Peng, J., Kong, S., & Kubatko, L. (2024). A Likelihood Ratio Test for Hybridization Under the Multispecies Coalescent. *Bulletin of the Society of Systematic Biologists*. https://doi.org/10.18061/bssb.v3i2.9560

Applications

A Likelihood Ratio Test for Hybridization Under the Multispecies Coalescent

Jing Peng[1], Sungsik Kong[2,3], Laura Kubatko[3,4]

[1] Center for Biostatistics, The Ohio State University, [2] Wisconsin Institute for Discovery, University of Wisconsin-Madison, [3] Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, [4] Department of Statistics, The Ohio State University

https://doi.org/10.18061/bssb.v3i2.9560

Bulletin of the Society of Systematic Biologists

Abstract

Several methods have been developed to carry out a statistical test for hybridization at the species level, including the ABBA-BABA test and *HyDe*. Here, we propose a new method for detecting hybridization and quantifying the extent of hybridization. Our test computes the likelihood of a species tree that is possibly subject to hybridization using site pattern frequencies from genomic-scale datasets under the multispecies coalescent. To do this, we extend the calculation of the likelihood for site pattern frequency data for the 4-taxon symmetric and asymmetric species trees proposed in Chifman and Kubatko (2015) by incorporating an inheritance parameter, resulting in efficient computation of the likelihood under a scenario of hybridization. We use this likelihood computation to construct a likelihood ratio test that a given species is a hybrid of two parental species. Simulations demonstrate that our test is more powerful than existing tests of hybridization, including *HyDe*, and that it achieves the desired type I error rate. We apply the method to two empirical data sets, one for which hybridization is believed to have occurred and one for which previous methods have failed to detect hybridization.

1 Introduction

The deluge of genomic data available for phylogenetic study has confirmed the ubiquity of interspecific hybridization across the tree of life. Processes such as hybridization are ideally represented by phylogenetic networks, which generalize phylogenetic trees to include reticulate evolutionary events that allow the possibility that the ancestry of a species is derived from two (or more) independent evolutionary lineages. Phylogenetic networks can thus be used to model processes such as hybridization, horizontal gene transfer, gene duplication and loss, and recombination (Linder & Rieseberg, 2004; Nakhleh, 2010). Despite the development of innovative inference algorithms (e.g., Kong et al., 2024; Solís-Lemus & Ané, 2016; Than et al., 2008), estimating a phylogenetic network is a challenging task because the methods currently available often scale poorly and are presently limited to the analysis of relatively small data sets (Hejase & Liu, 2016). An alternative approach is to employ methods that detect hybrids among a large number of genomic sequences, rather than attempting to estimate the phylogenetic network directly.

Model-based population genetic clustering approaches are widely used to identify hybrid individuals from genetic data, and serve as an important tool for understanding patterns of extant genetic variation. Often implemented within the maximum likelihood (Alexander et al., 2009) or Bayesian frameworks (Pritchard et al., 2000), these methods estimate contributions from a user-designated number of ancestral genetic pools (typically denoted by κ) to an individual's ancestry through estimation of probabilistic quantities called ancestry coefficients. Despite the popularity of these methods in studies of hybridization on phylogenetic time scales, population clustering methods were not originally designed for this task but rather for the task of identifying population structure in contemporary populations. As a result, interpretation of the ancestry coefficients in a phylogenetic context is inevitably subjective and prone to mis- or over-interpretation of the historical processes, because different evolutionary scenarios can result in indistinguishable patterns (Anderson & Dunham, 2008; Barilani et al., 2007; Lawson et al., 2018). For example, gene flow is often conjectured to be responsible for an intermediate ancestry coefficient, even though incomplete lineage sorting (ILS) can lead to very similar patterns. These methods are also sensitive to the choice of markers, the level of genetic differentiation between populations, and the amount of data utilized (Kalinowski, 2011; Latch et al., 2006; Vähä & Primmer, 2005). The algorithms implemented in some of these tools involve unsupervised clustering, and simulations (e.g., *structure*, Pritchard et al., 2000) have demonstrated that these methods can produce different outcomes

in replicate analyses due to label-switching or multimodality (Kopelman et al., 2015). While the former issue can be detected and eliminated through post-processing, the latter issue is much more difficult to assess. Therefore, we have recently recommended that the use of such methods be curtailed in favor of newer methods that are specifically designed to detect hybridization across phylogenetic time scales (Kong & Kubatko, 2021).

Simple and intuitive approaches based on site pattern frequencies have quickly gained popularity in detecting hybridization from genomic datasets. Some of the widely used methods include the f_3 and f_4 statistics (Reich et al., 2009) and Patterson's D-statistic (Patterson et al., 2012), also known as the ABBA-BABA test (Durand et al., 2011; Green et al., 2010). Patterson's D-statistic is used as the basis of a statistical test for which rejection of the null hypothesis indicates a history of introgression among the input taxa. Recent work in this area (e.g., Hibbins & Hahn, 2019) has included further development of the methodology in an attempt to quantify the direction and proportional contributions of the taxa involved in the introgression event. Kubatko & Chifman (2019) propose a coalescent-based method that uses phylogenetic invariants for detecting species that have arisen via hybridization, implemented in the computer program *HyDe* (Blischak et al., 2018). Unlike methods based on D-statistic, this method is not limited to the examination of a single individual per population and it has been shown to detect populations that may have arisen via hybrid speciation as well as their putative parental populations with statistical power that is similar to the D-statistic. In addition, *HyDe* estimates the inheritance parameter (γ) that quantifies the proportion of genomic contribution of each parental taxon to the hybrid species. Kong & Kubatko (2021) found that the accuracy of the γ estimates in *HyDe* is superior to the estimates of ancestry coefficient in population clustering methods, even when the amount of ILS is high, though large sample sizes may be required when extensive ILS is present. Another method that has been widely used to quantify γ is the f_4-ratio statistic, although it has been found to be sensitive to violations of the underlying population model (Patterson et al., 2012).

In this study, we develop a method for detecting hybridization and quantifying the extent of hybridization by computing the likelihood of a species tree that is possibly subject to hybridization using site pattern frequencies from genomic-scale datasets under the multispecies coalescent (MSC). To do this, we extend the calculation of the likelihood for site pattern frequency data for the 4-taxon symmetric and asymmetric species trees proposed in Chifman & Kubatko (2015) by incorporating γ. Because our method uses site pattern frequencies calculated from either multilocus or single nucleotide polymorphism (SNP) data, the likelihood can be evaluated in a computationally efficient manner. We use these likelihood computations to construct a likelihood ratio test (LRT) that a given species is a hybrid of two parental species. We use simulation to demonstrate that our test is more powerful than existing tests of hybridization, including *HyDe*, and that it achieves the desired type I error. We apply the method to two empirical data sets, one for which hybridization is believed to have occurred and one for which previous methods have failed to detect hybridization.

