RH: DETECTING INTROGRESSION FOR FIVE TAXA

Detection and Polarization of Introgression in a Five-taxon Phylogeny

James B. Pease[1,*] and Matthew W. Hahn[1,2]

[1] *Department of Biology, Indiana University, Bloomington, IN, 47405, USA;*

[2] *School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA*

[*] *Correspondence to be sent to: jbpease@indiana.edu.*

ABSTRACT.

In clades of closely related taxa, discordant genealogies due to incomplete lineage sorting (ILS) can complicate the detection of introgression. The *D*-statistic (a.k.a. the ABBA/BABA test) was proposed to infer introgression in the presence of ILS for a four-taxon clade. However, the original *D*-statistic cannot be directly applied to a symmetric five-taxon phylogeny, and the direction of introgression cannot be inferred for any tree topology. Here we explore the issues associated with previous methods for adapting the *D*-statistic to a larger tree topology, and propose new "D_{FOIL}" tests to infer both the taxa involved in and the direction of introgressions for a symmetric five-taxon phylogeny. Using theory and simulations, we find that previous modifications of the *D*-statistic to five-taxon phylogenies incorrectly identify both the pairs of taxa exchanging migrants as well as the direction of introgression. The D_{FOIL} statistics are shown to overcome this deficiency and to correctly determine the direction of introgressions. The D_{FOIL} tests are relatively simple and computationally inexpensive to calculate, and can be easily applied to various phylogenomic datasets. In addition, our general approach to the problem of introgression detection could be adapted to larger tree topologies and other models of sequence evolution.

(Keywords: introgression, incomplete lineage sorting, phylogenomics, hybridization)

In phylogenomic analyses of closely related taxa, conflicting phylogenetic signals among loci are a common occurrence. Discordant genealogies (ones that disagree with each other and possibly the true species topology) represent both a challenge in determining the true species phylogeny and a potential source of additional information about a clade's evolutionary history (Maddison 1997; Degnan and Rosenberg 2009; Edwards 2009). Rapid successive speciation events at any time before the present can lead to discordant genealogies via incomplete lineage sorting (ILS), where two lineages fail to coalesce within a population, making it possible for either lineage to coalesce first with a less related population (Hudson 1983; Tajima 1983; Pamilo and Nei 1988). Discordant genealogies of closely related species can also occur through various forms of hybridization, ranging in scope from the introgression of alleles between existing species to the formation of new hybrid species (Currat et al. 2008; Twyford and Ennos 2012). Even though ILS and introgression/hybridization cause discordant gene trees, the difference in their effects on tree topologies allows them to be disentangled. While several methods for the detection of hybrid speciation in clades with ILS have been previously proposed (Sang and Zhong 2000; Meng and Kubatko 2009; Yu et al. 2012; Yu et al. 2013), here we focus on the detection of introgression between distinct, non-hybrid species in a clade where ILS is present.

Several methods for distinguishing ILS from introgression in a four-taxon phylogeny (three species and an out-group; Fig. 1a) have been proposed previously. The *D*-statistic (a.k.a. the “ABBA/BABA test”) looks for an imbalance in the relative frequency of two minor, discordant gene trees (Green et al. 2010; Durand et al. 2011). Given a consensus phylogeny, ILS should produce the two minor topologies with equal frequency. However, introgression causes an imbalance toward a closer relationship between the two taxa exchanging alleles. Therefore, a statistically significant imbalance toward one discordant topology (indicated by the allele pattern ABBA or BABA) indicates that introgression has occurred.

The *4sp* algorithm (Garrigan et al. 2012) also uses the relative frequencies of allelic state patterns to determine regions of introgression. This algorithm uses a maximum-likelihood approach to estimate global parameters of the species tree, then calculates a local likelihood of introgression for each region of the genome (Garrigan et al. 2012). Joly et al. (2009; see also Joly 2012) use a similar method to look for loci that have sequences with unusually low genetic distances. A significantly lower minimum sequence distance for a pair of taxa indicates introgression at specific loci. Both of these methods, and the *D*-statistic, allow for fine-scale analysis of localized introgression when genomic spatial information is available (i.e., large chromosomal alignments).

When expanding phylogenetic tests for ILS and introgression beyond the simple four-taxon case, there are several challenges. In the presence of ILS, a four-taxon, rooted phylogeny (Fig. 1a) has only three different possible gene tree topologies. For a five-taxon species phylogeny (Fig. 1b) in the presence of ILS there are fifteen possible gene tree topologies both symmetric (Fig. 1c) and asymmetric (Fig. 1d). In addition to more possible gene trees, there are an increased number of possible introgression donor-recipient pairs, and the probability distribution of possible gene trees must also be considered (Degnan and Salter 2005; Degnan and Rosenberg 2006; Twyford and Ennos 2012).

Here, we propose a new set of statistical measures (the D_{FOIL} tests) that not only overcome the inherent challenges of additional phylogenetic information in five-taxon symmetric phylogenies, but also leverage this added data to infer both the taxa involved in and direction of introgression. We also show how these statistics can be used to infer the occurrence of introgression along ancestral branches, and present results that suggest a previous modification of the *D*-statistic (the “Partitioned” *D*-statistic; Eaton and Ree 2013) imprecisely identifies the taxa involved in introgression. The dynamics of all these statistics are shown theoretically and by simulation, followed by a discussion of their applications to data.

Materials and Methods

The Four-taxon D-statistic

In order to approach testing introgression in a five-taxon phylogeny, we first describe the four-taxon case. (Note that in much of the literature on the multispecies coalescent these are referred to as three-taxon and four-taxon trees, where the outgroup is not counted. For consistency with previous work on the *D*-statistic, we do not use this terminology here.) The four-taxon *D*-statistic for introgression was formalized to test ancestral admixture between human and Neanderthal populations (Green et al. 2010; Durand et al. 2011). This statistic applies to a four-taxon asymmetric phylogeny with three in-group taxa and an out-group, denoted $(((P_1,P_2),P_3),O)$ (Fig. 1a). All sites considered in the alignment of sequences from these taxa must be either mono- or biallelic, with the out-group defining the ancestral state (always named A) relative to the derived state (named B). Allelic state patterns for a given position in the alignment are given in the order $P_1P_2P_3O$ (e.g., ABBA; Fig. 1 and 2). Site pattern counts (e.g., n_{ABBA}) are the raw counts of these site types in a given region of the sequence alignment. The model generally assumes 0 or 1 substitutions at each site over the whole phylogeny, with a negligible number of reverse and convergent substitutions. This model also assumes 0 or 1 introgressions in a region. The true is tree is supported by the patterns BBAA and BBBA, and polyphyletic appearance of B in the discordant site patterns ABBA and BABA is attributed to either ILS or introgression (or both).

