# The timing and direction of introgression under the multispecies network coalescent

Mark S. Hibbins* and Matthew W. Hahn*,†

*Department of Biology and †Department of Computer Science, Indiana University, Bloomington, IN 47405

## ABSTRACT

Introgression is a pervasive biological process, and many statistical methods have been developed to infer its presence from genomic data. However, many of the consequences and genomic signatures of introgression remain unexplored from a methodological standpoint. Here, we develop a model for the timing and direction of introgression based on the multispecies network coalescent, and from it suggest new approaches for testing introgression hypotheses. We suggest two new statistics, $D_1$ and $D_2$, which can be used in conjunction with other information to test hypotheses relating to the timing and direction of introgression, respectively. $D_1$ may find use in evaluating cases of homoploid hybrid speciation, while $D_2$ provides a four-taxon test for polarizing introgression. Although analytical expectations for our statistics require a number of assumptions to be met, we show how simulations can be used to test hypotheses about introgression when these assumptions are violated. We apply the $D_1$ statistic to genomic data from the wild yeast *Saccharomyces paradoxus*, a proposed example of homoploid hybrid speciation, demonstrating its use as a test of this model. These methods provide new and powerful ways to address questions relating to the timing and direction of introgression.

## INTRODUCTION

The now-widespread availability of genomic data has demonstrated that gene flow between previously diverged lineages—also known as introgression—is a pervasive process across the tree of life (reviewed in Mallet et al. 2016). Whole-genome data has revealed the sharing of traits via gene flow between humans and an extinct lineage known as Denisovans (Huerta-Sánchez et al. 2014), between different species of *Heliconius* butterflies (The *Heliconius* Genome Consortium 2012), and between multiple malaria vectors in the *Anopheles gambiae* species complex (Fontaine et al. 2015; Wen et al. 2016a). Introgression can substantially alter the evolutionary trajectory of populations through adaptive introgression (Hedrick 2013), transgressive segregation (Rieseberg et al. 1999), and hybrid speciation (Schumer et al. 2014).

Species or populations exchanging migrants are often represented as a network, rather than as having strictly bifurcating relationships. The reticulations in such networks represent the histories of loci that have crossed species boundaries. However, phylogenetic networks are often conceived in different ways (Huson and Bryant 2006). Some representations imply specific evolutionary processes or directions of introgression, but these implications are not always intentional and/or properly addressed. For example, Figure 1 shows three ways in which introgression events can be depicted in a network, with species B and C involved in gene exchange in each. Figure 1a represents an introgression event between species B and C, after A and B have diverged and B has evolved independently for some period of time. This representation does not specify the direction of gene flow. Figure 1b suggests that lineage B is the result of hybridization between A and C, and therefore that the direction of gene flow is into B. Such a depiction is often used to represent the origin of admixed populations (e.g. Bertorelle and Excoffier 1998; Wang 2003), or hybrid speciation (e.g. Meng and Kubatko 2009), in which the hybridization event leads to the formation of a reproductively isolated lineage. Figure 1c suggests two speciation events that result in lineages sister to A and C, respectively, that then come together to form species B. This representation also implies the direction of introgression (into B), but differs from Figure 1b in that it could imply a period of independent evolution before hybridization (e.g. Patterson et al. 2012; Yu et al. 2014; Zhang et al. 2018).

Despite clearly representing different evolutionary scenarios—including in the timing of introgression relative to speciation, and the direction of introgression—species networks such as these are often used to represent introgression as a general process. One reason for this is that the three scenarios depicted in Figure 1 are difficult to distinguish. Popular methods for detecting introgression using SNPs, such as the $f_3$ and $f_4$ statistics (Reich et al. 2009; Patterson et al. 2012) and the related $D$ statistic (also known as the "ABBA-BABA" test; Green et al. 2010; Durand et al. 2011) can tell us whether introgression has occurred, and if so, between which taxa. These methods do not provide additional information about the introgression event, including its direction or its relationship to speciation. The same is true of phylogenetic methods that use gene trees without branch lengths to infer phylogenetic networks (e.g. Meng and Kubatko 2009; Yu et al. 2014), as the frequency of discordant topologies are the same under many different scenarios for the timing and direction of gene flow (Zhu and Degnan 2017).

Accurately inferring the direction of introgression and the timing of introgression is important in understanding the consequences of hybridization for evolution. One area where these inferences have become especially important is in evaluating the frequency of homoploid hybrid speciation (HHS). Schumer et al. (2014) proposed three criteria that must be met in order to label a species a homoploid hybrid: evidence for hybridization, evidence for reproductive isolation, and a causal link between the two. In applying these criteria, they suggested that few studies have been able to demonstrate a causal link between hybridization and reproductive isolation, and that HHS is likely a rare process. This has sparked a debate over how to characterize HHS and its frequency in nature (Feliner et al. 2017; Schumer et al. 2018a). Multiple studies of putative homoploid hybrids have tested general hypotheses of gene flow using genomic data, often in combination with morphological and reproductive isolation data (e.g. Elgvin et al. 2017; Barrera-Guzman et al. 2018). However, to date no population genetic models have been formulated that can provide explicit quantitative predictions of HHS. Such predictions would provide invaluable in characterizing the prevalence, causes, and consequences of HHS.

To address the aforementioned problems, here we develop an explicit model for the timing and direction of introgression based on the multispecies network coalescent (Yu et al. 2012; Yu et al. 2014; Wen et al. 2016b). The multispecies network coalescent model generalizes the multispecies coalescent (Hudson 1983; Rannala and Yang 2003) to allow for both incomplete lineage sorting and introgression (reviewed in Degnan 2018; Elworth et al. 2018). Under this model, a single sample taken from each of the extant lineages traces its history back through the network, following alternative paths produced by reticulations with a probability proportional to the amount of introgression that has occurred. The multispecies network coalescent model differs from population genetic approaches requiring multiple samples per population (such as the "IM" model; Nielsen and Wakeley 2001), though it does require that at least three species (and usually an outgroup) are sampled. However, the use of more than two lineages also makes it possible to more finely resolve the timing of migration, a problem that exists when analyzing sister species in the IM framework (Sousa et al. 2011).

Our study provides expectations for pairwise coalescence times under the multispecies network coalescent model. Two new statistics arise from these expectations, dubbed $D_1$ and $D_2$, which can be used alongside other information to test hypotheses regarding the timing and direction of intogression. We perform simulations to establish the power of these statistics when speciation and hybridization have occurred at various times in the recent past, with varying admixture proportions and effective population sizes. Finally, in order to demonstrate the use of the $D_1$ statistic, we apply it to a genomic dataset from the wild yeast species *Saccharomyces paradoxus*, a potential case of homoploid hybrid speciation (Leducq et al. 2016).

## MATERIALS AND METHODS

### A multispecies network coalescent model of introgression

Many statistics for detecting introgression, including the ABBA-BABA test, are based on expectations for the frequencies of different gene tree topologies. Our model, and the statistics that follow from it, are instead based on expected coalescence times—and the resulting levels of divergence—between pairs of populations or species. In what follows we present explicit expressions for these coalescence times, for genealogies evolving within a species network. Later in the paper we show how these times can complement and extend analyses based solely on the frequency of different topologies.

To make it easier to track the history of individual loci, we imagine that a species network can be separated into one or more "parent trees" (cf. Meng and Kubatko 2009; Liu et al. 2014). Every reticulation event in the network produces another parent tree, and loci with a history that follows such reticulate branches are considered to be produced by the corresponding parent tree. Embedded in a species network like the one shown in Figure 1a (where introgression is instantaneous) is one parent tree that represents the speciation history of lineages, which we define more formally here. Consider three taxa that have the phylogenetic relationship ((A,B),C). Let $t_1$ denote the time of speciation between A and B, and let $t_2$ denote the time of speciation between the common ancestor of A and B and lineage C (Figure 2). While speciation in nature is not an instantaneous process, in this idealized model these times represent the average split across loci. We additionally define $N$ as the effective population size of the ancestral population of all three taxa (i.e. the ancestor of taxa A, B, and C). These relationships, which depict the speciation history of the clade, will be referred to as "parent tree 1" (Figure 2a).

Because of the stochasticity of the coalescent process, parent tree 1 will generate one of four topologies at a particular locus: 1) A concordant tree in which A and B coalesce before $t_2$, denoted $AB1_1$ (Figure 2d); 2) A concordant tree in which A and B coalesce after $t_2$, denoted $AB2_1$ (Figure 2e); 3) A discordant tree where B and C are the first to coalesce after $t_2$, denoted $BC_1$ (Figure 2f); and 4) A discordant tree where A and C are the first to coalesce after $t_2$, denoted $AC_1$ (Figure 2g). The expected frequency of each of these topologies is a classic result from coalescent theory (Hudson 1983; Tajima 1983; Pamilo and Nei 1988). The expected coalescence times for each pair of species in each topology can also be found using straightforward properties of the coalescent model. There are three possible pairs of species, and therefore three times to coalescence in each of the four topologies. Fortunately, the symmetry of relationships means that many of these times are the same between pairs of species and across topologies. The equations that follow are presented in units of $2N$ generations for simplicity; see the Supplementary Materials for the full expressions.

For the gene tree $AB1_1$ in parent tree 1, the expected time to coalescence between A and B ($t_{\text{A-B}}$) is:

$$E[t_{A-B}|AB1_1] = t_1 + (1 - \frac{t_2 - t_1}{e^{(t_2 - t_1)} - 1}) \quad (1)$$

