Complex genetic admixture histories reconstructed with Approximate Bayesian Computations

Cesar Fortes-Lima[1], Romain Laurent[2], Valentin Thouzeau[3], Bruno Toupance[2], and Paul Verdu[2]

[1]Uppsala University
[2]MNHN
[3]PSL

October 1, 2020

Abstract

Admixture is a fundamental evolutionary process that has influenced genetic patterns in numerous species. Maximum-likelihood approaches based on allele frequencies and linkage-disequilibrium have been extensively used to infer admixture processes from genome-wide datasets, mostly in human populations. Nevertheless, complex admixture histories, beyond one or two pulses of admixture, remain methodologically challenging to reconstruct. We develop an Approximate Bayesian Computation (ABC) framework to reconstruct highly complex admixture histories from independent genetic markers. We built the software package MetHis to simulate independent SNPs or microsatellites in a two-way admixed population for scenarios with multiple admixture pulses, monotonically decreasing or increasing recurring admixture, or combinations of these scenarios; and draw model-parameter values from prior distributions set by the user. For each simulation, MetHis calculates 24 summary-statistics describing genetic diversity and moments of individual admixture fractions. We coupled MetHis with existing machine-learning ABC algorithms and investigate the admixture history of hybrid populations. Results show that Random-Forest ABC scenario-choice can accurately distinguish most complex admixture scenarios and errors are mainly found in regions of the parameter space where scenarios are highly nested, and, thus, biologically similar. We focus on African American and Barbadian populations as case studies. We find that Neural-Network ABC posterior parameter estimation is accurate and reasonably conservative under complex admixture scenarios. For both admixed populations, we find that monotonically decreasing contributions over time, from Europe and Africa, explain the observed data more accurately than multiple admixture pulses. This approach will allow for reconstructing detailed admixture histories when maximum-likelihood methods are intractable.

1 | INTRODUCTION

Hybridization between species and admixture between populations are powerful mechanisms influencing biological evolution. Genetic admixture patterns have thus been extensively studied to reconstruct past population migrations and understand admixture-related adaptation such as heterosis or post-admixture selection (Brandenburg et al., 2017; Hellenthal et al., 2014; Skoglund, Ersmark, Palkopoulou, & Dalen, 2015).

A long history of statistical developments in population genetics provided tools to identify and describe admixture patterns from genetic data (Bernstein, 1931; Cavalli-Sforza & Bodmer, 1971; Chakraborty & Weiss, 1988; Long 1991; Falush, Stephens, & Pritchard, 2003; Patterson et al., 2012), They enabled inferring the ancestral origins of admixed populations or investigate adaptive introgression in numerous species (e.g. Martin et al., 2013; Patin et al., 2017; Stryjewski & Sorenson, 2017).

1.1 | Maximum-likelihood methods to reconstruct admixture histories

Two classes of maximum-likelihood (ML) methods have been extensively deployed to infer admixture histories from genetic data. They rely on the moments of allelic frequency spectrum divergences among populations (Lipson et al., 2013; Patterson et al., 2012; Pickrell & Pritchard, 2012), and on admixture Linkage-Disequilibrium patterns – the distribution of LD within the admixed chunks of DNA inherited from the source populations in the genomes of admixed individuals (Chimusa et al., 2018; Gravel, 2012; Hellenthal et al., 2014; Loh et al., 2013; Moorjani et al., 2011). Notably, Gravel (2012) developed an approach to fit the observed curves of admixture-LD decay to those theoretically expected under admixture models involving one or two pulses of historical admixture. These approaches significantly improved our understanding of past admixture histories using genetic data (e.g. Baharian et al., 2016; Martin et al., 2013).

Despite these major achievements, ML admixture history inference methods suffer from inherent limitations acknowledged by the authors (Gravel, 2012; Hellenthal et al., 2014; Lipson et al., 2013). First, most ML approaches can only consider one or two pulses of admixture in the history of the hybrid population. Nevertheless, admixture processes are often expected to be much more complex, and it is not yet clear how ML methods behave when they can consider only simplified versions of the true admixture history underlying the observed data (Gravel, 2012; Hellenthal et al., 2014; Lipson et al., 2013; Loh et al., 2013; Medina, Thornlow, Nielsen, & Corbett-Detig, 2018; Ni et al., 2019). Second, it is possible to statistically compare ML values obtained from fitting models with different parameters to the observed data, as a guideline to find the "best" model. Nevertheless, formal statistical comparison of the success or failure of competing models to explain the observed data is often out of reach of ML approaches (Foll, Shim, & Jensen, 2015; Gravel, 2012; Ni et al., 2019). Finally, admixture-LD methods, in particular, rely on fine mapping of local ancestry segments in individual genomes and thus require substantial amounts of genomic data, and, sometimes, accurate phasing, which remain difficult in numerous case-studies.

1.2 | Approximate Bayesian Computation demographic inference

Approximate Bayesian Computation (ABC) approaches (Beaumont, Zhang, & Balding, 2002; Tavaré, Balding, Griffiths, & Donnelly, 1997), represent a promising alternative to infer complex admixture histories from genetic data. Indeed, ABC has been successfully used previously to formally test alternative demographic scenarios hypothesized to be underlying observed genetic patterns, and to estimate, *a posteriori* , the parameters of the winning models, when ML methods could not operate (Boitard, Rodriguez, Jay, Mona, & Austerlitz, 2016; Fraimout et al., 2017; Verdu et al., 2009).

ABC model-choice and posterior-parameter estimation rely on comparing observed summary statistics to the same set of statistics calculated from simulations produced under competing demographic scenarios (Beaumont et al., 2002; Blum & François, 2010; Csilléry, François, & Blum, 2012; Pudlo et al., 2016; Sisson, Fan, & Beaumont, 2018; Wegmann, Leuenberger, & Excoffier, 2009). Each simulation, and corresponding vector of summary statistics, is produced using model-parameters drawn randomly from prior distributions explicitly specified by the user. This makes ABC *a priori* particularly well suited to investigate highly complex historical admixture scenarios for which likelihood functions are very often intractable, but for which genetic simulations are feasible (Gravel, 2012; Pritchard et al., 1999; Verdu & Rosenberg, 2011; Buzbas & Verdu, 2018).

1.3 | An ABC framework for reconstructing complex admixture histories

In this paper, we show how ABC can be successfully applied to reconstruct, from genetic data, highly complex admixture histories beyond models with a single or two pulses of admixture classically explored with ML methods. To do so, we propose a novel forward-in-time genetic data simulator and a set of parameter-generator and summary-statistics calculation tools, embedded in an open source C software package called *MetHis* . It performs independent SNPs or microsatellites simulations under any two-source populations versions of the Verdu and Rosenberg (2011) general model of admixture; and is adapted to conduct ABC inferences with existing machine-learning ABC tools implemented in *R* (R Development Core Team, 2017).

We show that our *MetHis* -ABC framework can accurately distinguish major classes of complex historical admixture models, involving multiple admixture-pulses, recurring increasing or decreasing admixture over

time, or combination of these models, and provides conservative posterior parameter inference under chosen models. Furthermore, we introduce the quantiles and higher moments of the distribution of admixture fractions in the admixed population as highly informative summary-statistics for ABC model-choices and posterior-parameter estimations.

We exemplify our approach by reconstructing the complex admixture histories underlying observed genetic patterns separately for the African American (ASW) and Barbadian (ACB) populations. Both populations are known to be admixed populations of European and African descent in the context of the Transatlantic Slave Trade, whose histories of admixture remain largely unknown (e.g. Baharian et al., 2016; Martin et al., 2017). In this case-study, we find that the ACB and ASW populations' admixture histories are much more complex than previously inferred, and further reveal the diversity of histories undergone by these admixed populations during the TAST in the Americas.

