Syst. Biol. 64(4):651–662, 2015

DOI:10.1093/sysbio/syv023
Advance Access publication April 17, 2015

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, 1001 E. Third St., Bloomington, IN 47405, USA and* [2]*School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA*
[*]*Correspondence to be sent to: Department of Biology, Indiana University, 1001 E. Third St., Bloomington, IN, 47405, USA. E-mail: jbpease@indiana.edu*

Received 9 February 2015; reviews returned 17 March 2015; accepted 10 April 2015
Associate Editor: Laura Kubatko

Abstract.—When multiple speciation events occur rapidly in succession, discordant genealogies due to incomplete lineage sorting (ILS) can complicate the detection of introgression. A variety of methods, including the *D*-statistic (a.k.a. the "ABBA–BABA test"), have been proposed to infer introgression in the presence of ILS for a four-taxon clade. However, no integrated method exists to detect introgression using allelic patterns for more complex phylogenies. Here we explore the issues associated with previous systems of applying *D*-statistics to a larger tree topology, and propose new D_{FOIL} tests as an integrated framework to infer both the taxa involved in and the direction of introgression for a symmetric five-taxon phylogeny. Using theory and simulations, we show that the D_{FOIL} statistics correctly identify the introgression donor and recipient lineages, even at low rates of introgression. D_{FOIL} is also shown to have extremely low false-positive rates. The D_{FOIL} tests are computationally inexpensive to calculate and can easily be applied to phylogenomic data sets, both genome-wide and in windows of the genome. In addition, we explore both the principles and problems of introgression detection in even more complex phylogenies. [ABBA–BABA; *D*-statistics; genomics; hybridization; incomplete lineage sorting; introgression; phylogenetics; phylogenomics.]

In phylogenomic analyses, 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 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 can also occur through various forms of hybridization, ranging in scope from the introgression of alleles between species to the formation of new hybrid species (Currat et al. 2008; Twyford and Ennos 2012).

Since ILS and introgression/hybridization are both detected by the presence of discordant topologies, disentangling these two causes for discordance can be difficult. Many diverse approaches to this problem have been proposed previously. These methods have primarily been developed using the simplest case of a four-taxon phylogeny (three species and an outgroup; Fig. 1a). Given a consensus rooted four-taxon phylogeny under ILS alone, the two minor discordant trees should be sampled with equal frequency. In terms of sequence divergences, this means the relationship of each of the two paired taxa to the third in-group taxon should be equal (i.e., $P_1 - P_3 = P_2 - P_3$). However, introgression will cause an imbalance toward a closer relationship between the two taxa exchanging alleles. Most methods for detecting introgression rely on this basic principle, though they differ in how they quantify relationships between taxa and the imbalance of discordant phylogenies.

Some methods use sequence data to first construct gene trees, then reconcile the resulting topologies into a "reticulate phylogeny" to infer introgression, or to detect hybrid speciation events (Sang and Zhong 2000; Holder et al. 2001). Refinements of these tree topology-based methods continue to be explored (Meng and Kubatko 2009; Yu et al. 2012, 2013; Liu et al. 2014). Other methods distinguish ILS and introgression by calculating imbalances in mean or minimum pairwise-sequence distances without inferring a tree topology (Joly et al. 2009; Kulathinal et al. 2009; Joly 2012). The *4sp* algorithm (Garrigan et al. 2012) uses the relative frequencies of biallelic site patterns to determine regions of introgression. This algorithm first uses a maximum-likelihood approach to estimate global parameters of the species tree. Then the local likelihood of introgression for each region of the genome is calculated based on the relative proportions of various allelic site patterns.

In the *D*-statistic (a.k.a. the "ABBA–BABA test"; Huson et al. 2005; Green et al. 2010; Durand et al. 2011), a statistically significant imbalance in the number of sampled discordant biallelic site patterns "`ABBA`" and "`BABA`" gives evidence that introgression has occurred. Developed originally for use in hominids, this method has more recently been used to detect introgression in many other clades (e.g., Martin et al. 2013; Smith and Kronforst 2013; Jónsson et al. 2014; Fontaine et al. 2015). The *f*-statistics (f_4 for a four-population clade) are analogous to the *D*-statistic in form, but use population allele frequencies to estimate the proportion of admixture/introgression (Reich et al. 2009, 2011; Patterson et al. 2012). *D*-statistics only require

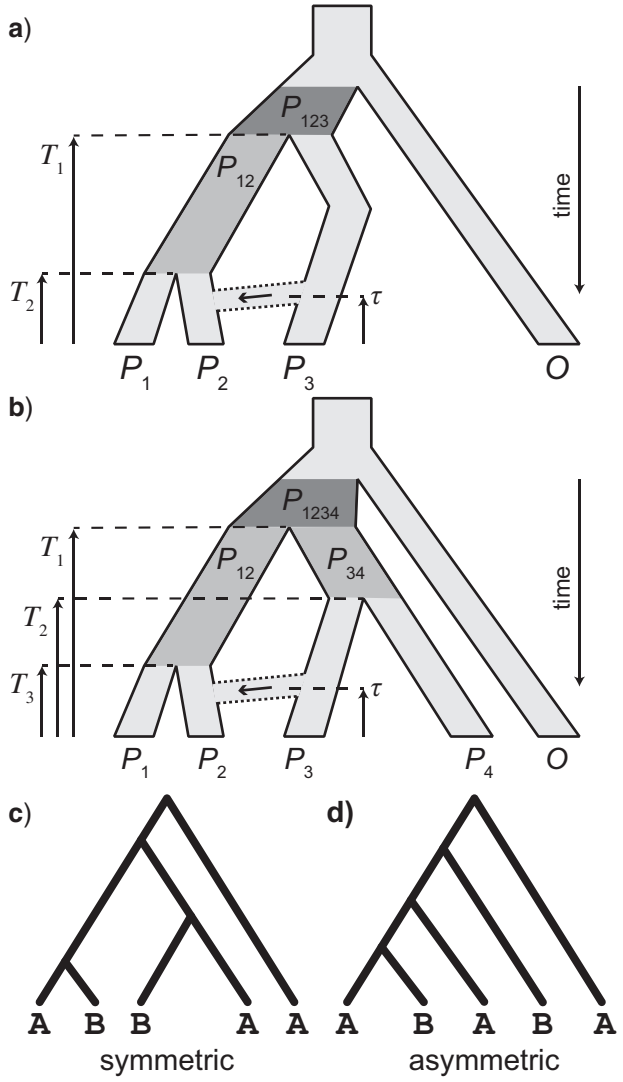

FIGURE 1. a) A four-taxon phylogeny with three in-group taxa ($P_1 - P_3$) and an out-group (O). b) A five-taxon phylogeny with four in-group taxa ($P_1 - P_4$) and an out-group (O). Ancestral branches are P_{12}, P_{34}, P_{123}, and P_{1234}. Introgressions (dotted region) are shown from $P_3 \Rightarrow P_2$ 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.

