Version dated: December 13, 2015

RH: INVARIANTS-BASED HYBRID DETECTION

An Invariants-based Method for Efficient Identification of Hybrid Species From Large-scale Genomic Data

Laura S. Kubatko[1,2,3] and Julia Chifman[1]

[1] *Department of Statistics,* [2] *Department of Evolution, Ecology, and Organismal Biology,*

[3] *Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, USA*

Corresponding author: Laura S. Kubatko, Department of Statistics,Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA; E-mail: lkubatko@stat.osu.edu.

Abstract.— Coalescent-based species tree inference has become widely used in the analysis of genome-scale multilocus and SNP datasets when the goal is inference of a species-level phylogeny. However, numerous evolutionary processes are known to violate the assumptions of a coalescence-only model and complicate inference of the species tree. One such process is hybrid speciation, in which a species shares its ancestry with two distinct species. Although many methods have been proposed to detect hybrid speciation, only a few have considered both hybridization and coalescence in a unified framework, and these are generally limited to the setting in which putative hybrid species must be identified in advance. Here we propose a method that can examine genome-scale data for a large number of taxa and detect those taxa that may have arisen via hybridization, as well as their potential "parental" taxa. The method is based on a model that considers both coalescence and hybridization together, and uses phylogenetic invariants to construct a test that scales well in terms of computational time for both the number of taxa and the amount of sequence data. We test the method using simulated data for up 20 taxa and 100,000bp, and find that the method accurately identifies both recent and ancient hybrid species in less than 30 seconds. We apply the method to two empirical datasets, one composed of *Sistrurus* rattlesnakes for which hybrid speciation is not supported by previous work, and one consisting of several species of *Heliconius* butterflies for which some evidence of hybrid speciation has been previously found.

(Keywords: hybrid speciation, coalescent, species tree, phylogenetic invariants)

Large-scale genomic data present many challenges in the inference of the evolutionary history of a collection of species. The most notable of these is the development of methods for inferring species-level phylogenetic relationships from multiple gene alignments that simultaneously incorporate the evolutionary processes that are known to contribute to variability in histories for the individual genes. Two important processes are incomplete lineage sorting (ILS) and hybridization (Maddison 1997). ILS results when two gene copies fail to coalesce in the most recent ancestral population and is commonly modeled by the coalescent process, which provides a link between the species tree and the gene trees that represent the phylogenetic history for each gene (Kingman 1982a,b; Tavaré 1984). In particular, multispecies coalescent theory models probabilities of rooted gene tree topologies within a given rooted species tree topology and has been used to derive the various probability distributions on gene trees given a particular species tree (Tajima 1983; Takahata and Nei 1985a; Pamilo and Nei 1988; Rosenberg 2002; Rannala and Yang 2003; Degnan and Salter 2005). To date, many methods have been proposed for estimation of species phylogeny from multi-locus data based on the coalescent process (e.g., BEST (Liu and Pearl 2007), *BEAST (Heled and Drummond 2010), STEM (Kubatko et al. 2009), MP-EST (Liu et al. 2010), SNAPP (Bryant et al. 2012), SVDquartets (Chifman and Kubatko 2014) (now implemented in PAUP* (Swofford 1998)), ASTRAL (Mirarb et al. 2014), among others).

Hybridization is another evolutionary process that can cause variability in gene trees within the containing species tree. It generally refers to the interbreeding of individuals from distinct populations, resulting in the production of a hybrid species that shares genetic information with both parental species. Hybridization between distinct species can occur for many generations with fertile offspring, making it possible for a new species to be formed. If the hybridization does not result in the formation of a new lineage, the process is termed introgression or introgressive hybridization (Dowling and DeMarais

1993; Roques et al. 2001; Thorsson et al. 2001; Salzburger et al. 2002; Weigel et al. 2002; Good et al. 2003; Grant et al. 2004; Mallet 2005, 2007; Baack and Rieseberg 2007). Despite the earlier belief that hybridization was rare, numerous recent studies have shown that hybrid speciation occurs in both plants and animals (Rieseberg 1997; Gross and Rieseberg 2005; Buerkle et al. 2000; Bullini 1994; Nolte et al. 2005; DeMarais et al. 1992; Gompert et al. 2006; Schwarz et al. 2005; Mavarez 2006; Meyer et al. 2006; Mallet 2007). Hybridization has been recognized as an important mechanism for the evolution of new species and recent estimates indicate that approximately 25% of plants and 10% of animals hybridize (Seehausen 2004; Mallet 2005, 2007; Baack and Rieseberg 2007; Mallet 2007). However, inference of hybridization cannot be based solely on observed gene tree variability since other processes (e.g., incomplete lineage sorting and gene duplication and loss) may contribute to disagreements in single-gene phylogenies (Maddison 1997).

Several models and methods have been developed to detect hybridization. One group of methods involves the identification and removal of hybrids prior to phylogenetic analysis, with the hybrids added to the inferred tree by connecting them to their parental species (Rieseberg and Morefield 1995; Posada 2002; Gauthier and Lapointe 2007). Joly et al. (2009) developed a method and software (JML; Joly (2012)) for identifying introgressed sequences by proposing that for some hybridization events the minimum distance between two sequences will be smaller than for incomplete lineage sorting. Another test that was originally developed to test ancient admixture is based on a relative abundance of ABBA or BABA single nucleotide patterns that can be evaluated using Patterson's D-statistic (Green et al. 2010; Durand et al. 2011; Patterson et al. 2012). Meng and Kubatko (2009) proposed a model for detecting hybridization under the coalescent model and used both a maximum likelihood and a Bayesian framework for inference. An extension to that model was later provided by Kubatko (2009) by utilizing gene tree densities for inference. Yu et al. (2014) also proposed a likelihood method that accounts for both reticulate evolutionary events and incomplete lineage sorting by providing methods for computing the likelihood of a phylogenetic network under the coalescent model. This method, as well as some earlier variations of it, is implemented in the software PhyloNet (Than et al. 2008).

In this paper we develop a method for detecting and quantifying the extent of hybridization using a coalescent-based model that is fast and accurate. At the heart of our method are special relations called *phylogenetic invariants*, which are functions (usually polynomials) in the site pattern probabilities that evaluate to zero on any probability distribution that is consistent with the tree topology and associated model. Invariants have been introduced by Cavender and Felsenstein (1987) and Lake (1987) as a means for phylogenetic reconstruction, and have recently been gaining popularity for use in phylogenetic tree inference (Eriksson 2005; Casanellas and Fernández-Sánchez 2011; Chifman and Kubatko 2014). Here we propose using a ratio between two linear invariants in site pattern probabilities to develop statistics that accurately identify hybrid taxa. Because these statistics are functions of site pattern probabilities across multi-locus or SNP data, they can be rapidly computed. In addition, we can derive the mean, variance, and asymptotic distribution of these invariants, enabling development of a hypothesis test for hybridization when the number of sites is large. We begin by giving the theoretical details of our model, and then evaluate the performance of several possible invariants-based statistics for four-taxon trees using simulation. The best-performing of these statistics, which we call the Hils statistic, is then evaluated for larger trees using simulation, with hybridization events at various “depths” of the tree (i.e., hybridization between tip species and hybridization between ancestral species). Finally, we apply our method to several empirical data sets, including the *Sistrurus* rattlesnakes and *Heliconius* butterflies.

