Hibbins, M. S., & Hahn, M. W. (2024). Distinguishing between histories of speciation and introgression using genomic data. *Bulletin of the Society of Systematic Biologists*. https://doi.org/10.18061/bssb.v3i1.9227

Investigations

Distinguishing Between Histories of Speciation and Introgression Using Genomic Data

Mark S. Hibbins[1], Matthew W. Hahn[2,3]

[1] Department of Ecology and Evolutionary Biology, University of Toronto, [2] Department of Biology , Indiana University, [3] Department of Computer Science, Indiana University

Keywords: introgression, speciation, supervised machine learning

https://doi.org/10.18061/bssb.v3i1.9227

Bulletin of the Society of Systematic Biologists

Abstract

Introgression creates complex, non-bifurcating relationships among species. At individual loci and across the genome, both introgression and incomplete lineage sorting interact to produce a wide range of different gene tree topologies. These processes can obscure the history of speciation among lineages, and, as a result, identifying the history of speciation vs. introgression remains a challenge. Here, we use theory and simulation to investigate how introgression can mislead multiple approaches to species tree inference. We find that arbitrarily low amounts of introgression may potentially mislead both gene tree and parsimony approaches to species tree inference if the level of incomplete lineage sorting is sufficiently high. We also show that an alternative approach based on minimum gene tree node heights is inconsistent and depends on the rate of introgression across the genome. To distinguish between speciation and introgression, we apply supervised machine learning models to a set of features that can easily be obtained from phylogenomic datasets. We find that multiple of these models are highly accurate in classifying the species history in simulated datasets. We also show that, if the histories of speciation and introgression can be identified, PhyloNet will return highly accurate estimates of the contribution of each history to the data (i.e. edge weights). Overall, our results highlight the promise of supervised machine learning as a potentially powerful complement to phylogenetic methods in the analysis of introgression from genomic data.

Introduction

Introgression, the process of hybridization and repeated back-crossing between previously isolated lineages, occurs frequently across the tree of life and is a common feature of modern phylogenomic datasets (Mallet et al., 2016; Taylor & Larson, 2019, Dagilis et al. 2021). From a phylogenetic perspective, histories of introgression are not consistent with a strictly bifurcating phylogeny and are therefore often represented using phylogenetic networks (Huson et al., 2010; Huson & Bryant, 2006; Solís-Lemus & Ané, 2016; Wen et al., 2018; Wen & Nakhleh, 2018). Phylogenetic networks contain additional horizontal "reticulation" edges, which are meant to display alternative histories among loci. Such networks imply that some parts of the genome follow the speciation history, while other parts follow the introgression history—a history that includes exchange between species post-speciation. Speciation and introgression often co-occur and can influence each other through processes such as reinforcement. Nonetheless, for these and other reticulate evolutionary processes, such as horizontal gene transfer and allopolyploidization, the presence of competing histories in the genome presents both conceptual and methodological challenges for inferring the history of speciation among lineages (Eckert & Carstens, 2008; Philippe & Douady, 2003, Thomas et al. 2017).

In phylogenomic studies, it is common to estimate a species tree using standard methods and then to interpret the results of introgression analyses using that species tree. Many popular methods for inferring introgression, such as the *D* statistic (Green et al., 2010), require a species tree to be specified *a priori*. Phylogenetic network methods co-estimate both histories, but they cannot explicitly label which history arose from speciation vs. introgression: they only infer that both histories exist in the data. Networks are often estimated with a previously constructed species tree in mind (often referred to as the "major" or "backbone" tree); this can be because the method requires a species tree be specified, or because the user has already estimated or assumed a bifurcating species tree before the network method is applied. Consequently, this pre-specified history is commonly assigned as the history of speciation implied by a

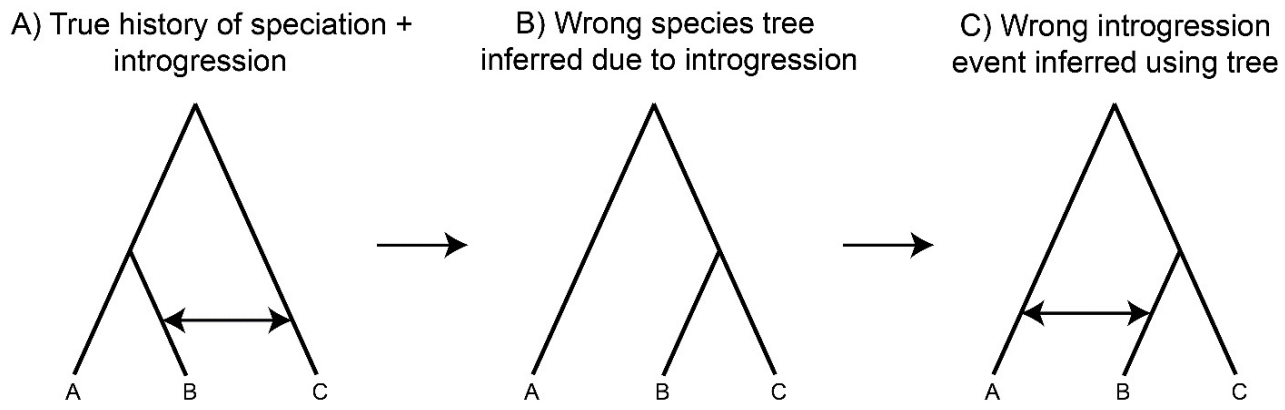

Figure 1. Mis-specifying the species tree affects downstream introgression analyses.

Sufficient introgression between species B and C (panel A) may cause the species tree to be incorrectly inferred, with species B and C becoming the most closely related (panel B). If this incorrect species tree is used for introgression analyses, introgression could erroneously be inferred between species A and B (panel C).

network. It is also common for software to assign the history of speciation to the network edge with the higher inheritance probability (e.g. Pickrell & Pritchard, 2012), under the assumption that the majority of the genome follows the speciation history. This approach to studying introgression typically ignores problems of uncertainty in species tree estimation, leading to erroneous inferences if the species tree is mis-estimated.

What makes this problem especially challenging is that introgression itself is a known source of species tree estimation error (Eckert & Carstens, 2008; Leaché et al., 2014; Long & Kubatko, 2018; Pang & Zhang, 2023; Solís-Lemus et al., 2016). Furthermore, several studies have now shown that introgression can be extensive enough to affect a majority of the genome (Fontaine et al., 2015; Forsythe et al., 2020; Li et al., 2019), so choosing the majority edge of a phylogenetic network as the species history can also lead to incorrect inferences. Consider, for example, a history of speciation of ((A,B),C), with post-speciation introgression between lineages B and C (Figure 1A). Sufficient levels of introgression between B and C could cause a species tree of ((B,C),A) to be erroneously inferred (Figure 1B). Subsequent application of this tree to introgression analyses would result in the incorrect inference of introgression between species A and B (Figure 1C). Distinguishing between the scenarios shown in Figures 1A and 1C remains a challenge in empirical datasets.

Efforts to distinguish histories of speciation and introgression are complicated by incomplete lineage sorting (ILS), a stochastic process in which lineages fail to coalesce in their most recent common ancestor (Hudson, 1983; Pamilo & Nei, 1988; Tajima, 1983). The stochastic nature of ILS means that both speciation and introgression histories can generate similar gene tree topologies at a locus. This makes it challenging to assign histories based solely on gene tree frequencies. For instance, the speciation history ((A,B),C) can generate a gene tree with the topology ((B,C),A) in the absence of introgression between B and C. A significant asymmetry in discordant gene tree frequencies is generally taken as evidence of introgression, but the wrong taxa can be implicated in introgression if the wrong species tree is used. Therefore, gene tree frequencies alone do not contain enough information to distinguish the introgression history from the species history, only enough to infer that introgression has occurred. One proposal has been to identify the introgression history as the one with the minimum average gene tree node height (Fontaine et al., 2015; Forsythe et al., 2020; Li et al., 2019). The logic behind this test is that post-speciation introgression is, by definition, more recent than speciation, leading to more recent coalescence. This rule should be reliable in cases with large amounts of introgression, but if introgression occurs at lower rates, then most gene trees matching the introgression history will be generated by ILS. In this case such trees will contain deeper nodes than the history of speciation. It is therefore unclear how much introgression is necessary for this approach to be useful.

While introgression and ILS reveal the complicated nonbifurcating ways that species are related to one another, understanding the history of speciation and introgression is still crucially important. The species tree provides a summary of the natural history of organisms, including their taxonomic groupings, divergence times, and dynamics of speciation and extinction (O'Meara, 2012). Understanding the species history is also central to understanding the role for introgression in evolution (Blair & Ané, 2020; Dowling & Secor, 1997; Harrison & Larson, 2014; Rhymer & Simberloff, 1996; Taylor & Larson, 2019). Misspecification of the species tree can significantly impact many downstream analyses, including inferences of introgression and the reconstruction of ancestral states. For these reasons, distinguishing histories of speciation from those of introgression remains an important and unsolved problem.