one sequence per taxon and are thus suitable for phylogenetic sampling, whereas f-statistics can only be used when robust population allele sampling data are available.

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 15 possible gene tree topologies, both symmetric (Fig. 1c) and asymmetric (Fig. 1d; see also Fig. 2). In addition to a greater number of gene trees, there are more 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).

Two previous methods have been proposed to address introgression among five taxa: The f_4-ratio test (Reich et al. 2009, 2011; Patterson et al. 2012) and the Partitioned D-statistics (Eaton and Ree 2013). These methods use a system of multiple four-taxon, f- or D-statistics, respectively, to test specific candidate introgression scenarios within a five-taxon phylogeny. However, neither of the methods presents a unified test that addresses all possible introgressions in a five-taxon phylogeny, making it possible that major introgression events could be missed. Here, we propose a new set of statistical measures (the D_{FOIL} tests) that comprise an integrated system to infer both the taxa involved in and the direction of introgression for all possible introgressions in a five-taxon symmetric phylogeny. The behavior of these statistics is explored theoretically and by simulation, followed by a discussion of their application.

MATERIALS AND METHODS

The Four-Taxon D-statistic

To approach testing for 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 out-group-rooted three- and four-taxon trees, where the out-group is not counted. For consistency with previous work on the D-statistic, we do not use this terminology here. Instead, we refer to "five-taxon trees" as trees with four in-group taxa rooted by a fifth out-group taxon.) The four-taxon D-statistic for introgression was formalized to test for ancestral introgression 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 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; Figs. 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 out-group cannot be involved in any introgression). The true 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).

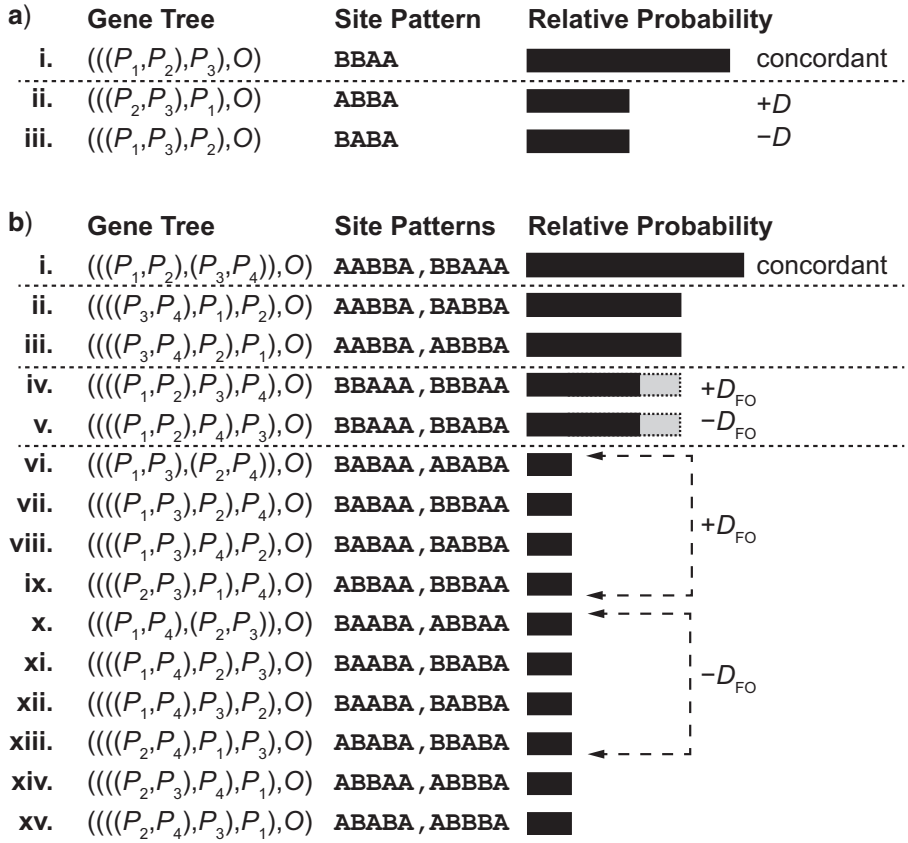

FIGURE 2. a) The three possible gene trees for a four-taxon phylogeny, including the gene tree concordant with the species phylogeny (i) and the two discordant gene trees (ii and iii). For each gene tree, the biallelic site patterns used in the *D*-statistic are shown. Note that the "left" (+*D*) and "right" (−*D*) terms of the *D*-statistic sample gene trees whose relative probability is equal. b) The 15 possible gene trees for an out-group-rooted five-taxon phylogeny, including the concordant gene tree (i) and discordant gene trees (ii–xv). The effect of discordant topologies on 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 shaded bars indicate the approximate relative rankings of expected frequency of gene trees, and on trees (iv) and (v) indicate that these gene trees can be equally as likely as (ii) and (iii) if T_2 and T_3 are equal.

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

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

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 $\Leftrightarrow$ denotes introgression of indeterminate direction and $\Rightarrow$ denotes a polarized introgression). Alternatively, a skew in *D* could indicate ancestral population structure (see section "Discussion"). 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, 2007; Degnan and Salter 2005; Degnan and Rosenberg 2006). Another practical concern is that the four-taxon *D*-statistic can only provide two answers (positive 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*-test expects that the biallelic site patterns ABBA and BABA are sampled with equal frequency under the null hypothesis (Equation (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 zero 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, they are expected to have equal relative probabilities regardless of their absolute probabilities (Fig. 2a). This feature will be important in designing *D*-statistics for five-taxon phylogenies.