(equation A.7 in Mendes and Hahn 2018). Implicit in this equation is the effective population size of the internal branch of parent tree 1 (i.e. the ancestor of taxa A and B), which determines the length of the branch in coalescent units; from here forward, this population size is referred to as $N_1$. For the time to coalescence between pairs B-C and A-C in topology $AB1_1$, the lineages

must coalesce after $t_2$ (looking backward in time), and then the lineage ancestral to A and B is expected to coalesce with C in $2N$ generations. Therefore

$$E[t_{B-C}|AB1_1] = E[t_{A-C}|AB1_1] = t_2 + 1 \quad (2)$$

For topology $AB2_1$, A and B now coalesce after $t_2$, but before either coalesces with another lineage. The time to coalescence is therefore:

$$E[t_{A-B}|AB2_1] = t_2 + \frac{1}{3} \quad (3)$$

For pairs B-C and A-C in topology $AB2_1$, first A and B must coalesce (which takes $t_2 + \frac{2}{3}N$ generations on average), and then the lineage ancestral to A and B can coalesce with C. This means that the total expected time for both pairs is:

$$E[t_{B-C}|AB2_1] = E[t_{A-C}|AB2_1] = t_2 + \frac{1}{3} + 1 \quad (4)$$

In the discordant topology $BC_1$, species B and C coalesce before any other pair of taxa, and must do so in the ancestral population of all three lineages (after $t_2$). The time to coalescence is therefore the same as in equation 3:

$$E[t_{B-C}|BC_1] = t_2 + \frac{1}{3} \quad (5)$$

Similarly, the time to the common ancestor of pairs A-B and A-C in topology $BC_1$ follow the same coalescent history as the two pairs in equation 4:

$$E[t_{A-B}|BC_1] = E[t_{A-C}|BC_1] = t_2 + \frac{1}{3} + 1 \quad (6)$$

Finally, we have the topology $AC_1$, in which species A and C coalesce first. Their time to coalescence is:

$$E[t_{A-C}|AC_1] = t_2 + \frac{1}{3} \quad (7)$$

Likewise, the times to the common ancestor of pairs A-B and B-C both take the whole height of the tree to coalesce, so:

$$E[t_{A-B}|AC_1] = E[t_{B-C}|AC_1] = t_2 + \frac{1}{3} + 1 \quad (8)$$

The above expectations were derived under "parent tree 1," which represents the species tree. If there is introgression from species C into species B, individual loci with a history of introgression will now follow an alternative route through the species network, and therefore a new parent tree. Given that an introgression event between species B and C occurs before $t_1$ (looking back in time), the topology of this parent tree will be ((B,C),A). We refer to this topology as "parent tree 2" (Figure 2b). The expected coalescence times of gene trees evolving inside parent tree 2 are similar to those from parent tree 1, with a few key differences. First, because this parent tree has the topology ((B,C),A), the two "concordant" gene trees also have this topology. In other words, the coalescent expectations for species pair A-B in parent tree 1 are the same as those for pair B-C in parent tree 2. Second, although $t_2$ remains the same in the

two trees, the first lineage-splitting event is determined by the timing of the instantaneous introgression event in parent tree 2—which we will denote as $t_m$—rather than by the speciation time, $t_1$ (Figure 2b). Lastly, parent tree 2 can have a different internal-branch effective population size, which we here denote as $N_2$.

With these differences in mind, the expected coalescence times for each pair of species in each gene tree topology evolving in parent tree 2 are as follows:

$$E[t_{B-C}|BC1_2] = t_\mathrm{m} + (1 - \frac{t_2 - t_\mathrm{m}}{e^{(t_2 - t_\mathrm{m})} - 1}) \quad (9)$$

$$E[t_{A-B}|BC1_2] = E[t_{A-C}|BC1_2] = t_2 + 1 \quad (10)$$

$$E[t_{B-C}|BC2_2] = E[t_{A-B}|AB_2] = E[t_{A-C}|AC_2] = t_2 + \frac{1}{3} \quad (11)$$

$$E[t_{A-B}|BC2_2] = E[t_{A-C}|BC2_2] = E[t_{B-C}|AB_2] = E[t_{A-C}|AB_2] = E[t_{A-B}|AC_2] = E[t_{B-C}|AC_2] =$$

$$= t_2 + \frac{1}{3} + 1 \quad (12)$$

Many parent trees can be defined within a species network, based on the direction of introgression and the taxa involved in introgression. Here we consider one additional tree, denoted parent tree 3 (Figure 2c), which again represents gene flow between species B and C, but in this case from B into C. In this parent tree, the speciation time between species A and B remains $t_1$, but now this time also represents the first point at which A and C can coalesce—$t_2$ is not relevant. This is because the presence of loci from lineage B in lineage C allows C to trace its ancestry through B going back in time, which in turn allows it to coalesce with A after $t_1$. The time to first coalescence for lineages from B and C inside this tree is again limited by the timing of introgression, which again predates $t_1$, and which we assume occurs at the same time as introgression in parent tree 2 (i.e. $t_m$). We define the internal-branch effective population size for this parent tree as $N_3$.

In general, then, the difference between parent tree 3 and parent tree 2 is that all $t_2$ terms are replaced with $t_1$. Therefore:

$$E[t_{B-C}|BC1_3] = t_\mathrm{m} + (1 - \frac{t_1 - t_\mathrm{m}}{e^{(t_1 - t_\mathrm{m})} - 1}) \quad (13)$$

$$E[t_{A-B}|BC1_3] = E[t_{A-C}|BC1_3] = t_1 + 1 \quad (14)$$

$$E[t_{B-C}|BC2_3] = E[t_{A-B}|AB_3] = E[t_{A-C}|AC_3] = t_1 + \frac{1}{3} \quad (15)$$

$$E[t_{A-B}|BC2_3] = E[t_{A-C}|BC2_3] = E[t_{B-C}|AB_3] = E[t_{A-C}|AB_3] = E[t_{A-B}|AC_3] = E[t_{B-C}|AC_3] =$$

$$= t_1 + \frac{1}{3} + 1 \quad (16)$$

**The $D_1$ statistic for the relative timing of introgression**

Given the expectations laid out above, we can now develop statistics that differentiate alternative biological scenarios. The first comparison we wish to make is between models of speciation followed by introgression (Figure 1a) and models where speciation and introgression are simultaneous (Figure 1b); the latter scenario corresponds to homoploid hybrid speciation or the creation of a new admixed population. We assume for now that introgression has occurred in the direction from C into B in both scenarios, as such cases will be the hardest to distinguish.

The distinguishing feature between these two biological scenarios is the timing of introgression relative to speciation or lineage-splitting, and therefore the expected coalescence times between sequences from species B and either species A or C. If introgression occurs after speciation, we expect that loci which follow the species tree embedded in the network (i.e. parent tree 1) will coalesce further back in time than loci that follow the introgression history in the network (i.e. parent tree 2). Figure 3 demonstrates this expected difference graphically. Explicitly, the expected coalescent times between A and B from parent tree 1 largely mirror the times between B and C from parent tree 2, except for genealogies that coalesce before $t_2$, where $E[t_{A-B}|AB1_1]$ depends on $t_1$ (equation 1), while $E[t_{B-C}|BC1_2]$ depends on $t_m$ (equation 9).

This difference captures information on the timing of introgression relative to speciation. If $t_1$ and $t_m$ are equal, speciation and introgression are effectively simultaneous, as would be the case in homoploid hybrid speciation. In this case, $E[t_{A-B}|AB1_1] = E[t_{B-C}|BC1_2]$, and the expected difference between these times is 0. If introgression has occurred significantly before speciation (going backward in time), $t_1 > t_m$, and therefore, $E[t_{A-B}|AB1_1] - E[t_{B-C}|BC1_2] > 0$. This difference will be larger the more recent the introgression event is relative to speciation.

Importantly, this expectation is conditional on $N_1$ and $N_2$ being equal; if they are not, the patterns produced by the alternative scenarios will depend both on the timing of introgression relative to speciation and the degree and direction of variation in $N$. Changing the value of $N_2$ affects the rate of coalescence after time $t_m$ and the height of genealogies coming from parent tree 2. However, the model presented in the previous section can incorporate this variation, and it can be incorporated into our test statistic (see below).

In developing a statistic to distinguish these scenarios that can be applied to real data, there are two important things to note. First, expected coalescent times can easily be used to model expected amounts of divergence through a simple multiplication by $2\mu$, where $\mu$ is the mutation rate per generation (assuming a constant mutation rate throughout the tree). As divergence can be measured directly from sequence data, the statistics presented here will be in terms of divergence. Second, we cannot know whether the gene tree topology at any given locus was generated by introgression or by incomplete lineage sorting. More importantly, we also do not know if the gene tree originates from parent tree 1 or parent tree 2. This affects both the theoretical expectation and empirical calculation of any statistic.

Ideally, in our formulation of a test statistic, we would like to ignore all irrelevant terms and simply take the difference $E[t_{A-B}|AB1_1] - E[t_{B-C}|BC1_2]$. However, due to the aforementioned practical constraints, this is not possible. In a dataset where the only available data for a given locus is the gene tree topology and pairwise genetic divergence, the most useful

test would measure genetic divergence conditional on a specific tree topology. While we cannot assign parent trees to each gene tree, we know from the coalescent process that the majority of gene trees with the topology ((A,B),C) will come from parent tree 1, and likewise that the majority of gene trees with the topology ((B,C),A) will come from parent tree 2. Therefore, if we measure the distance between A and B in all trees with the ((A,B),C) topology (denoted $d_{A\text{-}B}|AB$), and the distance between B and C in all trees with the ((B,C),A) topology ($d_{B\text{-}C}|BC$), we should be able to capture most of the difference in coalescence times caused by differences between $t_1$ and $t_m$.

Based on these considerations, we define a statistic to test the hypothesis that $t_1$=$t_m$ as:

$$D_1 = (d_{A-B}|AB) - (d_{B-C}|BC) \tag{17}$$

In terms of the coalescence times and genealogies defined above, weighting each expectation by the frequency of the relevant gene tree (denoted with *f* terms):

$$E[D_1] = (f_{AB1_1}E[t_{A-B}|AB1_1] + f_{AB2_1}E[t_{A-B}|AB2_1] + f_{AB_2}E[t_{A-B}|AB_2])$$
$$- (f_{BC1_2}E[t_{B-C}|BC1_2] + f_{BC2_2}E[t_{B-C}|BC2_2] + f_{BC_1}E[t_{B-C}|BC_1]) \tag{18}$$

These expectations explicitly recognize that the origin of any genealogy cannot be known, so that $D_1$ must average across all loci with the same genealogy (either *AB* or *BC*). Expectations for the frequencies of gene trees are given in the Supplementary Materials.

This definition of the $D_1$ statistic introduces another factor that adds complexity to the expected patterns of divergence: the admixture proportion, which we denote as γ. We call $\gamma_2$ the fraction of gene trees following the introgression history through the network (i.e. originating from parent tree 2). This parameter therefore determines the fraction of gene trees that come from parent tree 1 as well; see the Supplementary Materials for equations describing these expectations explicitly. The effect of γ on gene tree frequencies also interacts with variation in *N*; while γ describes the fraction of loci that originate from a particular parent tree, *N* determines the distribution of topologies that such a parent tree is expected to generate. Under the assumptions that $N_1 = N_2$ and $\gamma_2 = 0.5$, we expect $D_1$ to take on a value of 0 for the case of hybrid speciation and a positive value for the case of introgression after speciation. In all other circumstances, the expected value of $D_1$ under hybrid speciation will depend on the interplay of $N_1$, $N_2$ and $\gamma_2$ (see equation SXX in Supplementary Materials).

**The $D_2$ statistic for the direction of introgression**

We also wish to make the distinction between different directions of introgression in terms of our model. As explained above, introgression from C into B generates a different reticulation in the species network, and therefore a different parent tree, than introgression from B into C (compare Figure 2b to 2c). Gene flow from lineage B into lineage C allows loci sampled from C currently to trace their history back through B, which in turn allows lineages A and C to coalesce more quickly (specifically, after $t_1$ instead of after $t_2$). Conversely, gene flow from lineage C into lineage B does not change our expectations for the coalescence time of A and C relative to those expected from the species relationships. We can take advantage of these

expected differences in the amount of divergence between A and C to develop a statistic for inferring the direction of introgression.

