A three-sample test for introgression

Matthew W. Hahn,⇤,1,2 and Mark S. Hibbins1
1Department of Biology, Indiana University
2Department of Computer Science, Indiana University
⇤Corresponding author: E-mail: mwh@indiana.edu

Abstract

Many methods exist for detecting introgression between non-sister species, but the most commonly used require either a single sequence from four or more taxa or multiple sequences from each of three taxa. Here we present a test for introgression that uses only a single sequence from three taxa. This test, denoted D3, uses similar logic as the standard D-test for introgression, but by using pairwise distances instead of site patterns it is able to detect the same signal of introgression with fewer species. We use simulations to show that D3 has statistical power almost equal to D, demonstrating its use on a dataset of wild bananas (Musa). The new test is easy to apply and easy to interpret, and should find wide use among currently available datasets.

Key words: hybridization, ABBA-BABA, Patterson’s D, gene flow, admixture

Introduction

Genome-scale data have revealed extensive evidence for post-speciation introgression across the tree of life (reviewed in Mallet et al. 2016). Many of these analyses have been carried out in a phylogenetic context, using only a single sample from each population or species. Some methods use gene tree topologies themselves as input (e.g. Huson et al. 2005; Meng and Kubatko 2009; Yu et al. 2011; Edelman et al. 2018), while others use counts of shared derived alleles that reflect the underlying topologies (e.g. Green et al. 2010; Lohse and Frantz 2014). Importantly, the need to distinguish among topologies or between ancestral and derived sites using these methods means that at least four taxa must be sampled, and sometimes more (e.g. Pease and Hahn 2015; Elworth et al. 2018).

Here, we present a test for introgression that only requires a single sample from each of three taxa. With three taxa we cannot infer the frequencies of alternative gene tree topologies.
Instead, our test is based on a related prediction of the ILS model: that there is also an expected symmetry in the branch lengths among topologies. Using these branch lengths, we can develop a test statistic based on pairwise distances to detect the presence of introgression.

**New Approaches**

A test for introgression

Assume that lineages $A$ and $B$ are sister in the species tree, with divergence time $t_1$ (measured in units of $2N$ generations), and that the ancestor of $A$ and $B$ split from lineage $C$ at time $t_2$ (Figure 1a). We refer to gene trees having this topology as $AB$, such that the two discordant topologies are $AC$ and $BC$ (Figures 1b and 1c, respectively).

When ILS is the only cause of gene tree incongruence, topology $AB$ may be generated in two different ways, with different expected frequencies and branch lengths. Looking backwards in time, we refer to the topology in which lineages $A$ and $B$ coalesce before $t_2$ (this is the history shown in Figure 1a). Alternatively, the same topology can occur when these lineages coalesce in the ancestral population of all three lineages; we refer to this topology as $AB_2$.

The expected frequencies of these four topologies are (Hudson 1983):

\[ E[f_{AB}] = E[f_{AC}] = E[f_{BC}] = (1/3)e^{-(t_2-t_1)} \]  
\[ E[f_{AB1}] = 1 - e^{-(t_2-t_1)} \]  
\[ E[f_{AB2}] = E[t_{A-C}|AB1] = t_2+1 \]  
\[ E[t_{B-C}|AB2] = E[t_{A-C}|AB2] = \]  
\[ E[t_{B-C}|AC] = E[t_{A-C}|BC] = t_2+1/3+1 \]
Three-sample test · doi:XX/mst012

These times can be transformed into genetic distances between tip sequences by assuming an infinite sites mutation model and multiplying by two to account for mutations along both lineages since their common ancestor. Summing the weighted length of branches between any two taxa across all possible topologies leads to the following expected distances:

\[ E[d_{B-C}] = 2f_{AB}(t_2+1) + 2f_{AB}(t_2+1/3+1) + 2f_{BC}(t_2+1/3+3) + 2f_{AC}(t_2+1/3+1) \] (6)

\[ E[d_{A-C}] = 2f_{AB}(t_2+1) + 2f_{AB}(t_2+1/3+1) + 2f_{BC}(t_2+1/3+3) + 2f_{AC}(t_2+1/3) \] (7)

(leaving off the shared mutation parameter, \( \mu \), for clarity). Due to the underlying symmetries in topology frequencies and branch lengths under ILS, the expected values of \( d_{B-C} \) and \( d_{A-C} \) are exactly the same. Notably, these expectations hold for distances calculated without rooted gene trees or polarized substitutions (e.g., Figure 1d-f).