Methods

A Coalescent-based Model for Hybridization

We consider here the model originally proposed by Meng and Kubatko (2009) in which data arise along a phylogenetic species tree via an evolutionary process that allows for the possibility of both hybridization and incomplete lineage sorting, as modeled by the coalescent process. Hybridization cannot be modeled by a bifurcating phylogenetic tree, thus it is common to represent hybridization on a phylogeny by a horizontal line connecting two lineages of an otherwise-bifurcating phylogeny, as shown in the leftmost panel of Figure 1. This tree represents the evolutionary history of the species as a whole, and depicts a hybrid origin for taxon H. We refer to species H as the hybrid species, and to species P_1 and P_2 as the parental species. The times labeled by τ_i are speciation times, and in general we refer to the tree topology S_γ together with its vector of speciation times $\boldsymbol{\tau}$ by $(S_\gamma, \boldsymbol{\tau})$. The data arising along this phylogenetic species tree are a collection of site patterns. Letting $X_Y \in \{A, C, G, T\}$ denote the nucleotide observed for species Y at a specific location in the DNA sequence, we define a site pattern $\mathbf{X} = X_O X_{P_1} X_H X_{P_2}$ as an assignment of nucleotides to all species. We represent the site pattern probability on the species tree $(S_\gamma, \boldsymbol{\tau})$ for a particular observation $ijkl$ at the tips of the tree by

$$p_{ijkl|(S_\gamma,\boldsymbol{\tau})} = P(X_O = i, X_{P_1} = j, X_H = k, X_{P_2} = l|(S_\gamma, \boldsymbol{\tau})) \qquad (1)$$

for $i, j, k, l \in \{A, C, G, T\}$.

Our model defines the probability distribution on the space of all $4^4 = 256$ site patterns under a model that allows both ILS and hybridization via a three-stage process. First, the hybrid species is assigned one of its two putative parents, with probability γ of

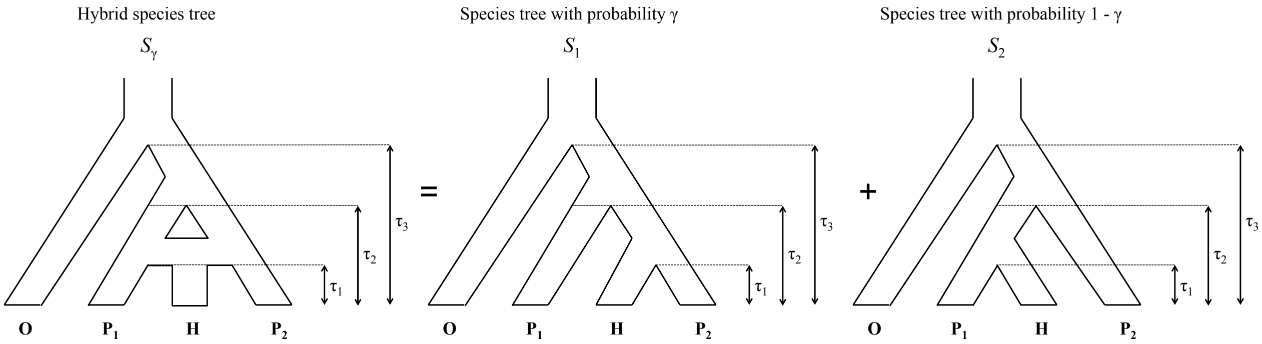


Figure 1: Model for the species-level relationships among four taxa under the coalescent model with hybridization. Here taxon H is a hybrid of taxa P_1 and P_2.

selecting parental species P_1 and probability $1-\gamma$ of selecting parental species P_2 (resulting in trees S_1 and S_2 in Figure 1 being the "parental species trees", respectively). Next, a gene tree is generated along the parental species tree from step 1 through the standard coalescent process (see, e.g., Kingman (1982a,b); Tavaré (1984); Rannala and Yang (2003); Tajima (1983); Takahata and Nei (1985b); Pamilo and Nei (1988); Wakeley (2009)). Finally, a site pattern is generated along the gene tree from step 2 according to one of the standard Markov substitution models (e.g., the GTR+I+Γ model (Lanave et al. 1984) or one of its sub-models). Combining steps 2 and 3, we see that the probability for site pattern $ijkl$ for a given species tree S_i, $i \in \{1,2\}$, is given by

$$p_{ijkl|(S_i,\boldsymbol{\tau})} = \sum_G \int_{\mathbf{t}} p_{ijkl|(G,\mathbf{t})} f((G,\mathbf{t})|(S_i,\boldsymbol{\tau}))d\mathbf{t},$$

where $(G,\mathbf{t})$ represents a gene tree with topology G and branch lengths $\mathbf{t}$, $p_{ijkl|(G,\mathbf{t})}$ is the probability of the particular observation $ijkl$ at the tips of gene tree $(G,\mathbf{t})$, and $f((G,\mathbf{t})|(S_i,\boldsymbol{\tau}))$ is the joint density of $(G,\mathbf{t})$ conditional on the species tree $(S_i,\boldsymbol{\tau})$. A full description of the computations required for this model are given in Chifman and Kubatko (2015), and we do not review them here. Finally, we write the site pattern probability on a hybrid species tree as

$$p_{ijkl|(S_\gamma,\boldsymbol{\tau})} = \gamma p_{ijkl|(S_1,\boldsymbol{\tau})} + (1-\gamma) p_{ijkl|(S_2,\boldsymbol{\tau})}. \tag{2}$$

For our purposes, it suffices to view the collection of site patterns observed in an empirical data set as a sample of observations from the probability distribution defined by the $\{p_{ijkl|(S_\gamma,\boldsymbol{\tau})} | i,j,k,l \in \{A,C,G,T\}\}$. We call data generated in this way "coalescent independent sites" and refer to this model as the "coalescent independent sites model".

Let $N_{\mathbf{X}}$ be the number of sites with site pattern $\mathbf{X}$ observed in a sample of N sites generated from hybrid species tree $(S_\gamma, \boldsymbol{\tau})$ under this coalescent-with-hybridization model. Define $\mathbf{p} = (p_{AAAA|(S_\gamma,\boldsymbol{\tau})}, p_{AAAC|(S_\gamma,\boldsymbol{\tau})}, \ldots, p_{TTTT|(S_\gamma,\boldsymbol{\tau})})$ and $\hat{\mathbf{p}} = (\hat{p}_{AAAA}, \hat{p}_{AAAC}, \ldots, \hat{p}_{TTTT})$, where $\hat{p}_{\mathbf{X}} = \frac{N_{\mathbf{X}}}{N}$. Then,

$$\hat{\mathbf{p}} \sim \frac{1}{N}\text{Multinomial}(N; \mathbf{p}). \tag{3}$$

When N is large, the $\hat{p}_{\mathbf{X}}$ are approximately normally distributed, and thus the sampling distributions of statistics based on the $\hat{p}_{\mathbf{X}}$ can be derived. We next describe how these ideas can be used to build tests for hybridization.

Invariants-based Hypothesis Tests for Hybridization

As mentioned in the Introduction, our tests are based on phylogenetic invariants, which are polynomials in the site patterns that evaluate to zero on one tree topology but do not evaluate to zero for at least one tree of a different topology. Consider four linear relationships that arise on the hybrid phylogenetic species tree $(S_\gamma, \boldsymbol{\tau})$ as described in the previous section:

$$f_1 = p_{iijj|(S_\gamma,\boldsymbol{\tau})} - p_{ijij|(S_\gamma,\boldsymbol{\tau})} \qquad f_3 = p_{ijii|(S_\gamma,\boldsymbol{\tau})} - p_{iiji|(S_\gamma,\boldsymbol{\tau})}$$

$$f_2 = p_{ijji|(S_\gamma,\boldsymbol{\tau})} - p_{ijij|(S_\gamma,\boldsymbol{\tau})} \qquad f_4 = p_{iiij|(S_\gamma,\boldsymbol{\tau})} - p_{iiji|(S_\gamma,\boldsymbol{\tau})}$$

where $i \neq j \in \{A, C, G, T\}$. It can be shown that f_2 and f_4 are zero when evaluated on site pattern probabilities that correspond to the species tree S_1, while f_1 and f_3 are non-zero (see Chifman and Kubatko (2015) for details). Similarly, f_1 and f_3 are zero when evaluated on site pattern probabilities that correspond to tree S_2, while f_2 and f_4 are not. However, when the site pattern probabilities correspond to the tree $(S_\gamma, \boldsymbol{\tau})$ with $\gamma \in (0, 1)$, none of the four linear relations are zero.