2 Methods

2.1 Likelihood of a 4-taxon network

Consider the rooted, 4-taxon network S in Figure 1, where the outgroup population is O, the two parental populations are $P1$ and $P2$, and H is the hybrid population with inheritance parameter γ. The species network S can be decomposed into two parental trees, S1 and S2, where the former can be obtained by removing the reticulation edge between $P1$ and H and the latter can be obtained by removing the edge between H and $P2$. In this case, sequences are assumed to evolve through gene trees, which arise either from species tree S1 with H as a sister taxon of $P2$ with probability γ, or from S2 with H as a sister taxon of $P1$ with probability $1-\gamma$. Note that we can summarize sequence data from the network S as site patterns. For example, a site pattern AGCC represents a position in the alignment for which species O, $P1$, H and $P2$ have nucleotides A, G, C and C, respectively. In a 4-taxon network, there are $4^4 = 256$ possible site patterns.

To define the site pattern probabilities, consider a 4-taxon species tree with species a, b, c, and d, and let $i_j \in \{A, C, G, T\}$ refer to the nucleotide observed for taxon j at the particular site under consideration. We refer to $i_a i_b i_c i_d$ as a site pattern for the species tree, and denote the probability of this site pattern by $p_{i_a i_b i_c i_d}$. Chifman & Kubatko (2015) derived explicit expressions for the site pattern probabilities under the multispecies coalescent model with the JC69 (Jukes & Cantor, 1969) substitution model and the assumption of a constant effective population size parameter θ with branch lengths $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$ in coalescent units. Under this model and the molecular clock assumption, the rooted symmetric 4-leaf species tree ((*a*, *b*), (*c*, *d*)); will have 9 distinct site patterns probabilities

$$\begin{aligned} &p_{xxxx}, \quad p_{xxxy} = p_{xxyx}, \quad p_{xyxx} = p_{yxxx}, \\ &p_{xyxy} = p_{yxxy}, \quad p_{xxyy}, \\ &p_{xyxz} = p_{yxxz} = p_{xyzx} = p_{yxzx} \\ &p_{xxyz}, \quad p_{yzxx}, \quad p_{xyzw}, \end{aligned}$$

while the rooted asymmetric 4-leaf species tree (*a*, (*b*, (*c*, *d*))); will have 11 distinct site patterns probabilities

$$\begin{aligned} &p_{xxxx}, \quad p_{xxxy} = p_{xxyx}, \quad p_{xyxx}, \quad p_{yxxx}, \\ &p_{xyxy} = p_{yxxy}, \quad p_{xxyy}, \quad p_{xyxz} = p_{xyzx}, \\ &p_{yxxz} = p_{yxzx}, \quad p_{xxyz}, \quad p_{yzxx}, \quad p_{xyzw}, \end{aligned}$$

where x, y, z and w denote different nucleotide states. Therefore, for the network S in Figure 1, we will have 15 site pattern probabilities, with each one being a weighted average of the probabilities from species trees S1 and S2. For example, the probability of site pattern $xxxy$ from the network S is

$$\begin{aligned} p_2 := p_{xxxy} &= \gamma p_{1xxxy} + (1-\gamma) p_{2xxxy} \\ &= \gamma p_{1xxxy} + (1-\gamma) p_{1xyxx}, \end{aligned}$$

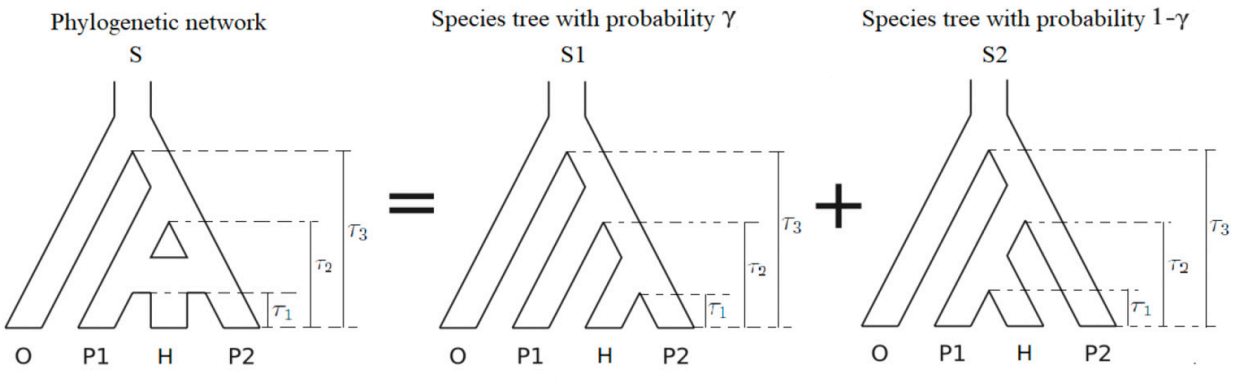

Figure 1. A rooted, 4-taxon network S with three edge (or branch) length parameters τ_1, τ_2, and τ_3. The inheritance parameter γ (and $1-\gamma$) can be used with the site pattern probabilities under species trees S1 and S2 (i.e., the parental trees of S) to derive site pattern probabilities arising from the network S.

where p_{1xxxy} and p_{2xxxy} are the probabilities of site pattern $xxxy$ from species trees S1 and S2, respectively. Similarly, we can write the other 14 site pattern probabilities

$$
\begin{aligned}
p_1 &:= p_{xxxx} = p_{1xxxx},\\
p_3 &:= p_{xxyx} = p_{1xxyx},\\
p_4 &:= p_{xxyy} = \gamma p_{1xxyy} + (1-\gamma) p_{1xyyx},\\
p_5 &:= p_{xxyz} = \gamma p_{1xxyz} + (1-\gamma) p_{1xzyx},\\
p_6 &:= p_{xyxx} = \gamma p_{1xyxx} + (1-\gamma) p_{1xxxy},\\
p_7 &:= p_{xyxy} = p_{1xyxy},\\
p_8 &:= p_{xyxz} = p_{1xyxz},\\
p_9 &:= p_{xyyx} = \gamma p_{1xyyx} + (1-\gamma) p_{1xxyy},\\
p_{10} &:= p_{yxxx} = p_{1yxxx},\\
p_{11} &:= p_{xyyz} = \gamma p_{1xyyz} + (1-\gamma) p_{1xzyy},\\
p_{12} &:= p_{zxyz} = \gamma p_{1zxyz} + (1-\gamma) p_{1zzyx},\\
p_{13} &:= p_{yxzx} = p_{1yxzx},\\
p_{14} &:= p_{yzxx} = \gamma p_{1yzxx} + (1-\gamma) p_{1yxxz},\\
p_{15} &:= p_{xyzw} = p_{1xyzw}.
\end{aligned}
$$