Given these results, a natural test of the ILS-only model can be formed using the statistic:

\[ D_3 = \frac{d_{B-C} - d_{A-C}}{d_{B-C} + d_{A-C}} \] (8)

Because the two terms in the numerator have the same expected values under ILS alone, the expectation of \( D_3 \) is 0. The denominator is a normalizing factor that bounds \( D_3 \) between -1 and +1.

\( D_3 \) can be significantly different from zero in the presence of gene flow. While the exchange of alleles between lineages A and B will have no effect on \( D_3 \), unequal amounts of introgression between either B and C (Figure 2a) or A and C (Figure 2b) can lead to deviations from zero. This occurs because gene flow between a pair of non-sister lineages leads to a breakdown in the symmetry of branch lengths predicted under ILS alone. In particular, introgression between B and C leads to both more trees with a BC topology and a shorter pairwise distance between these two lineages (Figure 2a). As a result, \( d_{B-C} \) will be smaller than \( d_{A-C} \), leading to a negative value of \( D_3 \). Conversely, gene flow between A and C leads to positive values of \( D_3 \). Exact expectations for \( D_3 \) in the presence of introgression are presented in the Appendix.

**Results and Discussion**

**Application of \( D_3 \)**

The \( D_3 \) test is straightforward to carry out, requiring only pairwise distances between three species. Ideally, distances should be calculated
from regions for which all three lineages have sequences present in the alignment. This will avoid biases that could possibly occur if regions with different ancestral effective population sizes (for example, in regions with different recombination rates; Pease and Hahn 2013) are sampled unequally for the two relevant distances. Otherwise, variation in either \( N \) or \( \mu \) across sites should not affect the expectation of \( D_3 \).

As an example application of this method, we calculated \( D_3 \) for whole-genome data from three sub-species within \textit{Musa acuminata} (wild bananas; the alignment can be found at https://doi.org/10.6084/m9.figshare.7924727.v1). As was found using the original \( D \)-test on these three taxa and an outgroup (Rouard et al. 2018), \( D_3 \) indicated gene flow between the subspecies \textit{M. a. malaccensis} and \textit{M. a. burmannica} (\( D_3 = 0.06; P < 0.0001 \)). The significance of \( D_3 \) was determined by a block bootstrap of the \textit{Musa} alignment, as is normally done for the \( D \)-test (Green et al. 2010).

**Statistical power of \( D_3 \) and comparison with \( D \)**

We tested the power of \( D_3 \) to detect gene flow with increasing levels of introgression (Figure 3a). As the fraction of the genome introgressed approaches 10\%, \( D_3 \) can detect gene flow in 94\% of simulated datasets (at \( P < 0.05 \)). This demonstrates that \( D_3 \) has good power to detect introgression. In contrast, when there is no gene flow (\( \gamma = 0 \)), the proportion of false positives is the number we would expect at this significance threshold (Figure 3a). We can also see that the expected values of \( D_3 \) under different levels of introgression (calculated according to the equations given in the Appendix) closely match the mean of simulated datasets (Figure 3a).

In order to directly compare these power calculations to the traditional \( D \)-test, we included an outgroup in the same simulated datasets (the outgroup was simply ignored for \( D_3 \) calculations).

As shown in Figure 3b, \( D \) has only slightly more statistical power, despite requiring more data than \( D_3 \). Our results match similar calculations for
Three-sample test · doi:XX/mst012

$D$ carried out previously (e.g. Good et al. 2015; Martin et al. 2015), demonstrating the general power of this class of tests to detect introgression between non-sister lineages.

