D_{GEN}: A Test Statistic for Detection of General Introgression Scenarios

Ryan A. Leo Elworth
Department of Computer Science, Rice University, 6100 Main Street, Houston, TX, USA
r.a.leo.elworth@rice.edu

Chabrielle Allen
Department of Computer Science, Rice University, 6100 Main Street, Houston, TX, USA

Travis Benedict
Department of Computer Science, Rice University, 6100 Main Street, Houston, TX, USA

Peter Dulworth
Department of Computer Science, Rice University, 6100 Main Street, Houston, TX, USA

Luay Nakhleh
Department of Computer Science and Department of BioSciences, Rice University, 6100 Main Street, Houston, TX, USA
nakhleh@rice.edu

Abstract

When two species hybridize, one outcome is the integration of genetic material from one species into the genome of the other, a process known as introgression. Detecting introgression in genomic data is a very important question in evolutionary biology. However, given that hybridization occurs between closely related species, a complicating factor for introgression detection is the presence of incomplete lineage sorting, or ILS. The D-statistic, famously referred to as the "ABBA-BABA" test, was proposed for introgression detection in the presence of ILS in data sets that consist of four genomes. More recently, D_{FOIL} – a set of statistics – was introduced to extend the D-statistic to data sets of five genomes.

The major contribution of this paper is demonstrating that the invariants underlying both the D-statistic and D_{FOIL} can be derived automatically from the probability mass functions of gene tree topologies under the null species tree model and alternative phylogenetic network model. Computational requirements aside, this automatic derivation provides a way to generalize these statistics to data sets of any size and with any scenarios of introgression. We demonstrate the accuracy of the general statistic, which we call D_{GEN}, on simulated data sets with varying rates of introgression, and apply it to an empirical data set of mosquito genomes.

We have implemented D_{GEN} and made it available, both as a graphical user interface tool and as a command-line tool, as part of the freely available, open-source software package ALPHA (https://github.com/chilleo/ALPHA).

2012 ACM Subject Classification Applied computing → Genomics, Applied computing → Computational biology

Keywords and phrases Introgression, genealogies, phylogenetic networks

Digital Object Identifier 10.4230/LIPIcs.WABI.2018.19

Funding This work was partially supported by NSF grants DBI-1355998, CCF-1302179, CCF-1514177, and DMS-1547433.

18th International Workshop on Algorithms in Bioinformatics (WABI 2018).
Editors: Laxmi Parida and Esko Ukkonen; Article No. 19; pp. 19:1–19:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

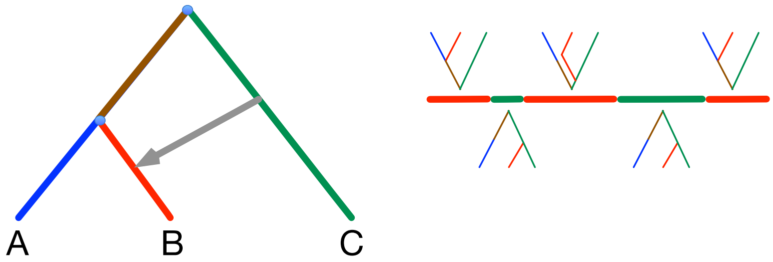

Figure 1 Hybridization and introgression. (Left) A phylogenetic network modeling the evolutionary history of three species (or, populations) A, B, and C. Species C split from the most recent common ancestor of A and B, and hybridization between (an ancestor of) C and (an ancestor of) B occurred. (Right) Due to hybridization and backcrossing, the genome of an individual in species B is a mosaic with different genomic segments having different genealogies. In particular, the genealogy of the middle segment involves incomplete lineage sorting.

1 Introduction

Hybridization – the interbreeding of individuals from two "different" species, or populations – has been recognized as an important evolutionary process underlying genomic diversification and species adaptation [1, 2, 22, 14, 15, 21, 8, 19, 16, 26]. Immediately upon interbreeding, each chromosome in the hybrid individual has a single source – the genome of one of the two parents. However, after multiple generations of backcrossing and recombination, the genomes of descendants of the hybrid individual turn into mosaics of genomic segments, each having a genealogy that could potentially differ from that of other segments (Fig. 1). The integration of genetic material from two different species into the genome of an individual is called *introgression*.

The discordance among the genealogies of different genomic segments could be used as a signal to detect introgression. For example, in the case of Fig. 1, the presence of some genealogies that place B closer to A than to C and others that place B closer to C than to A could indicate a potential hybridization event between B and C. However, a complicating factor in introgression detection is that incomplete lineage sorting, or ILS, could also be at play in cases where hybridization has occurred. ILS occurs when lineages from related populations fail to coalesce within the ancestral population, giving rise to the possibility that some lineages coalesce with others from farther populations. Mathematically, this process is often modeled by the multispecies coalescent [10, 24, 17, 5].

One class of methods for detecting hybridization and introgression, including in the presence of ILS, is to infer phylogenetic networks from the data of multiple unlinked loci sampled across the genomes. Indeed, several methods were introduced recently for this task [31, 29, 27, 23, 25, 32, 34, 33]. While providing accurate results, these methods are computationally very demanding.

