Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde

Iman Hamid[1*], Katharine Korunes[1], Sandra Beleza[2], Amy Goldberg[1*]
[1]Department of Evolutionary Anthropology, Duke University, 130 Science Drive, Durham, NC, 27708
[2]Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom

*corresponding authors:
Iman Hamid: imanhamid95@gmail.com
Amy Goldberg: amy.goldberg@duke.edu

Abstract:
Humans have undergone large migrations over the past hundreds to thousands of years, exposing ourselves to new environments and selective pressures. Yet, evidence of ongoing or recent selection in humans is difficult to detect. Many of these migrations also resulted in gene flow between previously separated populations. These recently admixed populations provide unique opportunities to study rapid evolution in humans. Developing methods based on distributions of local ancestry, we demonstrate that this sort of genetic exchange has facilitated detectable adaptation to a malaria parasite in the admixed population of Cabo Verde within the last ~20 generations. We estimate the selection coefficient is approximately 0.08, one of the highest inferred in humans. Notably, we show that this strong selection at a single locus has likely affected patterns of ancestry genome-wide, potentially biasing demographic inference. Our study provides evidence of adaptation in a human population on historical timescales.

Key words: natural selection; admixture; genetic ancestry; population genetics

Introduction:

Genetic studies have demonstrated the important role of adaptation throughout human evolution, including classic examples such as loci underlying pigmentation, and adaptation to high-altitude lifestyles and infectious disease (Norton et al., 2007; Pardis C. Sabeti et al., 2002; Fumagalli et al., 2011; Yi et al., 2010; Grossman et al., 2013; Lamason et al., 2005; Ohashi et al., 2004; Voight et al., 2006; Lachance & Tishkoff, 2013; Nielsen et al., 2007; Pickrell et al., 2009). Yet, we have a limited understanding of adaptation in human populations on historical timescales; that is, during the last tens of generations. The ongoing selective pressures shaping human genomic variation, and how quickly humans can adapt to strong selective pressures, remain unclear. Adaptation on these short timescales is of particular importance because large-scale migrations within the past thousands to hundreds of years have exposed human populations to new environments and diseases, acting as new selective pressures (Hellenthal et al., 2014; Mathias et al., 2016; G. B. Busby et al., 2016; Patin et al., 2017; Laso-Jadart et al., 2017; V. Fernandes et al., 2019; Nielsen et al., 2017).

Admixture—gene flow between previously diverged populations to form a new population with ancestry from both source populations—provides a particularly rapid opportunity for selection to act in a population by introducing alleles previously adapted in a source population into the admixed population (Huerta-Sánchez et al., 2014; Jeong et al., 2014; Norris et al., 2020; Racimo et al., 2015). Additionally, in recent human admixture, ancestry contributions from each source population are often large enough to introduce alleles at intermediate frequencies, potentially avoiding loss from drift (V. Fernandes et al., 2019; Fortes-Lima et al., 2019; Mathias et al., 2016; Ruiz-Linares et al., 2014). More generally, admixture is ubiquitous in human history (G. B. Busby et al., 2016; G. B. J. Busby et al., 2015; Hellenthal et al., 2014; Laso-Jadart et al., 2017; Moorjani et al., 2011; Patin et al., 2017; Triska et al., 2015); therefore, understanding the effects of selection in these often understudied populations is essential to the study of human evolution.

The admixture process may obscure the signals commonly used to detect selection by increasing linkage disequilibrium and changing the distribution of allele frequencies (Gravel, 2012; Lohmueller et al., 2010, 2011; Racimo et al., 2015). Further, common signatures of selection, such as deviation from neutral expectations of the allele frequency spectrum, may not be sensitive to adaptation on the scale of tens of generations (Field et al., 2016; P. C. Sabeti et al., 2006). Recent progress has given insight into human adaptation during the past few thousand years by using allele frequency trajectories from ancient DNA (Lindo et al., 2016) or the distribution of singletons in extremely large datasets (Field et al., 2016). We focus on admixed populations as an opportunity to detect rapid adaptation using modern populations and moderate sample sizes, allowing broader sets of populations to be studied. That is, within-genome ancestry patterns across multiple nearby loci may be easier to detect than single allele frequency shifts (Tang et al., 2007). Further, ancestry-based methods constrain the timing of potential selection to post-admixture, providing concrete information about the timing of selection, whereas non-ancestry-based summary statistics may detect selection in the source populations.

We test this concept developing new ancestry-based methods to characterize adaptation to malaria during the ~20 generations since the founding of the admixed human population of

Cabo Verde. The Republic of Cabo Verde is an archipelago off the coast of Senegal and was uninhabited before settlement in ~1460 by Portuguese colonizers and enslaved peoples from the Senegambian region of West Africa (Beleza et al., 2012; A. T. Fernandes et al., 2003; Korunes et al., 2020; Verdu et al., 2017), henceforth referred to as "European" and "West African" source populations, respectively. This is approximately 19-22 generations ago assuming a 25 to 28-year human generation time (Fenner, 2005). Recent analyses using genetic ancestry information alongside historical data confirmed that admixture in Cabo Verde likely began within the last ~20 generations (Korunes et al., 2020). In this study, we assume admixture occurred 20 generations ago, and we focus on three major island regions of Cabo Verde: Santiago, Fogo, and the Northwest Cluster (**Figure 1A**).

The malaria parasites *Plasmodium vivax*, *P. falciparum*, and *P. malariae* have been reported across the islands of Cabo Verde since settlement; recurrent malaria epidemics have primarily occurred in highly populated regions (DePina et al., 2019; World Health Organization et al., 2012). Santiago, which has consistently been the most densely populated of the Cabo Verde islands, has experienced the most substantial burden of malaria transmission since settlement ~20 generations ago. Personal and historical accounts of malaria incidence within Cabo Verde described the largest and most populous island, Santiago, as the most "sickly" and "malarious" (Patterson, 1988). In the last century, malaria epidemics of both *P. vivax* and *P. falciparum* have occurred primarily in Santiago (DePina et al., 2018; Ferreira et al., 2017; Snow et al., 2012; World Health Organization et al., 2012). It is not fully understood why Santiago has sustained a higher burden of malaria than the other Cabo Verdean islands (World Health Organization et al., 2012); however, it may be due to a combination of higher population density, climatic differences between islands, increased migration into Santiago, which has historically served as the main trading port for Cabo Verde, and the suitability of the island for the mosquito vector. The other two island regions we consider share ancestry components with Santiago, but largely lacked the selective pressure of recurrent malaria transmission, providing a unique opportunity to compare related populations with and without malaria as a selective pressure.

We hypothesized that admixture has facilitated rapid adaptation to the malaria parasite *Plasmodium vivax* via the malaria-protective *DARC* locus (*Duffy antigen receptor for chemokines*; also known as *Atypical Chemokine Receptor 1* (*ACKR1*)) in Santiago. The protective allele is almost fixed in West African populations and rare elsewhere (Gething et al., 2012; Howes et al., 2011). The malaria parasite *P. vivax* uses the chemokine receptor encoded by the *DARC* gene to enter and infect red blood cells. The Duffy-null allele (also known as FY*O, rs2814778) is protective against *P. vivax* infection via a SNP that disrupts binding of an erythroid-specific transcription factor in the promoter region (Gething et al., 2012; Mercereau-Puijalon & Ménard, 2010). Thus, individuals carrying the null allele have reduced expression of Duffy antigens on the surface of the blood cell, protecting against *P. vivax* infection. Duffy-null is a classic example of strong selection in the human lineage, and it has been estimated to be under one of the strongest selective pressures in human history (McManus et al., 2017; Kwiatkowski, 2005; Hamblin et al., 2002; Hamblin & Di Rienzo, 2000), suggesting it is a plausible selective pressure in Cabo Verde.

Interestingly, this hypothesis goes back to a voyage in 1721 in which Captain George Roberts reported that a disease in Santiago is "dangerous to strangers" during the rainy season

(Roberts, 1745). Consistent with ancestry-mediated protection from malaria, the record has been interpreted by medical historians to suggest that "foreign visitors and residents of European descent seem to have suffered more than the African and Afro-Portuguese majority" from malaria in Santiago (Patterson, 1988).

In this study, we combine ancestry-based summary statistics and simulations to identify and characterize selection at the malaria-protective *DARC* locus on the island of Santiago during the ~20 generations since the onset of admixture. Importantly, we also consider the genome-wide consequences of this mode of selection. That is, we find that strong selection at a single locus may shift genome-wide ancestry patterns, with potential to bias demographic inference. The results of this study provide evidence for rapid adaptation in human populations and advance our ability to detect and characterize selection in recently admixed populations.

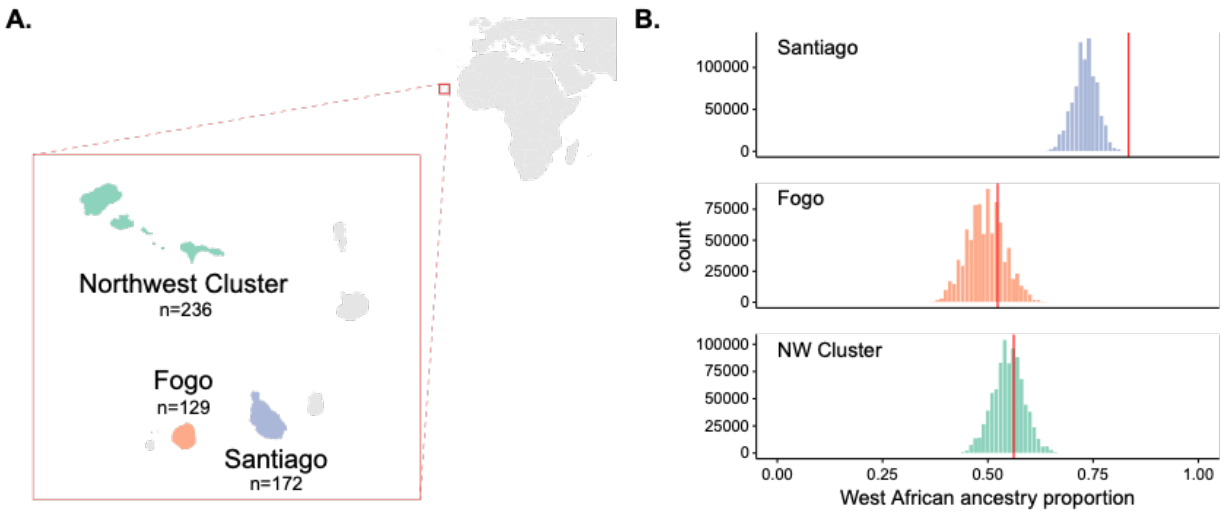

Figure 1. Enrichment of West African ancestry at *DARC* locus in Santiago, Cabo Verde. **A)** Map of Cabo Verde islands, and sample sizes for number of individuals from each island region. **B)** The distribution of West African-related local ancestry proportion across the genome by SNP (n=881,279) by island, with the *DARC* locus marked by vertical red lines. Local ancestry was estimated using RFMix (see **Methods**). The *DARC* locus is an outlier for high West African-related ancestry in Santiago, but not Fogo or the Northwest cluster.

Results:

Enrichment of West African Ancestry at the *DARC* locus on Santiago

Empirical studies of selection in admixed populations often look for regions of the genome that deviate from genome-wide patterns of genetic ancestry (G. Busby et al., 2017; V. Fernandes et al., 2019; Jeong et al., 2014; Jin et al., 2012; Laso-Jadart et al., 2017, p.; Lopez et al., 2019; Norris et al., 2019; Patin et al., 2017; Rishishwar et al., 2015; Tang et al., 2007; Triska et al., 2015; Vicuña et al., 2020; Zhou et al., 2016). Regions of the genome with substantially higher ancestry from one source than present on average in the rest of the genome are hypothesized to be enriched for genes under selection. For an allele at different frequencies in the source populations, selection will increase the frequency of the ancestry on which the adaptive allele occurs at that locus. We tested if the *DARC* locus was an outlier within the genome for

West African ancestry. We estimated local ancestry using the RFMix software (Maples et al., 2013) and calculated the proportion of individuals with West African ancestry at each SNP (see **Methods** for details on local ancestry assignment). **Figure 1B** and **Figure 1-figure supplement 1** show the distribution of West African ancestry for over 880,000 SNPs by island, with the value for the Duffy-null SNP position marked in red. Within Santiago, this locus has the highest frequency of West African ancestry in the population, occurring at 0.834 frequency compared to mean West African ancestry across SNPs for individuals from Santiago of 0.730. In contrast, the *DARC* locus is not an outlier in its frequency of West African ancestry on the other island regions, occurring at the 75th and 65th percentiles for Fogo and the NW Cluster, respectively. High West African local ancestry proportion at the *DARC* locus in Santiago is consistent with the expectation that the Duffy-null allele rapidly increased in frequency following admixture, simultaneously increasing the proportion of individuals with West African ancestry at that locus relative to the genome-wide average.

Under a simple population genetic model, we expect the frequency of a neutral allele in an admixed population to be a linear combination of the allele frequencies in each source population and their relative ancestry contributions. The Duffy-null allele is nearly fixed in the West African source population and largely absent in the European source population; therefore, under neutrality, the expected frequency of the allele in each Cabo Verdean population is approximately equal to the West African ancestry contribution. Using the observed global ancestry proportion inferred with ADMIXTURE (Alexander & Lange, 2011) as an estimate of the ancestry contribution from West Africa to each island, we found that the Duffy-null allele is at a higher frequency than expected under neutrality for the island of Santiago, but not the other regions of Cabo Verde (**Figure 1-figure supplement 2, Table 1**, binomial test, Santiago: $p = 2.193\times10^{-5}$; Fogo: $p = 0.1915$; NW Cluster: $p = 0.8172$).

Long, high-frequency West African ancestry tracts span the *DARC* locus on Santiago

The distribution of the lengths of ancestry tracts spanning a selected locus can provide information for detecting and characterizing selection beyond single-locus outlier tests. In the case of recent admixture and strong selection, we might generally expect to see a parallel increase in local ancestry proportion in the regions surrounding the beneficial locus because the beneficial allele increases before recombination can break up large surrounding ancestry blocks. This is analogous to the increase in linkage disequilibrium and homozygosity in non-admixed populations (Y. Kim & Nielsen, 2004; Pardis C. Sabeti et al., 2002; Voight et al., 2006). **Figure 2A** plots ancestry tracts that span the *DARC* locus for individuals from Santiago. As expected from source population allele frequencies, the Duffy-null allele is contained on all West African ancestry tracts spanning the locus and not found on any European ancestry tracts. Consistent with recent selection, West African ancestry tracts are longer and in higher frequency than European ancestry tracts covering the region. The median West African ancestry tract length spanning the locus is ~85 Mb, while the median European tract length is ~39 Mb.

In order to test if the observed local ancestry patterns are suggestive of selection beyond genome-wide ancestry proportion, we developed a summary statistic based on the length and frequency of the tract-length surrounding a locus. The integrated Decay in Ancestry Tract (*iDAT*) score compares the rate of decay of ancestry tract lengths as a function of distance from a site of

interest (**Figure 2B**; see **Methods** for details and performance of *iDAT* statistic under various demographic scenarios). The statistic is analogous to the commonly used integrated haplotype score (*iHS*) (Voight et al., 2006), but considers local ancestry tracts instead of haplotypes. Negative values for *iDAT* indicate longer West African ancestry tracts at higher frequencies compared to European ancestry tracts. Positive values indicate longer European ancestry tracts at higher frequencies compared to West African ancestry tracts. Windows that contain multiple extreme values of *iDAT* provide stronger evidence for recent selection. **Figure 2C** plots *iDAT* values along the genome for individuals from Santiago. Values are calculated by averaging over 20 Mb sliding windows (step size of 1 Mb) for 10000 random standardized *iDAT* scores. Notably, in Santiago, the *DARC* gene is contained in the window with the lowest *iDAT* score in the genome (**Figure 2C**; window *iDAT* score = -2.602). *iDAT* scores near the *DARC* locus are not outliers in other islands regions (**Figure 2-figure supplement 1**).