When introgression occurs in the C into B direction, only parent trees 1 and 2 are relevant. The coalescent expectations between A and C from parent tree 1 (equations 2, 4, 6, and 7) exactly mirror those from parent tree 2 (equations 10, 11 and 12). When introgression occurs from B into C, parent trees 1 and 3 become relevant. Coalescence times between A and C in parent tree 3 are all truncated, depending on $t_1$ instead of $t_2$ (equations 14, 15 and 16). An obvious distinction between histories is therefore the distance between A and C. In a manner similar to how we defined $D_1$, we can measure divergence between A and C conditional on alternative gene trees—either ((A,B),C) or ((B,C),A)—with the expectation that the former topology will arise primarily from parent tree 1, and the latter primarily from parent tree 2 or 3. Therefore, we define our statistic for inferring the direction of introgression as:

$$D_2 = (d_{A-C}|AB) - (d_{A-C}|BC) \tag{19}$$

The expected value of $D_2$ will depend on the direction in which gene flow has occurred. In the case of introgression from C into B, using the same formulation as for $D_1$, we have:

$$\begin{aligned} E[D_2|C \rightarrow B] &= (f_{AB1_1}E[t_{A-C}|AB1_1] + f_{AB2_1}E[t_{A-C}|AB2_1] + f_{AB_2}E[t_{A-C}|AB_2]) \\ &\quad - (f_{BC1_2}E[t_{A-C}|BC1_2] + f_{BC2_2}E[t_{A-C}|BC2_2] + f_{BC_1}E[t_{A-C}|BC_1]) \\ &= 0 \end{aligned} \tag{20}$$

This expression represents the expectation of $D_2$ when all of our assumptions have been met: when introgression has occurred only from C into B, the statistic should not be significantly different from 0 because the distance between A and C is the same in parent trees 1 and 2. When introgression has occurred from B into C, we have:

$$\begin{aligned} E[D_2|B \rightarrow C] &= (f_{AB1_1}E[t_{A-C}|AB1_1] + f_{AB2_1}E[t_{A-C}|AB2_1] + f_{AB_3}E[t_{A-C}|AB_3]) \\ &\quad - (f_{BC1_3}E[t_{A-C}|BC1_3] + f_{BC2_3}E[t_{A-C}|BC2_3] + f_{BC_1}E[t_{A-C}|BC_1]) \end{aligned} \tag{21}$$

In the expectation of our $D_2$ statistic, all the relevant gene trees coalesce after $t_2$, and therefore their rates of coalescence are determined by the same $N$. Because of this, we expect variation in $N$ to be less relevant to $D_2$. However, unequal ancestral population sizes will still affect the expectation of this statistic for introgression from B into C, and the expectations for both directions are expected to be affected by $\gamma_2$ or $\gamma_3$ (see Supplementary Materials). Therefore, like the $D_1$ statistic, we expect the null hypothesis to be 0 and the alternative to be a positive value under the assumptions of equal population sizes and $\gamma_2$ or $\gamma_3 = 0.5$. In other cases, the expected values of the statistics will come from an interplay of these parameters (see Supplementary Materials).

**Simulations**

To investigate the behavior of $D_1$ and $D_2$, we explored the parameter space of our model using simulated genealogies from the program *ms* (Hudson 2002). To simulate an introgression

event, we took two different approaches in order to ensure the robustness of our results. In the first approach, gene trees were simulated from parent trees separately and then combined in a single dataset, with the $t_m$ parameter specified for parent tree 2 or 3 representing the timing of the introgression event. In the second approach, we simulated gene trees directly from a species network by specifying an effectively instantaneous population splitting and merging event from the donor to the recipient at time $t_m$. The specific command lines used in *ms* for both approaches can be found in the Supplementary Materials.

For each statistic, we investigated the effects of the parameters defining the null and alternative hypotheses; for $D_1$ this is the difference in the timing of speciation and introgression ($t_1 - t_m$), and for $D_2$ it is the difference in the timing of lineage-splitting events ($t_2 - t_1$). Values of the statistics were calculated directly from the branch lengths of simulated genealogies. We also investigated the effects of variation in the effective population size (specifically, the value of $N_1$ relative to $N_2$ or $N_3$) and admixture proportion ($\gamma_2$ or $\gamma_3$). Lastly, we investigated the effect that introgression in both directions has on our statistics. We performed 100 simulations of 2000 genealogies each, for seven different values each of $t_1 - t_m$, $t_2 - t_1$, $N_1 / N_2$ or $N_1 / N_3$, and $\gamma_2$ or $\gamma_3$, unless specified otherwise. The relevant parameters were simulated for each direction for the $D_2$ statistic. We also performed a set of simulations for both statistics in which introgression occurred in both directions rather than only one.

We used a parametric bootstrap approach to evaluate the statistical significance of simulated $D$ statistics. We used a two-tailed significance test in which the observed value of the statistic is given a rank $i$ in relation to the simulated null distribution, and the $P$-value is calculated as:

$$P = 1 - 2 * |0.5 - i|$$

(Voight et al. 2005). We used the same approach to estimate the false negative rate and false positive rate for each combination of parameters. These values were, respectively, the proportion of simulated statistics that incorrectly accept or reject the null hypothesis at a particular significance level.

**Data from *Saccharomyces paradoxus***

To demonstrate the use of our model to test for homoploid hybrid speciation, we analyzed genomic data from three lineages (and an outgroup) of the North American wild yeast species *Saccharomyces paradoxus* (Leducq et al. 2016). This study identified three genetically distinct populations of *S. paradoxus*: two parent lineages dubbed *SpB* and *SpC*, and a hybrid lineage dubbed *SpC**. Analysis of whole-genome sequences from these populations shows that *SpC** has a mosaic genome, the majority of which is similar to *SpC*, with small genomic regions that are more similar to *SpB*. An investigation of reproductive isolation among all three lineages led Leducq et al. (2016) to conclude that *SpC** represents a case of homoploid hybrid speciation.

To evaluate this hypothesis, genomic data from all 161 strains was acquired from the authors. We then separated aligned genomic 5-kb windows into two categories, depending on the assigned topology in Leducq et al. (2016). Windows assigned "ANC" by Leducq et al. represent

loci where the topology has *SpC** sister to *SpC*; in terms of our definition of $D_1$, the distance between *SpC** and *SpC* at these loci corresponds to $d_{A-B}|AB$, under the assumption that the alternative history to hybrid speciation has *SpC** and *SpC* as sister species. There are a number of genomic windows in which the topology has *SpC** and *SpB* as sister lineages, but some of them are only found in particular groups of strains. We picked windows assigned as "H0" (found in all strains) and "H1b" (found in all strains but one) by Leducq et al. (2016) for our analysis. The distance between *SpC** and *SpB* in these windows corresponds to $d_{B-C}|BC$ in our formulation of $D_1$.

For all 161 strains in each of the three lineages (and an outgroup), the total dataset consists of sequence alignments from ANC-topology windows and H0/H1b-topology windows. To carry out calculations of $D_1$ we then randomly chose windows from 100 different combinations of a single strain from each of *SpC**, *SpC* and *SpB*. For each of these 100 samples, we calculated $d_{A-B}|AB$ and $d_{B-C}|BC$ in 5-kb windows simply as the proportion of nucleotide sites that differed between strains; levels of divergence in this system are low enough that no correction for multiple hits is necessary (see Results). For the ANC-topology windows, we calculated $d_{A-B}|AB$ in every other window to reduce the effects of autocorrelation between windows in close physical proximity; the H0/H1b-topology windows were sufficiently few in number and spaced far enough apart so that this was unnecessary for $d_{B-C}|BC$. $D_1$ was calculated as the difference in the mean of these two groups. Filtering and distance calculations from all genomic windows were carried out using the software package MVFtools (Pease and Rosenzweig 2018).

To evaluate our observed distribution of $D_1$ statistics with respect to the HHS hypothesis, we performed a set of simulations corresponding to an HHS scenario using parameters estimated from the study. First, we calculated average genome-wide per-site expected heterozygosity ($\pi$) as an estimator for the population-scaled mutation rate, $\theta$, in both *SpC* and *SpB*. We used the estimates of $\pi$ from these populations as proxies for $\theta$ along the internal branches of the ANC and H0/H1b topologies, respectively. To estimate the population parameters $N_1$, $N_2$, $t_2$, and $t_1$, we used our estimates of $\theta$ in conjunction with per-generation mutation rates and generation times from *S. cerevisiae* (Fay and Benavides 2005, Zhu et al. 2014), as well as divergence time estimates from Leducq et al. (2016). We simulated 10000 datasets under the assumption that $t_1$ and $t_m$ are equal (as would be the case under HHS). Each dataset consisted of 2002 "ANC" loci and 55 "H0/H1b" loci, sampled from the relevant parent tree.

**Data availability**

Code used to generate the simulated genealogical data is available in the Supplementary Materials. Genomic alignments used in the analysis of *S. paradoxus* are from Leducq et al. (2016), and are uploaded to figshare through the GSA portal at https://gsajournals.figshare.com/submit along with the Supplementary Materials.

## RESULTS

402 **Power of $D_1$ to distinguish alternative histories**

403 To determine the power of our new statistic, $D_1$, we asked whether it could distinguish
404 between a history of speciation followed by introgression and a history of homoploid hybrid
405 speciation (HHS), under different regions of the parameter space of our model. Under HHS,
406 speciation and introgression happen simultaneously (i.e. $t_1$=$t_m$), and the expected value of $D_1$ is 0
407 (equations 17 and 18), assuming that $N_1 = N_2$ and $\gamma_2 = 0.5$. As the time between speciation and
408 introgression increases, the value of $D_1$ should increase linearly.

409 As expected from our model, simulated values of $D_1$ are centered around a mean of 0
410 when $t_1$–$t_m$=0, $N_1 = N_2$, and $\gamma_2 = 0.5$, with values of $D_1$ increasing linearly as the introgression
411 event becomes more recent relative to speciation (Figure 4a). In this region of parameter space
412 introgression that occurs shortly after speciation may be difficult to distinguish from HHS; we
413 observe false negative rates (incorrectly accepting null hypothesis of HHS) of 58 and 70% when
414 $t_1 - t_m = 0.05$ (Supplementary Table S2). Values of $t_1 - t_m$ of 0.1 or greater can always be
415 distinguished from HHS under these assumptions (Supplementary Table S2).