What is special about these functions is that their ratio is a function of $\gamma \in (0, 1)$:

$$\begin{aligned} \frac{f_1}{f_2} &= \frac{p_{iijj|(S_\gamma,\boldsymbol{\tau})} - p_{ijij|(S_\gamma,\boldsymbol{\tau})}}{p_{ijji|(S_\gamma,\boldsymbol{\tau})} - p_{ijij|(S_\gamma,\boldsymbol{\tau})}} \\ &= \frac{\gamma\big(p_{iijj|(S_1,\boldsymbol{\tau})} - p_{ijij|(S_1,\boldsymbol{\tau})}\big) + (1-\gamma)\big(p_{iijj|(S_2,\boldsymbol{\tau})} - p_{ijij|(S_2,\boldsymbol{\tau})}\big)}{\gamma\big(p_{ijji|(S_1,\boldsymbol{\tau})} - p_{ijij|(S_1,\boldsymbol{\tau})}\big) + (1-\gamma)\big(p_{ijji|(S_2,\boldsymbol{\tau})} - p_{ijij|(S_2,\boldsymbol{\tau})}\big)} \\ &= \frac{\gamma\big(p_{iijj|(S_1,\boldsymbol{\tau})} - p_{ijij|(S_1,\boldsymbol{\tau})}\big) + (1-\gamma)(0)}{\gamma(0) + (1-\gamma)\big(p_{ijji|(S_2,\boldsymbol{\tau})} - p_{ijij|(S_2,\boldsymbol{\tau})}\big)} \\ &= \frac{\gamma}{1-\gamma}. \end{aligned} \tag{4}$$

Notice that the last equality holds because $p_{ijji|(S_2,\boldsymbol{\tau})} - p_{ijij|(S_2,\boldsymbol{\tau})} = p_{iijj|(S_1,\boldsymbol{\tau})} - p_{ijij|(S_1,\boldsymbol{\tau})}$. Using a similar argument we find that

$$\frac{f_3}{f_4} = \frac{\gamma}{1-\gamma} \quad \text{and} \quad \frac{f_1 + f_3}{f_2 + f_4} = \frac{\gamma}{1-\gamma}. \tag{5}$$

If we consider cumulative site pattern probabilities then the results in Equations (4) and (5) still hold. By a cumulative site pattern we mean, for example, $p_{ijji|(S_\gamma,\boldsymbol{\tau})} = \sum_{x\neq y\in\{A,C,G,T\}} p_{xyyx|(S_\gamma,\boldsymbol{\tau})}$. Under the JC69 model (Jukes and Cantor 1969), each of the terms in the sum will have the same value, regardless of the choice of x and y; under more complex models, these probabilities will vary depending on the particular x and y. We implement the JC69 version of the test here, though we use simulation to assess the performance under more complicated models.

Using the ratios in Equations (4) and (5) we construct formal significance tests of the following hypotheses:

$$H_0 : \gamma = 0 \text{ vs. } H_1 : \gamma > 0.$$

Here we consider the ratio $\frac{f_1}{f_2}$ to illustrate the procedure. First, we estimate this ratio using the site pattern probabilities observed in the sample,

$$\frac{\hat{f}_1}{\hat{f}_2} = \frac{\hat{p}_{iijj} - \hat{p}_{ijij}}{\hat{p}_{ijji} - \hat{p}_{ijij}} \tag{6}$$

To use this estimator as a test statistic in a hypothesis test, we need the distribution of the statistic when the null hypothesis is true. We first consider distributional results for the numerator and denominator separately. Using standard results for the multinomial distribution, we have

$$\begin{aligned}
\mu_{f_1} &:= E(\hat{p}_{iijj} - \hat{p}_{ijij}) = p_{iijj|(S_\gamma,\boldsymbol{\tau})} - p_{ijij|(S_\gamma,\boldsymbol{\tau})} & (7)\\
\mu_{f_2} &:= E(\hat{p}_{ijji} - \hat{p}_{ijij}) = p_{ijji|(S_\gamma,\boldsymbol{\tau})} - p_{ijij|(S_\gamma,\boldsymbol{\tau})} & (8)\\
\sigma^2_{f_1} &:= Var(\hat{p}_{iijj} - \hat{p}_{ijij}) = \frac{1}{N}(p_{iijj|(S_\gamma,\boldsymbol{\tau})}(1 - p_{iijj|(S_\gamma,\boldsymbol{\tau})}) & \\
&\quad + p_{ijij|(S_\gamma,\boldsymbol{\tau})}(1 - p_{ijij|(S_\gamma,\boldsymbol{\tau})}) + 2p_{iijj|(S_\gamma,\boldsymbol{\tau})}p_{ijij|(S_\gamma,\boldsymbol{\tau})}) & (9)\\
\sigma^2_{f_2} &:= Var(\hat{p}_{ijji} - \hat{p}_{ijij}) = \frac{1}{N}(p_{ijji|(S_\gamma,\boldsymbol{\tau})}(1 - p_{ijji|(S_\gamma,\boldsymbol{\tau})}) & (10)\\
&\quad + p_{ijij|(S_\gamma,\boldsymbol{\tau})}(1 - p_{ijij|(S_\gamma,\boldsymbol{\tau})}) + 2p_{ijji|(S_\gamma,\boldsymbol{\tau})}p_{ijij|(S_\gamma,\boldsymbol{\tau})}) & \\
\sigma_{f_1,f_2} &:= cov(\hat{p}_{iijj} - \hat{p}_{ijij}, \hat{p}_{ijji} - \hat{p}_{ijij}) & (11)\\
&\quad = \frac{1}{N}(-p_{iijj|(S_\gamma,\boldsymbol{\tau})}p_{ijji|(S_\gamma,\boldsymbol{\tau})} + p_{iijj|(S_\gamma,\boldsymbol{\tau})}p_{ijij|(S_\gamma,\boldsymbol{\tau})} & \\
&\quad + p_{ijji|(S_\gamma,\boldsymbol{\tau})}p_{ijij|(S_\gamma,\boldsymbol{\tau})} + p_{ijij|(S_\gamma,\boldsymbol{\tau})}(1 - p_{ijij|(S_\gamma,\boldsymbol{\tau})})). & (12)
\end{aligned}$$

Now, using the fact that when the sample size N is large we have $\hat{f}_1 \sim N(\mu_{f_1}, \sigma^2_{f_1})$ and $\hat{f}_2 \sim N(\mu_{f_2}, \sigma^2_{f_2})$, we apply the Geary-Hinkley transformation (Geary 1930; Hinkley

1969) to the ratio $\frac{\hat{f}_1}{\hat{f}_2}$ to get

$$\frac{(\mu_{f_2}\frac{\hat{f}_1}{\hat{f}_2} - \mu_{f_1})}{\sqrt{\sigma^2_{f_2}(\frac{\hat{f}_1}{\hat{f}_2})^2 - 2\sigma_{f_1,f_2}\frac{\hat{f}_1}{\hat{f}_2} + \sigma^2_{f_1}}} \sim N(0,1). \tag{13}$$

The term on the left-hand side of the above equation depends on several unknown quantities. We estimate these by substituting the observed site pattern frequencies into Equations (7) - (12) and re-arrange the expression in Equation (13) to obtain the test statistic

$$H := \frac{\hat{f}_2(\frac{\hat{f}_1}{\hat{f}_2} - \frac{\mu_{f_1}}{\mu_{f_2}})}{\sqrt{\hat{\sigma}^2_{f_2}(\frac{\hat{f}_1}{\hat{f}_2})^2 - 2\hat{\sigma}_{f_1,f_2}\frac{\hat{f}_1}{\hat{f}_2} + \hat{\sigma}^2_{f_1}}}. \tag{14}$$

We call the statistic H the Hils statistic, in honor of Professor Matthew H. Hils[1] . Under the null hypothesis that $\gamma = 0$, $H \sim N(0,1)$ for large N with $\frac{\mu_{f_1}}{\mu_{f_2}} = 0$, and the hypothesis test can be carried out by comparing the observed value of the test statistic with a standard normal distribution. Tests based on the ratios $\frac{f_3}{f_4}$ and $\frac{f_1+f_3}{f_2+f_4}$ can be derived analogously.