A different approach is to use the so-called D-statistic [9, 6], which infers the presence of introgression based on significant deviation from equality between the frequencies of two site patterns in a 3-taxon (plus an outgroup) data set (details below). More recently, Pease and Hahn [18] introduced D_{FOIL}, which extends the D-statistic to detect introgression in a 5-taxon scenario (4 taxa plus an outgroup). The extension from three to four taxa involved a detailed analysis of site patterns and resulted in a set of statistics that, when combined, would aid in the detection of introgression. However, as stated, that work of Pease and Hahn extended the D-statistic from three to four taxa. Both the D-statistic and D_{FOIL} are examples of

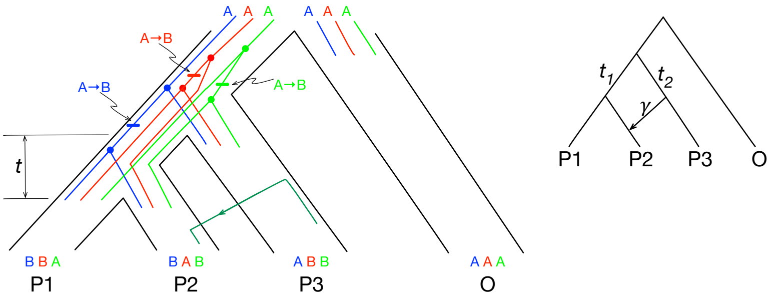

Figure 2 Illustration of the *D*-statistic. (Left) The demographic structure of three populations, P1, P2, and P3, along with an outgroup O, is shown. Patterns of the three parsimony-informative mutations (A→B) are shown for bi-allelic sites with states A and B, each mapped onto a different genealogy. The three genealogies give rise to patterns BBAA, BABA, and ABBA for the four taxa, respectively, when the taxa are listed in the order P1-P2-P3-O. The dark green arrow indicated gene flow from P3 to P2, which would result in excess of pattern ABBA. (Right) The phylogenetic network modeling the evolutionary history of the populations in the presence of gene flow from P3 to P2, where the gene flow is modeled as an instantaneous unidirectional event.

the use of phylogenetic invariants to detect deviation from the expected frequencies of site patterns under a neutral coalescent model with no gene flow. Similarly, the HyDe software package [3] implements an invariants-based method for identifying hybridization [13].

A major question is: Can one devise a statistic that is general enough to apply (the computational complexity issue aside) to data sets with any number of genomes and any set of postulated hybridization events?

In this paper, we address this question by showing that the phylogenetic invariants underlying both the D-statistic and D_{FOIL} could be generated automatically by contrasting gene tree distributions under the null multispecies coalescent [5] and the alternative multispecies network coalescent [30, 31]. Based on this observation, we devise an algorithm that automatically generates a statistic for detecting introgression in any evolutionary scenario. It is important to note, though, that as the number of genomes and number of postulated hybridization events increase, computing the statistic becomes computationally very demanding.

Our method, which we call D_{GEN}, is implemented in the publicly available, open-source software package ALPHA [7]. We demonstrate the accuracy of the method on simulated data sets, as well as its applicability to an empirical data set of mosquito genomes.

2 Methods

2.1 The *D*-statistic

Consider the species tree (((P1,P2),P3),O) in Fig. 2, which shows the evolutionary history of three species, or populations, P1, P2, and P3, along with an outgroup O. The significance of an outgroup in this scenario is that for any genomic site, the state that the outgroup has for that site is assumed to be the ancestral state of all three species P1, P2, and P3. We denote by A the ancestral state and by B the derived state.

Assuming all lineages from P1, P2, and P3 coalesce before any of them could coalesce with a lineage from O, there are three possible gene trees topologies, which are shown inside the branches of the species tree in Fig. 2, and are given by (((P1,P2),P3),O), (((P1,P3),P2),O), and (((P2,P3),P1),O). The probabilities of these three gene tree topologies when gene flow is excluded (but incomplete lineage sorting is accounted for) are, respectively, $1 - (2/3)e^{-t}$, $(1/3)e^{-t}$, and $(1/3)e^{-t}$, where t is the length, in coalescent units, of the branch that separates the splitting of P3 from the ancestor of P1 and P2 [4]. Clearly, the latter two gene tree topologies (those that are discordant with the species tree) have equal probabilities. Taking the two patterns BABA and ABBA to correspond to gene trees (((P1,P3),P2),O) and (((P2,P3),P1),O), then their expected frequencies in the absence of gene flow are equal.

However, when gene flow from P3 to P2 occurs and is modeled as an instantaneous event with probability γ (γ here is taken to represent the fraction of genomes in P2 that originated from P3 through gene flow), then the probabilities of the three gene tree topologies (((P1,P2),P3),O), (((P1,P3),P2),O), and (((P2,P3),P1),O) become, as derived in [28], $(1 - \gamma)(1 - (2/3)e^{-t_1}) + (1/3)\gamma e^{-t_2}$, $(1/3)(1-\gamma)e^{-t_1} + \gamma(1 - (2/3)e^{-t_2})$, and $(1/3)(1-\gamma)e^{-t_1} + (1/3)\gamma e^{-t_2}$, respectively, where t_1 and t_2 are the branch lengths, in coalescent units, in the phylogenetic network of Fig. 2. Now, with gene flow accounted for, when $\gamma \neq 0$ (and $t_2 > 0$), the expected frequencies of the two patterns BABA and ABBA are no longer equal. Thus, denoting by N_X the number of times site pattern X appears in a genomic data set, the D-statistic was defined as [6]