416 One factor that may reduce the power of $D_1$ is the presence of introgressed topologies
417 from additional parent trees. Our model predicts that $D_1$ should have reduced power when
418 introgression occurs in both directions (i.e. C→B and B→C), due to the presence of gene trees
419 generated by parent tree 3. The most common gene trees generated by parent tree 3 have the
420 same topology as those generated by parent tree 2, but the shorter coalescence time between
421 lineages A and B in this parent tree make tests based on $D_1$ conservative. To investigate the
422 magnitude of this effect, we simulated across the same values of $t_1$–$t_m$, now specifying equal
423 contributions of parent trees 2 and 3. Figure 4b shows that introgression shortly after speciation
424 in both directions results in negative values of $D_1$, and that $D_1$ can be centered around 0 when $t_1$–
425 $t_m$ is not 0. Here, the relative magnitude of $\gamma_2$ and $\gamma_3$ matters: increasingly large values of $\gamma_3$
426 relative to $\gamma_2$ will magnify this effect.

427 We expected two other parameters to affect the value of $D_1$: the difference in effective
428 population size ($N_1/N_2$) and the admixture proportion ($\gamma_2$). Differences in $N$ affect the
429 coalescence times of individual genealogies as well as the degree of incomplete lineage sorting
430 within each parent tree, while the admixture proportion interacts with the degree of incomplete
431 lineage sorting to determine the likely parent tree of origin for a particular gene tree topology. To
432 investigate these effects, we simulated across seven values both of $N_1/N_2$ and $\gamma_2$. Our results
433 (Figure 5; Supplementary Tables S1 and S2) show that the $D_1$ statistic is affected by variation in
434 both $N_1/N_2$ and $\gamma_2$. A two-fold difference in $N$ or a difference of 15% in $\gamma_2$ are enough to
435 essentially guarantee that, in the case of HHS, $D_1$ will deviate significantly from the idealized
436 expectation of 0. There are also regions of parameter space in which the statistic is not likely to
437 deviate from 0, even when there has been introgression after speciation. This occurs when the
438 effect of either $N_1$ / $N_2$ or $\gamma_2$ effectively "cancels out" the signals of divergence, leading to $D_1$
439 values close to 0. This risk appears to be highest when $N$ is smaller in parent tree 1 (Figure 5a),
440 or for values of $\gamma_2$ between 0.05 and 0.5 (Figure 5b). Our expectations for $D_1$ can also take into
441 account variation in $N$ and $\gamma$, if these quantities can be estimated (Supplementary equations 1 -

20). In such cases—and when simulations are used to generate the null (see below)—we do not have to use the expectation that $D_1 = 0$.

Finally, the results obtained from our two different simulation approaches are virtually identical (see Figure S1), confirming that they both reflect valid ways of simulating introgression.

**Power of $D_2$ to determine the direction of introgression**

We investigated the power of the $D_2$ statistic to distinguish between different directions of introgression, and how this power is affected by the time between speciation events, $t_2 - t_1$, in addition to the two other parameters described above. Our model predicts, under our previously stated assumptions, that introgression in the C→B direction should produce values of $D_2$ not different from 0, while introgression in the B→C direction should produce values significantly larger than 0. Furthermore, the magnitude of this difference should increase linearly as a function of $t_2 - t_1$, increasing the power of the test.

The results of our simulations, shown in Figure 6a, confirm the predictions of the model. In the B→C direction, the value of $D_2$ increases linearly as a function of $t_2 - t_1$, whereas $D_2$ values remain centered around 0 for introgression in the C→B direction. For the particular introgression scenario examined (introgression 0.6$N$ generations after speciation), our statistic always has excellent power to distinguish between the two directions (Supplementary Tables S3 and S4). The trend of our simulated results suggests that as the time between speciation events continues to decrease, the power of the $D_2$ statistic will be reduced. Therefore, our statistic may be prone to accepting the null hypothesis (C→B introgression) when two speciation events are followed very closely in time.

Introgression in both directions may reduce the power of $D_2$ to reject the null, again similarly to the behavior of $D_1$. This is due to the presence of ((B,C),A) gene trees concordant with parent tree 2, which share the same A-C divergence times as ((A,B),C) trees concordant with parent tree 1. We investigated this prediction, and how it interacts with the time between speciation events, with a set of simulations including equal contributions of parent trees 2 and 3, again 0.6$N$ generations after speciation. The simulated results confirm the predictions of our model (Figure 6b). The power of the statistic to reject C→B introgression alone is reduced compared to the same values of $t_2 - t_1$ when introgression is B→C, but power may still be good if the time between speciation events is high enough (Figure 6b).

We also tested whether variation in $N_1$ / $N_2$ and $\gamma_2$ would affect the $D_2$ statistic in the same ways they affect $D_1$. To do this, we simulated across the same range of parameters for the $D_2$ statistic. The results of these simulations suggest that $D_2$ is substantially more robust to variation in $N$ than $D_1$ (Figure 7a). Moderate false positive rates (incorrectly rejecting null of C → B introgression) of 10-20% begin to appear only when $N_1$ is 2/3 of $N_2$ or lower, and there appears to be no affect when $N_1$ is larger than $N_2$ (Supplementary Table S3). The statistic may be sensitive to false negatives (incorrectly accepting null of C → B) if $N_1$ is lower than $N_2$ and there is a short time between lineage-splitting events (ie. $t_2 - t_1$ is small); our simulations show a false negative rate of 34% when $N_1$ is half of $N_2$ and $t_2 - t_1$ is 0.5 (Supplementary Table S4). $D_2$ also

482 appears to be sensitive to variation in $\gamma_2$ (or $\gamma_3$, depending on the direction of introgression).
483 Values of $\gamma_2$ less than or equal to 0.2, or greater than or equal to 0.8, can lead to high false
484 positive rates (Supplementary Table S3). Unlike $D_1$, the false positive rate of $D_2$ may be robust to
485 some variation in $\gamma_2$ if $t_2 - t_1$ is large; in our simulations, the false positive rate remains at or
486 below standard significance at $\gamma_2 = 0.65$ when $t_2 - t_1$ is 0.6 (Supplementary Table S3). $D_2$ can
487 also have a high false negative rate in certain regions of parameter space: this is again caused by
488 the effects of $\gamma_3$ negating the divergence time signal. Our simulations suggest that $D_2$ is generally
489 at a higher risk for false negatives when $\gamma_3$ is small (0.35 or less), but this effect interacts with $t_2$
490 $- t_1$ in a somewhat complex way (Supplementary Table S4).

491 Our simulation results highlight the fact that the signal the $D_2$ statistic detects is that of
492 B→C introgression, regardless of whether introgression also happened in the other direction.
493 Therefore, a significantly positive value of $D_2$ cannot explicitly distinguish between B→C
494 introgression alone and B→C introgression coupled with some C→B introgression. Conversely,
495 a non-significant value of $D_2$ does not rule out the presence of some B→C introgression. As
496 with $D_1$, the relative magnitude of the contributions of parent trees 2 and 3 will affect this result,
497 as will the timing between speciation events. Therefore, the most accurate way to interpret any
498 value of $D_2$ is to state the primary direction of introgression, rather than stating with certainty
499 that introgression occurred only in one direction or another.

500 **Analysis of *S. paradoxus***

501 To demonstrate the use of our model to test the hypothesis of homoploid hybrid
502 speciation, we calculated $D_1$ from three lineages of the wild yeast *Saccharomyces paradoxus* and
503 an outgroup and obtained a value of -0.0004. If the hypothesis that the *SpC** lineage of *S.*
504 *paradoxus* is a hybrid species is correct (cf. Leducq et al. 2016), then we would expect $D_1$ to
505 follow the distribution obtained from our simulations under a homoploid hybrid speciation
506 scenario. A $D_1$ value significantly deviating from this expectation would indicate a bad fit to the
507 HHS hypothesis.

508 To interpret our empirical estimate of $D_1$ we simulated the *S. paradoxus* system under
509 HHS. We used our estimated values of $\theta$ of $2.23 \times 10^{-4}$ and $7.37 \times 10^{-4}$ for the *SpC* and *SpB*
510 populations, respectively, and values of the per-generation mutation rate and number of
511 generations per year of $1.84 \times 10^{-10}$ and 2920, respectively (Fay and Benavides 2005, Zhu et al.
512 2014). $N_1$ and $N_2$ were estimated at approximately $3 \times 10^5$ and $1 \times 10^6$, while divergence times in
513 units of $4N$ generations were estimated at 112 and 20.2 for $t_2$ and $t_1$, respectively (see
514 Supplementary Materials for details as to how all parameters were calculated.).

515 Using these population parameters, we simulated 10000 $D_1$ statistics under a homoploid
516 hybrid speciation scenario. Our simulations replicate the pattern observed in the data of a total
517 absence of incomplete lineage sorting, with all trees coming from a single parent tree having a
518 single topology. The simulations also show that under these parameters the value of $D_1$ is
519 expected be negative (Figure 8). However, the mean value of $D_1$ = -0.0004 observed in the *S.*
520 *paradoxus* system is highly unlikely to have arisen under our simulated hybrid speciation

521 scenario ($P < 1.0 \times 10^{-4}$, rank significance test described in Methods; Figure 8). This result
522 strongly rejects homoploid hybrid speciation for the *S. paradoxus* system.

523

524 **DISCUSSION**

525 There are now multiple methods that use only one sequence per lineage to detect the
526 presence of gene flow between species (reviewed in Elworth et al. 2018). These methods all take
527 advantage of the fact that expectations for the frequency of different gene tree topologies can
528 easily be calculated under ILS, with deviations from these expectations often indicating the
529 presence of gene flow. However, gene tree topologies alone, without branch lengths, cannot
530 distinguish among various biological scenarios involving introgression (Zhu and Degnan 2017).
531 In particular, the scenarios represented in Figure 1a and 1b cannot be distinguished solely using
532 tree topologies, leading to general confusion and a proliferation of claims about "hybrid species."
533 The goal of the current study is to model possible histories of gene flow under the multispecies
534 network coalescent, and to present two new test statistics to explicitly differentiate among such
535 histories. While neither of these statistics should be used as a test for the presence of
536 introgression itself, they complement other widely used statistics that can be used for this
537 purpose and that depend on the same sampling scheme. In what follows we discuss the
538 limitations and implications of this work.

539 ***Limitations of our model***: Our model of gene flow between lineages is simplified in
540 multiple ways, and application to real data will often have to confront several assumptions we
541 have made. We have assumed that a single, instantaneous introgression event leads to the
542 generation of a single alternative history to the species tree. If instead there are multiple
543 introgression events, each additional event will generate a new reticulation in the species
544 network (and an additional parent tree), in the extreme producing infinitely many parent trees for
545 continuous stretches of introgression. Under such scenarios we expect an increase in the
546 variance in coalescent times, but still expect our statistics to capture the main history of gene
547 flow (e.g. Figure 5b). One exception is discussed further in the next section when considering
548 hypotheses about homoploid hybrid speciation.

