Population genetic tests for the direction and relative timing of introgression

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

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

ARTICLE SUMMARY

Many current methods for investigating patterns of introgression do not provide explicit tests of alternative introgression scenarios. Here, we present two new statistics, dubbed D_1 and D_2, which infer the timing of introgression relative to speciation and the direction of introgression respectively. D_1 tests a null hypothesis of simultaneous introgression and speciation, consistent with homoploid hybrid speciation or the formation of an admixed population. D_2 has less stringent data requirements than other statistics designed to polarize introgression. These statistics provide powerful new approaches for testing specific introgression hypotheses using genomic data.

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 coalescent model of introgression using a parent-tree framework, and from it propose two new test statistics: D_1 and D_2. D_1 provides information on the timing of introgression relative to speciation, with the null hypothesis being that they are effectively simultaneous, and the alternative being that introgression occurs significantly afterwards. This statistic will prove useful for addressing hypotheses related to both the origin of admixed populations and the occurrence of homoploid hybrid speciation, in which speciation and introgression can be considered to occur simultaneously. D_2 provides a test for the direction of introgression, and has the advantages of being generally applicable and requiring data from only four taxa, similar to the ABBA-BABA test. Simulations show that D_1 and D_2 have the power to reject their null hypotheses in reasonable regions of parameter space. 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 statistics provide new and powerful ways to address questions relating to the timing and direction of introgression.

Keywords: homoploid hybrid speciation, admixture, ABBA-BABA test, multispecies network coalescent

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. 2016). 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).

The visual representation of gene flow between species or populations, and the resulting phylogenetic network, 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 are commonly depicted, 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 either the direction of introgression or whether hybridization led 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. 2018). 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 a coalescent model of introgression using a parent-tree framework, and propose two new test statistics from this model, dubbed D_1 and D_2. These statistics are intended to infer the timing of introgression relative to speciation and the direction of introgression, respectively. We perform simulations to establish the power of these statistics when speciation and hybridization have occurred at various times in the recent past. 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 parent-tree 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 one or more parent trees (cf. Meng and Kubatko 2009; Liu et al. 2014); this approach has also been referred to as the "multispecies network coalescent" (Wen et al. 2016). Later in the paper we show how these times can complement and extend analyses based solely on the frequency of different topologies.

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). We assume that all ancestral populations share a single effective size, which we denote as N. 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.

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

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

(Mendes and Hahn 2018). 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

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

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:

$$E[t_{A-B}|AB2_1] = t_2 + \frac{2}{3}N \quad (3)$$

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:

$$E[t_{B-C}|AB2_1] = E[t_{A-C}|AB2_1] = t_2 + \frac{2}{3}N + 2N \quad (4)$$

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:

$$E[t_{B-C}|BC_1] = t_2 + \frac{2}{3}N \quad (5)$$

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:

$$E[t_{A-B}|BC_1] = E[t_{A-C}|BC_1] = t_2 + \frac{2}{3}N + 2N \quad (6)$$

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

$$E[t_{A-C}|AC_1] = t_2 + \frac{2}{3}N \quad (7)$$

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:

$$E[t_{A-B}|AC_1] = E[t_{B-C}|AC_1] = t_2 + \frac{2}{3}N + 2N \quad (8)$$

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 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 introgression event in parent tree 2—which we will denote as t_m—rather than by the speciation time, t_1 (Figure 2b).

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:

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

$$E[t_{A-B}|BC1_2] = E[t_{A-C}|BC1_2] = t_2 + 2N \quad (10)$$

$$E[t_{B-C}|BC2_2] = E[t_{A-B}|AB_2] = E[t_{A-C}|AC_2] = t_2 + \frac{2}{3}N \quad (11)$$

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

$$= t_2 + \frac{2}{3}N + 2N \quad (12)$$

Many other parent trees can be formed, 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).

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:

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

$$E[t_{A-B}|BC1_3] = E[t_{A-C}|BC1_3] = t_1 + 2N \quad (14)$$

$$E[t_{B-C}|BC2_3] = E[t_{A-B}|AB_3] = E[t_{A-C}|AC_3] = t_1 + \frac{2}{3}N \quad (15)$$

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

$$= t_1 + \frac{2}{3}N + 2N \quad (16)$$

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 (i.e. parent tree 1) will coalesce further back in time than loci that are introgressed (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), $t_1 > t_{\mathrm{m}}$, 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.

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 2μ, where μ 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 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 $d_{\text{A-B}}|AB$), and the distance between B and C in all trees with the ((B,C),A) topology ($d_{\text{B-C}}|BC$), 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.

The D_1 statistic has a simple interpretation. A positive value significantly different than 0 suggests that introgression has occurred more recently in time than speciation (Figure 3a). Conversely, values not significantly different than 0 are consistent with speciation and introgression occurring effectively simultaneously (Figure 3b). Negative values are possible, but more difficult to interpret, as they indicate that introgression between species B and C occurred longer ago in the past than did speciation. This should not be possible if the species tree is ((A,B),C), but there are multiple reasons why negative values may be observed. In both the Results and the Discussion we consider the causes of such values, as well as multiple biological and statistical factors that may result in false positives or false negatives in D_1.

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 is expected to generate 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 null expectation of D_2: 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 this case we expect D_2 to take a value significantly greater than 0, because the terms $f_{BC1_2}E[t_{A-C}|BC1_3]$ and $f_{BC2_2}E[t_{A-C}|BC2_3]$ in the second half of the equation depend on t_1, which is smaller than t_2. (Once again, expectations for the frequencies of gene trees are given in the Supplementary Materials.) Unlike the D_1 statistic, it should not be possible to obtain a negative value of D_2. We address this and other considerations regarding D_2 in the Discussion.

Simulations

To demonstrate the behavior of D_1 and D_2 we performed a small exploration of 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 from the species tree (i.e. parent tree 1), but specified 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.

Not all parameters are relevant to the behavior of each of our statistics. For D_1, the most important value is $t_1 - t_m$, as this directly measures the difference in the timing of speciation and introgression. For D_2, the most important value is $t_2 - t_1$; this value reflects the degree of truncation of coalescent times in parent tree 3 when introgression has occurred from B into C. We therefore focused on exploring values in these differences in speciation and introgression times in our simulations. For each of the two approaches to simulating introgression described above, we performed 100 replicate simulations for each of seven different values of $t_1 - t_m$, and seven different values of $t_2 - t_1$, with 2000 independent genealogies simulated for each replicate. We then calculated D_1 and D_2 in the relevant scenarios.

Data from *Saccharomyces paradoxus*

To demonstrate the use of D_1 as a 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 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 composition 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. D_1 was then calculated as the difference in the mean of these two groups. Significance of an individual estimate of D_1 was evaluated using a two-sided approximate general independence test, as implemented in the R package *coin*. This test

evaluates whether there is a significant difference between two groups using label permutation. Filtering and distance calculations from all genomic windows were carried out using the software package MVFtools (Pease and Rosenzweig 2015).

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.

RESULTS

Power of D_1 to distinguish alternative histories

To determine the power of our new statistic, D_1, we asked whether it could distinguish between a history of speciation followed by introgression and a history of homoploid hybrid speciation (HHS). Under HHS, speciation and introgression happen simultaneously (i.e. t_1=t_m), and the expected value of D_1 is 0 (equations 17 and 18). As the time between speciation and introgression increases, the value of D_1 should increase linearly.

As expected from our model, simulated values of D_1 are centered around a mean of 0 when t_1–t_m=0, and increase linearly as the introgression event becomes more recent relative to speciation (Figure 4a). There is substantial overlap between the distribution of D_1-values at time 0 (when t_1=t_m) and at the first simulated time-step after 0 (t_1–t_m=0.05), suggesting that the D_1 statistic is unlikely to reject the null hypothesis of HHS in this scenario. Therefore, D_1 may be prone to false negatives (i.e. not rejecting the null hypothesis that t_1–t_m = 0) when introgression occurs very shortly after speciation. However, by the time t_1–t_m=0.1 essentially all values of D_1 are outside the range expected under the null hypothesis (Figure 4a), and there is consequently high power to reject the null. As this minimum time is equivalent to approximately 0.4N generations between speciation and introgression, we expect that the statistic should have high power when applied to examples from nature.

