Estimates of Introgression as a Function of Pairwise Distances

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

[1] Institute for Computer Science, Heinrich Heine University, Düsseldorf, Germany
[2] Department of Entomology and Center for Comparative Genomics, Institute for Biodiversity and Sustainability Science, California Academy of Sciences, San Francisco, USA

[*]**Corresponding author**: E-mail:
bastian.pfeifer@uni-duesseldorf.de
dkapan@calacademy.org

Abstract

We introduce a new method to detect introgressed loci by combining two widely used statistics: d_{xy} (*the average pairwise nucleotide diversity between population x and y*) and the four-taxon Patterson's *D* statistic. The result is a statistic that takes into account genetic distance across possible topologies named the Bd_f and is designed to detect and at the same time to quantify introgression. We also relate this new method to the recently published f_d estimate and incorporate all statistics into the powerful genomics R-package PopGenome. The updated PopGenome version is freely available on CRAN. The supplement material contains a wide range of simulation studies and a detailed manual how to perform the statistics within the PopGenome framework.

Introduction

Hybridization between species is an important evolutionary force. Although it has been well known to occur in plants, it has only recently been recognized as regularly occurring among animals (Mallet 2005). Generally thought to decrease differences between two species by sharing alleles across genomes, hybridization can paradoxically act as a ready source of variation for adaptation (Gilbert 2003, Hedrick 2013), aiding in evolutionary rescue (Stelkens *et al.* 2014), promoting range expansion (Pfennig *et al.* 2016), potentially leading to species divergence (Mallet 2007, Abbott *et al.* 2013) and ultimately fuelling adaptive radiation (Seehausen 2004, Meier *et al.* 2017). 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 2016).

One class of methods involves inspecting single nucleotide polymorphism (SNP) patterns across a number of taxa to detect a signal of hybridization between taxa. Here we focus on this class of tests involving four taxa (Kulathinal *et al.* 2009). The most widely used of these, Patterson's *D*, was first introduced by Green *et al.* (2010) and further developed by Durand et al. (2011). Patterson's *D* compares allele patterns of taxa with the Newick tree (((*P1*,*P2*),*P3*),*O*), to detect introgression between archaic taxon 3 and in-group taxa 1 or 2 (or vice-versa). In brief, assuming the outgroup *O* is fixed for allele A, derived alleles (B) in taxon 3, when shared with either taxon 2 or taxon 1, 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 expected *50:50* ratio based on 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 amount of hybrid ancestry by summing the ABBA or BABA pattern excess on a whole genome scale (Green et al. 2010; Racimo et al. 2015). In the past 7 years it has been widely applied to a variety of problems from those for which it was originally developed, understanding Neanderthal and human introgression (Green *et al.* 2010, Durand *et al.* 2011) to introgression of adaptations in butterfly mimicry (Dasmahapatra *et al.* 2012), introgression in plants (Eaton and Ree 2013) to a large variety of organisms more recently (e.g. Zinenkno et al. 2016).

Currently, Patterson's *D* is frequently used in sliding window scans of different regions of the genome (Fontaine *et al.* 2015; Kronforst *et al.* 2013; Zhang *et al.* 2016). However, intensive evaluations of the four-taxon ABBA-BABA statistics (Martin *et al.* 2015) showed that this approach can lead to many false positives in regions of low divergence and of low recombination. 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 (Pease et al. 2015), 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. 2012). On the other side, new statistics have been developed to better estimate the proportion introgression. Martin et al. (2015) recently proposed the f_d estimate which is based on the *f* estimates originally developed by Green et al. (2010) which measure the proportion of unidirectional introgression from P3 to P2. The *f* estimates are generally designed to relate the observed introgression to the maximal possible introgression retrieved from the present while Patterson's *D* (and methods proposed here) approaches to the ratio of introgression vs. non-introgression. More specifically, the f_d assumes that maximal introgression will lead to equally distributed derived allele frequencies in the donor and the recipient population and therefore propose to take the higher derived allele frequency at each variant site independently. This strategy aims to model a mixed population maximally affected by bidirectional introgression. However, this method has two major shortcomings: First, it is designed to measure the introgressed material into one potential population only. Second, the accuracy of measuring the fraction of introgression strongly depends on the time of gene-flow. Recently, a distance based method named *RNDmin* was introduced (Rosenzweig 2016) to detect introgression on a whole genome scale. However, this method requires phased genotypes and is designed to detect but not to quantify introgression.

Here we combine the approaches (D, f_d, and distance) to present a statistic the *basic distance fraction* (Bd_f) to estimate the proportion of introgression on a four-taxon tree which strictly ranges from -1 to 1, has symmetric solutions, and is less sensitive towards the time of gene-flow than f_d and can be applied to small genomic regions. While the Patterson's D statistic is 'tree-free' (Patterson et al. 2012) the Bd_f should only be applied to data where the species tree is known in order to detect and at the same time to quantify introgression.

New Approaches

To derive Bd_f we took a two-fold approach. First, we reformulated the statistics (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 genetic distances (d_{xy}) between taxa (Kronforst et al. 2013, Smith and Kronforst 2013).

First, following convention, the ancestral allele is A and the derived allele B. The derived allele frequencies of the four taxa are $p_{1k} \dots p_{4k}$ at variant site k. Second, the average pairwise nucleotide diversity between population x and y at variant site k is d_{xyk} (see supplementary information, section S1.2). Each genetic distance can be expressed as a sum of patterns in terms of ancestral and derived alleles (e.g. d_{12k} = BAXA + ABXA, see supplementary information, section S1.2) allowing the terms ABBA and BABA to be rewritten in terms of genetic distances, for instance:

$$ABBA = [(BBAA + ABBA) - (BBAA + BABA) + (BABA + ABBA)]/2$$

which can be expressed as a function of allele frequencies and distances

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

(for all equations see supplementary information, section S1.2).

With ABBA and BABA as distances in hand, we can reformulate any statistic based on these counts. For instance this leads to the following distance based Patterson's D equation for a region containing L variant positions:

$$D = \sum_{k=1}^{L} (p_{2k} \times d_{13k} - p_{1k} \times d_{23k}) / \sum_{k=1}^{L} (p_{3k} \times d_{12k}) \qquad (1)$$

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. (2015) 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.

We can apply the same distance logic to rewrite the f_d statistic (see supplementary information, section S1.4) leading to:

$$f_d = \sum_{k=1}^{L}(p_{2k} \times d_{13k} - p_{1k} \times d_{23k}) / \sum_{k=1}^{L} p_{Dk} \times d_{1Dk} - p_{1k} \times \pi_D \qquad (2)$$

where in the denominator, π_D is the average nucleotide diversity within population P_D, which is the population with the higher derived allele frequency in population P2 or P3 for each variant site *k*. The distance based formulations of the *f* estimates can be found in the supplemental material (section S1.4).