From the data, we observe the frequency with which site pattern k, $k = 1, 2, \cdots, 15$ is observed, which we denote d_k, in a sample of n sites. The entire data are then denoted by the vector $D = (d_1, \ldots, d_{15})$. The likelihood of network-associated parameters (τ_1, τ_2, τ_3, γ and θ) given network topology S is then given by

$$
\begin{aligned}
&\mathcal{L}(\tau_1, \tau_2, \tau_3, \gamma, \theta | D, S) = p_1^{d_1} p_2^{d_2} \cdots p_{15}^{d_{15}},\\
&with \sum_{k=1}^{15} d_k = M,
\end{aligned}
$$

where M is the total number of sites.

2.2 Likelihood ratio test of hybridization

Consider the 4-taxon species tree S2 in Figure 1, and note that a value of $\gamma = 0$ implies the absence of genetic contribution from $P2$ to H following the divergence of $P2$ from the ancestral species of $P1$ and H (i.e., no hybridization). We can thus develop a formal statistical hypothesis test for hybridization between species H and $P2$ by considering the following hypotheses

$$H_0 : \gamma = 0 \quad \text{vs.} \quad H_A : \gamma \in (0, 0.5] \tag{1}$$

Rejection of the null hypothesis above indicates support for the reticulation event between H and $P2$ for the tree S2, and H can be considered to be a hybrid of parental species $P1$ and $P2$. A failure to reject H_0 implies that there is not sufficient evidence from the data to prefer the network S over the species tree S2. Note that this test requires the identification of the "major tree" (i.e., the tree that specifies the species with which taxon H shares the majority of its ancestry – here, tree S1), and we test whether a minor portion of the ancestry of H is derived from a second identified species (here, species $P2$). We discuss this requirement in the Discussion section below.

We propose to test the hypothesis above using a LRT, which has well-established statistical properties in general contexts (Wilks, 1938). The test statistic is given b

$$\lambda_{LR} = -2 \ln \left[\frac{\sup_{\boldsymbol{\delta} \in \Omega_0} \mathcal{L}(\boldsymbol{\delta})}{\sup_{\boldsymbol{\delta} \in \Omega} \mathcal{L}(\boldsymbol{\delta})} \right],$$

where $\boldsymbol{\delta}$ is the vector of parameters, i.e., $\boldsymbol{\delta} = (\tau_1, \tau_2, \tau_3, \theta, \gamma)$, and Ω is the parameter space for the network model. Specifically, Ω is a 5-dimensional space defined by $\tau_1 \in (0, +\infty)$, $\tau_2 \in (\tau_1, \tau_3)$, $\tau_3 \in (\tau_2, +\infty)$, $\theta \in (0, +\infty)$, $\gamma \in [0, 0.5]$. The space Ω_0 is the subset of Ω defined by the null hypothesis, i.e., Ω_0 is the 4-dimensional subspace obtained by fixing $\gamma = 0$. Standard statistical theory can be applied to see that the asymptotic distribution of λ_{LR} under the null hypothesis is a 50:50 mixture of a χ^2_1 distribution and a point mass at 0 since the value of γ under the null hypothesis lies on the boundary of the parameter space (Self & Liang, 1987). Thus, the hypothesis test at level α can be carried out by comparing λ_{LR} with the $1-\alpha$ quantile of this mixture distribution.

We note here that to compute λ_{LR}, we have to find the value of $\boldsymbol{\delta}$ that maximizes the likelihood over both Ω and Ω_0. Because this optimization procedure is constrained by the bounds on the branch lengths, the population size parameter and the hybridization parameter, we use the following reparameterization, to develop an efficient optimization algorithm:

$$\tau_1' = log\left(\frac{\tau_1/\tau_2}{1+\tau_1/\tau_2}\right)$$
$$\tau_2' = log\left(\frac{\tau_2/\tau_3}{1+\tau_2/\tau_3}\right)$$
$$\tau_3' = log(\tau_3)$$
$$\theta' = log(\theta)$$
$$\gamma' = log\left(\frac{2\gamma}{1+2\gamma}\right).$$

Defining $\boldsymbol{\delta}'$ as the vector of transformed parameters, we must find the values of $\boldsymbol{\delta}' = (\tau_1', \tau_2', \tau_3', \theta', \gamma') \in \Omega' = R^5$ and $\boldsymbol{\delta}_0' = (\tau_1', \tau_2', \tau_3', \theta') \in \Omega_0' = R^4$ that maximize the likelihood. With this transformation, we perform a multidimensional optimization simultaneously for all parameters by applying the quasi-Newton method (BFGS) (Byrd et al., 1995; Fletcher & Reeves, 1964) for unconstrained multidimensional optimization, which uses function values and gradients to search parameter space. The BFGS method has better computational complexity than Newton's method, and because it uses an approximation of the gradient to carry out the search, it is expected to be more computationally efficient than gradient-free methods. However, because of correlation between θ and the τs, the density is relatively flat for values of θ, which sometimes causes the BFGS optimization process to terminate prematurely. We have noticed that a crucial step in developing a good implementation of this method is the selection of a good starting point, especially for the population size parameter θ.

We thus obtain a starting point, θ_0, by first setting a small lower bound (10^{-5} in our case), and then increasing it until at least one of the branch length moment estimates is smaller than 0. Using this value as the upper bound, we then find an initial interval $[a, b]$ for θ. This is a very wide interval, so we then implement a golden section search (Gill et al., 1981) to get a tighter interval for θ_0. The disadvantage of golden section search is its slow convergence, so we set a large stopping tolerance, terminating the search once $b - a < 0.01$. Constrained in this updated interval $[a_1, b_1]$, we finally use one-dimensional Brent optimization (Brent, 1973), and the optimal value is used as θ_0. Brent optimization can achieve superlinear convergence via a combination of golden section and parabolic interpolation steps. This procedure was motivated by the work of Peng et al. (2022) (see their Supplemental Information for details).