Another 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} (see Fig. 1a for an explanation of branch names). 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 ABAA. This means that when introgression is considered, both BABA and its "inverse pattern," ABAA, offer evidence of $P_3 \Leftrightarrow P_1$ introgression, and the same is true for ABBA/BAAA 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 parameters that affect the number of derived states are unequal among terminal branches of the phylogeny (see section "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 Fig. 2). Finally, both inverse patterns (e.g., BABA/ABAA) indicate the same potential introgressions.

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

From these principles, we developed a model to describe a clade of five taxa connected by a symmetric phylogeny, denoted as $(((P_1,P_2),(P_3,P_4)),O)$, with the in-group taxa arranged in two subpairs (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) no later than P_1 and P_2 did (at T_3), and that the two subpair lineages diverged at T_1. The labeling of the taxa (P_1-P_4) is arbitrary, as long as the subpairings are correct and the relationships between the three speciation times-before-present adhere to the relationships $T_1 > T_2 \geqslant T_3 > 0$.

As in the four-taxon case, we only sample biallelic sites (with the out-group always represented as A). Introgressions 0 or 1 are allowed per site (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 subpair as an intergroup introgression (of which there are eight possible pairings) and those between taxa in the same subpair as an intragroup introgression (four pairings; 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 (four pairings; Fig. 3a).

We propose a system of four *D*-statistics, 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:

$$D_{\mathrm{FO}} = \frac{\begin{array}{c}\left(n_{\mathrm{BABAA}}+n_{\mathrm{BBBAA}}+n_{\mathrm{ABABA}}+n_{\mathrm{AAABA}}\right)\\ -\left(n_{\mathrm{BAABA}}+n_{\mathrm{BBABA}}+n_{\mathrm{ABBAA}}+n_{\mathrm{AABAA}}\right)\end{array}}{\begin{array}{c}\left(n_{\mathrm{BABAA}}+n_{\mathrm{BBBAA}}+n_{\mathrm{ABABA}}+n_{\mathrm{AAABA}}\right)\\ +\left(n_{\mathrm{BAABA}}+n_{\mathrm{BBABA}}+n_{\mathrm{ABBAA}}+n_{\mathrm{AABAA}}\right)\end{array}} \qquad (2)$$

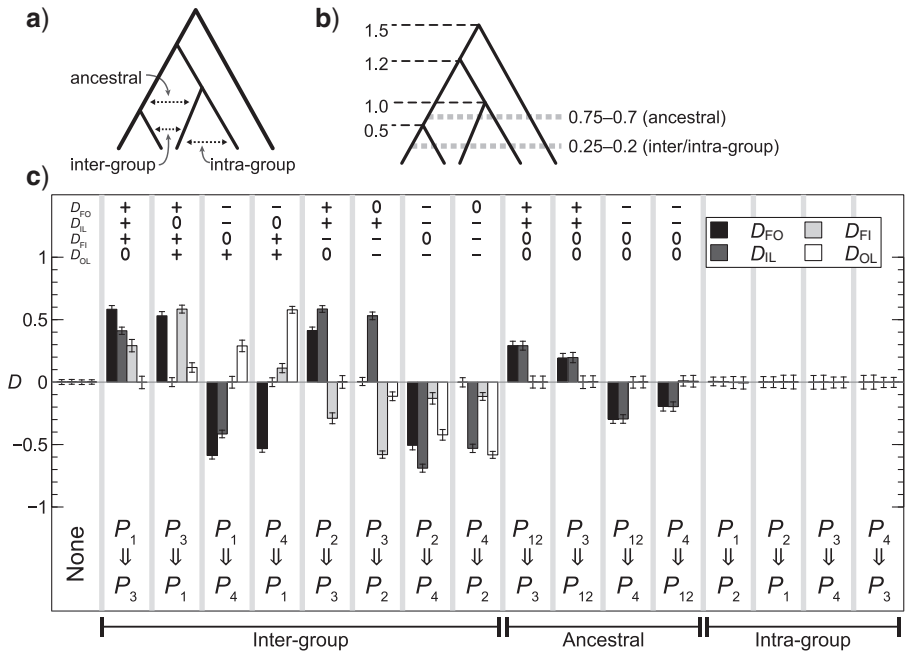

FIGURE 3. a) In a four-taxon symmetric topology, three introgression types can be defined: "intra-group" between taxa in the same subgroup, "inter-group" between taxa in different subgroups, and "ancestral" involving the ancestral population of one subgroup. b) Topology of tree used for simulations with species divergence times (given in $4N_e$ generations) next to their respective nodes. Intergroup and intragroup introgressions were simulated between 0.20 and $0.25 \cdot 4N_e$ generations ago and ancestral introgressions between 0.70 and 0.75 (gray dotted lines). c) Simulated 100 kb loci ($n = 100$) for no introgression and each of the 16 possible introgressions. Each of the eight intergroup introgressions (first eight nonshaded columns from left) shows a unique "signature" combination of $+/-/0$ signs for D_{FO}, D_{IL}, D_{FI}, and D_{OL} consistent with Table 1 (values shown are mean ± 1 SD). Ancestral introgressions (involving P_{12}) can be distinguished between $P_{12} \Leftrightarrow P_3$ and $P_{12} \Leftrightarrow P_4$, but direction cannot be determined. For the intragroup introgressions (rightmost four) and no-introgression treatments (left), all three D_{FOIL} statistics are ≈ 0 as expected.