These distance based interpretations suggests there exists a family of distance estimators of the proportion of introgression of varying complexity. 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 (supplementary information, section S1.5). The species tree pattern BBAA is the main source causing increased divergence between (P1,P2) and P3 in the absence of introgression. As a consequence within our Bd_f model, the divergence to P3 on the four-taxon tree will be explicitly included. The Bd_f statistic we propose here has the following form:

$$Bd_f = \sum_{k=1}^{L}(p_{2k} \times d_{13k} - p_{1k} \times d_{23k}) \ / \ \sum_{k=1}^{L}(p_{2k} \times d_{13k} + p_{1k} \times d_{23k}) \qquad (3)$$

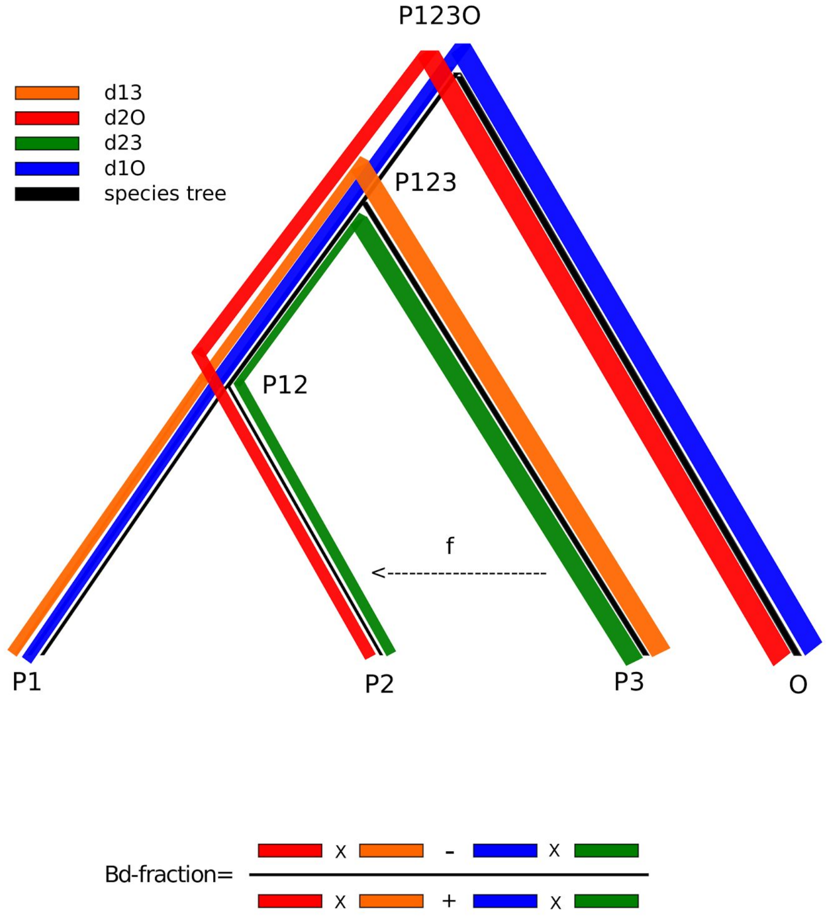

Fig. 1. A graphical interpretation of the *Bd$_f$* model. Here **f** *i*s the fraction of introgression from P3 to P2; reducing the distance between P2 and P3 and at the same time increasing the derived allele frequency in P2. *Bd$_f$* approximates the measure *f* by relating the differences of the connected path lengths (red, orange) and (blue, green) to the overall sum of connected path lengths. Note, when only sites are considered where the outgroup is mono-allelic the distance to the outgroup (d_{1O} and d_{2O}) simplifies to the derived allele frequencies p_x in population P1 and P2 (blue and red subtree).

In distance terms, *Bd$_f$* may be interpreted as the difference of the distances from P1 and P2 to the archaic population P3 which is caused by introgression. In section S2 of the supplementary information we also show mathematically that the *Bd$_f$* statistic belongs to the family of estimates which approaches to measure the real fraction of introgression.

Bd_f a Bayesian Model Selection Approach

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. Thus, we can transform the *Bd_f* motivation into a simple binomial model and use a Bayesian approach to compare models of introgression:

We are interested in comparing two models:
M_1 : Taxon 3 & 2 are sharing alleles due to introgression ($P3 \leftrightarrow P2$).
M_2 : Taxon 3 & 1 are sharing alleles due to introgression ($P3 \leftrightarrow P1$).

Both models can be explained by a binomial likelihood:

$$Pr(\,D|M_1) \,= \theta_1{}^{ABBA+BBAA}\,\theta_2{}^{BBAA} = \theta_1{}^{p_2\times\,d_{13}}\,\,\theta_2{}^{p_1p_2(1-p_3)}$$
$$Pr(\,D|M_2) \,= \theta_1{}^{BABA+BBAA}\,\theta_2{}^{BBAA} = \theta_1{}^{p_1\times\,d_{23}}\,\,\theta_2{}^{p_1p_2(1-p_3)}$$

where D is the data potentially influenced by introgression.

According to the *Bd_f* design *theta* (θ) includes information about the fraction of the data which is explained by ABBA and BABA relative to the species tree pattern BBAA. The BBAA pattern counts are used as an approximation to take the non-introgression signals into account. The model assumption is that the data *D* can be approximately explained by the BBAA species tree pattern plus the corresponding introgression pattern (ABBA or BABA). That is a sufficient assumption as we are interested in the relation of three different type of trees in the subspace of possible trees. The species tree (((P1,P2),P3),O) and the two introgression or *ILS* trees ((P1,(P2,P3)),O) and ((P2,(P1,P3)),O). It has been shown (Patterson et al. 2012, Durand et al. 2011) that instead of pattern counts, frequencies can be used as unbiased estimators and thus the pattern counts ABBA, BABA and BBAA can be simply replaced by the corresponding allele frequencies.

We use the conjugate Beta distribution as a prior:

$$Pr(M_1) = Beta(\overline{n}, \overline{n})$$
$$Pr(M_2) = Beta(\overline{n}, \overline{n})$$

where $\overline{n}$ is the average population size of P1, P2 and P3 included as a weighting factor to avoid unnecessary high uncertainty when only a few variant sites are available in a given region.

In order to form the posterior we propose the following update scheme; successive for each variant site k:

$$\alpha_1 = \sum_k^L p_{2k} \times d_{13k} \times \overline{n}$$
$$\alpha_2 = \sum_k^L p_{1k} \times d_{23k} \times \overline{n}$$
$$ß = \sum_k^L p_{1k} p_{2k} (1 - p_{3k}) \times \overline{n}$$

As a consequence the posterior distribution of each model M_1 and M_2 is:

$$Pr(M_1|D) = Beta(\overline{n} + \alpha_1, \overline{n} + ß)$$
$$Pr(M_2|D) = Beta(\overline{n} + \alpha_2, \overline{n} + ß)$$

The corresponding marginal log-likelihood can be calculated via the *gamma* function:

$$L(D|M_1) = log\ [\Gamma(\overline{n} + \alpha_1) \times \Gamma(\overline{n} + ß)] / \Gamma(\overline{n} + \alpha_1 + \overline{n} + ß)$$
$$L(D|M_2) = log\ [\Gamma(\overline{n} + \alpha_2) \times \Gamma(\overline{n} + ß)] / \Gamma(\overline{n} + \alpha_2 + \overline{n} + ß)$$

To compare the models via Bayes factors we propose the following transformation:

$$Bd_{bf} = 1 \qquad for\ Bd_f = 0$$
$$Bd_{bf} = 1 + exp(L(D|M_1) / L(D|M_2)) - exp(1) \qquad for\ Bd_f > 0$$
$$Bd_{bf} = 1 + exp(L(D|M_2) / L(D|M_1)) - exp(1) \qquad for\ Bd_f < 0$$

allowing researchers to judge the relative merit of the two competing models.