2 | MATERIALS AND METHODS

We evaluated how Approximate Bayesian Computation model-choice and posterior parameter estimation performed for reconstructing highly complex historical admixture processes from genetic data. To do so, we chose to work under the two source-populations version of the general mechanistic model of Verdu and Rosenberg (2011) briefly presented in**Supplementary Figure S1** . We introduce a novel software,*MetHis* , for genetic data simulation and summary-statistics calculation for machine-learning ABC inferences under this general model (**Supplementary Note S1**).

We conduct our proof of concept considering nine competing scenarios of complex admixture histories involving multiple admixture pulses, recurring decreasing or increasing admixture, and combinations of these processes (**Figure 1** , **Table 1**). We explore the recent admixture history of two enslaved-African descending populations in the Americas with genome-wide independent SNPs. Beyond this work, the*MetHis* -ABC framework can readily be used to study numerous histories of complex admixture using independent SNP or microsatellite markers (**Supplementary Note S1**).

2.1 | Nine competing complex admixture scenarios

2.1.1 | Founding of the admixed population H

For all scenarios (**Figure 1** , **Table 1**), we chose a fixed time for the founding (generation 0, forward-in-time) of the target admixed population H occurring 21 generations before present, with admixture proportions $s_{\text{Afr},0}$ and$s_{\text{Eur},0}$ from either source population S respectively, African and European in our case, with$s_{\text{Afr},0} + s_{\text{Eur},0} = 1$, and$s_{\text{Afr},0}$ in [0,1]. This corresponds to the first arrival of European permanent settlers in the Americas in the late 15^{th} century, considering 20 or 25 years per generation and the sampled generation born in the 1980s. Note that simulations considering a parameter $s_{\text{Afr},0}$ close to 0, or alternatively 1, correspond to founding of the population H from either one source population, therefore delaying the first "real" genetic admixture event to the next admixture event. Following founding, we consider three alternative scenarios for the admixture contribution of each source population S separately.

2.1.2 | Admixture-pulse(s) scenarios

For a given source population S, African or European, scenarios*S-2P* consider two possible pulses of admixture into population H occurring respectively at time $t_{\text{S,p1}}$ and$t_{\text{S,p2}}$ distributed in [1,20] with$t_{\text{S,p1}}$ [?] $t_{\text{S,p2}}$, with associated admixture proportion $s_{\text{S,tS,p1}}$ and$s_{\text{S,tS,p2}}$ in [0,1] satisfying, at all times$t$,$\sum_{S\in(Afr,Eur)} s_{S,t} \leq 1$(**Figure 1** , **Table 1**). Note that for one of either$s_{\text{S,t}}$ values close to 0, the two-pulse scenarios are equivalent to single pulse scenarios after the founding of H. Furthermore, for both $s_{\text{S,t}}$ values close to 0, scenarios *S-2P* are nested with scenarios where only the founding admixture pulse 21 generations ago is the source of genetic admixture. Alternatively, $s_{\text{S,t}}$ parameter values close to 1 consider a virtual complete replacement of population H by population S at that time, thus obliterating all previous admixture event.

2.1.3 | Recurring decreasing admixture scenarios

For a given source population S, scenarios *S-DE* consider a recurring monotonically decreasing admixture from population S at each generation between generation 1 (after founding at generation 0) and generation 20 (sampled population) (**Figure 1** , **Table 1**). In these scenario, $s_{S,g}$, with g in [1..20], are the discrete numerical solutions of a rectangular hyperbola function over the 20 generations of the admixture process until present as described in **Supplementary Note S2** . In brief, this function is determined by parameter u_S, the "steepness" of the curvature of the decrease, in [0,1/2],$s_{S,1}$, the admixture proportion from population S at generation 1 (after founding), in [0,1], and$s_{S,20}$, the last admixture proportion in the present, in [0,$s_{S,1}/3$]. Note that we chose the boundaries for $s_{S,20}$ in order to reduce the parameter space and nestedness among competing scenarios, by explicitly forcing scenarios *S-DE* into a substantially decreasing admixture process. Furthermore, note that parameter $u_S$values close to 0 create pulse-like scenarios of intensity$s_{S,1}$ occurring immediately after founding, followed by constant recurring admixture of intensity$s_{S,20}$ at each generation until present. Alternatively, parameter u_S values close to 1/2 create scenarios with a linearly decreasing admixture between$s_{S,1}$ and $s_{S,20}$ from population S at each generation after founding.

2.1.4 | Recurring increasing admixture scenarios

For a given source population S, scenarios *S-IN* mirrors the*S-DE* scenarios by considering instead a recurring monotonically increasing admixture from population S (**Figure 1** , **Table 1**). Here, $s_{S,g}$, with g in [1..20], are the discrete numerical solutions of the same function as in the S-DE decreasing scenarios (see above), flipped over time between generation 1 and 20. In these scenarios, $s_{S,20}$ is defined in [0,1] and $s_{S,1}$ in [0,$s_{S,20}/3$], and u in [0,1/2] parametrizes the "steepness" of the curvature of the increase. Note the analogous nestedness of recurring and pulse-like scenarios over the parameter space of u values as previously.

2.1.5 | Combining admixture scenarios from either source populations

We combine these three scenarios to obtain nine alternative scenarios for the admixture history of population H (**Figure 1** ,**Table 1**), with the only condition that, at each generation$g$ in [1..20], parameters satisfy$s_{Afr,g} + s_{Eur,g} + h_g = 1$, with h_g, in [0,1] being the remaining contribution of the admixed population H to itself at generation g .

Four scenarios (Afr2P-EurDE, Afr2P-EurIN, AfrDE-Eur2P, and AfrIN-Eur2P) consider a mixture of pulse-like and recurring admixture from each source. Three scenarios (Afr2P-Eur2P, AfrDE-EurDE, and AfrIN-EurIN), consider symmetrical classes of admixture scenarios from either source. Two scenarios (AfrIN-EurDE and AfrDE-EurIN) consider mirroring recurring admixture processes. Importantly, this scenario design considers nested historical scenarios in specific parts of the parameter space.

2.2 | Forward-in-time simulations with *MetHis*

Simulation of independent genetic markers under highly complex admixture histories is often not trivial under the coalescent and using classical existing software, as the coalescent generally assumes a different pedigree for each independent locus instead of a single pedigree having, in reality, produced all observed gene genealogies (see Wakeley, King, Low, & Ramachandran, 2012). In this context, and because pedigrees are rarely known *a priori* , we developed *MetHis* , a C open-source software package available at*https://github.com/romain-laurent/MetHis.MetHis* simulates independent SNPs or microsatellite markers in an admixed population H under any version of the two-source populations general model from Verdu and Rosenberg (2011), and calculates summary-statistics of interest to the study of complex admixture processes (**Supplementary Note S1**).

2.2.1 | Simulating the admixed population, Effective population size and sampling individuals

At each generation, *MetHis* performs simple Wright-Fisher (Fisher, 1922; Wright, 1931) forward-in-time simulations, individual-centered, in a panmictic population of diploid effective size N_g . For a given individual in the population H at the following generation (g + 1), *MetHis*independently draws each parent from the source populations with probability $s_{S,g}$ (**Figure 1** , **Table 1**), or from population H with probability$h_g = 1 - \sum_{\Sigma\ (Afr,Eur)} s_{S,g}$, randomly builds a haploid gamete of independent markers for each parent, and pairs the two constructed gametes to create the new individual.