For the branch lengths and the inheritance probability, we can simply use the moment estimators by adapting results in Kubatko et al. (2024) for networks. By solving the equation

$$p_i(\boldsymbol{\delta}) = d_i/M, \;\; i = 1, 2, \ldots, 15,$$

where M is the total number of sites, we obtain the following moment estimator for the branch lengths given θ_0

$$e^{-\mu\theta_0\tau_{0j}} = \frac{4(1+\mu\theta_0)}{72M} \times C_j' D, j = 1, 2, 3$$

where

$$C_1' = (18, 20, -34, 16, -8, 20, -26, -2, 16, 18, 18, -8, -30, 18, 6)$$
$$C_2' = (18, -4, 14, -8, -8, -4, 22, -2, -8, 18, -6, -8, 18, -6, -6)$$
$$C_3' = (18, 10, 10, 2, 2, 10, 2, 2, 2, -6, -6, 2, -6, -6, -6)$$

and μ is set to be 4/3 for the JC69 model. The moment estimator of γ can be obtained from any phylogenetic invariant; we use

$$\gamma_0 = \frac{d_4 - d_7}{d_4 + d_9 - 2d_7}.$$

2.3 Simulation study

We first use simulations to assess the type I error and statistical power of testing hybridization using our method (i.e., LRT), *HyDe*, and the ABBA-BABA test. Coalescent Independent Site (CIS) data is a natural fit for our method, because nucleotides are unlinked and assumed to arise from the coalescent model independently. Even though we are not generating CIS data in practice, these data are useful for verifying our method and theory under the correct model. To further simulate the performance of our method in application, we also implement this approach on multilocus data, in which all sites within a given locus are assumed to have evolved on the same genealogy and are not independent. A theoretical justification for this application can be found in Wascher and Kubatko (2021), which argues that methods developed for CIS data can also be applied to multilocus data, and we therefore consider both data types here.

To examine the performance of the three tests, we simulated two types of data: (1) unlinked CIS data (each site evolves on its own own tree drawn randomly from the distribution of gene trees expected for the true simulation parameters under the MSC model), and (2) multilocus data (a sequence of length l is simulated for each locus on an underlying gene tree drawn randomly from the expected gene tree distribution). The simulations were performed as follows:

1. Use **ms** (Hudson, 2002) to generate $N * \gamma$ gene tree samples under the MSC model based on the parental tree S1 and $N * (1 - \gamma)$ gene tree samples based on the parental tree S2 in Figure 1;
2. Use **seq-gen** (Rambaut & Grass, 1997) to generate DNA sequences of length l for each gene tree under the specified nucleotide substitution model ($l = 1$ for CIS data);
3. Count the 15 site pattern frequencies;
4. Compute the test statistics for LRT, *HyDe* and ABBA-BABA methods, and test the hypothesis (1).
5. Repeat steps 1–4 W times to obtain type I error and statistical power for the three methods.

All steps in the simulations were carried out in the R statistical software (R Core Team, 2018). In step 1, time is measured in coalescent units (number of generations scaled by $2N_e$, where N_e is the effective population size), and we set the population size parameter $\theta = 4N_e\mu = 0.002$ (constant throughout the tree), where μ is the mutation rate. For the speciation times in Figure 1, we assigned the vector $(\tau_1, \tau_2, \tau_3) = b \cdot (0.25, 0.5, 1.0)$, where different choices of b then involve stretching or shrinking the network; any choice of b results in trees that satisfy the molecular clock. We considered two choices for b: b = 1 and 2 to represent short and long branch trees, respectively. The hybridization

parameter γ is chosen to be 0 or to vary from 0.06 to 0.5 by 0.02. For CIS data, we simulate $N = 100K, 250K, 500K$ and $1M$ genes with one DNA site for each, while for multilocus data, we simulate $N = 1K, 2.5K, 5K$ and $10K$ genes with length $l = 100$ in step 2. In that case, we have the same length of simulated DNA alignments for CIS and multilocus data. In steps 5, we chose W=500.

Though the LRT and *HyDe* are both derived under the JC69 model, DNA sequence may evolve under more complex models. To evaluate performance of the three methods under more complex substitution models, we examined three additional scenarios for the substitution model used to generate the data. First, we used the general time reversible model (GTR) with substitution rates 1.0, 0.2, 2.5, 0.75, 3.2, 1.6 and base frequencies $\pi_A = 0.22, \pi_C = 0.28, \pi_G = 0.22$, and $\pi_T = 0.28$. Next, in order to mimic an empirical scenario, we selected the substitution model for the ATP gene from the dataset of Gerard et al. (2011), and use the estimated parameters for simulations in step 2. The model selected by AICc and BIC in PAUP* (Swofford, 2024) is the HKY85 model (Hasegawa et al., 1985) with nucleotide frequencies $A = 0.35, C = 0.32, G = 0.09$ and $T = 0.24$. The estimated transition/ transversion ratio is 4.97 with proportion of invariable sites 0.52 (labeled HKY+I hereafter). To further explore the effect of rate variation, we also simulated under this model with gamma distributed rates with $\alpha = 0.9$ (labeled HKY+I+Γ hereafter) added.

2.4 Application to real data

2.4.1 *Heliconius* butterflies

DNA sequence data from Martin et al. (2013) were downloaded for four populations of *Heliconius* butterflies (248,822,400 sites; available on Dryad). We selected a single individual from four populations, each of which represents a distinct species: *Heliconius melpomene rosina*, *H. m. timareta*, *H. cydno*, and the outgroup, *H. hecale*. Significant hybridization was detected in *H. cydno* at the population level in Martin et al. (2013) and using *HyDe* in Blischak et al. (2018) where they used multiple individuals per population. For counting the 15 site pattern frequencies, we only include sites with explicit nucleotides for all the species. Therefore, we have 128,321,514 sites for the four populations of *Heliconius* butterflies in the final analysis.

2.4.2 *Sistrurus* rattlesnakes

We applied our LRT to examine whether populations of *Sistrurus catenatus* rattlesnakes in northwest and central Missouri are of hybrid origin. These populations are found to include individuals with morphological characteristics intermediate between *S. c. catenatus* and *S. c. tergeminus* (Evans & Gloyd, 1948; Gloyd, 1940). While a hypothesis of hybridization is plausible based on morphological similarity, it is also possible that this similarity is due to evolutionary or ecological factors. Gibbs et al. (2011) did not find evidence of hybridization between *S. c. catenatus* and *S. c. tergeminus* based on microsatellite and mitochondrial markers, indicating that the individuals in Missouri were *S. c. tergeminus*. Gerard et al. (2011) developed and used a likelihood ratio test based on observed gene tree distributions to analyze these data, and also found no genetic evidence of hybridization between *S. c. catenatus* and *S. c. tergeminus*.