$$D = \frac{N_{ABBA} - N_{BABA}}{N_{ABBA} + N_{BABA}}, \tag{1}$$

and the significance of the deviation of D from 0 is assessed. Under no gene flow, we expect $D \approx 0$ (we do not write $D = 0$ since the counts in Eq. (1) are estimated from actual data and might not match the theoretical expectations exactly), and in the presence of gene flow, we expect D to deviate significantly from 0. Furthermore, when $D > 0$, it indicates introgression between P2 and P3 (in either or both directions), and when $D < 0$, it indicates introgression between P1 and P3.

To extend the D-statistic from the scenario depicted in Fig. 2 to the case of five taxa (four populations and an outgroup), Pease and Hahn [18] identified sets of site patterns that are expected to have equal frequencies under a no gene flow scenario but different frequencies when gene flow occurs. Next we show how to derive a general D-statistic that applies to a species phylogeny and any set of gene flow events, thus overcoming the need to derive a specialized D-statistic for individual evolutionary histories.

2.2 Towards the General Case

Let X be a set of taxa $X_1, X_2, \ldots, X_n$, where X_n is assumed to be an outgroup whose state A for a given bi-allelic marker is assumed to be the ancestral state. Then, for a given marker, a site pattern s is a sequence of length n where s_i $(1 \leq i < n)$, the state of the site in the genome of X_i, is either A or B.

Let $\mathcal{G}$ be the set of all rooted, binary gene trees on the n taxa $X_1, \ldots, X_n$. For a site pattern s, there might be multiple trees in $\mathcal{G}$ that are *compatible* with s; that is, trees on which the pattern s could have arisen in the presence of a single mutation (the infinite-sites assumption). We denote by $\mathcal{G}(s)$ the set of all trees in $\mathcal{G}$ that are compatible with pattern s. While the size of $\mathcal{G}$ only depends on the number of taxa n, the size of $\mathcal{G}(s)$ for a given s also depends on the number of ancestral versus derived alleles represented in s. For a given s with n total taxa and β taxa having the derived state (a 'B' instead of a 'A' in the site pattern), the size of $\mathcal{G}(s)$ will be the number of rooted, binary trees on β taxa times the number of rooted, binary trees on $n - \beta + 1$ taxa. Given a species phylogeny Ψ, we have

$$P(s|\Psi) = \sum_{g \in \mathcal{G}(s)} P(g|\Psi) \tag{2}$$

where $P(g|\Psi)$ is the probability mass function (pmf) of [4] when Ψ is a species tree and the pmf of [28] when Ψ is a phylogenetic network.

Assuming independence among sites, the expected number of occurrences of site pattern s in a genomic data set given species phylogeny Ψ, denoted by $\mathbb{E}(n(s))$, is given by $n \cdot P(s|\Psi)$. Using this notation, a general statistic for detecting introgression proceeds as follows. Let Ψ be the species tree that corresponds to the evolutionary scenario of no gene flow. Let Ψ' be the phylogenetic network that is obtained by adding to Ψ the gene flow events (instantaneous events represented by horizontal edges) to be tested on Ψ. For example, the phylogenetic network in Fig. 2 is obtained by adding the gene flow event from P3 to P2. A general D-statistic, D_{GEN}, is then computed as follows:

1. Let S be the set of all distinct parsimony-informative site patterns.
2. Parameterize Ψ and Ψ' so that they define probability distributions on gene tree topologies.
3. For every site pattern $s \in S$, compute $P(s|\Psi)$ and $P(s|\Psi')$.
4. Let $\mathcal{P}^{tree}(S)$ be the partition of set S induced by the equivalence relations $\{(s_1, s_2) : P(s_1|\Psi) = P(s_2|\Psi)\}$.
5. Let $\mathcal{P}^{network}(S)$ be the partition of set S induced by the equivalence relations $\{(s_1, s_2) : P(s_1|\Psi') = P(s_2|\Psi')\}$.
6. Let $S' \subseteq \mathcal{P}^{tree}(S)$ where $Y \in S'$ if and only if $Y \nsubseteq Z$ for any $Z \in \mathcal{P}^{network}(S)$. In other words, Y is an element of S' if it consists of a set of site patterns that all have equal probabilities under Ψ but not equal probabilities under Ψ'.
7. Let $U = \{(T_Y, B_Y, M_Y) : Y \in S', T_Y = \text{argmax}_{\{Y \cap Z : Z \in \mathcal{P}^{network}(S), Y \cap Z \neq \emptyset\}} P(s|\Psi')$ and $B_Y = \text{argmin}_{\{Y \cap Z : Z \in \mathcal{P}^{network}(S), Y \cap Z \neq \emptyset\}} P(s|\Psi')\}$ where s is an arbitrary element of $Y \cap Z$, and $M_Y = Y - (T_Y \cup B_Y)$. Put simply, site pattern probabilities that were previously equal in the tree case become totally ordered in the network case and can be divided into sets based on their new relation to one another. In other words, as an equivalence class $Y \in S'$ is refined by the elements of $\mathcal{P}^{network}(S)$, T_Y and B_Y are the two subsets of site patterns in Y with the highest and lowest probabilities, respectively, and M_Y is the set of remaining site patterns.
8. $D_{\mathrm{GEN}} = \left(\sum_{(T,B,M) \in U} N_T - N_B\right) / \left(\sum_{(T,B,M) \in U} N_T + N_B + 2N_M\right)$ where, as above, N_T is the number of times site patterns in T appear in the genomic data set (and similarly for N_B and N_M).
9. Similar to [18], calculate the χ^2 goodness of fit (df=1) using