Here, we decided to neglect mutation over the 21 generations of admixture considered. This is reasonable when studying relatively recent admixture histories and considering independent genotyped SNP markers. Nevertheless, for users interested in microsatellite variation and longer admixture histories, *MetHis* readily implements a standard General Stepwise Mutation Model allowing for insertion or deletion (Estoup, Jarne, & Cornuet, 2002), with parameters set by the user (**Supplementary Note S1**).

To focus on the admixture process itself without excessively inflating the parameter space, we consider, for each nine-competing model, the admixed population H with constant effective population size$N_g = 1000$ diploid individuals. Nevertheless, note that *MetHis* readily allows the user parametrization of stepwise or continuous changes in *Ne* (**Supplementary Note S1**).

After each simulation, we randomly draw individual samples matching sample-sizes in our observed dataset (see 2.4.3). We sample individuals until our sample set contains no individuals related at the 1^{st} degree cousin within each population and between population H and either source populations, based on explicit parental flagging during the last 2 generations of the simulations. Note that this is done to best mimic, *a priori* , the observed case-studies dataset, but excluding related individuals is an option set by the user in *MetHis* (**Supplementary Note S1**).

2.2.2 | Simulating source populations

MetHis , in its current form, does not allow simulating the source populations for the admixture process modeled in Verdu and Rosenberg (2011). Simulating source populations can be done separately using existing genetic data simulation software such as *fastsimcoal2*sequential coalescent (Excoffier, Dupanloup, Huerta-Sanchez, Sousa, & Foll, 2013; Excoffier & Foll, 2011).

Another possibility to simulate source populations emerges if genetic data is already available for the known source populations, as it is the case in our case studies of enslaved-African descendants in the Americas (see 2.4.3). We consider here that the African and European source populations are very large populations at the drift-mutation equilibrium, accurately represented by the Yoruban YRI and British GBR datasets here investigated (see 2.4.3). Therefore, we first build two separate datasets each comprising 20,000 haploid genomes of 100,000 independent SNPs, each SNP being randomly drawn in the site frequency spectrum (SFS) observed for the YRI and GBR datasets respectively. These two datasets are used as fixed gamete reservoirs for the African and European sources separately, at each generation of the forward-in-time admixture process. From these reservoirs, we build an effective individual gene-pool of diploid size N_{g}, by randomly pairing gametes avoiding selfing. These virtual source populations provide the parental pool for simulating individuals in the admixed population H with *MetHis* , at each generation. Thus, while our gamete reservoirs are fixed, the parental genetic pools are randomly built anew at each generation. Again, note that this is not necessary to the implementation of *MetHis* for investigating complex admixture histories; source populations can be simulated separately by the user at will.

2.3 | Summary Statistics

MetHis is designed to work in an ABC inference framework and, thus, can calculate numerous summary-statistics. A complete list of summary-statistics can be found in **Supplementary note S1** . Below are the summary-statistics considered in our case-study, in particular introducing the distribution of admixture fractions in population H, as summary-statistics for ABC inference.

2.3.1 | Distribution of admixture fractions as a set of summary-statistics

Most methods developed to estimate individual admixture fractions from genetic data (e.g. Alexander et al., 2009), are computationally intensive, which is out-of-reach when iterated for large sets of simulated genetic data. This explains why they are not routinely used in ABC inferences, despite being theoretically highly informative (Gravel, 2012; Verdu & Rosenberg, 2011).

Here, we propose, and implement in *MetHis* , an efficient way to use estimated individual admixture fractions as summary statistics for ABC inferences, based on allele-sharing-dissimilarity (ASD) (Bowcock et

al., 1994) and multidimensional scaling (MDS). For each simulated dataset, we first calculated a pairwise inter-individual ASD matrix using *asd*software (*https://github.com/szpiech/asd*) using all pairs of sampled individuals and all markers. Then we projected in two dimensions this pairwise ASD matrix with classical unsupervised metric MDS using the *cmdscale* function in *R* . We expect individuals in population H to be dispersed along an axis joining the centroids of the proxy source populations on the two-dimensional MDS plot. We projected population H individuals orthogonally onto this axis, and calculate each individual's relative distance to each centroid. We considered this measure as an estimate of individual average admixture level from either source. Note that by doing so, some individuals might show "admixture fractions" higher than one, or lower than zero, as they might be projected on the other side of the centroid when being genetically close to 100% from one source population or the other. Under an ABC framework, this is not a difficulty since this may happen also with the real data *a priori* , and ABC goal is to use summary statistics that mimic the observed ones.

This individual admixture estimation method has been shown to be highly concordant with cluster membership fractions as estimated with STRUCTURE or ADMIXTURE (Falush, Stephens, & Pritchard, 2003; Alexander, Novembre, & Lange, 2009) in real data analyses (e.g. Verdu et al., 2017). We confirm these previous findings since we obtain a Spearman correlation (calculated using the *cor.test* function in *R*), of rho = 0.950 (p-value $< 2.10^{-16}$) and rho= 0.977 (p-value $< 2.10^{-16}$) between admixture estimates based on ASD-MDS and on ADMIXTURE, for the two case-study datasets here explored (**Supplementary Figure S2**).

We used the mean, mode, variance, skewness, kurtosis, minimum, maximum, and all 10%-quantiles of the admixture distribution in population H, as 16 separate summary statistics for ABC inference.

2.3.2 |Within population summary statistics

We calculated marker by marker heterozygosities (Nei, 1978), and we considered the mean and variance of this quantity across markers in the admixed population as two separate summary statistics for ABC inference. In addition, we considered the mean and variance of ASD values across pairs of individuals within population H.

2.3.3 | Between populations summary statistics

We calculated multilocus pairwise F_{ST}(Weir & Cockerham, 1984) between population H and each source population respectively. Furthermore, we calculated the mean ASD between individuals in population H and, separately, individuals in either source population. Finally, we calculated the$f_3$statistics (Patterson et al., 2012).

2.4 | Approximate Bayesian Computation

MetHis provides, as outputs, scenarios-parameter vectors and corresponding summary-statistics vectors in reference tables ready to be used with the machine-learning ABC *abc*(Csilléry et al., 2012), and *abcrf*(Pudlo et al., 2016; Raynal et al., 2019) *R* packages.

2.4.1 | Simulating by randomly drawing parameter values from prior distributions

We performed *MetHis* simulations under each nine competing scenarios (**Figure 1**), drawing the corresponding model-parameters in prior-distributions detailed in **Table 1** and automatically generated by *MetHis* parameter-generator tools.

2.4.2 | Complex admixture model-choice with Random-Forest ABC

For ABC model-choice, we performed 10,000 independent *MetHis*simulations for each nine competing-scenarios. To mimic our case study datasets (see 2.4.3), we simulated 100,000 SNPs and sampled 50 individuals in population H, and 90 and 89 individuals respectively in the African and European source populations. Using 27 cores and the above design, we performed the 90,000 simulations with *MetHis* in four days, with 2/3 of that time for summary-statistics calculation only (**Supplementary Note S1**).

We used Random-Forest ABC for model-choice implemented in the*abcrf* function of the *abcrf* package to obtain the cross-validation table and associated prior error rate using an out-of-bag approach (**Figure 2**). We considered a uniform prior probability for the nine competing models. We considered 1,000 decision trees in the forest after visually checking that error-rates converged appropriately (**Supplementary Figure S3**), using the*err.abcrf* function. RF-ABC cross-validation procedures using groups of scenarios were conducted using the group definition option in the *abcrf* function (Estoup, Raynal, Verdu, & Marin, 2018). Finally, each summary statistics relative importance to the model-choice cross-validation was computed using the*abcrf* function (**Supplementary Figure S4**).

