Leveraging shared ancestral variation to detect local introgression

Lesly Lopez Fang[1,2], Diego Ortega-Del Vecchyo[3], Emily Jane McTavish[1*], Emilia Huerta-Sanchez[4*]

[1] Department of Life & Environmental Sciences, University of California, Merced, Merced, California, United States of America

[2] Quantitative & Systems Biology Graduate Group, University of California, Merced, Merced, California, United States of America

[3] Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Santiago de Querétaro, Querétaro, México

[4] Ecology, Evolution and Organismal Biology and Center for Computational Biology, Brown University, Providence, Rhode Island, United States of America

* Corresponding authors
ejmctavish@ucmerced.edu (EJM)
emilia_huerta-sanchez@brown.edu (EHS)

Abstract

Introgression is a common evolutionary phenomenon that results in shared genetic material across non-sister taxa. Existing statistical methods such as Patterson's D statistic can detect introgression by measuring an excess of shared derived alleles between populations. The D statistic is effective to detect genome-wide patterns of introgression but can give spurious inferences of introgression when applied to local regions. We propose a new statistic, D^+, that leverages both shared ancestral and derived alleles to infer local introgressed regions. Incorporating both shared derived and ancestral alleles increases the number of informative sites per region, improving our ability to identify local introgression. We use a coalescent framework to derive the expected value of this statistic as a function of different demographic parameters under an instantaneous admixture model and use coalescent simulations to compute the power and precision of D^+. While the power of D and D^+ is comparable, D^+ has better precision than D. We apply D^+ to empirical data from the 1000 Genome Project and *Heliconius* butterflies to infer local targets of introgression in humans and in butterflies.

Introduction

Analyses of both modern and ancient DNA have revealed that introgression is a common evolutionary process in the history of many species. Introgression has been found in swordtail fish [1], *Heliconius* butterflies [2,3], and from Neanderthals and Denisovans to modern-day non-African populations [4–8] as well as many other systems. These observations suggest that introgression is pervasive and thus

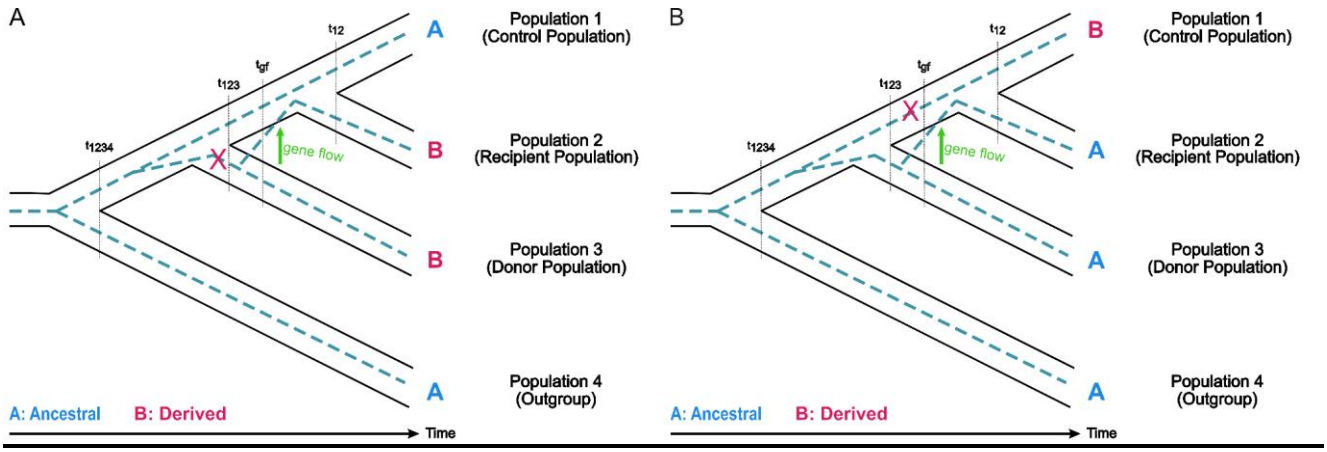
45 determining its relative contribution to the evolution of a species is of evolutionary
46 interest [9]. Therefore, detecting and quantifying introgressed segments in the genome
47 is necessary to begin measuring its biological importance. Introgression may introduce
48 both adaptive and deleterious variation in the recipient population. For example,
49 Tibetans inherited a beneficial haplotype at the *EPAS1* gene from Denisovans through
50 gene flow that facilitated high altitude adaptation to the hypoxic environment in the
51 Tibetan plateau [10–13] which is an example of adaptive introgression -- positive
52 selection acting on introgressed variants [10,14–16]. Similarly, purifying selection has
53 also acted on introgressed variation [17–20] to remove deleterious introgressed variants
54 and under specific conditions can mimic signatures of adaptive introgression [18,21].
55
56 The most widely-used method to detect introgression using data from one or more
57 individuals from each of four populations is the ABBA-BABA statistic, also known as
58 Patterson's D statistic [4,5]. This statistic has been used to detect introgression from
59 Neanderthals and Denisovans into modern humans ([4,22,23] as well as other systems.
60 The D statistic uses species tree and gene tree discordances within a 4-population tree
61 with two potential targets of introgression defined as population 1 (P_1) and population 2
62 (P_2); a donor population (P_3) as the source of gene flow to P_1 or P_2, and an outgroup
63 population (P_4, see Figs 1A and 1B). The patterns of biallelic single nucleotide
64 polymorphisms (SNP) generated by these gene trees (dotted lines in Figure 1a.b)
65 provide information on the shared ancestry between lineages in each population. The
66 D-statistic looks at patterns when the gene tree does not match the species/population
67 tree, which can be due to chance through Incomplete Lineage Sorting (ILS) or gene flow

from the donor population into P_1 or P_2. While ILS will generate an equal number of discordant sites shared between P_3 and P_1 and P_3 and P_2, introgression will result in an excess of shared sites between P_3 and either P_1 or P_2. D is a measure of this excess number of shared derived alleles.

The D statistic was designed to detect genome-wide gene flow but has also been used to look for signals of gene flow in local regions of the genome. However, studies have found that D produces spurious inferences of gene flow when applied to areas of the genome with low nucleotide diversity [24,25]. A previous study [25] partitioned butterfly genomes into small 5 kb windows and computed the D statistic in each window which showed that the D statistic becomes more unreliable when considering windows of low nucleotide diversity, because the variance of D is maximized in these windows. To improve inference of introgression in small windows [25] propose a new statistic, $\hat{f}_d$, that is a better estimator of the true introgression proportion. More recently [24] proposed to improve the D statistic by including the number of sites with an BBAA pattern — which is reduced in the presence of introgression— in the denominator of the D statistic.

In this study, we propose a new statistic, D^+, to detect introgression in genomic windows. In addition to using the shared derived variation measured in the D statistic, D^+ also leverages shared ancestral variation between the donor population and the recipient population. Introgression introduces not only mutations that accrued in the donor population before the gene flow event, but also re-introduces ancestral alleles in the recipient population. Following [5], we derive the theoretical expectations for the D^+

91 statistic under a coalescent framework to study its properties as a function of the
92 admixture proportion. We use simulations to measure its power, false positive rate and
93 precision compared to the D statistic. We also measure its performance by applying it to
94 humans and butterflies. We find that the D^+ statistic is more precise at detecting
95 introgressed regions than the D statistic due to its lower false positive rate in small
96 genomic regions, making it a useful statistic to identify local targets of introgression.
97



98

99 **Fig 1. Species and gene trees depicting informative sites due to gene flow.** (A)
100 Shared derived allele between population 2 and population 3, or ABBA site, and (B)
101 shared ancestral allele between population 2 and population 3, or BAAA site, due to
102 gene flow from population 3 to population 2. The ancestral allele is denoted A and the
103 derived allele is denoted B. t_{1234} is the time of divergence between population 4 and the
104 ancestral population of population 1, population 2 and population 3. t_{123} is the time of
105 divergence between population 1 and the ancestral population of population 1 and
106 population 2. t_{12} is the time of divergence between population 1 and population 2. t_{gf}
107 denotes the time of gene flow from donor population to recipient population.
108

Methods

D^+ statistic

Patterson's D statistic uses species and gene tree discordance within a 4-population tree with two populations as potential targets of introgression, population 1 (P_1) and population 2 (P_2). Population 3 (P_3) is a source of gene flow to either P_1 or P_2, and population 4 (P_4) serves as an outgroup (Fig 1). The patterns of biallelic single nucleotide polymorphisms (SNP) generated by the gene trees provide information on the shared ancestry between lineages in each population. Both the D and D^+ statistic look at site patterns yielded when the gene tree does not match the species tree. A mutation will convert an ancestral allele (A), determined by the allele present in the outgroup, into a derived allele (B). An ABBA site (Fig 1A) describes a derived allele shared between P_3 and P_2, while a BABA site occurs when a derived allele is shared between P_3 and P_1. An ABBA or BABA site could arise due to incomplete lineage sorting (ILS) or gene flow. Under coalescent expectations, incomplete lineage sorting will generate equal numbers of gene trees with ABBA or BABA sites. An ABBA site can only be generated in a gene tree where P_3 and P_2 coalesce first before they find a common ancestor with P_1. On the other hand, a BABA site only occurs on gene trees where P_1 and P_3 coalesce first before they find a common ancestor with P_2. We expect an excess of ABBA sites when there is gene flow from P_3 to P_2.

The D statistic measures an excess of ABBA or BABA sites [4,5]. D is the normalized difference between ABBA and BABA sites, $D = \frac{\sum_i ABBA_i - BABA_i}{\sum_i ABBA_i + BABA_i}$. The D statistic assumes

that the frequency of ABBA and BABA sites due to ILS is approximately equal. Therefore, an excess of shared derived sites between P_3 and P_2, or ABBA sites, indicates gene flow from P_3 to P_2 as shown in Fig 1A. Conversely, an excess of BABA sites indicates gene flow from P_3 to P_1.