549 We have modeled gene flow as a "horizontal" edge in the network, either because of
550 post-speciation introgression (Figure 1a) or hybrid speciation (Figure 1b). In such
551 representations the migrant individuals do not have an evolutionary history that is independent of
552 the donor lineage. In contrast, the representation in Figure 1c uses "non-horizontal" edges to
553 model gene flow, possibly indicating histories in which migrant individuals can evolve
554 independently (Degnan 2018). Such histories could possibly indicate biological scenarios
555 involving lineage fusion (e.g. Kearns et al. 2018) or where unsampled or extinct lineages are the
556 donor population, but seem much less relevant to most gene flow events. However, this
557 representation is commonly used in two settings. First, non-horizontal migration edges are often
558 used to indicate the direction of introgression after speciation, with the first bifurcation in Figure
559 1c representing speciation between species A and B, and the second representing gene flow from
560 C into B (e.g. Huson et al. 2010). This use does not imply an independent history, but instead

561 simply the direction of introgression. Second, some methods for inferring the topology of a
562 species network require or include non-horizontal edges (e.g. Yu et al. 2014; Solís-Lemus et al.
563 2017; Zhang et al. 2018). Although this choice adds parameters to these models, it also makes
564 some calculations easier. Despite the computational convenience, this seems to generally be a
565 less biologically realistic choice. Furthermore, the choice of horizontal edges, as is used here,
566 makes a clear distinction between the species history and any introgression histories; models that
567 use non-horizontal edges cannot distinguish between such alternatives (cf. Wen et al. 2016a).

568 Our model assumes that coalescence times and gene tree frequencies follow neutral
569 expectations. In the presence of either hitchhiking or background selection, $N_e$ will be reduced,
570 reducing coalescence times and increasing the concordance of gene trees with their respective
571 parent trees. This latter consequence should actually improve the power of our tests by reducing
572 the incomplete lineage sorting that occurs within each parent tree. A further complication may
573 be introduced because of the interaction between selection and introgression: in a number of
574 systems introgression appears to occur more frequently in regions of the genome with higher
575 recombination, which are less affected by linked selection (Geraldes et al. 2011; Brandvain et al.
576 2014; Aeschbacher et al. 2017; Schumer et al. 2018b). Because such regions will have larger $N_e$,
577 they may show more discordance and longer times to coalescence than average loci. In order to
578 overcome such confounding factors, it may be best to make comparisons among genomic
579 windows with similar recombination rates.

580 Finally, we have assumed that trees can be identified from individual non-recombining
581 windows of the genome. In reality, such windows are likely to be very small (e.g. Mendes et al.
582 2018), and averaging multiple non-recombining windows together may lead to the reconstruction
583 of an incorrect topology (Schierup and Hein 2000; Kubatko and Degnan 2007; Martin and Van
584 Belleghem 2017). While this implies that genomic windows should not be too big, neighboring
585 non-recombining segments will have correlated topologies, and inferred topologies from regions
586 on the order of the length of single genes are not very different from the true topology (Lanier
587 and Knowles 2012).

588 ***Implications for homoploid hybrid speciation***: It has become increasingly apparent that
589 HHS is a process that can generate new diversity quickly, with several charismatic examples
590 now known (e.g. in Darwin's finches; Lamichhaney et al. 2018). Characterizing the frequency
591 with which HHS occurs in nature is therefore important for our understanding of speciation and
592 the evolution of reproductive isolation, though strict criteria for identifying true cases of HHS
593 have been lacking. Schumer et al. (2014) proposed three pieces of evidence that are required to
594 demonstrate HHS: 1) evidence of introgression, 2) evidence of reproductive isolation of the
595 hybrid lineage from both parents, 3) evidence of a causal link between introgression and
596 reproductive isolation. While relatively standard methods exist for evaluating criteria 1 and 2, it
597 is much more difficult to explicitly evaluate criterion 3. Our $D_1$ statistic is unique in that it has a
598 specific distribution of expected values under a hybrid speciation scenario, which can be
599 predicted precisely using modelling and/or simulation. Therefore, it provides an explicit test of
600 criterion 3 by asking whether speciation and introgression are effectively simultaneous. Such a
601 relationship would strongly imply a causal link.

602 A commonly employed expectation for HHS is that there should be an approximately
603 50:50 split of two contrasting histories in the genome of the hybrid, as would be expected if each
604 parent species contributed equally. However, this pattern may be misleading for at least two
605 reasons. First, not all hybrid species are the result of isolation caused in the F1 generation of
606 crosses between two species. For example, the hybrid butterfly *Heliconius heurippa* likely arose
607 through two generations of backcrossing, resulting in an 82.5:12.5 pattern of ancestry (Mavarez
608 et al. 2006). Selection or drift may also cause deviations from 50:50 expectations in cases of true
609 HHS. Second, introgression without hybrid speciation can be extensive, affecting 50% of the
610 genome or more (e.g. in *Anopheles* mosquitoes; Fontaine et al. 2015; Wen et al. 2016a). Our $D_1$
611 statistic overcomes this limitation by explicitly allowing the admixture proportion to vary when
612 predicting its expected value under an HHS scenario.

613 There are several other biological scenarios in which the $D_1$ statistic in particular may be
614 misleading, resulting in either incorrect rejection or acceptance of HHS. If hybrid speciation is
615 followed by extinction of one parent lineage, then one sampled taxon will be more distantly
616 related to the hybrid than the other; this will lead to values of $D_1$ inconsistent with HHS even
617 though it has occurred. Similarly, if introgression occurs after hybrid speciation, the value of $D_1$
618 could be dominated by the more recent event, again leading to false rejection of HHS. While
619 both of these scenarios are problems for $D_1$, they would also be problems for any other methods
620 attempting to distinguish HHS from introgression-after-speciation. Lastly, if introgression has
621 occurred shortly after speciation, but is not causally related to it, there simply may not be enough
622 signal in the data to distinguish this scenario from one in which they are simultaneous.

623 ***Application of $D_1$ to an empirical example in yeast***: One clade of the yeast species
624 *Saccharomyces paradoxus* (denoted *SpC**) has been proposed to be a homoploid hybrid (Leducq
625 et al. 2016). Our estimate of $D_1$ from genomic data suggests that it is highly unlikely to have
626 arisen under a hybrid speciation scenario; here, we discuss the implications of this result for the
627 system and for attempts to infer HHS in general.

628 In their analysis of the genome of *SpC*,* Leducq et al. (2016) concluded that each locus
629 could be classified into one of two topologies: either [((*SpC**,*SpC*),*SpB*)], which comprises 92-
630 97% of the genome, and [((*SpC**,*SpB*),*SpC*)], which comprises the remainder. The absence of a
631 third gene tree topology, [((*SpC*,*SpB*),*SpC**)], leads us to believe that incomplete lineage sorting
632 is minimal or entirely absent in this system. Therefore, this represents a unique case in which the
633 genome is comprised of two specific gene trees whose estimated coalescence times can be
634 compared directly: a gene tree concordant with the species history (as described by equation 1)
635 and a gene tree concordant with a history of introgression (as described by equation 9). Our
636 simulations using population parameters from the *S. paradoxus* system replicate the observation
637 that incomplete lineage sorting is absent from the data.

638 Given the results of our analysis, the simplest explanation for the observed pattern is that
639 there was introgression between *SpC** and *SpB* after the split of *SpC* and *SpC**. It is also possible
640 that *SpC** originated as a hybrid species that has since undergone further introgression only with
641 *SpB*. While it may be difficult to distinguish these two hypotheses, it is clear from our results that

642 the primary signal carried in the data is one of introgression following speciation, rather than
643 homoploid hybrid speciation.

644 ***Considerations for the* $\mathbf{D_2}$ *statistic***: We have also introduced the $D_2$ statistic for
645 distinguishing the direction of introgression between two taxa (see Rouard et al. 2018 for an
646 empirical example using this test). Despite the importance of understanding the direction of
647 gene flow, relatively few studies have explicitly attempted to provide a solution to the problem
648 when only a single sequence is sampled from each lineage. Pease & Hahn (2015) developed a set
649 of $D_{\mathrm{FOIL}}$ statistics which can infer the direction of introgression in such cases; however, these
650 statistics require information from four ingroup taxa, and the taxa in question must have a
651 symmetrical tree topology. These considerations limit the generality of $D_{\mathrm{FOIL}}$ statistics.

652 The $D_2$ statistic can be calculated from three ingroup taxa and an outgroup; this is similar
653 to the sampling used for the original $D$ statistic (Green et al. 2010) and other related tests using
654 gene tree topologies (e.g. Huson et al. 2005). Using only the frequencies of topologies (or
655 nucleotide site patterns that reflect these underlying topologies), such tests cannot distinguish the
656 direction of introgression. While the internal branches on the two major tree topologies
657 produced by alternative directions of introgression do differ in length (i.e. from $t_2$ to $t_{\mathrm{m}}$ in parent
658 tree 2 vs. from $t_1$ to $t_{\mathrm{m}}$ in parent tree 3; Figure 2), this difference is not detectable using the $D$
659 statistic (Martin et al. 2015). There are of course population genetic methods for inferring the
660 direction of gene flow, and these can even be used between pairs of sister species (e.g. Lohse et
661 al. 2016). However, these methods both require more sequences from each taxon and are more
662 sensitive to violations of their assumptions due to linked selection because they are more
663 dependent on the variance in coalescence times. $D_2$ also appears to be more robust than $D_1$ to
664 variation in $N$ between populations (Figure 7a). Although such variation can be accounted for in
665 simulations, this robustness makes it less important that accurate estimates of $N$ be made.

666 ***Conclusions***: Here, we have developed a network coalescent model that can predict
667 coalescence times generated under various introgression scenarios. From this model, we propose
668 two new test statistics, named $D_1$ and $D_2$, which can be used to test hypotheses about the relative
669 timing and direction of introgression. $D_1$ evaluates the null hypothesis that lineage-splitting and
670 introgression occur simultaneously, which is expected to occur during homoploid hybrid
671 speciation or in the creation of admixed populations. This statistic builds on descriptive models
672 by providing a quantitative means for addressing hypotheses related to HHS. $D_2$ is designed to be
673 a test of the null hypothesis that introgression occurred primarily in the C→B direction (from an
674 unpaired species to a paired species), with rejection of the null indicating that introgression
675 primarily occurred in the B→C direction. Our model and statistics can be used with simulated
676 data to provide powerful hypothesis testing in a variety of systems; this is highlighted in our
677 application of the $D_1$ statistic to an empirical dataset from wild yeast.

678

679 **Acknowledgements**

680 We thank Nick Barton and two referees for their very helpful comments, as well as Christian
681 Landry and Jean-Baptiste Leducq for their assistance and for sharing their data with us.

682 Discussions with Rafael Guerrero, Fabio Mendes, and Ben Rosenzweig also helped to improve
683 the work. This research was supported by National Science Foundation grant MCB-1127059 to
684 M.W.H.

