METHODOLOGY ARTICLE

Estimates of introgression as a function of pairwise distances

Bastian Pfeifer[1,3] and Durrell D Kapan[2*]

*Correspondence: dkapan@calacademy.org
[2]Department of Entomology and Center for Comparative Genomics, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Dr., San Francisco, USA
Full list of author information is available at the end of the article

Abstract

Background: Research over the last 10 years highlights the increasing importance of hybridization between species as a major force structuring the evolution of genomes and potentially providing raw material for adaptation by natural and/or sexual selection. Fueled by research in a few model systems where phenotypic hybrids are easily identified, research into hybridization and introgression (the flow of genes between species) has exploded with the advent of whole-genome sequencing and emerging methods to detect the signature of hybridization at the whole-genome or chromosome level. Amongst these are a general class of methods that utilize patterns of single-nucleotide polymorphisms (SNPs) across a tree as markers of hybridization. These methods have been applied to a variety of genomic systems ranging from butterflies to Neanderthal's to detect introgression, however, when employed at a fine genomic scale these methods do not perform well to quantify introgression in small sample windows.
Results: We introduce a novel method to detect introgression by combining two widely used statistics: pairwise nucleotide diversity d_{xy} and Patterson's D. The resulting statistic, the *Basic distance fraction* (Bd_f), accounts for genetic distance across possible topologies and is designed to simultaneously detect and quantify introgression. We also relate our new method to the recently published f_d and incorporate these statistics into the powerful genomics R-package PopGenome, freely available on CRAN. The supplemental material contains a wide range of simulation studies and a detailed manual how to perform the statistics within the PopGenome framework.
Conclusion: We present a new distance based statistic Bd_f that avoids the pitfalls of Patterson's D when applied to small genomic regions and accurately quantifies the fraction of introgression (f) for a wide range of simulation scenarios.

Keywords: genomics; introgression; hybridisation; SNPs

Background

Hybridization between species is increasingly recognized as a major evolutionary force. Although long known to occur in plants, evidence is mounting that it regularly occurs in many animal groups [1]. Generally thought to decrease differences between two species by sharing alleles across genomes, hybridization can paradoxically act as a ready source of variation, impacting adaptation [2, 3], aiding in evolutionary rescue [4], promoting range expansion [5], leading to species divergence [6, 7] and ultimately fueling adaptive radiation [8, 9]. The advent of whole genome sequencing has prompted the development of a number of methods to detect hybridization across the genome (recently summarized in Payseur and Rieseberg [10])

One class of methods involves quantifying single nucleotide polymorphism (SNP) patterns to detect hybridization between taxa. Here we focus on this class of tests involving four taxa. The most widely used of these, Patterson's D, was first introduced by Green *et al.* [11] and further developed by Durand *et al.* [12]. Patterson's D compares allele patterns of taxa with the Newick tree (((P1,P2),P3),O), to detect introgression between archaic taxon 3 (P3) and in-group taxon 1 (P1) or 2 (P2 or vice-versa). In brief, assuming the outgroup O is fixed for allele A, derived alleles (B) in P3, when shared with either P2 or P1, act as a marker of introgression leading to the following patterns: ABBA or BABA respectively. An excess of either pattern, ABBA or BABA represents a difference from the 50 : 50 ratio expected from incomplete lineage sorting and thus represents a signal that can be used to detect introgression.

Since its introduction, Patterson's D has been used for a wide range of studies to estimate the overall amount of hybrid ancestry by summing the ABBA or BABA pattern excess on a whole genome scale starting with studies of Neanderthals and archaic humans [11, 12]. In the past 7 years, Patterson's D has been increasingly used to localize regions of hybrid ancestry, not only in archaic humans [13] but also in species including butterflies, plants and snakes [14–16].

Currently, Patterson's D is frequently used in sliding window scans of different regions of the genome [17–19]. However, intensive evaluations of the four-taxon ABBA-BABA statistics [20] showed that this approach can lead to many false positives in regions of low recombination and divergence. One of the main reasons is the presence of mainly one of the two alternative topologies as a consequence of a lack of independence of the positions [15], resembling an introgression event, which is exacerbated when analyzing smaller gene-regions. To circumvent this issue, several strategies have been developed. On one side, more sophisticated non-parametric methods have been used to reduce the number of false positives (e.g., Patterson *et al.* [21]). On the other side, new statistics have been developed to better estimate the proportion introgression. Martin *et al.* [20] recently proposed the f_d estimate based on the f estimates originally developed by Green *et al.* [11] which measure the proportion of unidirectional introgression from P3 to P2. Specifically, f_d assumes that maximal introgression will lead to equally distributed derived allele frequencies in the donor and the recipient population and therefore utilizes the higher derived allele frequency at each variant site. This strategy aims to model a mixed population maximally affected by introgression. However, this approach has two major shortcomings: First, it is designed to sequentially consider introgression between the archaic population P3 and only one ingroup taxa (P1 or P2). Second, the accuracy of measuring the fraction of introgression strongly depends on the time of gene-flow.

Here we combine the approaches of the four-taxon tests with genetic distance to derive a statistic, the *basic distance fraction* (Bd_f), that estimates the proportion of introgression on a four-taxon tree which strictly ranges from -1 to 1, has symmetric solutions, can be applied to small genomic regions, and is less sensitive to variation in the time of gene-flow than f_d .

Methods

To derive Bd_f we took a two-fold approach. First, we reformulated Patterson's D, and f_d in terms of genetic distances based on the hypothesis that past or recent

hybridization will leave a signature of reduced d_{xy} between taxa [18,22]. Second, we account for non-introgressed histories by incorporating distances from species tree patterns into the denominator.