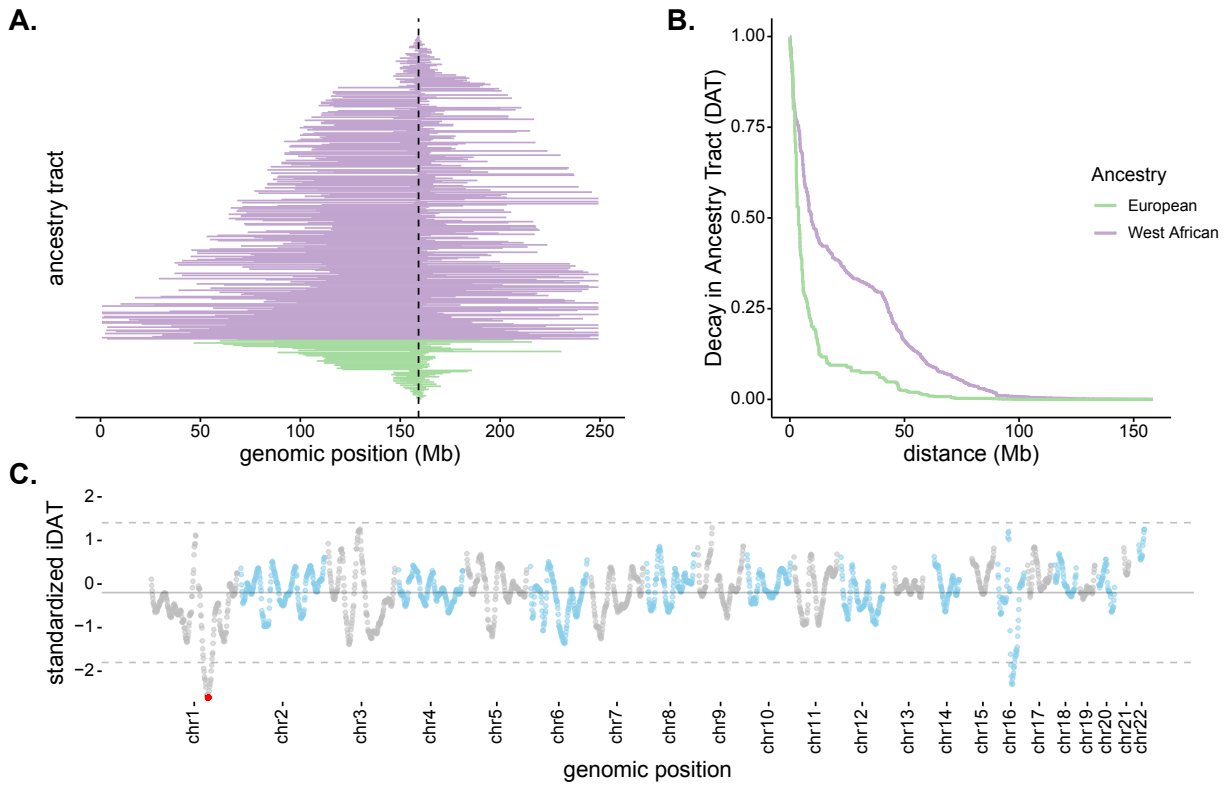

Figure 2. Long, high-frequency West African ancestry tracts span the DARC locus in Santiago. **A)** The distribution of West African (purple) and European (green) ancestry tract lengths spanning the DARC locus (dashed line). Each horizontal line represents a single chromosome in the population (n=343, one chromosome was excluded due to having unknown ancestry at the DARC locus). **B)** Decay in Ancestry Tract (DAT) as function of absolute distance from the Duffy-null allele for West African (purple) and European (green) ancestry tracts. **C)** Mean standardized integrated DAT (iDAT) score for 20 Mb sliding windows (step size = 1Mb), using standardized iDAT for 10000 random positions across the genome. Horizontal solid grey line indicates mean windowed standardized iDAT score (-0.196), and horizontal dashed grey lines indicates three standard deviations from the mean windowed score. The red dot is the most extreme windowed standardized iDAT score (-2.602), indicative of a larger area under the curve for West African DAT compared to European DAT. This 20 Mb window contains the Duffy-null SNP.

Ancestry-based signatures for Santiago cannot be explained by drift alone

In order to estimate the expected distribution of ancestry within the population and test if the values of various summary statistics for the *DARC* locus on Santiago can be explained by drift alone, we conducted neutral simulations in SLiM (Haller & Messer, 2019) (**Methods**). We calculated the following five summary statistics for each simulated population: the West African local ancestry proportion at *DARC*, the variance in the frequency of West African local ancestry across SNPs on chromosome 1, the median and mean West African ancestry tract length containing the Duffy-null allele, and the unstandardized *iDAT* score for the Duffy-null SNP. The variance in local ancestry along the chromosome provides a non-LD-based measure to capture high frequency and long tracts of West African ancestry using the population-wide measures of local ancestry for each SNP on the chromosome. Studies of demographic history and selection in recently admixed populations often assume constant population size and a single admixture event to simplify simulations. In order to confirm that assumptions about demographic history do not change our expectations, we considered multiple scenarios of population growth, differences in population size, and models of both constant contributions and single admixture events (**Table 2**). The values for the summary statistics for Santiago generally lie outside our expectations for all models, especially considered jointly (**Figure 2-figure supplement 2**).

Together, these summary statistics provide suggestive evidence that the *DARC* locus has been under positive selection on the island of Santiago since admixture started ~20 generations ago. To formally test this hypothesis, we extended the SWIF(r) framework developed by (Sugden et al., 2018). SWIF(r) is a machine learning classification framework that explicitly learns the joint distributions for a set of features and returns a posterior probability of positive selection at a site of interest. It is particularly useful for handling summary statistics that are correlated, such as the length and frequency of ancestry tracts. We trained SWIF(r) using data simulated in SLiM and estimated the posterior probability of positive selection at the *DARC* locus using the five ancestry-based measures (**Methods**). SWIF(r) returned a high posterior probability of positive selection at *DARC* on Santiago starting 20 generations ago ($P > 0.999$).

Classical haplotype-based signatures of selection not detected at *DARC* locus

The haplotype-based statistic, *iHS*, is often used to detect signatures of recent positive selection and partial selective sweeps (Voight et al., 2006), particularly in non-admixed populations. This statistic has been used as evidence of selection in recently admixed populations (V. Fernandes et al., 2019; Norris et al., 2020; Reynolds et al., 2019). However, the process of admixture results in the mixture of differentiated allele frequencies and diverged haplotypes, so interpretation of these statistics is difficult and applicability is limited. We demonstrate this by calculating *iHS* for all SNPs in our dataset for each island region and performing the common standardization based on allele frequencies, using the software *hapbin* (Maclean et al., 2015). **Figure 3** shows the distribution of absolute standardized *iHS* values along the genome for each island population, with the Duffy-null SNP indicated by the orange dot and flag. The absolute *iHS* values for all islands at the Duffy-null SNP are low. That is, the commonly-used statistic *iHS* does not detect significant signatures of selection at the Duffy-null SNP position. This analysis, and other summaries of variation that do not account for the allele frequency and LD changes

associated with admixture, may be detecting the high diversity in the African source populations rather than post-admixture selection. Without considering the process of admixture, we should be skeptical of the utility of these statistics in recently admixed populations. This emphasizes the importance of new methods that are admixture-aware.

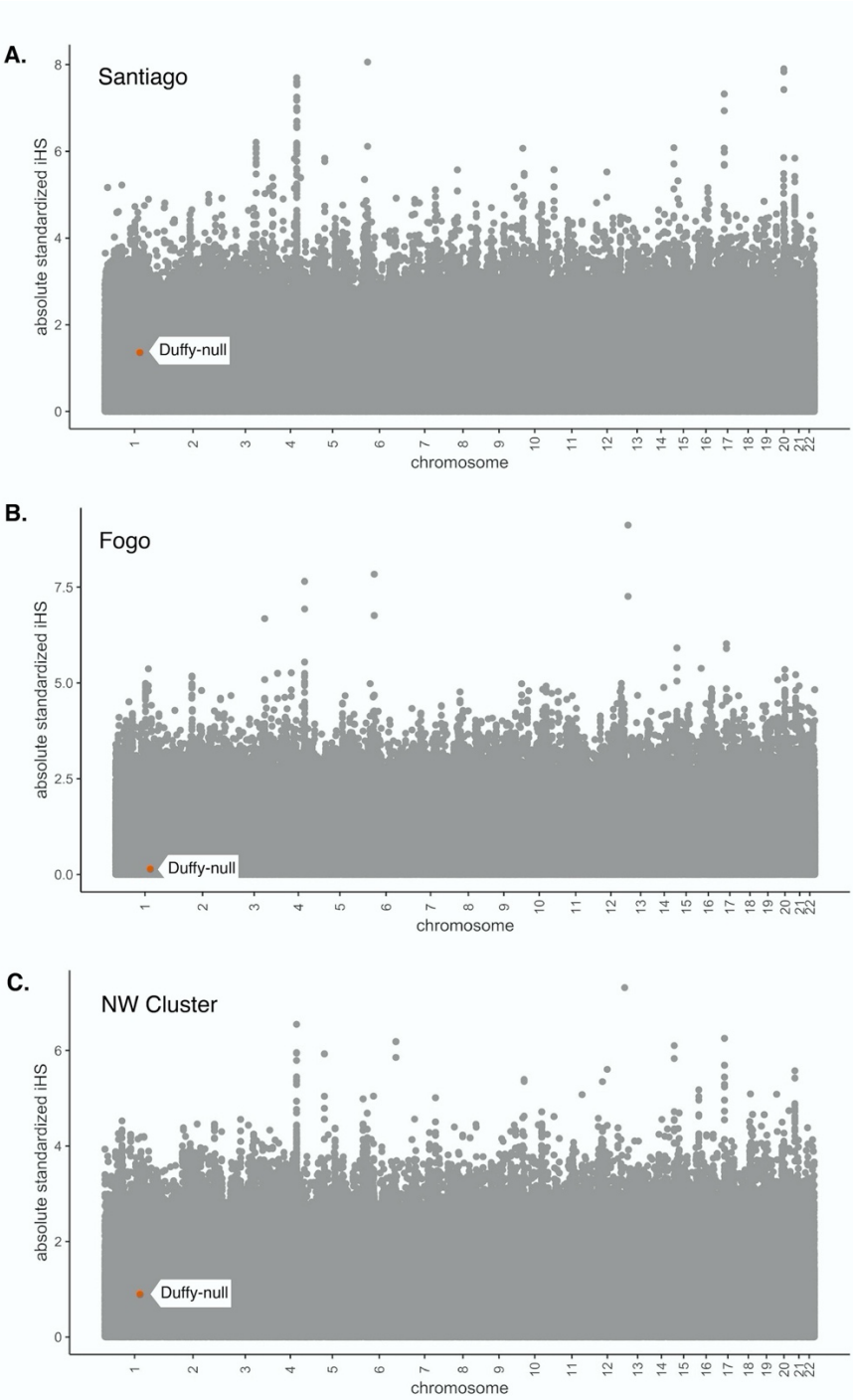

Figure 3. Absolute values of *iHS* for SNPs in the Cabo Verde dataset. *iHS* was calculated using the *hapbin* software, and standardized using the default method based on allele frequencies. **A)** Santiago, **B)** Fogo, and **C)** NW Cluster.

Value for Duffy-null SNP is indicated by orange dot and white label. Duffy-null *iHS* value is non-significant in all island regions.

Strong selection inferred at *DARC* locus in Santiago

Beyond identifying selection, inference of the strength of selection is informative about the evolutionary processes shaping human genomes. We used two complementary approaches to infer the strength of selection at the *DARC* locus. First, we considered a deterministic classical population-genetic model of selection based on the trajectory of allele frequencies over time on a grid of possible dominance and selection coefficients (**Methods**). The estimate of the selection coefficient depends on dominance; past studies have modeled Duffy-null as recessive (Hodgson et al., 2014), dominant (Pierron et al., 2018), and additive (McManus et al., 2017) when estimating selection strength in other human populations. **Figure 4A** plots the selection strength (s) as a function of the dominance coefficient (h) of the Duffy-null allele for a set of three realistic initial frequencies, assuming 20 generations of constant selection strength. Functional studies suggest that heterozygotes have at least partial protection against *P. vivax* infection (Cavasini et al., 2007; Gething et al., 2012; Kano et al., 2018; Sousa et al., 2007); while not an exact correlate for population-genetic model parameters, this suggests that the Duffy-null allele is unlikely to be fully recessive or fully dominant. Taking the mean of selection coefficients for $0.2 \leq h \leq 0.8$, we estimate the selection coefficient for each initial frequency, $s_{0.65} = 0.106$, $s_{0.70} = 0.082$, and $s_{0.75} = 0.056$, where s_{p} is the inferred selection coefficient for initial allele frequency p_{o}.

Second, we used a simulation and rejection framework, approximate Bayesian computation (ABC), to jointly infer the selection coefficient and initial West African contribution while allowing for drift (**Methods**). We used the five ancestry-based summary statistics described previously. We assumed an additive model, a single admixture event, and exponential growth in the population. Taking the median of the posterior distribution as the point estimate for selection coefficient, we estimated $s = 0.0795$ (**Figure 4B,** see **Figure 4-figure supplement 1** for estimates of s when modeling Duffy-null as either a dominant or a recessive mutation). This estimate of selection coefficient is consistent with those estimated under the deterministic population-genetic model.

Selection at a single locus impacts genome-wide ancestry estimates

Mean global ancestry proportion is often used as an estimate for initial ancestry contributions for admixed populations (Bryc et al., 2015; V. Fernandes et al., 2019; Hellenthal et al., 2014; Laso-Jadart et al., 2017; Mathias et al., 2016; Moreno-Estrada et al., 2013; Patin et al., 2017). However, our ABC estimates of the initial contributions from West Africa are lower than the mean ancestry currently observed in Santiago. The median of the posterior of initial contributions from West Africa is 0.690, with the middle fifty percentile of observed values in [0.682,0.697] (**Figure 5A**). In contrast, the observed mean ancestry in Santiago (**Figure 1B**) is 0.737. While this particular difference in observed and inferred founding contributions may be due to sampling biases or other neutral processes, it raises the more general question of how strong selection at a single locus may impact genome-wide ancestry patterns. We hypothesized

that selection at *DARC* may have increased the genome-wide West African ancestry proportions in the current population of Santiago.