We extend this idea by making use of the fact that introgressed regions are inherited in chunks that contain both shared derived alleles and ancestral alleles that are introduced into the recipient population. D^+ leverages the shared ancestral alleles between P_3 to P_2 to increase the amount of data about shared genetic variation in low nucleotide diversity regions. Sites where the ancestral allele is shared between P_3 and P_2 and the derived allele is only found in P_1 are BAAA sites (Fig 1B). In ABAA sites the ancestral allele is shared between P_3 and P_1 while P_2 has a derived allele. D^+ incorporates both shared derived alleles and ancestral alleles to strengthen our inferences of introgression.

$$D^+ = \frac{\sum_i (ABBA_i - BABA_i) + (BAAA_i - ABAA_i)}{\sum_i (ABBA_i + BABA_i) + (BAAA_i + ABAA_i)}$$

While in this paper, we mostly focus on comparisons between D^+ and D, note that we could also define a statistic $D_{ancestral}$ that measure the excess of shared ancestral alleles between P_3 and P_2 in a similar manner that the D statistic measures an excess of shared derived alleles between P_3 and P_2:

$$D_{ancestral} = \frac{\sum_i BAAA_i - ABAA_i}{\sum_i BAAA_i + ABAA_i}$$

152 $D_{ancestral}$ is normalized and ranges from -1 to 1, with $D_{ancestral} = 1$ indicating gene flow
153 from P_3 to P_2 and $D_{ancestral} = -1$ indicating gene flow from P_3 to P_1. $D_{ancestral}$
154 approximates zero under the null hypothesis of no gene flow.
155
156 [5] used a coalescent framework to derive the expectation of the D statistic under an
157 instantaneous admixture model (IUA). The probability of getting an ABBA or BABA site
158 is dependent on the mutation rate and the expected branch length of the branch where
159 a mutation yields an ABBA site (T_{ABBA}) or the branch where a mutation yields a BABA
160 site (T_{BABA}). The mutation rate μ is assumed to be constant. Therefore, the expected
161 number of ABBA or BABA sites can be estimated by calculating the expectation of
162 branch lengths of T_{ABBA} and T_{BABA} and multiplying by the mutation rate [5]. Similarly, we
163 can compute the probability of getting an ABAA or BAAA site (see S1 Appendix), and
164 we derived the expected lengths of T_{BAAA} and T_{ABAA} following the same framework (see
165 Fig 4). The full derivation of the expectation of T_{BAAA} and T_{ABAA} following is in S1
166 Appendix. We find that the analytical expectation of D^+ is $E[D^+] =$
167 $\frac{(\mu * E[T_{ABBA}] - \mu * E[T_{BABA}]) + (\mu * E[T_{BAAA}] - \mu * E[T_{ABAA}])}{(\mu * E[T_{ABBA}] + \mu * E[T_{BABA}]) + (\mu * E[T_{BAAA}] + \mu * E[T_{ABAA}])}$.
168
169 As is true of ABBA and BABA sites, the expected number of BAAA and ABAA sites are
170 equal when there is no gene flow. This is because, under no gene flow, we expect a
171 similar amount of ancestral allele sharing between P_1 and P_3 and between P_2 and P_3. In
172 the case of the BAAA and ABAA sites, we expect a similar amount of BAAA and ABAA
173 sites under no gene flow assuming the same mutation rate in P_1 and P_2. As the
174 admixture proportion from P_3 to P_2 increases, the number of BAAA sites exceeds the

number of ABAA sites. The expected difference is a function of the admixture proportion f and the branch lengths of t_{123} and t_{gf}.

$$E[T_{ABBA} - T_{BABA}] = E[T_{BAAA} - T_{ABAA}] = f(t_{123} - t_{gf})$$

Simulations to benchmark D^+

To evaluate D and D^+ we ran coalescent simulations using the software msprime [26]. The simulations followed the model depicting the evolutionary history of modern humans (Fig 2). The African and Eurasian populations are P_1 and P_2, respectively, and P_3 is the Neanderthal population. The outgroup (P_4) diverged 800,000 generations ago. The African-Eurasian and Neanderthal divergence time t_{123} was set 20,000 generations ago and the Eurasian and African divergence time t_{12} 16,000 generations ago [16]. The time of gene flow (t_{gf}) between Neanderthals and Eurasians was set 4,000 generations ago [16]. We use an admixture proportion (f) of 3%. All simulations had a constant N_e of 10,000, a mutation rate of 1.5^{-8} per bp per generation and a recombination rate of 10^{-8} per bp per generation following [16]. We ran 100 simulations, and, in each run, we sampled a single 20 MB genome from each population. The full code for simulations can be found in a GitHub repository (https://github.com/LeslyLopezFang/Dplus).

Introgressed regions were tracts of the genome of P_2 with ancestry from P_3. These introgressed tracts were used to quantify the number of introgressed bases in a window.

To calculate the expected branch lengths of T_{ABBA}, T_{BABA}, T_{BAAA} and T_{ABAA} and expectation of D and D^+ in Figs 4 and 5, we used msprime simulations with the same

parameters as in Fig 2 with a range of admixture proportions ($f = 0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\ and\ 1$). We ran 1,000,000 simulations to calculate the expected branch lengths and the number of ABBA, BABA, BAAA and ABAA sites to calculate D and D^+ per run. An example simulation command for 1,000,000 runs with 1 sample taken from each of the 4 populations and the time parameters listed above for an admixture proportion of 0.01 is:

```
mspms 4 1000000 -t 0.1 -I 4 1 1 1 1 -es 0.1 2 0.01 -ej f 5 3 -ej 0.25 1 2 -ej 0.5 2 3 -ej 20 3 4 -T
```

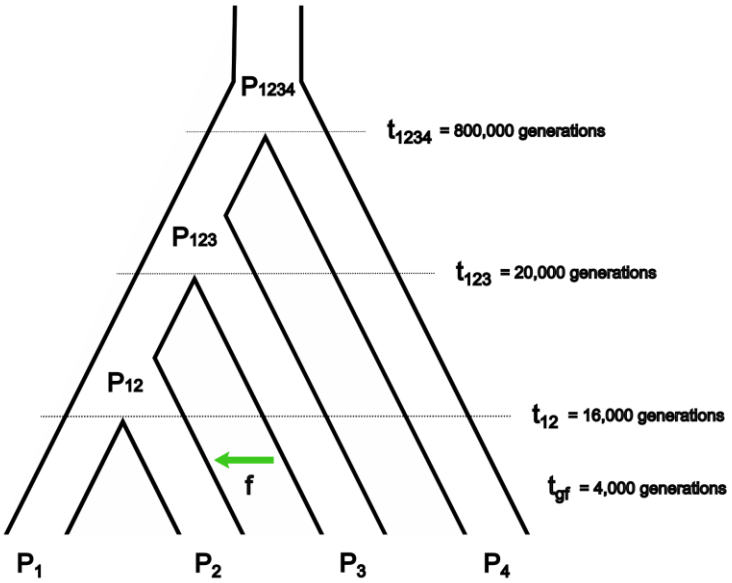


Fig 2. Demographic model for msprime simulations. (P_1) and (P_2) are sister populations that are closely related to (P_3). with (P_4) as the outgroup. There is gene flow from (P_3) to (P_2) at time t_{gf} 4,000 generations ago with an admixture proportion f. Divergence time of populations shown follow the demography of modern humans.

Calculating recall and precision in simulated human data.

We ran msprime simulations using the parameters shown in Fig 2 without an instance of admixture at t_{gf} to construct a null distribution for D and D^+ by sampling a genome from each population and computing D and D^+ in 50 kb non-overlapping windows. We take the significance threshold values for D and D^+ from their respective null distributions. For a p-value of 0.05, we get a signal of gene flow from P_3 to P_2 from the significance thresholds defined at the top 2.5% values from the null distribution of D and D^+. Undefined values (divided by 0) of D or D^+ where no informative sites were present in the window were dropped.

To find the true positives and false negatives we filter windows based on the percentage of bases overlapping introgressed segments. True positives are the introgressed windows that are statistically significant, while the false negatives are introgressed windows that are not statistically significant. The false positives for the simulated data are windows that have no introgressed bases but are statistically significant. Precision measures the probability of a window truly being introgressed given that its D^+ value is statistically significant. Precision is the percentage of true positives out of the sum of true positives and false positives. Recall measures how many of the introgressed windows are statistically significant and is the percentage of true positives out of the sum of true positives and false negatives.

Application of D^+ in modern-day humans

To evaluate the performance of D^+ at identifying introgressed regions in empirical data we apply D^+ to previously detected regions of Neanderthal introgression in modern-day humans. We assume that introgressed segments inferred in [7] from [27] are true introgressed segments. From the 1000 Genomes Project [28] we used an individual from the YRI (Yoruba in Ibadan, Nigeria) population for P_1 and an individual from the GBR (British from England and Scotland) population for P_2. P_3 is the Altai Neanderthal genome [27]. The ancestral allele of each position, or P_4, was taken from the ancestral allele listed in the 1000 Genome Project. For the GBR individual we used a Neanderthal introgression map including all the haplotypes inferred to be Neanderthal with a probability > 90% in [7]. We calculated D and D^+ in non-overlapping 50 kb windows using one autosomal chromosome of each individual from all three populations, discarding the first and last window of each chromosome. Each window had two D and D^+ values, one for each autosomal chromosome of the GBR individual but only the highest value was used.

To find significance thresholds, we treat the top 2.5% of D and D^+ values for the empirical distribution as thresholds for the statistically significant values. Introgressed windows were windows with a set minimum percentage of bases that overlap with the Neanderthal introgression map from [7]. A true positive for D or D^+ was an introgressed window equal to or greater than their corresponding statistical threshold. Recall was then calculated for introgressed windows. We assume that the introgression maps capture true positives or a subset of them; however, we cannot assume that regions not included in the introgression maps are true negatives. Therefore, we do not assess

false positives or precision. The full code can be found in a GitHub repository (https://github.com/LeslyLopezFang/Dplus).