Extension to Larger Trees

The hypothesis test derived in the previous section deals with the case in which four taxa are specified, with one of the four taxa identified as the putative hybrid species. In many settings, however, primary interest is in searching over a large collection of species with the goal of identifying which species might have arisen via a process that involved hybridization at some point in the past. To address this, we consider a large collection of sequences, and suppose that an outgroup sequence can be identified. For each subset of four sequences consisting of three sequences plus the outgroup, we carry out the above test

[1] Matthew H. Hils was a Professor of Biology at Hiram College until his untimely death due to cancer in June 2014. He served as academic advisor and research mentor to L.K. during her undergraduate studies, and contributed to her decision to pursue interdisciplinary graduate study tied to the biological sciences. See `http://news.hiram.edu/?p=10502`.

of hybridization for different assignments of the three ingroup sequences to the hybrid and parental taxa. Of the three possible choices for the hybrid taxon, we consider only two of those, eliminating from consideration the one for which $\hat{p}_{ijij} > max\{\hat{p}_{iijj}, \hat{p}_{ijji}\}$, since this implies that the two parental taxa are more closely related than either is to the putative hybrid. For a data set of $n+1$ sequences with one outgroup sequence, this results in $\binom{n}{3} \times 2$ hypothesis tests. To handle the issue of multiple comparisons, we use the Bonferroni correction, which is conservative in this case because the tests are correlated. Thus, if an overall α-level test is desired, we report significant evidence of hybridization when the p-value computed for a particular comparison is smaller than $\frac{\alpha}{\binom{n}{3}\times 2}$.

Our method is implemented in a program written in the C language called HyDe (**Hy**brid **De**tection), available at `http://www.stat.osu.edu/~lkubatko/software/`. The program takes a Phylip-formatted input alignment, with the outgroup sequence specified, and outputs a test statistic for every collection of three species in the input data.

Simulation Studies

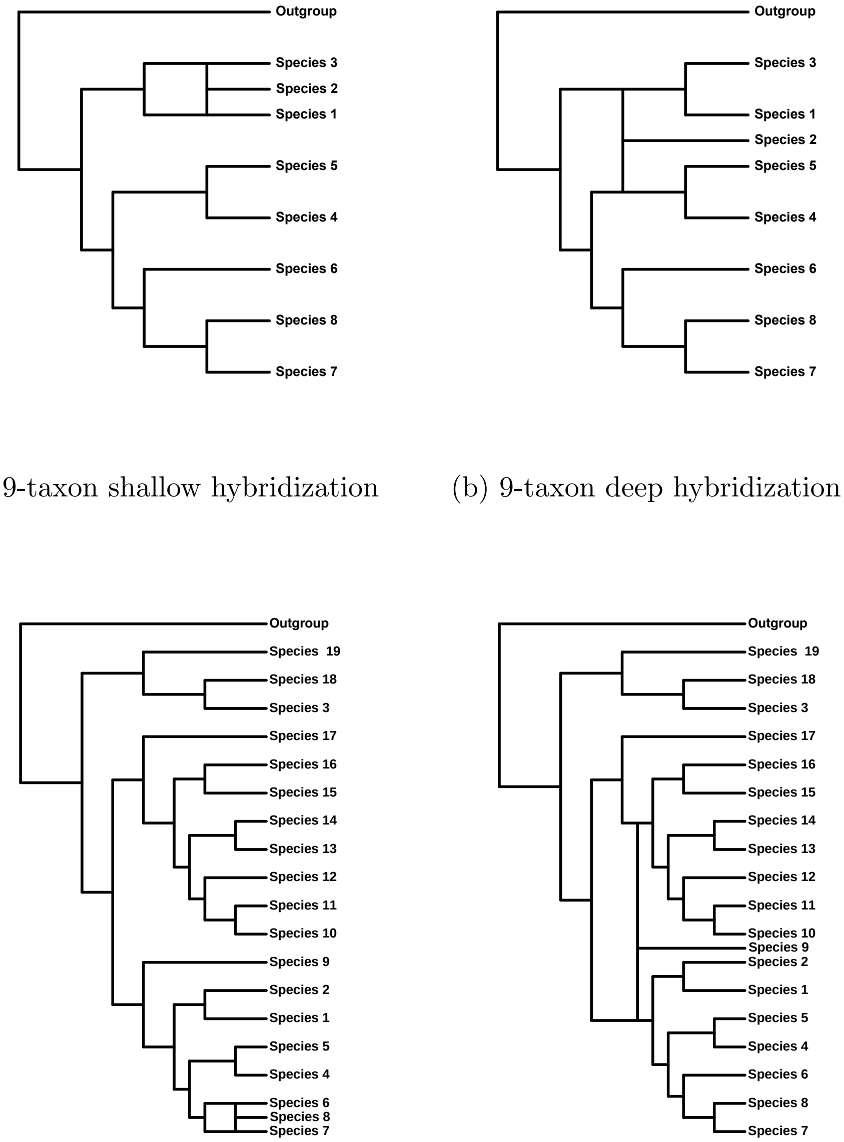
Four-taxon trees.— Our first set of simulation studies involves assessing the level and the power of the tests under various choices of the sample size, species trees branch lengths, and value of γ for four-taxon trees. We used the program COAL (Degnan and Salter 2005) to simulate gene trees from the two parental species trees in Figure 1 with γ values of 0, 0.1, 0.2, 0.3, 0.4, and 0.5 and for two sets of speciation times: $\tau_1 = 0.25, \tau_2 = 0.5, \tau_3 = 1.0$ (the "short" setting) and $\tau_1 = 0.5, \tau_2 = 1.0, \tau_3 = 2.0$ (the "long" setting). For each setting, we simulated $N = 50,000, 100,000, 250,000$ and $500,000$ coalescent independent sites under the GTR+I+Γ model using Seq-Gen (Rambaut and Grassly 1997)(Seq-Gen options: `-mGTR -r 1.0 0.2 10.0 0.75 3.2 1.6 -f 0.15 0.35 0.15 0.35 -i 0.2 -a 5.0 -g 3`). For each parameter setting, we generated 500 replicate data sets.

For each simulated data set, we tested the null hypothesis that $\gamma = 0$ using the test statistics corresponding to the ratios in Equations (4) and (5) at level $\alpha = 0.05$. We estimate the power of each test as the proportion of the 500 replicates for which the null hypothesis was rejected (when $\gamma = 0$, this gives an estimate of the level of the test). We also considered using each of the statistics to estimate the true hybridization parameter, γ. We report the mean of the estimated γ values, as well as the standard deviation and the mean squared error, for each parameter setting.