If introgression has not occurred at all, we expect D_1 to take on a negative value; this is because the signal of reduced divergence between B and C due to introgression does not exist. For similar reasons, we might also expect D_1 to be negative if the fraction of introgressed loci is low because the majority of ((B,C),A) gene trees in the genome will be discordant trees produced by parent tree 1, with a much smaller fraction harboring the signal of reduced B-C divergence seen in parent tree 2. Because there may be other causes of negative values, however, D_1 should not be used as a test for the presence of any introgression (see next paragraph).

One factor that can reduce the power of D_1 is the presence of introgressed topologies from additional parent trees. Our model predicts that D_1 should have reduced power when introgression occurs in both directions (i.e. C→B and B→C), due to the presence of gene trees generated by parent tree 3. The most common gene trees generated by parent tree 3 have the same topology as those generated by parent tree 2, but the shorter coalescence time between lineages A and B in this parent tree make tests based on D_1 conservative. To investigate the magnitude of this effect, we simulated across the same values of t_1–t_m, now specifying equal contributions of parent trees 2 and 3. Figure 4b shows that introgression shortly after speciation

in both directions results in negative values of D_1, and that D_1 can be centered around 0 when $t_1 - t_m$ is not 0. These results suggest that: 1) D_1 may be negative if introgression has occurred shortly after speciation and in both directions, and 2) when there is introgression in both directions, D_1 may be prone to false negatives even if introgression occurred substantially after speciation. Here, the relative magnitude of the contribution from parent trees 2 and 3 matters: increased contributions from parent tree 3 (i.e. gene flow from B→C) reduce the power of the test. If the main direction of introgression is from C→B—as would be the case if HHS is suspected—the power remains high.

Finally, the results obtained from our two different simulation approaches are virtually identical (see Figure S1 for data from population splitting/rejoining method), confirming that they 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, $t_2 - t_1$. Our model predicts 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 5a, 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 0.6N generations after speciation), our statistic always has excellent power to distinguish between the two directions. 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.

As with the D_1 statistic, D_2 is expected to be negative when no introgression has occurred in either direction. This is because the second term in equation 19 is dominated entirely by ((B,C),A) gene trees discordant with parent tree 1, which on average have larger A-C divergence than ((A,B),C) gene trees concordant with parent tree 1 (which dominate the first term in the equation). We expect the same pattern when introgression occurs at low levels, but other scenarios may also produce negative values (see next paragraph).

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 0.6N generations after speciation. The simulated results confirm the predictions of our model (Figure 5b). 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 5b).

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

Analysis of *S. paradoxus*

To demonstrate the use of the D_1 statistic as a test of homoploid hybrid speciation, we calculated D_1 from three lineages of the wild yeast *Saccharomyces paradoxus* and an outgroup. If the hypothesis that the *SpC** lineage of *S. paradoxus* is a hybrid species is correct (cf. Leducq et al. 2016), then we would expect D_1 to be 0. A D_1 value significantly different from 0 would indicate introgression that occurs significantly after speciation.

As there were multiple strains sequenced within the three relevant lineages (*SpB*, *SpC*, and *SpC**), we sampled 100 different combinations of one individual from each. The results of our analysis clearly show that the distribution of D_1 statistics are not centered around 0 (Figure 6), and in fact are mostly negative. The red line in Figure 6 shows the calculation of D_1 for a single combination of strains, including an evaluation of significance using a label permutation test. The D_1 value of -0.00034 obtained from this particular comparison is significantly different from 0 ($P = 0.0014$), and is close to the mean value taken from the 100 different combinations ($D_1 = -0.00047$). This suggests that the peak of the distribution itself is significantly non-zero. The calculation also demonstrates that label permutation tests are an effective and simple way to evaluate the significance of individual estimates of D_1.

Negative D_1 statistics are more difficult to interpret, but given the results presented here, we can reject with some confidence the hypothesis that the *SpC** population of *S. paradoxus* was formed by HHS. We provide a more detailed biological interpretation of this result, including the overall negative sign of the statistic, in the Discussion.

DISCUSSION

There are now multiple methods that use only one sequence per lineage to detect the presence of gene flow between species (e.g. Huson et al. 2005; Meng and Kubatko 2009; Green et al. 2010; Yu et al. 2012; Yu et al. 2014; Pease and Hahn 2015). These methods all take advantage of the fact that expectations for the frequency of different gene tree topologies can easily be calculated under ILS, with deviations from these expectations often indicating the presence of gene flow. However, gene tree topologies alone cannot distinguish among various

biological scenarios involving introgression (Zhu and Degnan 2017). In particular, the scenarios represented in Figure 1a and 1b cannot be distinguished using only tree topologies, leading to general confusion and a proliferation of claims about "hybrid species." More rarely, the representation in Figure 1c is used, possibly to indicate lineage fusion (e.g. Kearns et al. 2018), but more commonly to simply indicate the direction of introgression after speciation (with the first bifurcation representing speciation between species A and B, and the second representing gene flow from C into B; Huson et al. 2010). The goal of the current study is to explicitly differentiate among possible histories using two new test statistics. While neither of these statistics should be used as a test for the presence of introgression itself, they complement other widely used statistics that can be used for this purpose and that depend on the same sampling scheme. In what follows we discuss the assumptions and implications of this work.

Assumptions of our model: In order to develop powerful and accurate tests, we modeled the expected frequencies and coalescence times of gene trees produced under various introgression scenarios. Our model comes with a set of assumptions that could potentially be violated in natural systems, so understanding their effects when analyzing real data is important. The effect of these violations on D_1 and D_2 will likely depend on which assumption is being violated. First and foremost, our model assumes that the lineages in question share the same mutation rate (μ) and effective population size (N_e). Lineage-specific μ is unlikely to vary substantially in groups closely related enough to hybridize (Lynch 2010), and therefore it is unlikely that mutation rate variation will pose much of an issue. It is much more likely that N_e could potentially vary between lineages and that this assumption will be violated in real data. The effects of variation in N_e, however, are not easy to predict. The relevant branches that N_e might vary along are the internal branches of each parent tree. For instance, if N_e along the internal branch on parent tree 1 from time t_2 to t_1 (Figure 2a) differs from N_e along the internal branch on parent tree 2 from time t_2 to t_m (Figure 2b), then the coalescence times for the concordant trees generated by each may differ even when t_1=t_m. This could lead to false rejection of the null hypothesis of D_1. Depending on the particular pattern of variation in N_e, violations of this assumption could lead to either false rejection or false acceptance of null hypotheses for both of our statistics. If appropriate data are available, it will therefore be beneficial to estimate N_e or its proxies in either of these ancestral populations, or their descendant populations (i.e. species A and C).

We have also assumed a single, instantaneous introgression event, which leads to the generation of a single alternative history to the species tree. If instead there are multiple introgression events, each additional event would generate an additional parent tree, in the extreme producing infinitely many parent trees for continuous stretches of introgression. Under such scenarios we expect an increase in the variance in coalescent times, but still expect our statistics to capture the main history of gene flow (e.g. Figure 5b). One exception is discussed further in the next section when considering hypotheses about homoploid hybrid speciation. Lastly, our model assumes that coalescence times and gene tree frequencies follow neutral expectations. In the presence of either hitchhiking or background selection, N_e will be reduced, reducing coalescence times and increasing the concordance of gene trees with their respective parent trees. This latter consequence should actually improve the power of our tests by reducing

the ILS that occurs within each parent tree. The reduction in N_e may have more insidious effects (as discussed in the previous paragraph), but unless there are large differences in the overall role of selection between the two relevant ancestral populations we do not expect there to be a bias in the statistics introduced here.

Implications for homoploid hybrid speciation: It has become increasingly apparent that HHS is a process that can generate new diversity quickly, with several charismatic examples now known (e.g. in Darwin's finches; Lamichhaney et al. 2018). Characterizing the frequency with which HHS occurs in nature is therefore important for our understanding of speciation and the evolution of reproductive isolation, though strict criteria for identifying true cases of HHS have been lacking. Schumer et al. (2014) proposed three pieces of evidence that are required to demonstrate HHS: 1) evidence of introgression, 2) evidence of reproductive isolation of the hybrid lineage from both parents, 3) evidence of a causal link between introgression and reproductive isolation. While relatively standard methods exist for evaluating criteria 1 and 2, it is much more difficult to explicitly evaluate criterion 3. Our D_1 statistic provides an explicit test of criterion 3 by asking whether speciation and introgression are effectively simultaneous. Such a relationship would strongly imply a causal link.