Simulation study

To evaluate the performance of the *Bd*$_f$ we used a simulation set-up following Martin et al. (2015). The Hudson's ms program (Hudson 2002) was used to generate the topologies with different levels of introgression and the *seq-gen* program (Rambaut and Grass 1997) 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, … ,0.9, 1], we simulated 100 loci using 5kb windows to calculate three statistics: adjusted R^2 'goodness of fit', The euclidian 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 S3.1-S3.4) for a wide range of simulation parameters.

Results and Discussion

Simulations under a variety of background histories 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). Following behind Bd_f is f_d, which is more affected by changes in coalescent times. In this comparison, Patterson's *D* consistently overestimates the fraction of introgression (fig. 2). This known effect (Martin et al. 2015) is greatest when the coalescent times differ between ingroup taxa (P1,P2) and archaic taxon P3. This effect is also slightly impacted by the direction of introgression (fig. 2).

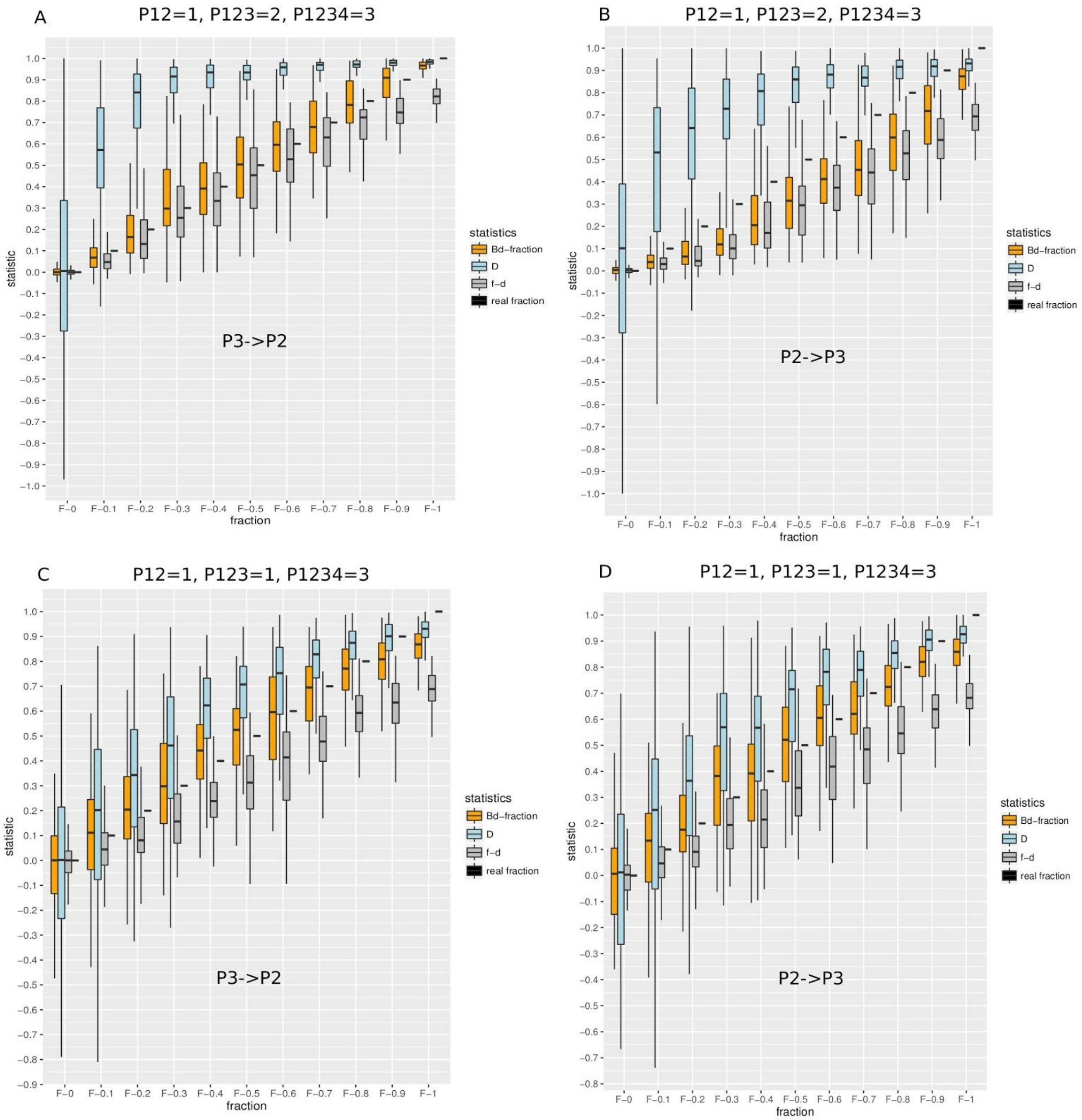

Fig. 2. Accuracy of statistics to measure the fraction of introgression. The comparison of simulated data with a known fraction of introgression using ms versus the statistics (y-axis). We simulated 100 loci for every fraction of introgression [0, 0.1, … 0.9, 1] and plotted the distribution of the corresponding 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→P2 (A,C) and P2→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).

We performed further simulations varying the distance to ancestral populations, time of gene flow, recombination, ancestral population sizes and the effect of low variability. We found that $Bd_{fraction}$ 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 (supplementary information, tables S3.1-S3.4).

Notably Bd_f is rarely affected by the time of gene-flow (table 1). This is due to the fact that, unlike f_d, Bd_f does not relate the introgression to its maximum calculated from the present. When gene flow occurs in the distant past the denominator of the f_d estimates increases leading to an underestimation of the fraction of introgression. Notably, the direction of gene-flow has an effect that synergizes with the time that it occurred with introgression between P2→P3 in the distant past showing lower values of the statistics overall.

Overall, 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.