Application of D^+ in *Heliconius* butterflies

D was applied to *Heliconius* butterflies and found to have high variance in areas of low nucleotide diversity [25]. To assess whether D^+ reduces variance in these areas of low nucleotide diversity we recreated Fig 3 from [25] using the same *Heliconius* genome data from [29]. They show values of D as a function of nucleotide diversity π for P_2 (the recipient population) in non-overlapping regions of 5 kb. Only biallelic alleles were used. D was computed using derived allele frequencies and we also use the frequencies from the four populations to compute D^+. The equation for D^+ can be written using the derived allele frequencies $\hat{p}_{ij}$ for population j (P_1, P_2, P_3 or P_4) at site i for L SNPs [4,5].

$$D^+ = \sum_{i=1}^{L} \frac{\left((1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) - \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})\right) + \left(\hat{p}_{i1}(1-\hat{p}_{i2})(1-\hat{p}_{i3})(1-\hat{p}_{i4}) - (1-\hat{p}_{i1})\hat{p}_{i2}(1-\hat{p}_{i3})(1-\hat{p}_{i4})\right)}{\left((1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) + \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})\right) + \left(\hat{p}_{i1}(1-\hat{p}_{i2})(1-\hat{p}_{i3})(1-\hat{p}_{i4}) + (1-\hat{p}_{i1})\hat{p}_{i2}(1-\hat{p}_{i3})(1-\hat{p}_{i4})\right)}$$

$\hat{f}_d$ [25] and d_f [24] were also computed for the 5 kb non-overlapping windows. $\hat{f}_d$ was only applied to windows where D is positive. The equation for $\hat{f}_d$ written in terms of derived allele frequencies with $\hat{p}_{iD}$ as the maximum of $\hat{p}_{i2}$ and $\hat{p}_{i3}$ is

$$\hat{f}_d = \sum_{i=1}^{L} \frac{\left((1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4})\right) - \left(\hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})\right)}{\left((1-\hat{p}_{i1})\hat{p}_{iD}\hat{p}_{iD}(1-\hat{p}_{i4})\right) - \left(\hat{p}_{i1}(1-\hat{p}_{iD})\hat{p}_{iD}(1-\hat{p}_{i4})\right)}$$

d_f incorporates BBAA sites where only P_1 and P_2 share a derived allele. The equation for d_f in terms of allele frequencies is

$$d_f = \sum_{i=1}^{L}\big((1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) + \hat{p}_{i1}\hat{p}_{i2}(1-\hat{p}_{i3})(1-\hat{p}_{i4})\big)$$

$$+\Big(\big(\hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})\big) + \hat{p}_{i1}\hat{p}_{i2}(1-\hat{p}_{i3})(1-\hat{p}_{i4})\Big)$$

Four samples were used, one each from *H. melpomene aglaope* (P_1), the recipient population *H.m. amaryllis* (P_2), the donor population *H. timareta thelxinoe* (P_3). The outgroup (P_4) consisted of a sample from species in the silvaniform clade including *H. hecale*, *H. ethilla*, *H. paradalinus sergestus* and *H. pardalinus ssp. nov.* The ancestral state of an allele was determined by the outgroup if the allele was fixed within the outgroup. Otherwise, it was the major allele of all four populations. The wing pattern loci *HmB* and *HmYb* are defined in [25]. Code was adapted from [25] with details in GitHub repository (https://github.com/LeslyLopezFang/Dplus).

Results

Theoretical results

The expectation for the values of D and D^+ is dependent on the branch lengths of the branches that produce each site pattern. T_{ABBA} is the length of the branch starting from the time of the most recent common ancestor of P_2 and P_3 until that lineage coalesces with P_1 (which happens in the ancestral population P_{123} under the instantaneous

309 admixture model). The average length of the T_{ABBA} branch increases with the migration
310 rate (Fig 3). A mutation on this branch produces an ABBA site pattern. T_{BABA} is then the
311 length of the branch from the time of the most recent common ancestor of P_1 and P_3
312 until that lineage coalesces with P_2. T_{BAAA} and T_{ABAA} are the external branches of P_1 and
313 P_2, respectively. When there is no gene flow, the average length of the external
314 branches of P_1 or P_2 are equal. With gene flow between P_2 and P_3, the external branch
315 of P_1 will be longer than the external branch of P_2; therefore, the expectation of T_{BAAA}
316 increases with the admixture proportion f.
317
318 The analytical and theoretical expectation of T_{ABBA}, T_{BABA}, T_{BAAA} and T_{ABAA} are shown in
319 Fig 3. The theoretical expectation of each branch takes into account all scenarios that
320 could produce each site pattern, including gene flow and no gene flow (S1 Appendix).
321 The simulated expected branch lengths approximate the theoretical expected branch
322 lengths at all the admixture proportions f calculated. When there is no admixture, the
323 number of ABBA sites is equal to the number of BABA sites as any sharing of derived
324 alleles between P_3 and P_2 (or P_3 and P_1) is due to incomplete lineage sorting. In the
325 case of ancestral sharing and under a model of no admixture, the number of BAAA sites
326 and ABAA sites will be equal because we assume equal mutation rates in P_1 and P_2.
327
328 For all values of migration between P_2 and P_3, the expected branch lengths that can
329 lead to a BAAA (T_{BAAA}) or a ABAA (T_{ABAA}) site are always greater than the expected
330 branch lengths that can lead to an ABBA (T_{ABBA}) or BABA site (T_{BABA}). Therefore, if we
331 assume a constant mutation rate, we expect to see more ABAA sites than BABA sites

and more BAAA sites than ABBA sites. In Fig 3, assuming a constant mutation rate multiplied with the analytical and simulated expected branch lengths, there are 5-6 times more BAAA and ABAA sites than ABBA and BABA sites.

Interestingly, our theoretical results also show that even though the number of BAAA and ABAA is higher (than ABBA or BABA), the difference between T_{BAAA} and T_{ABAA} (T_{BAAA} - T_{ABAA}) is equal to the difference (T_{ABBA} - T_{BABA}). Therefore, for all admixture proportions between P_2 and P_3, the expected difference of BAAA and ABAA sites (BAAA - ABAA) is equal to the expected difference of ABBA and BABA sites (ABBA - BABA). These observations suggest that leveraging ancestral shared variation can be informative about introgression and provides justification for defining D^+ which leverages both ancestral and derived allele sharing to maximize the number of informative sites used in a genomic window. This increase in informative sites can provide greater predictive accuracy for detecting local gene flow.

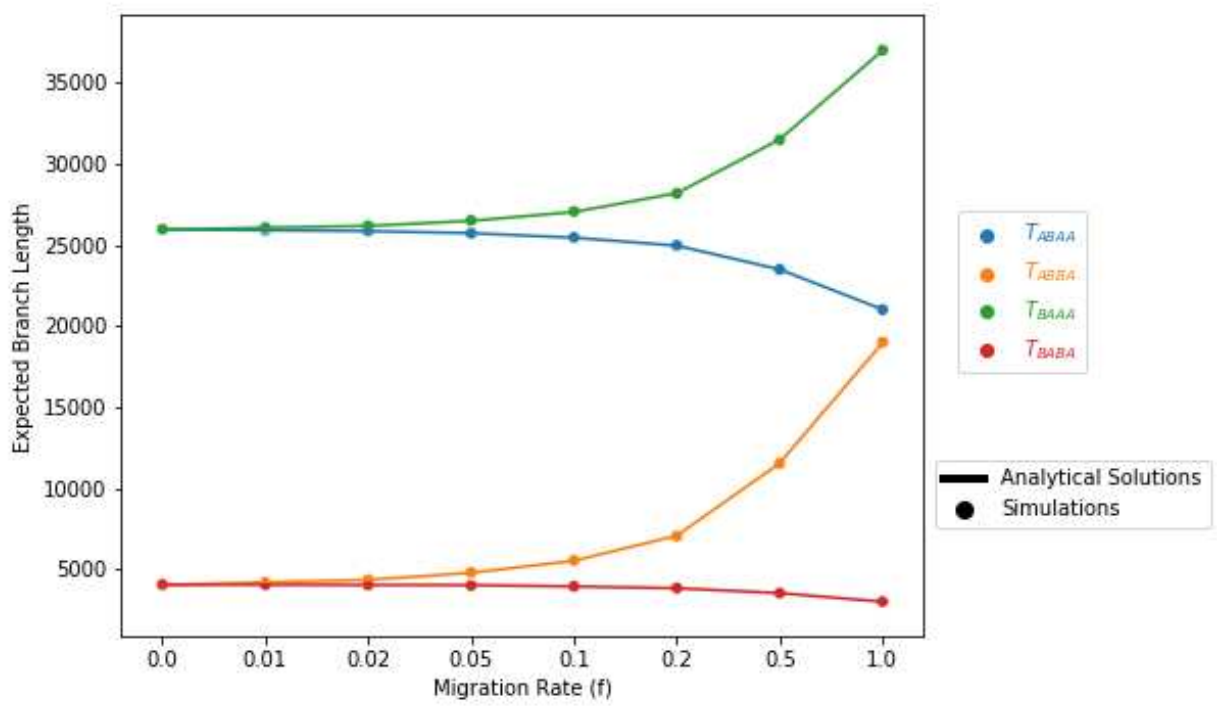


Fig 3. Analytical and simulated expected branch lengths of T_{ABBA}, T_{BABA}, T_{BAAA} and T_{ABAA}. The analytical (lines) and simulated (dots) expected branch lengths of T_{ABBA}, T_{BABA}, T_{BAAA} and T_{ABAA} for different proportions of admixture f between P_3 and P_2. The solutions to the analytical expectations match the simulated expectations. The branch length of T_{ABBA} is the branch that would produce an ABBA site pattern. The expectation of T_{ABBA} ($E[T_{ABBA}]$) can be used to calculate the expected number of ABBA sites. The same is true for T_{BABA}, T_{BAAA}, and T_{ABAA} for their respective site patterns. With no admixture ($f = 0$) the expected branch lengths for ABBA and BABA sites are equal ($E[T_{ABBA}] = E[T_{BABA}]$), as are the expected branch lengths for BAAA and ABAA sites ($E[T_{BAAA}] = E[T_{ABAA}]$) because the number of ABBA sites equals BABA sites and the number of BAAA sites equals the number ABAA sites due to ILS. As the admixture proportion increases, the expectation of T_{ABBA} and T_{ABBA} increases due to excess ABBA

and BAAA sites. The difference in T_{BAAA} and T_{ABAA} (T_{BAAA} - T_{ABAA}) is equal to the difference in T_{ABBA} and T_{BABA} (T_{ABBA} - T_{BABA}).