A commonly employed expectation for HHS is that there should be an approximately 50:50 split of two contrasting histories in the genome of the hybrid, as would be expected if each parent species contributed equally. However, this pattern may be misleading for at least two reasons. First, not all hybrid species are the result of isolation caused in the F1 generation of crosses between two species. For example, the hybrid butterfly *Heliconius heurippa* likely arose through two generations of backcrossing, resulting in an 82.5:12.5 pattern of ancestry (Mavarez et al. 2006). Selection or drift may also cause deviations from 50:50 expectations. Second, introgression without hybrid speciation can be extensive, affecting 50% of the genome or more (e.g. in *Anopheles* mosquitoes; Fontaine et al. 2015; Wen et al. 2016). Because D_1 contrasts the behavior of genealogies, rather than simply their frequencies, it is capable of testing whether genomic patterns of ancestry are due to hybrid speciation or extensive introgression without assumptions or knowledge of parental contributions.

However, there are also several scenarios in which the D_1 statistic may be misleading, resulting in either incorrect rejection or acceptance of HHS. False rejection may occur under several scenarios. If hybrid speciation is followed by extinction of one parent lineage, then one sampled taxon will be more distantly related to the hybrid than the other; this will lead to non-zero values of D_1 even though HHS has occurred. Similarly, if introgression occurs after hybrid speciation, the value of D_1 could be dominated by the more recent event, again leading to false rejection of HHS. While both of these scenarios are problems for D_1, they would also be problems for any other methods attempting to distinguish HHS from introgression-after-speciation. Conversely, our simulations show that introgression that occurs shortly after speciation—but that is not causally related to speciation—may lead to D_1 values not significantly different from 0 (Figure 4a). Because HHS is the null model of the D_1 test statistic, such false negatives may result in the incorrect acceptance of the HHS hypothesis. This effect is further magnified when introgression has occurred in both directions shortly after speciation (Figure 4b).

Application of D_1 to an empirical example in yeast: One clade of the yeast species *Saccharomyces paradoxus* (denoted *SpC**) has been proposed to be a homoploid hybrid (Leducq et al. 2016). We therefore expected D_1 to be 0 when applied to this species and its two purported parental species, *SpB* and *SpC*. Instead, D_1 applied to whole genomes from these three taxa is significantly negative, which implies that the hybrid speciation hypothesis is highly unlikely to explain the data. As discussed above, HHS could have occurred if it preceded extensive introgression among these lineages, but any signals of such an event have long been overwritten in the genome.

One question that arises, however, is why the value of D_1 is negative in this system. In our model and simulations we have assumed that we have correctly identified the species tree, and that topologies where species "A" and "B" are sister to one another are concordant with this tree. If we have incorrectly assigned the species tree, then the sign of D_1 can be negative (see equations 17 and 18). In our analysis of the *S. paradoxus* data we assumed that the majority topology [((*SpC**,*SpC*),*SpB*)] was the species tree. However, it may be the case that the minority topology [((*SpC**,*SpB*),*SpC*)] represents the species tree. Supplementary figure 9 of Leducq et al. (2016) supports this contention, as it shows that the minority topology includes a splitting event between *SpB* and *SpC** that is farther back in time than equivalent events in the majority tree (cf. Fontaine et al. 2015). Given this, a simple hypothesis to explain our result is that speciation between *SpB* and *SpC** was followed by extensive genome-wide introgression from *SpC* into *SpC**, such that ~92-97% of the genome now groups these two lineages as sister (Leducq et al. 2016). These results could be explained by more complex hypotheses, invoking additional introgression events, but they would also exclude the simple HHS hypothesis.

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

The D_2 statistic can be calculated from three ingroup taxa and an outgroup, similar to the sampling used for the original D statistic (Green et al. 2010) and other related tests using gene tree topologies (e.g. Huson et al. 2005). Using only the frequencies of topologies (or nucleotide site patterns that reflect these underlying topologies), such tests cannot distinguish the direction of introgression. While the internal branches on the two major tree topologies produced by alternative directions of introgression do differ in length (i.e. from t_2 to t_m in parent tree 2 vs. from t_1 to t_m in parent tree 3; Figure 2), this difference is not detectable using the D statistic (Martin et al. 2015). There are of course population genetic methods for inferring the direction of gene flow, and these can even be used between pairs of sister species (e.g. Lohse et al. 2016). However, these methods both require more sequences from each taxon and are more sensitive to

violations of their assumptions due to linked selection because they are more dependent on the variance in coalescence times.

Conclusions: Here, we have developed two new test statistics, named D_1 and D_2, which infer the relative timing and direction of introgression, respectively. D_1 evaluates the null hypothesis that lineage-splitting and introgression occur simultaneously, which is expected to occur during homoploid hybrid speciation or in the creation of admixed populations. This statistic builds on descriptive models by providing a quantitative means for addressing hypotheses related to HHS. D_2 is designed to be a test of the null hypothesis that introgression occurred primarily in the C→B direction (from an unpaired species to a paired species), with rejection of the null indicating that introgression primarily occurred in the B→C direction. Our simulations follow coalescent expectations from our model, demonstrating that both statistics have power to reject their null hypotheses for a reasonable range of parameter values. Application of D_1 as a test for HHS in *S. paradoxus* strongly rejected the hypothesis of a hybrid origin for this species.

Acknowledgements

We thank Christian Landry and Jean-Baptiste Leducq for their assistance and for sharing their data with us. Discussions with Rafael Guerrero, Fabio Mendes, and Ben Rosenzweig helped to improve the work. This research was supported by National Science Foundation grant MCB-1127059 to M.W.H.

Literature Cited

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. Proc Natl Acad Sci U S A 115**:** E218-E225.

Bertorelle, G., and L. Excoffier, 1998 Inferring admixture proportions from molecular data. Mol Biol Evol 15**:** 1298-1311.

Durand, E. Y., N. Patterson, D. Reich and M. Slatkin, 2011 Testing for ancient admixture between closely related populations. Mol Biol Evol 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.

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.

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. Mol Ecol 22**:** 4606-4618.

The *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. Mol Biol Evol 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.

Kearns, A. M., M. Restani, I. Szabo, A. Schroder-Nielsen, J. A. Kim *et al.*, 2018 Genomic evidence of speciation reversal in ravens. Nat Commun 9**:** 906.

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.

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. Nat Microbiol 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 Comput Biol 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.
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. Syst Biol 67**:** 158-169.
Meng, C., and L. S. Kubatko, 2009 Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol 75**:** 35-45.
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 *et al.*, 2012 Ancient admixture in human history. Genetics 192**:** 1065-1093.
Pease, J., and B. Rosenzweig, 2015 Encoding data using biological principles: the Multisample Variant Format for phylogenomics and population genomics. IEEE/ACM Trans Comput Biol Bioinform. doi: 10.1109/TCBB.2015.2509997
Pease, J. B., and M. W. Hahn, 2015 Detection and polarization of introgression in a five-taxon phylogeny. Syst Biol 64**:** 651-662.
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). BioRxiv. doi: https://doi.org/10.1101/306605
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, 2018 What do we mean when we talk about hybrid speciation? Heredity 120**:** 379-382.
Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105**:** 437-460.
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, 2016 Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol Ecol 25**:** 2361-2372.
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 Genet 8**:** e1002660.
Yu, Y., J. Dong, K. J. Liu and L. Nakhleh, 2014 Maximum likelihood inference of reticulate evolutionary histories. Proc Natl Acad Sci U S A 111**:** 16448-16453.
Zhang, C., H. A. Ogilvie, A. J. Drummond and T. Stadler, 2018 Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol 35**:** 504-517.
Zhu, S., and J. H. Degnan, 2017 Displayed trees do not determine distinguishability under the network multispecies coalescent. Syst Biol 66**:** 283-298.