Here, we use theory, simulation, and supervised machine learning analyses to investigate how histories of speciation and introgression may be disentangled. We ask how much introgression is necessary to potentially mislead several approaches to species tree inference, including summary gene tree approaches, parsimony, phylogenetic network approaches, and approaches based on minimum node depths. To try to disentangle these histories, we train supervised machine learning models on simulated datasets from com-

peting species tree topologies and introgression events, finding that the species tree can be recovered with a high degree of accuracy. Using feature-importance analyses, we find that the variances in coalescence times within the two competing gene tree topologies are the most informative features, but no single piece of information alone can accurately recover the species tree topology in all areas of parameter space. Based on our findings, we provide recommendations to researchers dealing with complex histories of reticulation in phylogenomic datasets.

Materials and Methods

Modelling the minimum amount of introgression needed to make gene trees and site patterns supporting the history of introgression the most frequent

We begin with an exploration of a theoretical question: how much introgression is necessary to make the gene tree topology matching the history of introgression the most common gene tree? This question has implications for summary approaches that infer species trees using the most common gene trees among quartets or rooted triplets (e.g. Liu et al., 2010; Zhang et al., 2018), and has been addressed in different ways in a handful of previous studies (Long & Kubatko, 2018; Pang & Zhang, 2023; Solís-Lemus et al., 2016; Zhu et al., 2016). We make use of the multispecies network coalescent framework, which models the effects of introgression and ILS on gene trees simultaneously by breaking histories of speciation and introgression into separate "parent tree" histories (Degnan, 2018; Hibbins & Hahn, 2019; Meng & Kubatko, 2009; Yu et al., 2014) (Figure 2). These parent trees are similar in concept to the displayed trees embedded in a phylogenetic network (Figure 1) (Zhu et al., 2016), but we use them as different histories within which to model the multispecies coalescent. We use "network" or "network model" to refer to the general model that combines multiple histories in a probabilistic framework (Figure 2A); "history" to refer to particular biological histories of speciation and introgression (i.e. parent trees, Figure 2B); and "topology" to refer to the relationships among species observed in a specific tree, regardless of its origin.

In what follows, we use "speciation" to refer to periods of divergence without gene flow; in phylogenies such events most commonly represent the evolution of reproductive isolation between species but could also be applied to populations undergoing periods of divergence. Consequently, the speciation history corresponds to the sequence of splitting events that result in lineages becoming reproductively isolated from one another, and therefore having independent coalescent histories. Introgression occurs after speciation between particular pairs of species, with histories of introgression putting the species involved as sister lineages (e.g. Figure 2B). Similar network models can be applied to other reticulate processes such as horizontal gene transfer and allopolyploidization, but we do not consider these processes here.

We model a rooted three-taxon tree with the species topology ((A,B),C) (Figure 2A). The internal branch shared by species A and B has a length of τ_s, in units of $2N$ generations. Post-speciation introgression occurs between species B and C in two possible directions: C → B and B → C. We model these two directions as separate possible introgression events that occur at rates of δ_2 and δ_3, respectively (Figure 2B). These δ parameters correspond to the proportion of the genome that has introgressed within each history and are equivalent to the inheritance probability parameters used in phylogenetic networks, often denoted γ (e.g. Meng & Kubatko, 2009; Yu et al., 2012). The histories produced by each direction of introgression also contain internal branches shared by species B and C, which we denote τ_{m2} and τ_{m3} for introgression from C → B and B → C, respectively (Figure 2B). The length of this internal branch is affected by the direction of introgression; for instance, when introgression is B → C, the presence of lineages from B in C allows C to coalesce with A more quickly, resulting in a smaller value of τ_m (Hibbins & Hahn, 2019, Figure 2B). Although processes of speciation and introgression occur over continuous time intervals, we model them as instantaneous "pulse" events for simplicity and consistency with the literature.

For the sake of mathematical tractability and visualization of results in the models presented here, we combine the two directions of introgression into averaged δ and τ_m parameters. As we treat each direction as an independent event, the total amount of introgression is simply the sum of the two δ parameters, giving

$$\delta = \delta_2 + \delta_3. \quad (1)$$

For the τ_m parameter, we sought to solve for a single parameter that incorporates the contributions of both directions of introgression to the overall rate of gene tree discordance. This gave us the following expression (see Section 1 of the Supplement for the complete derivation):

$$\tau_m = \ln\left(\frac{\delta_2 + \delta_3}{\delta_2\left(e^{-\tau_{m2}}\right) + \delta_3\left(e^{-\tau_{m3}}\right)}\right) \quad (2)$$

We use these formulations of δ and τ_m in all equations that follow.

Within a particular history, concordant trees (with respect to that history) arise with probability $1 - \frac{2}{3}e^{-\tau}$, and a given discordant gene tree will arise with probability $\frac{1}{3}(e^{-\tau})$, where τ is the length of that history's internal branch (Figure 2C). To get the overall expected frequencies of each gene tree topology, we weight the expected frequencies within each history by the probability that a locus follows that history ($1 - \delta$ for the species history, and δ for the introgression history). In what follows, we will refer to gene trees that match the species history as "AB" trees, and those that match the history of introgression as "BC" trees. For gene trees matching the species history (black tree in Figure 2B), their frequency, f_{AB}, is

$$f_{AB} = (1-\delta)\left(1 - \frac{2}{3}e^{-\tau_s}\right) + \delta\left(\frac{1}{3}\left(e^{-\tau_m}\right)\right) \quad (3)$$

and for gene trees matching the introgression history (blue tree in Figure 2B), their frequency, f_{BC}, is

$$f_{BC} = \delta\left(1 - \frac{2}{3}e^{-\tau_m}\right) + (1-\delta)\left(\frac{1}{3}\left(e^{-\tau_s}\right)\right). \quad (4)$$

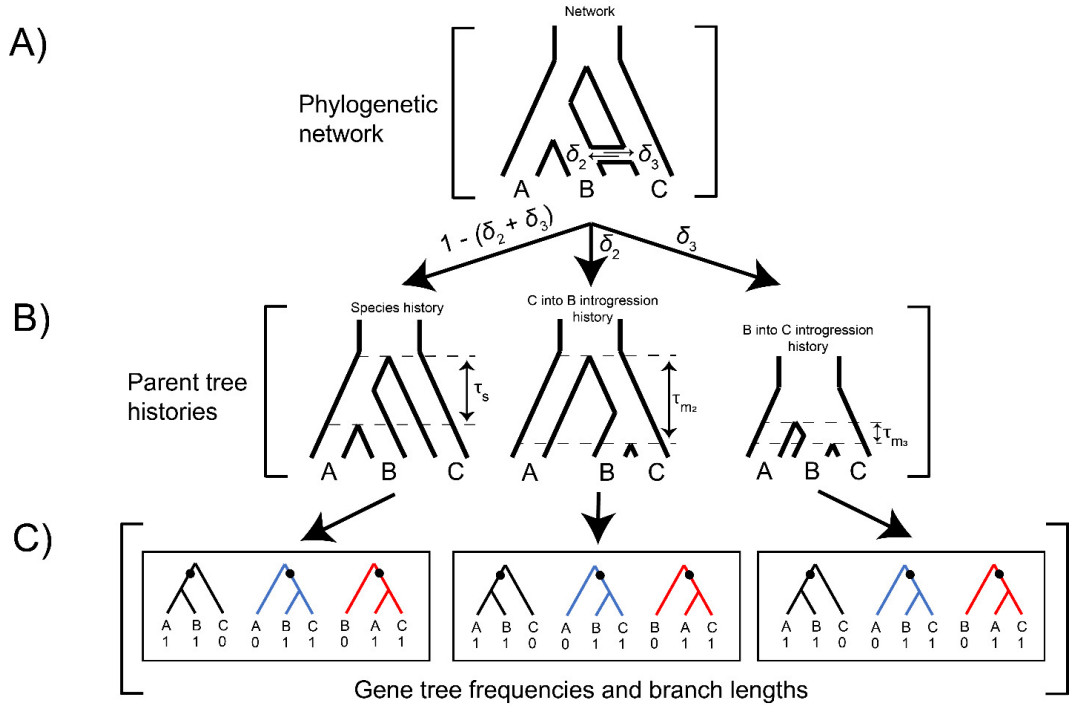

Figure 2. Modelling histories of speciation and introgression.

A) A phylogenetic network is used to model histories of speciation and introgression in a single probabilistic framework. B) The individual histories (speciation and introgression) modelled by a network can be separated into "parent trees" that describe the histories of speciation or introgression at individual loci. Loci follow histories of introgression with probability δ_2 (for C → B introgression) and δ_3 (for B → C introgression), corresponding to the proportion of introgression across the genome. The internal branch lengths τ_s and τ_m, for histories of speciation and introgression, respectively, are determined by the timing of introgression relative to speciation and the direction of introgression. Our model allows for arbitrary amounts of introgression in either or both directions, with τ_m and δ being weighted averages of the contributions of both directions. C) These three histories generate gene trees with expected frequencies and branch lengths under the standard multispecies coalescent model, which includes incomplete lineage sorting within each history. Mutations on the internal branches of these gene trees lead to parsimony-informative biallelic sites (shown as black dots and 0/1 ancestral/derived states at the tips).