D has a high false positive rate in small genomic windows.

We calculated D and D^+ for 50 kb windows on simulated genomes following the demography in Fig 2 with no admixture event at t_{gf} to get the null distribution of D and D^+ (Fig 4A). The null distriution of D is a multimodal distribution with large peaks at the tails as well as zero. The tails ($D = 1$ and $D = -1$) account for 12.2% of the distribution. These peaks at the tails cause a high false positive rate of 12.2% for D at p-values less than 0.13 (Fig 4B) because the significance threshold for D is 1 or -1. Therefore, we have low power to assess statistically significant values of D. In contrast D^+ has a null distribution centered on zero. The null distribution is much narrower than the null distribution of D and does not have peaks at the tails. As expected, the false positive rate of D^+ approximates the p-value set to find significant values of D^+ up until a significance threshold approaches 0 for high p-values (p-values >= 0.94) (Fig 4A).

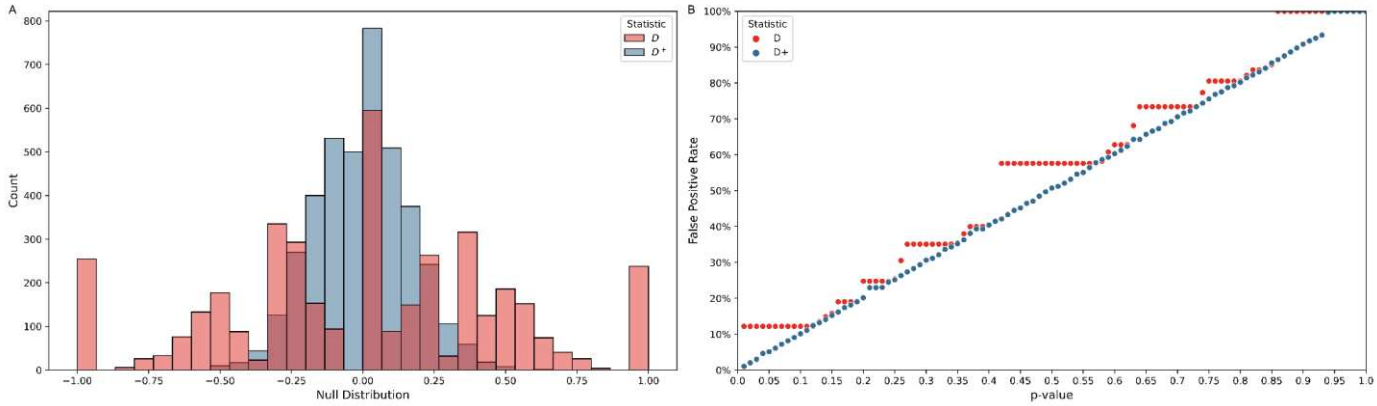


Fig 4. Null distribution and false positive rate for D and D^+ in simulations with no gene flow. D and D^+ were calculated in 50 kb windows of 100 runs of a 20 MB simulated genome under a model with no admixture. **(A)** The expectation of the null distribution of D and D^+ is zero. The null distribution for D (red) is multi-modal at the tails with the tails (-1 and 1) accounting for 12.2% of the values of D. The null distribution of D^+ (blue) is centered around zero. The null distribution of D^+ has a smaller variance than D. **(B)** False positive rates for D (red) and D^+ (blue) of null distribution. The p-value in the x-axis is used to set a significance threshold to get a false positive rate in the y-axis. D has a false positive rate of 12.2% with p-values less than 0.12. The false positive rate of D^+ is similar to the corresponding p-values.

D^+ has better precision than D in simulated data

We calculated precision and recall for 50 kb windows of 100 simulations with a 20 MB simulated genome shown in Fig 5 following the demography in Fig 2. Undefined values were dropped so more windows were analyzed for D^+ than D because D had more undefined values. While precision measures the accuracy of windows giving a signal of

gene flow from P_3 to P_2 through statistical significance, recall measures how many introgressed windows the statistic can detect without considering false positives.

We obtained precision and recall for p-values from 0.01-1 (Fig 5). Each p-value has a corresponding significant threshold value from the null distribution in Fig 4A in which values of D or D^+ greater than the threshold are statistically significant. For realistic p-values (i.e. p-values < 0.05), D^+ has better precision than D; At these realistic p-values, precision for D^+ ranges from 56.6% to 59.2% and the precision of D is 39.8% (Fig 5A). For these p-values, D has better recall than D^+ (Fig 5B). Precision and recall are the same, 39.8% and 6.7% respectively, for D at p-values < 0.13 because the threshold for a statistically significant D value is 1 since the null distribution is multimodal with peaks at the tails (Fig 4A).

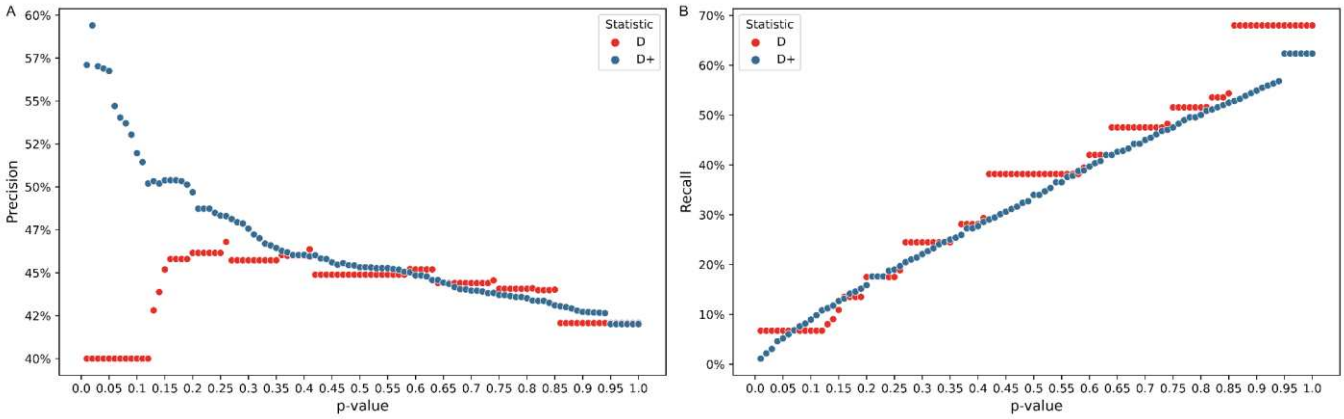


Fig 5. Precision and recall of D and D^+ in simulations. The Precision-Recall of D and D^+ for simulations with an admixture proportion of 3%. D (red) and D^+ (blue) were computed in non-overlapping 50 kb windows of 100 simulations of a 20 MB genome from each population with an admixture proportion of 3% ($f = 0.03$). (A) Precision and

(B) recall are shown as a function of the p-value (0.01-1) used to get a significant threshold value of D and D^+.

D^+ identifies Neanderthal introgressed regions in modern-day humans

To investigate the behavior of D^+ in real data, we applied D^+ to modern-day humans [28] and an Altai Neanderthal [27] to find if signals of gene flow corresponded to previously identified Neanderthal introgressed regions. Unlike simulated data, in real human genomes we do not know the ground truth, and to compare the performance of D and D^+, we assumed that the Neanderthal introgressed regions from [7] were true positives. We calculated D and D^+ in 50 kb non-overlapping windows and computed the recall of D and D^+ in introgressed windows (Fig 8). Introgressed windows are defined as windows with a minimum percentage of bases in the windows that overlap with introgressed segments from [7]. Statistical significance was computed using the genome-wide distribution as the null distribution. Recall is the number of these introgressed windows that were statistically significant over the total number of windows with a minimum percentage of introgressed bases. Recall for D^+ was consistently better than D across all windows tested with a minimum percentage of introgressed bases. The recall for both statistics decreases when the overlap between a window and an introgressed segment increases because the number of introgressed windows used to calculate recall decreases.

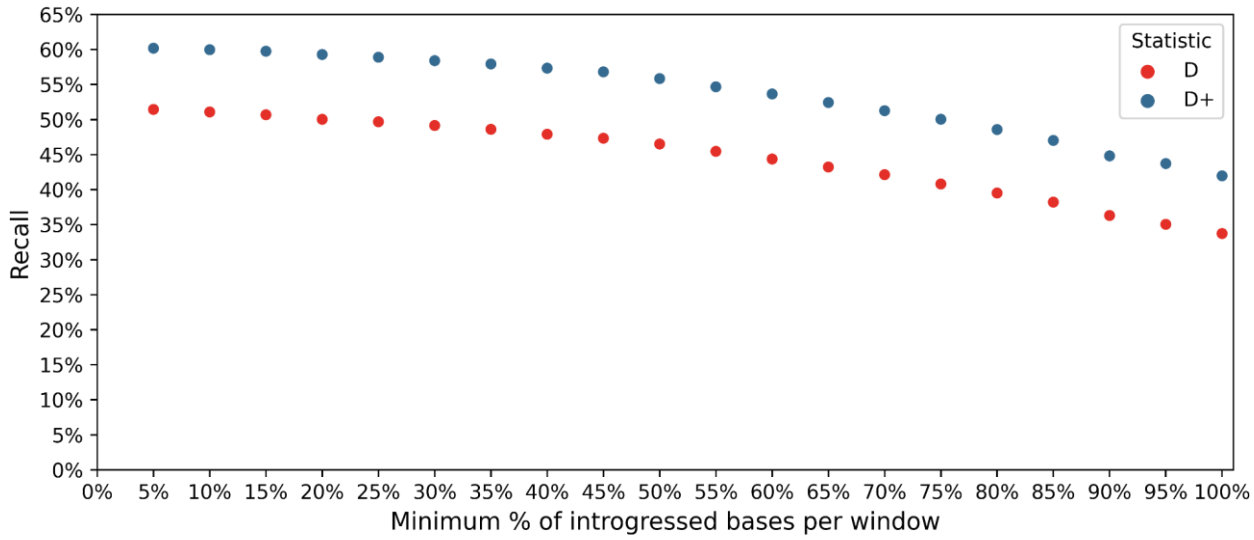


Fig 6. Recall of D and D^+ in human data. The recall of D and D^+ in non-overlapping 50 kb windows. Windows overlap with Neandertal introgression maps [7] from 5% to 100%. The populations are as follows: P_1: YRI, P_2: GBR, P_3: Altai Neandertal, P_4: Ancestral Alleles. Data for humans from 1000 Genomes Project [28] and data for Altai Neandertal from [27].