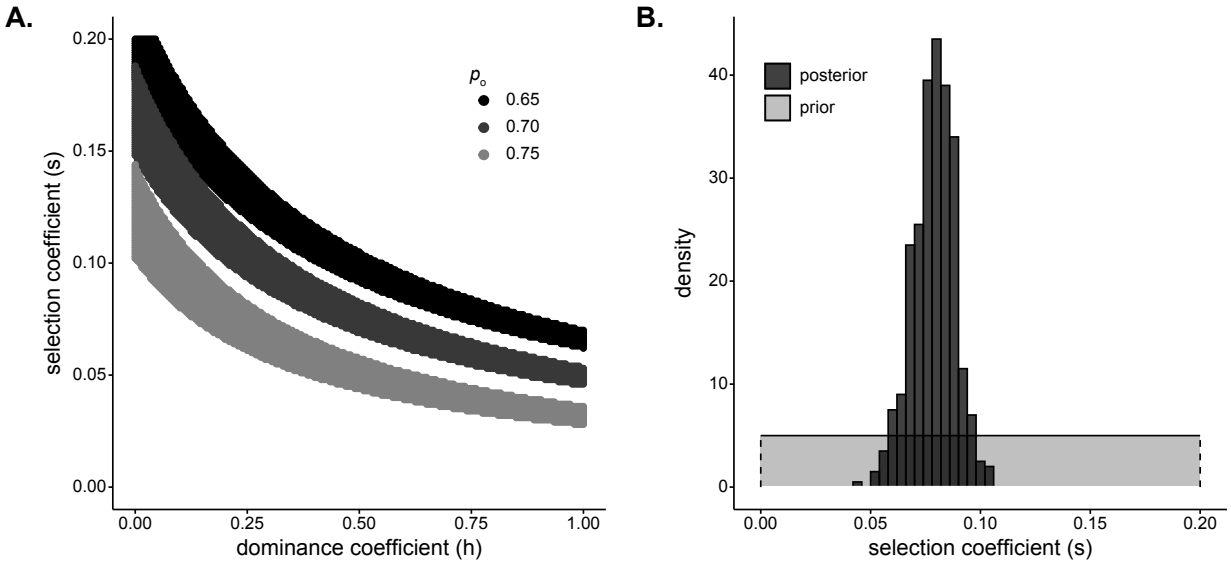

Figure 4. Strong selection inferred at *DARC* locus in Santiago. **A)** Pairs of *s* and *h* that result in a small difference in final allele frequency calculated under the model and the allele frequency observed in the Santiago genetic data, $\left|p_{20} - p_{\text{Duffy}}\right| < 0.01$ under a deterministic population genetic model. Colors indicate the initial Duffy-null frequency: p_o=0.65, black; p_o=0.70, dark grey; p_o=0.75, light grey. **B)** ABC estimates of the selection coefficient for Duffy-null on Santiago. Shaded grey area shows prior distribution of selection coefficient [$s \sim U(0, 0.2)$]. Dark grey histogram shows posterior distribution for selection coefficient (median = 0.0795), constructed from regression-adjusted values from accepted simulations.

To test the genomic consequences of post-admixture selection at a single locus, we simulated whole human autosomes under a model of exponential growth and a single admixture event with selection at a single locus. We first considered a model based on the history of Santiago, using the posterior distributions of selection coefficient and initial West African ancestry contribution as parameters for the simulations (**Methods**). **Figure 5A** plots the estimated posterior distribution of initial West African contribution in dark grey and the simulated distribution of global ancestry after 20 generations in white. The distribution of global ancestry in the populations simulated with selection (median 0.723, white) is noticeably higher than the initial contributions specified in the simulations (median 0.690, dark grey). This demonstrates that selection at a single locus is a plausible mechanism to increase mean global ancestry in an admixed population under a scenario similar to Santiago.

To explore the mechanism and relationship between selection strength at a single locus and genome-wide ancestry patterns, we simulated whole autosomes, assuming a single admixture event with initial West African ancestry contribution at 0.65 and selection coefficient varying from 0 to 0.2 at a single locus on chromosome 1. **Figure 5B** plots the mean ancestry for chr1 and the other 21 autosomes in each simulated population after 20 generations as a function of the selection coefficient at a single locus. Perhaps surprisingly, mean ancestry on chromosomes 2-22 also increases with selection strength (grey), indicating that global ancestry increases beyond the contribution of higher ancestry on the selected chromosome alone (black).

Together, this evidence suggests that strong selection at the *DARC* locus over 20 generations may have skewed global ancestry in Santiago and raises potential biases with a statistic that is often used to infer neutral demographic histories.

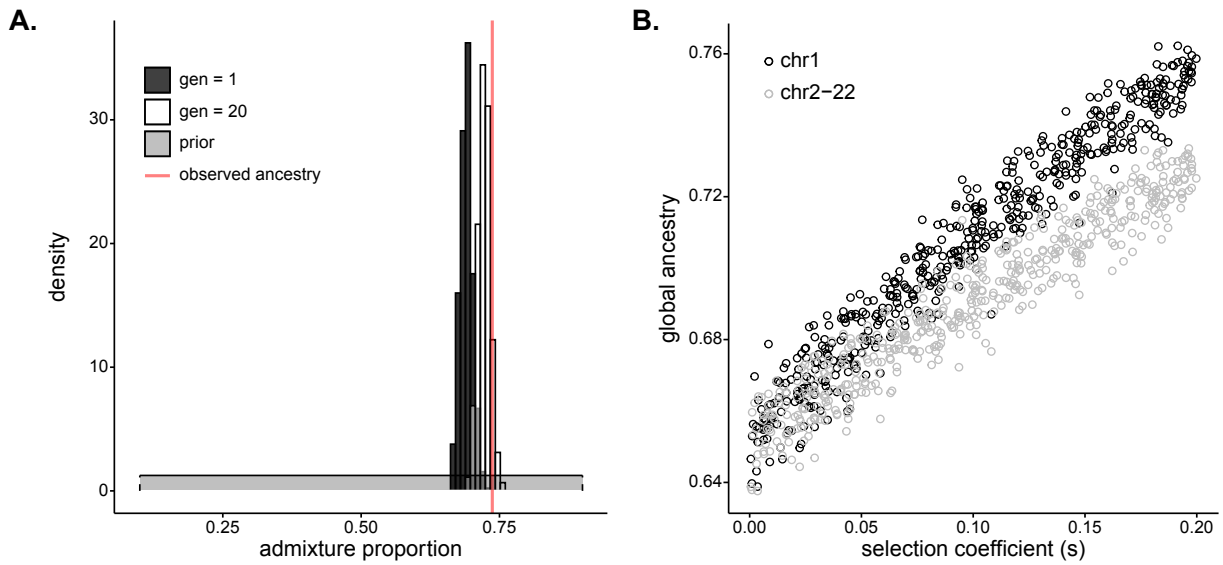

Figure 5. Selection at a single locus impacts genome-wide ancestry proportion. **A)** Inferred (dark grey), simulated (white), and observed (red) mean of global ancestry in Santiago over time. The dark grey histogram plots the posterior distribution for initial ($g = 1$) West African ancestry contribution inferred using ABC (median, 0.690); the prior distribution [$m \sim U(0.1, 0.9)$] is in light grey. The red line plots the mean global ancestry estimated by ADMIXTURE from modern genetic data from Santiago, 0.737. The observed global ancestry is higher than most values of the initial contributions inferred in dark grey. The white histogram plots the distribution of West African global ancestry proportion calculated after 20 generations in populations simulated with selection coefficients and initial ancestries drawn from the ABC-inferred values (median, 0.723). The global ancestry calculated after 20 generations of simulated selection (white) more closely matches that observed from Santiago genetic data (red line). **B)** West African mean global ancestry proportion calculated for 500 simulated populations after 20 generations under varying single-locus selection coefficients, *s.* We simulated whole autosomes, setting the initial West African ancestry contribution to 0.65. Black circles indicate mean ancestry on chromosome 1 alone. Grey circles indicate mean ancestry on the other autosomes (2-22). The increase in ancestry with selection for grey circles demonstrate that selection impacts global ancestry beyond the local effects of the chromosome under selection.

Discussion:

Using adaptation to malaria in the admixed population of Cabo Verde as a case study, we have demonstrated that admixture can facilitate adaptation in merely tens of generations in humans. Developing methods to identify and characterize post-admixture selection, we found that this rapid adaptation leaves detectable genomic signatures (**Figures 1 & 2**), with potential genome-wide consequences (**Figure 5**). Combining inference under two complementary methods, and under a range of possible dominance coefficients and initial allele frequencies, we estimated selection strength of $s \approx 0.08$ for the Duffy-null allele in Santiago (**Figure 4**). Our estimate is consistent with other studies which have inferred the strength of selection for Duffy-null ranging from ~0.04 (modeled under additive selection) in sub-Saharan African populations (McManus et al., 2017) and ~0.07 (modeled as recessive) to ~0.2 (modeled as dominant) in a Malagasy population with admixed African ancestry (Hodgson et al., 2014; Pierron et al., 2018).

Our estimated strength of selection for Duffy-null is among the highest inferred for a locus in any human population.

Introgression of an adaptive allele can facilitate adaptation on short timescales, particularly for traits with large effects from single loci. When ancestry contributions from multiple sources are high, such as is common in recent human admixture, selection post-admixture can be a faster mode of adaptation, similar to selection on standing variation (Hedrick, 2013; Hermisson & Pennings, 2005).

Commonly-used ancestry outlier approaches have identified candidate regions for admixture-enabled adaptation in many populations; for example, at the *HLA* and *LCT* regions in Bantu speaking populations (Patin et al., 2017), the *MHC* locus in Mexicans (Zhou et al., 2016), and adaptation to high-altitude at *EGLN1* and *EPAS1* in Tibetans (Jeong et al., 2014). However, results from this framework alone can be difficult to interpret because drift post-admixture may substantially change allele frequencies and the distribution of local ancestry within and between individuals (Belbin et al., 2018; Bhatia et al., 2014; Calfee et al., 2020). Further, recessive deleterious variation masked by heterosis can similarly cause a signal of increased introgressed ancestry, especially in regions of low recombination (B. Y. Kim et al., 2018). As a result, outlier approaches may have increased rates of false-positive detection of regions under selection. One recent approach modeled local ancestry deviations based on individual-level global ancestry distributions; however, determining a significance threshold for local ancestry deviations remains difficult (G. Busby et al., 2017). Outlier approaches also discard local haplotype information and are not informative about the selection strength or timing. Instead, we developed a suite of methods to identify and characterize selection post-admixture: the *iDAT* summary statistic, application of the SWIF(r) framework to estimate the probability a locus is under selection in an admixed population, and an ABC framework to jointly infer selection strength and initial admixture contributions (**Methods**).

Methods to detect adaptation driven by alleles introduced through gene flow in human populations have typically focused on ancient admixture between highly diverged populations, often with small contributions from one of the sources (Jagoda et al., 2018; Racimo et al., 2017; Setter et al., 2020). Recent advances leverage patterns of ancestry to consider recent admixture, though perform best for events at least hundreds of generations ago (Svedberg et al., 2020). Instead, we emphasize admixed populations as a model for adaptation on historical timescales, with selection dramatically changing genomic variation within tens of generations. This timescale is important for elucidating human history and has implications for conservation genetics and ecology in other organisms. Additionally, our summary statistic approach can be flexibly applied in a variety of inference methods. For example, our implementation in a likelihood-free ABC framework allows for flexible population history models fit to the population of interest. This approach moves beyond identification of loci under selection, allowing joint inference of selection and population history parameters.

We apply these methods to characterize post-admixture adaptation in the Cabo Verdean island of Santiago. The Cabo Verdean populations have a number of advantages for identifying and interpreting selection over the last ~500 years. First, the island geography minimizes within-population structure and provides comparison island populations with shared ancestry

components to partially account for demography. Second, historical records give a clear boundary for the earliest onset of selection in the admixed population, based on the initial occupation and admixture in the 1460s (~20 generations in the past). Further, the European and West African source populations have high levels of genetic divergence for human populations, improving local ancestry assignment accuracy. Though errors in phasing or local ancestry assignments are possible and should be considered if applied to other scenarios, it is unlikely that such errors would create signatures as extreme and long-ranging as we observe in Santiago.

We inferred that the Duffy-null allele rapidly increased in frequency after admixture as a result of its adaptive resistance to *P. vivax* infection. While we consider this to be the strongest candidate locus, given the large ancestry tracts, it is possible that selection at other nearby loci is responsible for the observed ancestry patterns. The Duffy-null allele shows extreme geographic differentiation, being nearly fixed in sub-Saharan African populations and mostly absent in non-African populations (Gething et al., 2012; Howes et al., 2011; Mercereau-Puijalon & Ménard, 2010), and multiple populations with ancestry from sub-Saharan Africa show evidence for admixture-enabled adaptation throughout human history at the *DARC* locus (G. Busby et al., 2017; V. Fernandes et al., 2019; Hodgson et al., 2014; Laso-Jadart et al., 2017; Pierron et al., 2018; Triska et al., 2015). Further, the Duffy-null allele has well-characterized functional protection against *P. vivax*, a malaria parasite with a documented record of recurrent transmission in Santiago since settlement.

We demonstrated how selection at the *DARC* locus may have affected patterns of ancestry genome-wide (**Figure 5**). The initial admixture contributions inferred under our model were lower than those observed in Santiago today, and we confirmed this pattern more generally in simulations. Population-genetic studies often use loci far from potentially selected sites as putatively neutrally-evolving loci, or treat many dispersed loci as neutral based on the assumption that a few selected sites will not dramatically change genome-wide distributions of summary statistics. However, we found that selection on a single locus may shape patterns of ancestry genome-wide. On short timescales, individuals with higher genome-wide proportions of ancestry from the source population carrying the beneficial allele are more likely to have a selective advantage in early generations post-admixture as recombination has yet to uncouple global ancestry proportion and the local ancestry of the selected allele (Pierron et al., 2018). Therefore, since global ancestry is often used to infer initial admixture contributions, these and related demographic inferences may be biased under the common assumption that genome-wide patterns of ancestry reflect demography alone.

Difference in observed and simulated global ancestry may be caused by a variety of statistical and evolutionary processes beyond selection, including sampling, estimation method, demographic model misspecification, different rates of migration over time, or drift. Regions aside from *DARC* may also have been under selection post-admixture in Santiago, and therefore affected the global ancestry patterns. For example, we also observed a cluster of extreme negative *iDAT* values on chromosome 16 [~chr16: 48000000-60000000] that may be of interest for future study.

The 10 annotated genes in this region and associated gene ontology terms are included in **Supplementary File 1**. A cursory literature search returned no known associations with malarial

response in this region, and it is more difficult to draw conclusions on selection history in this region without a prior hypothesis. That said, we note that this region does not show high proportions of West African ancestry compared to the genome-wide distribution (**Figure 1-figure supplement 1**). The extreme *iDAT* values in this region may be influenced by its adjacence to the centromere, which may affect the length of ancestry tracts.

Generally, we suggest that *iDAT* should be used as one line of evidence alongside other summaries of variation, such as those used in our ABC estimation and expected allele frequency calculations. Moreover, in the case of Duffy-null, we had a strong prior expectation for positive selection for the West African haplotype. Analogous to biallelic selection scenarios, with admixture between two source populations, it is difficult to distinguish between positive selection for one ancestry and negative selection for the other ancestry.

Another important limitation for ancestry-based methods of detecting selection in admixed populations is that they are best suited for scenarios wherein frequencies of a selected allele differ greatly between source populations, as is the case with Duffy-null. If an allele is present at similar frequencies in the source populations (i.e. regions of low *Fst* between source populations), selection will likely affect both ancestries in the admixed population similarly. Future studies into the adaptive histories of admixed populations should consider this limitation on the pool of potentially adaptive variants that can be detected using ancestry-based analyses.