To find the amount of introgression required to make f_{BC} equal to f_{AB}, we set the expected frequencies of the two topologies equal to one another, giving the following:

$$\begin{aligned}&\delta\left(1-\frac{2}{3}e^{-\tau_m}\right)+(1-\delta)\left(\frac{1}{3}\left(e^{-\tau_s}\right)\right)\\&=(1-\delta)\left(1-\frac{2}{3}e^{-\tau_s}\right)+\delta\left(\frac{1}{3}\left(e^{-\tau_m}\right)\right)\end{aligned}\tag{5}$$

Solving for δ, we obtain the following expression:

$$\delta_{\min}=\frac{1-e^{-\tau_s}}{2-e^{-\tau_m}-e^{-\tau_s}}\tag{6}$$

This same expression has been derived in Jiao et al. (2020) and Pang and Zhang (2023) for unidirectional introgression but can now be applied to arbitrary amounts of introgression in both directions using the definitions of δ and τ_m provided in Equations 1 and 2.

We applied the same theoretical framework to ask a different but related question: how much introgression is necessary to make the biallelic site pattern supporting the introgression history the most common? These parsimony-informative sites are important to the performance of many standard methods for phylogenetic inference. Biallelic sites arise from mutations on internal branches of gene trees (Figure 2C). Assuming an infinite-sites model (i.e. no multiple hits), the sum of the lengths of internal branches of relevant gene trees gives us the expected number of parsimony-informative sites supporting either the ((A,B),C) or ((B,C),A) gene trees (Mendes & Hahn, 2018). For gene trees that arise from lineage sorting in each history (which occurs with probability $1-e^{-\tau_s}$), the internal branch length is $\tau+\left(\frac{\tau}{e^{\tau}-1}\right)$ (Mendes & Hahn, 2018), while for all other gene trees (i.e. the ones due to incomplete lineage sorting in any history) the internal branch length is 1 (in units of $2N$ generations). To obtain the average internal branch length across histories, we weight the relevant branch lengths within each history by the frequencies of the gene trees, and then across histories by their probabilities at a locus. This gives the following expression for biallelic sites that support the species history:

$$\begin{aligned}S_s=(1-\delta)&\left(\left(\left(1-e^{-\tau_s}\right)\left(\tau_s+\left(\frac{\tau_s}{e^{\tau_s}-1}\right)\right)+\frac{1}{3}e^{-\tau_s}\right)\right)\\&+\delta\left(\frac{1}{3}e^{-\tau_m}\right)\end{aligned}\tag{7}$$

and this expression for sites supporting the history of introgression:

$$\begin{aligned}S_m=\delta&\left(\left(\left(1-e^{-\tau_m}\right)\left(\tau_m+\left(\frac{\tau_m}{e^{\tau_m}-1}\right)\right)+\frac{1}{3}e^{-\tau_m}\right)\right)\\&+(1-\delta)\left(\frac{1}{3}e^{-\tau_s}\right).\end{aligned}\tag{8}$$

To find the δ value required to make site patterns from both histories equally frequent, we set these two expressions equal, as before. After solving for δ and simplifying, we obtain the following expression:

$$\delta_{\min}=\frac{\tau_s}{\tau_m+\tau_s}.\tag{9}$$

To our knowledge, this relationship has not been found before.

Modelling the minimum gene tree node height

We also used our model to investigate the "minimum node height" criterion for distinguishing between histories (Fontaine et al., 2015; Forsythe et al., 2020; Li et al., 2019). For this analysis we make use of the expected coalescence times and conditional gene tree frequencies derived in Hibbins and Hahn (2019). Since gene trees in real data arise from a combination of different histories, we obtain general expressions here by averaging the time to first coalescence across all possible gene trees that share the same topology. These times to first coalescence are measured conditional on a gene tree topology, so they must be weighted by the frequency of a particular gene tree relative to the frequency of all other possible gene trees that share that topology, rather than simply the overall expected frequency. We used these expressions to calculate the expected node heights of AB and BC gene trees over a range of parameter values from our model, and asked in which parts of the parameter space each node height was the smallest. See Section 2 of the Supplementary Materials and Methods for the relevant expressions.

Simulating gene trees under the coalescent

To validate our theoretical results, and to conduct downstream analyses, we simulated gene trees under various speciation and introgression scenarios using the program *msprime* (Baumdicker et al., 2022). For each simulated dataset, we denoted the history of speciation among lineages as a Newick string and then used *Demography.from_species_tree()* to convert this into an *msprime* demography object. Introgression was added to these demography objects using the *demo.add_mass_migration* function with the specified timing, direction, and rate of introgression. We then simulated tree sequences using *sim_ancestry*(), specifying a single haploid sample from each lineage with a recombination rate of $1x10^{-8}$ and a sequence length of $1x10^{7}$. These tree sequences were converted into Newick trees for further parsing. We converted branch lengths to units of $2N$ generations assuming $N = 10000$ and sampled every third gene tree to reduce the effects of spatial autocorrelation, resulting in approximately 1200 gene trees per dataset (i.e. for each unique combination of parameters).

We simulated gene tree datasets in *msprime* under two competing network models: one where A and B are most closely related with post-speciation introgression between B and C (Figure 2A), and one where B and C are most closely related with post-speciation introgression between A and B (Supplementary Figure 1). The histories of speciation ("AB" for the scenario in Figure 2, "BC" for the scenario in Supplementary Figure 1) in these two scenarios were used as labels for binary classification. We simulated a 10x10x10 grid of τ_s, τ_m, and δ_2 / δ_3 values, in each of the two directions of introgression and for both possible network models, resulting in 4000 total simulated datasets. τ_s and τ_m both ranged from $0.1N$ to $2.2N$ generations, and δ_2 and δ_3 each ranged from 0.01 to 0.9 within each simulated direction of introgression. For each network model, we simulated introgression from the unpaired lineage into the paired lineage (C into B or A into B; δ_2 in Figure 2 and Supplementary Figure 1), and from the paired lineage into the unpaired lineage (B into C or B into A; δ_3 in Figure 2 and Supplementary Figure 1), but not in both directions simultaneously.

Supervised machine learning analyses to predict the history of speciation

To investigate whether it is possible to accurately distinguish histories of speciation and introgression, we applied supervised machine learning approaches to our simulated gene tree datasets using the Python package *scikit-learn* (Pedregosa et al., 2011). For each simulated gene tree dataset, we used *ETE3* (Huerta-Cepas et al., 2016) to parse 21 features, including the frequency of each of the three gene tree topologies (3 features), the time to coalescence of each pair of species within each of the gene trees (3x3 = 9 features), and the variances of these times to coalescence (3x3 = 9 features). See Supplementary Table 2 for a full description of each feature. The final dataframe therefore consisted of 4000 rows (one per set of simulated gene trees) and 21 columns (one for each feature). We split simulations into a training dataset of 3000 observations and a test dataset of 1000 observations, and then standardized all features using *scikit-learn*'s *StandardScaler()*. We trained five binary classification models on the training dataset, all using the default model parameters: logistic regression, support vector machine, Gaussian naïve Bayes, decision tree, and random forest. Finally, the performance of each of these trained models was scored on the test set.

We assessed the importance of each feature to each model's predictive power on the test set using the *permutation_importance()* function in *scikit-learn*. This analysis randomly permutes each feature in the dataset and scores its importance by how much this permutation decreases model performance. Permutation was repeated 30 times for each feature, and a feature was deemed significantly important for a model if its mean importance score across replicates was at least twice that score's standard deviation. We ranked the importance of each feature across models by taking the average importance score. To complement this analysis, we also performed stepwise model selection for a logistic regression model using the Akaike Information Criterion (AIC), as implemented in the R package *MASS*. This approach begins with the full model containing all variables, and incrementally removes each variable in a stepwise fashion until the model with the minimum AIC value, containing only the most informative variables, is reached.

Assessment of *PhyloNet* edge weight estimates

We ran *PhyloNet*'s *InferNetwork_ML* method (Yu et al., 2014) on simulated gene tree datasets to evaluate the accuracy of edge weight estimates against simulated introgression proportions. We simulated gene trees with *msprime* under the same parameter combinations as for the machine learning analyses and used a custom Python script to convert the

outputs of these simulations into NEXUS-formatted input files for *PhyloNet*. On the *PhyloNet* outputs, we used the Julia package *PhyloNetworks* (Solís-Lemus et al., 2017) and *ETE3* (Huerta-Cepas et al., 2016) to extract the weight of the edge in the network that corresponds to the known history of introgression. In cases where no introgression was inferred by *PhyloNet*, we assigned a weight of 0 when the estimated topology corresponded to the species tree, and a weight of 1 when it corresponded to the introgression history. Edge-weight estimates for each simulated value of the rate of introgression were averaged across all combinations of branch length parameters.