$D_3$ also has some obvious advantages over similar tests, as it does not require an outgroup (as does the $D$-test) or population samples from three taxa (as does the $f_3$-test; Reich et al. 2009). Even when data from outgroups are available, if there is either ILS or introgression involving these species the $D$-test may not be appropriate. $D_3$ can also detect introgression in both directions (i.e. from $B$ into $C$ and from $C$ into $B$), similar to $D$ but unlike $f_3$, which can only detect it in one direction (Peter 2016).

The $D$-test has been used with ancient DNA samples, as in the use of Neandertal sequences in the paper introducing this statistic (Green et al. 2010). Although the expectations of branch lengths for $D_3$ given here obviously assume that all sequences are sampled from the present (or are sampled contemporaneously from the past), all of the symmetry expectations hold if the ancient sample is the unpaired lineage (i.e. species $C$ in Figure 1). Therefore, there may also be limited cases in which $D_3$ can be applied to ancient samples.

Assumptions of $D_3$

Several points about the test introduced here merit further discussion and explanation. Although the expectations underlying $D_3$ require few assumptions, there are a few things to be cautious about. First, we have assumed that the pairwise distances used as input to $D_3$ accurately reflect coalescence times. This will only strictly be true for sequences evolving under an infinite sites model with the same shared mutation rate across lineages. Such conditions likely hold only for relatively closely related species, limiting the use of $D_3$ to recent divergences.

Second, while values of $D_3$ significantly different from zero can be interpreted as rejecting an ILS-only model (given the above assumptions), such results do not strictly mean that introgression is the cause of rejection. As with the $D$-test, population structure in the ancestor of all three lineages can produce deviations from the ILS-only expectations (Slatkin and Pollack 2008; Durand et al. 2011). In these cases additional analyses may be needed to distinguish among alternative causes of significant $D_3$ values (e.g. Lohse and Frantz 2014).

Finally, we have assumed here that the rooted species tree is known, even though the test does not require an outgroup. Of course it is often the case that the species tree can be inferred from either smaller amounts of sequence data or morphological characters, and so the species tree may be known despite the lack of genome-scale data from an outgroup taxon. However, if the species relationships are not known, a conservative approach would be to test all three combinations of pairwise distances (i.e. $d_{B-C}$, $d_{A-C}$, $d_{B-C} - d_{A-B}$, and $d_{A-C} - d_{A-B}$). If all three
are significantly different from zero, then it is likely that introgression has acted in the system.

**Materials and Methods**

In order to determine the statistical power of the tests discussed here, we simulated multi-locus datasets. For each of four different values of the admixture proportion (\(\gamma\)), we simulated 100 datasets consisting of 1000 non-recombining loci each using the coalescent simulator *ms* (Hudson 2002). The species tree used for all conditions had \(t_1=0.3\) and \(t_2=0.6\), and simulations with introgression had \(t_m=0.05\) (in units of \(4N\) generations). All simulations also included an outgroup taxon that diverged at \(t_o=4\), though data from the outgroup was only used for calculations involving \(D\). Gene trees from *ms* were passed to Seq-Gen (Rambaut and Grassly 1997) to simulate 1-kb alignments under the Jukes-Cantor model with \(\theta=0.01\). All simulation commands are provided in the Appendix.

The resulting datasets of 1000 loci were concatenated together to calculate either \(D\) or \(D_3\). Significance of each simulated dataset was determined by block bootstrapping 1000 times (with block size equal to 10-kb). The resulting values of \(D\) or \(D_3\) were used to generate a \(z\) distribution, and a nominal value of \(P<0.05\) was used as a threshold for significance.

**Acknowledgments**

The authors are grateful for Mathieu Rouard’s assistance with the *Musa* data, and for discussions with Jeff Good and Eric Stone that helped to improve the manuscript. This work was supported by the Precision Health Initiative of Indiana University and National Science Foundation grant DBI-1564611.

**References**


Three-sample test · doi:XX/mst012

MBE


Appendix

Model with introgression

When there is introgression, some loci have a history that takes a different path through the species network. For the simplest case with one introgression event, there is one reticulation and therefore one additional "parent tree" embedded in the species network (see Hibbins and Hahn 2019 for full explication). Here we describe the expectations for introgression from species C into species B (as in Figure 2A in the main text). Other introgression events follow the same logic as this one. The additional parent tree generated by this introgression has lineages C and B sister to one another, and is defined by two split times: \( t_m \) and \( t_2 \). The time \( t_2 \) is the same as in the species tree, but now lineages B and C can coalesce starting at time \( t_m \) (Figure 2A).