Larger trees.— To examine the performance of our method for larger taxon samples, we considered trees containing 8 species and an outgroup, and trees containing 19 species and an outgroup. We also considered both recent hybridization and more ancient hybridization in each case. Our model trees are shown in Figure 2. For each model tree, we generated 125 data sets containing 100,000 coalescent independent sites for $\gamma = 0, 0.1, 0.2, 0.3, 0.4$, and 0.5 as follows. First, 100000γ gene trees were generated from the species tree formed by connecting the hybrid taxon to the "left" parental lineage, and $100000(1 - \gamma)$ gene trees were generated from the species tree formed by connecting the hybrid taxon to the "right" parental lineage. For each gene tree, one coalescent independent site was generated using Seq-Gen (Rambaut and Grassly 1997) under the GTR+I+Γ model (Seq-Gen options: `-mGTR -r 1.0 0.2 10.0 0.75 3.2 1.6 -f 0.15 0.35 0.15 0.35 -i 0.2 -a 5.0 -g 3`). Each simulated data set was then given to our program with the outgroup specified, and the Hils statistic was computed for each possible combination of parents and hybrids. A cut-off for significance was determined using a Bonferroni correction with base level $\alpha = 0.05$, and the putative hybrid and parents were reported for any statistic whose p-value fell below α/M, where M was the total number of comparisons. We summarized results by counting the number of "True Positives" (data sets for which the correct hybrid and parental taxa are correctly identified), "Correct Sets" (data sets for which the correct hybrid and parental taxa are identified, but their assignment to which is the hybrid and which are the parental taxa is ambiguous), and "False Positives" (data sets for which an incorrect set of taxa are identified as being subject to hybridization).

Because many of the genome-scale datasets being generated today are multilocus datasets (rather than being generated under the coalescent independent sites model used here), we also simulated data under multilocus models. These simulations proceeded exactly as described above, except that rather than simulating 100,000 coalescent independent sites, we simulated 1,000 genes each of length 100bp. This choice was made to mimick the short read lengths generated by next-gen sequencing methods. We summarized these results in the same manner as described above. We justify application of our methodology to multilocus data in the Discussion section.

Empirical examples.— We have also explored the performance of our method on two empirical data sets; the *Sistrurus* rattlesnakes and *Heliconius* butterflies. The *Sistrurus* rattlesnakes are found across North America and are currently classified into two species, *Sistrurus catenatus* and *S. miliarius*, each with three putative subspecies. The dataset consists of 19 genes sampled from 26 rattlesnakes: 18 individuals within the species *Sistrurus catenatus* (with subspecies *S. c. catenatus* (Sca, 9 individuals), *S. c. edwardsii* (Sce, 4 individuals), and *S. c. tergeminus* (Sct, 5 individuals)); six within species *Sistrurus miliarius* (with subspecies *S. m. miliarius* (Smm, 1 individual), *S. m. barbouri* (Smb, 3 individuals), and *S. m. streckeri* (Sms, 2 individuals)); and two outgroup species, *Agkistrodon contortrix* and *A. piscivorus.* These data were originally analyzed by Kubatko et al. (2011) to determine species-level phylogenetic relationships. Prior to this analysis, the sequences were computationally phased, resulting in 52 sequences and 8,466 aligned nucleotide positions (data are available at TreeBase ID 11174). These data have been subsequently reanalyzed in several ways. For example, Chifman and Kubatko (2014) used



(a) 9-taxon shallow hybridization

(b) 9-taxon deep hybridization

(c) 20-taxon shallow hybridization

(d) 20-taxon deep hybridization

Figure 2: Model trees with 9 and 20 taxa and with either shallow or deep hybridization used for the simulation studies.

different methodology to infer the species phylogeny, and found agreement with the original analysis of Kubatko et al. (2011). Gerard et al. (2011) used a subset of the data to examine whether several specimens collected in Missouri and assigned to subspecies *S. c.*

catenatus were actually hybrid species. They did not find evidence of hybridization, in agreement with other results using different data (Gibbs et al. 2011).

The *Heliconius* butterflies are a diverse group tropical butterflies in the family *Heliconii* that are found throughout the southern United States and in Central and South America. We consider the study of Martin et al. (2013) in which genome-scale data for 31 individuals from seven distinct species were collected and evidence for gene flow between various species was assessed. We examine a subset of these data consisting of four individuals from each of the species *Heliconius cydno*, *H. melpomene rosina*, and *H. m. melpomene*, as well as one individual from the outgroup species *H. hecale*. Martin et al. (2013) found evidence that *H. m. rosina* is a hybrid of *H. m. melpomene* and *H. cydno*. We obtained the aligned genome-wide data from the complete study of Martin et al. (2013) from Dryad (`http://datadryad.org/resource/doi:10.5061/dryad.dk712`) (Martin et al. 2013a), and extracted the 13 sequences of interest. The resulting aligned sequences consisted of 248,822,400 base pairs.

Results

Four-taxon simulation studies.— The results of the four-taxon simulation studies are shown in Figure 3 and Table 1. Figure 3 shows that the various tests behave as we might expect in several ways. First, in all of the cases considered, the power increases as the sample size increases, reaching near 100% when alignments of length 500,000bp were used for many of the simulation conditions. Second, we note that as the value of γ increases from 0 (no hybridization) to 0.5 (equal contribution from both parental species), the power to detect hybridization increases as well, with near 100% power for the “long” branch length setting when $\gamma \geq 0.3$ for all three of the tests considered. Third, we note that all of the tests are more powerful for data simulated under the “long” branch length setting

(Figure 3 (d), (e), and (f)) than for data generated under the "short" branch length setting (Figure 3 (a), (b), and (c)). Finally, we note that all tests appear to achieve the nominal 0.05 level when data are simulated under the null hypothesis ($\gamma = 0$).

One unexpected result of the simulations designed to address the power was that the test based on $\frac{f_1}{f_2}$ is more powerful than the tests based on $\frac{f_3}{f_4}$ and $\frac{f_1+f_3}{f_2+f_4}$. This is most likely due to the variance associated with estimating the various site pattern probabilities that contribute to each invariant. We return to this point in the discussion. Based on this observation, we report results for only the ratio $\frac{f_1}{f_2}$ in what follows.