Data Availability

Scripts and data for all analyses are available at https://github.com/mhibbins/dist_histories.

Results

Arbitrarily small amounts of introgression can make gene trees and site patterns supporting the history of introgression the most frequent

To visualize the amount of introgression necessary to make discordant gene trees more common than concordant ones, we plotted the δ value derived in Equation 6 over the space of τ_s and τ_m (Figure 3A, Supplementary Figure 2). Overall, we find that the smaller τ_s is, the less introgression is necessary to make the introgression tree the most common gene tree. This occurs because smaller values of τ_s lead to more ILS (and therefore more discordance), even in the absence of introgression. As the value of τ_s approaches 0, the amount of discordance approaches 66% in the species tree history, and all tree topologies arise at approximately the same frequency. As this occurs, the amount of introgression required to tip the balance toward the gene tree topology matching the introgression history approaches 0. The minimum value of δ is also affected by the timing and direction of introgression, with more recent introgression (i.e. larger τ_m) from C into B resulting in a lower minimum value of δ required. Both the timing and direction of introgression affect the amount of ILS within introgressed histories, as more recent introgression and introgression from C into B both lead to longer internal branches in the introgressed history (Figure 2). Conversely, if ILS is low in the species history (large τ_s) and high in the introgression history (small τ_m and B into C introgression), rates of introgression can be very high (approaching 100%) and still not result in gene tree topologies matching the introgression history being most common. The histories of speciation and introgression in the extremes of the parameter space considered here are summarized in Figure 3D.

When plotting the minimum δ value for parsimony-informative sites derived in Equation 9 we find a similar overall pattern to the result for gene tree frequencies, but with a slightly different shape to the contours (Figure 3B). Once again, large amounts of ILS in the species history mean that little introgression is necessary to make sites supporting the introgression history the most common. The differences between Figures 3A and 3B imply that parsimony methods may perform better than summary methods under certain speciation and introgression scenarios, and vice versa. The logic behind this is that if biallelic sites, for example, have a larger minimum δ value than gene tree frequencies in a particular area of parameter space, then parsimony methods (which rely on biallelic sites) can tolerate more introgression than summary methods (which rely on gene tree frequencies) before the introgression history becomes most supported. To better visualize these differences, we plot the gene tree frequency-based δ value subtracted from the parsimony-based δ value in Figure 3C. In general, we find that methods that use gene tree frequencies should perform better (i.e. require more introgression to be misled) than parsimony methods when the δ value required is less than 50%, while parsimony methods should perform better when the δ value is greater than 50%. The areas where the two methods have the greatest difference in performance are not linear with the amount of discordance, but rather fall in regions with intermediate levels of discordance. In these parts of parameter space, the difference in minimum δ between methods is as high as approximately 11%. This means, for example, that when δ is less than 50%, summary methods can tolerate an up to 11% higher rate of introgression than parsimony methods before the introgression tree becomes most common. In either case, little introgression is needed to "hide" the species tree when there is a lot of ILS in the species history.

The results presented in this section suggest that, at least in the simplest three-taxon case, it may be difficult to distinguish between histories using gene tree frequencies or informative sites alone, even under relatively modest introgression scenarios. Previous studies have proposed using the minimum average node height among gene tree topologies to distinguish the species history from the introgression history (Fontaine et al., 2015; Forsythe et al., 2020; Li et al., 2019). Using the same modelling framework as in the previous section, we evaluated the space over which the "minimum node height" criterion may be effective at inferring the introgression history. We found that the minimum amount of introgression required to make the minimum average node height correspond to the history of introgression, rather than that of speciation, varied from 20% to approximately 40% (Figure 4). This variation depends primarily on the timing of introgression relative to speciation, with more recent introgression resulting in less introgression required. The direction of introgression has little effect; Figure 4 shows the result for equal amounts of introgression in both directions, while Supplementary Figure 3 shows the pattern for each direction separately. As the species history approaches a star phylogeny, the necessary rate can fall below 20% and approaches 0 (but this is a relatively small part of the parameter space). These results suggest that the average minimum node height can be informative when the rate of introgression is high but will fail in large parts of parameter space. Overall, it seems unlikely

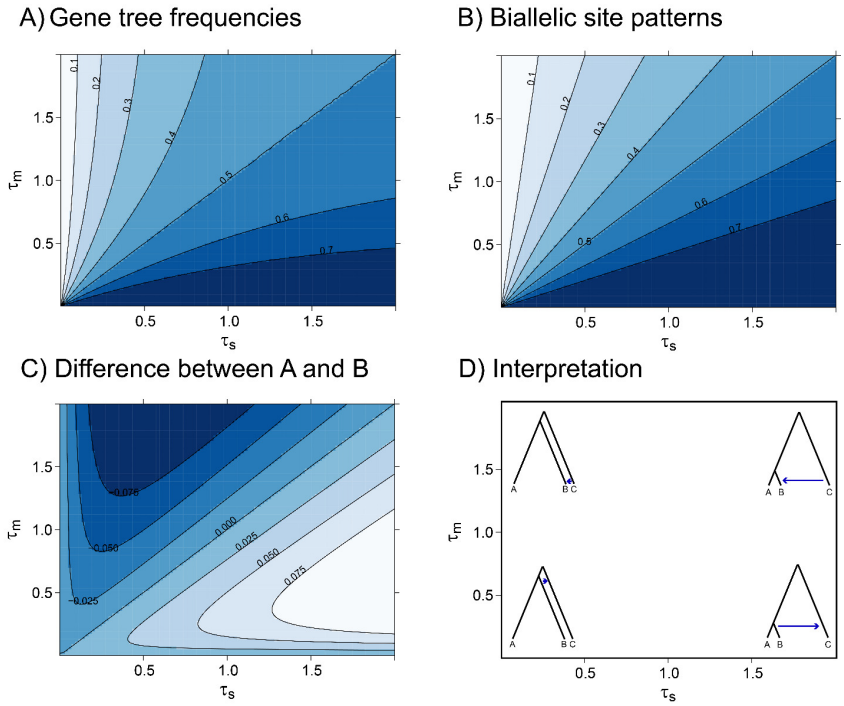

Figure 3. The minimum amount of the genome introgressing that is necessary to make gene trees (panel A) and site patterns (panel B) matching the introgression history the most common, as a function of values of τ_S and τ_m.

Values of δ are given next to the contour lines. C) The difference in the amount of introgression necessary to affect site patterns vs. gene trees; obtained by subtracting the contour values in panel A (for gene trees) from those in panel B (for site patterns) (i.e. B minus A). D) Interpreting the axes in panels A-C. Each corner is labelled with the introgression scenario that best describes that part of the parameter space, assuming a known speciation history of ((A,B),C). Note in the bottom two corners of panel D, the arrow is going from species B into species C, but very close to the time of B's common ancestor with A.

that any one feature alone will reliably help to identify one history over another.

Supervised machine learning recovers the history of speciation with high accuracy

Supervised machine learning is a powerful tool for data classification problems, and its applications in population genetics and phylogenetics are growing (Schrider & Kern, 2018). Supervised machine learning methods can efficiently use multidimensional inputs to make accurate classification of a dataset. Here, we made use of supervised machine learning for binary classification of the history of speciation in simulated datasets. We trained and tested five machine learning models for binary classification using *scikit-learn* (Pedregosa et al., 2011). We found that the supervised machine learning methods recovered the history of speciation with high accuracy, ranging from 76.3% for naïve Bayes to 93.3% for a random forest classifier (Table 1).

To understand which features were the most predictive of the correct species history across models, we conducted a permutation feature importance analysis as implemented in *scikit-learn*. This analysis randomizes individual features in the dataset one at a time and evaluates how this impacts the model's predictive accuracy. Each feature is scored by the amount that prediction accuracy is reduced when that feature is randomized before fitting a model. We ranked each feature by its overall importance score summed across the five models. We found that most features (17/21) had significant importance scores in at least one model, and there was notable variation in the importance of features across models (Figure 5). Across models, the most predictive features were the variances in coalescence times of the sister species in gene trees matching the two relevant histories: i.e. the variance in time to coalescence of A and B in AB gene trees (AB_AB_var in Figure 5) and B and C in BC gene trees (BC_BC_var). The next four most important features (in most models) were the node distances in those gene trees (AB_AB_dist and BC_BC_dist) and the frequencies of the gene trees (AB_freq and BC_freq). The importance of other features after these six is notably lower and varies between models. Four features did not have a significant importance score in any model: AB_AC_var, BC_AC_var, AB_BC_var, and BC_AB_var. See Supplementary Table 2 for complete descriptions of each feature.