$$\chi^2 = \left(\sum_{(T,B,M) \in U} N_T - N_B\right)^2 \Bigg/ \left(\sum_{(T,B,M) \in U} N_T + N_B + 2N_M\right).$$

Applying this algorithm to the case illustrated in Fig. 2, we have

- $\mathcal{P}^{tree}(S) = \{\{BBAA\}, \{BABA, ABBA\}\}$.
- $\mathcal{P}^{network}(S) = \{\{BBAA\}, \{BABA\}, \{ABBA\}\}$.
- Step (6) returns $S' = \{\{BABA, ABBA\}\}$.
- Step (7) returns $U = \{(\{BABA\}, \{ABBA\})\}$ which, indeed, is the D-statistic in the case of three taxa.

In Step (2) of the algorithm, we parameterize Ψ (and Ψ') by trying branch lengths (in coalescent units) in the set of values $\{0.5, 1.0, 2.0, 4.0\}$ and the set S' (in Step (6)) is determined based on the sets of site patterns whose equality does not break across the different settings of branch lengths. For the inheritance probability, we set it to 0.9.

In Step (7), if for an element (T, B, M) of U, we have $|T| \neq |B|$, we remove (arbitrarily) elements from the larger of the two sets to make them of equal size. Here, $|T|$ is distinguished from N_T as $|T|$ represents how many site patterns are contained in set T (the cardinality of set T), whereas N_T represents the occurrence of the site patterns in T in an actual multiple sequence alignment.

For the χ^2 test, we used a threshold of 0.01 on the p-value to determine significance. That is, if the p-value is smaller than 0.01 we considered support for introgression to be statistically significant; otherwise, it is not. Given that our formulation bases its determination of whether introgression is present or not off of significant deviations of D_{GEN} away from zero, sign changes are treated equivalently and thus one could equivalently choose to take the absolute value of D_{GEN}. In our implementation of D_{GEN} we have chosen to leave the sign. More information on this is given in the discussion section.

Why not contrast the site pattern distribution to the known distribution of gene trees? One question that might arise is: Why do we not use a χ^2 test to compare the two distributions – the empirical one and the theoretical one; that is,

$$\chi^2 = \sum_s \frac{(N_s - n_s)^2}{n_s},$$

where the sum is taken over all distinct site patterns s, N_s is the observed count of site pattern s, and $n_s = n \cdot P(\mathcal{G}(s)|\Psi)$. The problem with this approach is that to compute n_s, we need knowledge of the parameters (branch lengths) of the species phylogeny, which are unknown in this case. One potential remedy to this limitation is to first estimate the species tree parameters from the data, say under maximum likelihood, and then use this parameterized model to compute the n_s frequencies. However, it is unknown how the estimated parameters compare to the (unknown) true values when gene flow had occurred but the assumed topology in the estimation is a tree. In our solution above, this problem is remedied by not focusing on the parameter values in an absolute sense, but rather use arbitrary settings to find the site patterns whose relative frequencies change between a model of no gene flow and another with gene flow.

3 Results

3.1 Simulations

We first studied the performance of our method on the five-taxon scenario studied in [18] and given by species tree Ψ_1 in Fig. 3. All simulations share the same values for several parameters. As in [18], we have a constant fixed population size of $N_e = 10^6$ and recombination rate of $r = 10^{-8}$. We also use a fixed mutation rate of $\mu = 7 \times 10^{-9}$. In our simulation pipeline, we first generate gene trees for a 50kbp multiple sequence alignment using `ms` [11], followed by simulating the sequences under the Jukes-Cantor model of evolution [12] using `seq-gen` [20]. In other words, the sequences are evolved under a finite-sites model. The parameter values were chosen primarily to accomplish the two goals of being similar to relevant past work as well as being biologically relevant. An example of the full commands for this pipeline, before adding any reticulations is as follows:

```
 ms 5 1 -t 14000 -T -r 2000 50000 -I 5 1 1 1 1 1 -ej 1.0 2 1 -ej 1.0 4 3
-ej 1.2 3 1 -ej 1.5 5 1 | tail +4 | grep -v // > treefile
 seq-gen -mHKY -l 50000 -s .028 -p 50000 < treefile > seqfile
```

We then added a migration event between P1 and P3 at time 0.5, with varying migration rates, and calculated our D_{GEN} statistic on the resulting genomic data sets; results are shown

