*, *70*(4), 803–821. https://doi.org/10.1093/sysbio/syaa097

Bouckaert, T. G. A. B.-S., Remco AND Vaughan. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, *15*(4), 1–28. https://doi.org/10.1371/journal.pcbi.1006650

Cai, R., & Ané, C. (2020). Assessing the fit of the multi-species network coalescent to multi-locus data. *Bioinformatics*, *37*(5), 634–641. https://doi.org/10.1093/bioinformatics/btaa863

Degnan, J. H. (2013). Anomalous Unrooted Gene Trees. *Systematic Biology*, *62*(4), 574–590. https://doi.org/10.1093/sysbio/syt023

Frankel, L. E., & Ané, C. (2023). Summary tests of introgression are highly sensitive to rate variation across lineages. *bioRxiv*. https://doi.org/10.1101/2023.01.26.525396

Hahn, M. W., & Hibbins, M. S. (2019). A three-sample test for introgression. *Mol Biol Evol*, *36*(12), 2878–2882. https://doi.org/10.1093/molbev/msz178

Hamlin, J. A. P., Hibbins, M. S., & Moyle, L. C. (2020). Assessing biological factors affecting postspeciation introgression. *Evol Lett*, *4*(2), 137–154. https://doi.org/10.1002/evl3.159

Hibbins, M. S., & Hahn, M. W. (2021). Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, *220*(2). https://doi.org/10.1093/genetics/iyab173

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation . *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Kong, S., & Kubatko, L. S. (2021). Comparative performance of popular methods for hybrid detection using genomic data. *Systematic Biology*, *70*(5), 891–907. https://doi.org/10.1093/sysbio/syaa092

Kubatko, L. S., & Chifman, J. (2019). An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evolutionary Biology*, *19*(1), 112. https://doi.org/10.1186/s12862-019-1439-7

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, *46*(3), 523–536. https://doi.org/10.1093/sysbio/46.3.523

Markin, A., Wagle, S., Anderson, T. K., & Eulenstein, O. (2022). RF-Net 2: fast inference of virus reassortment and hybridization networks. *Bioinformatics*, *38*(8), 2144–2152. https://doi.org/10.1093/bioinformatics/btac075

Mitchell, J. D., Allman, E. S., & Rhodes, J. A. (2019). Hypothesis testing near singularities and boundaries. *Electron J Stat*, *13*(1), 2150–2193. https://doi.org/10.1214/19-ejs1576

Moret, B. M. E., Nakhleh, L., Warnow, T., Linder, C. R., Tholse, A., Padolina, A., Sun, J., & Timme, R. (2004). Phylogenetic networks: Modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *1*(1), 13–23. https://doi.org/10.1109/TCBB.2004.10

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., & Lanfear, R. (2019). The prevalence and impact of model violations in phylogenetic analysis. *Genome Biology and Evolution*, *11*(12), 3341–3352. https://doi.org/10.1093/gbe/evz193

Nguyen, L.-T., Schmidt, H. A., Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Pang, X.-X., & Zhang, D.-Y. (2022). Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. *Systematic Biology*. https://doi.org/10.1093/sysbio/syac047

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nat Rev Genet*, *16*(6), 359–371. https://doi.org/10.1038/nrg3936

Rambaut, A., & Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235–238. https://doi.org/10.1093/bioinformatics/13.3.235

Rannala, B., & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, *164*(4), 1645–1656. https://doi.org/10.1093/genetics/164.4.1645

Rhodes, J. A., Baños, H., Mitchell, J. D., & Allman, E. S. (2021). MSCquartets 1.0: Quartet methods for species trees and networks under the multispecies coalescent model in R. *Bioinformatics*, *37*(12), 1766–1768. https://doi.org/10.1093/bioinformatics/btaa868

Snir, S., & Rao, S. (2012). Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution*, *62*(1), 1–8. https://doi.org/10.1016/j.ympev.2011.06.021

Solís-Lemus, C., & Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, *12*(3), e1005896. https://doi.org/10.1371/journal.pgen.1005896

Solís-Lemus, C., Bastide, P., & Ané, C. (2017). PhyloNetworks: A package for phylogenetic networks. *Molecular Biology and Evolution*, *34*(12), 3292–3298. https://doi.org/10.1093/molbev/msx235

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stenz, N. W. M., Larget, B., Baum, D. A., & Ané, C. (2015). Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species Arabidopsis thaliana (L.) Heynh. *Systematic Biology*, *64*(5), 809–823. https://doi.org/10.1093/sysbio/syv039

Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'Agostino, E. R. R., Price, D. K., Waddell, P. J., Lang, M., Courtier-Orgogozo, V., David, J. R., Petrov, D., Matute, D. R., Schrider, D. R., & Comeault, A. A. (2022). Widespread introgression across a phylogeny of 155 Drosophila genomes. *Current Biology*, *32*(1), 111-123.e5. https://doi.org/10.1016/j.cub.2021.10.052

Tricou, T., Tannier, E., & Vienne, D. M. de. (2022). Ghost lineages highly influence the interpretation of introgression tests. *Systematic Biology*. https://doi.org/10.1093/sysbio/syac011

Wen, D., & Nakhleh, L. (2017). Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, *67*(3), 439–457. https://doi.org/10.1093/sysbio/syx085

Wen, D., Yu, Y., Zhu, J., & Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, *67*(4), 735–740.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628

Yu, Y., Degnan, J. H., & Nakhleh, L. (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLOS Genetics*, *8*(4), 1–10. https://doi.org/10.1371/journal.pgen.1002660

Zhang, C., & Mirarab, S. (2022). Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *bioRxiv*. https://doi.org/10.1101/2022.02.19.481132

Zhang, C., Ogilvie, H. A., Drummond, A. J., & Stadler, T. (2017). Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, *35*(2), 504–517. https://doi.org/10.1093/molbev/msx307

Zhang, C., Zhao, Y., Braun, E. L., & Mirarab, S. (2021). TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods in Ecology and Evolution*, *12*(11), 2145–2158. https://doi.org/10.1111/2041-210X.13696