We used the dataset of Gerard et al. (2011) to examine this question. The data consist of twelve genes (A, ATP, 1, 4, 11, 25, 31, 41, 61, 63, ETS, and GAPD) analyzed by Kubatko et al. (2011) (see their Table 2). The original dataset includes fourteen individuals: four of *S. c. catenatus*, four of *S. c. tergeminus*, four of the individuals in the putative hybrid zone, and two from outgroup populations of *Agkistrodon contortix* and *A. piscivorus*. We selected a single individual from the three *Sistrurus* populations and one outgroup species. After excluding ambiguous sites, the dataset included 4 individuals and 7,663 sites.

3 Results

3.1 Simulation study

We plot test sizes for testing hybridization using the ABBA-BABA test, *HyDe*, and the LRT under JC69, HKY+I, and HKY+I+Γ for the short and long branch trees for different values of γ between 0 and 0.5. (see the Supplemental Material for figures under all of the simulation settings). As a representative example, Figures 2 and 3 show plots of the test results for the short and long branch trees with sequence length of $100K$. From these plots, we observe that all methods exhibit increased power as γ gets closer to 0.5, a pattern that becomes more prominent with an increase in dataset size. Comparing the three methods, we see that the LRT is more powerful in all cases. Even when the nucleotide substitution model is misspecified, type I error rates were reasonably controlled by the LRT for CIS data, while *HyDe* tends to be a conservative test in some cases (see Figure 3 and Supplemental Material Sections S1 and S2).

For multilocus data, the type I error is a little inflated in most cases for the long branch tree and the two cases with smaller data sizes for the short branch tree (see Supplemental Material Sections S3.1 and S3.2, respectively.). In these cases, however, the ABBA-BABA test and *HyDe* have similar problems. This may be because the limited number of genes is not enough to fully characterize the gene tree distribution under the coalescent model. Although the statistical power is low overall for the short branch tree, we do not see much difference in the power or type I error for multilocus data with the same length for all methods in comparison with unlinked sites data. Similar patterns are observed when the number of loci is increased to 250K, 500K and 1M. As expected, the power for all three methods approaches 100% when the number of sites is 1M (see the Supplemental Material).

3.2 Empirical datasets

Table 1 shows the results of testing hybridization for the empirical data sets using the LRT, *HyDe* and the ABBA-

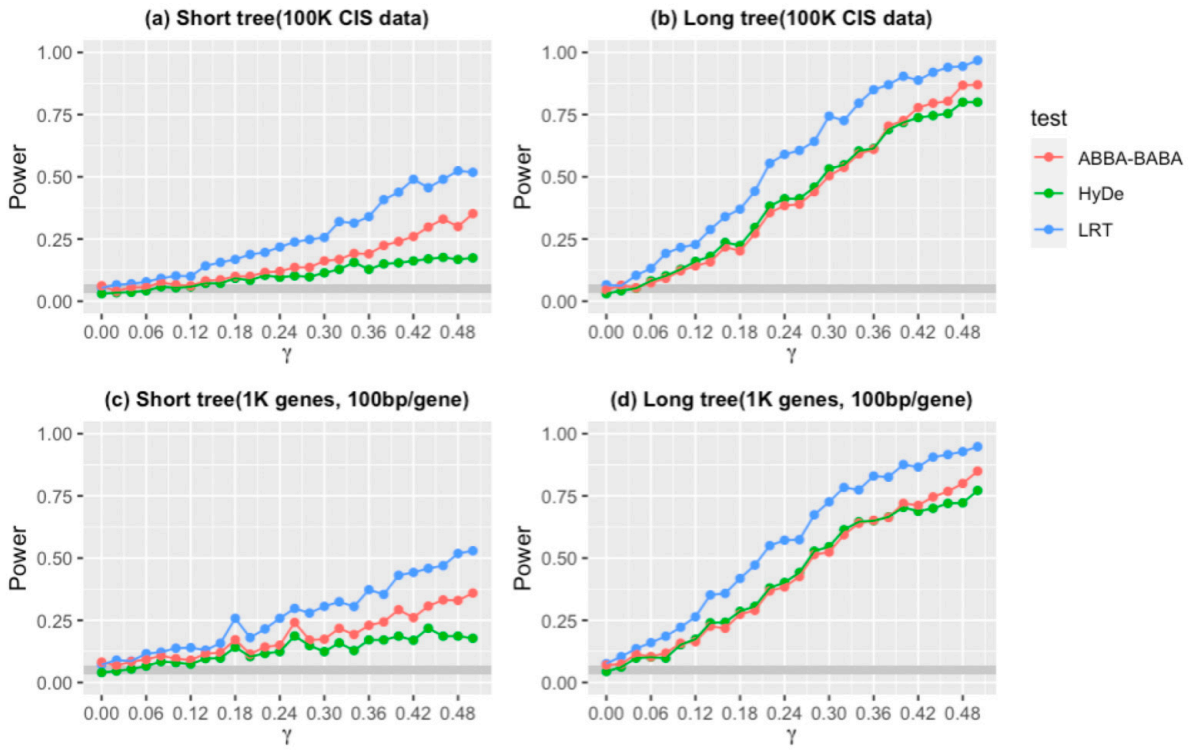

Figure 2. Hybrid detection power under JC69 in the ABBA-BABA test, $HyDe$, and the LRT for differing inheritance parameter values (γ) with sequence length 100 K . The shaded area gives the expected acceptance region $(0.031, 0.069)$ of the empirical type I error rate. (a) 500 CIS datasets from short branch tree, with type I error rates 0.054 (LRT), 0.030 ($HyDe$) and 0.062 (ABBA-BABA). (b) 500 CIS datasets from long branch tree, with type I error rates 0.066 (LRT), 0.030 ($HyDe$) and 0.046 (ABBA-BABA). (c) 500 multilocus datasets from short branch tree, with type I error rates 0.072 (LRT), 0.040 ($HyDe$) and 0.082 (ABBA-BABA). (d) 500 multilocus datasets from long branch tree, with type I error rates 0.076 (LRT), 0.044 ($HyDe$) and 0.070 (ABBA-BABA).