The *D*-statistic is calculated as (Green et al. 2010; Durand et al. 2011):

$$D = \frac{n_{\mathrm{ABBA}} - n_{\mathrm{BABA}}}{n_{\mathrm{ABBAA}} + n_{\mathrm{BABA}}} \qquad (1)$$

Under ILS and no introgression, the discordant tree patterns ABBA and BABA should occur with equal frequency (Hudson 1983; Tajima 1983; Pamilo and Nei 1988). Introgression between P_3 and either P_1 or P_2 will disproportionately increase the frequency of BABA or ABBA, respectively, since the introgressed pair of taxa should have relatively more shared derived (B) states. Therefore, the four-taxon *D*-statistic is a measure of inequality in the prevalence of site patterns that support the two possible discordant gene tree topologies. Positive values of *D* indicate $P_2 \Leftrightarrow P_3$ and negative values indicate $P_1 \Leftrightarrow P_3$, with values not differing significantly from zero not supporting either introgression (where ⇔ denotes introgression of indeterminate direction and ⇒ denotes a polarized introgression). This approach is not able to detect introgression between the two sister taxa, P_1 and P_2.

Considerations for Extending the D-statistic Beyond Four Taxa

As the number of taxa in a phylogeny increases, introgression testing becomes increasingly complex due to three factors: (1) the number of introgression donor-recipient pairs increases geometrically, (2) the number of possible gene tree topologies also increases geometrically, and (3) the probability distributions of discordant gene tree topologies under ILS become more complex with both the size and shape of the phylogeny (Rosenberg 2002; Degnan and Salter 2005; Degnan and Rosenberg 2006; Rosenberg 2007). Another practical concern is that the four-taxon *D*-statistic can only provide two answers (positive and or negative), corresponding to the introgressions $P_1 \Leftrightarrow P_3$ and $P_2 \Leftrightarrow P_3$. A single *D*-statistic is sufficient for the four-taxon case, but a larger phylogeny will require a system of multiple tests to distinguish among the (rapidly) increasing number of possible introgressions.

There are multiple perspectives from which the *D*-statistic can be understood. In all interpretations, the *D*-statistic has the null hypothesis that D=0 with or without ILS, and with no introgression. From the perspective of site patterns, the *D*-statistic states that the biallelic site patterns `ABBA` and `BABA` should be sampled with equal frequency under the null hypothesis (Eq. 1). The site patterns `ABBA` and `BABA` are a proxy for the two discordant gene trees $(((P_2,P_3),P_1),O)$ and $(((P_1,P_3),P_2),O)$, respectively (Fig. 2a, lines ii and iii), each of which is expected to occur with equal frequency. More generally, we can say that the *D*-statistic compares two sets of gene trees—the “left” and “right” terms of the numerator—whose sampling probabilities are expected to be equal given the probability distribution of gene trees for a particular species phylogeny. In a four-taxon phylogeny, the two sets of gene trees are each represented by only a single discordant gene tree (of the two possible discordant topologies). Since the two discordant gene trees are equally likely to be sampled under ILS, there is an equal probability of inferring that P_3 is more closely related to P_1 or P_2, and D equals 0 under the null hypothesis of no introgression. Importantly, in the *D*-statistic it is not necessary to calculate the exact probability of sampling either discordant gene tree. Because both discordant trees are due to ILS on the same ancestral branch, both are expected to have equal relative probability regardless of what the absolute probabilities are (Fig. 2a). This feature will be important in designing *D* statistics for five-taxon phylogenies.

Another important aspect of the four-taxon *D*-statistic is that it only uses shared derived states to infer introgression. However, introgression is equally capable of transferring the ancestral state (A) or the derived state (B). For example, consider a four-taxon phylogeny where a substitution A→B has occurred on the internal branch P_{12} (Fig. 1a). This would ordinarily lead to the site pattern BBAA. However, if $P_3 \Rightarrow P_1$ introgression occurs, then the A state is transferred resulting in pattern BAAA. This means that when introgression is considered, both ABBA and its "inverse pattern," BAAA, offer evidence of $P_3 \Leftrightarrow P_1$ introgression, and the same is true for BABA/ABAA and $P_3 \Leftrightarrow P_2$. As long as both terms in the numerator of *D* use these inverse site counts, the null hypothesis of D=0 is maintained. However, the inclusion of patterns with a single derived state (i.e. one B) may cause complications in some types of sequence data and in some biological scenarios where substitution rates are unequal among terminal branches of the phylogeny (see Discussion).

From these considerations of the four-taxon *D*-statistic, we can derive four general principles that can be used to test introgression in a five-taxon phylogeny. First, a system of multiple *D*-statistics will be required to distinguish among the larger number of possible donor-recipient combinations of introgression. Second, rather than designing a solution particular to this tree topology as a whole, we can discretize introgression testing into a system of taxon-by-taxon *D*-statistics by examining the relative relationships of a given taxon against two other (appropriately selected) taxa. Third, to maintain the null expectation of D=0, these two relative phylogenetic relationships must have an equal sampling probability across the distribution of all possible gene trees. As with the four-taxon statistic, it will not be necessary to calculate the actual probability values as long as gene trees can be selected in equally probable pairs (as illustrated in Figure 2). Finally, inverse patterns (e.g. BABA/ABAA) both indicate the same potential introgressions.

The Five-Taxon "Partitioned" D-statistics

We also use these principles to examine the "Partitioned" *D*-statistics previously proposed to infer inter-group introgression (Fig. 3a) in a symmetric five-taxon phylogeny (Eaton and Ree 2013). These statistics are:

$$D_1 = \frac{n_{\mathtt{ABBAA}} - n_{\mathtt{BABAA}}}{n_{\mathtt{ABBAA}} + n_{\mathtt{BABAA}}} \tag{2}$$

$$D_2 = \frac{n_{\mathtt{ABABA}} - n_{\mathtt{BAABA}}}{n_{\mathtt{ABABA}} + n_{\mathtt{BAABA}}} \tag{3}$$

$$D_{12} = \frac{n_{\mathtt{ABBBA}} - n_{\mathtt{BABBA}}}{n_{\mathtt{ABBBA}} + n_{\mathtt{BABBA}}} \tag{4}$$

Positive values of D_1 indicate $P_2 \Leftrightarrow P_3$ introgression, while negative values indicate $P_1 \Leftrightarrow P_3$ (Eaton and Ree 2013). Values not differing significantly from 0 indicate no introgression involving P_3. For D_2, positive and negative values indicate $P_2 \Leftrightarrow P_4$ and $P_1 \Leftrightarrow P_4$, respectively, while $D_2 = 0$ indicates no introgressions involving P_4. In addition, the D_{12} statistic was proposed as a means to determine the introgression donor and recipient taxa. D_{12} assumes that if both P_3 and P_4 exhibit the derived state, then the substitution must have occurred on the ancestral P_{34} branch (Fig. 1b). So P_3 or P_4 is assumed to be the introgression donor, and significant positive or negative values of D_{12} indicate the recipient is P_1 or P_2, respectively. Values not deviating significantly from 0 indicate P_1 or P_2 as the introgression donor and P_3 or P_4 as the recipient.