We explore model-choice erroneous assignation due to model nestedness in the parameter space, by considering 1,000 randomly chosen simulation per model as pseudo-observed data (**Supplementary Figure S5**). We train the RF algorithm based on the 9000 remaining simulations per model using the *abcrf* function similarly as above, which provides highly similar results as when considering 10,000 simulations per model (results not shown). We then use the *predict.abcrf* function to perform model choice independently for each 1000 simulated pseudo-observed data with known parameter vectors.

To empirically evaluate the power of the RF-ABC model-choice to distinguish complex admixture processes, we conducted similar cross-validations procedures based on additional 10,000 per scenario for 50,000 and, separately, 10,000 SNPs, instead of 100,000 SNPs (180,000 simulations in total, **Supplementary Figure S6A-B**).

Furthermore, using 100,000 SNPs, we produced 90,000 simulations and performed cross-validations (**Supplementary Figure S6C**), considering a five-times smaller sample set, with 10 sampled individuals in population H (instead of 50 as previously) and 18 individuals in each source population (instead of 90 and 89).

2.4.3 | Case-study population genetics datasets

We investigate, as two separate case studies, the admixture histories of the African American (ASW) and Barbadian (ACB) population samples from the 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium, 2015). Previous studies identified, within the same database, the West European Great-Britain (GBR) and the West African Yoruba (YRI) populations as reasonable proxies for the sources of both ACB and ASW, consistently with the macro-history of the Transatlantic Slave-Trade (Baharian et al., 2016; Martin et al., 2017; Verdu et al. 2017).

Samples in the 1000 Genomes Project were *a priori* sampled to be family unrelated. To avoid confounding factors due to cryptic relatedness in our sample compared to *MetHis* simulations, we excluded individuals more closely related than first-degree cousins in the four populations separately using RELPAIR (Epstein, Duren, & Boehnke, 2002), as previously done (Verdu et al. 2017). We also excluded the three ASW individuals showing traces of Native American or East-Asian admixture, as reported in previous studies (Martin et al., 2017). Among the remaining individuals we randomly drew 50 individuals in the targeted admixed ACB and ASW, respectively, and included the remaining 90 YRI individuals and 89 GBR individuals.

We extracted biallelic polymorphic sites (SNPs as defined by the 1000 Genomes Project Phase 3) from the merged ACB+ASW+GBR+YRI data set, excluding singletons. Since *MetHis* only simulates independent markers, we LD-pruned the ACB and ASW SNP-sets using the PLINK (Purcell et al., 2007) –indep-pairwise option with a sliding window of 100 SNPs, moving in increments of 10 SNPs, and r^2 threshold of 0.1. Finally, we randomly drew 100,000 SNPs from the remaining SNP-set.

2.4.4 | Prior-checking of simulations' fit to the ACB and ASW data

We plotted each prior summary statistics distributions and visually verified that the observed summary statistics for the ACB and ASW respectively fell within the simulated distributions (**Supplementary Figure S7**). Then, we explored the first four PCA axes computed with the *princomp* function in *R* , based on the 24 summary statistics and all 90,000 simulations, and visually checked that observed summary statistics were

within the cloud of simulated statistics (**Supplementary Figure S8**). Finally, we performed a goodness-of-fit approach using the *gfit* function from the *abc* package in *R* , with 1,000 replicates and tolerance level 0.01 (**Supplementary Figure S9**).

2.4.5 | RF-ABC model-choice for the admixture history of ACB and ASW populations

For the ACB and ASW observed data separately, we performed model-choice prediction and estimation of posterior probabilities of the winning model using the *predict.abcrf* function in the *abcrf* package, using the complete simulated reference table for training the Random-Forest algorithm (100,000 SNPs, 50 individuals in population H, 90 and 89 individuals in the African and European sources respectively) (**Figure 3** , **Supplementary Table S1**).

2.4.6 | Posterior parameter estimation with Neural-Network ABC

It is difficult to estimate jointly the posterior distribution of all model parameters with RF-ABC (Raynal et al., 2019). Furthermore, although RF-ABC performs satisfactorily well with an overall limited number of simulations under each model (Pudlo et al., 2016), posterior parameter estimation with other ABC approaches, such as simple rejection (Pritchard et al., 1999), regression (Beaumont et al., 2002; Blum & François, 2010) or Neural-Network (NN) (Csilléry et al., 2012), require substantially more simulations *a priori* . Therefore, we performed, for posterior parameter estimations, 90,000 additional simulations, for a total of 100,000 simulations under the best scenarios identified with RF-ABC for the ACB and ASW separately. For comparison purposes, we performed 100,000 simulations under the loosing scenario Afr2P-Eur2P (see **Results**), and conducted anew the below parameter estimation and error evaluation procedures for this scenario.

2.4.7 | Neural-Network tolerance level and number of neurons in the hidden layer

We determined empirically the NN tolerance level (i.e. the number of simulations to be included in the NN training), and number of neurons in the hidden layer. Indeed, while the NN needs a substantial amount of simulations for training, there is also a risk of overfitting posterior parameter estimations when considering too large a number of neurons in the hidden layer. However, there are no absolute rules for choosing both numbers (Csilléry et al., 2012; Jay, Boitard, & Austerlitz, 2019).

Therefore, we tested four different tolerance levels to train the NN for parameter estimation (0.01, 0.05, 0.1, and 0.2), and a number of neurons ranging between four and seven (the number of free parameters in the winning scenarios, see **Results**). For each pair of tolerance level and number of neurons, we conducted cross-validation with 1,000 randomly chosen simulated datasets in turn used as pseudo-observed data with the "*cv4abc* " function in the package *abc* . We considered the median point-estimate of each posterior parameter$\left(\hat{\theta}_i\right)$ to be compared with the true parameter value used for simulation (θ_i). The cross-validation parameter prediction error was then calculated across the 1,000 separate posterior estimations for pseudo-observed datasets for each pair of tolerance level and number of neurons, and for each parameter θ_i, as$\frac{\sum_1^{1000}\left(\hat{\theta}_i-\theta_i\right)^2}{(1000\times Variance(\theta_i))}$, using the *summary.cv4abc* function in *abc*(Csilléry et al., 2012). Results showed that, *a priori* , all numbers of neurons considered performed very similarly for a given tolerance level (**Supplementary Table S2**). Furthermore, results showed that considering 1% closest simulations to the pseudo-observed ones reduced the average error for each tested number of neurons. Thus, we decided to opt for four neurons in the hidden layer and a 1% tolerance level for training the NN in all subsequent parameter inference, in order to avoid overfitting.

2.4.8 | Estimation of model-parameters posterior distributions for ACB and ASW

We jointly estimated model-parameters posterior distributions for the ACB and ASW admixed population separately, using NN-ABC *neuralnet*methods' option in the package *abc* , based on the logit-transformed ("*logit* " transformation option) summary statistics using a 1% tolerance level and four neurons in the hidden layer (**Figure 4** , **Table 2**).