This parent tree can also produce all three possible topologies. Because the BC topology is now the one that matches the parent tree, there are two histories with this topology; we denote these BC12 and BC22. The two topologies discordant with the parent tree from an introgression history are denoted AB2 and AC2.

The expected frequencies of these topologies are:

\[
E[f_{BC2}]=E[f_{AB2}]=E[f_{AC2}]=\frac{1}{3}e^{-(t_2-t_m)}
\]

(1)

\[
E[f_{BC12}]=1-e^{-(t_2-t_m)}
\]

(2)

Our goal here is to find the expectation for \( D_3 \) (as given in equation 8 in the main text) in the presence of introgression. We therefore require the expected coalescence times \( t_B-C \) and \( t_A-C \) for each topology from the second parent tree:

\[
E[t_{B-C}|BC12]=t_m+(1-\frac{t_2-t_m}{e^{(t_2-t_m)}-1})
\]

(3)

\[
E[t_{B-C}|BC22]=t_2+1/3
\]

(4)

\[
E[t_{B-C}|AB2]=E[t_{B-C}|AB2]=t_2+1/3+1
\]

(5)

and

\[
E[t_{A-C}|BC12]=t_2+1
\]

(6)

\[
E[t_{A-C}|BC22]=E[t_{A-C}|AB2]=t_2+1/3+1
\]

(7)

\[
E[t_{A-C}|AC2]=t_2+1/3
\]

(8)

The expected distances between lineages across all loci will be comprised of the average distance across trees with both introgressed and non-introgressed histories. Therefore, we must weight the contributions of each history by the admixture proportion, \( \gamma \), which describes the fraction of the genome following the introgression history (with \( 1-\gamma \) following the species history). Combining results on the expected time to coalescence for the species history (given in the main text, and denoted with the subscript "1" here) with the expected times for the introgression history (supplementary equations 1-8), we have:

\[
E[d_{B-C}]= (1-\gamma)*2\mu[f_{AB12}(t_2+1)
\]

\[
+ f_{AB2}(t_2+1/3+1)+f_{BC1}(t_2+1/3)
\]

\[
+ f_{AC1}(t_2+1/3+1)+
\]

\[
\gamma*2\mu[f_{BC12}(t_m+(1-\frac{t_2-t_m}{e^{(t_2-t_m)}-1}))
\]

\[
+ f_{BC2}(t_2+1/3)+f_{AB2}(t_2+1/3+1)
\]

\[
+ f_{AC2}(t_2+1/3+1)]
\]

(9)
and

\[ E[d_{A-C}] = (1 - \gamma) \ast 2\mu [f_{AB1}(t_2 + 1) + f_{AB2}(t_2 + 1/3 + 1) + f_{AC1}(t_2 + 1/3) + \gamma \ast 2\mu [f_{BC12}(t_2 + 1) + f_{BC21}(t_2 + 1/3 + 1) + f_{BC2}(t_2 + 1/3) + f_{AC2}(t_2 + 1/3)] + f_{AC3}(t_2 + 1/3)] \] (10)

These expected values can be used to find the value of \( D_3 \) for any amount of introgression at any time in the past (see, for example, Figure 3 in the main text).

Simulated alignments with introgression

To quantify the power of \( D_3 \) and compare it to the \( D \)-test, we simulated 1-kb alignments of three focal species plus an outgroup, under four different introgression scenarios. For the scenario with no introgression, we used the following call to ms:

```
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
```

For the three scenarios with introgression, we used:

```
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 -es 0.05 3 tbs -ej 0.05 5 2
```

We passed values of 0.99, 0.95, and 0.9 to “tbs”, corresponding to admixture proportions of 0.01, 0.05, and 0.1, respectively. All gene trees simulated in ms were passed to Seq-Gen to generate alignments with the following:

```
seq-gen -m HKY -l 1000 -s 0.01
```

These alignments were then concatenated and used to estimate the statistics.