Direction of gene-flow	Time of Gene-flow	*D*	f_d	Bd_f	
P3→P2	0.1	0.3905 1.407465 0.4838746	0.7978 0.0928434 0.1930699	0.8115 0.0048354 0.2361434	1 2 3
P3→P2	0.3	0.3918 1.146151 0.6805372	0.7681 0.4038538 0.1529574	0.787 0.0255170 0.2659381	1 2 3
P3→P2	0.5	0.3805 1.027749 0.7240085	0.7291 0.7815143 0.1250093	0.7525 0.0782484 0.2924232	1 2 3
P3→P2	0.7	0.4084 0.7600799 0.785031	0.7308 1.144712 0.0895343	0.762 0.1341295 0.2750616	1 2 3
P2→P3	0.1	0.3952 0.6901206 0.4838746	0.7778 0.5375956 0.1930699	0.7691 0.3026357 0.1894856	1 2 3
P2→P3	0.3	0.3702 0.4257417 0.8134292	0.6938 1.246362 0.1077545	0.7003 0.5154848 0.2378773	1 2 3

P2→P3	0.5	0.3069 0.2797494 0.9632419	0.5779 2.030634 0.0690975	0.5936 0.9802573 0.2355684	1 2 3
P2→P3	0.7	0.1639 0.4375968 1.350988	0.4283 2.732302 0.0431646	0.4617 1.708893 0.1957117	1 2 3

Table 1. Effect of the time of gene-flow. For each direction of 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) [1] the adjusted R^2 'goodness of fit'. [2] SSLF 'sum of squares due to lack of fit' divided by the sample size n=100. [3] 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→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→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*

To further test Bd_f, we evaluated the performance to detect introgression by simulating 10.000 neutral loci and 1.000 locus subject to introgression, interpreting the results using 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* (Robin et al. 2011). 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 S4); but both, in agreement with Martin et al. (2015), greatly outperform the Patterson's D statistic especially for smaller genomic regions. We also included the recently published *RNDmin* method (Rosenzweig et al. 2016) in this latter analysis, this alternative only gives good results when the signal of introgression is very strong (supplementary information, section S4).

To ensure good performance also on real data we calculated Bd_f for 50kb consecutive windows on the 3L arm of malaria vectors in the *Anopheles gambiae* species complex (fig. 3A) confirming the recently detected region of introgression (Fontaine et al. 2015). Figure 3C shows that some extreme negative Bd_f values are caused by the the lack of information in the window considered; reducing the Bayes factor compared to the corresponding Bd_f value. Notably, the Bd_{bf} values just show weak evidence ($Bd_{bf} < 3$) of introgression for the majority of windows. The maximum value for the *Anopheles gambiae* 3L arm is $Bd_{bf} = 7.15$ (at genomic region:21.850.000-21.900.000 bp).

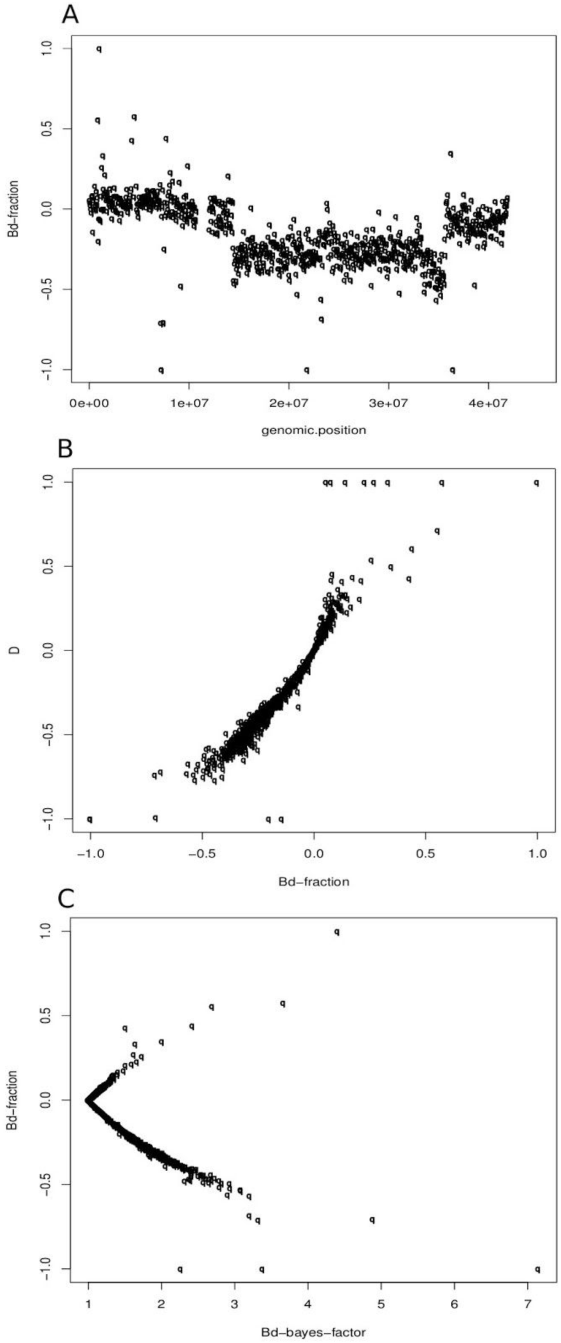

Fig. 3. *Anopheles gambiae* 3La inversion. Confirming introgression on the 3L arm of the malaria vector *Anopheles gambiae* (Fontaine *et al.* 2015, fig. 4). We used the R-package PopGenome to scan the chromosome with 50kb consecutive windows and plotted **A:** the Bd_f values along the chromosome **B:** Bd_f vs Patterson's *D*. Bd_f eliminates some extreme false positive *D* values and suggest slightly lower signals of introgression on the whole scale. **C:** A bivariate plot Bd_f vs Bd_{bf} for the same region, clearly showing the effects of included uncertainty introduced by the binomial Bayesian model as seen from some extreme negative Bd_f values having a reduced Bayes factor.

Strengths and Weaknesses of Approach

The distance based approach has the following strengths: First, the distance approach points to a family of statistics that can directly identify changes in genetic distances due to 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 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 fitting particular scenario therefore not directly applicable to other situations.

Overall, the distance based interpretation of introgression statistics suggests a general framework for estimation of the fraction of introgression on a known tree and can be extended using *Bayes factors* to aid in outlier identification and potentially model selection. The distance based framework introduced here could 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

In the last 8 years there has been an explosion of SNP based 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 not best applied to smaller genomic regions or *gene-scans* across entire regions. 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 provide all of these statistics, Bd_f , f_d, *RNDmin*, and the original Patterson's *D* in the powerful genomics R-package *PopGenome* and thus can easily be applied to individual loci, sets of loci and whole-genome sequencing data as well as subsites such as coding regions, genes and synonymous and non-synonymous sites.

Acknowledgments

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

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., et al. 2013. Hybridization and speciation. *Journal of Evolutionary Biology*, *26*, 229–246.

Dasmahapatra et al. (Heliconius Genome Consortium). 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature(London), 487(7405), 94–98.

Durand EY, 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). Systematic Biology, 62(5), 689–706.

Fontaine MC, Pease JB, Steele A, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science (New York, NY). 347(6217):1258524.

Gilbert LE. 2003. Adaptive novelty through introgression in Heliconius wing patterns: evidence for a shared genetic "tool box" from synthetic hybrid zones and a theory of diversification. In: Boggs CL, Watt WB, Ehrlich PR , editors. Ecology and evolution taking flight: Butterflies as model systems. Chicago: University of Chicago Press. pp. 281–318.