## Literature Cited


Aeschbacher, S., J. P. Selby, J. H. Willis and G. Coop, 2017 Population-genomic inference of the strength and timing of selection against gene flow. Proceedings of the National Academy of Sciences of the United States of America 114: 7061-7066.

Barrera-Guzman, A. O., A. Aleixo, M. D. Shawkey and J. T. Weir, 2018 Hybrid speciation leads to novel male secondary sexual ornamentation of an Amazonian bird. Proceedings of the National Academy of Sciences of the United States of America 115: E218-E225.

Bertorelle, G., and L. Excoffier, 1998 Inferring admixture proportions from molecular data. Molecular Biology and Evolution 15: 1298-1311.

Brandvain, Y., A. M. Kenney, L. Flagel, G. Coop and A. L. Sweigart, 2014 Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genetics 10: e1004410.

Degnan, J. H., 2018 Modeling hybridization under the network multispecies coalescent. Systematic Biology 67: 786-799.

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.

Elgvin, T. O., C. N. Trier, O. K. Torresen, I. J. Hagen, S. Lien *et al.*, 2017 The genomic mosaicism of hybrid speciation. Science Advances 3: e1602996.

Elworth R.A.L., O. H. A., Zhu J., Nakhleh L., 2018 Advances in computational methods for phylogenetic networks in the presence of hybridization, arXiv: 1808.08662.

Fay, J. C., and J. A. Benavides, 2005 Evidence for domesticated and wild populations of Saccharomyces cerevisiae. PLoS Genetics 1: 66-71.

Feliner, G. N., I. Alvarez, J. Fuertes-Aguilar, M. Heuertz, I. Marques *et al.*, 2017 Is homoploid hybrid speciation that rare? An empiricist's view. Heredity 118: 513-516.

Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey *et al.*, 2015 Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347: 1258524.

Geraldes, A., P. Basset, K. L. Smith and M. W. Nachman, 2011 Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. Molecular Ecology 20: 4722-4736.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel *et al.*, 2010 A draft sequence of the Neandertal genome. Science 328: 710-722.

Hedrick, P. W., 2013 Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology 22: 4606-4618.

*Heliconius* Genome Consortium, 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487: 94-98.

Hudson, R. R., 1983 Testing the constant-rate neutral allele model with protein sequence data. Evolution 37: 203-217.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Huerta-Sanchez, E., X. Jin, Asan, Z. Bianba, B. M. Peter *et al.*, 2014 Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194-197.

Huson, D. H., and D. Bryant, 2006 Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23: 254-267.

Huson, D. H., T. Klopper, P. J. Lockhart and M. A. Steel, 2005 Reconstruction of reticulate

networks from gene trees. Research in Computational Molecular Biology, Proceedings 3500: 233-249.
Huson, D. H., R. Rupp and C. Scornavacca, 2010 *Phylogenetic networks : concepts, algorithms and applications*. Cambridge University Press, Cambridge, UK ; New York.
Kearns, A. M., M. Restani, I. Szabo, A. Schroder-Nielsen, J. A. Kim *et al.*, 2018 Genomic evidence of speciation reversal in ravens. Nature Communications 9: 906.
Kubatko, L. S., and J. H. Degnan, 2007 Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology 56: 17-24.
Lamichhaney, S., F. Han, M. T. Webster, L. Andersson, B. R. Grant *et al.*, 2018 Rapid hybrid speciation in Darwin's finches. Science 359: 224-228.
Lanier, H. C., and L. L. Knowles, 2012 Is recombination a problem for species-tree analyses? Systematic Biology 61: 691-701.
Leducq, J. B., L. Nielly-Thibault, G. Charron, C. Eberlein, J. P. Verta *et al.*, 2016 Speciation driven by hybridization and chromosomal plasticity in a wild yeast. Nature Microbiology 1: 15003.
Liu, K. J., J. Dai, K. Truong, Y. Song, M. H. Kohn *et al.*, 2014 An HMM-based comparative genomic framework for detecting introgression in eukaryotes. PLoS Computational Biology 10: e1003649.
Lohse, K., M. Chmelik, S. H. Martin and N. H. Barton, 2016 Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics 202: 775-786.
Lynch, M., 2010 Evolution of the mutation rate. Trends in Genetics 26: 345-352.
Mallet, J., N. Besansky and M. W. Hahn, 2016 How reticulated are species? Bioessays 38: 140-149.
Martin, S. H., J. W. Davey and C. D. Jiggins, 2015 Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Molecular Biology and Evolution 32: 244-257.
Martin, S. H., and S. M. Van Belleghem, 2017 Exploring evolutionary relationships across the genome using topology weighting. Genetics 206: 429-438.
Mavarez, J., C. A. Salazar, E. Bermingham, C. Salcedo, C. D. Jiggins *et al.*, 2006 Speciation by hybridization in *Heliconius* butterflies. Nature 441: 868-871.
Mendes, F. K., and M. W. Hahn, 2018 Why concatenation fails near the anomaly zone. Systematic Biology 67: 158-169.
Mendes, F. K., A. P. Livera and M. W. Hahn, 2018 The perils of intralocus recombination for inferences of molecular convergence. Philosophical Transactions of the Royal Society B. in press.
Meng, C., and L. S. Kubatko, 2009 Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theoretical Population Biology 75: 35-45.
Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158: 885-896.
Pamilo, P., and M. Nei, 1988 Relationships between gene trees and species trees. Molecular Biology and Evolution 5: 568-583.
Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland *et al.*, 2012 Ancient admixture in human history. Genetics 192: 1065-1093.
Pease, J., and B. Rosenzweig, 2018 Encoding data using biological principles: the Multisample Variant Format for phylogenomics and population genomics. IEEE/ACM Transactions on Computational Biology and Bioinformatics 15:1231-1238.

Pease, J. B., and M. W. Hahn, 2015 Detection and polarization of introgression in a five-taxon phylogeny. Systematic Biology 64: 651-662.

Rannala, B., and Z. Yang, 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price and L. Singh, 2009 Reconstructing Indian population history. Nature 461: 489-494.

Rieseberg, L. H., M. A. Archer and R. K. Wayne, 1999 Transgressive segregation, adaptation and speciation. Heredity 83: 363-372.

Rouard, M., G. Droc, G. Martin, J. Sardos, Y. Hueber *et al.*, 2018 Three new genome assemblies support a rapid radiation in *Musa acuminata* (wild banana). Genome Biology Evolution. in press.

Schierup, M. H., and J. Hein, 2000 Consequences of recombination on traditional phylogenetic analysis. Genetics 156: 879-891.

Schumer, M., G. G. Rosenthal and P. Andolfatto, 2014 How common is homoploid hybrid speciation? Evolution 68: 1553-1560.

Schumer, M., G. G. Rosenthal and P. Andolfatto, 2018a What do we mean when we talk about hybrid speciation? Heredity 120: 379-382.

Schumer, M., C. Xu, D. L. Powell, A. Durvasula, L. Skov *et al.*, 2018b Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science 360: 656-660.

Solís-Lemus, C., P. Bastide and C. Ané, 2017 PhyloNetworks: A package for phylogenetic networks. Molecular Biology and Evolution 34: 3292-3298.

Sousa, V. C., A. Grelaud and J. Hey, 2011 On the nonidentifiability of migration time estimates in isolation with migration models. Molecular Ecology 20: 3956-3962.

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

Voight, B. F., A. M. Adams, L. A. Frisse, Y. D. Qian, R. R. Hudson *et al.*, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proceedings of the National Academy of Sciences of the United States of America 102: 18508-18513.

Wang, J., 2003 Maximum-likelihood estimation of admixture proportions from genetic data. Genetics 164: 747-765.

Wen, D., Y. Yu, M. W. Hahn and L. Nakhleh, 2016a Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Molecular Ecology 25: 2361-2372.

Wen, D., Y. Yu and L. Nakhleh, 2016b Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genetics 12: e1006006.

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

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 of the United States of America 111: 16448-16453.

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

Zhu, S., and J. H. Degnan, 2017 Displayed trees do not determine distinguishability under the

823 network multispecies coalescent. Systematic Biology 66**:** 283-298.
824 Zhu, Y. O., M. L. Siegal, D. W. Hall and D. A. Petrov, 2014 Precise estimates of mutation rate
825 and spectrum in yeast. Proceedings of the National Academy of Sciences of the United
826 States of America 111: E2310-2318.

827

**Figures**

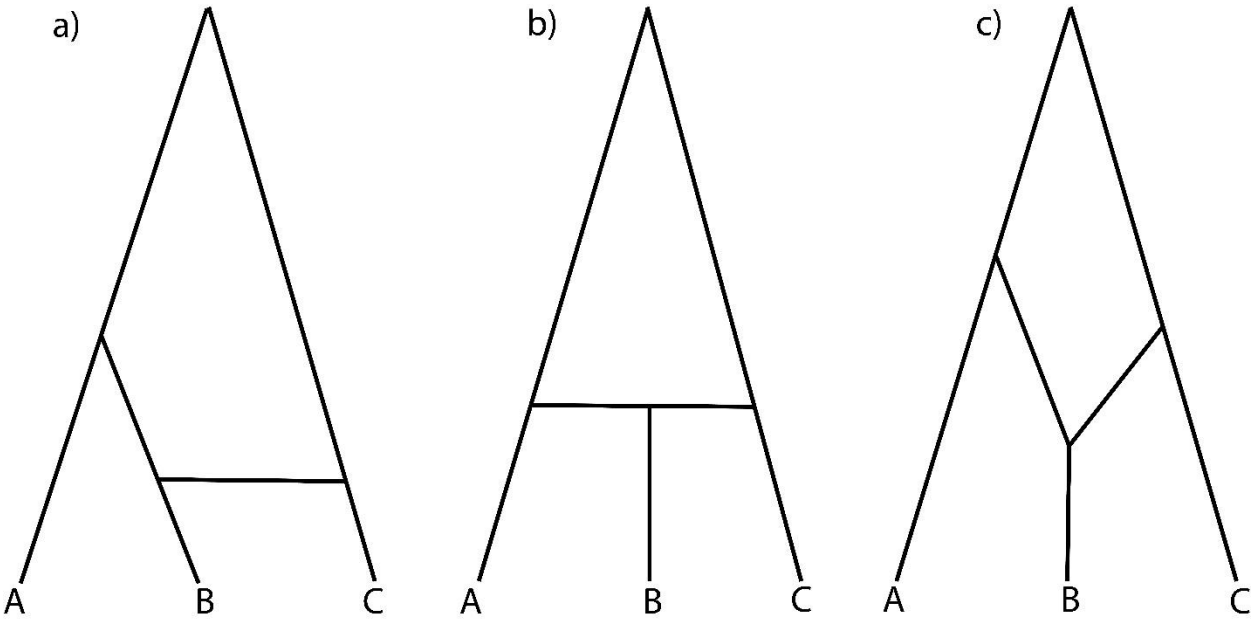


Figure 1: Different phylogenetic network representations of species relationships. a) Speciation between lineages A and B, followed by introgression between lineages B and C (with unspecified direction). b) Homoploid hybrid speciation. Lineage B is created by a hybridization event between A and C. c) This representation is used to denote either speciation between lineages A and B followed by introgression from C into B, or two speciation events followed by the merging of two lineages to form species B.