First, following convention, A and B denote ancestral and derived alleles respectively. Derived allele frequencies of the four taxa are $p_{1k} \dots p_{4k}$ at variant site k. Second, d_{xyk} is the average pairwise nucleotide diversity between population x and y at variant site k. Each genetic distance can be expressed as a sum of patterns in terms of ancestral and derived alleles allowing the terms ABBA and BABA to be rewritten in terms of genetic distances.

Patterson's D Statistic as a Function of Pairwise Distances

Here we derive the Patterson's D statistic as a function of pairwise genetic distance between taxon x and taxon y (d_{xy}). Following [23] the genetic distance d_{xy} is defined as

$$d_{xyk} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \pi_{ij}$$

at a given variant site k, where n_x is the number of individuals in population x n_y is the number of individuals in population y. Then at site k, $\pi_{ij} = 1 \vee 0$ is the boolean value indicating that the individual i of population x and the individual j of population y contain the same variant (0) or not (1). The genetic distances d_{xy} in terms of derived allele frequencies (p) are as follows:

$$\begin{aligned} d_{12k} &= p_{1k}(1 - p_{2k}) + (1 - p_{1k})p_{2k} \\ d_{13k} &= p_{1k}(1 - p_{3k}) + (1 - p_{1k})p_{3k} \\ d_{23k} &= p_{2k}(1 - p_{3k}) + (1 - p_{2k})p_{3k} \end{aligned}$$

Following [12, 21] instead of pattern counts, allele frequencies can be used as an unbiased estimator. According to that we define A as the ancestral allele frequency $(1 - p)$ and B as the derived allele frequency (p) allowing the terms

$$\begin{aligned} d_{12k} &= BAXA + ABXA \\ d_{13k} &= BXAA + AXBA \\ d_{23k} &= XBAA + XABA \end{aligned}$$

at site k. Here X is A+B = 1 and the position of the letter indicates the population order. The terms ABBA and BABA can then be expressed in terms of distances. If:

$$\begin{aligned} ABBA &= [(BBAA + ABBA) - (BBAA + BABA) + (BABA + ABBA)]/2 \\ BABA &= [(BBAA + BABA) - (BBAA + ABBA) + (BABA + ABBA)]/2 \end{aligned}$$

they can be expressed as:

$$ABBA = [p_{2k} \times d_{13k} - p_{1k} \times d_{23k} + p_{3k} \times d_{12k}] \times (1 - p_{4k})/2$$
$$BABA = [p_{1k} \times d_{23k} - p_{2k} \times d_{13k} + p_{3k} \times d_{12k}] \times (1 - p_{4k})/2$$

This leads to the following distance based Patterson's D equation for a region containing L variant positions:

$$D = \frac{\sum_{k=1}^{L} p_{2k} \times d_{13k} - p_{1k} \times d_{23k}}{\sum_{k=1}^{L} p_{3k} \times d_{12k}} \tag{1}$$

where d_{xyk} is the average pairwise nucleotide diversity between population x and y at variant position k; and p_{xk} the derived allele frequency in population x. In the context of distances $p_{2k} \times d_{13k}$ may be seen as the contribution of the variation contained between the lineages 1 to 3 (d_{13k}) to population 2.

Visualized by equation (1) the Patterson's D denominator (ABBA + BABA) simplifies to an expression of the derived allele frequency of the archaic population P3 times the average pairwise nucleotide diversity (d_{xy}) between population P1 and P2. This interpretation highlights the original difficulty that Patterson's D has handling regions of low diversity since the denominator will be systematically reduced in these areas due to the d_{12k} variable; increasing the overall D value. This effect intensifies when at the same time the divergence to the donor population P3 is high. Martin *et al.* [20] proposed f_d which corrects for this by considering the higher derived allele frequency (P2 or P3) at each given variant position; systematically increasing the denominator.

Martin's f_d Estimator

We can apply the same distance logic to rewrite the f_d statistic. Following the same logic as D we start with the definition of f_{hom} [11].

$$f_{hom} = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_3, P_3, O)}$$

where

$$S(P_1, P_2, P_3, O) = \sum_{k}^{L} p_{2k} \times d_{13k} - p_{1k} \times d_{23k}$$

Substituting P_2 with P_3,

$$S(P_1, P_3, P_3, O) = \sum_{k}^{L} p_{3k} \times d_{13k} - p_{1k} \times \pi_{3k}$$

where π_{3k} is the average pairwise nucleotide diversity within population P3 at site k. $p_{3k} \times d_{13k}$ may be interpreted as the contribution of population 3 to the variation contained between the lineages 1 to 3 (subtracting the contribution of population

1 contained in population 3). Here it is assumed that introgression goes from P3 to P2. Following Martin *et al.* [20] f_d is defined as $f_d = \frac{S(P_1,P_2,P_3,O)}{S(P_1,P_D,P_D,O)}$ where P_D is the population (2 or 3) with the highest frequency at each variant position. Here the denominator is:

$$S(P_1, P_D, P_D, O) = \sum_k^L p_{Dk} \times d_{1Dk} - p_{1k} \times d_{DDk} = \sum_k^L p_{Dk} \times d_{1Dk} - p_{1k} \times \pi_{Dk}$$

Leading to the statistic:

$$f_d = \frac{\sum_{k=1}^L p_{2k} \times d_{13k} - p_{1k} \times d_{23k}}{\sum_{k=1}^L p_{Dk} \times d_{1Dk} - p_{1k} \times \pi_{Dk}} \tag{2}$$

where in the denominator, π_{Dk} is the average nucleotide diversity within population P_D, which is the population with the higher derived allele frequency in population P_2 or P_3 for each variant site k. The difference between the f_d statistic versus f_{hom} is that there is no assumption in the former about the direction of introgression. These distance based interpretations suggest there exists a family of related distance estimators for the proportion of introgression. Here we propose a very simple version, we call Bd_f, that makes direct use of the distance based numerator of the Patterson's D statistic and relates the differences of distances to the total distance considered (fig. 1) by incorporating the BBAA species tree pattern into the denominator. The species tree pattern BBAA contributes to increased divergence between (P1,P2) and P3 in the absence of introgression. As a consequence within our Bd_f framework, we explicitly include the divergence to P3 on the four-taxon tree.