The framework developed here is broadly applicable to detect and characterize selection in other recently admixed human and non-human populations. For most analyses, we assumed a single admixture event 20 generations in the past. The simulation software we use, SLiM, makes it straightforward to consider other models of population history specific to populations of interest for future studies. Further research into the interaction of selection and demography will refine inference. Additionally, while the timing of admixture in Cabo Verde is well documented, our ABC approach may be extended to infer the timing of selection and admixture as well, as the tract length distribution is informative about these parameters (Gravel, 2012; Korunes et al., 2020; Liang & Nielsen, 2014). Indeed, we demonstrated the need to jointly infer selection and demographic histories.

Code & data availability:

Scripts for analyses, simulations, and to reproduce figures can be found at https://github.com/agoldberglab/CV_DuffySelection. Sampling consent forms do not allow for public release of genotype data. Inferred local ancestry information can be found at https://doi.org/10.5281/zenodo.4021277.

Acknowledgements:

We thank Joshua Schraiber and Kelley Harris for useful discussions. We thank Hua Tang and Greg Barsh for generating genetic data used in this study. We acknowledge support from NIH grant R35 GM133481 to AG and NIH NIGMS grant F32 GM139313 awarded to KLK.

Methods:

Genetic data & ancestry inference

For this study, we used SNP array data from (Beleza et al., 2013) which included 564 admixed individuals across the island regions of Cabo Verde. From this dataset, we filtered individuals with greater than 5% missing calls overall or greater than 10% missing calls on a single chromosome. This resulted in removal of one individual with high missingness on chromosome 14 (11.36%). We merged genotypes for the remaining 563 individuals with genotypes from 107 IBS (Iberian Population in Spain) and 107 GWD (Gambian in Western Division - Mandinka) samples from high-coverage resequencing data released through the International Genome Sample Resource (Clarke et al., 2017; Fairley et al., 2020). Our analyses considered autosomal chromosomes only. We selected biallelic SNPs occurring in both the Cabo Verde samples and the reference samples. The final merged dataset contained 884,656 autosomal SNPs. Average missingness by SNP was 0.0017.

Using the 884,656 autosomal SNP dataset, we performed phasing with SHAPEIT2 using the Phase 3, NCBI build 37 (hg19) reference panel of haplotypes and associated genetic map in IMPUTE2 format (Delaneau et al., 2013). Following the SHAPEIT documentation, we first ran SHAPEIT -check to exclude sites not contained within the reference map, followed by SHAPEIT phasing to yield phased genotypes at 881,279 SNPs. For local ancestry inference, we ran RFMix v1.5 on the phased samples using a two-way admixture model (Maples et al., 2013). We used the RFMix PopPhased program with default window size, the --use-reference-panels-in-EM option, -e = 2 (2 EM iterations), and --forward-backward. Ancestry references for European and West African source populations, respectively, were IBS and GWD.

We observed a low overall proportion of regions within a given individual's genome assigned as "unknown" local ancestry by RFMix. The mean and median genome-wide proportion of unknown ancestry in our dataset were 0.0089 and 0.0083, respectively. Typically, more than 99% of each individual's genome could be assigned as either West African or European ancestry. Coupled with the fact that there is little historical or genetic evidence of migration from non-West African or non-European populations to Cabo Verde (Carreira, 1983; Korunes et al., 2020; Verdu et al., 2017), we find it is unlikely that individuals in our dataset have high proportions of recent ancestry from a non-European or non-African source population.

To validate our phase imputation and RFMix local ancestry calls, we also performed local ancestry assignment using a second method, ELAI, which performs its own independent phasing prior to calling local ancestry (Guan, 2014). We ran ELAI under a two-way admixture model, again using IBS and GWD genotypes as references for the source populations. We set the parameters -mg (number of generations) to 20, -s (EM steps) to 30, -C (upper clusters) to 2, and -c (lower clusters) to 10, based on the 5 x C recommendation from the ELAI documentation. The estimates of each individual's average genome-wide ancestry are highly correlated between ELAI and RFMix (Pearson's $R = 0.9964$; $p < 1\times10^{-8}$).

Global ancestry inferred in ADMIXTURE (Alexander & Lange, 2011) was averaged over ten independent runs with randomly chosen seeds using supervised cluster mode with the GWD

and IBS individuals specified as the reference populations. Our estimates of global ancestry by individual from ADMIXTURE are consistent with those of RFMix (Pearson's $R = 0.9973$; $p < 1\times10^{-8}$).

The island of Boa Vista was excluded from analyses due to our small sample from the region (26 individuals), leaving a final set of 537 individuals across the three island regions considered in this study (Santiago: 172, Fogo: 129, NW Cluster: 236; **Figure 1A**).

Local ancestry calls can be found at https://doi.org/10.5281/zenodo.4021277.

Integrated Decay in Ancestry Tract (*iDAT*) score

In order to account for global ancestry patterns that contribute to the ancestry tract-length distribution when identifying loci under selection post-admixture, we developed the integrated Decay in Ancestry Tract (*iDAT*) score. This statistic quantifies the length and homozygosity of the tract-length distribution by comparing the decay of tract lengths from alternate ancestries at increasing distance from a site of interest.

First, we describe the Decay in Ancestry Tract (*DAT*) feature, which is calculated similarly to expected haplotype homozygosity (*EHH*) (Pardis C. Sabeti et al., 2002), using local ancestry tracts rather than haplotypes. For each source population *i* we calculate:

$$DAT_i(x) = \left(\frac{n_x}{n}\right)^2,$$

where n_x is the number of ancestry tracts that extend some absolute distance x from a position of interest, and n is the total number of ancestry tracts (extending in either direction) that contain the site of interest. We calculate *DAT* at increasing distances from the site of interest. Similar to integrated haplotype homozygosity (*iHH*) (Voight et al., 2006), we then calculate the area under the curve for *DAT* as a function of distance from the position of interest, producing an integrated *DAT* (*iDAT*) (**Figure 2B**). For this study, we calculated *iDAT* only for distances where $DAT \geq 0.25$; that is, where at least half of the ancestry tracts extend that absolute distance from the site of interest. We compare the difference in order of magnitude between *iDAT* for each ancestry, analogous to the integrated haplotype score (*iHS*), we have,

$$iDAT\ score = \ln\left(\frac{iDAT_2}{iDAT_1}\right),$$

where $iDAT_i$ is the *iDAT* calculated for source population *i*. The length of ancestry tracts, and therefore the *iDAT* score, will be influenced by relative ancestry contributions from each source population. So, when possible, we standardized the *iDAT* score using the empirical distribution of *iDAT* scores for 10,000 random positions across the genome. In this way, we deviate from the calculation of standardized *iHS*, which is standardized by the empirical distribution of SNPs with the same allele frequency because the length of haplotypes will be affected by the age of a variant (Voight et al., 2006). In the case of admixed populations, we instead need to account for the effect of global ancestry proportion on ancestry tract lengths, where the majority ancestry

will tend to have longer contiguous tracts. We standardized by the genome-wide empirical distribution of *iDAT* scores, rather than *iDAT* scores for variants with the same local ancestry proportion, because the variance in local ancestry across the genome can be heavily influenced by drift. That is, to standardize the *iDAT* score, we calculate

$$\frac{\ln\left(\frac{iDAT_2}{iDAT_1}\right) - E\left[\ln\left(\frac{iDAT_2}{iDAT_1}\right)\right]}{SD\left[\ln\left(\frac{iDAT_2}{iDAT_1}\right)\right]}.$$

By standardizing against a genome-wide distribution of *iDAT* values, we can account for demographic parameters, such as admixture proportions and timing since admixture, that may affect the global *iDAT* distribution. Single-locus deviations from the genome-wide expectations may then be indicative of selection at that site, and may warrant further study.

Single-chromosome simulations

We used SLiM forward simulations with tree-sequence recording to track local ancestry (Haller et al., 2019; Haller & Messer, 2019). We considered eight different demographic scenarios: combinations of initial population size (N=10000 or N=1000) with either a constant population size or exponential growth at a rate of 0.05 per generation, and either a single pulse of admixture at the start of simulation or continuous admixture at 1% total new migrants per generation (**Table 2**). The proportion of new migrants from each source population was weighted by the respective initial admixture contributions.

For each demographic scenario, we generated 1000 simulations of the human chromosome 1. For realistic recombination rates, we used the population-averaged human genetic map provided by IMPUTE2 (Delaneau et al., 2013) (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html). We simulated admixture from two source populations to form a third admixed population (similar to recipe 17.5 in the SLiM manual) (Haller & Messer, 2016). One source population, representing the West African source population, was fixed for a neutral variant at the same position as the Duffy-null allele (chr1:159174683; GRCh37 coordinates in accordance with genetic map). Because the true West African ancestry contribution is unknown, for neutral simulations this parameter was drawn from a uniform distribution with lower and upper bounds at 0.65 and 0.75, respectively. We simulated the admixed population for 20 generations.

Using the tree-sequence files output from the SLiM simulations, we calculated the five ancestry-based summary statistics for each simulation: West African local ancestry proportion at *DARC*, variance in the distribution of West African local ancestry proportion across SNPs along the chromosome, mean and median West African ancestry tract length containing Duffy-null, and unstandardized *iDAT* score for the Duffy-null variant. We used the unstandardized *iDAT* score because there is no genome-wide distribution of *iDAT* scores for single-chromosome simulations. *iDAT* scores also could not be standardized using the distribution of simulated Duffy-null *iDAT* scores because each simulation had a different starting admixture proportion.

We sampled 172 individuals from each simulation and compared the simulated distribution to the observed values of the statistics for the 172 individuals from Santiago (**Figure 2-figure supplement 2**). The genetic map provided by IMPUTE2 is population-averaged. To determine whether population-specific differences in recombination rate may affect our ancestry-based statistics, we performed the same neutral simulations and comparison to empirical data using one of three genetic maps: GWD or IBS maps from (Spence & Song, 2019), or an African American genetic map from (Hinch et al., 2011). The expectations for mean and median tract length are affected by population-specific differences in recombination rate; however, because recombination rate affects both ancestries equally, the choice of genetic map does not change the expectation for *iDAT* score (**Figure 2-figure supplement 3**). Of note, the African American genetic map contains regions with extremely high recombination rate, resulting in the extreme differentiation in expectation of tract lengths between simulations using this and the other genetic maps, and, in particular, shorter expected tract lengths. Using this map would inflate our estimates of selection strength under the ABC framework; we chose to use the more conservative and general-use IMPUTE2 genetic map for population-averaged recombination rates.

Ancestry tracts extend over large proportions of the chromosome at the timescale of interest for this study (~20 generations). Therefore, in this case, fine-scale recombination rate differences are not expected to significantly affect our expectations for ancestry-based statistics.

Performance of *iDAT*

The impact of different population size and migration scenarios on *iDAT* is summarized in the Methods, under Single-chromosome simulations (**Figure 2-figure supplement 2**). These scenarios are relevant for the population history of Cabo Verde. Here, in order to understand the general behavior and applicability of *iDAT* for future analyses, we extend the scenarios considered beyond those likely to represent Cabo Verdean history. We consider combinations varying the generations since admixture, the selection coefficient, the initial contribution from the source populations, different chromosome lengths for the position of the adaptive allele, and different *DAT* cut-offs.

First, we considered the demographic history of the admixed population. Modifying the single-chromosome simulations described above, we conducted simulations setting the admixture timing to 10, 100, or 1000 generations in the past and admixture contribution from the source population with the adaptive allele to 0.1, 0.5 or 0.9. This source population was fixed for the variant at the Duffy-null position. We assumed a constant population size of N=10000 and a single-pulse of admixture. For each combination of admixture timing and admixture proportion, we generated 1000 simulations for each selection strength of $s = 0$, $s = 0.01$, $s = 0.05$, or $s = 0.1$. This resulted in 36 different scenarios and 36000 simulations. For each simulation, we calculated *iDAT* for the variant at the Duffy-null position.

To interpret *iDAT* performance under these scenarios, we plot the proportion of Duffy-null *iDAT* values from the selection simulations that are in the bottom fifth percentile of the respective neutral Duffy-null *iDAT* distribution (**Figure 2-figure supplement 4).** We compare within demographic models because admixture proportion has a strong influence on the expectation and possible range of *iDAT* values. We also note that *iDAT* values cannot be

calculated when an allele has been fixed in a population, as observed in the older admixture scenarios with high selection strength and high starting admixture proportion from the selected ancestry. As such, this statistic may be more useful under recent admixture and selection (i.e., fewer than 100 generations in the past) with substantial admixture contributions from both source populations of interest.

We next sought to assess how chromosome size and choice of *DAT* cut-off value affects the expected distribution of *iDAT* values. For this, we performed simulations of human chromosomes 1 (~250 Mb), 7 (~160 Mb), 15 (~100 Mb), and 22 (~50 Mb), to capture a range of reasonable chromosome sizes. We used the associated IMPUTE2 genetic map for each of these simulations. For consistency across simulations, one source population was fixed for a variant at a position at 80% of the physical length of the chromosome. We again assumed a constant population size at N=10000 and a single pulse of admixture. Based on the demographic history of Cabo Verde, we assumed the modern-day West African ancestry proportion to be the initial admixture contribution of 0.73 from the source population providing the variant of interest. We simulated the admixed population for 20 generations. We generated 1000 simulations each of neutral ($s = 0$) and strong selection scenarios ($s = 0.05$). For each simulation, we calculated *iDAT* for the tracked variant for distances where $DAT \geq 0.25$ (following our main analyses), $DAT \geq 0.125$, $DAT \geq 0.0625$, and $DAT \geq 0.01$. We show the proportion of *iDAT* values for the simulated variant under selection that are in the bottom fifth percentile of the neutral distribution of *iDAT* values, for each chromosome size and *DAT* cut-off (**Fig. 2-figure supplement 5**).

Generally, on this timescale, selection strength and admixture proportion, *iDAT* performs well across chromosome sizes and cut-off values. That said, we note that the cut-off of 0.25 works slightly better for the smallest human chromosome (chr 22), though we emphasize that its performance is not very different from the other cut-off values. Further, using lower cut-offs requires more computation time, and depending on the SNP density of the dataset, this may be an important consideration. We encourage future studies looking to detect signatures of selection using this statistic on smaller chromosomes or different admixture histories, particularly for non-human populations, to consider testing *iDAT* performance under their specific model of interest.

SWIF(r) implementation

We incorporated ancestry-based summary statistics and admixture simulations into the SWIF(r) framework developed by (Sugden et al., 2018). For the simulated training set, we followed the single-chromosome simulation framework described above. We considered a single realistic demographic scenario for training: starting population size N=10000, exponential growth at a rate of 0.05 per generation, and single-pulse admixture with starting West African ancestry contribution randomly drawn from a uniform distribution from 0.65 to 0.75.

For simulations with selection at a single locus, we assumed an additive selection model (h=0.5). Selection coefficient was randomly drawn from a uniform distribution from 0 to 0.2. We calculated the five ancestry-based summary statistics for each simulation. Under this model, we generated 50000 neutral simulations and 100 positive selection simulations for training; these training proportions are to reflect a prior probability of selection scenarios at 0.002. Since

positive selection is relatively rare compared to neutral scenarios, SWIF(r) calibrates posterior probabilities according to this designated prior probability.