$$D_{IL} = \frac{\left(n_{ABBAA} + n_{BBBAA} + n_{BAABA} + n_{AAABA}\right) - \left(n_{ABABA} + n_{BBABA} + n_{BABAA} + n_{AABAA}\right)}{\left(n_{ABBAA} + n_{BBBAA} + n_{BAABA} + n_{AAABA}\right) + \left(n_{ABABA} + n_{BBABA} + n_{BABAA} + n_{AABAA}\right)} \quad (3)$$

$$D_{FI} = \frac{\left(n_{BABAA} + n_{BABBA} + n_{ABABA} + n_{ABAAA}\right) - \left(n_{ABBAA} + n_{ABBBA} + n_{BAABA} + n_{BAAAA}\right)}{\left(n_{BABAA} + n_{BABBA} + n_{ABABA} + n_{ABAAA}\right) + \left(n_{ABBAA} + n_{ABBBA} + n_{BAABA} + n_{BAAAA}\right)} \quad (4)$$

$$D_{OL} = \frac{\left(n_{BAABA} + n_{BABBA} + n_{ABBAA} + n_{ABAAA}\right) - \left(n_{ABABA} + n_{ABBBA} + n_{BABAA} + n_{BAAAA}\right)}{\left(n_{BAABA} + n_{BABBA} + n_{ABBAA} + n_{ABAAA}\right) + \left(n_{ABABA} + n_{ABBBA} + n_{BABAA} + n_{BAAAA}\right)} \quad (5)$$

Each of these Equations (2)–(5) 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 subpair; 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 (Fig. 2b, "$+D_{FO}$" and "$-D_{FO}$"). Under ILS alone, the relative 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 toward $D_{FO} > 0$. Alternatively, $P_1 \Leftrightarrow P_4$ introgression will shift toward $D_{FO} < 0$. 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 subpair.

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, whereas 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 intergroup introgressions, specifically identifying both the donor and recipient taxa. When an intergroup introgression occurs in a symmetric five-taxon phylogeny, the relationship changes not only between the recipient taxa and the donor taxa, but also 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 the 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} test is separately assessed to be significantly positive, significantly negative, or not different from zero by a χ^2 goodness-of-fit test. The sum of site pattern counts in the "left" and "right" terms should be equal for each D_{FOIL} *D*-test ($n_{\mathrm{L}} = n_{\mathrm{R}}$), and thus the expected value is the average $[n_{\mathrm{L}} = n_{\mathrm{R}} = (n_{\mathrm{L}} + n_{\mathrm{R}})/2]$. Using these expectations, a χ^2 goodness of fit test (df = 1) can be constructed to determine the significance of deviations in each D_{FOIL} component:

$$\chi^2 = \sum_i \frac{\left(\mathrm{Observed}_i - \mathrm{Expected}_i\right)^2}{\mathrm{Expected}_i}$$

$$= \frac{\left(n_{\mathrm{L}} - \frac{n_{\mathrm{L}}+n_{\mathrm{R}}}{2}\right)^2}{\frac{n_{\mathrm{L}}+n_{\mathrm{R}}}{2}} + \frac{\left(n_{\mathrm{R}} - \frac{n_{\mathrm{L}}+n_{\mathrm{R}}}{2}\right)^2}{\frac{n_{\mathrm{L}}+n_{\mathrm{R}}}{2}} = \frac{\left(n_{\mathrm{L}} - n_{\mathrm{R}}\right)^2}{\left(n_{\mathrm{L}}+n_{\mathrm{R}}\right)} \quad (6)$$

TABLE 1. Expected signs of the D_{FOIL} components under all combinations of introgressing taxa.

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

Statistically significant positive or negative values of each D_{FOIL} component ($P < \alpha$) are assigned a sign of "+" or "−," while nonsignificant values are "0." The signs of the four tests in the order $\{D_{\mathrm{FO}}, D_{\mathrm{IL}}, D_{\mathrm{FI}}, D_{\mathrm{OL}}\}$ form the D_{FOIL} signature (Table 1).

Simulations

We tested the D_{FOIL} statistics on a set of simulated regions of different lengths, each of which was allowed to evolve over a phylogeny with or without introgression. All simulated regions were generated using *ms* (Hudson 2002). A population size of $N_e = 10^6$, mutation rate of $\mu = 7 \times 10^{-9}$ per site per generation, and recombination rate of $r = 1 \times 10^{-8}$ per site per generation were used. For computational efficiency in *ms*, we simulated 100 kb and 150 kb loci with $r = 1 \times 10^{-8}$ per site per generation by specifying 10 kb and 15 kb loci with $r = 1 \times 10^{-7}$ per site per generation (which have equivalent ρ values). Introgression was simulated by a temporary period of migration from the donor population to the recipient population at various fixed rates. Biallelic site patterns were counted for each region. To simulate the possible effects of convergent mutations in a finite-sites model, five additional substitutions per sequence window were added randomly to each region at sites with one or more existing derived states. D_{FOIL} and Partitioned *D*-statistics were calculated from the site pattern counts for each simulated window. To call the sign of each *D*-test as significantly positive or negative, the same χ^2 goodness-of-fit test described above was used with a cutoff of $P < 0.01$. Plots were generated using *Veusz* (http://home.gna.org/veusz/). The *dfoil* program used for all calculations, and *dfoil_sim* used for all simulations, simulation commands, and site count data sets are available for public use as Supplementary Material on Dryad (at http://dx.doi.org/10.5061/dryad.4h462). The *dfoil* program and future updates are available on GitHub (http://www.github.com/jbpease/dfoil).

Results

Application and Accuracy of the D_{FOIL} Method

To demonstrate the effectiveness of the D_{FOIL} method, we tested it on 100 sequences of 100 kb in length under simulation conditions with no introgression as well as all 16 possible introgressions for a five-taxon symmetric phylogeny (Fig. 3). In all 16 cases and the control (i.e., no introgression), the D_{FOIL} signature (signs of the four D_{FOIL} tests) that was observed in the simulations matched the theoretical expectations (Fig. 3c; Table 1). For example, the introgression $P_1 \Rightarrow P_3$ (Fig. 3c, leftmost introgression) 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 intergroup introgressions (Fig. 3c, leftmost eight introgressions), a unique D_{FOIL} signature clearly distinguishes each case.