The Bd_f Estimator

In distance terms we may interpret the ABBA and BABA patterns as polarized shared distances (shared distance between two taxa caused by the derived alleles) on a 4-taxon tree. ABBA for example can be interpreted as the polarized shared distance between (P2,P3) and P1, where BABA is the polarized shared distance between (P1,P3) and P2. Thus, ABBA is a signal of shared increased distance to P1 and BABA is a signal of shared increased distance to P2. However, in order to relate those distances to the distances which are not a signal of introgression, the BBAA pattern must to be taken into account, because the species tree captures the third way in which two populations can share derived alleles. According to the interpretations given above, the BBAA species tree pattern can be seen as the polarized shared distances of (P1,P2) to P3. We propose to include this pattern to each introgression class (class 1: P1 is sharing with P3 and class 2: P2 is sharing with P3) in order to relate the distances to the total distances given within the shared distance system discussed here.

- For $P3 \leftrightarrow P2$ (P3 and P2 sharing); this simply is ABBA+BBAA.
- For $P3 \leftrightarrow P1$ (P3 and P1 sharing); this simply is BABA+BBAA.

A decreased BBAA polarized shared distance and an increased polarized shared distance ABBA is a signal of $P3 \leftrightarrow P2$ introgression. When at the same time the

BABA signal reduces we have a maximal support for the ABBA signal. Thus, the denominator of the Bd_f can be written as:

$$(ABBA) + BBAA) + (BABA + BBAA) = \sum_{k=1}^{L} p_{2k} \times d_{13k} + p_{1k} \times d_{23k}$$

For a given region including L variant sites.

Thus, the Bd_f statistic we propose here has the following form:

$$Bd_f = \frac{\sum_{k=1}^{L} p_{2k} \times d_{13k} - p_{1k} \times d_{23k}}{\sum_{k=1}^{L} p_{2k} \times d_{13k} + p_{1k} \times d_{23k}} \tag{3}$$

In distance terms, Bd_f may be interpreted as the difference of the distances from P1 and P2 to the archaic population P3 that is caused by introgression. The transformation of the denominator back into the basic Patterson's D statistic form suggests adding the given species tree BBAA pattern to the ABBA and BABA class respectively; which can be reasonably assumed to be the most likely pattern in the absence of introgression for a given species tree (((P1,P2),P3),O). With these patterns in hand it becomes possible to distinguish between signals of introgression and non-introgression. It should be noticed, however, that the Bd_f equation still produces some extreme false positives when e.g the derived allele frequency p_1 or p_2 is zero (often true when block-size is small). Thus, we encourage the user to apply *laplace smoothing* in genomic scan applications. In this case the derived allele frequency p is simply replaced by $p = (\sum_{k=1}^{n+2} \pi + 1)/(n + 2)$ for population 1 & 2 and d_{xy} is updated accordingly. The parameter π is a boolean variable and equals to 1 when a derived allele is present. We have implemented *laplace smoothing* for Bd_f as a feature in PopGenome.

Simulation study

To evaluate the performance of the Bd_f we used a simulation set-up following Martin *et al.* [20]. The Hudson's ms program [24] was used to generate the topologies with different levels of introgression and the seq-gen program [25] to generate the sequence alignments upon which to compare the performance of the three main statistics discussed in this paper, Patterson's D (D), f_d and Bd_f while varying the distance to ancestral populations, time of gene flow, recombination, ancestral population sizes and the effect of low variability. These simulations had the following settings in common: for each fraction of introgression $[0, 0.1, \ldots, 0.9, 1]$, we simulated 100 loci using 5kb windows to calculate three statistics: adjusted R^2 'goodness of fit', The euclidean distance (sum of squared distances) of the mean values to the real fraction of introgression, also called the 'sum of squares due to lack of fit' (SSLF) and the 'pure sum of squares error' (SSPE). The accuracy of the statistics is shown in fig. 2 and in the supplementary material (tables S1.1-S1.4) for a wide range of simulation parameters.

All of these analyses were done in the R-package PopGenome [26], that efficiently calculates Bd_f (and other statistics including f_d, $RNDmin$ [27], and the original Patterson's D) from the scale of individual loci to entire genomes.

Results

We performed extensive simulations varying the distance to ancestral populations, time of gene flow, recombination, ancestral population sizes and mutation rates. We found that Bd_f outperforms or is essentially equivalent to the f_d estimate to measure the real fraction of introgression for most of the studied ranges of simulation cases. Overall, because it captures natural variation in the denominator, Bd_f has slightly higher variances compared to f_d while the mean values are often the least biased as shown by the sum of squares due to lack of fit, yet it provides the best (or nearly equivalent) estimates to f_d as judged by the goodness of fit in almost all cases (supplementary information, section S1).