2.4.9 | Posterior parameter estimation error

We wanted to evaluate the posterior error performed by the NN-ABC approach in the vicinity of our observed data rather than randomly on the entire parameter space. To do so, we first identified the 1,000 simulations closest to the real data with a tolerance level of 1%, for the ACB and ASW respectively. Then, we performed 1,000 separate NN-ABC parameter estimations similarly parameterized as above, using in turn the other 99,999 simulations as reference table, and recorded the median point estimate for each parameter. We then compared each parameter estimate with the true parameter used for each 1,000 pseudo-observed target and provide three types of error measurements in **Table 3** . The mean-squared error scaled by the variance of the true parameter $\frac{\sum_1^{1000}\left(\hat{\theta}_i-\theta_i\right)^2}{(1000\times Variance(\theta_i))}$ as previously (Csilléry et al. 2012); the mean-squared error $\frac{\sum_1^{1000}\left(\hat{\theta}_i-\theta_i\right)^2}{1000}$, to compare errors for a given scenario-parameter between the ACB and ASW analyses; and the mean absolute error $\frac{\sum_1^{1000}\left|\hat{\theta}_i-\theta_i\right|}{1000}$, which provides a more intuitive parameter estimation error. For comparison, we conducted the above analysis using instead parameters estimated under the loosing scenario Afr2P-Eur2P (**Supplementary Table S3**).

2.4.10 | 95% credibility interval accuracy

We evaluated *a posteriori* if, in the vicinity of the two observed datasets respectively, the lengths of the estimated 95% credibility intervals (CI) for each parameter was accurately estimated or not (e.g. Jay et al., 2019). To do so, we calculated how many times the true parameter(θ_i) was found inside the estimated 95% CI [2.5%quantile$\left(\hat{\theta}_i\right)$; 97.5%quantile$\left(\hat{\theta}_i\right)$], among the 1,000 out-of-bag NN-ABC posterior parameter estimations (**Supplementary Table S4**). For each parameter, if less than 95% of the true parameter values are found inside the 95% CI estimated for the observed data, we consider the length of this credibility interval as underestimated indicative of a non-conservative behavior of the parameter estimation. Alternatively, if more than 95% of the true parameter-values are found inside the estimated 95% CI, we consider its length as overestimated, indicative of an excessively conservative behavior of parameter estimation. For comparison, we conducted the above analysis using instead parameters estimated under the loosing scenario Afr2P-Eur2P (**Supplementary Table S5**).

2.4.11 | Comparing the accuracy of posterior parameters estimations using NN, RF, or Rejection ABC

We compared four ABC posterior parameter estimation methods: NN-ABC estimation of the parameters taken jointly as a vector (similarly as in the above procedures), NN-ABC estimation of the parameters taken in turn separately, RF-ABC estimation of the parameters which also considers parameters in turn and separately (Raynal et al., 2019), and simple Rejection ABC estimation for each parameter separately (Pritchard et al., 1999). For each method, we used in turn the 1,000 simulations closest to the real data as pseudo-observed data and the 99,999 remaining simulations as reference table. We consider the same parameters for the NN, and we 500 decision trees for the RF to limit the computational cost at little accuracy cost *a priori* (**Supplementary Figure S3**). We computed the three types of errors and the accuracies of the 95% CI for each ABC method similarly as previously (**Table 4**).

3 | RESULTS

3.1 | Complex admixture scenarios cross-validation with RF-ABC

We trained the RF-ABC model-choice algorithm using 1,000 trees, which guaranteed the convergence of the model-choice prior error rates (**Supplementary Figure S3**). Based on this training, the complete out-of-bag cross-validation matrix showed that the nine competing scenarios of complex historical admixture could be relatively reasonably distinguished despite the high level of nestedness of the scenarios here considered (**Figure 2**). Indeed, we calculated an out-of-bag prior error rate of 32.41%, considering each 90,000 simulation, in turn, as out-of-bag pseudo-observed target dataset, compared to a prior probability of 88.89% to erroneously select a scenario. Furthermore, we found the posterior probabilities of identifying the correct scenario ranging from 55.17% (prior probability = 11.11% for each competing scenario), for the two-pulses scenarios from both the African and European sources (Afr2P-Eur2P), to 77.71% for the scenarios considering monotonically decreasing recurring admixture from both sources (AfrDE-EurDE).

Importantly, the average probability, for a given admixture scenario, of choosing any one alternative (wrong) scenario were on average 4.05% across the eight alternative scenarios, ranging from 2.79% for the AfrDE-EurDE scenario, to 5.60% for the Afr2P-Eur2P scenario (**Figure 2**). This shows that our approach did not systematically favor one or the other competing scenario when wrongly choosing a scenario instead of the true one. Furthermore, note that Afr-DE-EurDE scenarios were rarely confused (3.8%) with other recurring admixture scenarios containing at least one recurring admixture increase (AfrIN-EurDE, AfrDE-EurIN, AfrIN-EurIN), which shows a strong discriminatory power of RF-ABC model-choice *a priori* , even among complex recurring admixture scenarios.

In cross-validation analyses of groups of scenarios (Estoup et al., 2018), monotonically recurring admixture scenarios (AfrDE-EurDE, AfrDE-EurIN, AfrIN-EurDE, AfrIN-EurIN) can be well distinguished from scenarios considering two possible pulses after the founding event (Afr2P-Eur2P, Afr2P-EurDE, Afr2P-EurIN, AfrDE-Eur2P, AfrIN-Eur2P). Indeed, we found an out-of-bag prior error rate of 13.85%, and posterior cross-validation probabilities of identifying the correct group of scenarios of 86.08% and 86.23% respectively for the two groups.

Detailed investigation of cross-validation results shows that inaccuracies of RF-ABC model-choices occur mainly in parameter-spaces where scenarios are highly nested and, in fact, close biologically (**Figure 2**). As expected, model-choice increasingly mistakes the AfrDE-EurDE scenarios for scenarios containing two admixture pulses (Afr2P-Eur2P, Afr2P-EurIN, AfrIN-Eur2P) as values ofu_{Afr} and u_{Eur} are closer to 0, regardless of introgression rates values (**Supplementary Figure S5A**). Intuitively, the closer these parameter values are to 0, the more peaked the decrease of recurring admixture are, which increases model-choice confusion with pulse-like scenarios. Instead,u -values closer to 0.5 correspond to linearly decreasing admixture over time which are hardly confounded with pulse-like scenarios. Furthermore, the model-choice increasingly confuses, as expected regardless of introgression values, Afr2P-Eur2P scenarios with recurring increasing admixture scenarios (AfrIN-EurIN, AfrDE-EurIN, AfrIN-EurDE), as the time of the second admixture pulses from Europe or Africa are recent (**Supplementary Figure S5B**).

Most importantly, RF-ABC model-choice power to discriminate among complex admixture processes *a priori* was not strongly affected by the numbers of markers considered. Indeed, we found an out-of-bag prior error of 33.53% and 37.93% (instead of 32.41%), considering respectively 50,000 and 10,000 SNPs, instead of 100,000, together with a very similar distribution of correct and mistaken predictions among scenarios (**Supplementary Figure S6A-B**). Finally, dividing by five the sample sizes in population H and each source populations increased, as expected, the cross-validation error rate (48.39%). Nevertheless, all scenarios continue to be correctly identified three to six times more often than expected *a priori* , and the distribution of erroneous predictions remained similar to previously (**Supplementary Figure S6C**). Altogether, these results showed that RF-ABC model-choice can be successfully used to distinguish highly complex admixture models even when substantially less genetic and sample data are considered.

3.2 | Simulating the observed data with *MetHis*

Using *MetHis* , we produced 90,000 vectors of 24 summary statistics each, overall highly consistent with the observed ones for the ACB and the ASW populations respectively. First, each observed statistic is visually reasonably well simulated under the nine competing scenarios here considered (**Supplementary Figure S7**). Second, the observed data each fell into the simulated sets of 24 summary statistics projected in the first four PCA dimensions (**Supplementary Figure S8**). Finally, the observed summary statistics vectors were not significantly different (p-value = 0.468 and 0.710, for the ACB and ASW respectively) from the simulated ones using a goodness-of-fit approach (**Supplementary Figure S9**). Therefore, we successfully simulated datasets producing sets of summary statistics reasonably close to the observed ones, despite considering constant effective population sizes, fixed virtual source population genetic pool-sets, and neglecting mutation during the admixture process.