Ancestral introgressions involving branch P_{12} were distinguishable between cases of introgression with either P_3 or P_4, but the direction of these introgressions cannot be determined by the D_{FOIL} signature. As expected, the average value for all $D-$statistics when there is no introgression (Fig. 3c, leftmost column), or when there are only intragroup introgressions (Fig. 3c, rightmost four), is zero. Therefore, the D_{FOIL} tests can distinguish the taxa involved in all introgression cases, and can additionally determine the direction of introgression for the eight intergroup introgressions.

To determine the window size for which the significance of each D-statistic can be accurately assessed, we simulated 10,000 loci of lengths 5, 25, 70, and 100 kb; and 5000 loci of length 150 kb over a common phylogeny with recombination rate of 1×10^{-8} per site per generation, and no introgression (Fig. 4). We find that the variance in D in very small windows (5–25 kb, under the recombination rates used here) drives up the false-positive rate of D (Fig. 4), consistent with previous findings (Martin et al. 2015). However, as the size of the window increases, more independently recombining subregions are incorporated into the sampled sequences, causing the variance in D to decrease. At a window size of ~100 kb, we find the distribution of D values becomes χ^2-distributed (Fig. 4). Therefore, we conclude that D-statistics, including the four components of D_{FOIL}, are usable for genomic regions where the population recombination parameter, "ρ" ($\rho = 4N_e rL$), is greater than ≈ 4000 (where L is length of the sampled window and r is the recombination rate per site per generation).

We also tested the power of the D_{FOIL} method by simulating 10,000 loci of 100 kb with introgression occurring at different times relative to speciation times, and with different strengths of introgression (Fig. 5a and b, respectively). For all tests shown in Figure 5a, $P_1 \Rightarrow P_3$ introgression was simulated at a rate of $m = 5$ individuals per generation for 50,000 generations. We find that regardless of the time of introgression, the power to detect any introgression event remains relatively constant (Fig. 5a). However, as the time of introgression approaches the time of speciation (i.e., moving from $0.1 \cdot 4N_e$ generations ago to $0.9 \cdot 4N_e$ generations ago),

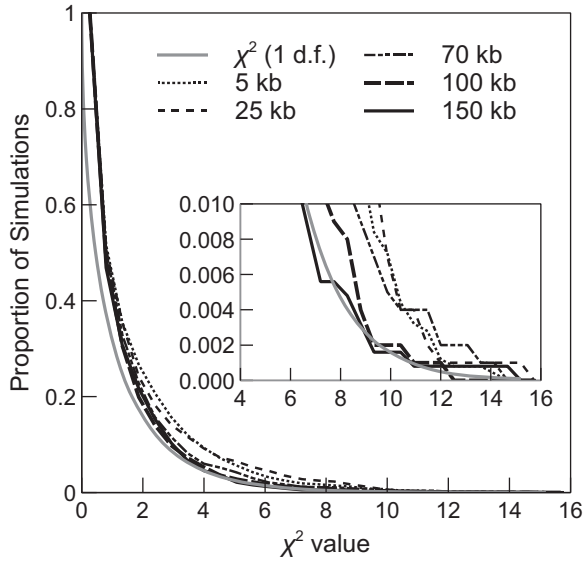

FIGURE 4. As the sampled window size increases, D-statistics (including the four components of D_{FOIL}) better fit a χ^2 distribution (df = 1, gray line). At window sizes $\geq$100 kb ($\rho \geq 4000$), the false-positive rate equals α. D_{FI} is shown, though all four D_{FOIL} components had equivalent distributions. The inset graph shows the right tail of the distribution in detail.

D_{FOIL} more often infers ancestral $P_{12} \Leftrightarrow P_3$ introgression. In these cases, the recipient taxon (P_3) is correctly identified, but the low sequence divergence makes it more difficult to determine whether P_1 or P_2 is the donor population. In a very rare number of cases (~0.2%), D_{FOIL} inferred introgression between the correct pair of taxa but in the opposite direction ($P_3 \Rightarrow P_1$). This incorrect inference was the result of stochastic error, caused by an introgression occurring when species split times T_2 and T_3 are equal; simulations that have different T_2 and T_3 values (as shown in Fig. 3) did not exhibit this reversal in the inferred direction of introgression (data not shown).

D_{FOIL} also has strong power to detect introgression at a variety of introgression strengths (as measured by the migration rate), identifying the correct introgression in 82.4% of simulations with $m = 500$ and 76.9% of simulations with $m = 50$ (Fig. 5b). Even with a migration rate of only $m = 5$, the correct introgression was inferred in 37.5% of simulations. At lower rates ($m = 0.1$–1), D_{FOIL} had lower power and inferred ancestral introgression in a small minority of cases (0.8–2%). Across all simulated rates of migration, introgression was never detected between the wrong pair of lineages or in the wrong direction. Therefore, even at low rates of introgression and at a range of relative introgression times, D_{FOIL} can accurately infer the correct pair of taxa and the correct direction of introgression.

The "Partitioned D-statistics"

We also examined the "Partitioned D-statistics" previously proposed to infer intergroup introgression in

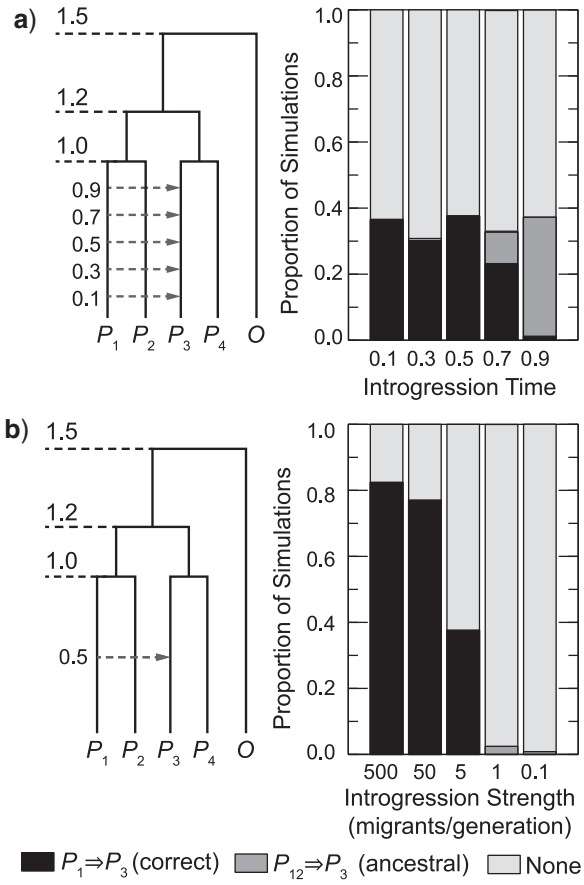

FIGURE 5. a) D_{FOIL} infers the correct introgression under a variety of introgression times relative to speciation. When the time of introgression approaches the time of speciation, the low sequence divergence causes D_{FOIL} to indicate introgression of the alleles from the ancestral branch instead of the correct $P_1 \Rightarrow P_3$. b) Even at low rates of migration, D_{FOIL} still has the power to identify the correct introgression (with no misidentifications). $P < 0.01$ cutoff used for all four D_{FOIL} components in all simulations.