Green RE, Krause J, Briggs AW, et al. (56 co-authors). 2010. A draft sequence of the Neandertal genome. Science. 328(5979):710–722.

Hedrick, P. W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology, 22(18), 4606–4618.

Hudson RR 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

Kronforst MR, Hansen MEB, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, Kapan DD, Mullen SP. 2013. Hybridization reveals the evolving genomic architecture of speciation. Cell Rep. 5:666–677.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends in Ecology & Evolution, 20(5), 229–237.

Mallet, J. 2007. Hybrid speciation. Nature(London), 446(7133), 279–283.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. 32:244–257

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*, 14363.

Patterson N, 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.

Payseur, B. A., & Rieseberg, L. H. 2016. A Genomic Perspective on Hybridization and Speciation. Molecular Ecology, 25(11), 2337–2360.

Pfeifer B, Wittelsbuerger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31:1929–1936

Pfennig, K. S., Kelly, A. L., & Pierce, A. A. 2016. Hybridization as a facilitator of species range expansion. *Proceedings of the Royal Society of London Series B*, *283*(1839).

Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13:235–238.

R Core Team. 2013. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: http://www.R-project.org/.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. 2015. Evidence for archaic adaptive introgression in humans. Nature reviews Genetics.16(6):359-371.

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., & Hahn, M. W. 2016. Powerful methods for detecting introgressed regions from population genomic data. Molecular Ecology.

Seehausen O. 2004. Hybridization and adaptive radiation, Trends in Ecology & Evolution, 16(4):198-207.

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

Stelkens, R. B., Brockhurst, M. A., Hurst, G. D., & Greig, D. 2014. Hybridization facilitates evolutionary rescue. Evolutionary Applications, 7(10), 1209.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77.

Zhang W, Dasmahapatra KK, Mallet J, Moreira GRP, Kronforst MR. 2016. Genome-wide introgression among distantly related Heliconius butterfly species.Genome Biology.17:25.

Zinenko, O., Sovic, M., Joger, U., & Gibbs, H. L. 2016. Hybrid origin of European Vipers (Vipera magnifica and Vipera orlovi) from the Caucasus determined using genomic scale DNA markers. BMC Evolutionary Biology, 16, 76.

Supplementary Material

1 Material and Methods: Estimates of Introgression

1.1 Patterson's D statistic

Following (Durand et al. 2011) Patterson's *D* (*D*) is defined as:

$$D(P_1, P_2.P_3, O) = (\sum C_{ABBA}(k) - C_{BABA}(k)) / (\sum C_{ABBA}(k) + C_{BABA}(k))$$

Where the ABBA-BABA counts, at variant site *k*, are as follows:

$$C_{ABBA}(k) = (1 - p_{1k})p_{2k}p_{3k}(1 - p_{4k})$$
$$C_{ABBA}(k) = p_{1k}(1 - p_{2k})p_{3k}(1 - p_{4k})$$

Where for each taxon (P1, P2, P3 & P4):

$$p_{1k} = derived\ allele\ frequency\ of\ P_1$$
$$p_{2k} = derived\ allele\ frequency\ of\ P_2$$
$$p_{3k} = derived\ allele\ frequency\ of\ P_3$$
$$p_{4k} = derived\ allele\ frequency\ of\ O$$

1.2 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 (Wakeley 1996) 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 = the number of individuals in population x
n_y = the number of individuals in population y

Then at site *k*

π_{ij} = (1 or 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:

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

Following (Patterson *et al.* 2012, Durand *et al.* 2011) 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

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

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 be expressed in terms of distances.

If:

$$ABBA = \ [(BBAA \ + \ ABBA) \ - (BBAA \ + \ BABA) \ + (BABA \ + \ ABBA)]/2$$
$$BABA = \ [(BBAA \ + \ BABA) \ - (BBAA \ + \ ABBA) \ + (BABA \ + \ ABBA)]/2$$

they can be expressed as:

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

Thus, the Patterson’s *D* can be written as:

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

1.3 Martin’s f$_d$ estimator

Similar to Patterson’s *D,* Martin *et al.* (2015) derive the f_d estimator as follows.

$$S_1 \, (P_1, P_2.P_3, O) = \sum C_{ABBA}(k) \ - C_{BABA}(k)$$

$$S_2\,(P_1, P_D.P_D, O) = \sum C_{A max\{p_{2k},p_{3k}\} max\{p_{2k},p_{3k}\} A}(k)\ - C_{B(1-max\{p_{2k},p_{3k}\}) max\{p_{2k},p_{3k}\} A}(k)$$

$$f_d(P_1, P_2.P_3, O)\ = S_1\,(P_1, P_2.P_3, O)\,/\,S_2\,(P_1, P_D.P_D, O)$$

Where P_D is the maximum derived allele frequency of P3 and P2 at a given variant position k.

1.4 Martin's f_d estimator as a function of pairwise distances:

Following the same logic as *D* we start with the definition of f_{hom} (Green 2010).

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

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

Substituting P_2 by P_3,

$$S(P_1,P_3,P_3,O)\ = \sum_{k=1}^{L} p_{3k} \times d_{13k}\ -\ p_{1k} \times d_{33k}$$

$$S(P_1,P_3,P_3,O)\ = \sum_{k=1}^{L} p_{3k} \times d_{13k}\ -\ p_{1k} \times \pi_3.$$

where π_3 is the average pairwise nucleotide diversity within population P_3 .

$p_{3k} \times d_{13k}$ 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 P_3 to P_2.

Following Martin *et al.* (2015) 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

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

$$S(P_1,P_D,P_D,O)\ = \sum_{k=1}^{L} p_{Dk} \times d_{1Dk}\ -\ p_{1k} \times \pi_D.$$

Leading to the statistic:

$$f_d\ =\ \sum_{k=1}^{L} p_{2k} \times d_{13k}\ -\ p_{1k} \times d_{23k}\,/\sum_{k=1}^{L} p_{Dk} \times d_{1Dk}\ -\ p_{1k} \times \pi_D$$

The difference of f_d statistic versus f_{hom} is that there is no assumption in the former about the direction of introgression.

1.5 The *Bd*$_f$ estimator

In distance terms we may interpret the ABBA and BABA patterns as *polarized shared distances* based on the derived allele frequencies on a *4-taxon tree*. ABBA for example can be interpreted as the polarized shared distance between the (P2,P3) clade 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; the third way in which two populations can share the 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.

2 A Brief Analysis of the *Bd*$_f$

It has been shown (Green 2010) that the f_{hom} (supplementary section 1.4) approaches to the real fraction of introgression because the denominator provides an limit of the maximal possible introgression by assuming that maximal introgression is leading to complete homozygosity between *P2* and *P3*. However, Martin *et al.* (2015) stated that this estimate can produce f_{hom} values greater than one when e.g the derived allele frequency in population P2 is higher than in P3. Thus, the authors suggest the f_d as an alternative. Given these concerns about f_{hom} we can assume that the f_{hom} denominator is just a lower limit of the real maximal possible introgression.