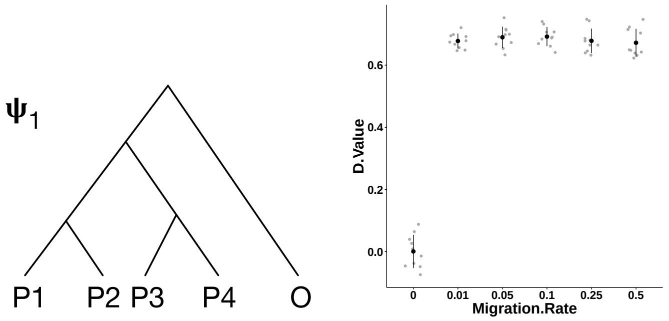

Figure 3 **5-taxon simulation results.** (Left) A 5-taxon species tree. The two most recent divergence events are set at 1.0 coalescent units, the divergence time of the ancestor of all in-group taxa (P1–P4) is set to 1.2 coalescent units, and the time of the root node is set to 1.5 coalescent units. A migration event between P1 and P3 at time 0.5 was added to species tree. (Right) Values of D_{GEN} on data sets with varying migration rates. Each point corresponds to a D_{GEN} value whose p-value was lower than 0.01 obtained from a different data set simulated under the same settings. The dark dots correspond to the mean and the lines correspond to 1 standard deviation around the mean.

in Fig. 3. As the results show, D_{GEN} performs very well at determining the presence of introgression in data sets. In particular, when the data evolved with no migration (migration rate 0), the D_{GEN} values hardly deviate from 0, and when the migration rate is non-zero, the method detects the presence of introgression with a strong deviation from 0. These results are consistent with the performance of D_{FOIL} [18].

Next, we considered cases beyond that of five taxa (i.e., cases not possible with either the D-statistic or D_{FOIL}). We conducted simulations that show the effect of migration rate and time of the migration event on the performance of D_{GEN}, as shown in Fig. 4.

As the results show, the D_{GEN} statistic performs very well at detecting introgression in this case as well. In particular, as the migration rate increases, so does the accuracy of the method. For a migration rate of 10^{-6} or higher, the method detects, with high significance, the presence of introgression. In the cases of extremely low migration rates (10^{-7} and 10^{-8}), the method tends to indicate slight deviation from a no-introgression scenario.

As for varying the time of the migration event, the accuracy of the method is what one would expect. As the time between the migration event and the divergence event increases (the time of the migration event decreases), the power to detect introgression is much higher. That power starts decreasing as the migration event becomes more ancient and, as a result, less signal is present for its detection.

3.1.1 Multiple Reticulations

The question we set out to investigate next is: Given that the D-statistic is designed to work under the assumption of a single gene flow event, how does it perform when there is more than one event? Fig. 5 shows a typical scenario for the D-Statistic, Scenario S1, in which the standard 4-taxon backbone tree has a single reticulation from P3 to P2.

The following four scenarios add an extra reticulation with a high migration rate. The effect of adding these reticulations on the value of the D-Statistic are shown in Fig. 6.

As expected, the S1 case yields the best results, followed by the S2 case with a weaker D value. All other statistic values demonstrate that even in the presence of a significant migration with introgression from P3 to P2, multiple introgressions can cause that information to be lost from inference. These results show that it is important to account for multiple