D^+ can detect introgression events in regions of low nucleotide diversity

One of the main reasons the D statistic is not useful for detecting introgression in small regions of the genome is that the variance of D is high in areas of low nucleotide diversity [25]. To address this [25] proposed $\hat{f}_d$ as an alternative approach to quantify and detect introgression in small genomic regions. The numerator of $\hat{f}_d$ is in the same form as that of D; however, the denominator of $\hat{f}_d$ replaces the derived allele frequency of P_2 and P_3 with the maximum derived allele frequency of P_2 and P_3. This leads to $\hat{f}_d$

449 having a lower variance in areas of low nucleotide diversity, thus reducing spurious
450 results in comparison to D. Like $\hat{f}_d$, d_f is also designed to quantify the admixture
451 proportion of small genomic regions [24]. The approach in d_f is to incorporate BBAA
452 sites as fewer sites with this pattern are expected when introgression occurs between
453 P_2 and P_3 or between P_1 and P_3.
454
455 Both $\hat{f}_d$, d_f are estimates of the admixture proportion while D and D^+ are used to detect
456 and not quantify introgression. To compare D^+ to $\hat{f}_d$ and d_f we used the same
457 *Heliconius* genome data from [29]. *Heliconius* butterflies have strong evidence for both
458 genome-wide and adaptive introgression between species, including mimicry loci for
459 wing patterns [14,29,30]. We use these data to compute these statistics in windows as
460 a function of nucleotide diversity, since the relationship between D and nucleotide
461 diversity observed in [29] inspired the developments of new statistics to detect and
462 quantify introgression in small windows of the genome. For the four populations, we use
463 *H. melpomene aglaope* as P_1, *H. melpomene amaryllis* as P_2, *H. timareta thelxinoe* as
464 P_3 and the *H. hecale*, *H. ethilla*, *H. paradalinus sergestus* and *H. pardalinus ssp. nov.*
465 species in the silvaniform clade as the outgroup (P_4). We compute nucleotide diversity
466 π, $\hat{f}_d$, d_f, D and D^+ in non-overlapping 5 kb windows. Windows from the introgressed
467 loci responsible for the red wing pattern (*HmB*) and the yellow and white wing pattern
468 (*HmYb*) are shown in red and yellow, respectively, in Fig 7. We find similar results as
469 [25]; D has a high variance and a wide distribution in regions of low nucleotide diversity
470 (Fig 9A). As nucleotide diversity increases the distribution of D narrows. $\hat{f}_d$ reduces the
471 high variance of values in areas of low nucleotide diversity (Fig 7B). d_f also reduces

variance with most of the d_f values centered around zero, including windows with the *HmB* and *HmYb* loci (Fig 7C). D^+ has smaller variance with fewer outliers than D and similar variance to d_f (Fig 7D). Many of the highest positive values of D^+ are in windows with the *HmB* and *HmYb* loci. We also computed $D_{ancestral}$ which only uses the ancestral shared patterns (ABAA and BAAA), and it has surprisingly low variance as well (S2 Fig).

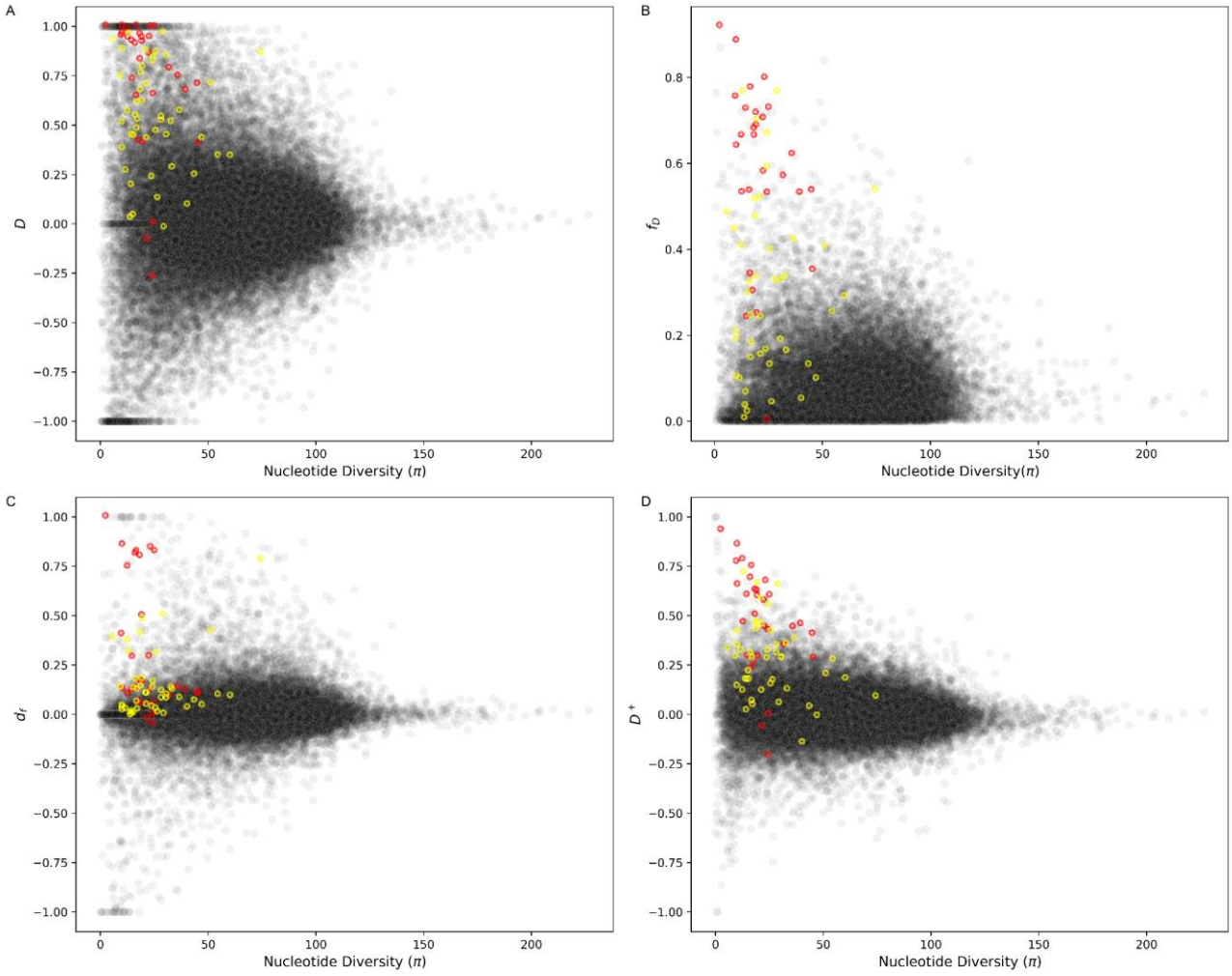


Fig 7. Application of D, $\hat{f}_d$, d_f and D^+ in *Heliconius* butterfly. (A) D, (B) $\hat{f}_d$, (C) d_f and (D) D^+ as a function of nucleotide diversity in P_2 in non-overlapping 5 kb windows.

P_1: *H. melpomene aglaope*, P_2: *H. melpomene amaryllis,* P_3: *H. timareta thelxinoe,* P_4: *H. hecale*, *H. ethilla*, *H. paradalinus sergestus* and *H. pardalinus ssp. nov.* from the silvaniform clade. Red and yellow circles correspond to windows with introgressed loci HmB and HmYb, respectively. Methods follow Fig 3 from [25] with *Helicionius* genome data from [29].

Discussion

Multiple studies have found that introgression plays an important evolutionary role as it introduces new genetic variation in a population that can be targeted by natural selection; this is an accelerated process of accumulating new alleles compared to a *de novo* mutation process. Therefore, detecting what regions of the genome exhibit signatures of introgression is an important step to evaluate its relative contribution to evolution. To date, Patterson's D statistic is the most widely used metric for detection of introgression genome wide. While D works well at detecting introgression at the genome-wide scale, some studies have shown that D might not be the best choice to detect introgression in small regions of the genome. In this paper, we define a new statistic, D^+, that leverages sites with both shared ancestral and shared derived alleles to improve detection of introgression in small genomic windows. We use coalescent theory to understand its theoretical properties and derive the expectation of D^+ as a function of gene flow. We show that the expected counts of BAAA sites and ABAA sites are equal under a model of no introgression. As the proportion of admixture increases one of these two site patterns increases suggesting that BAAA and ABAA sites are

505 informative to detect introgression. Interestingly, our theoretical results also show that
506 the expected difference in counts of BAAA and ABAA sites equals the expected
507 difference of ABBA and BABA sites (Fig 3). However, in general there are more BAAA
508 and ABAA sites than ABBA and BABA sites.
509
510 D^+ is more conservative than D with a smaller expectation and variance than D (Fig 4
511 and S1 Fig). As a result, D^+ has less false positives than D, likely because D^+ includes
512 more informative sites (Fig 6). Therefore, D^+ also has better precision than D in
513 simulated data under the Neanderthal admixture model presented in Fig 2 (Fig 5A).
514 While D had a slightly higher recall in simulated human data (Fig 5B), D^+ had slightly
515 higher recall in human empirical data despite D having generally more extreme values
516 (frequently reaching a maximum value of 1 across windows). Overall, D^+ has statistical
517 properties that make it more stable than D at detecting introgression in small genomic
518 windows and provides an alternative method to detect introgression.

519 Other methods such as $\hat{f}_d$ [25] and d_f [24] have been derived from Patterson's D to
520 quantify the introgression proportion, f, in small genomic regions. $\hat{f}_d$ leverages ABBA
521 and BABA sites, d_f leverages ABBA, BABA and BBAA sites, and D^+ leveraged ABBA,
522 BABA, BAAA and ABAA sites. To compare with these methods, we applied them to a
523 *Heliconius* butterflies data set, and we found that similarly to $\hat{f}_d$ and d_f, the variance of
524 D^+ is reduced in regions of low nucleotide diversity. This suggests that like $\hat{f}_d$ and d_f,
525 D^+ will also not lead to a high number of false positives, especially in regions of low

nucleotide diversity. In fact, just using the ancestral site patterns alone is better behaved than the D statistic (S2 Fig), which shows the utility of using ancestral shared variation.