Simulations

We tested the new D_{FOIL} statistics described below on a set of simulated chromosomes divided into 10 kb regions, each of which was allowed to evolve over a common phylogeny (Fig. 3b and 3c). All simulated chromosomes were generated using the Python libraries for the EggLib coalescent simulation engine (v2.1.7; De Mita and Siol 2012). A simulated population size of N_e=10^6 and per-base-per-generation mutation rate of μ=10^{-9} were used. Introgression was simulated by a temporary period of migration from the introgression donor population to the recipient population at a fixed rate (500 individuals per generation). Biallelic site patterns were counted for each region. To simulate the possible effects of convergent mutations in a finite sites mode, five additional substitutions were added randomly to each 10 kb region at sites with one or more existing derived states. The four D_{FOIL} and three Partitioned D-statistics were then calculated from the site pattern counts for each 10 kb window. Custom Python scripts were used for tabulation of site patterns and calculation of Partitioned D-statistics and D_{FOIL}. Plots were generated with the *matplotlib* library (http://www.matplotlib.org). The simulation generation script and site count data sets are available through Dryad (URL PENDING). The DFOIL program used for all calculations and plots is for available for public use (http://www.bitbucket.org/jbpease/dfoil).

RESULTS

The D_{FOIL} Test for a Symmetric Five-taxon Phylogeny

Our model describes a clade of five taxa connected by a symmetric phylogeny, denoted $(((P_1,P_2),(P_3,P_4)),O)$, with the in-group taxa arranged in two sub-pairs (P_1/P_2 and P_3/P_4) and an out-group taxon (O; Fig. 1b). We define that P_3 and P_4 diverged (at time-before-present T_2) before P_1 and P_2 (at T_3), and also that the two sub-pair lineages diverged at T_1. The labeling of the taxa (P_1–P_4) is arbitrary, as long as the sub-pairings are correct and the relationships between the three speciation times-before-present adhere to the relationships $T_1 > T_2 \geq T_3 > 0$.

As in the four-taxon case, we only sample mono- or biallelic sites (with the out-group always represented as A). 0 or 1 introgressions are allowed per region (at time-before-present τ). Reverse and convergent substitutions are expected to occur in negligible amounts, and are expected to affect all topologies equally (Durand et al. 2011). We refer to introgressions occurring between one taxon from each sub-pair as an inter-group introgression (of which there are $n = 8$ possible pairings; Fig. 3a) and those between taxa in the same sub-pair as an intra-group introgression ($n = 4$; Fig. 3a). Additionally, introgression between the ancestral branch P_{12} and P_3 or P_4 is possible when $T_2 > \tau > T_3$, which will be referred to as an ancestral introgression ($n = 4$; Fig. 3a).

We propose a system of four D-statistics, collectively named D_{FOIL}, to distinguish among the 16 possible introgressions in a symmetric five-taxon phylogeny. The name D_{FOIL} borrows from the "FOIL method," a grade school mnemonic for multiplying two binomials ("First, Outer, Inner, Last"). We apply these labels to the four in-group taxa, and name the four D_{FOIL} statistics D_{FO} ("first"=P_1/P_3 *vs.* "outer" =P_1/P_4), D_{IL} ("inner"=P_2/P_3 *vs.* "last"=P_2/P_4), D_{FI} ("first" *vs.* "inner"), and D_{OL} ("outer" *vs.* "last"). These are defined as:

$$D_{FO} = \frac{(n_{\mathrm{BABAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{ABABA}} + n_{\mathrm{AAABA}}) - (n_{\mathrm{BAABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{ABBAA}} + n_{\mathrm{AABAA}})}{(n_{\mathrm{BABAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{ABABA}} + n_{\mathrm{AAABA}}) + (n_{\mathrm{BAABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{ABBAA}} + n_{\mathrm{AABAA}})} \quad (5)$$

$$D_{IL} = \frac{(n_{\mathrm{ABBAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{BAABA}} + n_{\mathrm{AAABA}}) - (n_{\mathrm{ABABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{BABAA}} + n_{\mathrm{AABAA}})}{(n_{\mathrm{ABBAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{BAABA}} + n_{\mathrm{AAABA}}) + (n_{\mathrm{ABABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{BABAA}} + n_{\mathrm{AABAA}})} \quad (6)$$

$$D_{FI} = \frac{(n_{\mathrm{BABAA}} + n_{\mathrm{BABBA}} + n_{\mathrm{ABABA}} + n_{\mathrm{ABAAA}}) - (n_{\mathrm{ABBAA}} + n_{\mathrm{ABBBA}} + n_{\mathrm{BAABA}} + n_{\mathrm{BAAAA}})}{(n_{\mathrm{BABAA}} + n_{\mathrm{BABBA}} + n_{\mathrm{ABABA}} + n_{\mathrm{ABAAA}}) + (n_{\mathrm{ABBAA}} + n_{\mathrm{ABBBA}} + n_{\mathrm{BAABA}} + n_{\mathrm{BAAAA}})} \quad (7)$$

$$D_{OL} = \frac{(n_{\mathtt{BAABA}} + n_{\mathtt{BABBA}} + n_{\mathtt{ABBAA}} + n_{\mathtt{ABAAA}}) - (n_{\mathtt{ABABA}} + n_{\mathtt{ABBBA}} + n_{\mathtt{BABAA}} + n_{\mathtt{BAAAA}})}{(n_{\mathtt{BAABA}} + n_{\mathtt{BABBA}} + n_{\mathtt{ABBAA}} + n_{\mathtt{ABAAA}}) + (n_{\mathtt{ABABA}} + n_{\mathtt{ABBBA}} + n_{\mathtt{BABAA}} + n_{\mathtt{BAAAA}})} \quad (8)$$

Each of these equations calculates the *D*-statistic for one of the four in-group taxa using the principles described previously. For example, D_{FO} tests P_1 and describes the relative support for P_1 being more closely related to P_3 or P_4 (i.e., the two taxa from the opposite sub-pair; Fig 1b). These two relationships are inferred by sampling two sets of gene trees (indicated by biallelic site patterns) that have an equal total probability of being sampled under the null hypothesis given the distribution of gene trees for a symmetric five-taxon phylogeny ("$+D_{FO}$" and "$–D_{FO}$" in Fig. 2b). Under ILS alone, the relative apparent strength of the two relationships represented on either side of the numerator of D_{FO} should be equal. Therefore, $P_1 \Leftrightarrow P_3$ introgression will lead to more sampling of sites that support a closer relationship between P_1 and P_3, and therefore shift towards $D_{FO} > 0$. Alternatively, $P_1 \Leftrightarrow P_4$ introgression will shift towards $D_{FO} < 0$. We also apply the principle of inverse patterns (e.g. `BABBA` and `ABAAA`), and so both terms of all D_{FOIL} tests include two pairs of inverse patterns. D_{IL}, D_{FI}, and D_{OL} are calculated identically to D_{FO}, with P_2, P_3, and P_4, respectively, as the focal taxon instead of P_1, and P_1/P_2 used for comparison in D_{FI} and D_{OL}. For all four tests, significant positive or negative values support a hypothesis of introgression for the focal taxon with one of the two taxa from the other sub-pair.