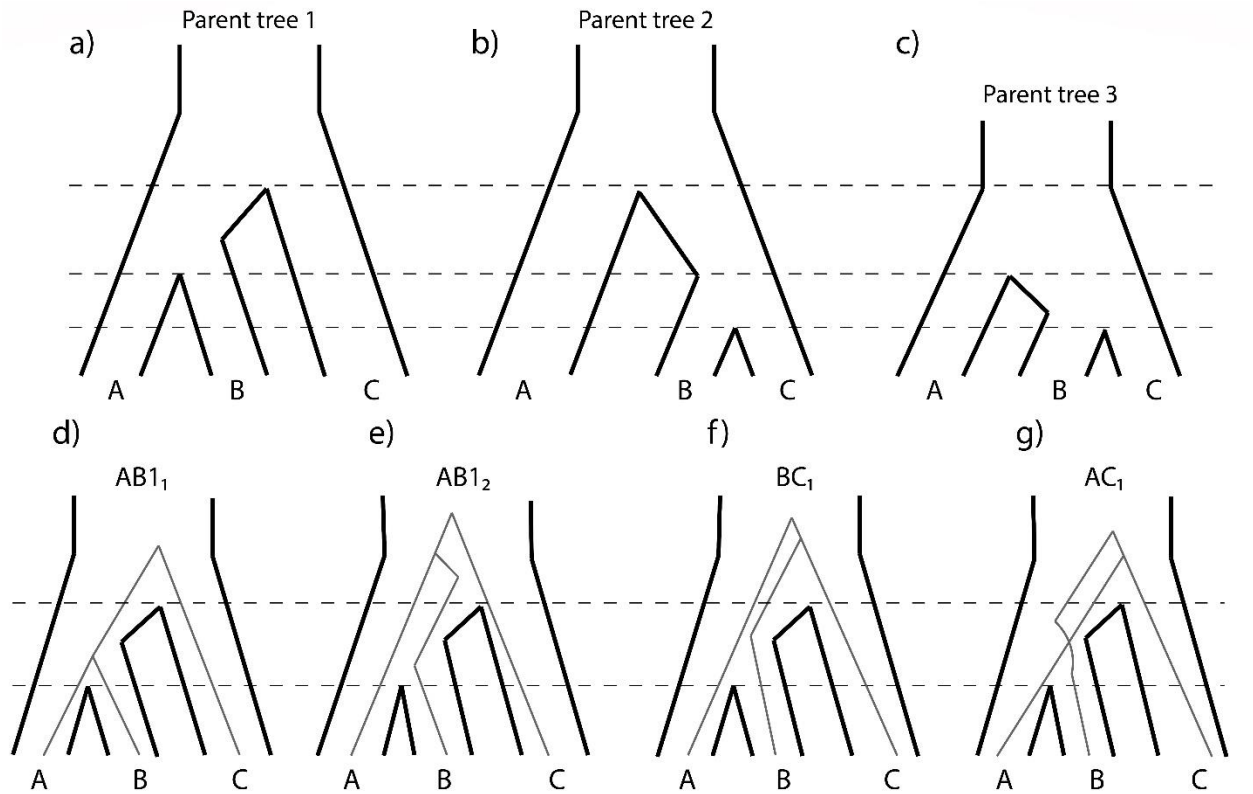


Figure 2: Trees forming the conceptual foundations of our coalescent model, labelled with relevant time parameters. Panels a-c depict the parent trees generated from different introgression scenarios, within which gene trees will sort at particular loci. Panels d-g show the possible ways that gene trees can sort within parent tree 1, demonstrating how gene trees are expected to sort within their respective parent trees according to the coalescent process.

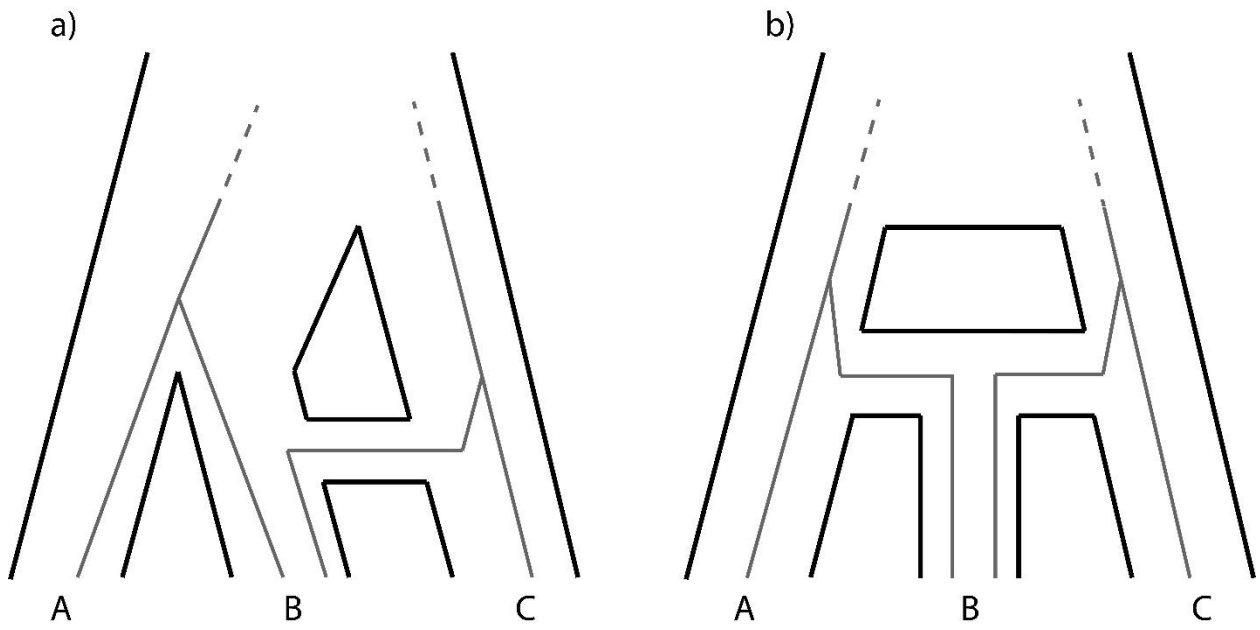


Figure 3: Coalescent times depend on the underlying reticulation history. Grey lines show example genealogical histories (for both the *AB* and *BC* topologies), focusing on the timing of the first coalescent event. a) A reticulation history where speciation and introgression do not occur at the same time. b) A reticulation history where speciation and introgression occur at the same time (as in homoploid hybrid speciation).

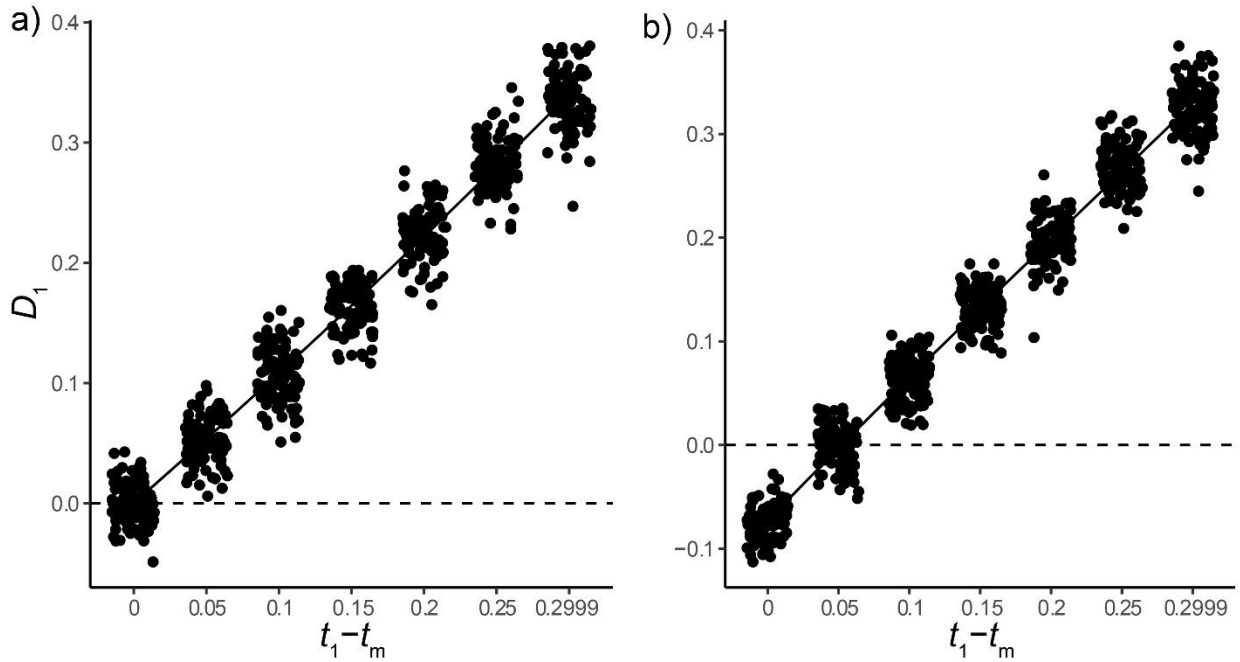


Figure 4: $D_1$ as a function of the difference in the timing of speciation and introgression, for a) introgression in the C→B direction only, and b) introgression in both directions. Dots represent values obtained from simulations, with jitter added for clarity. The solid line shows the expected values from our coalescent model, while the dashed line shows the null hypothesis of $t_1$ - $t_m$ = 0. Time on the x-axis is measured in units of 4*N*.