a symmetric five-taxon phylogeny (Eaton and Ree 2013). These statistics are:

$$D_1 = \frac{n_{\text{ABBAA}} - n_{\text{BABAA}}}{n_{\text{ABBAA}} + n_{\text{BABAA}}} \quad (7)$$

$$D_2 = \frac{n_{\text{ABABA}} - n_{\text{BAABA}}}{n_{\text{ABABA}} + n_{\text{BAABA}}} \quad (8)$$

$$D_{12} = \frac{n_{\text{ABBBA}} - n_{\text{BABBA}}}{n_{\text{ABBBA}} + n_{\text{BABBA}}} \quad (9)$$

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 zero 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 polarize 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). P_3 or P_4 is assumed as 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 zero indicate P_1 or P_2 as the introgression donor and P_3 or P_4 as the recipient. The Partitioned D-statistics can be applied in the reverse, exchanging P_1/P_2 and P_3/P_4 to test for introgression in the opposite direction.

When applied to a symmetric five-taxon tree with ILS, a problem arises with the Partitioned $D-$statistics due to the principles of D-statistics. As noted previously, inverse pairs of biallelic patterns (e.g., ABBAA/BAABA) both indicate the same underlying gene tree when introgression is considered (see section "Materials and Methods"). This means that the site pattern counts used in D_1 and D_2 of the Partitioned D-statistic are not specifically indicative of the introgressions they propose to test. The left term of D_1 (n_{ABBAA}) and right term of D_2 (n_{BAABA}) indicate the same relationships between the four in-group taxa, as do the right term of D_1 (n_{BABAA}) and left term of D_2 (n_{ABABA}). Therefore, intergroup introgressions between any two taxa should change both D_1 and D_2 in opposite directions because these inverse patterns indicate the same relationships.

The 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. 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. 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 preintrogression 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 intergroup pair of taxa, introgressions in both directions will either both raise or both lower the value of D_{12}.

In all cases of intergroup introgression, simulated data show an inverse relationship between D_1 and D_2 and a direct relationship between D_1 and D_{12}, confirming the mirror effect (Fig. 6). For example, when the introgressions $P_1 \Rightarrow P_3$ and $P_3 \Rightarrow P_1$ (Fig. 6, left side) have occurred, the intended values for the Partitioned-D statistics are $D_1 < 0$ and $D_2 \approx 0$ (Equations (7) and (8)). 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. D_{12} also showed evidence of correlation with D_1 (Fig. 6). When the intended value of D_{12} was zero ($P_1 \Rightarrow P_3$, $P_1 \Rightarrow P_4$), the mean value of D_{12} was closer to zero with a higher variance, but it still deviates in the same direction regardless of direction of introgression between a pair of taxa. Therefore, we can conclude from both theory and simulations that D_1 and D_2 are inversely

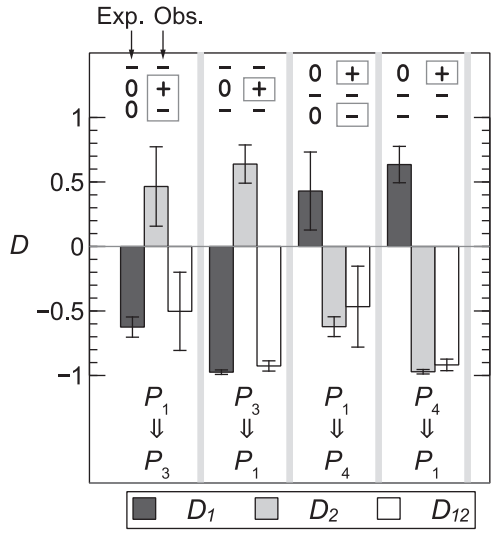

FIGURE 6. Partitioned *D*-statistics calculated from the same 100 kb regions as shown in Figure 3. In these cases, the Partitioned *D*-statistics exhibit a "mirror effect" due to an inverse relationship between D_1 and D_2. For example, in the region being introgressed from $P_1 \Rightarrow P_3$ (leftmost column), D_1 is negative as expected, but D_2 is positive where a value of zero is expected. In addition, D_{12} has the same sign for introgressions in both directions for each pair of taxa, though its distribution is closer to zero in cases where its expected value is zero ($P_1 \Rightarrow P_4$ and $P_1 \Rightarrow P_3$).

related and D_{12} is nonzero for introgression in both directions, though it remains possible that there are some parameter combinations for which these problems are less severe.

DISCUSSION

D_{FOIL} *is Accurate and Has High Power to Detect Introgression*

We have demonstrated through both theory and simulation that the D_{FOIL} test can correctly identify the taxa involved in and direction of introgression. Introgression is inferred with high power under a range of introgression rates and times, and with extremely low rates of false-positives. Furthermore, D_{FOIL} tests a five-taxon symmetric phylogeny for all detectable introgressions simultaneously, and does not require a specific candidate introgression.