Importantly, it should be noted that the constraint of equal probability for each of the two terms in the numerators of the D_{FOIL} statistics means that an analogous set of tests for an asymmetric five-taxon phylogeny cannot be constructed. In the asymmetric phylogeny ((((P_1,P_2),P_3),P_4)O) the relationship between P_1 and P_3 is not expected to be equal to P_1 and P_4 under the null. This occurs because a closer relationship between P_1 and P_3 only requires ILS to have occurred on one branch, while a closer relationship between P_1 and P_4 requires ILS on two ancestral branches. This highlights the special property that the four-taxon asymmetric and five-taxon symmetric phylogenies share with respect to ILS. In both topologies, all discordant coalescences must occur in the root, and therefore the exact probability distributions of discordant topologies are not needed to construct a test of the null hypothesis of no introgression.

Direction of Introgression and Significance

As described thus far, the four D_{FOIL} statistics are simply individual applications of the *D*-statistic to each of the four in-group taxa. However, all four D_{FOIL} statistics considered collectively contain more information than the sum of the individual *D*-tests. In addition to identifying the taxa involved in introgression, the D_{FOIL} statistics can also provide information about the direction of inter-group introgressions, specifically identifying both the donor and recipient taxa. When an inter-group introgression occurs in a symmetric five-taxon phylogeny, the relationship changes not only between the donor and recipient taxa change, but also the recipient taxon and the donor's sister taxon. For example, if $P_3 \Rightarrow P_1$ occurs in an otherwise concordant gene tree, then the resultant topology becomes $(((P_1,P_3),P_4),P_2),O)$ (Fig 2b, line viii). The shared history of P_3 and P_4 means that P_4 also changes its relationship to P_1. Conversely, if $P_1 \Rightarrow P_3$ occurred the resultant topology is $(((P_1,P_3),P_2),P_4),O)$ (Fig 2b, line vii), and now the relationship of P_2 with P_3 and P_4 is altered by association. Even though both relationships change, the introgressing taxon should change more strongly than its sister taxon, and this disparity informs the overall assignment of donor and recipient. In this way, the signs (+, –, or 0) of the four D_{FOIL} tests collectively form a signature that provides information beyond the sum of its individual components. The D_{FOIL} tests individually indicate which taxa are introgressing, but collectively can also identify the donor and recipient taxa (or ancestral lineage) for a given introgression.

Each D_{FOIL} tests is separately assessed to be significantly positive, significantly negative, or not different than 0 by a two-tailed binomial test of the raw site counts for the left and right terms with a 50:50 expectation. Statistically significant positive and negative values are assigned a sign of "+" or "−," while non-significant values are "0." The four tests in the order (D_{FO}, D_{IL}, D_{FI}, D_{OL}) form the D_{FOIL} signature (Table 1). The appropriate P-value cutoffs for significance are determined by simulation, using the tested topology with no introgression. The 0.05 tail of simulated components of D_{FOIL} from regions with no introgression are used as the empirical P-value cutoffs. Individual P-values for each D_{FOIL} component can be calculated; however, it is expected that D_{FO} and D_{IL} will have approximately equal significance thresholds, as will D_{FI} and D_{OL}. The difference in the thresholds between these two pairs of statistics should be approximately proportional to the difference in divergence times T_2 and T_3.

Application and accuracy of the D_{FOIL} method

To demonstrate the effectiveness of the D_{FOIL} method, we simulated the evolution of chromosomes over a common phylogeny (Fig. 3a and 3b). Along the chromosome, we simulated sixteen sets of n=500 loci of length 10 kb for each of the sixteen possible introgressions (eight inter-group, four ancestral branch, four intra-group). Between each introgressed region, n=100 regions of no-introgression control were included for visual contrast (Fig. 3d, shaded). In all sixteen cases and the control, the D_{FOIL} signature (signs of the four D_{FOIL} tests) that was observed in the simulations matched the theoretical expectations (Fig. 3d and Table 1). For example, the introgression $P_1 \Rightarrow P_3$ (Fig. 3d, leftmost region) shows positive values for $D_{FO,}$ D_{IL}, and D_{FI}, while $D_{OL} \approx 0$, in agreement with the expected +/+/+/0 signature. In each of the eight inter-group introgressions (Fig. 3d, leftmost eight regions), a unique D_{FOIL} signature clearly distinguishes each case. Ancestral introgressions involving branch P_{12} (Fig. 3c and middle of 3d) were distinguishable from cases of introgression with P_3 or P_4, but the direction of these introgression cannot be determined by the D_{FOIL} signature alone (see Discussion). As expected, the average value for all four D_{FOIL} tests in intra-group introgressions (Fig. 3d, rightmost four regions) is 0. Therefore, the D_{FOIL} tests can distinguish the taxa involved in inter-group and ancestral introgressions, and additionally the direction of introgression for inter-group introgressions.

The Five-Taxon "Partitioned" D-statistics

We also used the D-statistic framework described above to examine the Partitioned D-statistics (Eq. 2-4; Eaton and Ree 2013). As previously noted, inverse pairs of biallelic patterns (e.g. ABBAA/BAABA) both indicate the same underlying gene tree when introgression is considered. This means that the site pattern counts used in D_1 and D_2 of the Partitioned D-statistic are not necessarily unique to the introgressions they propose to test. Specifically, the left term of D_1 (n_{ABBAA}) and right term of D_2 (n_{BAABA}) are directly related, as are the right term of D_1 (n_{BABAA}) and left term of D_2 (n_{ABABA}). Therefore, an inter-group introgression between any two taxa should change both D_1 and D_2 in opposite directions.

This inverse relationship between D_1 and D_2 means that introgression between one pair of taxa creates a "mirror effect" that makes the other pair of taxa falsely exhibit evidence of introgression. Therefore, while D_1 and D_2 in this form can detect the occurrence of an inter-group introgression, the mirror effect makes it unclear which two taxa are actually exchanging alleles. This effect is a consequence of using biallelic patterns where two (out of four) taxa exclusively share one state, since, by default, the other two taxa share the opposite state.

In all cases of inter-group introgression, the simulation data showed an inverse relationship between D_1 and D_2, confirming the mirror effect (Fig. 4). For example, the introgressions $P_1 \Rightarrow P_3$ and $P_3 \Rightarrow P_1$ (Fig. 4, left side) are expected to be $D_1 < 0$ and $D_2 \approx 0$. However, in both cases D_2 showed positive values that mirrored the negative values of D_1, making it falsely seem as though $P_2 \Leftrightarrow P_4$ was also occurring in both cases. A similar result was observed for the introgressions $P_1 \Rightarrow P_4$ and $P_4 \Rightarrow P_1$ (Fig. 4, right side). In both cases, the expected value of D_1 is 0, but instead D_1 mirrors D_2, with positive values of D_1 that make it appear as though $P_2 \Leftrightarrow P_3$ has also occurred. We can conclude both from theory and simulations that D_1 and D_2 are inversely related, and therefore are non-specific in determining which pair of taxa have introgressed.