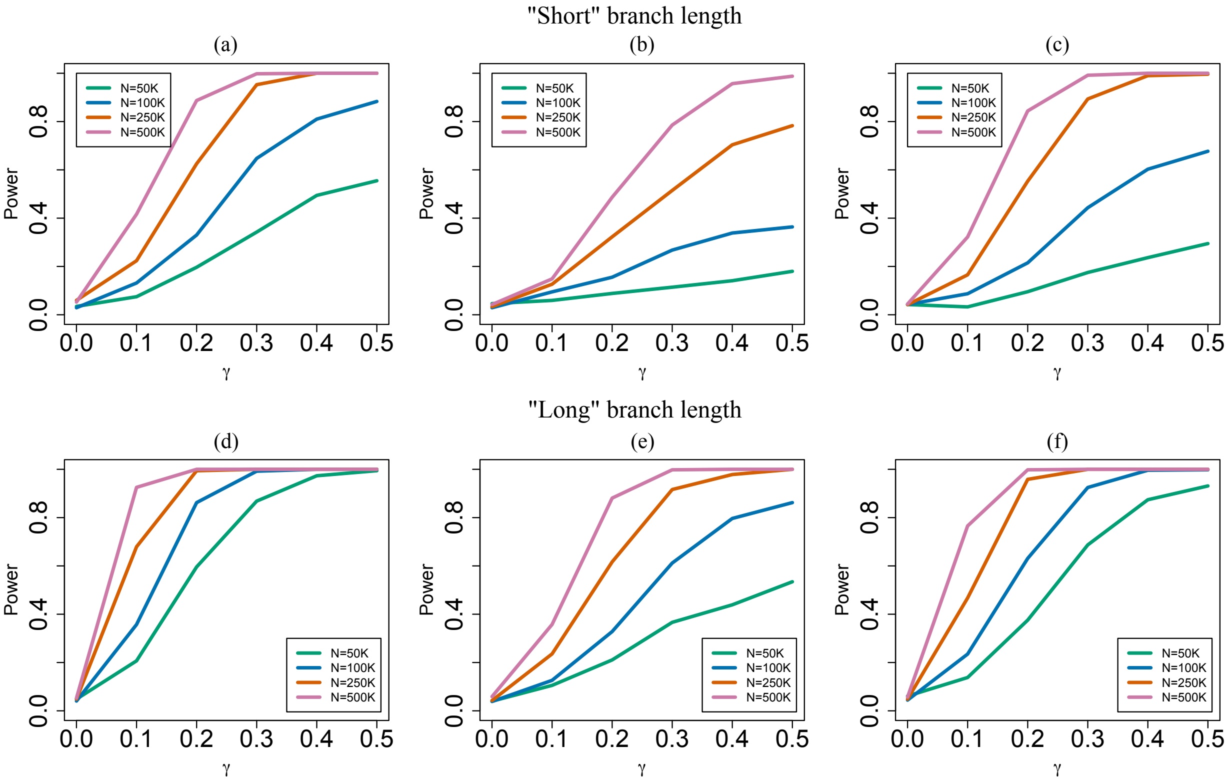


Figure 3: Results of the power simulations for the four-taxon hybrid species tree in Figure 1. Plots (a), (b), and (c) correspond to data simulated for the "short" branch length setting, and plots (d), (e), and (f) correspond to data simulated for the "long" branch length settings. Plots (a) and (d) give results for the test based on $\frac{f_1}{f_2}$; plots (b) and (e) give results for the test based $\frac{f_3}{f_4}$; and plots (c) and (f) give results for the test based on $\frac{f_1+f_3}{f_2+f_4}$.

Table 1 gives the results of the four-taxon simulation studies designed to estimate γ

	True mixing parameter γ											
	"Short" branch length						"Long" branch length					
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
500K												
Mean	-0.0078	0.0943	0.1942	0.2992	0.4021	0.4992	-0.0011	0.0969	0.199	0.2998	0.3999	0.4999
SD	0.0634	0.0538	0.0492	0.0408	0.0353	0.038	0.0289	0.0247	0.0235	0.0217	0.0204	0.0211
MSE	0.0041	0.0029	0.0025	0.0017	0.0013	0.0014	0.0008	0.0006	0.0006	0.0005	0.0004	0.0004
250K												
Mean	-0.008	0.0868	0.1869	0.2982	0.3963	0.503	-0.0026	0.0965	0.1966	0.3003	0.3983	0.5011
SD	0.0974	0.0895	0.0726	0.0613	0.0549	0.0554	0.041	0.0375	0.034	0.0329	0.0282	0.0288
MSE	0.0096	0.0082	0.0054	0.0038	0.003	0.0031	0.0017	0.0014	0.0012	0.0011	0.0008	0.0008
100K												
Mean	-0.0373	0.072	0.1755	0.2924	0.3865	0.502	-0.007	0.0924	0.1977	0.2977	0.3993	0.4972
SD	0.187	0.1409	0.1369	0.1114	0.1092	0.0908	0.0653	0.0614	0.0535	0.0507	0.0467	0.046
MSE	0.0364	0.0206	0.0194	0.0125	0.0121	0.0083	0.0043	0.0038	0.0029	0.0026	0.0022	0.0021
50K												
Mean	-0.081	0.0603	0.1251	0.2871	0.3774	0.4835	-0.0202	0.0913	0.1972	0.2937	0.3994	0.5004
SD	0.8257	0.6497	0.3777	0.8711	0.2149	0.197	0.1152	0.0864	0.0735	0.0742	0.0719	0.0639
MSE	0.6884	0.4237	0.1482	0.759	0.0467	0.0391	0.0137	0.0075	0.0054	0.0055	0.0052	0.0041

Table 1: Estimates of the parameter γ using the ratio $\frac{f_1}{f_2}$ for data simulated on the four-taxon hybrid species tree in Figure 1 with the "short" and "long" branch length settings.

using the ratio $\frac{f_1}{f_2}$. These results also match our intuition about how the method should perform. As the sample size increases, the estimates become closer to the true values used to generate the data, and the variance decreases as the sample size increases. In general, the estimates obtained from the "long" branch length setting are slightly better than those obtained from data generated under the "short" branch length setting. Overall, the method seems to provide very reasonable estimates of γ.

Simulation studies for larger trees.— The results of the simulation studies for larger trees are given in Tables 2 and 3. For the 9-taxon simulations, we note first that for data generated under the coalescent independent sites model, when $\gamma = 0$ approximately 5% of the data sets give significant results, and thus the test appears to attain the desired significance level in this case. For the multilocus data sets, however, the type I error rate is larger than the specified 0.05 level, and thus the test appears to reject the null hypothesis more often than it should. When $\gamma > 0$, we see that the test is powerful for both the shallow and the deep hybridization events and for both types of data, with the power

	Coalescent Independent Sites						Multi-locus Data					
	Shallow Hybridization			Deep Hybridization			Shallow Hybridization			Deep Hybridization		
γ	False Pos.	True Pos.	True Sets	False Pos.	True Pos.	True Sets	False Pos.	True Pos.	True Sets	False Pos.	True Pos.	True Sets
0	0.064	–	–	0.048	–	–	0.216	–	–	0.224	–	–
0.1	0.024	0.384	0.192	0.024	0.664	0.520	0.176	0.424	0.312	0.088	0.719	0.552
0.2	0.032	0.968	0.920	0.000	0.952	0.944	0.000	0.968	0.864	0.000	1.000	0.896
0.3	0.032	0.976	0.976	0.000	0.976	0.976	0.000	1.000	0.968	0.000	1.000	1.000
0.4	0.008	0.976	0.144	0.000	1.000	0.448	0.000	1.000	0.248	0.000	1.000	1.000
0.5	0.016	0.960	0.000	0.000	0.952	0.000	0.000	1.000	0.000	0.000	1.000	1.000

Table 2: Results of the simulation study for 9 taxa. The columns labeled "False Pos." refer to the proportion of data sets for which a triplet of taxa were incorrectly identified as involving a hybridization event (false positives); the columns labeled "True Pos." refer to the proportion of data sets for which the correct triplet of taxa involving the hybridization event was identified *and* the hybrid taxon was correctly identified (true positives); and the columns labeled "True Sets" refer to the proportion of data sets for which the correct triplet of taxa was identified but the hybrid taxa was specified incorrectly. All data sets for which the true set was identified also identified the triplet with the correct hybrid assignment, and thus this proportion is always a fraction of the proportion of true positives.

above 90% in both cases when $\gamma \geq 0.2$. Furthermore, the test almost always selects the correct assignment of hybrid and parental taxa, with the proportion of times that this is exclusively generated increasing toward 100% as γ increases for the coalescent independent sites data. One observation we made that is not reflected in the results in Table 2 is that for data simulated from the tree involving the deep hybridization event, many sets appear as significant when some true relationship is detected. For example, it is common to have the hybrid correctly assigned, but the parental species assigned as belonging to a taxon from the sister clade of the true parent. This is especially true for the multilocus data sets with the deep hybridization event. In other words, this test is good at picking out the hybrid taxon, but not as good at unambiguously picking out its parents when the hybridization event occurs deeper in the tree. This was not the case for the shallow event, where it often got exactly the correct relationships and only those in most cases.