The effect of background history

Simulations under a variety of coalescent times show that Bd_f is the most accurate approximation of the real fraction of introgression, including under the different coalescent events simulated for both directions of introgression (fig. 2, table 1). Following behind Bd_f is f_d, which is more affected by differences in coalescent times. In this comparison, Patterson's D consistently overestimates the fraction of introgression (fig. 2, table 1). This known effect [20] is greatest in the most common case where the coalescent times differ between ingroup taxa (P1,P2) and the archaic taxon P3. This effect is also slightly impacted by the direction of introgression (fig. 2, table 1). However, for the more unrealistic case where the ingroup taxa (P1,P2) and the archaic taxon P3 are evolutionary very close it should be noticed that Bd_f essentially differs from the f_d estimate. In this specific case the 'pure sum of squares error' (SSPE) of Bd_f increases leading to a lower 'goodness of fit' value compared to f_d , while the 'sum of squares due to lack of fit' (SSLF) are still notably low signifying a very precise mean estimate of the real fraction of introgression. From Figure 2 we see that the Bd_f related SSPE values are high only and only if the signal of introgression is very low. So, we expect Bd_f to quantify stronger signals of introgression more precisely.

The effect of the time of gene-flow

One advantage of Bd_f compared to the other methods studied in this paper is that it is rarely affected by the time of gene-flow (fig. 3). This is due to the fact that, unlike f_d, Bd_f does not relate the signal of introgression to its maximum calculated from the present. When gene flow occurs in the distant past the denominator of f_d estimates increases leading to an underestimation of the fraction of introgression. The 'goodness of fit' of Bd_f is consistently higher than f_d (fig 3A), but more importantly, at the same time the SSLF values are almost unaffected by the time of gene-flow (fig. 3B). Notably, the direction of gene-flow has an effect that synergizes with the time that it occurred, with introgression between $P2 \rightarrow P3$ in the distant past overall showing lower values of the statistics.

The effect of recombination

We found that all three methods Bd_f, f_d and Patterson's D become more accurate with increasing recombination rates. This is due to the increase of independent sites of a region analyzed. While Bd_f tends to have higher variances when the recombination rate is low it's variance is comparable to f_d as soon as the recombination rate increases (supplementary table S1.2).

On the ability to detect introgression

To further test Bd_f, we evaluated the performance to detect introgression by simulating 10,000 neutral loci and 10 loci subject to introgression, interpreting the results using a receiver operating characteristic curve (ROC) analysis that evaluates the area under the curve (AUC) a measure that summarizes model performance, the ability to distinguish introgression from the neutral case, calculated with the R-package pROC [28]. For this simulation scenario Bd_f and the f_d estimate show nearly the same utility (higher is better) for the fraction of introgression and distance to ancestral population (supplementary information, section S2); but both, in agreement with Martin *et al.* [20], greatly outperform the Patterson's D statistic especially for smaller genomic regions. We also included the recently published $RNDmin$ method in this latter analysis; this alternative only gives good results when the signal of introgression is very strong (supplementary information, section S2). In addition, unlike f_d, Bd_f is able to quantify the proportion of admixture symmetrically ($P3 \leftrightarrow P2$ and $P3 \leftrightarrow P1$) it simplifies the analysis of real genomic data on a 4-taxon system.

Application

To ensure test with real data we calculated Bd_f for 50kb consecutive windows on the 3L arm of malaria vectors in the *Anopheles gambiae* species complex (fig. 4) confirming the recently detected region of introgression found in an inversion [17]. Notably, the Bd_f shows a weak signal of introgression for the majority of windows. As expected, some extreme false positive values of D, caused by the lack of information within a window, are assigned to reduced Bd_f values (fig. 4A). We used a weighted jackknife approach with a block size of 50kb to test for significance of the inverted region (15Mb-35Mb). For both, the Bd_f with a z value of -2.06 ($P = 0.0394$) and the Patterson's D with a z value of -3.87 ($P = 0.0001$) the null hypotheses can be rejected suggesting significant introgression. In addition, in order to detect chromosome-wide outliers we tested the null hypotheses (Bd_f, f_d and $D = 0$) on the basis of 50 kb consecutive windows using a normal distribution with the standard deviation calculated from the whole chromosome. The corresponding P values were corrected by multiple testing using the Benjamin-Hochberg false discovery rate (FDR) method [29] (see thresholds in Figure 4). Overall the majority of the inverted region shows a smoothed value that is below the the threshold of significance ($Bd_f \leq -0.125$, Figure 4). Furthermore, the largest absolute value for the 3L arm is $Bd_f = -0.98$ (at genomic region: 7,250,000 - 7,300,000 bp) strongly supporting introgression. The f_d value for that region is 1 and Patterson's D=-1. A second window at genomic position: 21,850,000-21,900,000 points to strong introgression. Here $Bd_f = -0.90$, $f_d = 0.85$ and $D = -1$. We further analyzed the SNP-specific Bd_f values in these two regions using a *drop-one* jackknife procedure: The first region contains 2 SNPs with Bd_f values of -0.98 and thus no z scores were calculated due to zero variance. The resulting z value for the second region is $z = -5.86$ ($P = 4.48^{-9}$) confirming strong introgression in that region. In this case the null hypotheses based on the 0.05 significance level can be rejected.

These analyses were all performed within the R package PopGenome [26] and can be easily reproduced with the code given in the supplementary material section S3.

Discussion

In the last 8 years there has been an explosion of population genomic methods to detect introgression. The Patterson's D method, based on patterns of alleles in a four-taxon comparison, has been widely applied to a variety of problems that differ from those for which it was originally developed. This statistic can be used to assess whether or not introgression is occurring at the whole genome scale, however, Patterson's D is best not applied to smaller genomic regions or gene-scans as noted by Martin *et al.* 2015.