To further investigate which features had the most predictive power, we took advantage of the statistical properties of logistic regression to conduct stepwise model selection using the Akaike Information Criterion. The final model contained 9 of the 21 features (Supplementary Table 1), including 5 of the 6 most important features across machine learning models discussed in the previous paragraph—the frequency of BC gene trees (BC_freq) was not included. Curiously, the variance in time to coalesce of A and B in BC gene trees (BC_AB_var) did not have a significant importance score in any of the machine learning models, and yet was included by the stepwise selection procedure. Aside from the omission of BC_freq and inclusion of BC_AB_var, this analysis was generally in agreement with

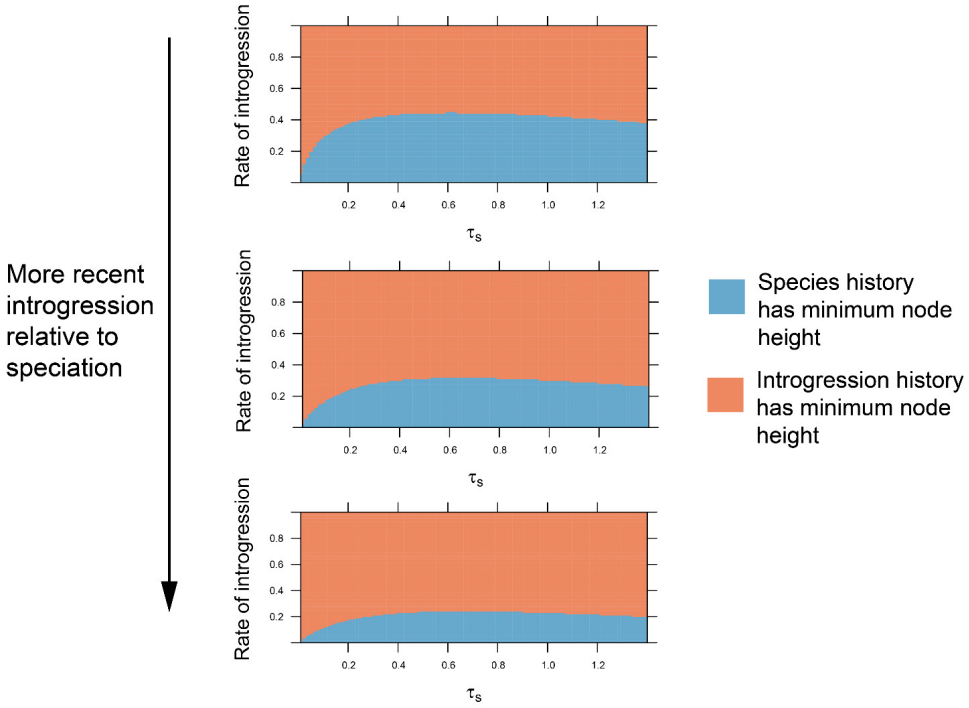

Figure 4. Distinguishing histories of speciation and introgression using the minimum gene tree node height.

In blue areas, the minimum node height (averaged across all gene trees sharing a topology) is in gene trees matching the species history, while in orange it is in gene trees matching the introgression history. The boundary between the two denotes the minimum amount of introgression necessary to make the introgression history the one with the shortest node height. This minimum amount of introgression decreases as the species tree approaches a star tree (i.e. as τ_s approaches 0) and as introgression becomes more recent relative to speciation (from top to bottom, t_m values of 0.45, 0.3, 0.15, respectively). In most of parameter space, a significant amount of introgression is necessary, ranging from 20% (bottom) to 40% (top). Patterns shown are for equal rates of introgression in both directions.

Table 1. Performance of our supervised machine learning classifiers on simulated test datasets.

Model	Performance score on test set
Logistic regression	0.792
Support vector machine	0.913
Gaussian Naïve Bayes	0.763
Decision tree	0.878
Random forest	0.933

the feature permutation analysis about the most important features for distinguishing histories.

To help uncover general patterns that might be useful for distinguishing between competing histories in real datasets, we plotted the behavior of the six most important features discussed in the previous paragraph in each of the two histories in our simulated datasets (Figure 6). Each variable is plotted over the space of δ values for each direction of introgression and averaged across all values of τ_s and τ_m within a value of δ. Generally, the gene tree with the lower variance in the time to first coalescence corresponds to the species history. This pattern remains true except when rates of introgression are very low (less than approximately 5%) or very high (greater than approximately 75%). The consistency of this pattern across a wide range of possible δ values partially explains why these features emerged as the most important in our machine learning models (Figure 5). The other features behave as expected from our theoretical results, with increasing rates of introgression decreasing the minimum node height and increasing the frequency of trees that match the introgression history. In general, the minimum node height corresponds to the introgression history when the rate of introgression is higher than ~25%, and the most common gene tree matches the species history when the rate of introgression is less than 50%, but there is significant deviation from these patterns depending on rates of discordance within each history, as well as the direction of introgression (Supplementary Figure 4). Nonetheless, the large differences in values between histories in most of the space makes these features highly informative together or when combined with other information.

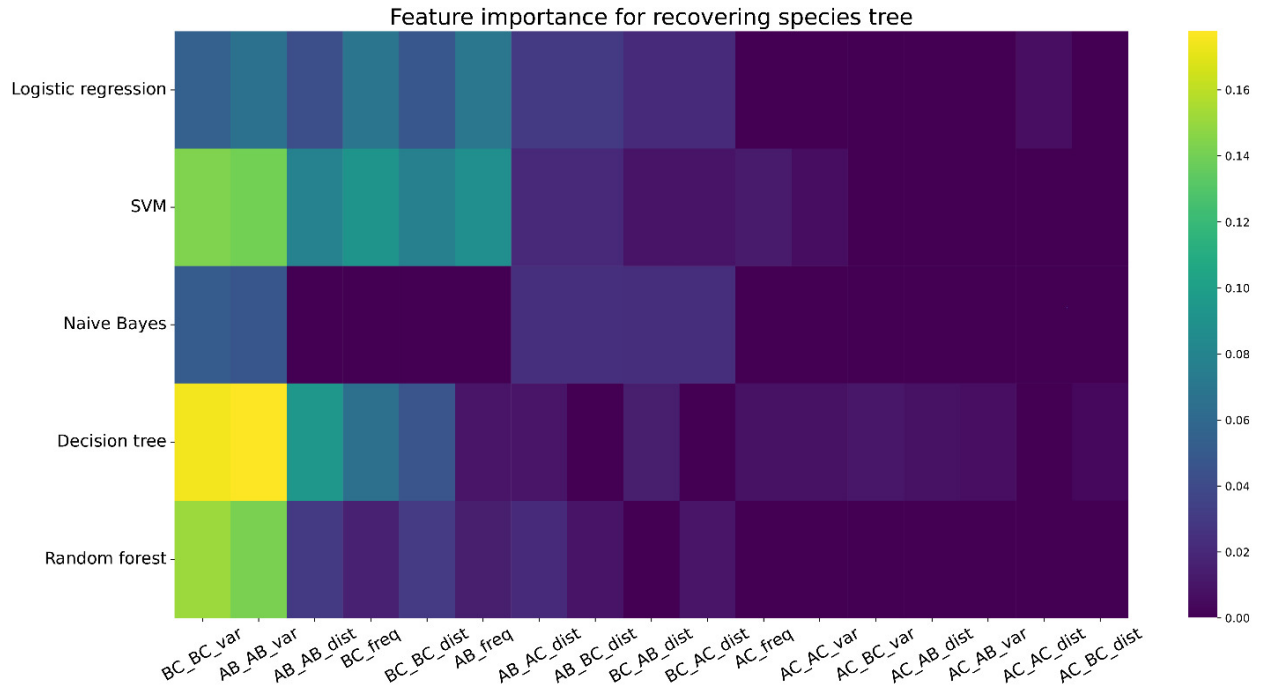

Figure 5. Model features (x-axis) ranked by importance score (color scale) across all our trained classification models (y-axis).

Lighter colors on the scale indicate a higher importance score for that feature. The importance score reports the degree to which accuracy is reduced in the fitted model when that feature is randomized (i.e. a score of 0.16 means prediction accuracy is reduced by 16%). In feature names, the first pair of letters indicate the gene tree topology, and the second pair indicate the pair of species within that gene tree topology, so the feature "AB_AC_dist" is the branch length distance between species A and C in gene trees where A and B are sister taxa. Four features are not shown which did not have a significant importance score in any of our trained models.

PhyloNet accurately estimates edge weights for different histories despite low signal in gene tree frequencies

While the machine learning results in the previous section provide possibilities for distinguishing histories of speciation and introgression, they do not provide information on how much of the genome corresponds to each history. Phylogenetic network methods do not explicitly label histories, but they may nonetheless be able to accurately estimate the edge weights corresponding to these histories (one of which should correspond to δ), in addition to correctly identifying the paired lineage involved in introgression. These estimates may be especially useful in parts of the parameter space where gene tree frequencies do not correspond intuitively to introgression probabilities. For example, in the bottom right space of Figure 3A-C, introgression probabilities can be as high as 70% or more, though the gene tree matching the species history is still the most common. To assess edge weights, we simulated gene tree datasets with *msprime* and then passed them to the *InferNetwork_ML* method of *PhyloNet* (Yu et al., 2014). We asked whether the estimated reticulation edge weights for the edge corresponding to the simulated introgression history reflected gene tree frequencies or the true simulated rates of introgression.