The results for the 20-taxon trees are largely the same. The test still demonstrates good power to detect the hybridization event, though the power does not rise above 90%

	Coalescent Independent Sites						Multi-locus Data					
	Shallow Hybridization			Deep Hybridization			Shallow Hybridization			Deep Hybridization		
γ	False Pos.	True Pos.	True Sets	False Pos.	True Pos.	True Sets	False Pos.	True Pos.	True Sets	False Pos.	True Pos.	True Sets
0	0.048	–	–	0.040	–	–	0.064	–	–	0.072	–	–
0.1	0.008	0.072	0.008	0.000	1.000	0.240	0.112	0.008	0.064	0.008	0.648	0.352
0.2	0.000	0.704	0.096	0.000	0.936	0.936	0.016	0.688	0.160	0.000	0.984	0.960
0.3	0.000	0.952	0.080	0.000	0.928	0.928	0.000	1.000	0.168	0.000	1.000	1.000
0.4	0.000	0.960	0.000	0.000	0.968	0.968	0.000	1.000	0.000	0.000	1.000	0.976
0.5	0.000	0.952	0.000	0.000	0.984	0.928	0.000	1.000	0.000	0.000	1.000	0.896

Table 3: Results of the simulation study for 20 taxa. Column headings are as in Table 2.

for all settings until $\gamma \geq 0.3$, rather than 0.2 as in the 9-taxon case. In addition, the proportion of data sets with "Correct Sets" decreases for the shallow hybridization events in this case, meaning that when a hybridization event is identified, it nearly always involved correct identification of which species was the hybrid and which were the parental species. Though there is a hint of an elevated type I error rate when multilocus data were simulated, the problem is not as dramatic as in the 9-taxon case. Overall, the method maintains its good ability to detect hybrid species.

Empirical data: Sistrurus rattlesnakes.— Recall that this dataset contains two species, each containing three subspecies, as well as two outgroup species, for a total of eight tips in the species phylogeny of interest. When analyzing empirical data of this nature, for which several individuals are sampled within each species, our main interest will be in detecting individuals that show evidence of hybrid origin from parental individuals that are members of two different species. The current version of our software will output the test statistic for all assignments of hybrid and parental taxa for a given outgroup, but this output can easily be examined to consider only the comparisons of interest. For the rattlesnake data for a particular choice of outgroup, we can consider all choices of one individual allele from each of three subspecies, and for each such choice, one individual will be assigned to be the hybrid and the other two assigned to be the parental taxa. For example, we can select one

Sca individual, one Sce individual, and one Sct individual, and carry out the Hils test for each possible choice of hybrid among these three. Thus, for our particular data set consisting of 18 Sca alleles, 8 Sce alleles, 10 Sct alleles, 2 Smm alleles, 6 Smb alleles, and 4 Sms alleles, there will be $\sum_{n_i \in \{0,1\}, \sum n_i = 3} \binom{18}{n_1}\binom{8}{n_2}\binom{10}{n_3}\binom{2}{n_4}\binom{6}{n_5}\binom{4}{n_6} = 7,840$ possible choices of three alleles, and two test statistics will be computed on each, resulting in $2 * 7840 = 15,680$ possible comparisons for each choice of outgroup sequence. We carry out the Bonferroni correction within the analysis for each outgroup, and thus each comparison uses significance level $\alpha = 0.05/15680 = 0.0000032$.

An additional practical issue that arose with our empirical data but was not observed with simulated data was that for some choices of three alleles, one or more of the site pattern frequencies p_{iijj}, p_{ijij}, and p_{ijji} was observed to be 0. To correct for this, we added a small count (0.005) to each observed site pattern count in all cases before computing estimated site pattern frequencies and carrying out the test. With this modification, we find no evidence of hybrid origin for any of the sequences with any choice of outgroup sequence, consistent with other analyses in this group (Gerard et al. 2011; Gibbs et al. 2011).

Empirical data: Heliconius butterflies.— This dataset consists of 3 species with 4 individuals sampled per species, plus an outgroup. Thus, the number of comparisons of interest is $4 * 4 * 4 * 2 = 128$ and the Bonferroni-corrected level of the tests is $0.05/128 = 0.00039$. The analysis of all possible hybrid/parental combinations for the alignment of length $\approx$ 248 million bp took 16 minutes on a 2x Quad Core Xeon E5520 / 2.26GHz / 32GB desktop linux machine. All comparisons were statistically significant at the 0.00039 level. This result is not surprising, given the previous evidence of hybridization as described in Martin et al. (2013), and given the large sample size. What is interesting, however, is the strength of the evidence for hybridization. For example, across all comparisons in which an *H. m. rosina* individual was specified as the hybrid, the smallest test statistic was 172.6143, indicating overwhelming evidence for hybridization (recall that we are comparing to a standard normal distribution). In contrast, when one of the other species was identified as the hybrid and *H. m. rosina* was (incorrectly) identified as a parental taxon, the values of the test statistic ranged from ~ 55 to 76, again indicating strongly significant deviation from the expected patterns under no gene flow, but not as strong as the case in which the hybrid is correctly identified as *H. m. rosina*. Overall, these results are in agreement with the work of Martin et al. (2013) on this group, and demonstrate the utility of our method in rapidly identifying hybrid taxa from genome-scale data.

Discussion

We have proposed a method for detecting hybrid species using a model of hybrid speciation that incorporates coalescent stochasticity. The test is based on observed site pattern frequencies, which leads to several convenient properties. First, the computations required for the test can be carried out very rapidly, as all that is required is to obtain counts of observed site pattern frequencies for four taxa of interest. This computation is so rapid that there are essentially no limits on the length of sequences that can be handled by the method, and it is thus appropriate for genome-scale data. Second, observed site pattern frequencies arise from a multinomial distribution under the coalescent hybridization model used here, which allows derivation of the asymptotic distribution of the estimators of the site pattern frequencies. This ultimately leads to a null distribution for testing the hypothesis of interest that is asymptotically normally distributed which provides a straightforward test of the hypothesis of interest. Finally, we note that our method is derived under the assumption that each site has its own underlying gene tree, an experimental design that we propose calling "coalescent independent sites". The method is thus clearly appropriate for genome-wide SNP data, whether biallelic or not. We argue that the method is also appropriate for multilocus data, in that as the number of loci becomes large and provided that alignment lengths are not biased toward certain gene tree topologies, the proportion of sites observed from a particular gene tree will approach the proportion expected under the coalescent independent sites model. We thus carry out simulations for both multilocus and coalescent independent sites data, and we test our method on an empirical multilocus dataset.

Our simulations show that the method is powerful for detecting hybridization for both recent and ancient hybridization events, although for ancient hybridization events it may be more difficult to pinpoint the precise parental species for the detected hybrids. In addition, the proportional contribution of the two parental species to the genome of the hybrid species can be estimated accurately and unbiasedly. The simulations also show that the method scales extremely well: for 20-taxon trees with 100,000 sites, computations can be completed in less than 30 seconds, while for a dataset with 13 sequences and over 248 million sites, the analysis took less than 20 minutes on an older desktop linux machine. To the extent of our knowledge, this method is thus the only technique available for exploratory hybrid identification for large numbers of sequences using genome-scale data.

The method is based on phylogenetic invariants, and we note that the particular choice of invariants used here was somewhat arbitrary. Indeed, the ABBA-BABA test (Green et al. 2010; Durand et al. 2011; Patterson et al. 2012) is based on the difference of ABBA and BABA patterns similar to our invariant f_2 and it too has been shown to be useful in detecting hybridization. However their statistic is normalized by the total number of observations whereas our method is based on the ratio of two linear invariants leading to a function that depends only on the mixing parameter γ. Based on this crucial observation we were able to derive Hils statistic for accurate detection of hybridization. We have also noticed that the ratio between f_3 and f_4 was not as powerful, thus it is possible that other invariants may be identified that work as well or better than the ones we have chosen here. It is also possible that invariants that operate on more than four taxa at a time could be determined, with potential improvements in the localization of hybrid and parental taxa for more ancient hybridization events. It is also possible that a set of linear invariants specific to species trees under the coalescent exists and can be classified, and if such a set exists, these species invariants may improve the performance. We suggest that exploring these directions is appealing, as site pattern-based methods provide the possibility of both rapid computation and convenient asymptotic distributions, making them suitable for processing the large genome-scale datasets that are becoming increasingly available. In fact, the performance of these methods improves with sequence length, since site pattern probabilities can be more accurately estimated, with little associated computational cost.