The distance based approach proposed here has the following strengths: First, the distance approach points to a family of statistics that can directly identify changes in genetic distances that are a natural consequence of introgression. Second, distance measured by d_{xy} allows direct comparisons of quantities that are easily interpreted. Third, a simple member of this family based on these distances, Bd_f, accurately predicts the fraction of introgression over a wide-range of simulation parameters. Furthermore, the Bd_f statistic is symmetric (like Patterson's D) which makes it easy to implement and interpret. However, Bd_f also outperforms Patterson's D in all cases (the latter shows a strong positive bias) and Bd_f also outperforms or is equivalent to f_d in nearly all cases by showing both higher goodness of fit, a lower sum of squares due to lack of fit than f_d. Furthermore, unlike f_d, Bd_f does not vary strongly with the time of gene-flow.

There are several areas where further improvements could be made. Although the distance based derivation of all three statistics is sound, and Bd_f is empirically supported by simulation, further mathematical analysis for this general class of distance estimators is desired. Like other statistics under consideration in this paper, Bd_f depends on resolved species tree with a particular configuration of two closely related species, a third species and an outgroup, and therefore it is not directly applicable to other scenarios. In addition, both the f_d and Bd_f perform less accurately when measuring the proportion of admixture when the gene-flow occurs from P2 to P3. Moreover, our simulations revealed (Figure 5) the asymmetrical affect of gene-flow direction on genetic distance: gene-flow from P3 to P2 does not affect the distance between taxon 1 & 3 (d_{13}), however, the opposite it true when introgression from P2 to P3 occurs, the distance between taxon 1 & 2 (d_{12}) is not affected. This suggests comparisons of d_{xy} within given genomic regions may contain signal to infer the direction of introgression and therefore more accurately measure the proportion of admixture.

Overall, the distance based interpretation of introgression statistics suggests a general framework for estimation of the fraction of introgression on a known tree and may be extended in a few complementary directions including the use of model based approaches to aid in outlier identification and potentially model selection. The distance based framework introduced here may lead to other further improvements by measuring how genetic distance changes between different taxa as a function of hybridization across different parts of the genome.

Conclusion

Here we present both a simplified distance based interpretation for Patterson's D and Martin *et al.'s* f_d and a new distance based statistic Bd_f that avoids the pitfalls

of Patterson's D when applied to small genomic regions and is more accurate and less prone to vary with variation in the time of gene flow than f_d. We propose Bd_f as an estimate of introgression which can be used to simultaneously detect and quantify introgression. We implement Bd_f (as well as the other four-taxon statistics, f_d, and the original Patterson's D) in the powerful R-package, PopGenome [26], now updated to easily calculate these statistics for individual loci to entire genomes.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
BP and DDK designed the project. BP developed the methods and performed the simulations. BP and DDK wrote the manuscript.

Acknowledgements
We would like to thank Bettina Harr, Jim Henderson, Karl Lindberg, Paul Staab and Sebastian E. Ramos-Onsins for helpful discussions.

Author details
[1]Institute for Computer Science and Department of Biology, Heinrich Heine University, Düsseldorf, Germany. [2]Department of Entomology and Center for Comparative Genomics, Institute for Biodiversity Science and Sustainability, California Academy of Sciences, 55 Music Concourse Dr., San Francisco, USA. [3]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria.

References
1. Mallett, J.: Hybridization reveals the evolving genomic architecture of speciation. Trends in Ecology & Evolution **20**, 229–237 (2005)
2. Gilbert, L.E.: Adaptive novelty through introgression in heliconius wing patterns: evidence for a shared genetic "tool box" from synthetic hybrid zones and a theory of diversification. University of Chicago Press, 281–318 (2003)
3. Hedrick, P.W.: Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology **22**, 4606–4618 (2013)
4. Stelkens, R.B., *et.al*: Hybridization facilitates evolutionary rescue. Evolutionary Applications **7**, 1209 (2014)
5. Pfennig, K.S., Kelly, A.L., Pierce, A.A.: Hybridization as a facilitator of species range expansion. Proceedings of the Royal Society of London Series B **283** (2016)
6. Mallett, J.: Hybrid speciation. Nature **446**, 279–283 (2007)
7. Abbott, R., Albach, S., Arntzen, J.W., Baird, S.J.E., Bierne, N., *et. al*: Hybridization and speciation. Journal of Evolutionary Biology **26**, 229–246 (2013)
8. Seehausen, O.: Hybridization and adaptive radiation. Trends in Ecology & Evolution **16**, 198–207 (2004)
9. Meier, J.I., Marques, D.A., Mwaiko, S., et.al: Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nature Comm. **8**, 14363 (2017)
10. Payseur, B.A., Rieseberg, L.H.: A genomic perspective on hybridization and speciation. Molecular Ecology **25**, 2337–2360 (2016)
11. Green, R.E., Krause, J., Briggs, A.W., *et. al*: A draft sequence of the neandertal genome. Science **328**, 710–722 (2010)
12. Durand, E.Y., Patterson, N., Reich, D., M, S.: Testing for ancient admixture between closely related populations. Mol Biol Evol **28**, 2239–2252 (2011)
13. Racimo, F., Sankararaman, S., Nielsen, R., Huerta-Sanchez, E.: Evidence for archaic adaptive introgression in humans. Nature reviews Genetics **16**, 359–371 (2015)
14. Dasmahapatra, *et. al* (Heliconius Genome Consortium): Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature **487**, 94–98 (2012)
15. Eaton, D.A.R., Ree, R.H.: Inferring phylogeny and introgression using radseq data: An example from flowering plants (pedicularis: Orobanchaceae). Systematic Biology **62**, 689–706 (2013)
16. Zinenko, O., Sovic, M., Joger, U., Gibbs, H.L.: Hybrid origin of european vipers (vipera magnifica and vipera orlovi) from the caucasus determined using genomic scale dna markers. BMC Evolutionary Biology **16**, 76 (2016)
17. Fontaine, M.C., Pease, J.B., Steele, A., *et. al*: Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science **347**, 1258524 (2015)
18. Kronforst, M.R., Hansen, M.E.B., Crawford, N.G., *et. al*: Hybridization reveals the evolving genomic architecture of speciation. Cell Rep. **5**, 666–677 (2013)
19. Zhang, W., Dasmahapatra, K.K., Mallet, J., Moreira, G., Kronforst, M.R.: Genome-wide introgression among distantly related heliconius butterfly species. Genome Biology **17** (2016)
20. Martin, S.H., Davey, J.W., Jiggins, C.D.: Evaluating the use of abba-baba statistics to locate introgressed loci. Mol. Biol. Evol. **32**, 244–257 (2015)
21. Patterson, N., Moorjani, P., Luo, Y., et.al: Ancient admixture in human history. Genetics **192**, 1065–1093 (2012)
22. Smith, J., Kronforst, M.R.: Do heliconius butterfly species exchange mimicry alleles ? Biol Lett. **9**, 20130503 (2013)