In general, we found that *PhyloNet*'s estimated edge weights correspond to the actual proportions of speciation and introgression histories, rather than to gene tree frequencies (Figure 7). We observed substantial variation in the frequency of BC gene trees, especially at high rates of introgression, where essentially any frequency is possible depending on the values of other parameters. This is especially true when introgression was from B into C: for example, with a rate of introgression of 90%, the average frequency of BC gene trees was only ~45%, with a large standard error (Figure 7, bottom right). This is in line with the theoretical results plotted in Figure 3A and 3B. Despite this, *PhyloNet* is generally able to accurately estimate edge weights, with a nearly 1:1 correspondence between the simulated amount of introgression and the mean estimated amount (though there is also significant variation in these estimates). In addition, *PhyloNet* was almost always able to correctly identify species B as one of the lineages involved in introgression. An important caveat to these results, however, is that at very high rates of B into C introgression *PhyloNet* often failed to infer the presence of introgression at all, instead returning a bifurcating tree matching the history of introgression (Supplementary Figure 5). This occurred in approximately 20% of simulations at a simulated rate of 70%, and up to approximately 55% of simulations when the simulated rate of introgression was 90%. Nonetheless, these results suggest that if the histories of speciation and introgression can be correctly identified, *PhyloNet* is generally able to accurately estimate what proportion of the genome corresponds to each history.

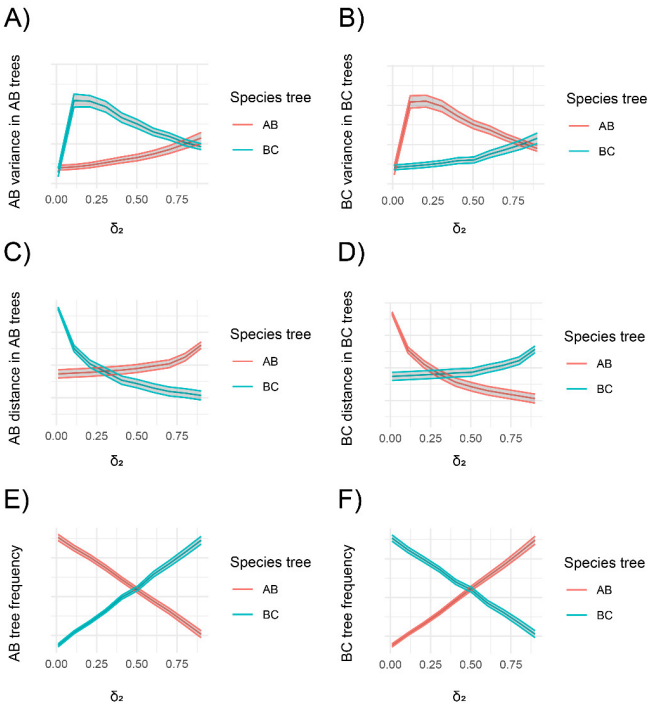

Figure 6. Behavior of the six most informative features identified from our machine learning classifiers, for C → B / A → B introgression (depending on the underlying network model).

Color denotes the true history of speciation underlying the simulated data, and the error bands (transparent areas) denote the standard error introduced by variation in τ_s and τ_m within each value of δ on the x-axis. A) The variance in the time to coalescence of A and B in AB gene trees (denoted AB_AB_var in Figure 5). B) The variance in the time to coalescence of B and C in BC gene trees (denoted BC_BC_var in Figure 5). C) The average time to coalescence of A and B in AB gene trees (denoted AB_AB_dist in Figure 5). D) The average time to coalescence of B and C in BC gene trees (BC_BC_dist in Figure 5). E) The frequency of AB gene trees (denoted AB_freq in Figure 5). F) The frequency of BC gene trees (denoted BC_freq in Figure 5).

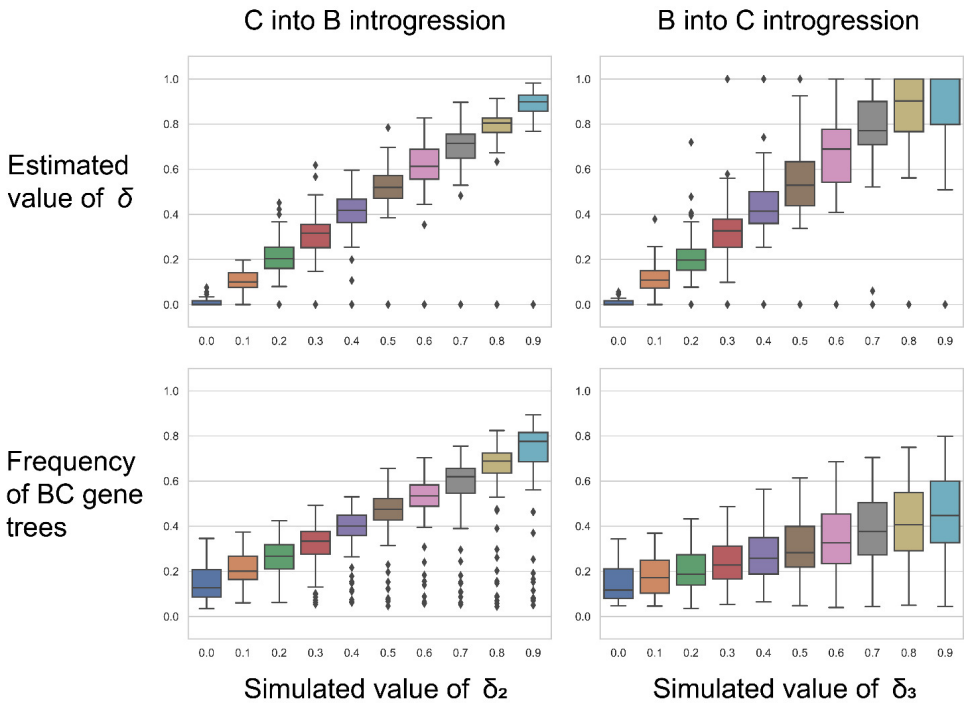

Figure 7. Ability of *PhyloNet* to estimate the proportion of the genome that has introgressed (top row) relative to observed gene tree frequencies (bottom row) when the true history of introgression is known.

For each value of δ, estimates are grouped across all simulated values of τ_s and τ_m. Individual points show outlier estimates; estimates of 0 are assigned when the species tree is inferred with no reticulations and estimates of 1 are assigned when the introgression history is inferred with no reticulations.

Discussion

Genomes often contain a complex mosaic of different relationships among the same set of species. Understanding the historical processes that generate these different histories, such as speciation and post-speciation introgression, is critical to obtaining a complete understanding of the natural history of organisms. Here, we have used theory, simulation, and a supervised machine learning approach to show that while currently available approaches to species tree inference may often be misled in the presence of introgression, it should be possible in principle to distinguish among these histories using the information contained in genome-scale datasets. By exploring the behavior of the most informative features in our machine learning models, in addition to highlighting the accuracy of edge-weight estimates in phylogenetic network approaches, we can now provide recommendations for future analyses and methods, as well as discussing the assumptions and limitations of our work.

For the sake of tractability, our theoretical model (and simulation analyses) makes simplifying assumptions that have implications for patterns in genomic data. First, we assume a single, discrete, post-speciation introgression event between one pair of non-sister taxa. There are a multitude of ways that real introgression scenarios could be more complicated, including multiple discrete events between different pairs of species, or continuous periods of gene flow rather than discrete pulses of hybridization. So long as introgression is primarily between one pair of lineages, the general patterns of genomic features we have reported here should hold, but the power of particular features to distinguish between different histories may be affected. Alternatively, if other lineages are involved in introgression scenarios, this may affect gene tree patterns in ways that are not accounted for by our model. For example, introgression between ancestral sister lineages can result in *both* discordant gene tree topologies becoming more common than the species tree topology, rather than just one (Jiao & Yang, 2021; Long & Kubatko, 2018; Solís-Lemus et al., 2016). Additionally, introgression from a more distantly related ghost lineage can cause an incorrect history of introgression to be inferred from available data (Tricou et al., 2022), resulting in loci matching the "introgression history" having longer branches than expected. Lastly, introgression events between multiple pairs of lineages will present a more complex problem than simply distinguishing between two competing histories, as new histories are introduced by each event; future work will be needed to address such scenarios.