Acknowledgements

This work was supported in part by the National Science Foundation under award DMS-1106706 (J.C., L.K.) and NIH Cancer Biology Training Grant T32-CA079448 at Wake Forest School of Medicine (J.C.).

*

References

Baack, E. and L. Rieseberg. 2007. A genomic view of introgression and hybrid speciation. Curr. Opin. Genet. Devel. 17:1–6.

Bryant, D., R. Bouckaert, J. Felsenstein, N. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Buerkle, C. A., R. Morris, M. Asmussen, and L. Rieseberg. 2000. The likelihood of homoploid hybrid speciation. Heredity 84:441–451.

Bullini, L. 1994. Origin and evolution of animal hybrid species. Trends Ecol. Evol. 9:422–426.

Casanellas, M. and J. Fernández-Sánchez. 2011. Relevant phylogenetic invariants of evolutionary models. Journal de Mathématiques Pures et Appliquées 96:207 – 229.

Cavender, J. A. and J. Felsenstein. 1987. Invariants of phylogenies in a simple case with discrete states. Journal of Classication 4:57–71.

Chifman, J. and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Chifman, J. and L. Kubatko. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. In print, Journal of Theoretical Biology .

Degnan, J. and L. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

DeMarais, B. D., T. Dowling, M. Douglas, W. Minckley, and P. Marsh. 1992. Origin of gila seminuda (teleostei: Cyprinidae) through introgressive hybridization: implications for evolution and conservation. Proc. Natl Acad. Sci. USA 89:2747–2751.

Dowling, T. E. and B. D. DeMarais. 1993. Evoltionary significance of introgressive hybridization in cyprinid fishes. Nature 362:444–446.

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

Eriksson, N. 2005. Tree construction using Singular Value Decomposition. chap. 19, Pages 347–358 *in* Algebraic Statistics for Computational Biology (L. Pachter and B. Sturmfels, eds.). Cambridge University Press.

Gauthier, O. and F. J. Lapointe. 2007. Hybrid and phylogenetics revisited: a statistical test of hybridization using quartets. Syst. Botany 32:8–15.

Geary, R. C. 1930. The frequency distribution of the quotient of two normal variates. Journal of the Royal Statistical Society 93:pp. 442–446.

Gerard, D., H. L. Gibbs, and L. Kubatko. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evolutionary Biology 11:291.

Gibbs, H. L., M. Murphy, and J. E. Chiucchi. 2011. Genetic identity of endangered massasauga rattlesnakes (*sistrurus* sp.) in missouri. Conservation Genetics 12:433–439.

Gompert, Z., J. Fordyce, M. Forister, A. Shapiro, and C. Nice. 2006. Homoploid hybrid speciation in an extreme habitat. Science 314:1923–1925.

Good, J. M., J. R. Dembroski, D. W. Nagorsen, and J. Sullivan. 2003. Phylogeography and introgressive hybridization: Chipmunks (genus *tamias*) in the northerm rocky mountains. Evolution 57:1900–1916.

Grant, P. R., B. R. Grant, J. A. Markert, L. F. Keller, and K. Petren. 2004. Convergence evolutino of Darwin's finches caused by introgressive hybridization and selection. Evolution 58:1588–1599.

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

Gross, B. and L. Rieseberg. 2005. The ecological genetics of homoploid hybrid speciation. J. Hered. 96:241–252.

Heled, J. and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hinkley, D. V. 1969. On the ratio of two correlated normal random variables. Biometrika 56:635–639.

Joly, S. 2012. Jml: testing hybridization from species trees. Molecular Ecology Resources 12:179–184.

Joly, S., P. A. McLenachan, and P. J. Lockhart. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. The American Naturalist 174:pp. E54–E70.

Jukes, T. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Kingman, J. F. C. 1982a. On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Kingman, J. F. C. 1982b. The coalescent. Stoch. Proc. Appl. 13:235–248.

Kubatko, L. S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.

Kubatko, L. S., B. C. Carstens, and L. L. Knolwes. 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

Kubatko, L. S., H. L. Gibbs, and E. W. Bloomquist. 2011. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in sistrurus rattlesnakes. Systematic Biology .

Lake, J. A. 1987. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony . Molecular Biology and Evolution 4:167–191.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86–93.

Liu, L. and D. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Liu, L., L. Yu, and S. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.

Mallet, J. 2007. Hybrid speciation. Nature 446:279–283.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins. 2013a. Data from: Genome-wide evidence for speciation with gene flow in heliconius butterflies. Dryad Digital Repository .

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins. 2013b. Genome-wide evidence for speciation with gene flow in heliconius butterflies. Genome Research 23:1817–1828.

Mavarez, J. e. a. 2006. Speciation by hybridization in heliconius butterflies. Nature 441:868–871.

Meng, C. and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theor. Pop. Biol. 75:35–45.

Meyer, A., W. Salzburger, and M. Schartl. 2006. Hybrid origin of a swordtail species (teleostei: Xiphophorus clemenciae) driven by sexual selection. Mol. Ecol. 15:721–730.

Mirarb, S., R. Reaz, M. S. Bayzid, T. Zimmerman, M. S. Swenson, and T. Warnow. 2014. Astral: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

Nolte, A. W., J. Freyhof, K. Stemshorn, and D. Tautz. 2005. An invasive lineage of sculpins, cottus sp. (pisces, teleostei) in the rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. Proc. R. Soc. Lond. B 272:2379–2387.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. 2012. Ancient admixture in human history. Genetics 192:1065–1093.

Posada, D. 2002. Evalution of methods for detecting recombination from DNA sequences:empirical data. Mol. Biol. Evol. 19:708–717.

Rambaut, A. and N. Grassly. 1997. SeqGen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. in Biosci. 13:235–238.

Rannala, B. and Z. Yang. 2003. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 164:1645–1656.

Rieseberg, L. 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28:359–389.

Rieseberg, L. H. and J. D. Morefield. 1995. Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. Pages 333–353 *in* Experimental and Molecular Approaches to Plant Biosystematics (P. Hoch and A. G. Stephenson, eds.). Missouri Botanical Garden, St. Louis.

Roques, S., J.-M. Sevigny, and L. Bernatchez. 2001. Evidence for a broadscale introgressive hybridization between two redfish (genus *sebastes*) in the North-west Atlantic: a rare marine example. Mol. Ecol. 10:149–165.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.

Salzburger, W., S. Baric, and C. Sturmbauer. 2002. Speciation via introgressive hybridization in East African cichilds? Mol. Ecol. 11:619–625.

Schwarz, D., B. Matta, N. Shakir-Botteri, and B. McPheron. 2005. Host shift to an invasive plant triggers rapid animal hybrid speciation. Nature 436:546–549.

Seehausen, O. 2004. Hybridization and adaptive radiation. Trends Ecol. Evol. 19:198–206.

Swofford, D. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Takahata, N. and M. Nei. 1985a. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.

Takahata, N. and M. Nei. 1985b. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

Than, C., D. Ruths, and L. Nakhleh. 2008. Phylonet: A software package for analyzing and reconstructing reticulate evolutionary histories. BMC Bioinformatics 9:322.

Thorsson, A., E. Salmela, and K. Anamthawat-Jonsson. 2001. Morphological, cytogenetic, and molecular evidence for inrogressive hybridization in birch. J. Hered. 92:404–408.

Wakeley, J. 2009. Coalescent Theory: An Introduction. Roberts and Company.

Weigel, D. E., J. T. Peterson, and P. Spruell. 2002. A model using phenotypic characteristics to detect introgressive hybridizations in wild westslope cutthroat trout and rainbow trout. Transactions of the American Fisheries Society 141:389–403.

Yu, Y., J. Dong, K. J. Liu, and L. Nakhleh. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences 111:16448–16453.