3.3 | Random-Forest ABC scenario-choice for the history of ACB and ASW populations

We performed RF-ABC model-choice separately for the admixture history of the Barbadian (ACB) and the

African American (ASW) populations, to evaluate whether our *MetHis* -ABC method could identify subtle differences in the history of both populations having experienced the TAST under the British colonial empire (Martin et al. 2017; Baharian et al. 2016). For the ACB, **Figure 3** shows that the majority of votes (53.1%) went to an admixture scenario AfrDE-EurDE with a posterior probability of the winning scenario of 60.28%. This posterior probability is above the mean posterior-probability obtained when the wrong scenario is chosen for the 1000 AfrDE-EurDE simulations closest to the observed one (56.8%, SD=11.6%, for 37 simulations wrongly assigned in total). The second most chosen scenario was the AfrDE-Eur2P scenario. However, this scenario is voted for 3.5 times less often than the winning scenario AfrDE-EurDE, gathering 15.1% of the 1,000 votes, only slightly above the 11.11% prior probability for each nine-competing scenario (**Figure 3** ; **Supplementary Table S1**).

RF-ABC scenario-choice results were less segregating for the ASW.**Figure 3** shows that the AfrDE-EurDE scenario also gathered the majority of votes, albeit with lower posterior probability than for the ACB (33.5% of 1,000 votes, with posterior probability = 48.0%). This posterior-probability is slightly below the average posterior-probability obtained when the wrong scenario is chosen for the 1000 AfrDE-EurDE simulations closest to the ASW (50.7%, SD = 7.9%, for 192 simulations wrongly assigned). The second most chosen scenario, AfrDE-Eur2P, was only slightly less chosen with 31.7% of the votes (**Figure 3** , **Supplementary Table S1**). Altogether these results denote an ambiguity of the RF-model choice in the part of the parameter-space occupied by the ASW.

Considering only these two best scenarios to train the RF and re-conducting scenario-choice improved the scenario discrimination in favor of the AfrDE-EurDE scenario. While we found, again, only a slight majority of votes (51.8%) in favor of the AfrDE-EurDE scenario, the posterior probability for this model was substantially increased to 57.9%, thus above the average posterior-probability threshold calculated above (50.7%). This indicated that the AfrDE-EurDE scenario best explained the ASW observed genetic patterns, despite overall limited discriminatory power of our approach in the ambiguous part of the summary-statistics space occupied by this population.

3.4 | Neural-Network ABC parameter inference accuracy for the ACB and ASW populations

For the ACB under the AfrDE-EurDE scenario (**Figure 4A** ,**Table 2**), we found that the two recent admixture intensities from Africa and Europe ($s_{\text{Afr,20}}$ and$s_{\text{Eur,20}}$, respectively) and the steepness of the European recurring introgression decrease (u_{Eur}) had sharp posterior densities clearly distinct from their respective priors. Note that the cross-validation error on these parameters in the vicinity of our real data were low (average absolute error 0.02744, 0.0044, and 0.1084, respectively for $s_{\text{Afr,20}}$,$s_{\text{Eur,20}}$, and u_{Eur}) (**Table 3**), and lengths of 95% CI reasonably accurate (96.4%, 94.4%, 94.1% of 1,000 cross-validation true parameter values fell into estimated 95% CI, **Supplementary Table S4**).

Furthermore, the two ancient admixture intensities from Africa and Europe at generation 1 ($s_{\text{Afr,1}}$ and$s_{\text{Eur,1}}$, respectively), also had posterior densities apparently distinguished from their prior distributions, but both had much wider 95% CI (**Figure 4A** , **Table 2**). Consistently, we found a slightly increased posterior parameter error in this part of the parameter space for both parameters, with average absolute error equal to 0.121 and 0.095 respectively for$s_{\text{Afr,1}}$ and $s_{\text{Eur,1}}$(**Table 3**). Nevertheless, note that 95.8% and 94.7% of 1,000 cross-validation true values for those two parameters fell into the estimated 95% CI (**Supplementary Table S4)** . This shows a reasonably conservative behavior of our method for these estimations, albeit indicating that information is lacking in our data or set of summary statistics for a more accurate estimation of these parameters, rather than an inherent inaccuracy of our approach.

Interestingly (**Figure 4A** , **Table 2**), we found that accurate posterior estimation of the steepness of the African recurring introgression decrease (u_{Afr}) is difficult. Indeed, the posterior density of this parameter showed a tendency towards small values only slightly departing from the prior, indicative of a limit of our method to estimate this parameter (**Figure 4A** ,**Table 2**). Finally (**Figure 4A** , **Table 2**), we found that we had virtually no information to estimate the founding admixture proportions from Africa and Europe at generation 0, as our posterior estimates barely departed from the prior and associated mean absolute error

was high (0.2530, **Table 3**). Nevertheless, our method seemed to be performing reasonably conservatively for these two latter parameters (95.6% and 95.3% of 1,000 cross-validation true parameter values fell into estimated 95% CI, **Supplementary Table S4**). This indicates that information is strongly lacking in our data or summary statistics for successfully capturing these parameters, rather than inherent inaccuracy of our approach.

For the ASW under the AfrDE-EurDE model, our posterior parameter estimation results were overall less accurate compared to those obtained for the ACB population, as indicated by overall larger CI and cross-validation errors (**Figure 4B** , **Table 2** ,**Table 3** , **Supplementary Table S4**). This was consistent with the more ambiguous RF-ABC model-choice results obtained for this population (**Figure 3**).

Note that, we conducted the above analyses under the loosing scenario Afr2P-Eur2P instead, for comparison. We find, as expected, that parameters and 95% CI are very poorly estimated for all parameters under this model (**Supplementary Table S3 and S5**). This indicates, consistently, that no information is available in the ACB or ASW data for accurate and conservative estimation of the loosing scenario Afr2P-Eur2P parameters using ABC.

3.5 | Comparing NN, RF, and Rejection ABC posterior parameter estimation accuracy

The three types of posterior parameter estimation errors (scaled mean-square error, mean-square error, absolute error) were systematically lower for the two NN methods (joint or independent posterior parameter estimations) than for the RF and Rejection independent posterior parameter estimations (**Table 4**). Altogether, these results showed that considering the NN estimation for parameters taken jointly as a vector is overall preferable, since it further allowed the joint interpretation of parameter values estimated *a posteriori* , with little accuracy loss.

The lengths of 95% CI estimated with NN joint parameter estimation were, across all parameters, more accurate than those obtained with all other methods with, on average, 95.1% and 95.2% of true parameter values falling within the estimated 95% CI, for the ACB and ASW respectively (**Supplementary Table S3**). Furthermore, the lengths of 95% CI estimated with NN and RF independent posterior parameter estimations were systematically under-estimated, with less than 94% of the true parameter values falling into the estimated 95% CI. Finally, the lengths of 95% CI estimated with the Rejection method were also rather accurately estimated, although on average slightly over-estimated compared to the NN joint estimation with, on average, 95.5% of the 1,000 cross-validation true parameter values within the estimated 95% CI for the ACB, and 95.8% for the ASW.

3.6 | Admixture histories of the African American ASW and Barbadian ACB

Figure 5 visually synthesized the estimated posterior parameters of the complex admixture scenarios reconstructed with the*MetHis* – ABC framework, and associated 95% CI (**Table 2**).