23. Wakeley, J.: The variance of pairwise nucleotide differences in two populations with migration. THEORETICAL POPULATION BIOLOGY **49**, 39–57 (1996)
24. Hudson, R.R.: Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18**, 337–338 (2002)
25. Rambaut, A., Grass, N.: Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. Bioinformatics **13**, 235–238 (1997)
26. Pfeifer, B., Wittelsbuerger, U., Ramos-Onsins, S.E., Lercher, M.: Popgenome: an efficient swiss army knife for population genomic analyses in r. Mol. Biol. Evol. **31**, 1929–1936 (2014)
27. Rosenzweig, B.K., Pease, J.B., Besansky, N.J., Hahn, M.W.: Powerful methods for detecting introgressed regions from population genomic data. Molecular Ecology (2016)
28. Robin, X., *et. al*: proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics **12**, 77 (2011)
29. Benjamin, J., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol **57**, 289–300 (1995)

Figures

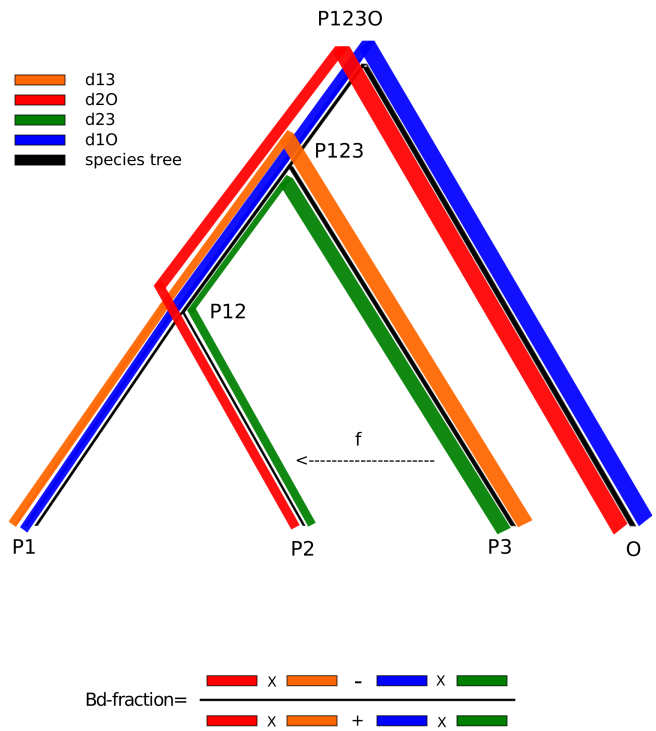

Figure 1 A graphical interpretation of the Bd_f model. A four-taxon tree illustrating the distance based Bd_f model in terms of connecetd path lengths differences. Here f is the fraction of introgression from P3 to P2; reducing the distance between P2 and P3 and at the same time increasing the derived allele frequency in P2. Bd_f approximates the measure f by relating the differences of the connected path lengths (red, orange) and (blue, green) to the overall sum of connected path lengths. When only sites are considered where the outgroup is mono-allelic the distance to the outgroup (d1O and d2O) simplifies to the derived allele frequencies p in population P1 and P2 (blue and red subtree).