Networks are sometimes thought of as consisting of a species tree-like backbone with edges added on (Francis & Steel, 2015), and some approaches to network inference fix this backbone to simplify estimation of other parameters (Flouri et al., 2020; Molloy et al., 2021). This backbone tree is also often visualized as the tree corresponding to the majority edge in a phylogenetic network. In introgression analyses, it is common to construct this backbone species tree using standard methods, which are then used to guide subsequent investigations. Our theoretical results suggest that standard methods for species tree inference will often fail to infer the correct species tree in the presence of introgression (Figure 3). If levels of ILS in the species history are sufficiently high, very little introgression is necessary to mislead both summary gene tree (Figure 3A) and parsimony (Figure 3B) methods into inferring the introgression history as the species tree (Eckert & Carstens, 2008; Long & Kubatko, 2018; Pang & Zhang, 2023; Solís-Lemus et al., 2016; Zhu et al., 2016). We also found that sufficiently high rates of introgression can mislead network methods to infer the wrong backbone tree (Supplementary Figure 5). This is important because both summary gene tree and parsimony approaches to species tree inference have been shown to outperform maximum-likelihood phylogeny inference in the presence of discordance due to ILS (Mendes & Hahn, 2018; Mirarab et al., 2016) and are often applied to datasets with high rates of gene tree discordance. Even network methods, which explicitly model ILS and introgression simultaneously, can fail to detect both histories at high rates of introgression.

Very high rates of gene tree discordance, on the order of 50-60% or higher, coupled with evidence of at least some introgression have been found in many empirical systems, including tomatoes (Pease et al., 2016), darters (MacGuigan & Near, 2019), butterflies (Edelman et al., 2019), monkeyflowers (Nelson et al., 2021), primates (Vanderpool et al., 2020), and suboscine birds (Singhal et al., 2021), suggesting the problems identified here may be common. If an incorrect species tree or backbone tree is used for downstream analyses, introgression inferences involving the placement and directionality of the hybrid edge and estimated inheritance probabilities will also be misled. In empirical studies where species relationships are uncertain due to introgression, one classic approach has been to identify regions of the genome that are expected to be resistant to introgression. These can be genes directly involved in the evolution of reproductive isolation ("speciation genes"; Cutter, 2013; Ting et al., 2000; Zachos, 2009), or regions of low recombination. While multiple empirical studies have now demonstrated a positive correlation between recombination and introgression (Brandvain et al., 2014; Geraldes et al., 2011; Martin et al., 2019; Nelson et al., 2021; Schumer, Rosenthal, et al., 2018), it is often a weak relationship and may have limited predictive power in isolation. Furthermore, genes involved in reproductive incompatibilities may be more likely to have discordant topologies due to ILS (R. J. Wang & Hahn, 2018), making them less informative for determining the species history.

In addition to its implications for species tree inference, our model provides a cautionary note on the utility of gene tree frequencies in other introgression-related inferences. The frequency of gene trees matching the introgression history does not correspond neatly to the rate of introgression across the genome, owing to ILS at introgressed loci. This is highlighted by the fact that the *D* statistic (Green et al., 2010), a test based on gene tree frequencies, is a poor estimator of the rate of introgression (Hamlin et al., 2020; Martin et al., 2015). In addition, the lack of information

in gene tree frequencies complicates inferences of homoploid hybrid speciation, a controversial process (Nieto Feliner et al., 2017; Schumer et al., 2014; Schumer, Xu, et al., 2018) in which hybridization is proposed to cause speciation without a change in ploidy. One potential piece of evidence for homoploid hybrid speciation is equal frequencies of the two most common gene tree topologies across the genome, which one might expect if reproductive isolation begins immediately in F1 hybrids. Our results show that it is possible to have two equally frequent majority gene trees at any rate of introgression (Figure 3), not only with δ=0.50. Conversely, a value of δ=0.50 does not necessarily imply that hybrid speciation has occurred, or that gene trees corresponding to the two histories occur at equal frequencies (Figure 7). This highlights the difficulty in devising tests for hybrid speciation based on only gene tree frequencies. Fortunately, we found that *PhyloNet* accurately estimates the rate of introgression across parameter space of our model (Figure 7), demonstrating the utility of additional branch-length information in estimating the amount of introgression.

Our machine learning models, especially those based on decision trees, were able to classify the species history of simulated datasets with high accuracy (Table 1). This suggests that it may be possible to build a similar classifier that can be applied to real datasets in cases where there are two competing histories in the data. However, one hurdle to developing such an approach is the necessity of generating simulated training datasets. These simulations require knowledge of demographic parameters that cannot necessarily be accurately estimated if the histories of speciation and introgression are not known. One solution is to simulate over a large space of possible parameter combinations that covers the entire range of biologically plausible values, avoiding issues of extrapolation or model misspecification in empirical datasets, and then to build a classifier from this data for application to empirical data. Generating this training dataset would not be computationally prohibitive using standard coalescent simulation in the three-species case. Such a classifier could easily be applied to larger phylogenies if there is only a single introgression event; for multiple intogression events, it may be necessary to divide larger trees into a set of classification tasks for three-taxon subtrees. Similar simulation approaches to training machine learning classifiers have found success in application to real datasets for other introgression-related questions. For example, Schrider et al. (2018) identified introgressed loci in two *Drosophila* species by simulating genomic windows and training a classifier on summary statistics. Burbrink and Gehara (2018) identified ancient introgression in the phylogeny of New World kingsnakes using a classifier trained on pairwise distance features obtained from simulated gene trees. One other possibility is to use unsupervised or semi-supervised machine learning approaches, such as a generative adversarial network (e.g. Smith & Hahn, 2023; Z. Wang et al., 2021), to generate histories of speciation and introgression with gene tree features that most closely resemble those observed in the empirical data.

Our feature importance analysis revealed general patterns in genomic summary statistics that may be useful for distinguishing among histories. The behavior of the six most informative features is in line with expectations from our mathematical model and other previous work and demonstrates how these features can each reveal different pieces of information about the histories of the data (Figure 6). The most useful features were the variances in coalescent times of sister taxa within gene trees, as they are informative over a large part of the space of possible δ values. Gene trees matching the introgression history tended to have a higher variance in coalescence times; this is consistent with introgression introducing a more recent lower bound to coalescence, increasing the range of possible coalescence times. This lower bound also explains the informativeness of gene tree pairwise distances, which can take smaller possible values in gene trees matching the introgression history.

Nevertheless, no single feature is completely informative in all areas of parameter space. This observation explains why most of our features had significant importance scores in at least one model (Figure 5). The four features with no significant importance scores all involved the variance in the timing of the second coalescence event in gene trees; this quantity should not be directly predictive of the underlying species history but may be marginally useful for inferring the direction of introgression within a network model, a signal that is likely highly correlated with other features in our dataset. Finally, the behavior of our model features suggests that machine learning approaches may be useful for other introgression-related inference tasks. For example, the simulated direction of introgression broadly affected the patterns observed in our simulated datasets (Supplementary Figure 3, Figure 6 vs. Supplementary Figure 4, Figure 7A vs 7B), so a classifier for the direction of introgression might have similarly high accuracy. These results highlight the power of machine learning to combine different pieces of information to make accurate predictions.

Overall, our work highlights the challenges of distinguishing between speciation and introgression histories, but also provides promising paths forward that may eventually lead to the development of methods for carrying out this task. While massive amounts of introgression may break down the distinctions between contemporary species, understanding the historical processes of speciation and hybridization underlying these lineages is still crucial for studying macroevolutionary processes such as speciation and extinction, as well as the evolution of traits over time.

Acknowledgements

We would like to thank Claudia Solís-Lemus and Marianne Bjorner for help with *PhyloNetworks*, as well as Claudia Solís-Lemus and four anonymous reviewers for their constructive comments. We would especially like to thank one reviewer who identified a key error in our initial model formulation and helped us to derive the correct solution. This

work was supported by National Science Foundation grant DEB-1936187 awarded to Matthew Hahn, and the EEB Postdoctoral Fellowship awarded to Mark Hibbins by the Department of Ecology and Evolutionary Biology at the University of Toronto.

Submitted: September 08, 2023 EDT, Accepted: March 20, 2024 EDT

References

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), Article iyab229. https://doi.org/10.1093/genetics/iyab229

Blair, C., & 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. https://doi.org/10.1093/sysbio/syz056

Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G., & Sweigart, A. L. (2014). Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genetics*, *10*(6), e1004410. https://doi.org/10.1371/journal.pgen.1004410

Burbrink, F. T., & Gehara, M. (2018). The biogeography of deep time phylogenetic reticulation. *Systematic Biology*, *67*(5), 743–744. https://doi.org/10.1093/sysbio/syy019

Cutter, A. D. (2013). Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. *Molecular Phylogenetics and Evolution*, *69*(3), 1172–1185. https://doi.org/10.1016/j.ympev.2013.06.006

Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Systematic Biology*, *67*(5), 786–799. https://doi.org/10.1093/sysbio/syy040

Dowling, T. E., & Secor, C. L. (1997). The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics*, *28*, 593–619. https://doi.org/10.1146/annurev.ecolsys.28.1.593

Eckert, A. J., & Carstens, B. C. (2008). Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. *Molecular Phylogenetics and Evolution*, *49*(3), 832–842. https://doi.org/10.1016/j.ympev.2008.09.008

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., … Mallet, J. (2019). Genomic architecture and introgression shape a butterfly radiation. *Science*, *366*(6465), 594–599. https://doi.org/10.1126/science.aaw2090

Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2020). A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution*, *37*(4), 1211–1223. https://doi.org/10.1093/molbev/msz296

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., … Besansky, N. J. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, *347*(6217), 1258524. https://doi.org/10.1126/science.1258524

Forsythe, E. S., Nelson, A. D. L., & Beilstein, M. A. (2020). Biased gene retention in the face of introgression obscures species relationships. *Genome Biology and Evolution*, *12*(9), 1646–1663. https://doi.org/10.1093/gbe/evaa149

Francis, A. R., & Steel, M. (2015). Which phylogenetic networks are merely trees with additional arcs? *Systematic Biology*, *64*(5), 768–777. https://doi.org/10.1093/sysbio/syv037

Geraldes, A., Basset, P., Smith, K. L., & Nachman, M. W. (2011). Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Molecular Ecology*, *20*(22), 4722–4736. https://doi.org/10.1111/j.1365-294X.2011.05285.x

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

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

Harrison, R. G., & Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, *105*(1), 795–809. https://doi.org/10.1093/jhered/esu033

Hibbins, M. S., & Hahn, M. W. (2019). The timing and direction of introgression under the multispecies network coalescent. *Genetics*, *211*(3), 1059–1073. https://doi.org/10.1534/genetics.118.301831

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, *37*(1), 203–217. https://doi.org/10.1111/j.1558-5646.1983.tb05528.x

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Huson, D. H., Rupp, R., & Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms, and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511974076

Jiao, X., Flouri, T., Rannala, B., & Yang, Z. (2020). The impact of cross-species gene flow on species tree estimation. *Systematic Biology*, *69*(5), 830–847. https://doi.org/10.1093/sysbio/syaa001

Jiao, X., & Yang, Z. (2021). Defining species when there is gene flow. *Systematic Biology*, *70*(1), 108–119. https://doi.org/10.1093/sysbio/syaa052

Leaché, A. D., Harris, R. B., Rannala, B., & Yang, Z. (2014). The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, *63*(1), 17–30. https://doi.org/10.1093/sysbio/syt049

Li, G., Figueiro, H. V., Eizirik, E., & Murphy, W. J. (2019). Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. *Molecular Biology and Evolution*, *36*(10), 2111–2126. https://doi.org/10.1093/molbev/msz139

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, *10*, 302. https://doi.org/10.1186/1471-2148-10-302

Long, C., & Kubatko, L. (2018). The effect of gene flow on coalescent-based species-tree inference. *Systematic Biology*, *67*(5), 770–785. https://doi.org/10.1093/sysbio/syy020

MacGuigan, D. J., & Near, T. J. (2019). Phylogenomic signatures of ancient introgression in a rogue lineage of darters (Teleostei: Percidae). *Systematic Biology*, *68*(2), 329–346. https://doi.org/10.1093/sysbio/syy074

Mallet, J., Besansky, N., & Hahn, M. W. (2016). How reticulated are species? *Bioessays*, *38*(2), 140–149. https://doi.org/10.1002/bies.201500149

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, *32*(1), 244–257. https://doi.org/10.1093/molbev/msu269

Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology*, *17*(2), e2006288. https://doi.org/10.1371/journal.pbio.2006288

Mendes, F. K., & Hahn, M. W. (2018). Why concatenation fails near the anomaly zone. *Systematic Biology*, *67*(1), 158–169. https://doi.org/10.1093/sysbio/syx063

Meng, C., & 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. https://doi.org/10.1016/j.tpb.2008.10.004

Mirarab, S., Bayzid, M. S., & Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, *65*(3), 366–380. https://doi.org/10.1093/sysbio/syu063

Molloy, E. K., Durvasula, A., & Sankararaman, S. (2021). Advancing admixture graph estimation via maximum likelihood network orientation. *Bioinformatics*, *37*(1), i142–i150. https://doi.org/10.1093/bioinformatics/btab267

Nelson, T. C., Stathos, A. M., Vanderpool, D. D., Finseth, F. R., Yuan, Y. W., & Fishman, L. (2021). Ancient and recent introgression shape the evolutionary history of pollinator adaptation and speciation in a model monkeyflower radiation (*Mimulus* section *Erythranthe*). *PLoS Genetics*, *17*(2), e1009095. https://doi.org/10.1371/journal.pgen.1009095

Nieto Feliner, G., Alvarez, I., Fuertes-Aguilar, J., Heuertz, M., Marques, I., Moharrek, F., … Villa-Machio, I. (2017). Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity*, *118*(6), 513–516. https://doi.org/10.1038/hdy.2017.7

O'Meara, B. C. (2012). Evolutionary inferences from phylogenies: a review of methods. *Annual Review of Ecology, Evolution, and Systematics*, *43*, 267–285. https://doi.org/10.1146/annurev-ecolsys-110411-160331

Pamilo, P., & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, *5*(5), 568–583. https://doi.org/10.1093/oxfordjournals.molbev.a040517

Pang, X. X., & Zhang, D. Y. (2023). Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. *Systematic Biology*, *72*(1), 35–49. https://doi.org/10.1093/sysbio/syac047

Pease, J. B., Haak, D. C., Hahn, M. W., & Moyle, L. C. (2016). Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology*, *14*(2), e1002379. https://doi.org/10.1371/journal.pbio.1002379

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: machine learning in python. *arXiv*, 1201.0490. https://doi.org/10.48550/arXiv.1201.0490

Philippe, H., & Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology*, *6*(5), 498–505. https://doi.org/10.1016/j.mib.2003.09.008

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Rhymer, J. M., & Simberloff, D. (1996). Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, *27*, 83–109. https://doi.org/10.1146/annurev.ecolsys.27.1.83

Schrider, D. R., Ayroles, J., Matute, D. R., & Kern, A. D. (2018). Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genetics*, *14*(4), e1007341. https://doi.org/10.1371/journal.pgen.1007341

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Schumer, M., Rosenthal, G. G., & Andolfatto, P. (2014). How common is homoploid hybrid speciation? *Evolution*, *68*(6), 1553–1560. https://doi.org/10.1111/evo.12399

Schumer, M., Rosenthal, G. G., & Andolfatto, P. (2018). What do we mean when we talk about hybrid speciation? *Heredity*, *120*(4), 379–382. https://doi.org/10.1038/s41437-017-0036-z

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., … Przeworski, M. (2018). Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, *360*(6389), 656–660. https://doi.org/10.1126/science.aar3684

Singhal, S., Derryberry, G. E., Bravo, G. A., Derryberry, E. P., Brumfield, R. T., & Harvey, M. G. (2021). The dynamics of introgression across an avian radiation. *Evolution Letters*, *5*(6), 568–581. https://doi.org/10.1002/evl3.256

Smith, M. L., & Hahn, M. W. (2023). Phylogenetic inference using generative adversarial networks. *Bioinformatics*, *39*(9), Article btad543. https://doi.org/10.1093/bioinformatics/btad543

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

Solís-Lemus, C., Yang, M., & Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology*, *65*(5), 843–851. https://doi.org/10.1093/sysbio/syw030

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, *105*(2), 437–460. https://doi.org/10.1093/genetics/105.2.437

Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology and Evolution*, *3*(2), 170–177. https://doi.org/10.1038/s41559-018-0777-y

Ting, C. T., Tsaur, S. C., & Wu, C. I. (2000). The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5313–5316. https://doi.org/10.1073/pnas.090541597

Tricou, T., Tannier, E., & de Vienne, D. M. (2022). Ghost lineages highly influence the interpretation of introgression tests. *Systematic Biology*, *71*(5), 1147–1158. https://doi.org/10.1093/sysbio/syac011

Vanderpool, D., Minh, B. Q., Lanfear, R., Hughes, D., Murali, S., Harris, R. A., … Hahn, M. W. (2020). Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biology*, *18*(12), e3000954. https://doi.org/10.1371/journal.pbio.3000954

Wang, R. J., & Hahn, M. W. (2018). Speciation genes are more likely to have discordant gene trees. *Evolution Letters*, *2*(4), 281–296. https://doi.org/10.1002/evl3.77

Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H. H., Mathieson, I., & Mathieson, S. (2021). Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, *21*(8), 2689–2705. https://doi.org/10.1111/1755-0998.13386

Wen, D., & Nakhleh, L. (2018). 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. https://doi.org/10.1093/sysbio/syy015

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), e1002660. https://doi.org/10.1371/journal.pgen.1002660

Yu, Y., Dong, J., Liu, K. J., & Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(46), 16448–16453. https://doi.org/10.1073/pnas.1407950111

Zachos, F. E. (2009). Gene trees and species trees - mutual influences and interdependences of population genetics and systematics. *Journal of Zoological Systematics and Evolutionary Research*, *47*(3), 209–218. https://doi.org/10.1111/j.1439-0469.2009.00541.x

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhu, J., Yu, Y., & Nakhleh, L. (2016). In the light of deep coalescence: revisiting trees within networks. *BMC Bioinformatics*, *17*(14), 415. https://doi.org/10.1186/s12859-016-1269-1