We found a virtual complete replacement of the ACB and ASW populations at generation 1, thus consistent with our inability to accurately estimate the founding proportions from the African and European sources at generation 0. Furthermore, we found an increasingly precise posterior estimation of introgression rates forward-in-time. This is also consistent with the nature of recurrent admixture processes, where older information may be lost or replaced when more recent admixture events occur.

Interestingly, we found that the recurring introgression from the European gene pool rapidly decreases after generation 1, for both the ACB and ASW, albeit with substantial differences (**Figure 5**). Indeed, we found that, for the ACB, European introgression falls below 10% at generation 9 until no more than 1% in the present. Comparatively, the European contribution diminished substantially less rapidly for the ASW, going below 10% only after generation 12 until roughly 2% in the present. Therefore, it seems that neither sustained European migrations, nor the relaxation of social and legal constraints on admixture between descendant communities subsequent to the abolition of slavery and the end of segregation, have translated into increased European genetic contribution to the gene-pool of admixed populations in the Americas.

Finally, we found substantial recurring contributions from the African source for both admixed populations (**Figure 5**). For the ACB, we found a progressive decrease of the African recurring introgression until a virtually constant recurring admixture close to 28% from generation 10 and onward. For the ASW, our results showed a sharper decrease of the African contribution after founding until a virtually constant recurring admixture process close to 24% from generation 5 until present. Nevertheless, the ACB occupy an ambiguous region of the parameter-space, and results should be considered cautiously, as another complex admixture model might more accurately explain this data. Altogether, the signal of substantial ongoing admixture from Africa could stem from the known importance of African recurring forced migrations during the TAST into the Americas; further prompts the influence of African slave descendants migrations within the Americas before and after the end of slavery (Baharian et al., 2016).

4 | DISCUSSION

Our novel *MetHis* forward-in-time simulator and summary-statistics calculator coupled with RF-ABC scenario-choice can distinguish among highly complex admixture histories using genetic data. As expected, scenario-choice errors are particularly made in regions of the parameter space for which models are highly nested (Robert, Mengersen, & Chen, 2010), and, thus, biologically similar. Furthermore, we found that NN-ABC provided accurate and reasonably conservative posterior parameter estimation for numerous parameters of the winning scenario, using human population data as a case-study. Finally, we empirically demonstrated that the moments of the distribution of admixture fractions in the admixed population were highly informative for ABC inference, as expected theoretically (Gravel, 2012; Verdu & Rosenberg, 2011).

Altogether, our results for the two recently-admixed human populations illustrate how our *MetHis* – ABC framework can bring fundamental new insights into the complex demographic history of admixed populations; a framework that can easily be adapted, using *MetHis*(**Supplementary Note S1**), for investigating complex admixture histories when maximum-likelihood methods are intractable.

We considered nine competing scenarios all deriving from the general mechanistic admixture model of Verdu and Rosenberg (2011). While the two-source version of this model can readily be simulated with*MetHis* , it considers $2g$ -1 model parameters (with g the duration of the admixture process), plus effective population sizes parameters and mutation parameters. Estimating jointly all these parameters is out of reach of ML methods, and further likely out of reach of ABC posterior-parameter estimation procedures. However, conducting ABC model-choice for disentangling major classes of relatively simplified admixture processes followed by ABC parameter estimation under the winning model, is flexible enough to bring new insights into the evolutionary history of admixed populations, far beyond all admixture scenarios that can be explored with existing ML methods (Gravel 2012; Hellenthal et al. 2014).

The sample and SNP-set explored here is often out of reach in non-model species. Nevertheless, our results considering vastly reduced SNP or sample sets demonstrate that ABC can remain remarkably accurate to disentangle highly complex admixture processes with much less genetic or sample data. This is due to the fact that ABC relies on the amount of information carried by summary-statistics about model parameters, rather than the absolute amount of genetic data investigated. Therefore, the*MetHis* -ABC framework remains promising to reconstruct complex admixture histories, provided that summary-statistics considered by the user are, *a priori* , informative about model parameters, and that summary-statistics are reasonably well estimated with the observed data. Altogether, large parameter and summary-statistics spaces, lack of information from summary statistics, and scenario nestedness, are well known to affect ABC performances and, thus, imperatively need to be thoroughly evaluated case by case (Csilléry, Blum, Gaggiotti, & Francois, 2010; Robert et al., 2010; Sisson et al., 2018).

To further increase the range of applicability of our *MetHis* -ABC framework, our software readily implements microsatellite markers together with a general stepwise mutation model (Estoup, Jane, & Cornuet, 2002), fully parameterizable by the user (**Supplementary Note S1**). This will allow investigating numerous complex admixture histories, much older than the one here explored, and from non-model species. Even if prior knowledge of the founding date is lacking,*MetHis* users can simply set the founding of the population

in a remote past and implement a second founding event with variable date to be estimated, together with later additional admixture events and other parameters of interest, in the ABC inference. Nevertheless, it is not trivial to predict how old an admixture processes should be to be successfully investigated with ABC (Buzbas and Verdu 2018). Indeed, ancient admixture processes can leave scarcely identifiable signatures in the observed data, if obliterated by more recent admixture events. This was theoretically expected (Buzbas, & Verdu, 2018), and future studies combining ancient and modern DNA samples may bring further information into the ancient admixture history reconstruction.

Importantly, the computational cost of our study depends, for 2/3, on summary statistics calculation at the end of the admixture process, as is often the case in ABC. Considering much longer admixture processes than the ones here investigated will mechanically increase computation time but will not increase summary-statistics calculation time. Furthermore, note that the computational cost of simulating data with *MetHis* does not rely excessively on the number of generations considered (within reason), nor on the absolute number of markers used, but rather on the effective population size in the admixed population set by the user.

Although *MetHis* readily allows considering changes of effective population size in the admixed population at each generation as a parameter of interest to ABC inference (**Supplementary Note S1**), we did not, for simplicity, investigate here how such changes affected our results. Future work using *MetHis* will specifically investigate how effective size changes may influence genetic patterns in admixed populations, a question of major interest as numerous admixed populations have experienced founding events and/or bottlenecks during their genetic history (e.g. Browning et al., 2018).

The current *MetHis* – ABC approach does not make use of admixture linkage-disequilibrium patterns in the admixed population, and only relies on independent SNP or microsatellite markers. Nevertheless, admixture LD has consistently proved to bring massive information about complex admixture histories in populations where large genomic datasets are available (Gravel, 2012; Hellenthal et al., 2014; Malinsky et al., 2018; Medina et al., 2018; Ni et al., 2019; Stryjewski & Sorenson, 2017). However, existing methods to calculate admixture LD patterns remain computationally intensive and require both dense marker-sets and accurate phasing, which is difficult under ABC where such statistics have to be calculated for each one of the numerous simulated datasets. In this context, RF-ABC (Pudlo et al., 2016; Raynal et al., 2019), or AABC (Buzbas & Rosenberg, 2015), methods allow substantially diminishing the number of simulations required for satisfactory ABC inference. This makes both approaches promising tools for using, in the future, admixture-LD patterns to reconstruct complex admixture processes from genomic data.

Finally, future developments of the *MetHis* -ABC framework will focus on implementing sex-specific admixture models, as these processes are known to affect genetic diversity patterns in a specific way, and are of interest to numerous study-cases (Goldberg, Verdu, & Rosenberg, 2014). Furthermore, the *MetHis* forward-in-time simulator represents an ideal tool to further investigate admixture-related selection forces, and admixture-specific assortative matting processes, as these processes can simply be modeled by specifically parameterizing individual reproduction and survival in the simulations, unlike most coalescent-based simulators.