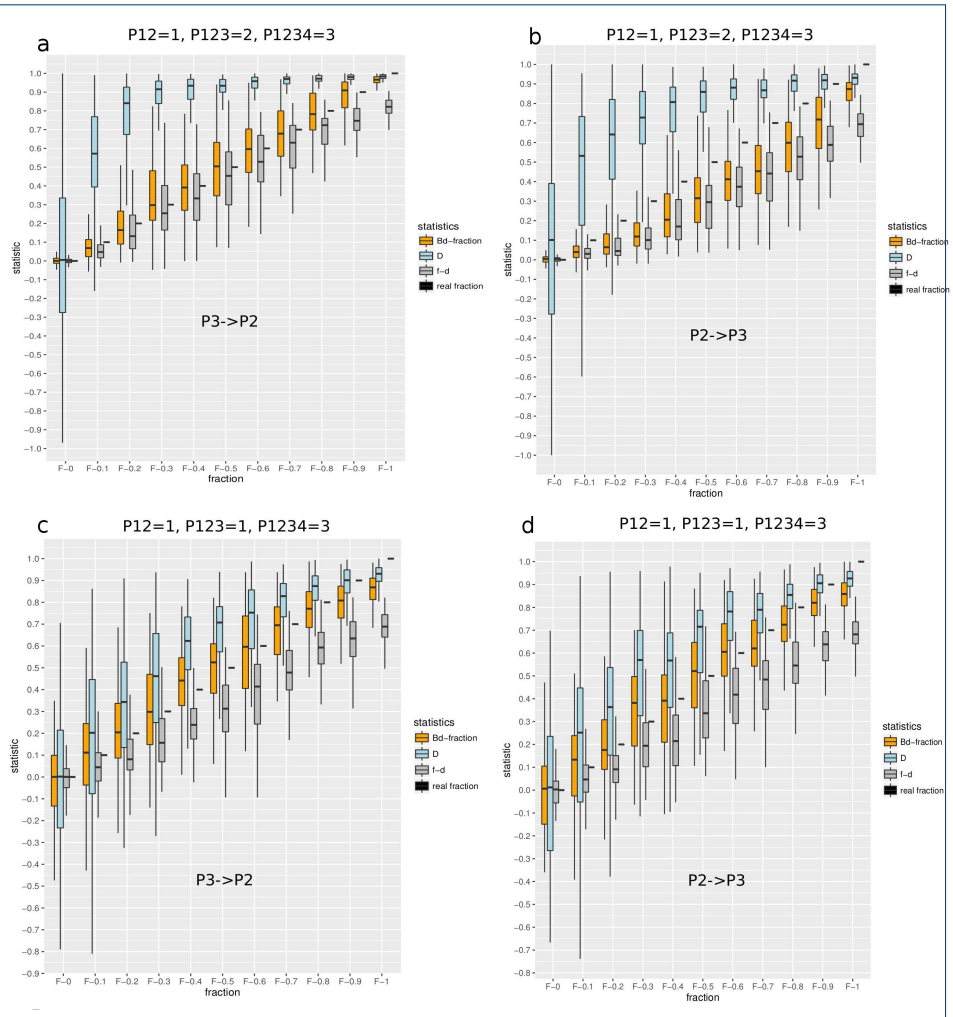

Figure 2 Accuracy of statistics to measure the fraction of introgression. The comparison of simulated data with a known fraction of introgression using ms versus the statistics (y-axis). We simulated 100 loci for every fraction of introgression $[0, 0.1, \ldots 0.9, 1]$ and plotted the distribution of the corresponding statistic outcomes. A window size of 5kb and a recombination rate of r=0.01 was used. The background histories (coalescent events) are **a:** P12=$1N_e$, P123=$2N_e$, P1234=$3N_e$ generations ago. **b:** P12=$1N_e$, P123=$2N_e$, P1234=$3N_e$ generations ago. **c:** P12=$1N_e$, P123=$1N_e$, P1234=$3N_e$ generations ago. **d:** P12=$1N_e$, P123=$1N_e$, P1234=$3N_e$ generations ago. Introgression directions are $P3 \rightarrow P2$ (A,C) and $P2 \rightarrow P3$ (B,D). Colors: f_d (grey), Bd_f (orange) Patterson's D (light blue) and the real fraction of introgression (black boxes). The boxplots shown here are created with the R-package ggplot2 (version 2.2.1).

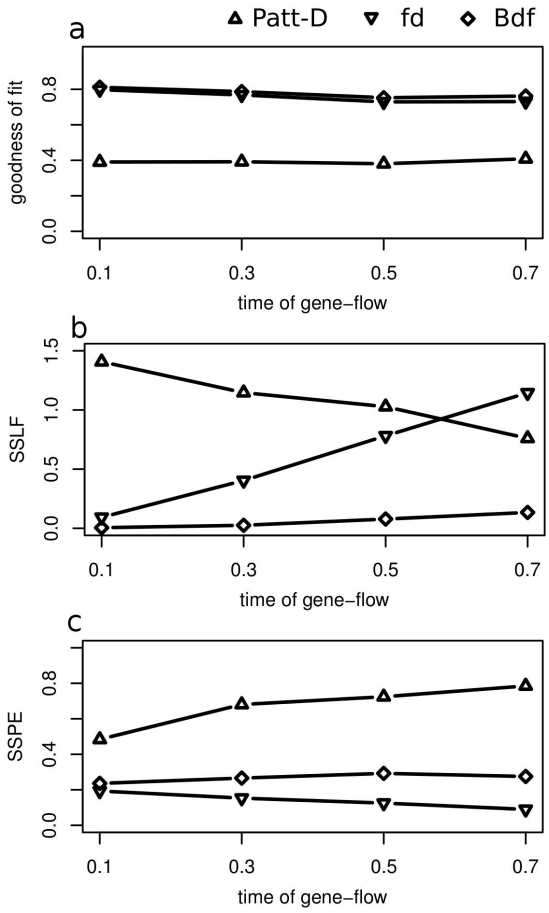

Figure 3 The effect of time of gene-flow. For $P3 \rightarrow P2$ introgression we varied the time of gene-flow (0.1, 0.3, 0.5, 0.7 N_e) and calculated for each statistic (D, f_d and Bd_f) **a:** the adjusted R^2 'goodness of fit'. **b:** SSLF 'sum of squares due to lack of fit' divided by the sample size n=100. **c:** SSPE 'pure sum of squares error'. Scaled recombination rate is $N_e r$=50 $(r = 0.01)$. The background history is: P12=1N_e, P123=2N_e and P1234=3N_e generations ago. The calls to ms are: $P3 \rightarrow P2$: ms 32 1 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1 -es Gene-flow 2 Fraction -ej Gene-flow 5 3 -r 50 5000. $P2 \rightarrow P3$: ms 32 1 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1 -es Gene-flow 3 Fraction -ej Gene-flow 5 2 -r 50 5000