Here we show that the denominator of Bd_f belongs to the family of measuring the maximal possible introgression and thus the Bd_f statistic approaches the real fraction of introgression.

The denominator of the f_{hom} estimate is:

$$S(P_1, P_3, P_3, O) = (1-p_1)p_3p_3 - p_1(1-p_3)p_3$$

We set $corr = p_1(1-p_3)p_3$, which can be seen as a correction factor which comes into play only and only if the archaic population P3 is heterozygous. In case of introgression between *P3* and *P2*, $(1-p_1)p_3p_3$ is the asymptotically more increasing function of the $S(P_1, P_3, P_3, O)$ denominator.

When we are able to show that under complete lineage sharing between *P2* and *P3* ($p_2 = p_3$) the following equation holds:

$$S(P_1, P_3, P_3, O) = denom(BD_{fraction}) - \alpha * corr$$

Where *alpha* is a scaling factor >1.

We than would have a validation that the denominator of Bd_f belongs to the family of statistics estimating the fraction of introgression:

First resolving the left side:

$$(1-p_1)p_3p_3 - p_1(1-p_3)p_3 = p_1(1-p_2)p_3 + (1-p_1)p_2p_3 + 2p_1p_2(1-p_3) - \alpha * corr$$
$$\Leftrightarrow$$
$$p_3[(1-p_1)p_3 - p_1(1-p_3)] = p_1(1-p_2)p_3 + (1-p_1)p_2p_3 + 2p_1p_2(1-p_3) - \alpha * corr$$
$$\Leftrightarrow$$
$$p_3[p_3 - p_1p_3 - p_1 + p_1p_3] = p_1(1-p_2)p_3 + (1-p_1)p_2p_3 + 2p_1p_2(1-p_3) - \alpha * corr$$
$$\Leftrightarrow$$
$$p_3[p_3 - p_1] = p_1(1-p_2)p_3 + (1-p_1)p_2p_3 + 2p_1p_2(1-p_3) - \alpha * corr$$
$$\Leftrightarrow$$
$$p_3^2 - p_1p_3 = p_1(1-p_2)p_3 + (1-p_1)p_2p_3 + 2p_1p_2(1-p_3) - \alpha * corr$$

Now resolving the right side:

$$p_3^2 - p_1p_3 = p_1p_3 - p_1p_2p_3 + p_2p_3 - p_1p_2p_3 + 2p_1p_2 - 2p_1p_2p_3 - \alpha * corr$$
$$\Leftrightarrow$$
$$p_3^2 - p_1p_3 = p_1p_3 + p_2p_3 + 2p_1p_2 - 4p_1p_2p_3 - \alpha * corr$$

Now substituting $p_2 = p_3$ (complete lineage sharing between P3 and P2).

$$p_3^{\,2} - p_1 p_3 \quad = p_1 p_3 + p_3^{\,2} + 2p_1 p_3 - 4p_1 p_3^{\,2} \; - \alpha * corr$$

$$\Leftrightarrow$$

$$p_3^{\,2} - p_1 p_3 \quad = 3p_1 p_3 + p_3^{\,2} - 4p_1 p_3^{\,2} \; - \alpha * corr$$

$$\Leftrightarrow$$

$$0 \quad = 4p_1 p_3 - 4p_1 p_3^{\,2} \; - \alpha * corr$$

$$\Leftrightarrow$$

$$0 \quad = 4[p_1 p_3 - p_1 p_3^{\,2}] \; - \alpha * corr$$

$$\Leftrightarrow$$

$$0 \quad = 4p_1 p_3(1 - p_3) \; - \; \alpha * p_1 p_3(1 - p_3)$$

$$\Leftrightarrow \; \alpha = 4$$

$$0 \; = 0$$

This shows that the correction factor used for the f_{hom} is just scaled by a factor *alpha* for the Bd_f denominator and thus the Bd_f statistic approaches to measure the real fraction of introgression.

3 Simulation Results: On the Accuracy to Measure the Real Fraction of Introgression

Distance of ancestral population: Starting with the following topology (((P1,P2),P3),P4) we simulate varying depths to common ancestors of P1 & P2, and at the root (P1234), where recombination rate is fixed with $N_e r$ =50 (r=0.01) for each direction of gene flow (see supplementary table 1 below).

Direction of gene-flow	Distance to ancestral population (P12-P123-P1234)	*D*	f_d	Bd_f	
P3→P2	0.3-1-3	0.4388 0.7060372 0.5175175	0.7662 0.2892631 0.166674	0.7739 0.1883595 0.2209828	1 2 3
P3→P2	0.5-1-3	0.4916 0.4894583 0.5618507	0.7622 0.32828 0.1675572	0.7837 0.1277751 0.2180397	1 2 3
P3→P2	0.7-1-3	0.5158 0.2927739 0.6544165	0.7675 0.366725 0.1694591	0.7698 0.0733439 0.2486593	1 2 3
P3→P2	1-1-3	0.5846	0.7733	0.7014	1

		0.1177175 0.6033176	0.4002708 0.1667211	0.0379529 0.3540125	2 3
P3→P2	0.5-2-3	0.3314 1.622079 0.602628	0.7901 0.0814349 0.2004631	0.8064 0.0797936 0.2252406	1 2 3
P3→P2	0.5-3-3	0.2704 1.973632 0.612887	0.8048 0.0298935 0.2086517	0.8077 0.0884342 0.2254532	1 2 3
P3→P2	1-2-3	0.3905 1.407465 0.4838746	0.7978 0.0928434 0.1930699	0.8115 0.0048354 0.2361434	1 2 3
P3→P2	1-3-3	0.3267 1.702683 0.6095603	0.8142 0.0342730 0.2039458	0.8212 0.0268799 0.2297386	1 2 3
P3→P2	2-2-3	0.5858 0.4688205 0.4950339	0.8115 0.1919717 0.1760974	0.7254 0.1339255 0.3274011	1 2 3
P2→P3	0.3-1-3	0.1593 0.8881787 0.7641099	0.4724 1.837565 0.1412364	0.432 2.205826 0.1071995	1 2 3
P2→P3	0.5-1-3	0.366 0.2165403 0.7531743	0.6895 1.080729 0.1313857	0.6465 1.044189 0.166181	1 2 3
P2→P3	0.7-1-3	0.4925 0.1415166 0.631658	0.7283 0.6810742 0.1551363	0.7001 0.3466498 0.2417637	1 2 3
P2→P3	1-1-3	0.5745 0.1211462 0.5876195	0.7569 0.4168669 0.1742366	0.704 0.0472772 0.333956	1 2 3
P2→P3	0.5-2-3	0.2525 0.3320516 0.8885661	0.6755 1.225495 0.1285244	0.6183 1.460954 0.1333156	1 2 3
P2→P3	0.5-3-3	0.2361 0.37603 0.9413875	0.6709 1.321473 0.1249192	0.5852 1.690323 0.1160202	1 2 3