BABA test. All three methods detect hybridization for the *Heliconius* data set with p-values less than 0.0001, consistent with previous work. For the *Sistrurus* rattlesnakes, all three methods fail to reject the null hypothesis of no hybridization at $\alpha = 0.05$, again consistent with previous studies. Table 1 also shows the estimates of the inheritance parameter ($\hat{\gamma}$) from *HyDe* and the LRT, which can be used to help decide if a network is preferable to a binary tree as a representation of the speciation history of a group. For example, the estimates of γ for the *Heliconius* data set are very close to 0.5 with small p-values, strongly suggesting a role for hybridization in the history of *H. cydno*. Conversely, for the *Sistrurus* data set, the estimates of γ are less than 0.1, suggesting little to no gene flow following speciation.

4 Discussion

In this article, we develop the likelihood for a 4-taxon network under JC69 and the multispecies coalescent model. Based on that, we propose a likelihood ratio test for hybridization given a 4-taxon species tree topology. Simulation studies demonstrate that our method achieves higher statistical power than the other two popular methods of testing hybridization, *HyDe* and the ABBA-BABA test, with reasonable type I error when sequence data are simulated under JC69, HKY+I, and HKY+I+Γ. The increase in power is especially evident when the number of independent sites is limited and the species are recently diverged. Our simulations demonstrate that the test performs well for both coalescent independent sites and for multilocus data without much difference in power given the same overall sequence length. We used two empirical data sets to highlight the performance of the method in practice. We note that for all of the methods examined, the type I error may be inflated when the number of genes is limited. Thus, we encourage users to consider the estimate of γ when drawing conclusions. From the biological perspective, an estimated value of the hybridization parameter close to 0 is likely to indicate a lack of signal for hybridization. From a model selection perspective, we may not want to consider a network over a binary tree in such cases.

A primary innovation of our method is that we are more likely to detect the hybridization with less data than we are using current methods. In addition, the method provides maximum likelihood estimates (MLEs) for all the branch lengths, for the population size parameter, and for the hybridization parameter. That means the estimates share all the desirable statistical properties of MLEs, like consistency, asymptotically normality, and asymptotic efficiency. We are currently working to extend this method to larger networks.

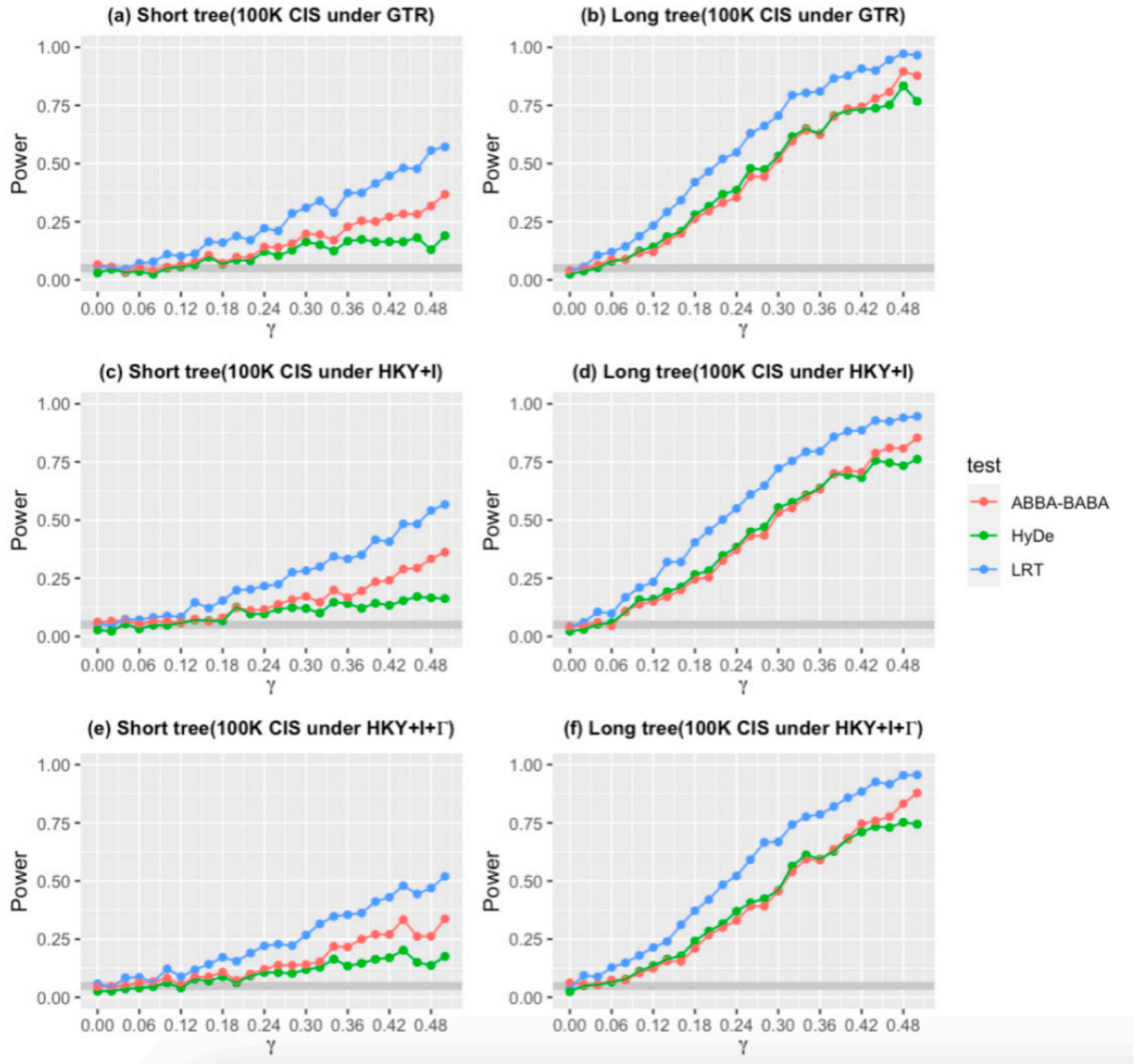

Figure 3. Hybrid detection power under GTR, HKY+I, and HKY+I $+\Gamma$ in the ABBA-BABA test, *HyDe*, and the LRT in different inheritance parameter values (γ). The shaded area gives the expected acceptance region $(0.031, 0.069)$ of the empirical type I error rate. (a) Type I error rates are 0.058 (LRT), 0.030 ($HyDe$) and 0.066 (ABBA-BABA). (b) Type I error rates are 0.040 (LRT), 0.024 ($HyDe$) and 0.040 (ABBA-BABA). (c) Type I error rates are 0.054 (LRT), 0.028 ($HyDe$) and 0.062 (ABBA-BABA). (d) Type I error rates are 0.044 (LRT), 0.022($HyDe$) and 0.042 (ABBA-BABA). (e) Type I error rates are 0.058 (LRT), 0.026 ($HyDe$) and 0.046 (ABBA-BABA). (f) Type I error rates are 0.040 (LRT), 0.024 ($HyDe$) and 0.062 (ABBA-BABA).