To validate this extension of SWIF(r), we generated a new set of 1000 neutral simulations and 1000 positive selection simulations. **Figure 6** shows a precision-recall plot for a SWIF(r) implementation using the five ancestry-based statistics, a prior positive selection probability of 0.2% (reflecting the training set proportions above), and two classes (neutral or positive selection). **Figure 6-figure supplement 1** demonstrates SWIF(r) performance for each class and across values of admixture contribution and selection coefficient. This SWIF(r) implementation had a low rate of false-negative classification of neutral simulations as positive selection scenarios.

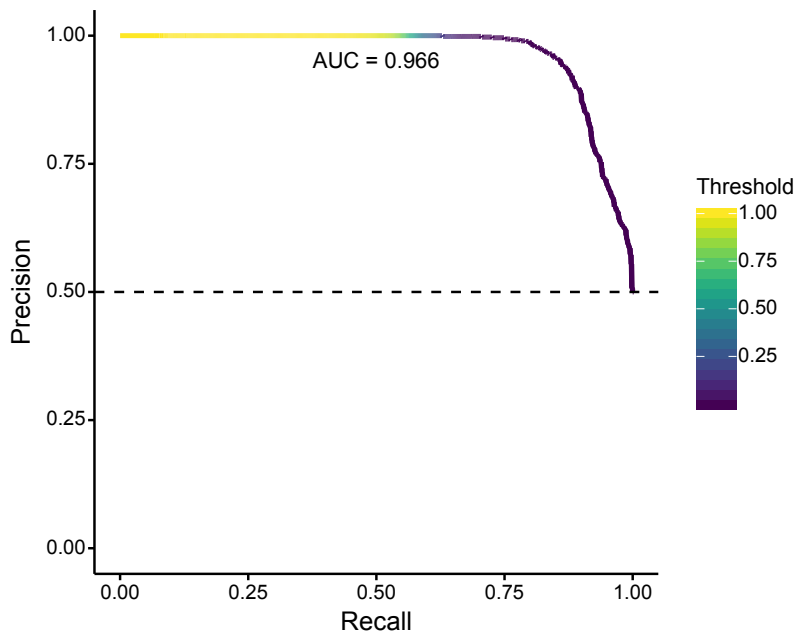

Figure 6. Precision-Recall curve for validation of SWIF(r) classification of neutral and positively selected variants, using 1000 neutral and 1000 positive selection simulations. With our ancestry-based measures, SWIF(r) achieved an area under the curve (AUC) of 0.966, where an AUC of 1 represents a classifier with perfect skill. Horizontal dashed line indicates the no-skill classifier for this dataset.

Inference of selection under a deterministic population-genetic model

To estimate the selection coefficient for the Duffy-null allele based on the dominance and the allele frequency trajectory over 20 generations, we used a deterministic population-genetic model of selection. We used the following recursion equation (Coop, 2020):

$$p_{t+1} = p_t + \frac{p_t hs + q_t s(1-h)}{1 - 2p_t q_t sh - {q_t}^2 s} p_t q_t,$$

where p_t is the frequency of Duffy-null in a given generation t, q_t is the frequency of the alternate allele in generation t, s is the selection coefficient which is constant over time, and h is the

dominance coefficient (0 if Duffy-null is dominant, 1 if recessive, and 0.5 in an additive selective model). We calculated the allele frequency over a grid of values for pairs of $h \in [0,1]$ and $s \in [0,0.2]$ in 0.005 and 0.001 increments, respectively. **Figure 4A** shows the combination of h and s that produce $\left|p_{20} - p_{\text{Duffy}}\right| < 0.01$, where p_{20} is the calculated frequency of the selected allele after 20 generations and p_{Duffy} is the observed frequency of Duffy-null in Santiago (0.834). Because the true initial allele frequency p_o is not known, we performed the analysis for three reasonable starting allele frequencies: $p_o = 0.65$, $p_o = 0.70$, and $p_o = 0.75$.

Inference of selection coefficient using approximate Bayesian computation

We used an approximate Bayesian computation (ABC) framework with non-linear regression (neural network) adjustment to jointly estimate the posterior distributions for selection strength acting on the Duffy-null allele and the initial West African ancestry contribution (*R* package '*abc*', with 'neuralnet' method) (Csilléry et al., 2012).

We followed the same simulation framework we previously described for our SWIF(r) positive selection simulations of chromosome 1, with the following demographic scenario for training: starting population size N=10000, exponential growth at a rate of 0.05 per generation, and single-pulse admixture. We assumed an additive selection model (h=0.5). The selection coefficient (s) and the initial West African ancestry contribution (m) were drawn from uniform prior distributions, $s \sim U(0,0.2)$ and $m \sim U(0.1,0.9)$. We generated 10000 simulations for ABC inference and calculated the five ancestry-based summary statistics for each simulation.

We chose the tolerance and hidden layer sizes for ABC estimation based on the best combination of *RMSE* and R^2 values for leave-one-out cross-validation under different combinations of these hyperparameters. We performed cross-validation for 1000 simulations. Estimation accuracy for this dataset was similar under a variety of hyperparameter choices. We set the number of units in the hidden layer to 'sizenet=2' and the acceptance rate to 'tol=0.05'. For the cross-validation, our estimates of the selection coefficient had an *RMSE* = 0.0083 and R^2= 0.9785; our estimates of starting admixture proportion had an *RMSE* = 0.0090 and R^2= 0.9985 (**Figure 4-figure supplement 2**).

Global Ancestry Simulations

To assess how selection at a single locus impacts genome-wide patterns of ancestry, we simulated changes in global ancestry over 20 generations (**Figure 5**). Whole autosome (22 chromosome) simulations with a realistic recombination map is computationally intensive. Instead, we took two complementary approaches.

First, we performed whole autosome simulations (22 independently segregating chromosomes) in SLiM by specifying a "crossover rate" between chromosomes of 0.5 per generation. We used the total lengths of autosomal chromosomes from the Human Genome Assembly GRCh37.p13 (https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37.p13). We used a uniform recombination rate within each chromosome of 1×10^{-8} crossovers per base position per generation. We considered the following demographic scenario: starting population

size N=10000, exponential growth at a rate of 0.05 per generation, and single-pulse admixture. Global ancestry was calculated by taking the mean West African ancestry proportion across the autosome.

Second, we performed two-chromosome simulations (modeled based on human chromosomes 1 and 2) in order to incorporate a human genetic map for realistic recombination rates. We again used the genetic maps provided by IMPUTE2 for chromosomes 1 and 2. Two-chromosome simulations were performed under the same demographic model as whole autosome simulations. Global ancestry was calculated by taking a weighted mean of ancestry for the two chromosomes, where chromosome 2 represented the contribution for the ~92% of the genome that segregates independently from chromosome 1.

To test how selection at a single locus affects global ancestry, we considered both whole autosome simulations and two-chromosome simulations, using a West African ancestry contribution of 0.65 for all simulations in the founding generation. We varied the selection coefficient for the simulated Duffy-null variant on chromosome 1 from 0 to 0.2. We assumed an additive selection model (h=0.5). The results are similar across simulation methods (**Figure 5-figure supplement 1**).

To determine how our ABC estimates of initial West African ancestry contribution differ from final global ancestry after 20 generations of selection at the Duffy-null locus, we next simulated whole autosomes admixing for 20 generations drawing parameters from the previously inferred posterior distributions of the selection coefficient and initial ancestry contribution. Specifically, we used the paired estimates of selection coefficient and West African admixture contribution from each accepted simulation from the ABC analysis and passed those parameters as input for each whole autosome simulation; this produced a distribution of estimates of global ancestry for populations simulated with realistic selection and initial ancestry contributions (**Figure 5A)**.

References:

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246. https://doi.org/10.1186/1471-2105-12-246

Belbin, G. M., Nieves-Colón, M. A., Kenny, E. E., Moreno-Estrada, A., & Gignoux, C. R. (2018). Genetic diversity in populations across Latin America: Implications for population and medical genetic studies. *Current Opinion in Genetics & Development*, *53*, 98–104. https://doi.org/10.1016/j.gde.2018.07.006

Beleza, S., Campos, J., Lopes, J., Araújo, I. I., Almada, A. H., Silva, A. C. e, Parra, E. J., & Rocha, J. (2012). The Admixture Structure and Genetic Variation of the Archipelago of Cape Verde and Its Implications for Admixture Mapping Studies. *PLOS ONE*, *7*(11), e51103. https://doi.org/10.1371/journal.pone.0051103

Beleza, S., Johnson, N. A., Candille, S. I., Absher, D. M., Coram, M. A., Lopes, J., Campos, J., Araújo, I. I., Anderson, T. M., Vilhjálmsson, B. J., Nordborg, M., Silva, A. C. e, Shriver, M. D., Rocha, J., Barsh, G. S., & Tang, H. (2013). Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *PLOS Genetics*, *9*(3), e1003372. https://doi.org/10.1371/journal.pgen.1003372

Bhatia, G., Tandon, A., Patterson, N., Aldrich, M. C., Ambrosone, C. B., Amos, C., Bandera, E. V., Berndt, S. I., Bernstein, L., Blot, W. J., Bock, C. H., Caporaso, N., Casey, G., Deming, S. L., Diver, W. R., Gapstur, S. M., Gillanders, E. M., Harris, C. C., Henderson, B. E., … Price, A. L. (2014). Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture. *The American Journal of Human Genetics*, *95*(4), 437–444. https://doi.org/10.1016/j.ajhg.2014.08.011

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. (2015). The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, *96*(1), 37–53. https://doi.org/10.1016/j.ajhg.2014.11.010

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., Amenga-Etego, L. N., Enimil, A., Apinjoh, T., Ndila, C. M., Manjurano, A., Nyirongo, V., Doumba, O., Rockett, K. A., Kwiatkowski, D. P., Spencer, C. C., & Malaria Genomic Epidemiology Network. (2016). Admixture into and within sub-Saharan Africa. *ELife*, *5*, e15266. https://doi.org/10.7554/eLife.15266

Busby, G. B. J., Hellenthal, G., Montinaro, F., Tofanelli, S., Bulayeva, K., Rudan, I., Zemunik, T., Hayward, C., Toncheva, D., Karachanak-Yankova, S., Nesheva, D., Anagnostou, P., Cali, F., Brisighelli, F., Romano, V., Lefranc, G., Buresi, C., Ben Chibani, J., Haj-Khelil, A., … Capelli, C. (2015). The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Current Biology*, *25*(19), 2518–2526. https://doi.org/10.1016/j.cub.2015.08.007

Busby, G., Christ, R., Band, G., Leffler, E., Le, Q. S., Rockett, K., Kwiatkowski, D., & Spencer, C. (2017). Inferring adaptive gene-flow in recent African history. *BioRxiv*, 205252. https://doi.org/10.1101/205252

Calfee, E., Agra, M. N., Palacio, M. A., Ramírez, S. R., & Coop, G. (2020). Selection and hybridization shaped the Africanized honey bee invasion of the Americas. *BioRxiv*, 2020.03.17.994632. https://doi.org/10.1101/2020.03.17.994632

Carreira, A. (1983). *Cabo Verde: Formação e extinção de uma sociedade escravocrata (1460-1878)*. CEE e Instituto Cabo-Verdeano do Livro.

Cavasini, C. E., de Mattos, L. C., Couto, Á. A. D., Couto, V. S. D., Gollino, Y., Moretti, L. J., Bonini-Domingos, C. R., Rossit, A. R., Castilho, L., & Machado, R. L. (2007). Duffy blood group gene polymorphisms among malaria vivax patients in four areas of the Brazilian Amazon region. *Malaria Journal*, *6*(1), 167. https://doi.org/10.1186/1475-2875-6-167

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tassé, A.-M., & Flicek, P. (2017). The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Research*, *45*(D1), D854–D859. https://doi.org/10.1093/nar/gkw829

Coop, G. (2020, May). Population and Quantitative Genetics. *Github*. https://github.com/cooplab/popgen-notes/blob/master/popgen_notes.pdf

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. https://doi.org/10.1038/nmeth.2307

DePina, A. J., Dia, A. K., de Ascenção Soares Martins, A., Ferreira, M. C., Moreira, A. L., Leal, S. V., Pires, C. M., Moreira, J. M. G., Tavares, M. F., da Moura, A. J. F., Pereira, J. M., Faye, O., Seck, I., & Niang, E. H. A. (2019). Knowledge, attitudes and practices about malaria in Cabo Verde: A country in the pre-elimination context. *BMC Public Health*, *19*(1), 850. https://doi.org/10.1186/s12889-019-7130-5

DePina, A. J., Niang, E. H. A., Barbosa Andrade, A. J., Dia, A. K., Moreira, A., Faye, O., & Seck, I. (2018). Achievement of malaria pre-elimination in Cape Verde according to the data collected from 2010 to 2016. *Malaria Journal*, *17*(1), 236. https://doi.org/10.1186/s12936-018-2376-4

Fairley, S., Lowy-Gallego, E., Perry, E., & Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, *48*(D1), D941–D947. https://doi.org/10.1093/nar/gkz836

Fenner, J. N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, *128*(2), 415–423. https://doi.org/10.1002/ajpa.20188

Fernandes, A. T., Velosa, R., Jesus, J., Carracedo, A., & Brehm, A. (2003). Genetic Differentiation of the Cabo Verde Archipelago Population Analysed by STR Polymorphisms. *Annals of Human Genetics*, *67*(4), 340–347. https://doi.org/10.1046/j.1469-1809.2003.00029.x

Fernandes, V., Brucato, N., Ferreira, J. C., Pedro, N., Cavadas, B., Ricaut, F.-X., Alshamali, F., & Pereira, L. (2019). Genome-Wide Characterization of Arabian Peninsula Populations: Shedding Light on the History of a Fundamental Bridge between Continents. *Molecular Biology and Evolution*, *36*(3), 575–586. https://doi.org/10.1093/molbev/msz005

Ferreira, M., Moreira, A., Correia, A., & De Pina, A. (2017, February 20). *Malaria in Cabo Verde: Pre-elimination versus the Elimination* [Poster]. Keystone Symposia Conference B5: Malaria: From Innovation to Eradication. https://doi.org/10.13140/RG.2.2.28878.72004

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., & Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science*, *354*(6313), 760–764. https://doi.org/10.1126/science.aag0776

Fortes-Lima, C. A., Laurent, R., Thouzeau, V., Toupance, B., & Verdu, P. (2019). Complex genetic admixture histories reconstructed with Approximate Bayesian Computations. *BioRxiv*, 761452. https://doi.org/10.1101/761452

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLOS Genetics*, *7*(11), e1002355. https://doi.org/10.1371/journal.pgen.1002355

Gething, P. W., Elyazar, I. R. F., Moyes, C. L., Smith, D. L., Battle, K. E., Guerra, C. A., Patil, A. P., Tatem, A. J., Howes, R. E., Myers, M. F., George, D. B., Horby, P., Wertheim, H. F. L., Price, R. N., Müeller, I., Baird, J. K., & Hay, S. I. (2012). A Long Neglected World Malaria Map: Plasmodium vivax Endemicity in 2010. *PLOS Neglected Tropical Diseases*, *6*(9), e1814. https://doi.org/10.1371/journal.pntd.0001814

Gravel, S. (2012). Population Genetics Models of Local Ancestry. *Genetics*, *191*(2), 607–619. https://doi.org/10.1534/genetics.112.139808

Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D. J., Griesemer, D., Karlsson, E. K., Wong, S. H., Cabili, M., Adegbola, R. A., Bamezai, R. N. K., Hill, A. V. S., Vannberg, F. O., Rinn, J. L., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2013). Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell*, *152*(4), 703–713. https://doi.org/10.1016/j.cell.2013.01.035

Guan, Y. (2014). Detecting Structure of Haplotypes and Local Ancestry. *Genetics*, *196*(3), 625–642. https://doi.org/10.1534/genetics.113.160697

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, *19*(2), 552–566. https://doi.org/10.1111/1755-0998.12968

Haller, B. C., & Messer, P. W. (2016). *SLiM: An Evolutionary Simulation Framework*.