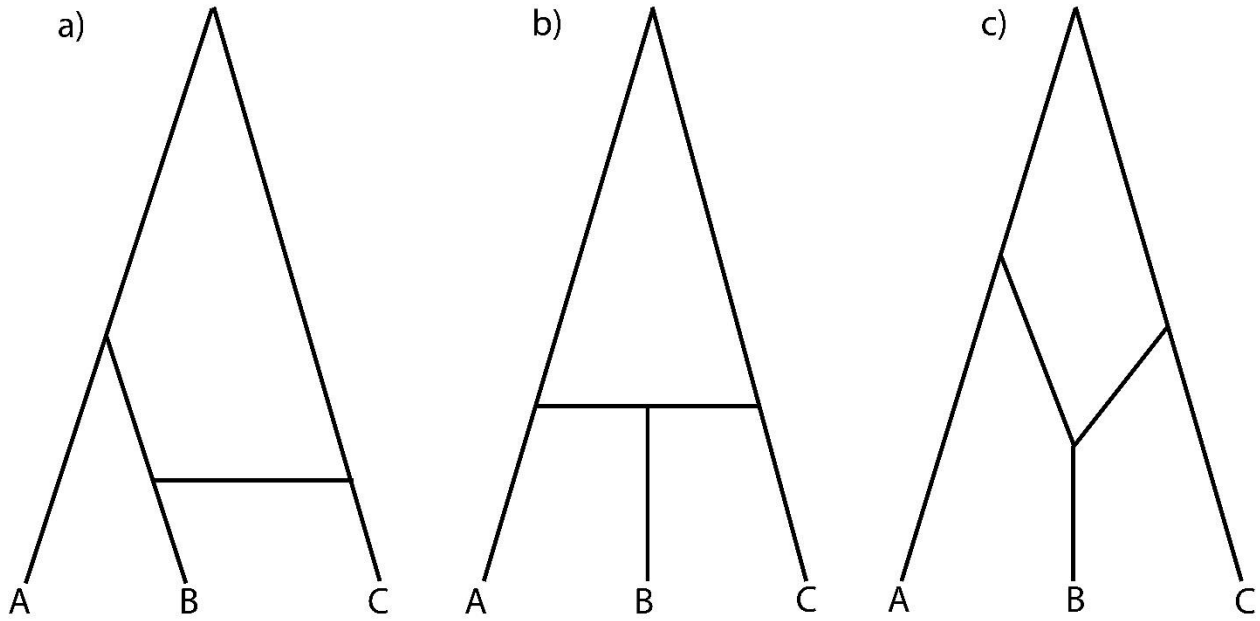

Figure 1: Different phylogenetic network representations of species relationships.

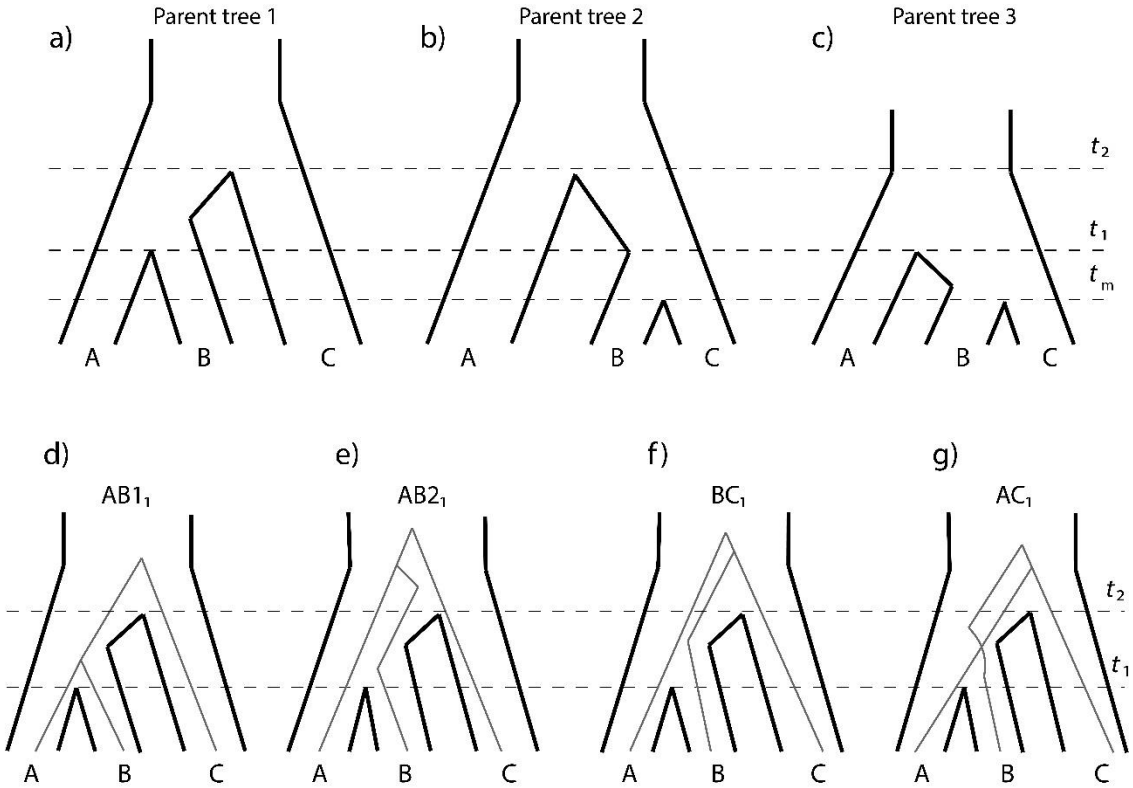

Figure 2: Trees forming the conceptual foundations of our coalescent model, labelled with relevant time parameters. Panels a-c depict the parent trees generated from different introgression scenarios, within which gene trees will sort at particular loci. Panels d-g show the possible ways that gene trees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

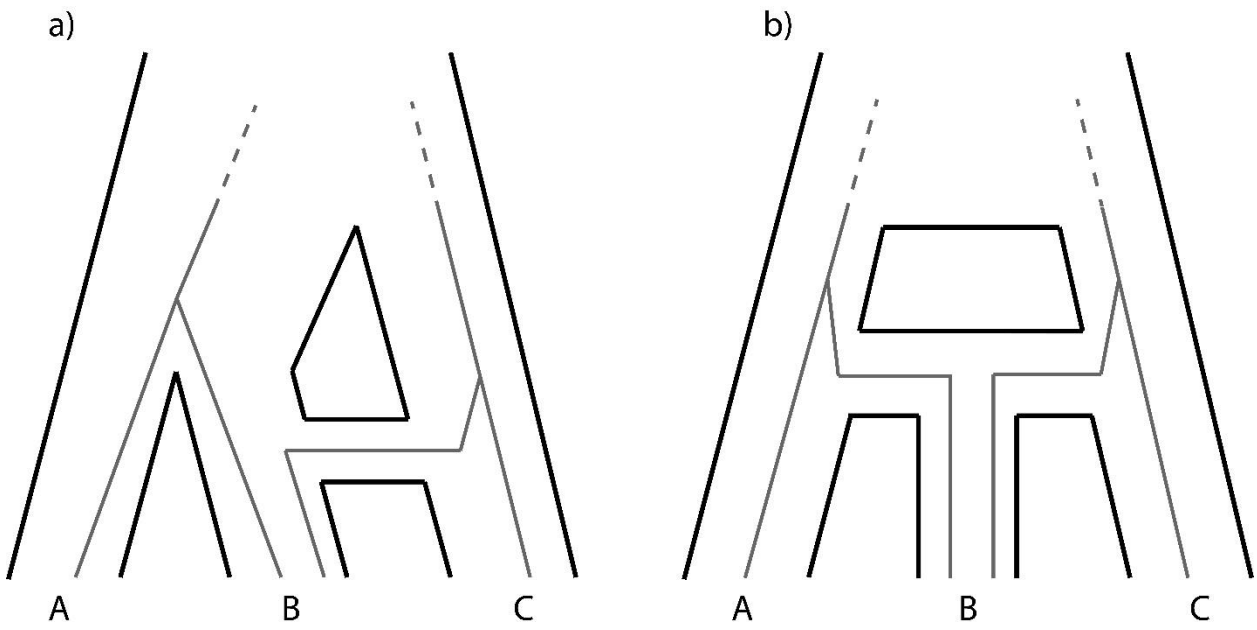

Figure 3: Coalescent times depend on the underlying reticulation history. Grey lines show example genealogical histories (for both the *AB* and *BC* topologies), focusing on the timing of the first coalescent event. a) A reticulation history where speciation and introgression do not occur at the same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