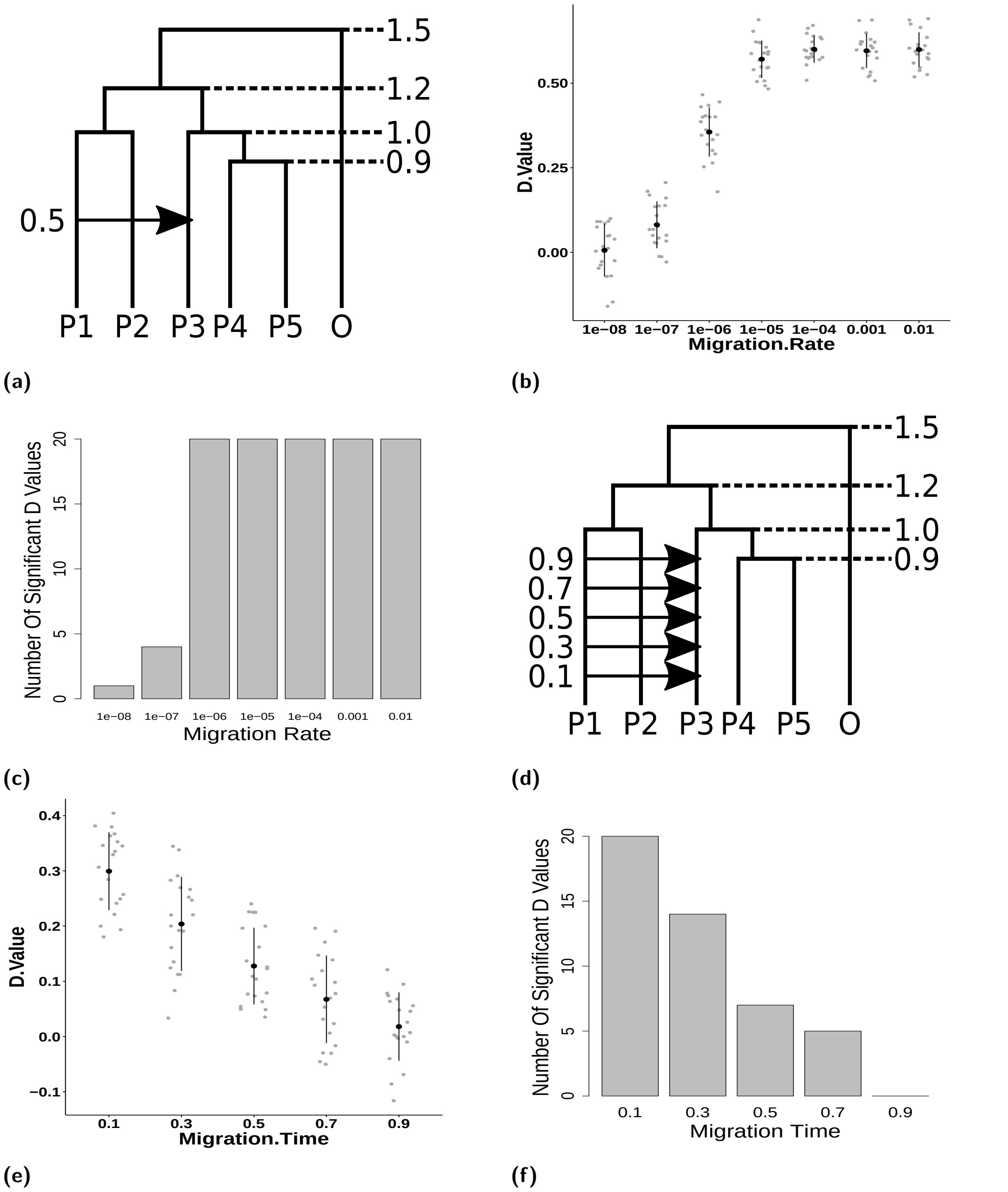

Figure 4 Simulation results on 6-taxon scenarios. (a) The network used for analyzing the effects of varying migration rate of a reticulation. The results of the corresponding *DG* values and the number of data sets where the D_{GEN} values were significant (p-value smaller than 0.01) are shown in (b) and (c), respectively. (d) The network used for analyzing the effects of varying the time of the migration event. The results of the corresponding *DG* values and the number of data sets where the D_{GEN} values were significant (p-value smaller than 0.01) are shown in (e) and (f), respectively.

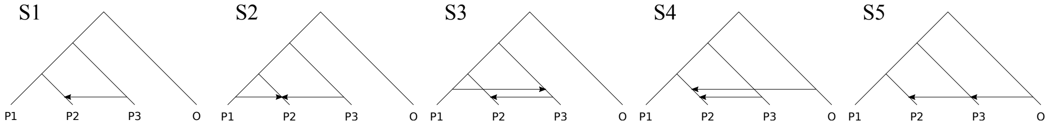

Figure 5 The standard D-Statistic scenario followed by four scenarios where an additional reticulation is added. S1 adds the first reticulation which is held constant throughout all scenarios and has M=0.1. The added reticulations in S2 through S5 have M=0.5.

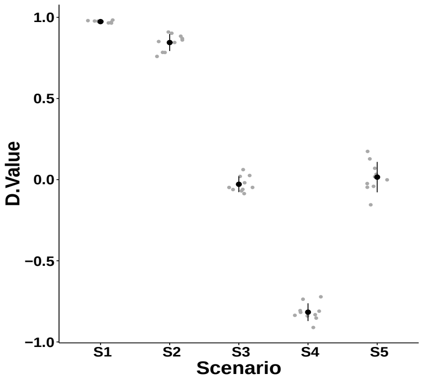

Figure 6 ***D*-Statistic values for scenarios with a "hidden" reticulation.** The scenarios S1–S5 are shown in Fig. 5.

reticulations simultaneously, which our D_{GEN} statistic allows for given that by its design it is not restricted to any specific number of reticulations.

3.1.2 D-Statistic Subsetting

When data sets with more than four taxa are to be analyzed by the D-statistic, a workaround is to subset the set of taxa into groups of four genomes (one outgroup and three in-group taxa) and conduct D-statistic analyses on each subset independently. Our method, being general, allows for analyzing the data set without any subsetting. The question we set out to investigate here is: Does subsetting and running the D-statistic on individual 4-taxon subsets equate to running D_{GEN} on the full data set? To answer this question, we considered the evolutionary scenario of Fig. 7.

In the case of the D-statistic, only two out of the ten simulations recovered a significant non-zero D value, whereas D_{GEN} inferred significant D values for all runs. This further demonstrates the need for and significance of a method that works directly on a full data set and accounting for multiple migration events.

3.2 Analysis Of a Mosquito Genomic Data Set

Finally, we present results from a real biological data set with six taxa. In both [8] and [26], the evolutionary history of the *Anopheles gambiae* species complex was found to be reticulate. Both studies found particularly strong signals of introgression in the 3L chromosome in an area known as the 3La inversion. This reticulation between *Anopheles quadriannulatus* (Q) and *Anopheles merus* (R) is shown in the network of Fig. 8(a) with the other species of *An. coluzzii* (C), *An. arabiensis* (A), *An. melas* (L), and *An. christyi* (O).