The patterns counted by D_{12} could also be arrived at through introgressions other than the ones intended to be tested. For example, ABBBA is assumed by Eaton and Ree (2013) to be the result of a $P_3 \Rightarrow P_2$ or $P_4 \Rightarrow P_2$ transfer of the B state from a pre-introgression pattern of AABBA. However, ABBBA can also be formed from the patterns ABABA and ABBAA by $P_2 \Rightarrow P_3$ or $P_2 \Rightarrow P_4$, respectively. This means that for any given inter-group pair of taxa, introgressions in both directions will both raise or both lower the value of D_{12}. This occurs because Eaton and Ree (2013) assume that the substitution A→B occurred on the ancestral P_{34} branch, which also forces P_3 and P_4 to both be state B. This effectively collapses P_3 and P_4 to a single branch and makes D_{12} a four-taxon D-statistic.

Simulations also confirmed these conclusions, showing that D_{12} values for a given pair of taxa are either positive or negative regardless of direction of introgression. For example, among the four introgressions shown in Figure 4, $P_3 \Rightarrow P_1$ and $P_4 \Rightarrow P_1$ exhibit negative values of D_{12}, respectively, as expected. However, $P_2 \Rightarrow P_3$ and $P_2 \Rightarrow P_4$ are expected to be $D_{12} \approx 0$, but while they appear closer to zero, both cases still show a non-zero value of D_{12} (approximately at the same level as D_1) with the same sign as introgression in the opposite direction. These simulations confirm that D_{12} will have the same sign for an introgression between given pairs of taxa regardless of direction.

Discussion

Robustness of the D_{FOIL} Tests to Special Cases

There are several special cases in which introgression occurs at particular points in the phylogeny and for which detection of these events might be more problematic. However, the triangulation approach of the D_{FOIL} tests allows them to still function even in these special cases. If speciation of the two subgroups occurred at approximately the same time (i.e., $T_2 \approx T_3$), such that the difference between the two is not clearly discernible, the signs of all four statistics would remain unaffected. This would, however, negate the possibility of ancestral introgression, since both speciation events occur at approximately the same time.

If introgression occurs very close to the second speciation event (i.e., $\tau \approx T_3$) this would cause D_{FI} and D_{OL} to reduce to 0 in the cases where introgression is $P_1/P_2 \Rightarrow P_3/P_4$. In this instance, D_{FI} and D_{OL} are only reflecting the reality that the introgression involves a population of P_1 or P_2 that is so recently diverged that they are practically indistinguishable genetically. So as τ approaches T_3, the D_{FOIL} statistics effectively convert to the signature of an ancestral introgression (+/+/0/0 or −/−/0/0).

One assumption that may be violated in real datasets is the expectation of uniform substitution rates across the tree. If a particular taxon has a much overall higher substitution rate than the others, then all distances involving that taxon would be relatively higher due to an increased rate of substitution rather than an introgression involving its sister taxon. There are two straightforward solutions to these lineage-specific effects. The first would be to weight the expected *D*-statistic values away from 50:50 when there is a prior expectation that substitution rates are not equal between sister taxa in a subgroup. The second would be to simply exclude the terminal-branch-substitution site patterns (AAABA, AABAA, ABAAA, and BAAAA) from all calculations. As long as these counts are excluded from both sides of the equation, the expectation of equality of the left and right terms of each D_{FOIL} statistic is not violated.

Ancestral Introgressions

Ancestral introgressions can also be detected by the D_{FOIL} framework, though the direction of this introgression cannot be detected strictly by the D_{FOIL} signature. When calculating a single, genome-wide estimate for each of the D_{FOIL} tests, determining the direction of ancestral introgressions is not possible. However, if, for example, both $P_{12}\Rightarrow P_3$ and $P_3\Rightarrow P_{12}$ are present at different locations in the genome, the difference in their expected values for D_{FO} and D_{IL} could be used to distinguish between these two cases in the different regions.

More Taxa, More Models

In order to detect introgression in a phylogeny of six or more taxa, the simplest option is simply to subsample a part of the tree in the appropriate configuration and either to use the four-taxon *D*-statistic or the five-taxon D_{FOIL} tests. The D_{FOIL} tests offer the added feature of information about the direction of introgression in the subgroup of interest, when a five-taxon subset is available in the symmetric configuration. Aside from the subsampling option, we envision that a formal expansion of the model for more than five taxa would also be possible for certain tree topologies using the previously described principles of the *D*-statistic. The symmetric five-taxon tree represents a special case with respect to ILS, since the topology dictates that all discordant coalescences must occur in the root. This leads to a simple distribution of gene trees and provides a topology where each of the four in-group taxa can be compared in a straightforward manner with the two taxa in the opposite sub-pairs. Beyond this special case though, the probability distribution of gene trees becomes far more complex due to increasing topological constraints (Rosenberg 2002; Degnan and Salter 2005; Degnan and Rosenberg 2006). Therefore, while an explicit solution in larger phylogenies is not impossible, the more straightforward solution is simply to subsample four taxa of interest and an appropriate outgroup.

Some care must be taken when subsampling from a larger tree, or when planning which taxa to sample from nature. Introgression from taxa not included in the sample—what Durand et al. (2011) refer to as “ghost taxa”—could cause incorrect results using D_{FOIL} or the *D*-statistic. When a pair of taxa introgress within the sampled clade, we infer introgression from closer relationships between these two taxa. However, if we only sampled the recipient of introgressed alleles, this taxon will appear to be unusually divergent from its sister taxon. Alternatively, a false positive of introgression could be detected if the ghost taxon has shared ancestry with one of the sampled taxa, or two taxa in the sample were recipients of introgressed alleles from the same ghost taxon. In general, however, we would expect that introgressions from taxa not in the sample would simply result in a noisier signal of introgression or increased rate of false negatives.

In the D_{FOIL} tests, only biallelic site patterns are used to determine the phylogenetic relationships between taxa. This simple distance measure is particularly ideal for closely related species with few sequence differences and many biallelic sites. Since the site patterns used in D_{FOIL} fundamentally derive from relative phylogenetic relationships and gene trees, these same underlying phylogenetic relationships could be used to build a D_{FOIL} method using other models of sequence evolution. The conceptual framework would remain the same, only the model of inferring phylogenetic relationships between taxa would be altered.

Conclusions

In clades of closely related species, where ILS is prevalent or there are few sequence differences (or both), it can be difficult to detect introgression. The D_{FOIL} system offers a simple means to infer introgression in a symmetric five-taxon clade, requires little computational power, and functions even with relatively little sequence divergence and high levels of ILS. When computed on multiple alignments of whole-chromosomes, the added spatial context will allow for the detection of localized introgressions throughout the genome. Spatial context also makes possible the detection of introgressions among different combinations of taxa at various locations throughout the genome. The D_{FOIL} statistic is also designed specifically to detect the direction of introgression, when sequence data for sufficient taxa are available.

The D_{FOIL} tests can be used to test for general introgression across the genome in datasets without reference genomes, using RNA-Seq, RAD-Seq, or other targeted sequencing technologies. Since such loci may not have a known order along chromosomes, these data are not suitable for locus-by-locus testing of introgression. For these data, a single genomic mean value for each of the four D_{FOIL} tests can be computed (as was done by Green et al. 2010), and the average direction of introgression can also be determined. This implementation of D*FOIL* offers a more diffuse — but still informative — look at introgression from a smaller subset of data without spatial context.