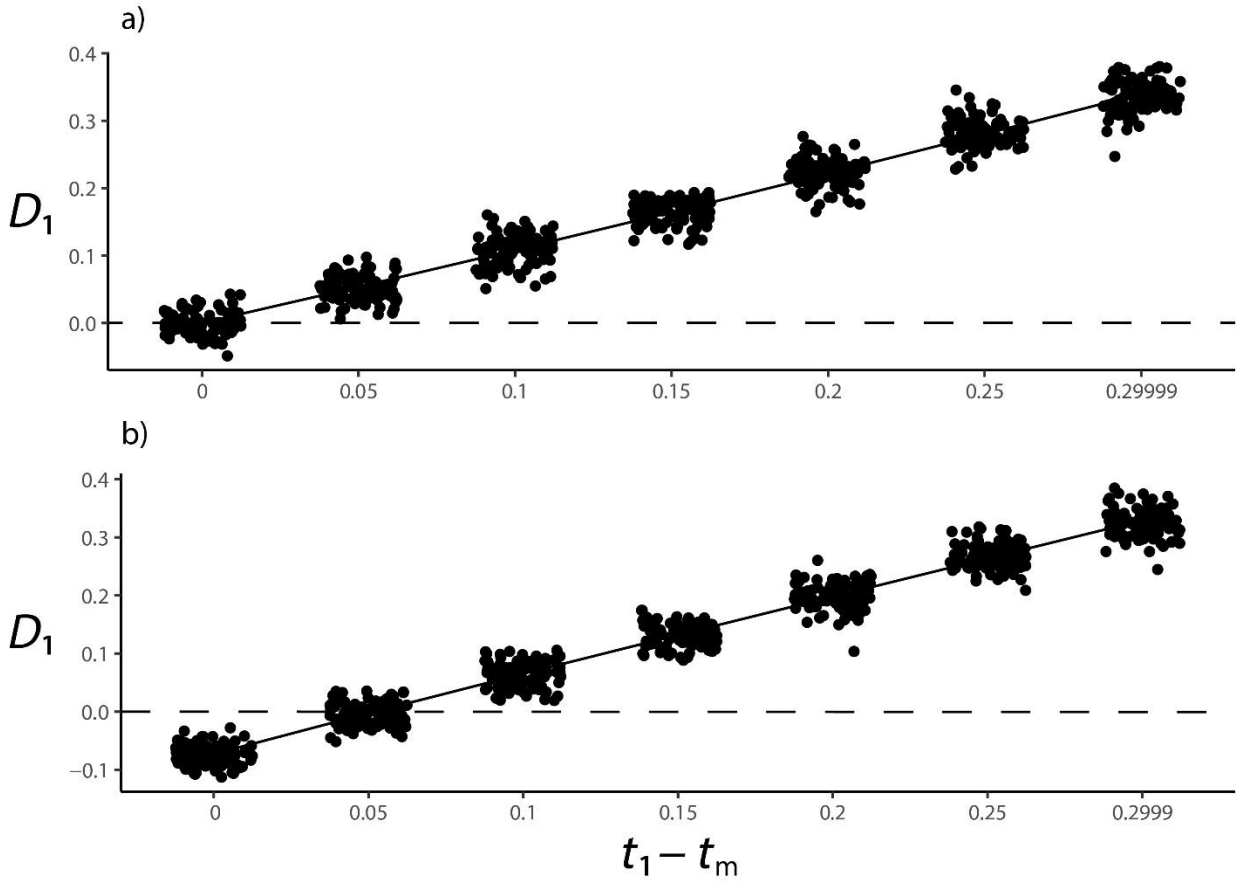

Figure 4: D_1 as a function of the difference in the timing of speciation and introgression, for a) introgression in the C→B direction only, and b) introgression in both directions. Dots represent values obtained from simulations, with jitter added for clarity. The solid line shows the expected values 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*.

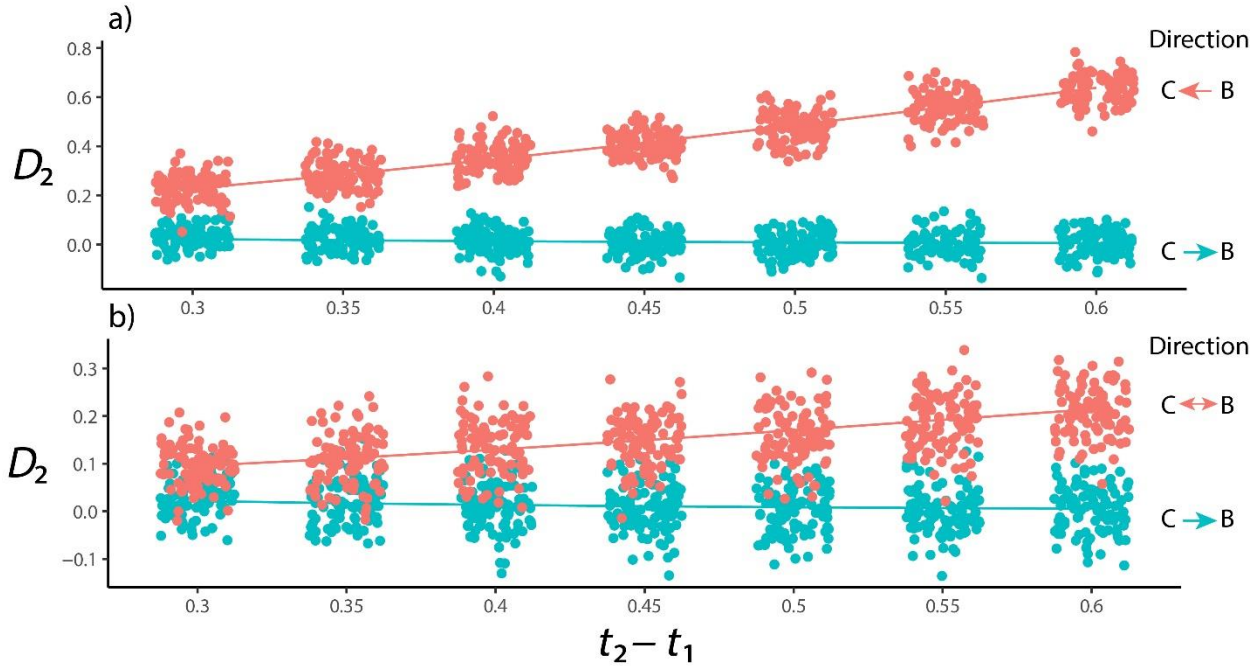

Figure 5: D_2 as a function of the time between speciation events, $t_2 - t_1$. The color legend denotes the direction of introgression, with the solid line showing expected values from our model, and the points showing simulated values. Panel a) contrasts the two directions individually, while panel b) contrasts C→B introgression with introgression in both directions. Time on the x-axis is measured in units of $4N$.