All these statistics have both shared and distinct aspects in how they leverage genetic patterns, and future studies might focus on integration of these approaches to improve the detection and quantification of introgression. We recognize that all these statistics have been benchmarked to detect or quantify introgression under very specific and simple demographic scenarios that may not closely reflect the true demographic histories of actual species or populations. Future studies that compare and contrast how different statistics - that detect and quantify introgression [24,25,31–33] - behave under more complex demographic scenarios and under different evolutionary time scales will help characterize the behavior of these statistics and expand our understanding of the power and limitations of each method.

Here, we have shown that ancestral shared variation between a donor and recipient population is influenced by the introgression proportion. Notably, in humans, there is evidence that archaic introgression may have re-introduced ancestral alleles with regulatory effects [34] pointing to the importance of studying ancestral shared variation. Beyond their functional effects, leveraging ancestral information may be informative on ghost admixture events from uncharacterized ghost populations [27]. Patterns of ancestral shared variation may also help address how pervasive introgression is across the tree of life, and D^+ which leverages both derived and shared ancestral variation, provides a new way to detect introgression that can help answer this question.

Acknowledgements

We thank the E.H.S. laboratory at Brown University and the Blois-McTavish group at UC Merced.

References

1. Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 2018;360: 656–660.

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

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

4. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328: 710–722.

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

6. Vernot B, Akey JM. Resurrecting surviving Neandertal lineages from modern human genomes. Science. 2014;343: 1017–1021.

7. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507: 354–357.

8. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018;173: 53–61.e9.

9. Dagilis AJ, Peede D, Coughlan JM, Jofre GI, D'Agostino ERR, Mavengere H, et al. 15 years of introgression studies: quantifying gene flow across Eukaryotes. bioRxiv. 2021. p. 2021.06.15.448399. doi:10.1101/2021.06.15.448399

10. Huerta-Sánchez E, Casey FP. Archaic inheritance: supporting high-altitude life in Tibet. J Appl Physiol. 2015;119: 1129–1134.

11. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512: 194–197.

12. Zhang P, Zhang X, Zhang X, Gao X, Huerta-Sanchez E, Zwyns N. Denisovans and Homo sapiens on the Tibetan Plateau: dispersals and adaptations. Trends Ecol Evol. 2022;37: 257–267.

13. Zhang X, Witt K, Ko A, Yuan K, Xu S, Nielsen R, et al. The history and evolution of the Denisovan-EPAS1haplotype in Tibetans. doi:10.1101/2020.10.01.323113

14. Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, et al. Adaptive introgression across species boundaries in Heliconius butterflies. PLoS Genet. 2012;8: e1002752.

15. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. 2015;16: 359–371.

16. Racimo F, Marnetto D, Huerta-Sánchez E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Mol Biol Evol. 2017;34: 296–317.

17. Harris K, Nielsen R. The Genetic Cost of Neanderthal Introgression. Genetics. 2016;203: 881–891.

18. Kim BY, Huber CD, Lohmueller KE. Deleterious variation shapes the genomic landscape of introgression. PLoS Genet. 2018;14: e1007741.

19. Petr M, Pääbo S, Kelso J, Vernot B. Limits of long-term selection against Neandertal introgression. Proc Natl Acad Sci U S A. 2019;116: 1639–1644.

20. Telis N, Aguilar R, Harris K. Selection against archaic hominin genetic variation in regulatory regions. Nature Ecology & Evolution. 2020;4: 1558–1566.

21. Zhang X, Kim B, Lohmueller KE, Huerta-Sánchez E. The Impact of Recessive Deleterious Variation on Signals of Adaptive Introgression in Human Populations. Genetics. 2020;215: 799–812.

22. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 2011;89: 516–528.

23. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 2012 [cited 10 Sep 2021]. Available: https://science.sciencemag.org/content/338/6104/222.abstract

24. Pfeifer B, Kapan DD. Estimates of introgression as a function of pairwise distances. BMC Bioinformatics. 2019;20: 207.

614 25. Martin SH, Davey JW, Jiggins CD. Evaluating the Use of ABBA–BABA Statistics to
615 Locate Introgressed Loci. Molecular Biology and Evolution. 2015. pp. 244–257.
616 doi:10.1093/molbev/msu269

617 26. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and
618 genealogical analysis for large sample sizes. doi:10.1101/033118

619 27. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The
620 complete genome sequence of a Neanderthal from the Altai Mountains. Nature.
621 2014;505: 43–49.

622 28. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP,
623 Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:
624 68–74.

625 29. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et
626 al. Genome-wide evidence for speciation with gene flow inHeliconiusbutterflies.
627 Genome Research. 2013. pp. 1817–1828. doi:10.1101/gr.159426.113

628 30. Consortium THG, The Heliconius Genome Consortium. Butterfly genome reveals
629 promiscuous exchange of mimicry adaptations among species. Nature. 2012. pp.
630 94–98. doi:10.1038/nature11041

631 31. Hibbins MS, Hahn MW. The Timing and Direction of Introgression Under the
632 Multispecies Network Coalescent. Genetics. 2019;211: 1059–1073.

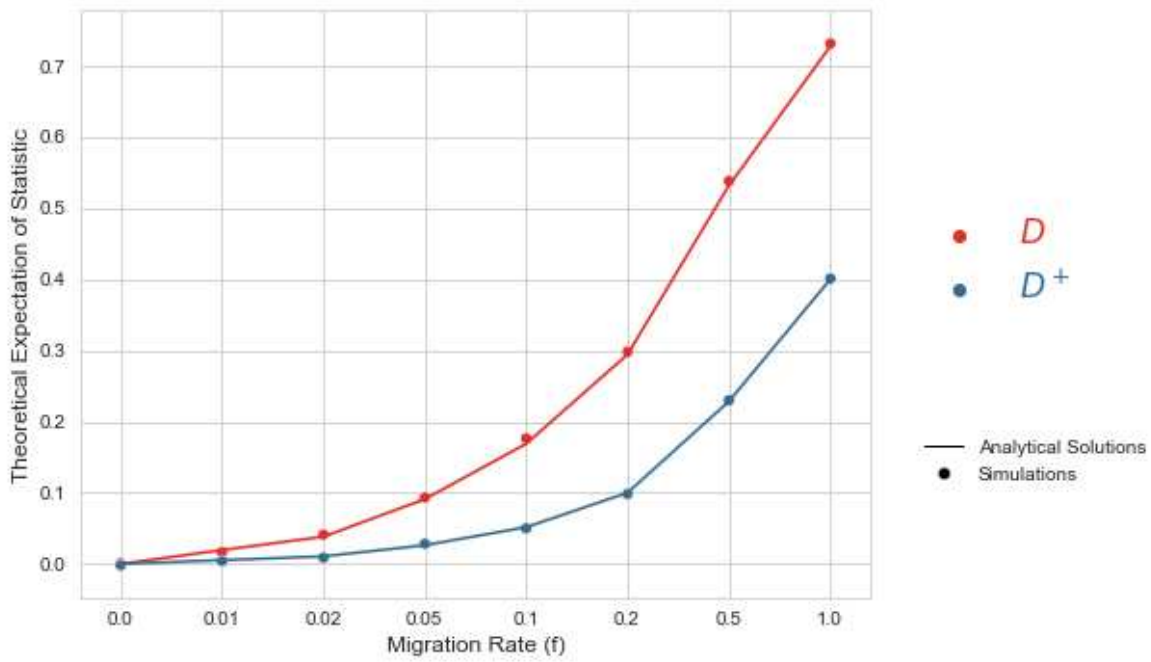
633 32. Hamlin JAP, Hibbins MS, Moyle LC. Assessing biological factors affecting
634 postspeciation introgression. Evol Lett. 2020;4: 137–154.

635 33. Hibbins MS, Hahn MW. The effects of introgression across thousands of
636 quantitative traits revealed by gene expression in wild tomatoes. PLoS Genet.
637 2021;17: e1009892.

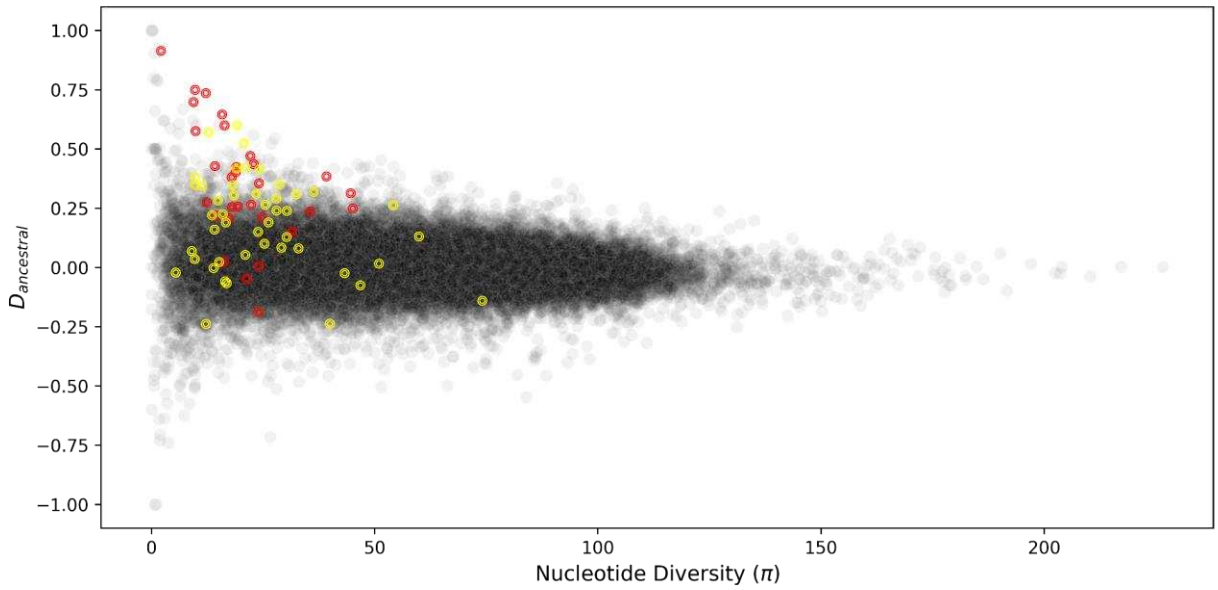
638 34. Rinker DC, Simonti CN, McArthur E, Shaw D, Hodges E, Capra JA. Neanderthal
639 introgression reintroduced functional ancestral alleles lost in Eurasian populations.
640 Nat Ecol Evol. 2020;4: 1332–1341.