Supplementary Material

Simulation scripts, simulated site pattern counts and D-statistics are available through Dryad (URL PENDING). The DFOIL script is available on BitBucket (http://www.bitbucket.org/dfoil)

Funding

This work was supported by National Science Foundation grant MCB-1127059.

Acknowledgements

We thank Julien Dutheil, Michael Fontaine, Daniel Neafsey, and Nora Besansky for helpful discussion, and Gregg Thomas for comments on the manuscript.

REFERENCES

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. Evolution, 62:1908-1920.

De Mita S, Siol M. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet, 13:27.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet, 2:e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol, 24:332-340.

Degnan JH, Salter LA. 2005. Gene tree distributions under the coalescent process. Evolution, 59:24-37.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol, 28:2239-2252.

Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst Biol, 62:689-706.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? Evolution, 63:1-19.

Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC. 2012. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. Genome Res, 22:1499-1511.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai WW, Fritz MHY, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S. 2010. A draft sequence of the Neandertal genome. Science, 328:710-722.

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

Joly S. 2012. JML: testing hybridization from species trees. Mol Ecol Resour, 12:179-184.

Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. Am Nat, 174:E54-E70.

Maddison WP. 1997. Gene trees in species trees. Syst Biol, 46:523-536.

Meng C, Kubatko LS. 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, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol, 5:568-583.

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

Rosenberg NA. 2007. Counting coalescent histories. J Comput Biol, 14:360-377.

Sang T, Zhong Y. 2000. Testing hybridization hypotheses based on incongruent gene trees. Syst Biol, 49:422-434.

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

Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. Heredity, 108:179-189.

Yu Y, Barnett RM, Nakhleh L. 2013. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Syst Biol, 62:738-751.

Yu Y, Degnan JH, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet, 8:456-465.

TABLE 1. The expected sign (+, −, or 0) of D_{FO}, D_{IL}, D_{FI}, and D_{OL} for no introgression, eight inter-group introgressions, four ancestral introgressions, and two pairs of intra-group introgressions. Each inter-group introgression can be determined by a unique D_{FOIL} "signature."

Introgression	D_{FO}	D_{IL}	D_{FI}	D_{OL}
None	0	0	0	0
$P_1 \Rightarrow P_3$	+	+	+	0
$P_3 \Rightarrow P_1$	+	0	+	+
$P_1 \Rightarrow P_4$	–	–	0	+
$P_4 \Rightarrow P_1$	–	0	+	+
$P_2 \Rightarrow P_3$	+	+	–	0
$P_3 \Rightarrow P_2$	0	+	–	–
$P_2 \Rightarrow P_4$	–	–	0	–
$P_4 \Rightarrow P_2$	0	–	–	–
$P_{12} \Rightarrow P_3$	+	+	0	0
$P_3 \Rightarrow P_{12}$	+	+	0	0
$P_{12} \Rightarrow P_4$	–	–	0	0
$P_4 \Rightarrow P_{12}$	–	–	0	0
$P_1 \Leftrightarrow P_2$	0	0	0	0
$P_3 \Leftrightarrow P_4$	0	0	0	0

Figure Captions

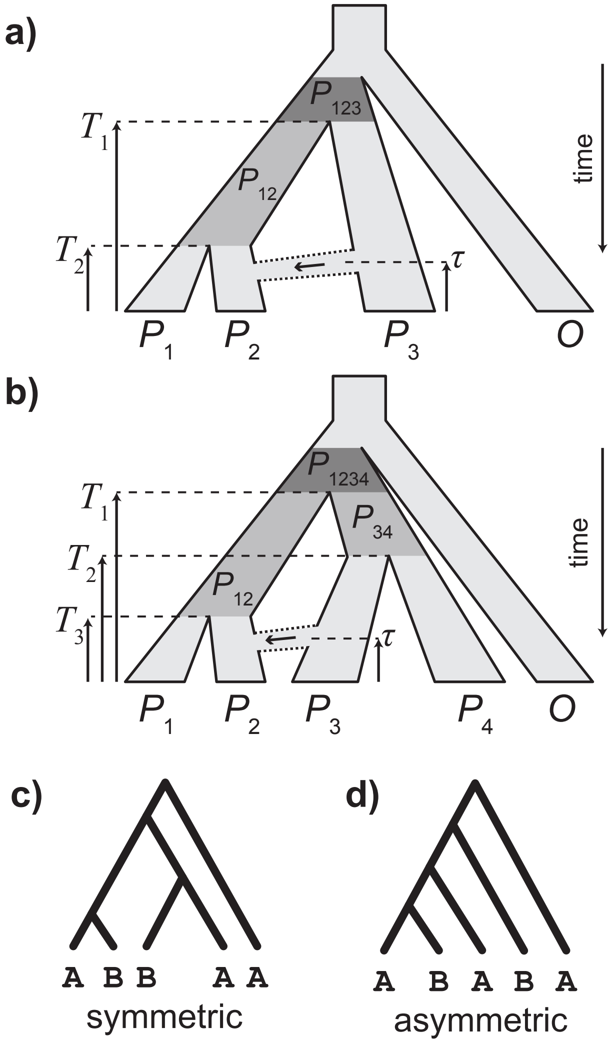
Figure 1. (a) A four-taxon phylogeny with three in-group taxa (P_1−P_3) and an out-group (O). (b) A five-taxon phylogeny with four in-group taxa (P_1–P_4) and an out-group (O). Ancestral branches are P_{12}, P_{34}, P_{123}, and P_{1234}. Introgressions (dotted region) are shown from $P_3 \Rightarrow P_2$ in both (a) and (b). The time-before-present of the two or three speciations, respectively, are T_1, T_2, and T_3, and time-since-introgression is τ. Five-taxon trees can have (c) symmetric or (d) asymmetric topologies, shown with examples of allelic state patterns ABBAA and ABABA, respectively.