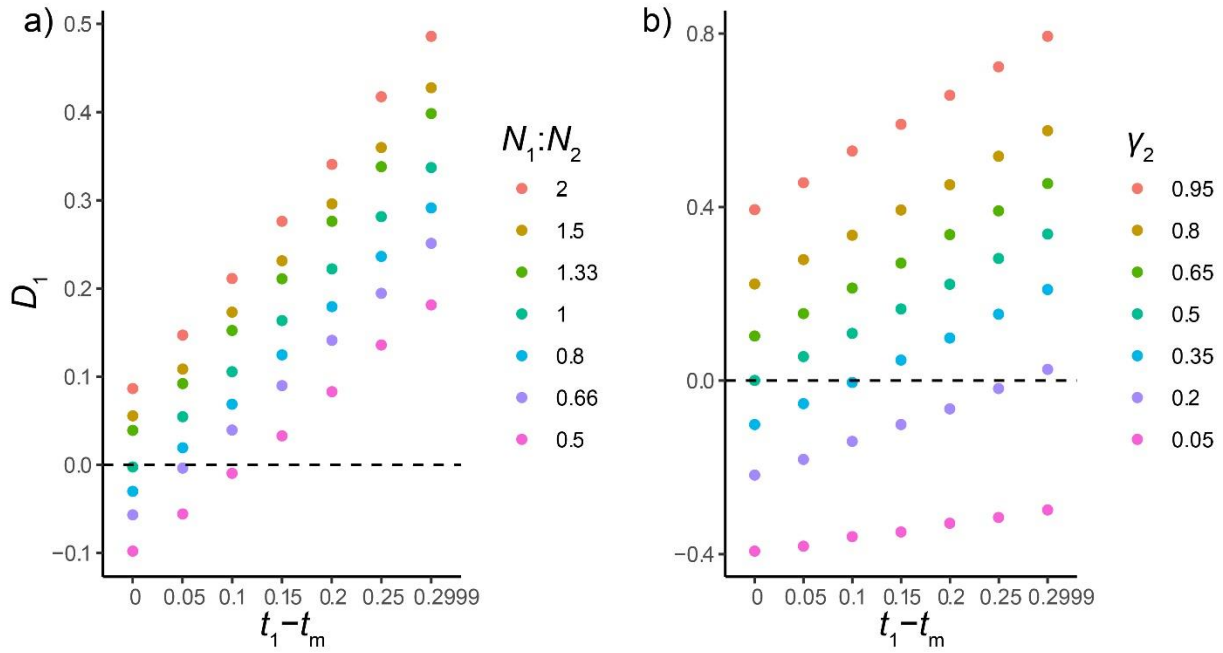


Figure 5: $D_1$ as a function of variation in $N_1$ / $N_2$ and $\gamma_2$. a) Simulations show that variation in $N_1$ / $N_2$ (color legend) changes the mean value of $D_1$. Each point represents the mean value of $D_1$ across 100 simulated datasets. b) Simulations show that variation in $\gamma_2$ (color legend) changes the mean value of $D_1$. Each point represents the mean value of $D_1$ across 100 simulated datasets.

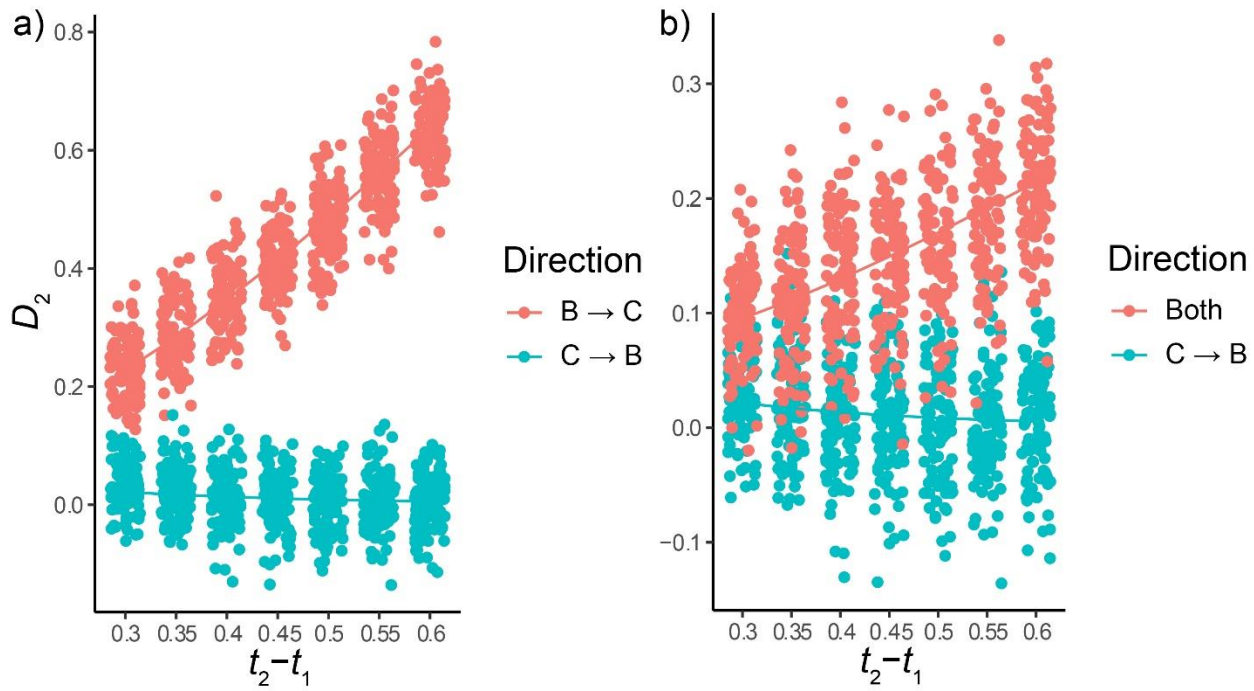


Figure 6: $D_2$ as a function of the time between speciation events, $t_2 - t_1$. The color legend denotes the direction of introgression, with the solid line showing expected values from our model, and the points showing simulated values. Panel a) contrasts the two directions individually, while panel b) contrasts C→B introgression with introgression in both directions. Time on the x-axis is measured in units of $4N$.

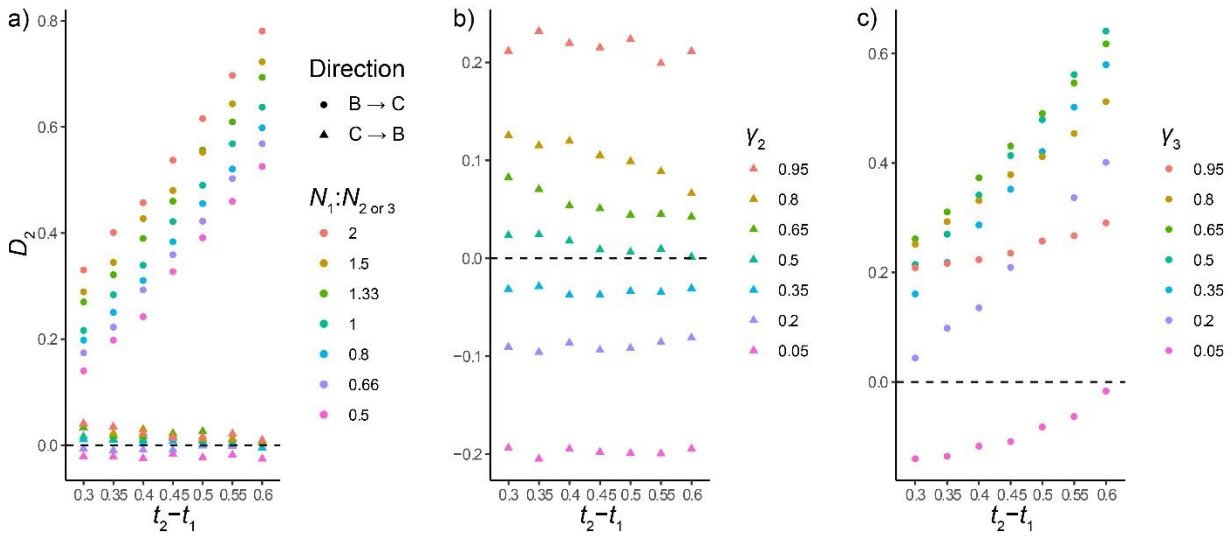


Figure 7: $D_2$ as a function of variation in $N_1$ / $N_2$ and $\gamma_2$. a) Effects of variation in $N$ (color legend) on $D_2$ for both directions of introgression (shape legend). Each point represents the mean of 100 simulated datasets. b) Effects of variation in $\gamma_2$ (color legend) on $D_2$ for C → B introgression. Each point represents the mean of 100 simulated datasets. c) Effects of variation in $\gamma_3$ (color legend) on $D_2$ for B → C introgression. Each point represents the mean of 100 simulated datasets.

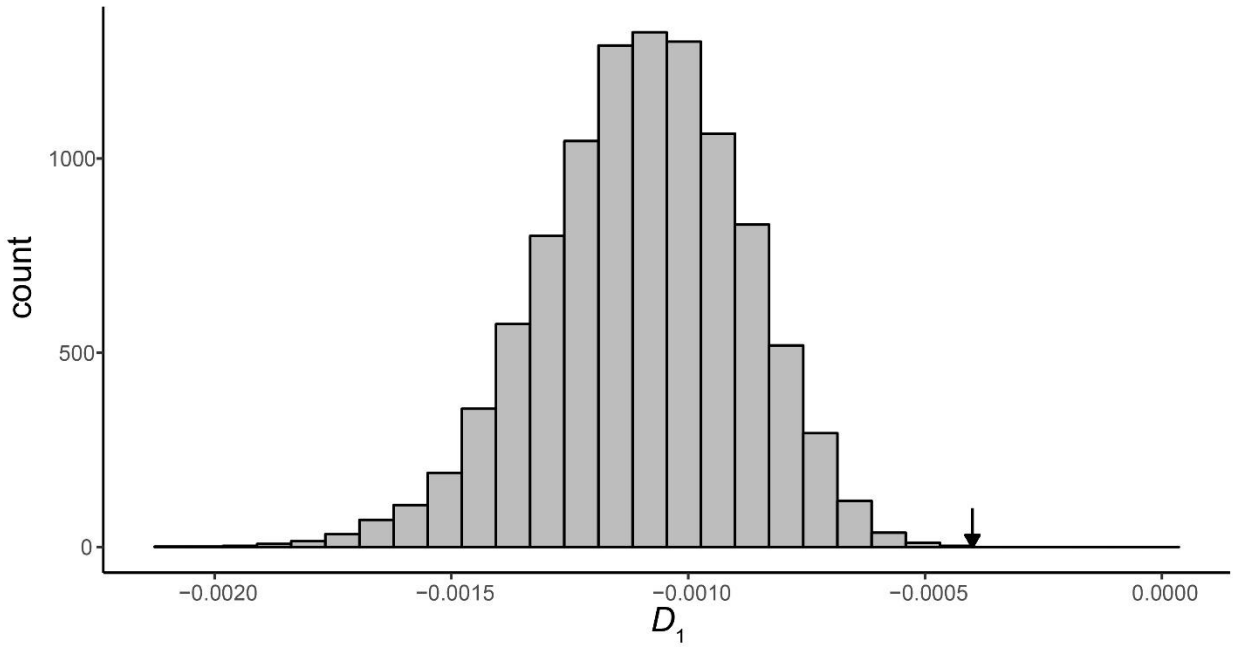


Figure 8: Distribution of 10,000 $D_1$ statistics simulated under a hybrid speciation scenario using *S. paradoxus* demographic parameters. The arrow indicates the mean value of $D_1$ estimated for the *S. paradoxus* system, which was significant by a rank significance test ($D_1$ = -0.0004, $p < 1x10^{-4}$).