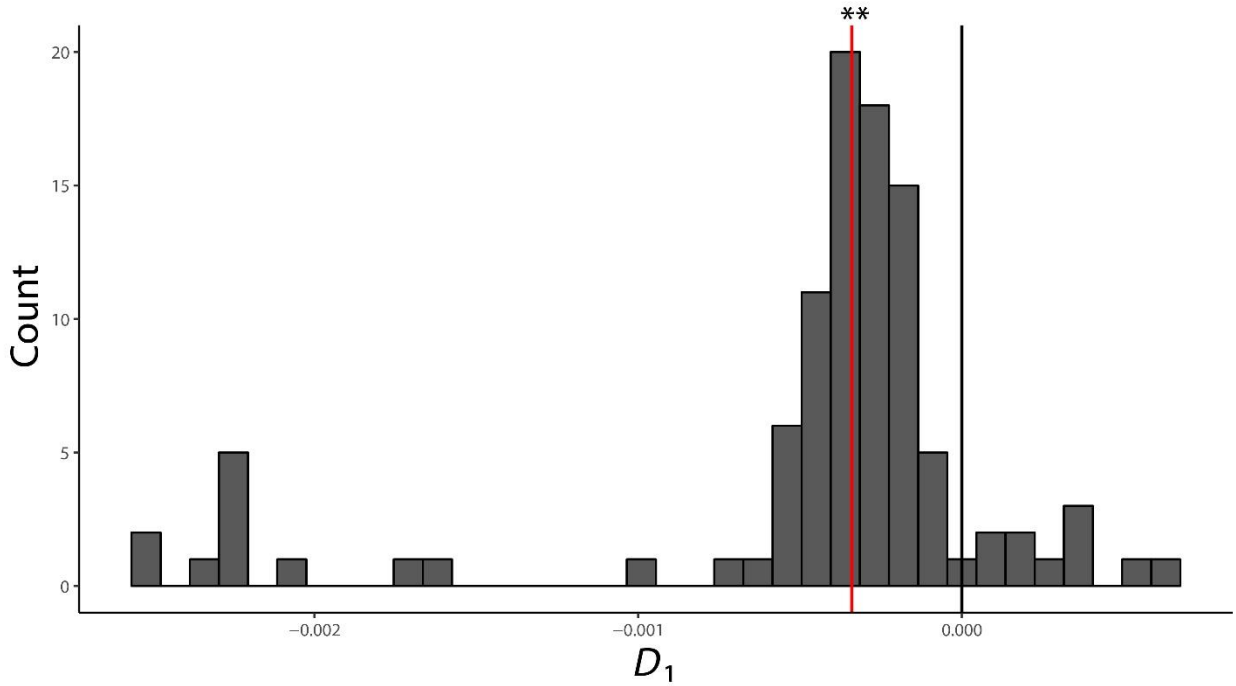

Figure 6: Distribution of D_1 statistics calculated from 100 random samples of strains of *S. paradoxus*. The black line through 0 shows the null hypothesis that $t_1 - t_m = 0$, while the red line represents a single estimate of D_1 with significance calculated via permutation ($P = 0.0014$).

Supplementary Materials and Methods

May 22, 2018

Calculation of expectations for D_1

The D_1 statistic is defined, under the assumption that only parent trees 1 and 2 are relevant, in terms of coalescent expectations as:

$$E[D_1] = (f_{AB1_1}E[t_{A-B}|AB1_1] + f_{AB2_1}E[t_{A-B}|AB2_1] + f_{AB_2}E[t_{A-B}|AB_2]) \\ - (f_{BC1_2}E[t_{B-C}|BC1_2] + f_{BC2_2}E[t_{B-C}|BC2_2] + f_{BC_1}E[t_{B-C}|BC_1]) \quad (1)$$

(This is equation 18 in the main text). If we consider coalescent time in units of $2N$ generations, substituting each time term with its mathematical expectation yields the following:

$$E[D_1] = (f_{AB1_1}(t_1 + (1 - \frac{t_2 - t_1}{e^{t_2 - t_1} - 1})) + f_{AB2_1}(t_2 + \frac{1}{3}) + f_{AB_2}(t_2 + \frac{1}{3}) \\ - (f_{BC1_2}(t_\mathrm{m} + (1 - \frac{t_2 - t_\mathrm{m}}{e^{t_2 - t_\mathrm{m}} - 1})) + f_{BC2_2}(t_2 + \frac{1}{3}) + f_{BC_1}(t_2 + \frac{1}{3})) \quad (2)$$

It is important to note that, due to the conditional nature with which our statistic is defined, the frequency terms in equations S.1 and S.2 cannot simply be taken as gene tree frequency expectations from individual parent trees. Rather, each component is conditional on particular gene tree topologies, which can come from different parent trees. Therefore, we must: 1) weight each frequency by the fraction of loci in the genome that evolved within that parent history; 2) normalize each frequency so that it is with respect to all the gene trees of a particular topology, rather than with respect to all gene trees in the dataset.

First, the expected frequencies of each genealogy within a parent tree are a classical result from the coalescent:

$$P(\text{tree}|\text{lineage sorting}) = 1 - e^{-\tau} \quad (3)$$

$$P(\text{tree}|\text{ILS}) = \frac{1}{3}e^{-\tau} \quad (4)$$

Where τ represents the internal branch lengths — that is, the time between lineage splitting events (i.e. between t_1, t_2, and t_m) — in each parent tree. Equation S.3 represents the expected frequency for gene trees $AB1_1$, $BC1_2$, and $BC1_3$ in our model, while equation S.4 can be used as the frequency for the other gene trees (all conditional on their parent trees).

Let us define the percentage contribution of parent tree 2 to the history of the species we are considering as γ, and the contribution from parent tree 1 as $1-\gamma$. To normalize each gene tree frequency with respect to all the trees sharing its topology, we multiply the expected frequency given above by the parental contribution of that gene tree, divided by the sum of the expectation for all frequencies sharing its topology. With these two considerations, and substituting the appropriate times from our model for τ, we can define each of the frequency terms used in our statistic:

$$f_{AB1_1} = \frac{(1-\gamma)(1-e^{-(t_2-t_1)})}{(1-\gamma)(1-e^{-(t_2-t_1)}) + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)} + \gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})}} \quad (5)$$

$$f_{AB2_1} = \frac{(1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}}{(1-\gamma)(1-e^{-(t_2-t_1)}) + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)} + \gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})}} \quad (6)$$

$$f_{AB_2} = \frac{\gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})}}{(1-\gamma)(1-e^{-(t_2-t_1)}) + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)} + \gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})}} \quad (7)$$

$$f_{BC1_2} = \frac{\gamma(1-e^{-(t_2-t_\mathrm{m})})}{\gamma(1-e^{-(t_2-t_\mathrm{m})}) + \gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})} + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}} \tag{8}$$

$$f_{BC2_2} = \frac{\gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})}}{\gamma(1-e^{-(t_2-t_\mathrm{m})}) + \gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})} + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}} \tag{9}$$

$$f_{BC_1} = \frac{(1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}}{\gamma(1-e^{-(t_2-t_\mathrm{m})}) + \gamma\frac{1}{3}e^{-(t_2-t_\mathrm{m})} + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}} \tag{10}$$

In the special case where $t_1 - t_\mathrm{m} = 0$, all terms in equation S.5 exactly cancel out, leaving us with $E[D_1] = 0$, as stated in the main text. The expected value of D_1 is a linear function of $t_1 - t_\mathrm{m}$.

We also considered values of D_1 under a scenario in which introgression has occurred in both directions. The expectation of D_1 in this circumstance must also now include the contributions from parent tree 3 in addition to 2. With this additional consideration of the gene tree frequencies from parent tree 3, the expectation for D_1 is as follows:

$$\begin{aligned} E[D_1] = (f_{AB1_1}(t_1 + (1 - \frac{t_2 - t_1}{e^{t_2-t_1} - 1})) + f_{AB2_1}(t_2 + \frac{1}{3}) + f_{AB_2}(t_2 + \frac{1}{3}) + f_{AB_3}(t_1 + \frac{1}{3}) \\ - (f_{BC1_2}(t_\mathrm{m} + (1 - \frac{t_2 - t_\mathrm{m}}{e^{t_2-t_\mathrm{m}} - 1})) + f_{BC2_2}(t_2 + \frac{1}{3}) + f_{BC_1}(t_2 + \frac{1}{3}) \\ + f_{BC1_3}(t_\mathrm{m} + (1 - \frac{t_1 - t_\mathrm{m}}{e^{t_1-t_\mathrm{m}}})) + f_{BC2_3}(t_1 + \frac{1}{3})) \end{aligned} \tag{11}$$