P2→P3	1-2-3	0.3952 0.6901206 0.4838746	0.7778 0.5375956 0.1930699	0.7691 0.3026357 0.1894856	1 2 3
P2→P3	1-3-3	0.3767 0.8946288 0.6853883	0.7589 0.6210276 0.1569502	0.7226 0.6119493 0.192384	1 2 3
P2→P3	2-2-3	0.5628 0.4821917 0.5399023	0.802 0.1947744 0.1839213	0.7188 0.1419037 0.3309354	1 2 3

Supplementary Table 1: Distance of ancestral population: For each direction of gene flow and distance to ancestral populations (see above) we calculated for each statistic (*D*, f_d and Bd_f) [1] the adjusted R^2 'goodness of fit'. [2] *SSLF* 'sum of squares due to lack of fit' divided by the sample size n=100. [3] SSPE 'pure sum of squares error'. The time of gene-flow was a constant at $0.1N_e$, the scaled recombination rate is $N_e r$=50 (*r*=0.01), and the calls to ms are as follows:
P3→P2: ms 32 1 -I 4 8 8 8 8 -ej P12 2 1 -ej P123 3 1 -ej P1234 4 1 -es 0.1 2 Fraction -ej 0.1 5 3 -r 50 5000
P2→P3: ms 32 1 -I 4 8 8 8 8 -ej P12 2 1 -ej P123 3 1 -ej P1234 4 1 -es 0.1 3 Fraction -ej 0.1 5 2 -r 50 5000

Recombination: To test the impact of recombination on these statistics we varied the recombination rates from (*r* = 0 - .08). With increasing recombination rates the accuracy to measure the real fraction of introgression increases for f_d and Bd_f while the Patterson's *D* is rarely affected by varying this parameter.

Direction of gene-flow	Recombination rate	*D*	f_d	Bd_f	
P3→P2	0/5000	0.3307 0.7697069 1.488486	0.6335 0.1065987 0.4484599	0.6062 0.0139072 0.6535393	1 2 3
P3→P2	50/5000	0.3905 1.407465 0.4838746	0.7978 0.0928434 0.1930699	0.8115 0.0048354 0.2361434	1 2 3
P3→P2	100/5000	0.4072 1.490496 0.3725337	0.8676 0.1114164 0.1155275	0.8758 0.0080236 0.1467002	1 2 3
P3→P2	200/5000	0.4046 1.511661 0.3639124	0.8986 0.1163041 0.0864206	0.902 0.0121368 0.1096251	1 2 3

P3→P2	300/5000	0.4069 1.564095 0.3190911	0.9249 0.1209162 0.0624568	0.9257 0.0150437 0.0819618	1 2 3
P3→P2	400/5000	0.4041 1.577837 0.3185764	0.9383 0.1222677 0.0509774	0.936 0.0169133 0.0691337	1 2 3

Supplementary Table 2: Recombination: For recombination rates varying from (0 to .08) we calculated for each statistic (*D*, f_d and Bd$_f$) [1] the adjusted R^2 'goodness of fit'. [2] *SSLF* 'sum of squares due to lack of fit' divided by the sample size n=100 . [3] SSPE 'pure sum of squares error'. The time of gene-flow was a constant at 0.1N_e, background history is P12=1N_e, P123=2N_e and P1234=3N_e, and the calls to ms are as follows:

P3→P2: ms 32 1 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1 -es 0.1 2 ***Fraction*** *-ej 0.1 5 3 -r* ***Ne*r*** *5000* only one direction of gene-flow shown.

Ancestral population sizes: We varied ancestral population sizes at the nodes P12 and P123 and simulated the impact on the (*D*, f_d and Bd_f) statistics (see supplementary table 3 below).

Direction of gene-flow	Ancestral population size P12-P123	*D*	f_d	Bd_f	
P3→P2	1-2	0.4133 1.147567 0.6070287	0.8275 0.0668421 0.1808706	0.8342 0.0304105 0.2064996	1 2 3
P3→P2	1-10	0.4606 1.047664 0.4794316	0.8031 0.0084466 0.2410548	0.8062 0.100913 0.2165126	1 2 3
P3→P2	2-10	0.5955 0.5244162 0.4275634	0.795 0.0094751 0.2554508	0.8022 0.0473944 0.2420035	1 2 3
P3→P2	2-1	0.4857 0.9736148 0.474269	0.8177 0.1291925 0.1743487	0.8237 0.0036228 0.2196956	1 2 3
P3→P2	10-1	0.579 0.61373 0.4147273	0.8 0.1866553 0.1840674	0.7677 0.0649301 0.2693273	1 2 3
P3→P2	10-2	0.6208 0.406646 0.4288073	0.7957 0.1198737 0.2109028	0.7778 0.0146505 0.2804412	1 2

					3

Supplementary Table 3: Ancestral population sizes: For different ancestral population sizes (multiples of 1, 2 and 10N_e) at nodes P12 and P123 we calculated for each statistic (D, f_d and Bd_f) and present ¹ the adjusted R^2 'goodness of fit'. ² *SSLF* 'sum of squares due to lack of fit' divided by the sample size n=100. ³ SSPE 'pure sum of squares error'. The time of gene-flow is 0.1N_e, scaled recombination rate is $N_e r$=50 (r=0.01), and background history is: P12=1N_e, P123=2N_e and P1234=3N_e, and the calls to *ms* are:

P3→P2: ms 32 1 -I 4 8 8 8 8 -ej 1 2 1 -en 1.01 1 ***P12*** *-ej 2 3 1 -en 2.01 1* ***P123*** *-ej 3 4 1 -es 0.1 2* ***Fraction*** *-ej 0.1 5 3 -r 50 5000*

The effect of low variability: We varied the nucleotide diversity θ to test the effect of low variability on the statistics D, f_d and Bd_f.

Direction of gene-flow	Variability theta (θ)	*D*	f_d	Bd_f	
P3→P2	3/5000	0.396 1.308409 0.5742832	0.8153 0.1108171 0.1771698	0.8197 0.0129878 0.2252038	1 2 3
P3→P2	5/5000	0.4015 1.323871 0.5361765	0.8083 0.121785 0.1832012	0.8205 0.0191775 0.2206314	1 2 3
P3→P2	25/5000	0.4035 1.201857 0.6192605	0.7991 0.1310123 0.1922683	0.8102 0.0228752 0.2379141	1 2 3
P3→P2	50/5000	0.4092 1.357399 0.4953229	0.8068 0.1009227 0.1866151	0.8217 0.006412517 0.2276914	1 2 3

Supplementary Table 4 The effect of low variability: For unidirectional gene-flow P3→P2, we varied θ from 3/5000 - 50/5000 and calculated for each statistic (D, f_d and Bd$_f$) and present ¹ the adjusted R^2 'goodness of fit'. ² *SSLF* 'sum of squares due to lack of fit' divided by the sample size n=100. ³ SSPE 'pure sum of squares error'. The time of gene-flow was a constant at 0.1N_e, the scaled recombination rate is N_e*r=50 (r=0.01) and background history is P12=1N_e, P123=2N_e and P1234=3N_e. The calls to *ms* are:

P3→P2: ms 32 1 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1 -es 0.1 2 ***Fraction*** *-ej 0.1 5 3 -r 50 5000 -t* ***variability***

4 Simulation Results: Detecting Introgression from Whole Genome Data

To test the performance of the various statistics (*D, f_d, RNDmin, Bd_f*) to distinguish neutral models from models with varying levels of introgression or varying distances to ancestral populations, we performed two simulations on 1kb windows. For each statistic of interest we present the area under the curve (AUC) values. All simulations start with 10.000 loci under the neutral scenario (*f=0*) and 1.000 locus with subject to introgression. The Recombination rate is fixed at *r=0.01*.