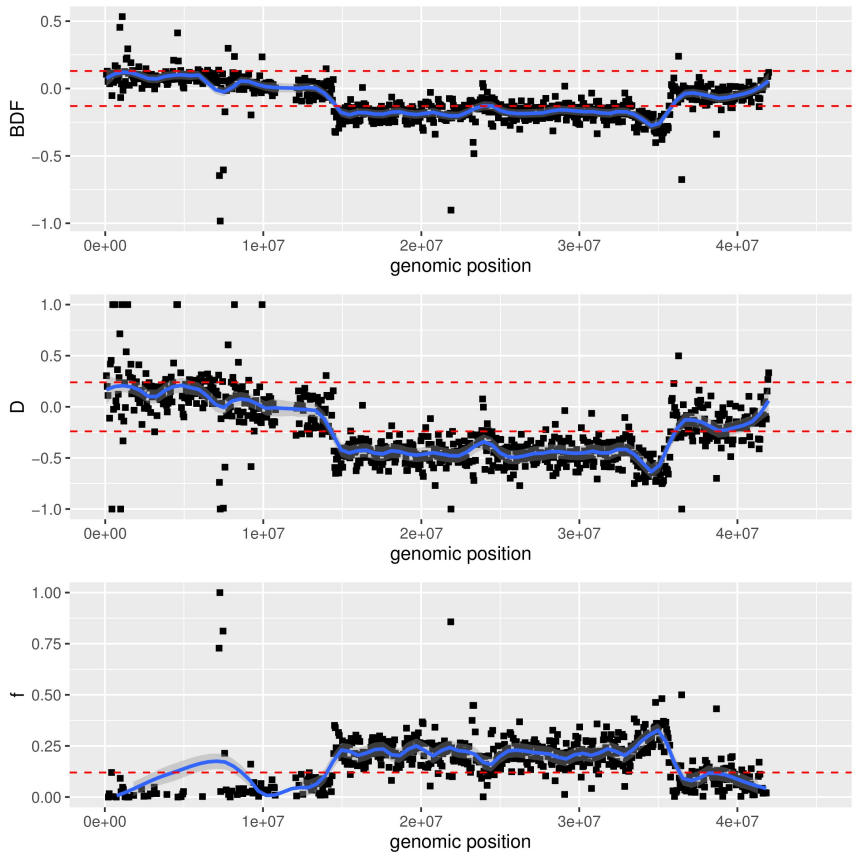

Figure 4 Anopheles gambiae 3La inversion. Confirming introgression on the 3L arm of the malaria vector Anopheles gambiae (Fontaine *et al.* 2015, fig. 4). We used the R-package PopGenome to scan the chromosome with 50kb consecutive windows and plotted **a:** the Bd_f values along the chromosome (*laplace smoothed*) **b:** Patterson's D along the chromosome. **c:** f_d is plotted along the chromosome with the y-axis ranging from 0 to 1. Bd_f eliminates some extreme false positive D values and suggest slightly lower signals of introgression on the whole scale. Bd_f detects the same outlier regions subject to strong introgression as f_d but it does this in a symmetric fashion. Red dashed horizontal line corresponds to FDR 5% significance level threshold.

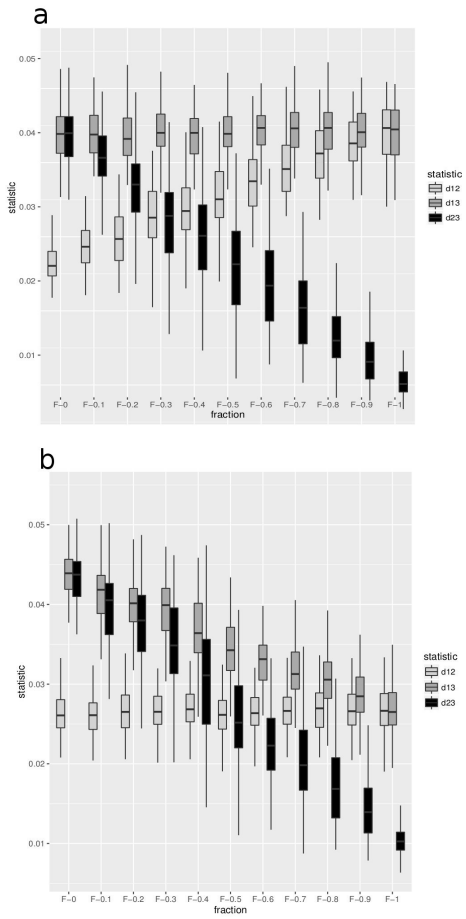

Figure 5 The effect of introgression on pairwise distances. The effect of the fraction of introgression on the average pairwise distance measurements d_{12}, d_{13} and d_{23}. **a:** The effect is shown for $P3 \rightarrow P2$ introgression. **b:** Shows the effect in case of $P2 \rightarrow P3$ introgression. The background history is: P12=$1N_e$, P123=$2N_e$ and P1234=$3N_e$ generations ago.

Tables

Table 1 The effect of the distance to ancestral population. This table refers to Figure 2 and displays some supporting values.

Direction of gene-flow	Distance to ancestral	D	f_d	Bd_f	
$P3 \rightarrow P2$	1-1-3	0.5846	0.7733	0.7014	[a]
		0.1177	0.4003	0.0380	[b]
		0.6033	0.1667	0.3540	[c]
$P3 \rightarrow P2$	1-2-3	0.3905	0.7978	0.8115	[a]
		1.4075	0.0928	0.0048	[b]
		0.4839	0.1931	0.2361	[c]
$P2 \rightarrow P3$	1-1-3	0.5745	0.7569	0.7040	[a]
		0.1211	0.4169	0.0473	[b]
		0.5876	0.1742	0.3340	[c]
$P2 \rightarrow P3$	1-2-3	0.3952	0.7778	0.7691	[a]
		0.6901	0.5376	0.3026	[b]
		0.4839	0.1931	0.1895	[c]

[a] the adjusted R^2 'goodness of fit'.
[b] SSLF 'sum of squares due to lack of fit' divided by the sample size n=100.
[c] SSPE 'pure sum of squares error'.