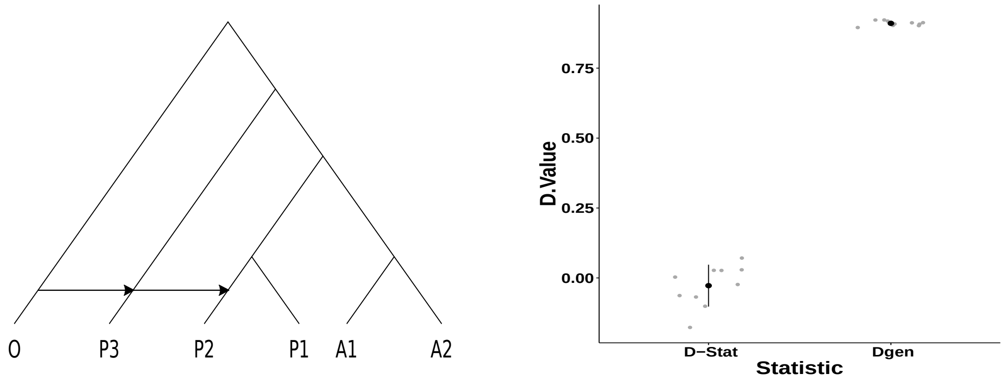

Figure 7 The effect of subsetting on the detectability of introgression. (Left) A 6-taxon evolutionary history with two migration events. (Right) The values of the D-statistic on subsets of four taxa, and D_{GEN} on the full data set.

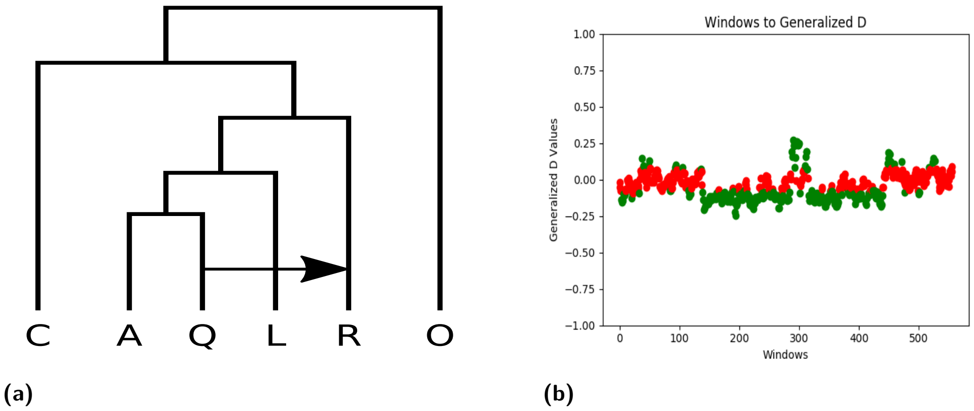

Figure 8 D_{GEN} values for the 3L mosquito chromosome. (a) Evolutionary history of six mosquito genomes. (b) D_{GEN} analysis of the 3L chromosome with window length set to 500kbp and with a 100kbp offset between windows.

The results from Fig. 8 (b) show that D_{GEN} does recover the introgressed region around the 3La inversion in comparison to the rest of the 3L chromosome, consistent with previous studies. Fig. 8(b) is an example of a figure that can be generated directly through the graphical user interface of the ALPHA toolkit [7] and is presented as generated directly from ALPHA. The figures output by ALPHA can be run on the full genome or on variable sized windows with variable sized offsets between windows. Here the window size used was 500kbp with a 100kbp offset between windows. The software can also vary the significance cutoff with which to display values as significant (green) or not significant (red). Here a significance cutoff value of 0.01 was used, as is used throughout the paper.

4 Discussion and Conclusions

In this paper, we extended the popular D-statistic to general cases of evolutionary histories of any number of taxa and any number and placement of migration events. What enabled this extension is the observation that the “ABBA-BABA” phylogenetic invariant underlying the D-statistic can be derived automatically by making use of the probability mass function of gene tree topologies under the multispecies coalescent and multispecies network coalescent models.

Our simulation results show that the new statistic D_{GEN} and method for deriving and computing it are very powerful for detecting introgression in various settings. In particular, we demonstrated that hidden migration events could negatively affect the performance of the D-statistic, which operates under the assumption of a single migration event. Furthermore, subsetting a data set of more than four taxa into data sets with four taxa is problematic. Our D_{GEN} statistic addresses these two issues by enabling the analysis of data sets with more than four taxa and more than a single migration event. While analyses in the style of the D-statistic make major assumptions, such as assuming the infinite sites model as well as ignoring dependence between sites, they are resilient to violations in these assumptions. Our results further support this given that our simulations violate both of these assumptions, having been performed under the full coalescent with recombination model with a mutation model allowing for recurrent mutation.

It is important to note that the D-statistic provides values that could be positive or negative. The sign of these values give an indication on the directionality of the migration in the case of four taxa. However, in the case of larger data sets, the sign of the D_{GEN} values is not easily interpretable in terms of directionality. It is also important to note that the actual D value is not the quantity of interest; rather, it is the statistical significance of its deviation from 0. There is of course, however, a strong correlation between the two.

As stated above, the method has been implemented in the ALPHA toolkit, which allows for conducting D_{GEN} analyses on the command-line as well as through a graphical user interface.