Table 1. Results of testing hybridization for the empirical dataset $\hat{\gamma}$ is the estimate of inheritance parameter γ.

Datasets	# of sites	LRT		*HyDe*		ABBA-BABA test
		p-value	$\hat{\gamma}$	p-value	$\hat{\gamma}$	p-value
Heliconius butterflies	128,321,514	< 0.0001	0.493	< 0.0001	0.415	< 0.0001
Sistrurus rattlesnakes	7,663	0.064	0.066	0.380	0.036	0.382

The assumptions that (1) nucleotide sites evolve according to the JC69 substitution model and (2) effective population sizes are constant throughout the tree permit the use of formulas from Chifman and Kubatko (2015) for computing the 15 site pattern probabilities. Empirical data, however, may evolve under a nucleotide substitution model more complex than JC69. Our simulation studies indicate that our method is robust to data arising from the GTR model, though we did not exhaustively check all possible substitution rate matrices. When the true nucleotide sub-

stitution model differs substantially from JC69, our method may lead to inflated type I error. In these cases, the estimate of the hybridization parameter may provide a meaningful complement to the p-value in terms of biological interpretation. We are currently investigating approaches for extending our method to larger networks with multiple hybridization events. In that case, *HyDe* has multiple testing problems, while our method provides the possibility of using a likelihood-related score for model selection.

As we mention above, our test also assumes that the "major tree" (i.e., the tree that identifies both putative parents of the hybrid taxon, as well as which parent has contributed the majority of the hybrid's genome) has been identified prior to application of the test. This allows us to formulate the test in the formal statistical framework of a likelihood ratio test, which then provides desirable statistical properties for the test. In practice, however, the major tree may be unknown, and previous work (e.g., Leaché et al. (2014)) suggests that inference of the major tree can sometimes be biased by even small amounts of gene flow. When there is uncertainty in the major tree, the method of Haque and Kubatko (2024) could be combined with the test we propose here to carry out multiple tests with differing assumptions for the major tree to provide a global test that incorporates an appropriate correction for multiple testing.

An additional assumption is that within the major tree, the species that contributes the majority of the genomic information to the hybrid species is identified (i.e., the alternative hypothesis is $\gamma \in (0, 0.5]$). The primary reason for this assumption is that restricting γ to lie in $[0, 0.5]$ allows us to identify the null distribution as a 50:50 mixture of a χ^2_1 distribution and a point mass at 0. Expanding the possible interval for γ to $[0, 1]$ would necessitate a composite null hypothesis (i.e., $\gamma = 0$ and $\gamma = 1$ would both correspond to no hybridization) and the null distribution would be a complicated multi-component mixture distribution as a result. Additionally, the test in its present form is comparable to existing tests (e.g., *HyDe* and the ABBA-BABA test). If the major tree topology can be identified but the "major parental species" is uncertain, the test could be run twice with each parental species designated to be the "major parent" and the correction of Haque and Kubatko (2024) (or even a simple Bonferroni correction, which will be conservative) could be applied.

Justison et al. (2023) present three macroevolutionary patterns of hybridization, namely lineage generative, neutral, and degenerative hybridization. The three scenarios differ in the change in the number of lineages before and after the hybridization event. As shown in Figure 1, our method assumes a lineage-generative hybridization scenario in which a reticulation event results in the gain of a lineage, and both parental lineages continue to exist to the present. However, it is important to stress that our method should be applicable to lineage neutral and degenerative hybridization as well. Kong & Kubatko (2021) showed that *HyDe* and the ABBA-BABA test, which assume lineage generative and neutral scenarios, respectively, can detect hybridization reliably even when the true scenario contradicts the assumed model. In particular, *HyDe* is robust when the node ages of two parental lineages (i.e., τ_1 in S1 and S2 in Figure 1) differ, allowing us to expect a similar pattern in the proposed method. Nevertheless, our method may not perform desirably under some common, but extreme, biological scenarios, like an excessively high amount of ILS in the data, incomplete taxon sampling, and/or the presence of "ghost" (i.e., extinct, unknown, or unsampled) lineages that played a role in hybridization. Further exploration of performance under these scenarios is needed.

Overlooking the existence of ghost lineages, especially when these lineages have played a role in past hybridization events, can result in coalescent-based hybrid detection methods producing erroneous results. While we have not explicitly evaluated the performance of the proposed method under such scenarios in this study, it is predictable that ghost lineages would pose difficulty in detecting hybridization since our model does not account for such phenomenon. A number of recent studies show the negative influence of ghost lineages on the performance of hybrid detection methods. For example, using the D_3 method (Hahn & Hibbins, 2019), Tricou, Tannier, & De Vienne (2022) show that the detection of hybridization in published studies can be veiled or even reversed when ghost lineages outnumber the sampled lineages. Tricou, Tannier, & Vienne (2022) also show that the ABBA-BABA test can misidentify the parental lineages (donors) and hybrid lineage (recipient) when ghost lineages are not taken into account. Similarly, Bjørner et al. (2023) report reduced precision and power of *HyDe* in the presence of ghost lineages. Furthermore, Pang & Zhang (2024) report reduced power in popularly used population clustering methods (e.g., *structure*) in the same biological context, although Kong & Kubatko (2021) previously demonstrated their incompetency in detecting hybrids in general.

Code and functions that were used to carry out the simulation study and empirical analysis are presented in the Supplemental Material. They can also be obtained by contacting author Jing Peng at cathelena03@gmail.com.

Funding

The authors have no funding to acknowledge.

Supporting Information

Supplemental Material is available at https://github.com/lkubatko/LRT.