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Hamblin, M. T., & Di Rienzo, A. (2000). Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. *The American Journal of Human Genetics*, *66*(5), 1669–1679. https://doi.org/10.1086/302879

Hamblin, M. T., Thompson, E. E., & Di Rienzo, A. (2002). Complex Signatures of Natural Selection at the Duffy Blood Group Locus. *The American Journal of Human Genetics*, *70*(2), 369–383. https://doi.org/10.1086/338628

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. https://doi.org/10.1111/mec.12415

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A Genetic Atlas of Human Admixture History. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518

Hermisson, J., & Pennings, P. S. (2005). Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics*, *169*(4), 2335–2352. https://doi.org/10.1534/genetics.104.036947

Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., Wang, K., Buxbaum, S. G., Akylbekova, E. L., Aldrich, M. C., Ambrosone, C. B., Amos, C., Bandera, E. V., Berndt, S. I., Bernstein, L., Blot, W. J., Bock, C. H., Boerwinkle, E., … Myers, S. R. (2011). The landscape of recombination in African Americans. *Nature*, *476*(7359), 170–175. https://doi.org/10.1038/nature10336

Hodgson, J. A., Pickrell, J. K., Pearson, L. N., Quillen, E. E., Prista, A., Rocha, J., Soodyall, H., Shriver, M. D., & Perry, G. H. (2014). Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1789), 20140930. https://doi.org/10.1098/rspb.2014.0930

Howes, R. E., Patil, A. P., Piel, F. B., Nyangiri, O. A., Kabaria, C. W., Gething, P. W., Zimmerman, P. A., Barnadas, C., Beall, C. M., Gebremedhin, A., Ménard, D., Williams, T. N., Weatherall, D. J., & Hay, S. I. (2011). The global distribution of the Duffy blood group. *Nature Communications*, *2*(1), 1–10. https://doi.org/10.1038/ncomms1265

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., … Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, *512*(7513), 194–197. https://doi.org/10.1038/nature13408

Jagoda, E., Lawson, D. J., Wall, J. D., Lambert, D., Muller, C., Westaway, M., Leavesley, M., Capellini, T. D., Mirazón Lahr, M., Gerbault, P., Thomas, M. G., Migliano, A. B., Willerslev, E., Metspalu, M., & Pagani, L. (2018). Disentangling Immediate Adaptive Introgression from Selection on Standing Introgressed Variation in Humans. *Molecular Biology and Evolution*, *35*(3), 623–630. https://doi.org/10.1093/molbev/msx314

Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D. B., Pritchard, J. K., Beall, C. M., & Rienzo, A. D. (2014). Admixture facilitates genetic adaptations to high altitude in Tibet. *Nature Communications*, *5*(1), 1–7. https://doi.org/10.1038/ncomms4281

Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B., & Jin, L. (2012). Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Research*, *22*(3), 519–527. https://doi.org/10.1101/gr.124784.111

Kano, F. S., de Souza, A. M., de Menezes Torres, L., Costa, M. A., Souza-Silva, F. A., Sanchez, B. A. M., Fontes, C. J. F., Soares, I. S., de Brito, C. F. A., Carvalho, L. H., & Sousa, T. N. (2018). Susceptibility to Plasmodium vivax malaria associated with DARC (Duffy antigen) polymorphisms is influenced by the time of exposure to malaria. *Scientific Reports*, *8*(1), 1–14. https://doi.org/10.1038/s41598-018-32254-z

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2018). Deleterious variation shapes the genomic landscape of introgression. *PLOS Genetics*, *14*(10), e1007741. https://doi.org/10.1371/journal.pgen.1007741

Kim, Y., & Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics*, *167*(3), 1513–1524. https://doi.org/10.1534/genetics.103.025387

Korunes, K., Soares-Souza, G., Bobrek, K., Tang, H., Araújo, I., Goldberg, A., & Beleza, S. (2020). Sex-biased admixture and assortative mating influence demographic inference and shape genetic variation in admixed Cabo Verdeans. *BioRxiv*. BIORXIV/2020/422766

Kwiatkowski, D. P. (2005). How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *The American Journal of Human Genetics*, *77*(2), 171–192. https://doi.org/10.1086/432519

Lachance, J., & Tishkoff, S. A. (2013). Population Genomics of Human Adaptation. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 123–143. https://doi.org/10.1146/annurev-ecolsys-110512-135833

Lamason, R. L., Mohideen, M.-A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'Donnell, D., Kittles, R., … Cheng, K. C. (2005). SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science*, *310*(5755), 1782–1786. https://doi.org/10.1126/science.1116238

Laso-Jadart, R., Harmant, C., Quach, H., Zidane, N., Tyler-Smith, C., Mehdi, Q., Ayub, Q., Quintana-Murci, L., & Patin, E. (2017). The Genetic Legacy of the Indian Ocean Slave Trade: Recent Admixture and Post-admixture Selection in the Makranis of Pakistan. *The American Journal of Human Genetics*, *101*(6), 977–984. https://doi.org/10.1016/j.ajhg.2017.09.025

Liang, M., & Nielsen, R. (2014). The Lengths of Admixture Tracts. *Genetics*, *197*(3), 953–967. https://doi.org/10.1534/genetics.114.162362

Lindo, J., Huerta-Sánchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., Cybulski, J. S., Willerslev, E., DeGiorgio, M., & Malhi, R. S. (2016). A time transect of exomes from a Native American population before and after European contact. *Nature Communications*, *7*(1), 13175. https://doi.org/10.1038/ncomms13175

Lohmueller, K. E., Bustamante, C. D., & Clark, A. G. (2010). The Effect of Recent Admixture on Inference of Ancient Human Population History. *Genetics*, *185*(2), 611–622. https://doi.org/10.1534/genetics.109.113761

Lohmueller, K. E., Bustamante, C. D., & Clark, A. G. (2011). Detecting Directional Selection in the Presence of Recent Admixture in African-Americans. *Genetics*, *187*(3), 823–835. https://doi.org/10.1534/genetics.110.122739

Lopez, M., Choin, J., Sikora, M., Siddle, K., Harmant, C., Costa, H. A., Silvert, M., Mouguiama-Daouda, P., Hombert, J.-M., Froment, A., Le Bomin, S., Perry, G. H., Barreiro, L. B., Bustamante, C. D., Verdu, P., Patin, E., & Quintana-Murci, L. (2019). Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest. *Current Biology*, *29*(17), 2926-2935.e4. https://doi.org/10.1016/j.cub.2019.07.013

Maclean, C. A., Chue Hong, N. P., & Prendergast, J. G. D. (2015). hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. *Molecular Biology and Evolution*, *32*(11), 3027–3029. https://doi.org/10.1093/molbev/msv172

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Mathias, R. A., Taub, M. A., Gignoux, C. R., Fu, W., Musharoff, S., O'Connor, T. D., Vergara, C., Torgerson, D. G., Pino-Yanes, M., Shringarpure, S. S., Huang, L., Rafaels, N., Boorgula, M. P., Johnston, H. R., Ortega, V. E., Levin, A. M., Song, W., Torres, R., Padhukasahasram, B., … Barnes, K. C. (2016). A continuum of admixture in the Western

Hemisphere revealed by the African Diaspora genome. *Nature Communications*, *7*(1), 1–10. https://doi.org/10.1038/ncomms12522

McManus, K. F., Taravella, A. M., Henn, B. M., Bustamante, C. D., Sikora, M., & Cornejo, O. E. (2017). Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLOS Genetics*, *13*(3), e1006560. https://doi.org/10.1371/journal.pgen.1006560

Mercereau-Puijalon, O., & Ménard, D. (2010). Plasmodium vivax and the Duffy antigen: A paradigm revisited. *Transfusion Clinique et Biologique*, *17*(3), 176–183. https://doi.org/10.1016/j.tracli.2010.06.005

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. L., & Reich, D. (2011). The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLOS Genetics*, *7*(4), e1001373. https://doi.org/10.1371/journal.pgen.1001373

Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., Ortiz-Tello, P. A., Martínez, R. J., Hedges, D. J., Morris, R. W., Eng, C., Sandoval, K., Acevedo-Acevedo, S., Norman, P. J., Layrisse, Z., Parham, P., Martínez-Cruzado, J. C., Burchard, E. G., Cuccaro, M. L., … Bustamante, C. D. (2013). Reconstructing the Population Genetic History of the Caribbean. *PLOS Genetics*, *9*(11), e1003925. https://doi.org/10.1371/journal.pgen.1003925

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302–310. https://doi.org/10.1038/nature21347

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, *8*(11), 857–868. https://doi.org/10.1038/nrg2187

Norris, E. T., Rishishwar, L., Chande, A. T., Conley, A. B., Ye, K., Valderrama-Aguirre, A., & Jordan, I. K. (2019). Admixture-enabled selection for rapid adaptive evolution in the Americas. *BioRxiv*, 783845. https://doi.org/10.1101/783845

Norris, E. T., Rishishwar, L., Chande, A. T., Conley, A. B., Ye, K., Valderrama-Aguirre, A., & Jordan, I. K. (2020). Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biology*, *21*(1), 29. https://doi.org/10.1186/s13059-020-1946-2

Norton, H. L., Kittles, R. A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V. A., Bradley, D. G., McEvoy, B., & Shriver, M. D. (2007). Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. *Molecular Biology and Evolution*, *24*(3), 710–722. https://doi.org/10.1093/molbev/msl203

Ohashi, J., Naka, I., Patarapotikul, J., Hananantachai, H., Brittenham, G., Looareesuwan, S., Clark, A. G., & Tokunaga, K. (2004). Extended Linkage Disequilibrium Surrounding the Hemoglobin E Variant Due to Malarial Selection. *The American Journal of Human Genetics*, *74*(6), 1198–1208. https://doi.org/10.1086/421330

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J. B., Fernandes, V., Pereira, L., … Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, *356*(6337), 543–546. https://doi.org/10.1126/science.aal1988

Patterson, K. D. (1988). Epidemics, Famines, and Population in the Cape Verde Islands, 1580-1900. *The International Journal of African Historical Studies*, *21*(2), 291–313. JSTOR. https://doi.org/10.2307/219938

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., & Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, *19*(5), 826–837. https://doi.org/10.1101/gr.087577.108

Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-loth, V., Sanchez, J., Alva, O., Arachiche, A., Boland, A., Olaso, R., Deleuze, J.-F., Ricaut, F.-X., Rakotoarisoa, J.-A., Radimilahy, C., Stoneking, M., & Letellier, T. (2018). Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nature Communications*, *9*(1), 1–9. https://doi.org/10.1038/s41467-018-03342-5

Racimo, F., Marnetto, D., & Huerta-Sánchez, E. (2017). Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Molecular Biology and Evolution*, *34*(2), 296–317. https://doi.org/10.1093/molbev/msw216

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, *16*(6), 359–371. https://doi.org/10.1038/nrg3936

Reynolds, A. W., Mata-Míguez, J., Miró-Herrans, A., Briggs-Cloud, M., Sylestine, A., Barajas-Olmos, F., Garcia-Ortiz, H., Rzhetskaya, M., Orozco, L., Raff, J. A., Hayes, M. G., & Bolnick, D. A. (2019). Comparing signals of natural selection between three Indigenous North American populations. *Proceedings of the National Academy of Sciences*, *116*(19), 9312–9317. https://doi.org/10.1073/pnas.1819467116

Rishishwar, L., Conley, A. B., Wigington, C. H., Wang, L., Valderrama-Aguirre, A., & Jordan, I. K. (2015). Ancestry, admixture and fitness in Colombian genomes. *Scientific Reports*, *5*(1), 1–16. https://doi.org/10.1038/srep12376

Roberts, G. (1745). Account of a Voyage to the Islands of the Canaries, Cape de Verde, and Barbados, in 1721. In *A New General Collection of Voyages and Travels* (Thomas Astley, Ed., Vol. 1, p. 643). reprinted London, 1968; 1st printed 1745-47.

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., Avila, F. de, Gómez-Valdés, J., León-Mimila, P., Hunemeier, T., Ramallo, V., Cerqueira, C. C. S. de, Burley, M.-W., Konca, E., Oliveira, M. Z. de, Veronez, M. R., … Gonzalez-José, R. (2014). Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLOS Genetics*, *10*(9), e1004572. https://doi.org/10.1371/journal.pgen.1004572

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive Natural Selection in the Human Lineage. *Science*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309

Sabeti, Pardis C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. https://doi.org/10.1038/nature01140

Setter, D., Mousset, S., Cheng, X., Nielsen, R., DeGiorgio, M., & Hermisson, J. (2020). VolcanoFinder: Genomic scans for adaptive introgression. *PLOS Genetics*, *16*(6), e1008867. https://doi.org/10.1371/journal.pgen.1008867

Snow, R. W., Amratia, P., Kabaria, C. W., Noor, A. M., & Marsh, K. (2012). The changing limits and incidence of malaria in Africa: 1939-2009. *Advances in Parasitology*, *78*, 169–262. https://doi.org/10.1016/B978-0-12-394303-3.00010-4

Sousa, T. N., Sanchez, B. a. M., Cerávolo, I. P., Carvalho, L. H., & Brito, C. F. A. (2007). Real-time multiplex allele-specific polymerase chain reaction for genotyping of the Duffy antigen, the Plasmodium vivax invasion receptor. *Vox Sanguinis*, *92*(4), 373–380. https://doi.org/10.1111/j.1423-0410.2007.00902.x

Spence, J. P., & Song, Y. S. (2019). Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, *5*(10), eaaw9206. https://doi.org/10.1126/sciadv.aaw9206

Sugden, L. A., Atkinson, E. G., Fischer, A. P., Rong, S., Henn, B. M., & Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature Communications*, *9*(1), 1–14. https://doi.org/10.1038/s41467-018-03100-7

Svedberg, J., Shchur, V., Reinman, S., Nielsen, R., & Corbett-Detig, R. (2020). Inferring Adaptive Introgression Using Hidden Markov Models. *BioRxiv*, 2020.08.02.232934. https://doi.org/10.1101/2020.08.02.232934

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G., & Risch, N. J. (2007). Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. *The American Journal of Human Genetics*, *81*(3), 626–633. https://doi.org/10.1086/520769

Triska, P., Soares, P., Patin, E., Fernandes, V., Cerny, V., & Pereira, L. (2015). Extensive Admixture and Selective Pressure Across the Sahel Belt. *Genome Biology and Evolution*, *7*(12), 3484–3495. https://doi.org/10.1093/gbe/evv236

Verdu, P., Jewett, E. M., Pemberton, T. J., Rosenberg, N. A., & Baptista, M. (2017). Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population. *Current Biology*, *27*(16), 2529-2535.e3. https://doi.org/10.1016/j.cub.2017.07.002

Vicuña, L., Klimenkova, O., Norambuena, T., Martinez, F. I., Fernandez, M. I., Shchur, V., & Eyheramendy, S. (2020). Post-Admixture Selection on Chileans Targets Haplotype Involved in Pigmentation and Immune Defense Against Pathogens. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evaa136

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLOS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072

World Health Organization, Global Malaria Programme, University of California, S. F., Global Health Sciences, Global Health Group, Cabo Verde, & Ministério da Saúde. (2012). *Moving towards sustainable elimination in Cape Verde*. World Health Organization.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., … Wang, J. (2010). Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, *329*(5987), 75–78. https://doi.org/10.1126/science.1190371