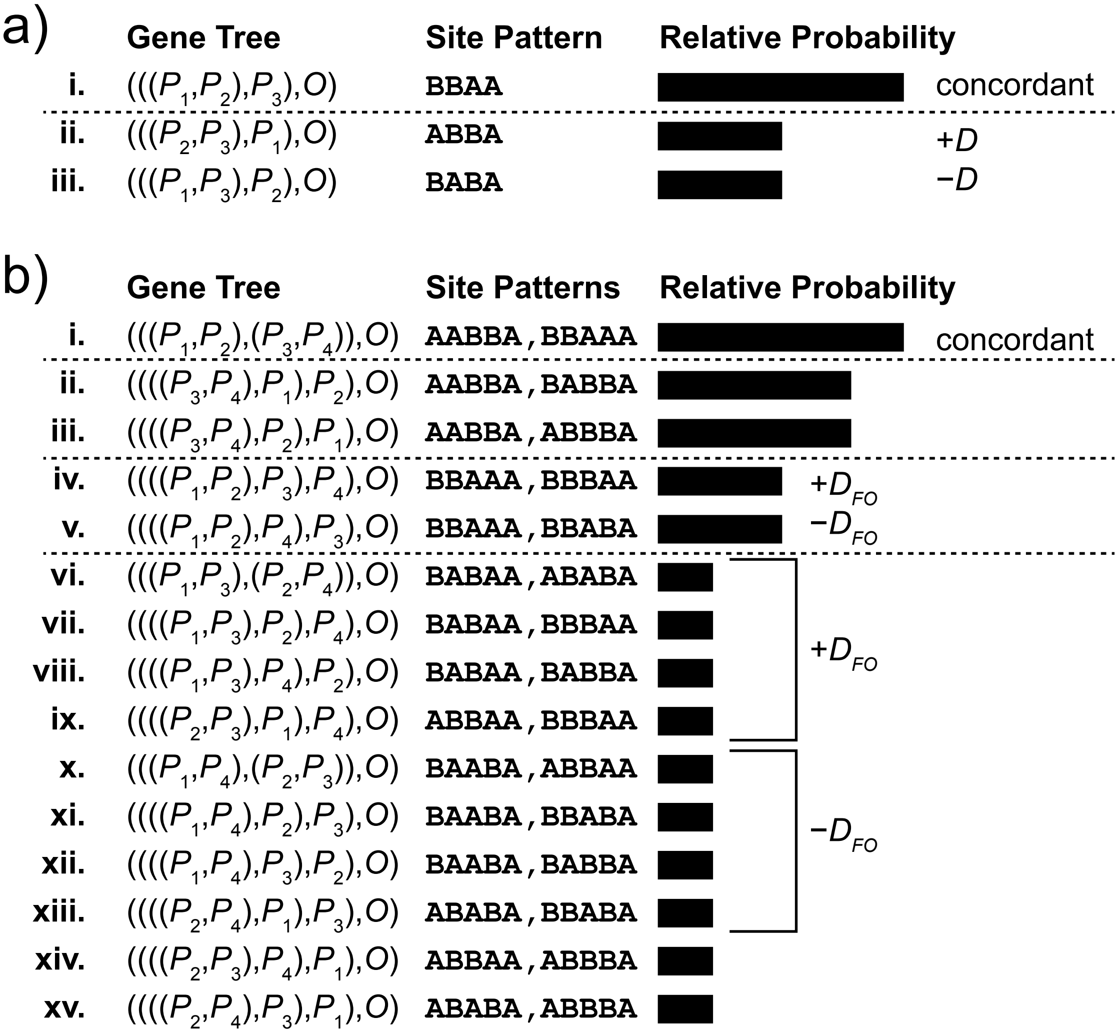
Figure 2. (a) The three possible gene trees for a four-taxon phylogeny, including the gene tree concordant with the species phylogeny (i) and the two discordant gene trees (ii and iii). For each gene tree the biallelic site patterns used in the D-statistic are shown. Note that the "left" (+D) and "right" (−D) terms of the D-statistic sample gene trees whose relative probability is equal. (b) The fifteen possible gene trees for a five-taxon phylogeny, including the concordant gene tree (i) and discordant gene trees (ii-xv). The effects of discordant topologies for the D_{FO} statistic is shown as an example. D_{FO} samples sets of gene trees for the "left" (+D_{FO}) and "right" (−D_{FO}) terms, whose relative probabilities are equal regardless of their absolute probabilities. The relative values of gene tree pairs where ILS only occurs on a single branch (ii/iii and iv/v) depend on the relative length of the ancestral branches P_{12} and P_{34}.

Figure 3. The D_{FOIL} statistics can detect both the taxa involved in and direction of introgression. (a) In a four-taxon symmetric topology, three classes of introgression can be defined: "intra-group" between taxa in the same subgroup, "inter-group" between taxa in different subgroups,

and "ancestral" involving the ancestral population of one subgroup. (b) Topology of tree used for simulations with species divergence times given in terms of millions of years ago (Ma) listed next to their respective nodes and inter-group introgression occurring between 0.20 and 0.25 Ma (dashed line). (c) The same topology showing the time of ancestral introgression between 0.53 and 0.58 Ma. (d) Simulated chromosomes with *n*=500 10 kb regions for each of the sixteen possible introgressions. Between each introgression are *n*=100 simulated 10 kb sequences without introgression. Mean values over 20 consecutive regions for each D_{FOIL} statistic are shown by solid and dashed lines. Each of the eight inter-group introgressions (first eight non-shaded columns from left) all show a unique D_{FOIL} signature (i.e., a unique combination of +/−/0 signs for D_{FO}, D_{IL}, D_{FI}, and D_{OL}) consistent with Table 1. The ancestral introgressions (involving P_{12}) can be distinguished between P_{12}⇔P_3 and P_{12}⇔P_4, but the direction cannot be determined by D_{FOIL} signature alone. For the intra-group introgressions (rightmost four non-shaded columns) and no-introgression treatments (narrow, shaded regions throughout), all three D_{FOIL} statistics are approximately 0 as expected.

FIGURE 4. Partitioned *D*-statistics calculated for the same simulated chromosome and phylogeny as shown in Figure 3b. The Partitioned *D*-statistics suffer from a "mirror effect" due to the inverse relationship of D_1 and D_2. For example, in the region under introgression P_1⇒P_3 (leftmost column), D_1 is negative as expected, but D_2 is positive (and should =0). Also, D_{12} has the same sign for introgressions in both directions for each pair of taxa.


a)
P_{123}
T_1
P_{12}
T_2
τ
time
P_1
P_2
P_3
O
b)
P_{1234}
T_1
P_{34}
T_2
P_{12}
T_3
τ
time
P_1
P_2
P_3
P_4
O
c)
A B B A A
symmetric
d)
A B A B A
asymmetric