Our normalized gene tree frequencies must now also account for the presence of these additional gene trees. We define γ_2 as the contribution of parent tree 2, γ_3 as the contribution of parent tree 3, and $1 - \gamma_2 - \gamma_3$ as the contribution of parent tree 1. Using the same normalization approach as above, we obtain:

$$f_{AB1_1} = \frac{(1-\gamma_2-\gamma_3)(1-e^{-(t_2-t_1)})}{(1-\gamma_2-\gamma_3)(1-e^{-(t_2-t_1)}) + (1-\gamma_2-\gamma_3)\frac{1}{3}e^{-(t_2-t_1)} + \gamma_2\frac{1}{3}e^{-(t_2-t_\mathrm{m})} + \gamma_3\frac{1}{3}e^{-(t_1-t_\mathrm{m})}} \tag{12}$$

$$f_{AB2_1} = \frac{(1-\gamma_2-\gamma_3)\frac{1}{3}e^{-(t_2-t_1)}}{(1-\gamma_2-\gamma_3)(1-e^{-(t_2-t_1)}) + (1-\gamma_2-\gamma_3)\frac{1}{3}e^{-(t_2-t_1)} + \gamma_2\frac{1}{3}e^{-(t_2-t_\mathrm{m})} + \gamma_3\frac{1}{3}e^{-(t_1-t_\mathrm{m})}} \tag{13}$$

$$f_{AB_2} = \frac{\gamma_2\frac{1}{3}e^{-(t_2-t_\mathrm{m})}}{(1-\gamma_2-\gamma_3)(1-e^{-(t_2-t_1)}) + (1-\gamma_2-\gamma_3)\frac{1}{3}e^{-(t_2-t_1)} + \gamma_2\frac{1}{3}e^{-(t_2-t_\mathrm{m})} + \gamma_3\frac{1}{3}e^{-(t_1-t_\mathrm{m})}} \tag{14}$$

$$f_{AB_3} = \frac{\gamma_3\frac{1}{3}e^{-(t_1-t_\mathrm{m})}}{(1-\gamma_2-\gamma_3)(1-e^{-(t_2-t_1)}) + (1-\gamma_2-\gamma_3)\frac{1}{3}e^{-(t_2-t_1)} + \gamma_2\frac{1}{3}e^{-(t_2-t_\mathrm{m})} + \gamma_3\frac{1}{3}e^{-(t_1-t_\mathrm{m})}} \tag{15}$$

$$f_{BC1_2} = \frac{\gamma_2(1-e^{-(t_2-t_\mathrm{m})})}{\gamma_2(1-e^{-(t_2-t_\mathrm{m})}) + \gamma_2(\frac{1}{3}e^{-(t_2-t_\mathrm{m})}) + (1-\gamma_2-\gamma_3)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma_3(1-e^{-(t_1-t_\mathrm{m})}) + \gamma_3(\frac{1}{3}e^{-(t_1-t_\mathrm{m})})} \tag{16}$$

$$f_{BC2_2} = \frac{\gamma_2(\frac{1}{3}e^{-(t_2-t_\mathrm{m})})}{\gamma_2(1-e^{-(t_2-t_\mathrm{m})}) + \gamma_2(\frac{1}{3}e^{-(t_2-t_\mathrm{m})}) + (1-\gamma_2-\gamma_3)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma_3(1-e^{-(t_1-t_\mathrm{m})}) + \gamma_3(\frac{1}{3}e^{-(t_1-t_\mathrm{m})})} \tag{17}$$

$$f_{BC_1} = \frac{(1-\gamma_2-\gamma_3)(\frac{1}{3}e^{-(t_2-t_1)})}{\gamma_2(1-e^{-(t_2-t_\mathrm{m})}) + \gamma_2(\frac{1}{3}e^{-(t_2-t_\mathrm{m})}) + (1-\gamma_2-\gamma_3)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma_3(1-e^{-(t_1-t_\mathrm{m})}) + \gamma_3(\frac{1}{3}e^{-(t_1-t_\mathrm{m})})} \tag{18}$$

$$f_{BC1_3} = \frac{\gamma_3(1-e^{-(t_1-t_\mathrm{m})})}{\gamma_2(1-e^{-(t_2-t_\mathrm{m})}) + \gamma_2(\frac{1}{3}e^{-(t_2-t_\mathrm{m})}) + (1-\gamma_2-\gamma_3)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma_3(1-e^{-(t_1-t_\mathrm{m})}) + \gamma_3(\frac{1}{3}e^{-(t_1-t_\mathrm{m})})} \tag{19}$$

$$f_{BC2_3} = \frac{\gamma_3(\frac{1}{3}e^{-(t_1-t_\mathrm{m})})}{\gamma_2(1-e^{-(t_2-t_\mathrm{m})}) + \gamma_2(\frac{1}{3}e^{-(t_2-t_\mathrm{m})}) + (1-\gamma_2-\gamma_3)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma_3(1-e^{-(t_1-t_\mathrm{m})}) + \gamma_3(\frac{1}{3}e^{-(t_1-t_\mathrm{m})})} \tag{20}$$

Calculation of expectations for D_2

Using the same conventions described above for D_1, we can define the expectation of D_2 in each introgression scenario. For the direction $C \rightarrow B$:

$$E[D_2|C \rightarrow B] = (f_{AB1_1}E[t_{A-C}|AB1_1] + f_{AB2_1}E[t_{A-C}|AB2_1] + f_{AB_2}E[t_{A-C}|AB_2]) \\ - (f_{BC1_2}E[t_{A-C}|BC1_2] + f_{BC2_2}E[t_{A-C}|BC2_2] + f_{BC_1}E[t_{A-C}|BC_1]) \quad (21)$$

This corresponds to equation 20 in the main text. Substituting the expected coalescent times:

$$E[D_2|C \rightarrow B] = (f_{AB1_1}(t_2+1)] + f_{AB2_1}(t_2+\frac{1}{3}+1) + f_{AB_2}(t_2+\frac{1}{3}+1) \\ - (f_{BC1_2}(t_2+1) + f_{BC2_2}(t_2+\frac{1}{3}+1) + f_{BC_1}(t_2+\frac{1}{3}+1)) \quad (22)$$

The normalized gene tree frequencies used in S.21 and S.22 are defined above in equations S.5 - S.10. For the direction $B \rightarrow C$ we have:

$$E[D_2|B \rightarrow C] = (f_{AB1_1}E[t_{A-C}|AB1_1] + f_{AB2_1}E[t_{A-C}|AB2_1] + f_{AB_3}E[t_{A-C}|AB_3]) \\ - (f_{BC1_3}E[t_{A-C}|BC1_3] + f_{BC2_3}E[t_{A-C}|BC2_3] + f_{BC_1}E[t_{A-C}|BC_1]) \quad (23)$$

Corresponding to equation 21 in the main text. Substituting:

$$E[D_2|B \rightarrow C] = (f_{AB1_1}(t_2+1)] + f_{AB2_1}(t_2+\frac{1}{3}+1) + f_{AB_3}(t_1+\frac{1}{3}+1) \\ - (f_{BC1_3}(t_1+1) + f_{BC2_3}(t_1+\frac{1}{3}+1) + f_{BC_1}(t_2+\frac{1}{3}+1)) \quad (24)$$

Here, we need to define new normalized gene tree frequencies for the case when only gene trees from parent trees 1 and 3 are present. They are as follows:

$$f_{AB1_1} = \frac{(1-\gamma)(1-e^{-(t_2-t_1)})}{(1-\gamma)(1-e^{-(t_2-t_1)}) + (1-\gamma)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})})} \quad (25)$$

$$f_{AB2_1} = \frac{(1-\gamma)(\frac{1}{3}e^{-(t_2-t_1)})}{(1-\gamma)(1-e^{-(t_2-t_1)}) + (1-\gamma)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})})} \quad (26)$$

$$f_{AB_3} = \frac{\gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})})}{(1-\gamma)(1-e^{-(t_2-t_1)}) + (1-\gamma)(\frac{1}{3}e^{-(t_2-t_1)}) + \gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})})} \quad (27)$$

$$f_{BC1_3} = \frac{\gamma(1-e^{-(t_1-t_{\mathrm{m}})})}{\gamma(1-e^{-(t_1-t_{\mathrm{m}})}) + \gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})}) + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}} \quad (28)$$

$$f_{BC2_3} = \frac{\gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})})}{\gamma(1-e^{-(t_1-t_{\mathrm{m}})}) + \gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})}) + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}} \quad (29)$$

$$f_{BC_1} = \frac{(1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}}{\gamma(1-e^{-(t_1-t_{\mathrm{m}})}) + \gamma(\frac{1}{3}e^{-(t_1-t_{\mathrm{m}})}) + (1-\gamma)\frac{1}{3}e^{-(t_2-t_1)}} \quad (30)$$

Lastly, we define the expectation of D_2 when introgression occurs in both directions:

$$E[D_2|C \to B, B \to C] = (f_{AB1_1}(t_2+1) + f_{AB2_1}(t_2+\frac{1}{3}+1) + f_{AB_2}(t_2+\frac{1}{3}+1) + f_{AB_3}(t_1+\frac{1}{3}+1))$$
$$- (f_{BC1_2}(t_2+1) + f_{BC2_2}(t_2+\frac{1}{3}+1) + f_{BC_1}(t_2+\frac{1}{3}+1) + f_{BC1_3}(t_1+1) + f_{BC2_3}(t_1+\frac{1}{3}+1)) \quad (31)$$