641

Supplementary Information



S1 Fig. Theoretical and analytical expectations of D and D^+. Analytical (lines) and simulated (dots) expectation of D (red) and D^+ (blue) as a function of the admixture proportion (f) of 0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5 and 1. The simulated expectations of D and D^+ concur with the analytical expectations. The expectation of D and D^+ are both zero when there is no gene flow and both expectations increase as f increases.



S2 Fig. Application of $D_{ancestral}$ in *Heliconius* butterfly. $D_{ancestral}$ as a function of nucleotide diversity in P_2 in non-overlapping 5 kb windows. P_1: *H. melpomene aglaope*, P_2: *H. melpomene amaryllis,* P_3: *H. timareta thelxinoe,* P_4: *H. hecale*, *H. ethilla*, *H. paradalinus sergestus* and *H. pardalinus ssp. nov.* from the silvaniform clade. Red and yellow circles correspond to windows with introgressed loci HmB and HmYb, respectively. Methods follow Figure 3 from (15) with *Helicionius* genome data from (20).

S1 Appendix. Derivation of D^+.

[5] uses the instantaneous admixture model (IUA) to propose a test that infers patterns of gene flow (Fig 1)

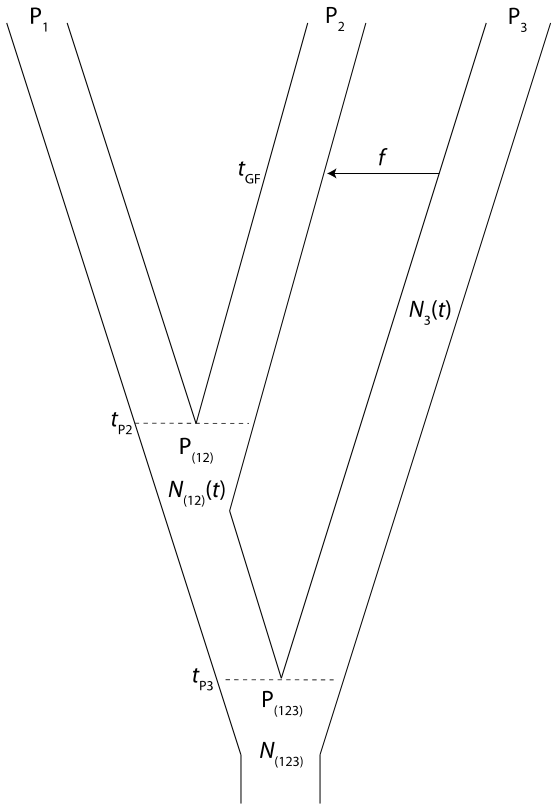


Fig 1. Taken from [5]. Instantaneous admixture model (IUA).

Under the IUA, (4) and [5] propose the D statistic is to infer patterns of gene flow. It quantifies differences in the number of site patterns $N(ABBA)$ and $N(BABA)$:

$$D = \frac{N(ABBA) - N(BABA)}{N(ABBA) + N(BABA)}$$

Where $N(ABBA)$ and $N(BABA)$ are the number of sites that have an ABBA or a BABA

pattern. In an ABBA pattern, the lineages P_2 and P_3 share a derived site. Under the BABA pattern, the lineages P_1 and P_3 share a derived site.

To estimate D, [5] assumed that the effective population sizes are equal across the whole demographic scenario. Therefore $N_1 = N_2 = N_3 = N_{12} = N_{123}$. [5] derived the probability of obtaining an ABBA or a BABA site, where the probability of obtaining those sites is equal to the product of the mutation rate times the expected length of the branch where a mutation would produce an ABBA or BABA site, respectively. Based on [5], the expected length of the branch T_{ABBA} where a mutation would produce an ABBA site is equal to:

$$E[T_{ABBA}] = f(T_{P3} - T_{GF}) + (1-f)\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{P2}}\frac{2N}{3} + f\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{GF}}\frac{2N}{3}$$

And:

$$E[T_{BABA}] = (1-f)\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{P2}}\frac{2N}{3} + f\left(1-\frac{1}{2N}\right)^{T_{P3}-T_{GF}}\frac{2N}{3}$$

Using those expected branch lengths, the expected value of the D statistic can be calculated as:

$$E[D] = \frac{E[T_{ABBA}] - E[T_{BABA}]}{E[T_{ABBA}] + E[T_{BABA}]}$$

Now we will derive the expected lengths of the branches where a mutation would create a BAAA or an ABAA site. A BAAA site is one where there is a derived allele in the P_1 individual and an ABAA site only contains a derived allele in the P_2 individual.

BAAA sites

In this section we show how to estimate the expected lengths of branches that produce a BAAA site under the IUA. The expected branch lengths are equal to the sum of the contributions from six different scenarios that could lead to the coalescence of the lineage P_1:

1) There was no gene flow from P_3 to P_2. The P_1 lineage coalesces with the P_2 lineage between times T_{P3} and T_{P2}:

$$(1-f) * \sum_{i=1}^{T_{P3}-T_{P2}} (Branch\ length\ at\ generation\ i) * P(Coalescence\ at\ generation\ i)$$

$$(1-f) * \left(\sum_{i=1}^{T_{P3}-T_{P2}} (T_{P2}+i) * \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1} \right)$$

2) There was no gene flow from P_3 to P_2. The P_1 lineage coalesces with either P_2 or P_3 in the first coalescent event that takes place after T_{P3} going backwards into the past.

$$(1-f) * P(No\ coalescence\ of\ P_1\ and\ P_2\ before\ T_{P3})$$

$$* E[Branch\ length\ in\ first\ coalescent\ event\ between\ lineages\ P_1, P_2\ and\ P_3]$$

$$* P(P_1\ lineage\ coalesces\ in\ first\ coalescent\ event)$$

$$(1-f) * \left(1 - \sum_{i=1}^{T_{P3}-T_{P2}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(\frac{2N}{3} + T_{P3}\right) * \frac{2}{3}$$

3) There was no gene flow from P_3 to P_2. The P_1 lineage coalesces with the ancestral lineage of P_2 and P_3 in the second coalescent event that takes place after T_{P3} going backwards into the past.

$$(1-f) * P(no\ coalescence\ of\ P_1\ and\ P_2\ before\ T_{P3})$$

$$* E\left[\begin{array}{c} Branch\ length\ in\ second\ coalescent\ event\ between\ lineages\ P_1 \\ and\ the\ ancestral\ lineage\ of\ P_2\ and\ P_3 \end{array}\right]$$

$$* P(P_1\ lineage\ coalesces\ in\ second\ coalescent\ event)$$

$$(1-f) * \left(1 - \sum_{i=1}^{T_{P3}-T_{P2}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(2N + \frac{2N}{3} + T_{P3}\right) * \frac{1}{3}$$

4) There was gene flow from P_3 to P_2. The P_3 and P_2 lineages did not coalesce between times T_{P3} and T_{GF}. The lineage P_1 coalesces in the first coalescent event after T_{P3} going backwards into the past.

$$f * P(no\ coalescence\ for\ P_2\ and\ P_3\ before\ T_{GF})$$

$$* E[branch\ length\ in\ first\ coalescent\ event\ between\ lineages\ P_1, P_2\ and\ P_3]$$

$$* P(P_1\ lineage\ coalesces\ in\ first\ coalescent\ event)$$

$$f * \left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(\frac{2N}{3} + T_{P3}\right) * \frac{2}{3}$$

5) There was gene flow from P_3 to P_2. The P_3 and P_2 lineages did not coalesce between times T_{P3} and T_{GF}. The P_1 lineage coalesces with the ancestral lineage of P_2 and P_3 in the second coalescent event that takes place after T_{P3} going backwards into the past.

$$f * P(no\ coalescence\ for\ P_2\ and\ P_3\ before\ T_{GF})$$

$$* E\left[\begin{array}{c} branch\ length\ in\ second\ coalescent\ event\ between\ lineages\ P1 \\ and\ the\ ancestral\ lineage\ of\ P_2\ and\ P_3 \end{array}\right]$$

$$* P(P_1\ lineage\ coalesces\ in\ second\ coalescent\ event)$$

$$f * \left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(2N + \frac{2N}{3} + T_{P3}\right) * \frac{1}{3}$$

6) There was gene flow from P_3 to P_2. P_2 and P_3 coalesce between T_{P3} and T_{GF}. The lineage P_1 coalesces with the lineage ancestral to P_2 and P_3 after T_{P3} going backwards into the past.

$$f * P(coalescence\ for\ P_2\ and\ P_3\ before\ T_{GF})$$

$$* E[branch\ length\ in\ coalescent\ event\ between\ lineages\ P_1\ and\ lineage\ (P_2, P_3)]$$

$$f * \left(\sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^{i-1} \right) * (2N + T_{P3})$$

If we sum those six contributions, we get:

$$E[T_{BAAA}] = (1-f) * \left(\left(\sum_{i=1}^{T_{P3}-T_{P2}} (T_{P2} + i) * \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^{i-1} \right) + \left(\left(1 - \sum_{i=1}^{T_{P3}-T_{P2}} \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^{i-1} \right) * \left(\frac{2N}{3} + T_{P3}\right) * \frac{2}{3} \right) + \left(\left(1 - \sum_{i=1}^{T_{P3}-T_{P2}} \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^{i-1} \right) * \left(2N + \frac{2N}{3} + T_{P3}\right) * \frac{1}{3} \right) \right) + f * \left(\left(\left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^{i-1} \right) * \left(\frac{2N}{3} + T_{P3}\right) * \frac{2}{3} \right) + \left(\left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^{i-1} \right) * \left(2N + \frac{2N}{3} + T_{P3}\right) * \frac{1}{3} \right) + \left(\left(\sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^{i-1} \right) * (2N + T_{P3}) \right) \right)$$