D_{FOIL} was used in the analysis of the highly discordant phylogeny of the *Anopheles gambiae* species complex (Fontaine et al. 2015). In these data, D_{FOIL} inferred large-scale ancestral introgression between the *A. gambiae*+*coluzzii* ancestral lineage and *Anopheles arabiensis*. This large-scale introgression was consistent with the general model of the phylogeny of this group, which shows evidence of massive introgression on the autosomes. The extremely recent divergence (and likely ongoing gene flow) of *A. gambiae* and *A. coluzzii* did limit the power of D_{FOIL} testing in this phylogeny, consistent with the results of our simulations and the discussion below on ancestral introgression. Therefore, we have demonstrated that D_{FOIL} can be applied successfully to a large-scale genomic data set and can correctly infer introgression, even when the introgressed regions represent a large portion of the genome.

Considerations for Using D_{FOIL} *on Phylogenomic Data Sets*

Originally, the *D*-statistic and Partitioned *D*-statistics were designed for detection of relatively low levels of genome-wide admixture. However, the use of *D*-statistics and D_{FOIL} on subregions of the genome arranged spatially on chromosomes can provide a powerful approach for detecting introgression that is confined to a specific locus. Used in this way, D_{FOIL} can detect introgression either between different pairs of taxa or between the same two taxa in opposite directions in different parts of the genome.

Through simulations, we have demonstrated that *D* will have an acceptable false-positive rate as long as sites are sampled over a large enough region, relative to the amount of recombination. At a recombination rate of 10^{-8} per site per generation and $N_e = 10^6$, D_{FOIL} follows a χ^2 distribution for genomic regions of approximately 100 kb. We also find that the false-positive rate of introgression detected in smaller genome windows (5–25 kb) is too high for accurate D_{FOIL} application at this scale. This finding is consistent with the conclusions of Martin et al. (2015), who find that *D*-statistics have too much variance to be accurate in small genomic windows (they define these as 5 kb regions). These results imply that the physical size of the windows used for analysis may have to vary across the genome, with regions of lower recombination requiring much larger sequence windows. Furthermore, entities with extremely low recombination, such as the sex-limited Y- or W-chromosomes, mitochondrial or plastid genomes, or inversions, cannot use the same statistical cutoffs as previously described, since increased linkage will introduce higher variance in *D*-statistics. In such regions only a single, independent genealogy may be sampled, possibly leading to increased statistical support for a discordant topology as larger regions are sampled even though no introgression has occurred.

Since D_{FOIL} always counts biallelic patterns in inverse pattern pairs (i.e., `ABABA` and `BABAA`), determining which allele is ancestral or derived is not strictly necessary. This means that, mathematically, D_{FOIL} can be used without an out-group. However, this is not recommended in actual data analysis for two reasons. First, determining the relative substitution rates for each branch is important for testing for aberrant substitution/error rates on the terminal branches (as described in the next section). Second, a carefully chosen out-group is recommended for determining the consensus phylogeny in the first place. Even with these stipulations, strictly speaking, D_{FOIL} could be applied to two-related pairs of taxa even if an out-group is unavailable.

Possible Causes of Erroneous Introgression Inference

One assumption that may be violated in real data sets 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 number of substitutions rather than an introgression involving its sister taxon. This rate increase can be due to biological causes or simply a result of disproportionate error in one sequence or genome. There are two straightforward solutions to these lineage-specific effects. The first would be to change the expected *D*-statistic values away from a 1:1 ratio when there is a prior expectation that substitution rates are not equal between sister taxa in a subgroup. The second would be to exclude the terminal-branch-substitution site patterns (AAABA, AABAA, ABAAA, and BAAAA) from all calculations and instead to use them to calculate linage-specific error rates (as was done in the original *D*-statistic analysis; Green et al. 2010). As long as corresponding pairs of counts are excluded from both sides of the equation, the expectation of equality of the left and right terms for each D_{FOIL} statistic is not violated. We also note that four-taxon patterns containing a single derived allele (BAAA and ABAA) could also be used in the four-taxon *D*-statistic without violating any assumptions of this test (as is done in the *4sp* method; Garrigan et al. 2012).

In addition to variation in substitution rates, ancestral population structure can lead to scenarios where two taxa appear more related than expected. While the possibility of this has been shown in theory (Durand et al. 2011; Eriksson and Manica 2012), investigations of this phenomenon in human–Neanderthal data still support introgression over population structure (Yang et al. 2012; Lohse and Frantz 2014). For a five-taxon symmetric phylogeny, we can consider the case where the ancestral P_{1234} population had structure such that the ancestral subpopulation that would later lead to P_1 had a closer relationship to the ancestor of P_3 than the ancestor of P_4. In this case, the closer relationship between P_1 and P_3 might be inferred to be post-speciation introgression instead of population structure. However, D_{FOIL} requires a closer relationship between P_1 and both P_3 and P_4 in the case of $P_3 \Rightarrow P_1$, or between P_3 and both P_1 and P_2 in the case of $P_1 \Rightarrow P_3$ (corresponding to D_{FOIL} signatures of +0++ and +++0, respectively). Therefore, it does not seem likely that simple population structure could lead to an incorrect inference of introgression. Alternatively, if there was ancestral structure in P_{1234}, such that P_{12}, P_3, and P_4 were three structured subpopulations with closer relationships between P_{12} and P_3, this scenario might be inferred as ancestral introgression (++00 or −−00). Further work will be needed to explore all of the possible combinations of ancestral population structure and their consequences.

We also note that all of the D_{FOIL} signatures require at least two significant positive or negative components for ancestral introgression and at least three for intergroup introgressions. While the four *D*-tests in D_{FOIL} are not independent, and thus the *P*-values are not multiplicative, D_{FOIL}'s design requirement of multiple significant values for detection of introgression means that a single erroneous D_{FOIL} component is not enough to imply introgression when none has occurred (i.e., more than one ± sign is required to detect introgression). Additionally, no two intergroup introgressions differ by a single change of a ± sign to a zero. This means that if a single D_{FOIL} component is not statistically significant, this will not lead to inference of the wrong introgression. Instead, the signature will either default to an ancestral pattern (e.g., +++0 to ++00, see next section) or be an invalid pattern (e.g., +++0 to +0+0). Thus, the design of D_{FOIL} protects against inference of the wrong introgression.