<u>Fraction of introgression:</u> To test the impact of varying the fraction of introgression we simulated the fraction of introgression for the 'alternative model' from *f*=0.1 to *f*=1 and compared this to the neutral scenario where the fraction of introgression is zero (*f=0*). See below (supplementary table 5).

Direction of gene-flow	Fraction of introgression	*D*	f_d	*RNDmin*	Bd_f
P3→P2	0.1	0.6252	0.7128	0.5579	0.7065
P3→P2	0.2	0.6846	0.8426	0.6043	0.833
P3→P2	0.3	0.7163	0.9221	0.6479	0.9139
P3→P2	0.4	0.7293	0.9541	0.7196	0.9478
P3→P2	0.5	0.738	0.981	0.7588	0.9753
P3→P2	0.6	0.7466	0.9922	0.8246	0.989
P3→P2	0.7	0.7585	0.9979	0.851	0.9961
P3→P2	0.8	0.7607	0.9988	0.9216	0.998
P3→P2	0.9	0.7748	1	0.9659	0.9996
P3→P2	1	0.7871	1	1	0.9998

Supplementary Table 5: <u>Fraction of introgression:</u> The effect of varying fractions of introgression on the model utility in the ROC analysis as indicated for values of AUC. The background history (coalescent times) is: *P12=1N_e*, *P123=2N_e* and *P1234=3N_e* generations ago. The calls to *ms* are:
Neutral model: ms 32 1 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1 -r 10 1000
Alternative model: P3→P2: ms 32 1 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1 -es ***0.1*** *2* ***Fraction*** *-ej* ***0.1*** *5 3 -r 10 1000*

Distance of ancestral population:

Finally, by varying the distance to ancestral populations we tested the impact of a low amount of introgression on the various statistics.

Direction of gene-flow	Distance to ancestral population (P12-P123-P1234)	*D*	f_d	*RNDmin*	Bd_f
P3→P2	0.3-1-3	0.6366	0.686	0.5308	0.6782
P3→P2	1-1-3	0.5884	0.6077	0.5107	0.5969
P3→P2	0.5-2-3	0.6314	0.7292	0.5489	0.7272
P3→P2	0.5-3-3	0.6454	0.7493	0.5498	0.7488
P3→P2	1-2-3	0.6252	0.7128	0.5482	0.7065
P3→P2	1-3-3	0.6465	0.7604	0.569	0.7575
P3→P2	2-2-3	0.6016	0.6406	0.5312	0.622
P3→P2	1.5-2-3	0.6065	0.668	0.5526	0.6573
P2→P3	0.3-1-3	0.5221	0.5515	0.8464	0.5464
P2→P3	1-1-3	0.541	0.6126	0.9361	0.6157
P2→P3	0.5-2-3	0.5652	0.6197	0.8717	0.5998
P2→P3	0.5-3-3	0.5842	0.6195	0.9653	0.5888
P2→P3	1-2-3	0.5919	0.6615	0.5595	0.6538
P2→P3	1-3-3	0.6202	0.6799	0.731	0.6566
P2→P3	2-2-3	0.562	0.6588	0.9821	0.6677

Supplementary Table 6: Distance of ancestral population: 10.000 loci under the neutral scenario (*f*=0). Fraction of introgression for the 'alternative model' simulations is *f*=0.1 (1.000 locus). Recombination rate is r=0.01. Time of gene-flow is 0.1N_e.

Neutral model:

*ms 32 1 -I 4 8 8 8 8 -ej **P12** 2 1 -ej **P123** 3 1 -ej **P1234** 4 1 -r 10 1000*

Alternative model:

P3→P2: ms 32 1 -I 4 8 8 8 8 -ej ***P12*** *2 1 -ej* ***P123*** *3 1 -ej* ***P1234*** *4 1 -es 0.1 2* ***0.9*** *-ej 0.1 5 3 -r 10 1000*
P2→P3: ms 32 1 -I 4 8 8 8 8 -ej ***P12*** *2 1 -ej* ***P123*** *3 1 -ej* ***P1234*** *4 1 -es 0.1 3* ***0.9*** *-ej 0.1 5 2 -r 10 1000*

5 PopGenome Usage

```
# Install the PopGenome package from CRAN within R
install.packages(“PopGenome”)

# Load the package
library(PopGenome)

# Read the data (Fontaine et. al, 2015)
genome <- readVCF("AGC_refHC_bialSNP_AC2_2DPGQ.3L_V2.CHRcode2.DRYAD.vcf.gz",
10000,"4",1,45000000, include.unknown=TRUE)

# Define the populations

Aquad <- c("SRS408143", "SRS408145", "SRS408151", "SRS408155", "SRS408966",
"SRS408969", "SRS408972", "SRS408973", "SRS408983", "SRS420578")

Amela <- c("SRS408142", "SRS408185", "SRS408994")

Ameru <- c("SRS408186", "SRS408187", "SRS408967", "SRS408974", "SRS408992",
"SRS410266","SRS410284", "SRS410286", "SRS410290", "SRS420577")

# Define the outgroup
Chris <- c("CHRISTYI")

# Set the populations
genome <- set.populations(genome, list(Aquad,Amela,Ameru),diploid=TRUE)

# Set the outgroup
genome  <- set.outgroup(genome, Chris, diploid=TRUE)

# Transform the data into 50kb consecutive windows
slide <- sliding.window.transform(genome,50000,50000,type=2)

# Perform the Bdf
slide <- introgression.stats(slide, do.BDF=TRUE)
```

```
head(slide@BDF)
head(slide@BDF_bayes)

BDF <- slide@BDF
BDF_bayes <- slide@BDF_bayes

# Perform the D and f_d
slide <- introgression.stats(slide, do.D=TRUE)

head(slide@D)
head(slide@f)

D <- slide@D
f <- slide@f

# Get the genomic positions
genome.pos <- sapply(slide@region.names, function(x){
split <- strsplit(x," ")[[1]][c(1,3)]
val <- mean(as.numeric(split))
return(val)
})

# Plot the results (Bd-fraction)
plot(genome.pos, BDF, pch=19, ylab="Bd-fraction", xlab="genomic position",
main="3La inversion")

# Plot the results (Bd-fraction bayes factor)
plot(genome.pos, BDF_bayes, pch=19, ylab="Bd-fraction (Bayes factor)",
xlab="genomic position", main="3La inversion")
```