We can replace some of the terms in that equation using an exponential function. This

748 simplifies the past equation to:

749 $$E[T_{BAAA}] = (1-f)$$

750 $$* \left(\left(\int_{i=0}^{T_{P3}-T_{P2}} (T_{P2}+i)\frac{1}{2N} e^{\frac{-i}{2N}} \right) + \left(\left(e^{-\frac{TP3-TP2}{2N}} \right) * \left(\frac{2N}{3} + T_{P3} \right) * \frac{2}{3} \right) \right.$$

751 $$\left. + \left(\left(e^{-\frac{TP3-TP2}{2N}} \right) * \left(2N + \frac{2N}{3} + T_{P3} \right) * \frac{1}{3} \right) \right)$$

752 $$+ f \left(\left(\left(e^{-\frac{TP3-TGF}{2N}} \right) * \left(\frac{2N}{3} + T_{P3} \right) * \frac{2}{3} \right) \right.$$

753 $$\left. + \left(\left(e^{-\frac{TP3-TGF}{2N}} \right) * \left(2N + \frac{2N}{3} + T_{P3} \right) * \frac{1}{3} \right) + \left(\left(1 - e^{-\frac{TP3-TGF}{2N}} \right) * (2N + T_{P3}) \right) \right)$$

754 And solving the integral from the first term, we get:

755 $$E[T_{BAAA}] = (1-f)$$

756 $$* \left(\left((-e^{\frac{-(TP3-TP2)}{2N}} (2N + (TP3 - TP2) + TP2) + 2\mathrm{N} + \mathrm{TP2}) \right) \right.$$

757 $$\left. + \left(\left(e^{-\frac{TP3-TP2}{2N}} \right) * \left(\frac{2N}{3} + T_{P3} \right) * \frac{2}{3} \right) + \left(\left(e^{-\frac{TP3-TP2}{2N}} \right) * \left(2N + \frac{2N}{3} + T_{P3} \right) * \frac{1}{3} \right) \right)$$

758 $$+ f \left(\left(\left(e^{-\frac{TP3-TGF}{2N}} \right) * \left(\frac{2N}{3} + T_{P3} \right) * \frac{2}{3} \right) \right.$$

759 $$\left. + \left(\left(e^{-\frac{TP3-TGF}{2N}} \right) * \left(2N + \frac{2N}{3} + T_{P3} \right) * \frac{1}{3} \right) + \left(\left(1 - e^{-\frac{TP3-TGF}{2N}} \right) * (2N + T_{P3}) \right) \right)$$

760 We can simplify this equation to get:

761 $$E[T_{BAAA}] = (1-f) * \left((2\mathrm{N} + \mathrm{T_{P2}}) + \left(\left(-e^{-\frac{T_{P3}-T_{P2}}{2N}} \right) * \left(\frac{2N}{3} \right) \right) \right)$$

762 $$+ f \left(\left(-e^{-\frac{T_{P3}-T_{GF}}{2N}} \right) * \left(\frac{2N}{3} \right) + 2N + T_{P3} \right)$$

ABAA sites

To calculate the branch lengths of the ABAA sites, we need to also calculate the contributions from six different scenarios. The calculations of the three scenarios without gene flow are equal to those of the BAAA sites:

$$(1-f)*\left(\left(\sum_{i=1}^{T_{P3}-T_{P2}}(T_{P2}+i)*\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right)\right.$$

$$+\left(\left(1-\sum_{i=1}^{T_{P3}-T_{P2}}\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right)*\left(\frac{2N}{3}+T_{P3}\right)*\frac{2}{3}\right)$$

$$\left.+\left(\left(1-\sum_{i=1}^{T_{P3}-T_{P2}}\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right)*\left(2N+\frac{2N}{3}+T_{P3}\right)*\frac{1}{3}\right)\right)$$

The contributions of the three scenarios with gene flow are:

1) There was gene flow from P_3 to P_2. The P_3 and P_2 lineages did not coalesce between times T_{P3} and T_{GF}. The lineage P_2 coalesces in the first coalescent event after T_{P3} going backwards into the past.

$$f * P(no\ coalescence\ for\ P_2\ and\ P_3\ before\ T_{GF})$$

$$* E[branch\ length\ in\ first\ coalescent\ event\ between\ lineages\ P_1, P_2\ and\ P_3]$$

$$* P(P_2\ lineage\ coalesces\ in\ first\ coalescent\ event)$$

779 $$f * \left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(\frac{2N}{3} + T_{P3}\right) * \frac{2}{3}$$

780

781 2) There was gene flow from P_3 to P_2. The P_3 and P_2 lineages did not coalesce between

782 times T_{P3} and T_{GF}. The P_2 lineage coalesces with the ancestral lineage of P_1 and P_3 in

783 the second coalescent event that takes place after T_{P3} going backwards into the past.

784

785 $$f * P(no\ coalescence\ for\ P_2\ and\ P_3\ before\ T_{GF})$$

786 $$* E\left[\begin{matrix} branch\ length\ in\ second\ coalescent\ event\ between\ lineages\ P2 \\ and\ the\ ancestral\ lineage\ of\ P_1\ and\ P_3 \end{matrix}\right]$$

787 $$* P(P_2\ lineage\ coalesces\ in\ second\ coalescent\ event)$$

788 $$f * \left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(2N + \frac{2N}{3} + T_{P3}\right) * \frac{1}{3}$$

789

790 3) There was gene flow from P3 to P2. The lineages P2 and P3 coalesce between TGF

791 and TP3.

792 $$f * \left(\sum_{i=1}^{TP3-TGF} (TGF + i) * \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right)$$

793

794 Therefore, when we put it all together, we get:

795 $$E[T_{ABAA}] = (1-f)$$

796 $$* \left(\left(\sum_{i=1}^{T_{P3}-T_{P2}} (T_{P2}+i) * \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1} \right)\right.$$

797 $$\left. + \left(\left(1 - \sum_{i=1}^{T_{P3}-T_{P2}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(\frac{2N}{3}+T_{P3}\right) * \frac{2}{3} \right)\right.$$

798 $$\left. + \left(\left(1 - \sum_{i=1}^{T_{P3}-T_{P2}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(2N+\frac{2N}{3}+T_{P3}\right) * \frac{1}{3} \right)\right)$$

799 $$+ f \left(\left(\left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(\frac{2N}{3}+T_{P3}\right) * \frac{2}{3} \right)\right.$$

800 $$\left. + \left(\left(1 - \sum_{i=1}^{T_{P3}-T_{GF}} \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1}\right) * \left(2N+\frac{2N}{3}+T_{P3}\right) * \frac{1}{3} \right)\right)$$

801 $$\left. + \left(\sum_{i=1}^{TP3-TGF} (TGF+i) * \frac{1}{2N}\left(1-\frac{1}{2N}\right)^{i-1} \right)\right)$$

802 Replacing some of the terms in that equation using an exponential function, we obtain:

803 $$E[T_{ABAA}] = (1-f)$$

804 $$*\left(\left(\int_{i=0}^{T_{P3}-T_{P2}} (T_{P2}+i)\frac{1}{2N}e^{\frac{-i}{2N}}\right)+\left(\left(e^{-\frac{TP3-TP2}{2N}}\right)*\left(\frac{2N}{3}+T_{P3}\right)*\frac{2}{3}\right)\right.$$

805 $$\left.+\left(\left(e^{-\frac{TP3-TP2}{2N}}\right)*\left(2N+\frac{2N}{3}+T_{P3}\right)*\frac{1}{3}\right)\right)$$

806 $$+f\left(\left(\left(e^{-\frac{TP3-TGF}{2N}}\right)*\left(\frac{2N}{3}+T_{P3}\right)*\frac{2}{3}\right)\right.$$

807 $$\left.+\left(\left(e^{-\frac{TP3-TGF}{2N}}\right)*\left(2N+\frac{2N}{3}+T_{P3}\right)*\frac{1}{3}\right)+\left(\int_{i=0}^{T_{P3}-T_{GF}} (T_{GF}+i)\frac{1}{2N}e^{\frac{-i}{2N}}\right)\right)$$

808 After solving the integrals, we get:

809 $$E[T_{ABAA}] = (1-f)$$

810 $$*\left(\left((-e^{\frac{-(TP3-TP2)}{2N}}(2N+(TP3-TP2)+TP2)+2\mathrm{N}+\mathrm{TP2})\right)\right.$$

811 $$\left.+\left(\left(e^{-\frac{TP3-TP2}{2N}}\right)*\left(\frac{2N}{3}+T_{P3}\right)*\frac{2}{3}\right)+\left(\left(e^{-\frac{TP3-TP2}{2N}}\right)*\left(2N+\frac{2N}{3}+T_{P3}\right)*\frac{1}{3}\right)\right)$$

812 $$+f\left(\left(\left(e^{-\frac{TP3-TGF}{2N}}\right)*\left(\frac{2N}{3}+T_{P3}\right)*\frac{2}{3}\right)\right.$$

813 $$\left.+\left(\left(e^{-\frac{TP3-TGF}{2N}}\right)*\left(2N+\frac{2N}{3}+T_{P3}\right)*\frac{1}{3}\right)\right.$$

814 $$\left.+\left((-e^{\frac{-(TP3-TGF)}{2N}}(2N+(TP3-TGF)+TGF)+2\mathrm{N}+\mathrm{TGF})\right)\right)$$

815 If we simplify this equation, we get:

816 $$E[T_{ABAA}] = (1-f)*\left((2\mathrm{N}+\mathrm{T_{P2}})+\left(\left(-e^{-\frac{T_{P3}-T_{P2}}{2N}}\right)*\left(\frac{2N}{3}\right)\right)\right)$$

817 $$+f\left(-\left(e^{-\frac{T_{P3}-T_{GF}}{2N}}\right)\frac{2N}{3}+2N+T_{GF}\right)$$