Ancestral Introgressions

Ancestral introgressions can also be detected by the D_{FOIL} framework, though the direction of this introgression cannot be detected by the D_{FOIL} signature alone. When calculating a single, genome-wide estimate for each of the D_{FOIL} tests, determining the direction of ancestral introgressions is not possible. This is because at the time of ancestral introgression, there are only four lineages (P_{12}, P_3, P_4, and O) making this test more similar to the four-taxon *D*-statistic, which cannot explicitly polarize introgressions. In data sets where the phylogeny or order of divergence times may be unclear, the appearance of a 00++ or 00− D_{FOIL} signature may indicate $P_{34} \Rightarrow P_1$ or $P_{34} \Rightarrow P_2$, respectively. Since this introgression is not possible unless P_1 and P_2 diverged before P_3 and P_4, this may indicate that the labeling of taxa pairs may need to be swapped in accordance with the required labeling for D_{FOIL} (see section "Materials and Methods"). Note also that all intergroup D_{FOIL} signatures are interpretable independent of which pair of taxa was the first to diverge.

If introgression occurs very close to the second speciation event (i.e., $\tau \approx T_3$) this can cause D_{FI} and D_{OL} to reduce to zero 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 it is practically indistinguishable from its sister taxon. So as τ approaches T_3, the D_{FOIL} statistics effectively converge on the signature of an ancestral introgression (++00 or −−00). Also, if speciation of the two subgroups occurred at approximately the same time (i.e., $T_2 \approx T_3$), then there is no possibility to detect ancestral introgression because there are only three lineages (P_{12}, P_{34}, and O).

More Taxa, More Models

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 (though see the next section for caveats with this approach). 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 subpairs. 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).

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. For example, a Kimura two-parameter model could be used such that transitions and transversions are weighted differently when computing the *D*-value. The conceptual framework would remain the same, only the weighting scheme by which the left and right terms of *D* are calculated would be altered.

Introgression from Unsampled "Ghost" Lineages

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—commonly known as "ghost taxa"—could cause misleading results using D_{FOIL} or the *D*-statistic (Beerli 2004; Slatkin 2005). We infer introgression between two taxa by detecting a closer phylogenetic relationship than would be expected, given their relationships to other taxa. In the case of a "distal ghost" donor from a lineage entirely outside the sampled clade, the recipient of introgressed alleles is sampled, but not the donor. Therefore, distal ghost introgression makes the recipient taxon appear to be unusually divergent from its sister taxon without a corresponding convergence with a donor taxon from within the sampled sequences. This implies that the resulting pattern will not correspond to one of the D_{FOIL} signatures. Alternately, a "proximal ghost" donor from within the sampled group will have shared ancestry with at least two of the sampled lineages. This could result in a false-positive of introgression from any lineages related to the proximal ghost taxon. In general, however, we would expect that introgressions from taxa not in the sample (proximal or distal) would simply result in a noisier signal of any introgressions, leading to an increased rate of false-negatives.

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 data sets 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 (analogous to the application of the *D*-statistic in 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

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.4h462.

Funding

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

Acknowledgements

We thank Deren Eaton, Julien Dutheil, Michael Fontaine, Daniel Neafsey, and Nora Besansky for helpful discussion, and Gregg Thomas, Andy Anderson, Laura Kubatko, and three referees for comments on the manuscript.

REFERENCES

Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol. Ecol. 13:827–836.

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

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

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

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

Eaton D.A.R., Ree R.H. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62:689–706.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Eriksson A., Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. USA. 109:13956–13960.

Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.-C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

Garrigan D., Kingan S.B., Geneva A.J., Andolfatto P., Clark A.G., Thornton K.R., Presgraves D.C. 2012. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. Genome Res. 22:1499–1511.

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

Holder M.T., Anderson J.A., Holloway A.K. 2001. Difficulties in detecting hybridization. Syst. Biol. 50:978–982.

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.

Huson D.H., Klöpper T., Lockhart P.J., Steel M.A. 2005. Reconstruction of reticulate networks from gene trees. In: Miyano S., Mesirov J., Kasif S., Istrail S., Pevzner P.A., Waterman M., editors. Research in computational molecular biology. Berlin Heidelberg: Springer. p. 233–249.

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

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

Jónsson H., Schubert M., Seguin-Orlando A., Ginolhac A., Petersen L., Fumagalli M., Albrechtsen A., Petersen B., Korneliussen T.S., Vilstrup J.T., Lear T., Myka J.L., Lundquist J., Miller D.C., Alfarhan A.H., Alquraishi S.A., Al-Rasheid K.A.S., Stagegaard J., Strauss G., Bertelsen M.F., Sicheritz-Ponten T., Antczak D.F., Bailey E., Nielsen R., Willerslev E., Orlando L. 2014. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc. Natl Acad. Sci. USA. 111:18655–18660.

Kulathinal R.J., Stevison L.S., Noor M.A.F. 2009. The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. 5:e1000550.

Liu K.J., Dai J., Truong K., Song Y., Kohn M.H., Nakhleh L. 2014. An HMM-based comparative genomic framework for detecting introgression in eukaryotes. PLoS Comput. Biol. 10:e1003649.

Lohse K., Frantz L.A.F. 2014. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. Genetics 196:1241–1251.

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

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

Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol. Biol. Evol. 32: 244–257.

Meng C., Kubatko L.S. 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.

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

Reich D., Patterson N., Kircher M., Delfin F., Nandineni Madhusudan R., Pugach I., Ko Albert M.-S., Ko Y.-C., Jinam Timothy A., Phipps Maude E., Saitou N., Wollstein A., Kayser M., Pääbo S., Stoneking M. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am. J. Hum. Genet. 89:516–528.

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

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

Rosenberg N.A. 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.

Slatkin M. 2005. Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. Mol. Ecol. 14:67–73.

Smith J., Kronforst M.R. 2013. Do *Heliconius* butterfly species exchange mimicry alleles? Biol. Lett. 9:20130503.

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

Twyford A.D., Ennos R.A. 2012. Next-generation hybridization and introgression. Heredity 108:179–189.

Yang M.A., Malaspinas A.-S., Durand E.Y., Slatkin M. 2012. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. Mol. Biol. Evol. 29:2987–2995.

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

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