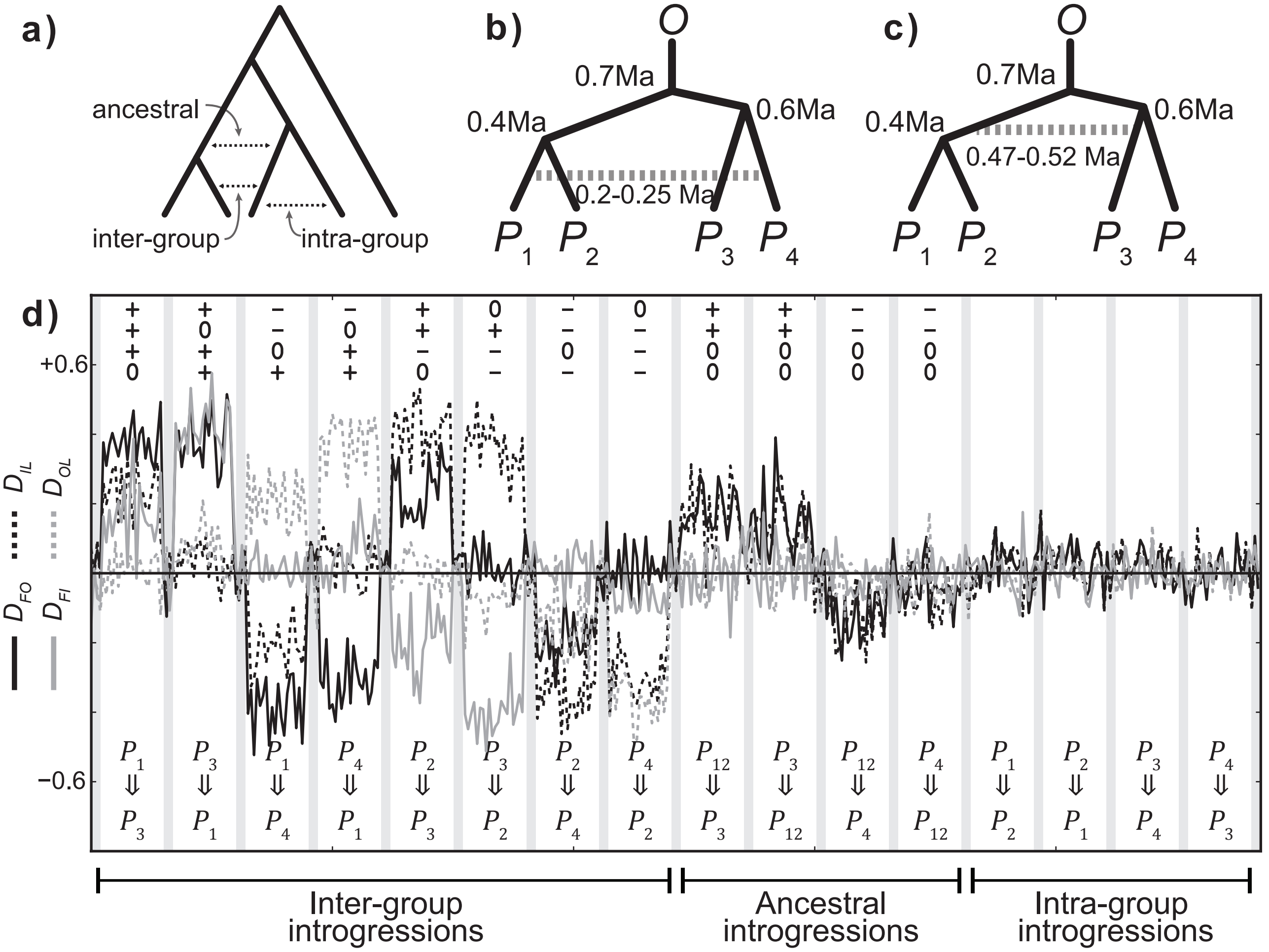


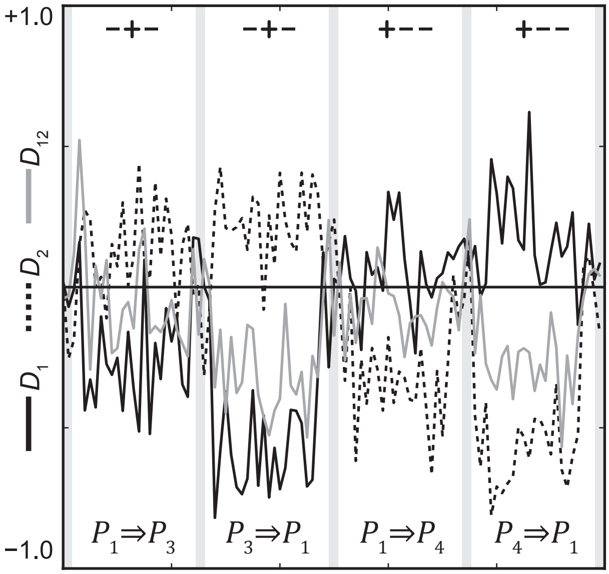
a)

	Gene Tree	Site Pattern	Relative Probability	
i.	$(((P_1,P_2),P_3),O)$	BBAA	█████████	concordant
ii.	$(((P_2,P_3),P_1),O)$	ABBA	█████	$+D$
iii.	$(((P_1,P_3),P_2),O)$	BABA	█████	$-D$

b)

	Gene Tree	Site Patterns	Relative Probability	
i.	$(((P_1,P_2),(P_3,P_4)),O)$	AABBA, BBAAA	█████████	concordant
ii.	$((((P_3,P_4),P_1),P_2),O)$	AABBA, BABBA	███████	
iii.	$((((P_3,P_4),P_2),P_1),O)$	AABBA, ABBBA	███████	
iv.	$((((P_1,P_2),P_3),P_4),O)$	BBAAA, BBBAA	█████	$+D_{FO}$
v.	$((((P_1,P_2),P_4),P_3),O)$	BBAAA, BBABA	█████	$-D_{FO}$
vi.	$(((P_1,P_3),(P_2,P_4)),O)$	BABAA, ABABA	██	$+D_{FO}$
vii.	$((((P_1,P_3),P_2),P_4),O)$	BABAA, BBBAA	██	
viii.	$((((P_1,P_3),P_4),P_2),O)$	BABAA, BABBA	██	
ix.	$((((P_2,P_3),P_1),P_4),O)$	ABBAA, BBBAA	██	
x.	$(((P_1,P_4),(P_2,P_3)),O)$	BAABA, ABBAA	██	$-D_{FO}$
xi.	$((((P_1,P_4),P_2),P_3),O)$	BAABA, BBABA	██	
xii.	$((((P_1,P_4),P_3),P_2),O)$	BAABA, BABBA	██	
xiii.	$((((P_2,P_4),P_1),P_3),O)$	ABABA, BBABA	██	
xiv.	$((((P_2,P_3),P_4),P_1),O)$	ABBAA, ABBBA	██	
xv.	$((((P_2,P_4),P_3),P_1),O)$	ABABA, ABBBA	██	


a)
ancestral
inter-group
intra-group
b)
O
0.7Ma
0.4Ma
0.6Ma
0.2-0.25 Ma
P1 P2 P3 P4
c)
O
0.7Ma
0.4Ma
0.6Ma
0.47-0.52 Ma
P1 P2 P3 P4
d)
+0.6
−0.6
DIL DOL
DFO DFI
+ + + 0
+ 0 + +
− − 0 +
− 0 + +
+ + − 0
0 + − −
− − 0 −
0 − − −
+ + 0 0
+ + 0 0
− − 0 0
− − 0 0
P1 ⇓ P3
P3 ⇓ P1
P1 ⇓ P4
P4 ⇓ P1
P2 ⇓ P3
P3 ⇓ P2
P2 ⇓ P4
P4 ⇓ P2
P12 ⇓ P3
P3 ⇓ P12
P12 ⇓ P4
P4 ⇓ P12
P1 ⇓ P2
P2 ⇓ P1
P3 ⇓ P4
P4 ⇓ P3
Inter-group introgressions
Ancestral introgressions
Intra-group introgressions


+1.0
−1.0
D_1
D_2
D_{12}
−+−
−+−
+−−
+−−
$P_1 \Rightarrow P_3$
$P_3 \Rightarrow P_1$
$P_1 \Rightarrow P_4$
$P_4 \Rightarrow P_1$