Zhou, Q., Zhao, L., & Guan, Y. (2016). Strong Selection at MHC in Mexicans since Admixture. *PLOS Genetics*, *12*(2), e1005847. https://doi.org/10.1371/journal.pgen.1005847

Supplemental Material

Table 1. Expected and observed Duffy-null allele frequencies for each island and source population. Expected Duffy-null frequencies are approximated by mean West African global ancestry proportion for each island, calculated using the ADMIXTURE software.

population	n (sampled individuals)	expected frequency	observed frequency	binomial test p-value
Santiago	172	0.737	0.834	2.193×10^{-5}
Fogo	129	0.498	0.539	0.192
NW Cluster	236	0.552	0.557	0.817
GWD	107	0.997	1.000	-
IBS	107	0.002	0.019	-

Table 2. Demographic models used for single-chromosome neutral simulations relevant to Cabo Verde demographic history.

initial population size (N)	population growth model	population growth rate (per generation)	admixture type	proportion of new migrants (per generation)	scenario number
1000	constant size	-	single-pulse	-	1
			continuous	0.01	2
	exponential	0.05	single-pulse	-	3
			continuous	0.01	4
10000	constant size	-	single-pulse	-	5
			continuous	0.01	6
	exponential	0.05	single-pulse	-	7
			continuous	0.01	8

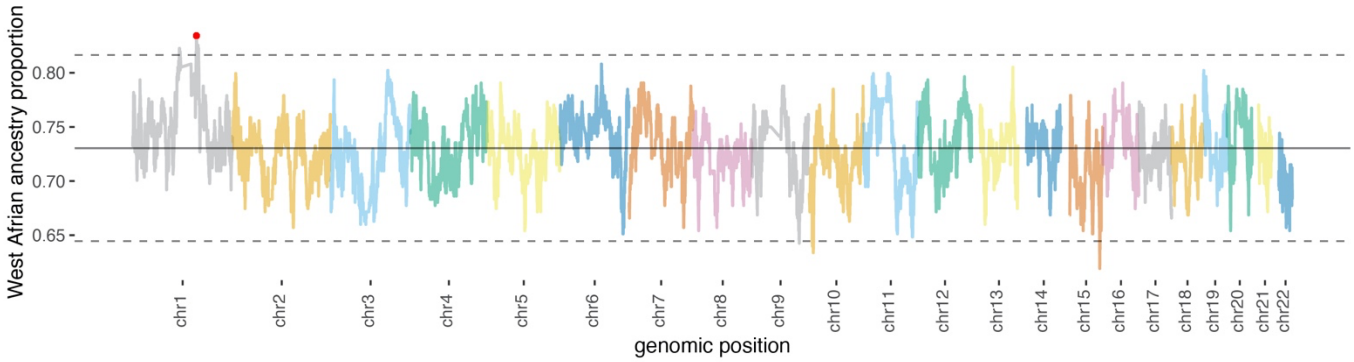

Figure 1-figure supplement 1. Local ancestry proportion along the genome. The mean is indicated by the solid horizontal line, and dashed horizontal lines represent 3 standard deviations from the mean. Again, this plot demonstrates Duffy-null (red dot) as the highest value for West African ancestry proportion.

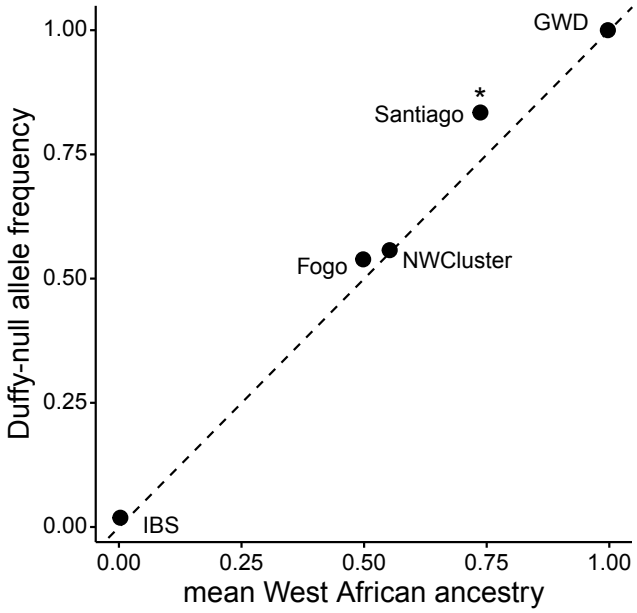

Figure 1-figure supplement 2. The observed frequency of Duffy-null for each island vs neutral expectation based on mean global ancestry (as estimated by ADMIXTURE). *indicates significant p-value<0.001 for binomial test (see **Table 1** for sample sizes and further details).

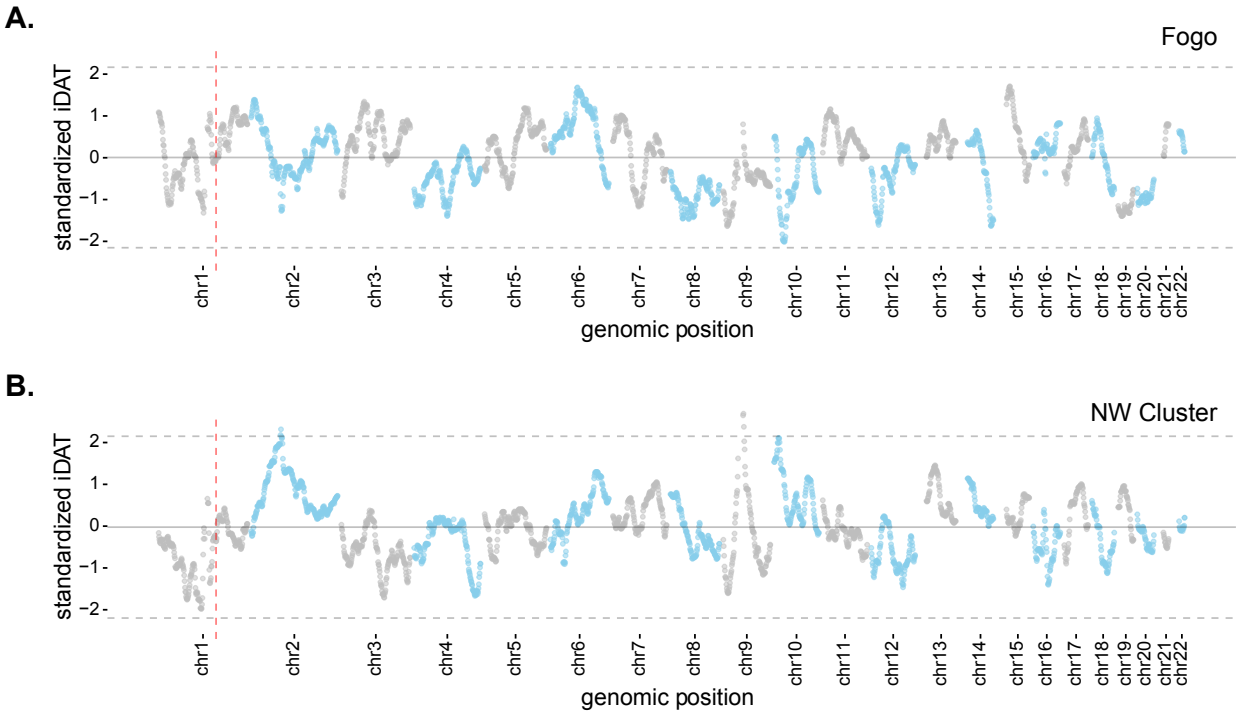

Figure 2-figure supplement 1. Mean standardized integrated *DAT* (*iDAT*) score for 20 Mb sliding windows (step size = 1Mb), using standardized *iDAT* for 10000 random positions across the genome for **A)** Fogo and **B)** the Northwest Cluster. Solid grey lines indicate mean windowed standardized *iDAT* score for each island (Fogo, 0.006; NW Cluster, -0.024) and dashed grey lines indicate three standard deviations from the mean. Vertical dashed red lines indicate the *DARC* locus, which is not an outlier for either Fogo or the NW Cluster.

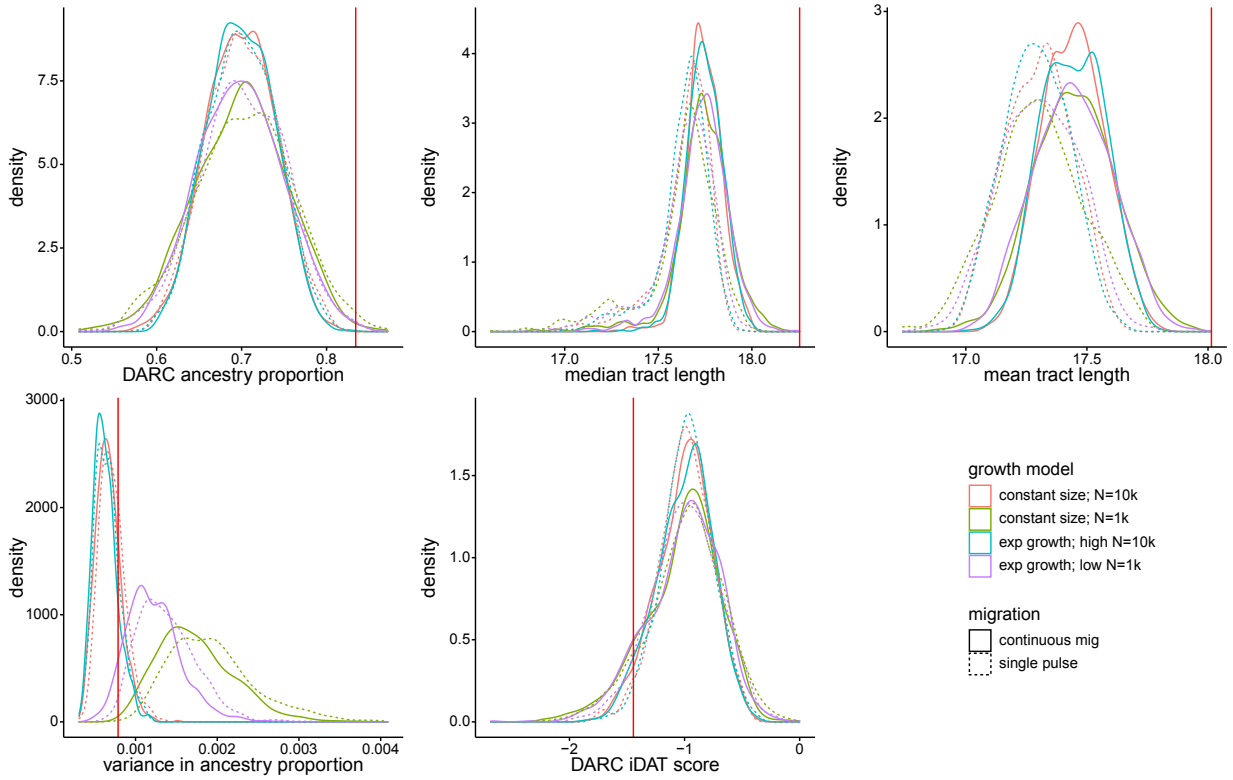

Figure 2-figure supplement 2. Density distributions for five ancestry-based statistics under eight neutral models. Summary statistics were calculated from a random sample of 172 individuals from each simulated population, matching the number of individuals from Santiago included in our analyses. High population size models correspond to initial N=10000, low population size (high drift) models correspond to initial N=1000. Exponential growth model corresponds to a rate of 0.05 per generation. Continuous migration refers to 1% total new migrants each generation, at the same proportions as initial admixture contributions for each source population. Vertical red line represents each measure's observed value for Santiago.

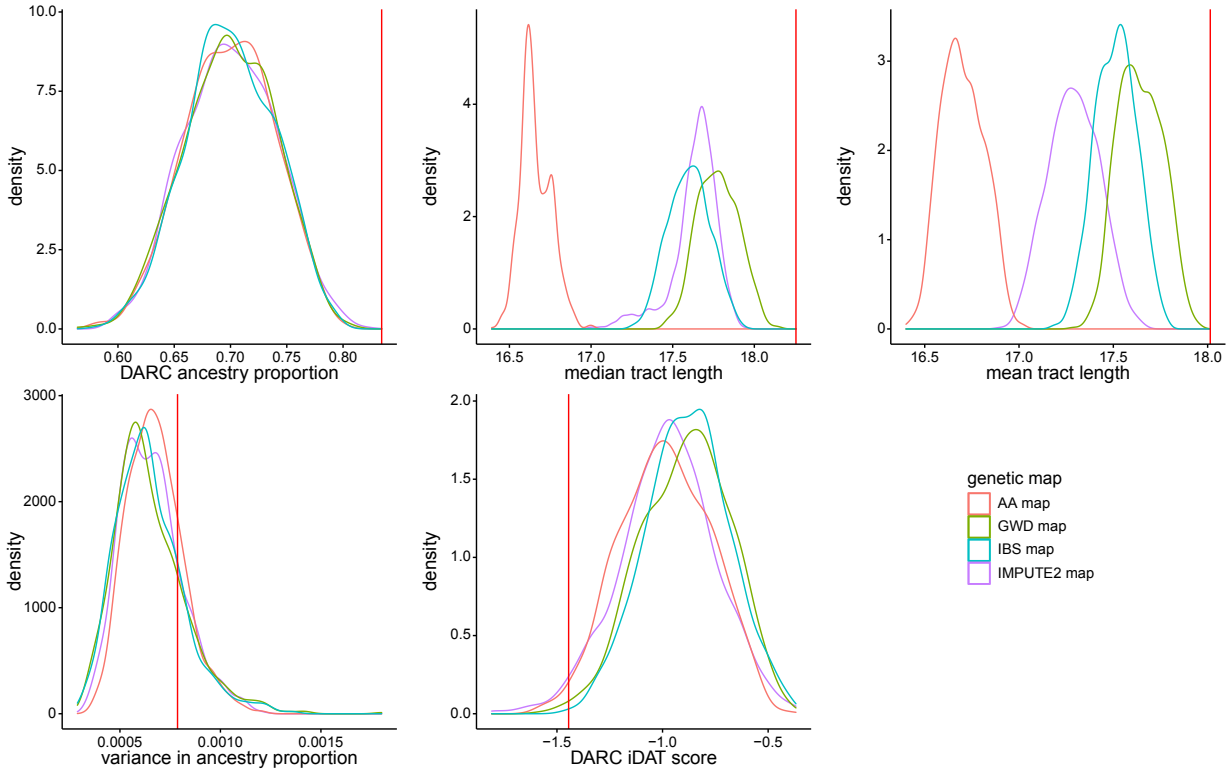

Figure 2-figure supplement 3. Density distributions for five ancestry-based statistics under simulations using different genetic maps. Simulations shown assumed a single pulse of admixture with a constant population size of N=10000. Initial admixture contributions were drawn from a uniform distribution from 0.65 to 0.75. Summary statistics were calculated from a random sample of 172 individuals from each simulated population, matching the number of individuals from Santiago included in our analyses. Genetic maps correspond to the population-averaged IMPUTE2 map, IBS-specific genetic map, GWD-specific genetic map, and African American (AA)-specific genetic map. Vertical red line represents each measure's observed value for Santiago.