Submitted: March 10, 2024 EST, Accepted: September 30, 2024 EST

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Anderson, E. C., & Dunham, K. K. (2008). The influence of family groups on inferences made with the program Structure. *Molecular Ecology Resources*, *8*(6), 1219–1229. https://doi.org/10.1111/j.1755-0998.2008.02355.x

Barilani, M., Sfougaris, A., Giannakopoulos, A., Mucci, N., Tabarroni, C., & Randi, E. (2007). Detecting introgressive hybridisation in rock partridge populations (*Alectoris Graeca*) in Greece through Bayesian admixture analyses of multilocus genotypes. *Conservation Genetics*, *8*(2), 343–354. https://doi.org/10.1007/s10592-006-9174-1

Bjørner, M., Molloy, E. K., Dewey, C. N., & Solís-Lemus, C. (2023). Detectability of varied hybridization scenarios using genome-scale hybrid detection methods. *arXiv*. https://doi.org/10.48550/arXiv.2211.00712

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

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Prentice-Hall.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.2172/204262

Chifman, J., & Kubatko, L. (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, *374*, 35–47. https://doi.org/10.1016/j.jtbi.2015.03.006

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Evans, P. D., & Gloyd, H. K. (1948). The subspecies of the massasauga, *Sistruus catenatus*, in Missouri. *Bull. Chicago Acad. Sci.*, *8*, 225–232.

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, *7*, 149–154. https://doi.org/10.1093/comjnl/7.2.149

Gerard, D., Gibbs, H. L., & Kubatko, L. (2011). Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evolutionary Biology*, *11*, 291. https://doi.org/10.1186/1471-2148-11-291

Gibbs, H. L., Murphy, M., & Chiucchi, J. E. (2011). Genetic identity of endangered massasauga rattlesnakes (*Sistrurus* sp.) in Missouri. *Conservation Genetics*, *12*, 433–439. https://doi.org/10.1007/s10592-010-0151-3

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical optimization*. Academic Press.

Gloyd, H. K. (1940). The rattlesnakes, genera *Sistrurus* and *Crotalus*. *Special Publ. Chicago Acad. Sci.*, *4*, 104–118.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., & others. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

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

Haque, M. R., & Kubatko, L. (2024). A global test of hybrid ancestry from genome-scale data. *Statistical Applications in Genetics and Molecular Biology*, *23*(1), 20220061. https://doi.org/10.1515/sagmb-2022-0061

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*, 160–174. https://doi.org/10.1007/BF02101694

Hejase, H. A., & Liu, K. J. (2016). A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics*, *17*(1), 422. https://doi.org/10.1186/s12859-016-1277-1

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. (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

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, *3*, 21–132. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7

Justison, J. A., Solís-Lemus, C., & Heath, T. A. (2023). SiPhyNetwork: An R package for simulating phylogenetic networks. *Methods in Ecology and Evolution*, *14*(7), 1687–1698. https://doi.org/10.1111/2041-210X.14116

Kalinowski, S. T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: Simulations and implications for human population structure. *Heredity*, *106*(4), 625–632. https://doi.org/10.1038/hdy.2010.95

Kong, S., & Kubatko, L. (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

Kong, S., Swofford, D. L., & Kubatko, L. (2024). Inference of phylogenetic networks from sequence data using composite likelihood. *Systematic Biology, to Appear*. https://doi.org/10.1101/2022.11.14.516468

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across *K*. *Molecular Ecology Resources*, *15*(5), 1179–1191. https://doi.org/10.1111/1755-0998.12387

Kubatko, L., & 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

Kubatko, L., Gibbs, H. L., & Bloomquist, E. W. (2011). Inferring species-level phylogenies and taxonomic distinctiveness using multi-locus data in *Sistrurus* rattlesnakes. *Systematic Biology*, *60*(4), 393–409. https://doi.org/10.1093/sysbio/syr011

Kubatko, L., Leonard, A., & Chifman, J. (2024). Identifiability of speciation times under the multispecies coalescent. *Journal of Theoretical Biology*, *595*, 111927. https://doi.org/10.1016/j.jtbi.2024.111927

Latch, E. K., Dharmarajan, G., Glaubitz, J. C., & Rhodes, O. E. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, *7*(2), 295–302. https://doi.org/10.1007/s10592-005-9098-1

Lawson, D. J., Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 3258. https://doi.org/10.1038/s41467-018-05257-7

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

Linder, C. R., & Rieseberg, L. H. (2004). Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany*, *91*(10), 1700–1708. https://doi.org/10.3732/ajb.91.10.1700

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., & Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, *23*, 1817–1828. https://doi.org/10.1101/gr.159426.11

Nakhleh, L. (2010). Evolutionary phylogenetic networks: Models and issues. In L. S. Heath & N. Ramakrishnan (Eds.), *Problem solving handbook in computational biology and bioinformatics* (pp. 125–158). Springer US. https://doi.org/10.1007/978-0-387-09760-2_7

Pang, X.-X., & Zhang, D.-Y. (2024). A cautionary note on using STRUCTURE to detect hybridization in a phylogenetic context. *bioRxiv*. https://doi.org/10.1101/2024.02.06.579057

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

Peng, J., Swofford, D., & Kubatko, L. (2022). Estimation of speciation times under the multispecies coalescent. *Bioinformatics*, *38*(23), 5182–5190. https://doi.org/10.1093/bioinformatics/btac679

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489–494. https://doi.org/10.1038/nature08365

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610. https://doi.org/10.2307/2289471

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

Swofford, D. L. (2024). *Phylogenetic analysis using parsimony (*and other methods). Version 4.* https://paup.phylosolutions.com

Than, C., Ruths, D., & Nakhleh, L. (2008). PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, *9*(1), 322. https://doi.org/10.1186/1471-2105-9-322

Tricou, T., Tannier, E., & De Vienne, D. M. (2022). Ghost lineages can invalidate or even reverse findings regarding gene flow. *PLOS Biology*, *20*(9), e3001776. https://doi.org/10.1371/journal.pbio.3001776

Tricou, T., Tannier, E., & 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

Vähä, J.-P., & Primmer, C. R. (2005). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, *15*(1), 63–72. https://doi.org/10.1111/j.1365-294X.2005.02773.x

Wascher, M., & Kubatko, L. (2021). Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. *Systematic Biology*, *70*(1), 33–48. https://doi.org/10.1093/sysbio/syaa039

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, *9*(1), 60–62. https://doi.org/10.1214/aoms/1177732360

Supplementary Materials

PengKongKubatko_Final.tex

Download: https://ssbbulletin.scholasticahq.com/article/124365-a-likelihood-ratio-test-for-hybridization-under-the-multispecies-coalescent/attachment/250401.zip?auth_token=eOrT52FqvhZVf34geE00