Equation S.31 uses the gene tree frequencies defined in S.12 - S.20.

ms commands

D_1 simulations, with parent trees simulated separately and then combined

Parent tree 1

ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3

Parent tree 2

ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.3 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.25 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.2 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.1 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.05 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.0001 3 2

Parent tree 3

ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.3 4 2 -ej 0.2999 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.3 4 2 -ej 0.25 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.3 4 2 -ej 0.2 3 2, ms 4 1000 -T -I 4 1 1 1 1- ej 4.0 2 1 -ej 0.3 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.3 4 2 -ej 0.1 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.3 4 2 -ej 0.05 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.3 4 2 -ej 0.0001 3 2

D_2 simulations, with parent trees simulated separately and then combined

Parent tree 1

ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.65 3 2 -ej 0.3 4 3, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.7 3 2 -ej 0.3 4 3, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.75 3 2 -ej 0.3 4 3, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.8 3 2 -ej 0.3 4 3, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.85 3 2 -ej 0.3 4 3, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.9 3 2 -ej 0.3 4 3

Parent tree 2

ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.65 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.7 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.75 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.8 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.85 4 2 -ej 0.15 3 2, ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.9 4 2 -ej 0.15 3 2

Parent tree 3

ms 4 1000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.3 4 2 -ej 0.15 3 2

D_1 simulations, with splitting and rejoining of lineages

C into B direction

ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.2999 3 0.5 -ej 0.2999 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.25 3 0.5 -ej 0.25 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.2 3 0.5 -ej 0.2 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.1 3 0.5 -ej 0.1 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.05 3 0.5 -ej 0.05 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.0001 3 0.5 -ej 0.0001 5 2

D_2 simulations, with splitting and rejoining of lineages

B into C direction

ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.15 2 0.5 -ej 0.15 5 3, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.65 3 2 -ej 0.3 4 3 -es 0.15 2 0.5 -ej 0.15 5 3, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.7 3 2 -ej 0.3 4 3 -es 0.15 2 0.5 -ej 0.15 5 3, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.75 3 2 -ej 0.3 4 3 -es 0.15 2 0.5 -ej 0.15 5 3, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.8 3 2 -ej 0.3 4 3 -es 0.15 2 0.5 -ej 0.15 5 3, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.85 3 2 -ej 0.3 4 3 -es 0.15 2 0.5 -ej 0.15 5 3, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.9 3 2 -ej 0.3 4 3 -es 0.15 2 0.5 -ej 0.15 5 3

C into B direction

ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.6 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.65 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.7 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.75 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.8 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.85 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2, ms 4 2000 -T -I 4 1 1 1 1 -ej 4.0 2 1 -ej 0.9 3 2 -ej 0.3 4 3 -es 0.15 3 0.5 -ej 0.15 5 2

Supplementary figures

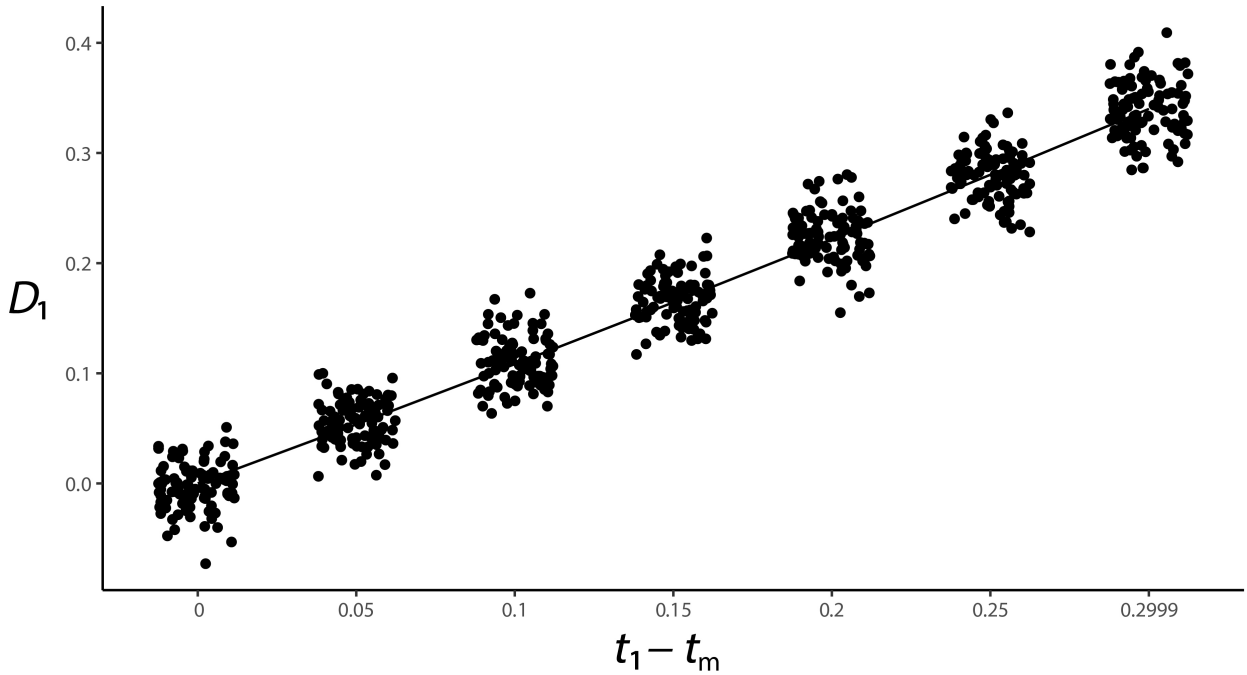

Figure S1: D_1 as a function of the difference in timing of speciation and introgression (units of $4N$), using an instantaneous population split and rejoin event to simulate introgression. Solid line indicates expected values from the mathematical model, while dots indicate simulated values at a particular timestep.

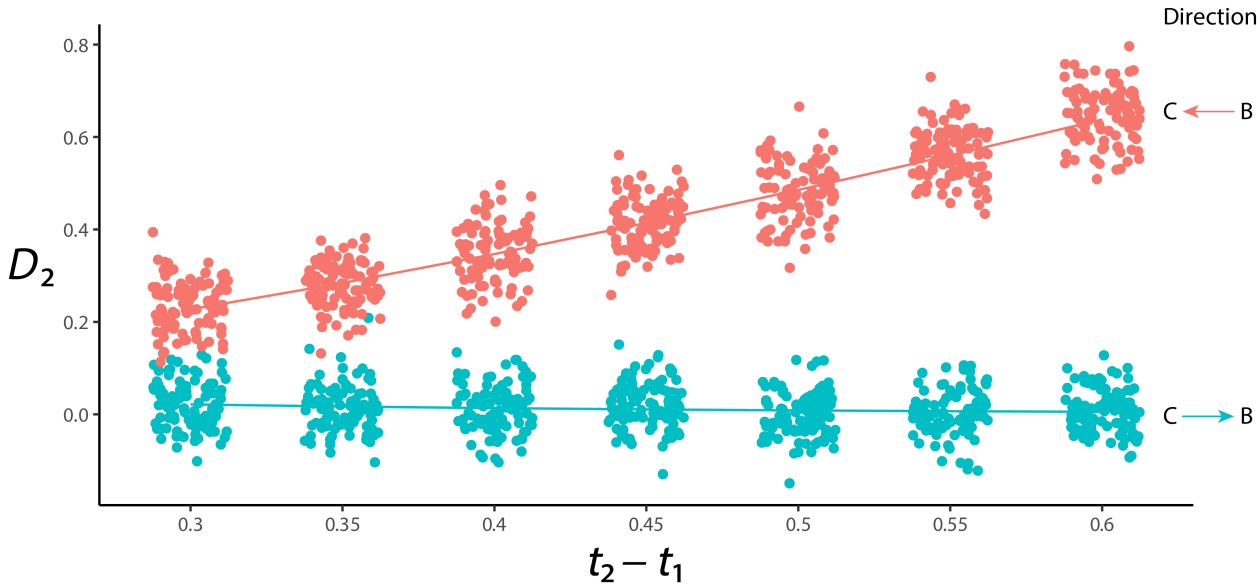

Figure S2: D_2 as a function of the time between speciation events in units of $4N$, using an instantaneous population split and rejoin event to simulate introgression. Color indicates the direction of introgression, solid line indicates mathematical expectation, and dots indicate simulated values.