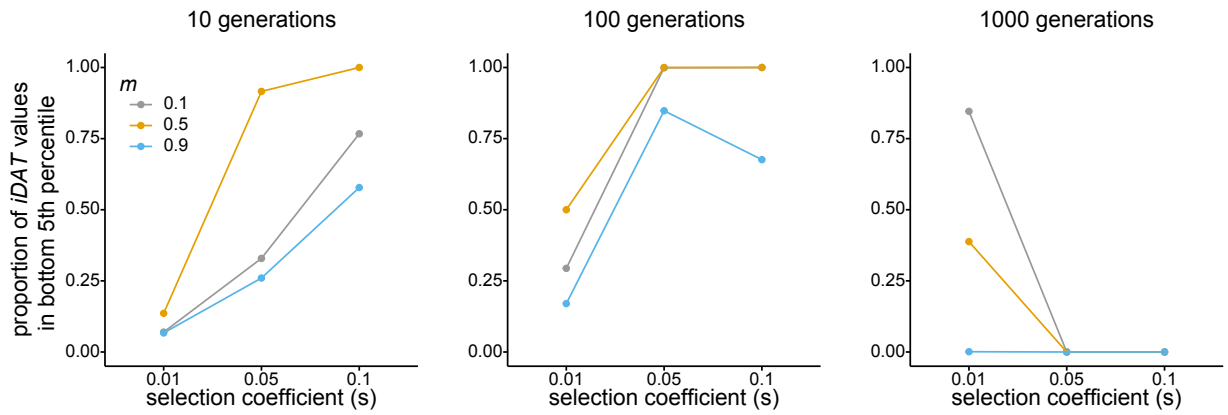

Figure 2-figure supplement 4. Performance of *iDAT* under various scenarios. Each plot corresponds to number of generations since admixture (10 – left; 100 – middle; 1000 – right). Line and point colors correspond to source population 1 admixture contribution at $m = 0.1$ (grey), $m = 0.5$ (yellow), $m = 0.9$ (blue). Within each plot, the x-axis shows selection strength for the simulated variant at the Duffy-null position, and the y-axis shows the proportion of Duffy-null *iDAT* values from the selection simulations that are in the bottom fifth percentile of the simulated neutral Duffy-null *iDAT* distribution. Notably, *iDAT* cannot be calculated for variants that are fixed in the population, as was the case for many simulations of older admixture (100 or 1000 generations) and high admixture proportion (m=0.9) and/or stronger selection. This is reflected in the statistic's performance under these scenarios.

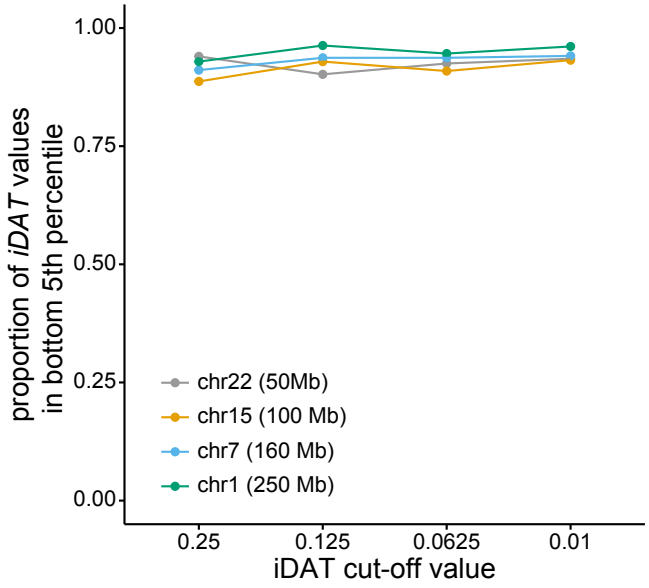

Figure 2-figure supplement 5. Performance of *iDAT* for various chromosome sizes and cut-off values. Line and point colors correspond to simulated human chromosome and corresponding size (chr 1 – green; chr 7 – blue; chr 15 – yellow; chr 22 – grey). X-axis shows *DAT* cut-off values, and y-axis shows proportion of *iDAT* values at the simulated variant under selection that are in the bottom fifth percentile of simulated neutral *iDAT* values.

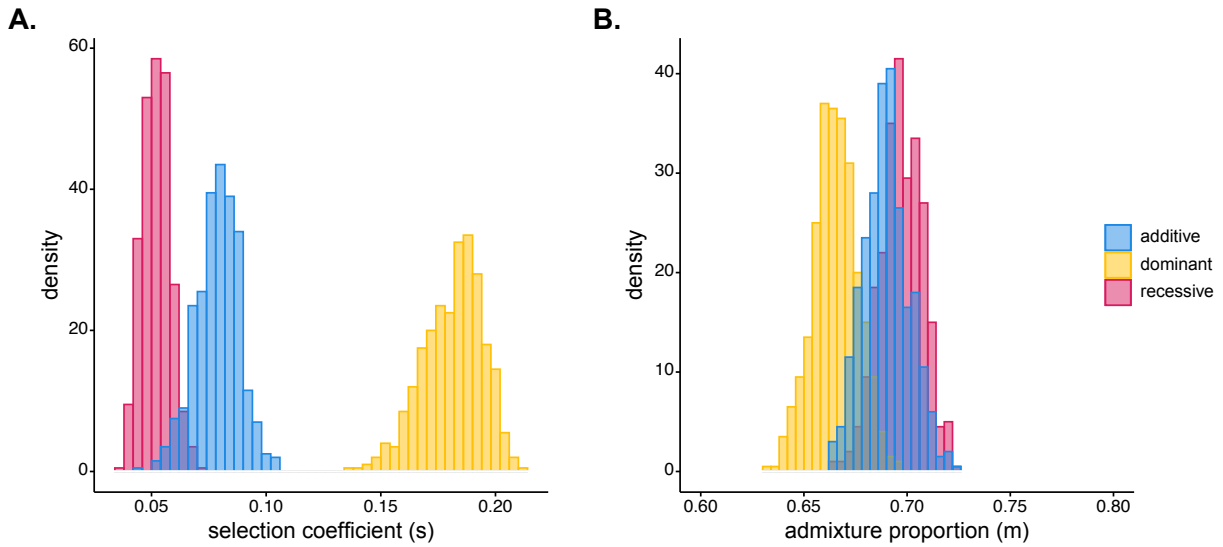

Figure 4-figure supplement 1. Results of ABC estimation of posterior distributions for **A)** selection coefficient for Duffy-null and **B)** initial West African ancestry contribution for Santiago. Duffy-null allele was modeled as additive (blue; h=0.5), dominant (yellow; h=1 in SLiM), or recessive (pink; h=0 in SLiM). Posterior median estimates for selection coefficient: s_{rec}= 0.052, s_{add}= 0.0795, s_{dom}= 0.183; initial ancestry contribution: m_{rec}= 0.697, m_{add}= 0.690, m_{dom}= 0.665. Prior distributions were s~U(0, 0.2) and m~U(0.1, 0.9).

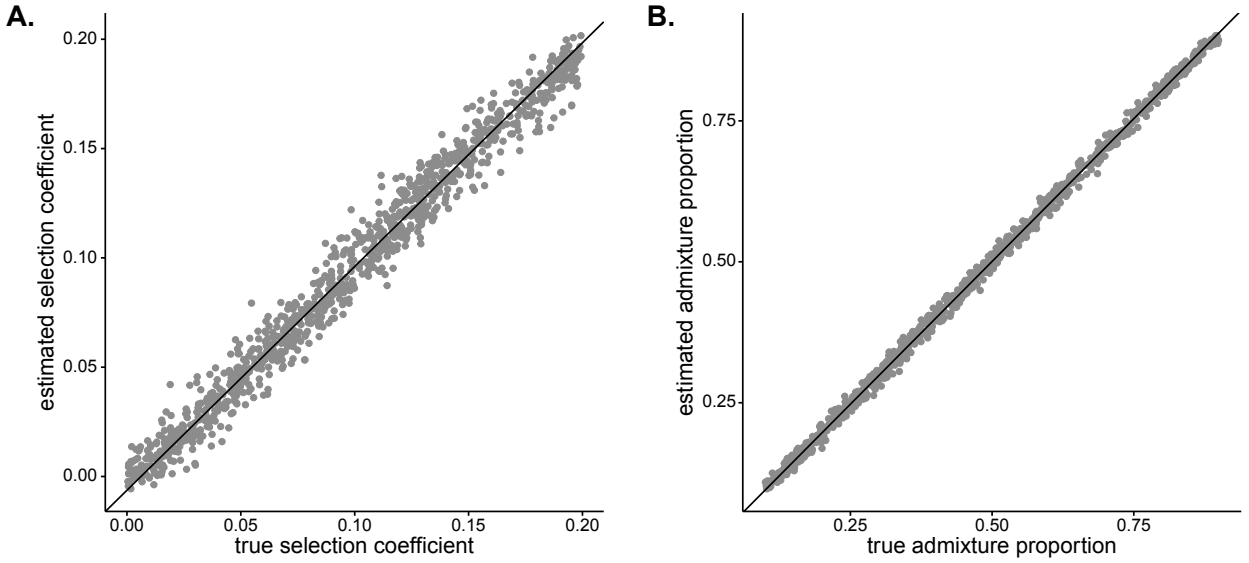

Figure 4-figure supplement 2. Results of leave-one-out cross-validation of ABC joint estimation. **A)** selection coefficient **(***RMSE* = 0.0083, R^2=0.9785) and **B)** initial West African admixture contribution (*RMSE* = 0.0090, R^2=0.9985).

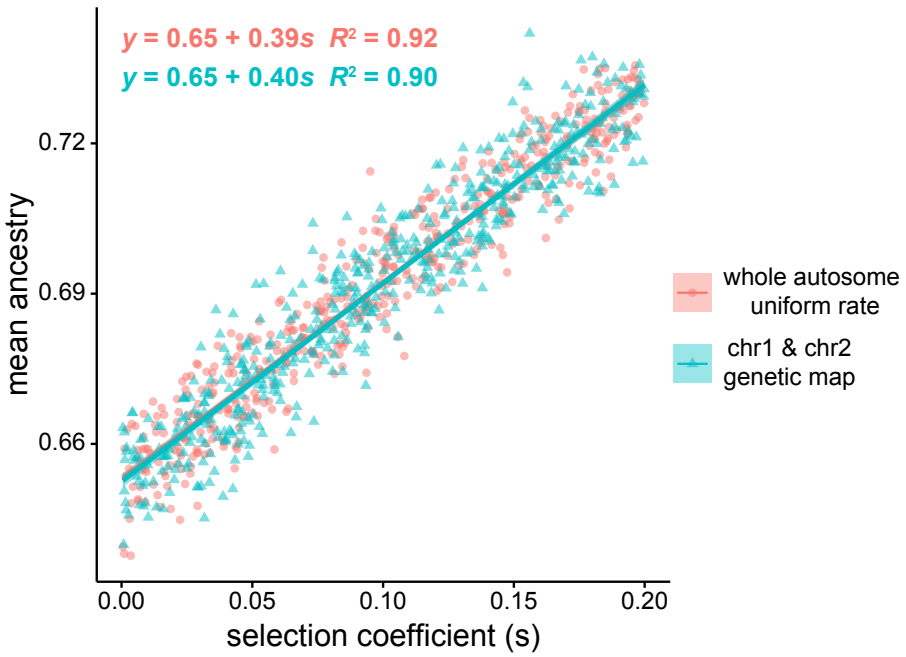

Figure 5-figure supplement 1. Effect of selection on global ancestry across simulation methods. Pink circles indicate West African mean global ancestry after 20 generations versus selection coefficient for whole autosome (22 chromosome) simulations, using a uniform recombination rate within each chromosome. Green triangles represent mean weighted ancestry for chromosome 1 and chromosome 2, with chromosome 2 representing the 92% of the genome that segregates independently from chromosome 1, using a human genetic map for recombination rates. We performed 500 simulations for each model, and all simulations started with West African ancestry contribution of 0.65. The estimated slope and intercept for the two methods of simulating global ancestry are highly similar. ANCOVA results suggest there is a significant effect of selection coefficient on global ancestry: $F(1,997)=1.0519\times10^{4}$, $p < 2\times10^{-6}$, but there is no significant effect of recombination rate and simulation model on global ancestry estimate after controlling for selection coefficient: $F(1,997) = 1.6350\times10^{-1}$, $p = 0.686$.

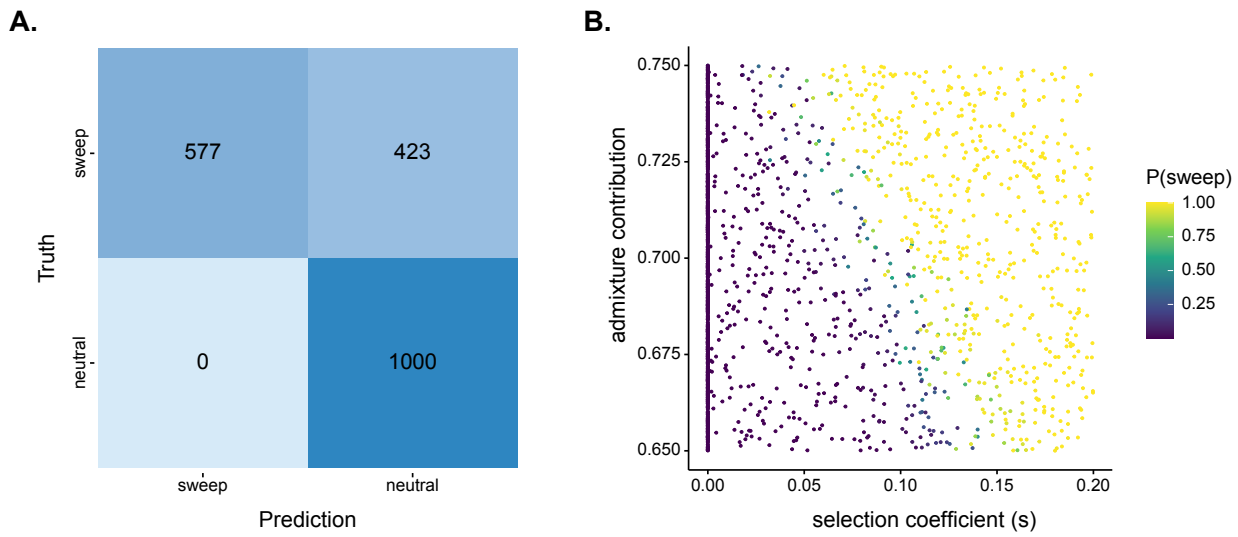

Figure 6-figure supplement 1. SWIF(r) classification results for 1000 neutral and 1000 positive selection simulations used for the test set based on Santiago's demographic history. **A)** Confusion matrix with threshold P(selection) > 0.5. There are no false positives in test set and a high rate of false negatives. **B)** Scatterplot of initial admixture contribution vs selection coefficient. Points colored by P(selection) as estimated by SWIF(r). The majority of false-negative classifications (i.e. classifying selection scenarios as neutral) occurred with low starting admixture proportion or low selection strength. If we consider ABC estimates for admixture proportion (~0.7) and selection strength (~0.08), Duffy-null on Santiago would sit around the transition from high to low rate of false negatives. SWIF(r) returned high probability for Duffy-null, P(selection) > 0.999.

Supplementary Files:

Supplementary File 1 – Chromosome 16:46582888-60359576 GO terms

File containing ENSEMBL gene IDs and associated GO terms for the 10 genes that overlap with region showing extreme *iDAT* signatures.