Finally, while we present the D_{GEN} statistic and its computation as a way of analyzing introgression under general evolutionary scenarios, computational complexity will become prohibitive for increasingly large, complex data sets. In particular, Step (3) in the algorithm above for computing D_{GEN} entails computing the probabilities of all gene tree topologies under a species tree and a phylogenetic network model. This calculation is very demanding, especially in the case of the phylogenetic network. For example, while generating D_{GEN} for four or five taxa takes approximately ten and forty seconds, respectively, generating it for six taxa takes thirteen minutes and for seven taxa thirty-eight hours. Fortunately, our implementation allows a D_{GEN} statistic to only ever need to be generated once for a particular evolutionary scenario, as the statistic itself is saved to a file that can be used on all current and future data sets for that scenario. This process of running a previously generated statistic on a new data set is, of course, computationally trivial. It will be important future work, however, to address the computational limits of D_{GEN} when going to arbitrarily large numbers of taxa.

References

1 M.L. Arnold. *Natural Hybridization and Evolution.* Oxford U. Press, 1997.

2 N.H. Barton. The role of hybridization in evolution. *Molecular Ecology*, 10(3):551–568, 2001.

3 P.D. Blischak, J. Chifman, A.D. Wolfe, and L.S. Kubatko. HyDe: a Python package for genome-scale hybridization detection. *Systematic Biology*, 2018.

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

5 J.H. Degnan and N.A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, 24(6):332–340, 2009.

6 Eric Y. Durand, Nick Patterson, David Reich, and Montgomery Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.

7 RA Leo Elworth, Chabrielle Allen, Travis Benedict, Peter Dulworth, and Luay Nakhleh. ALPHA: A toolkit for automated local phylogenomic analyses. *Bioinformatics*, 1:3, 2018.

8 Michael C Fontaine, James B Pease, Aaron Steele, Robert M Waterhouse, Daniel E Neafsey, Igor V Sharakhov, Xiaofang Jiang, Andrew B Hall, Flaminia Catteruccia, Evdoxia Kakani, Sara N. Mitchell, Yi-Chieh Wu, Hilary A. Smith, R. Rebecca Love, Mara K. Lawniczak, Michel A. Slotman, Scott J. Emrich, Matthew W. Hahn, and Nora J. Besansky. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524, 2015.

9 Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prafer, Matthias Meyer, Hern A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Hober, Barbara Hoffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Oeljko Kucan, Ivan Guic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Paabo. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, 2010.

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

11 Richard R Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

12 T. Jukes and C. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, NY, 1969.

13 Laura Kubatko and Julia Chifman. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *bioRxiv*, page 034348, 2015.

14 J. Mallet. Hybridization as an invasion of the genome. *TREE*, 20(5):229–237, 2005.

15 J. Mallet. Hybrid speciation. *Nature*, 446:279–283, 2007.

16 J. Mallet, N. Besansky, and M.W. Hahn. How reticulated are species? *BioEssays*, 38(2):140–149, 2016.

17 P. Pamilo and M. Nei. Relationship between gene trees and species trees. *Mol. Bio. Evol.*, 5:568–583, 1998.

18 James B Pease and Matthew W Hahn. Detection and polarization of introgression in a five-taxon phylogeny. *Systematic biology*, 64(4):651–662, 2015.

19 Fernando Racimo, Sriram Sankararaman, Rasmus Nielsen, and Emilia Huerta-Sánchez. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371, 2015.

20 Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13(3):235–238, 1997.

21 L. H. Rieseberg. Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, 28:359–389, 1997.

22 Loren H Rieseberg, Olivier Raymond, David M Rosenthal, Zhao Lai, Kevin Livingstone, Takuya Nakazato, Jennifer L Durphy, Andrea E Schwarzbach, Lisa A Donovan, and Chris-

tian Lexer. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, 301(5637):1211–1216, 2003.

23 Claudia Solís-Lemus and Cécile Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet*, 12(3):e1005896, 2016.

24 N. Takahata. Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.

25 Dingqiao Wen and Luay Nakhleh. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. *Systematic Biology*, 67(3):439–457, 2018.

26 Dingqiao Wen, Yun Yu, Matthew W Hahn, and Luay Nakhleh. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25(11):2361–2372, 2016.

27 Dingqiao Wen, Yun Yu, and Luay Nakhleh. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genetics*, 12(5):e1006006, 2016.

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

29 Y. Yu and L. Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16:S10, 2015.

30 Yun Yu, James H Degnan, and Luay Nakhleh. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet*, 8(4):e1002660, 2012.

31 Yun Yu, Jianrong Dong, Kevin J Liu, and Luay Nakhleh. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46):16448–16453, 2014.

32 Chi Zhang, Huw A Ogilvie, Alexei J Drummond, and Tanja Stadler. Bayesian inference of species networks from multilocus sequence data. *Molecular biology and evolution*, 35(2):504–517, 2018.

33 Jiafan Zhu and Luay Nakhleh. Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. *Bioinformatics*, 2018. (to appear).

34 Jiafan Zhu, Dingqiao Wen, Yun Yu, Heidi M Meudt, and Luay Nakhleh. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS Computational Biology*, 14(1):e1005932, 2018.