5 | Acknowledgements

We thank Frédéric Austerlitz, Erkan O. Buzbas, Antoine Cools, Flora Jay, Evelyne Heyer, Margueritte Lapierre, Guillaume Laval, Nina Marchi, Etienne Patin, Noah A. Rosenberg, and Zachary A. Szpiech for useful comments and discussions. We warmly thank Olivier Hardy for help designing the microsatellite mutation model implemented in*MetHis* . We thank three anonymous reviewers and the editor for recommendations having improved the article. This project was funded in part by the French Agence Nationale de la Recherche project METHIS (ANR 15-CE32-0009-01). CFL was funded in part by the Sven and Lilly Lawski's Foundation (N2019-0040).

6 | REFERENCES

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature, 526* (7571), 68-74. doi:10.1038/nature15393

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19* (9), 1655-1664. doi:10.1101/gr.094052.109

Baharian, S., Barakatt, M., Gignoux, C. R., Shringarpure, S., Errington, J., Blot, W. J., . . . Gravel, S. (2016). The Great Migration and African-American Genomic Diversity.*PLoS Genet, 12* (5), e1006059. doi:10.1371/journal.pgen.1006059

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics., 162* (4), 2025-2035.

Bernstein, F. (1931). Die geographische Verteilung der Bludgruppen und ihre anthropologische Bedeutung. In *Comitato Italiano per o studio dei problemi della populazione* (pp. 227-243). Roma: Instituto Poligraphico dello Stato.

Blum, M. G. B., & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing, 20* , 63-67. doi:https://doi.org/10.1007/s11222-009-9116-0

Boitard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLoS Genet, 12* (3), e1005877. doi:10.1371/journal.pgen.1005877

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature, 368* (6470), 455-457. doi:10.1038/368455a0

Brandenburg, J. T., Mary-Huard, T., Rigaill, G., Hearne, S. J., Corti, H., Joets, J., . . . Tenaillon, M. I. (2017). Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. *PLoS Genet, 13* (3), e1006666. doi:10.1371/journal.pgen.1006666

Browning, S. R., Browning, B. L., Daviglus, M. L., Durazo-Arvizu, R. A., Schneiderman, N., Kaplan, R. C., & Laurie, C. C. (2018). Ancestry-specific recent effective population size in the Americas. *PLoS Genet, 14* (5), e1007385. doi:10.1371/journal.pgen.1007385

Buzbas, E. O., & Rosenberg, N. A. (2015). AABC: approximate approximate Bayesian computation for inference in population-genetic models. *Theor Popul Biol, 99* , 31-42. doi:10.1016/j.tpb.2014.09.002

Buzbas, E. O., & Verdu, P. (2018). Inference on admixture fractions in a mechanistic model of recurrent admixture. *Theor Popul Biol, 122* , 149-157. doi:10.1016/j.tpb.2018.03.006

Cavalli-Sforza, L. L., & Bodmer, W. F. (1971). *The genetics of human populations* . San Francisco,: W. H. Freeman.

Chakraborty, R., & Weiss, K. M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci U S A, 85* (23), 9119-9123.

Chimusa, E. R., Defo, J., Thami, P. K., Awany, D., Mulisa, D. D., Allali, I., . . . Mazandu, G. K. (2018). Dating admixture events is unsolved problem in multi-way admixed populations. *Brief Bioinform* . doi:10.1093/bib/bby112

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & Francois, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol, 25* (7), 410-418. doi:10.1016/j.tree.2010.04.001

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution, 3* , 475-479.

Estoup, A., Raynal, L., Verdu, P., & Marin, J. M. (2018). Model choice using Approximate Bayesian Computation and Random Forests: analyses based on model grouping to make inferences about the genetic history of Pygmy human populations. *Journal of the Sfds, 159* (3), 167-190.

Estoup, A., Jarne, P., Cornuet, J. M., (2002). Homoplasy and mutation model at microsatellite

loci and their consequences for population genetics analysis. *Mol Ecol 11: 1591-1604.*

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data.*PLoS Genet, 9* (10), e1003905. doi:10.1371/journal.pgen.1003905

Excoffier, L., & Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics, 27* (9), 1332-1334. doi:10.1093/bioinformatics/btr124

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.*Genetics, 164* (4), 1567-1587.

Fisher, R. A. (1922). Darwinian evolution of mutations. *Eugen Rev, 14* (1), 31-34.

Foll, M., Shim, H., & Jensen, J. D. (2015). WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Mol Ecol Resour, 15* (1), 87-98. doi:10.1111/1755-0998.12280

Fraimout, A., Debat, V., Fellous, S., Hufbauer, R. A., Foucaud, J., Pudlo, P., . . . Estoup, A. (2017). Deciphering the Routes of invasion of Drosophila suzukii by Means of ABC Random Forest. *Mol Biol Evol, 34* (4), 980-996. doi:10.1093/molbev/msx050

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics, 191* (2), 607-619. doi:10.1534/genetics.112.139808

Goldberg, A., Verdu, P., & Rosenberg, N.A., (2014). Autosomal admixture levels are informative about sex bias in admixed populations. *Genetics 198* (3), 1209-1229

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry.*Genetics, 196* (3), 625-642. doi:10.1534/genetics.113.160697

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. doi:10.1126/science.1243518

Jay, F., Boitard, S., & Austerlitz, F. (2019). An ABC Method for Whole-Genome Sequence Data: Inferring Paleolithic and Neolithic Human Expansions. *Mol Biol Evol, 36* (7), 1565-1579. doi:10.1093/molbev/msz038

Lipson, M., Loh, P. R., Levin, A., Reich, D., Patterson, N., & Berger, B. (2013). Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol Biol Evol, 30* (8), 1788-1802. doi:10.1093/molbev/mst099

Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics, 193* (4), 1233-1254. doi:10.1534/genetics.112.147330

Long, J. C. (1991). The genetic structure of admixed populations. *Genetics, 127* (2), 417-428.

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADpainter and fineRADstructure: Population Inference from RADseq Data. *Mol Biol Evol, 35* (5), 1284-1290. doi:10.1093/molbev/msy023

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., . . . Kenny, E. E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet, 100* (4), 635-649. doi:10.1016/j.ajhg.2017.03.004

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., . . . Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res, 23* (11), 1817-1828. doi:10.1101/gr.159426.113

Medina, P., Thornlow, B., Nielsen, R., & Corbett-Detig, R. (2018). Estimating the Timing of Multiple Admixture Pulses During Local Ancestry Inference. *Genetics, 210* (3), 1089-1107. doi:10.1534/genetics.118.301411

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., . . . Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet, 7* (4), e1001373. doi:10.1371/journal.pgen.1001373

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics, 89* (3), 583-590.

Ni, X., Yuan, K., Liu, C., Feng, Q., Tian, L., Ma, Z., & Xu, S. (2019). MultiWaver 2.0: modeling discrete and continuous gene flow to reconstruct complex population admixtures.*Eur J Hum Genet, 27* (1), 133-139. doi:10.1038/s41431-018-0259-3

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., . . . Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science, 356* (6337), 543-546. doi:10.1126/science.aal1988

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., . . . Reich, D. (2012). Ancient admixture in human history. *Genetics, 192* (3), 1065-1093. doi:10.1534/genetics.112.145037

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet, 8* (11), e1002967. doi:10.1371/journal.pgen.1002967

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol, 16* (12), 1791-1798. doi:10.1093/oxfordjournals.molbev.a026091

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2016). Reliable ABC model choice via random forests.*Bioinformatics, 32* (6), 859-866. doi:10.1093/bioinformatics/btv684

R Development Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from*http://www.R-project.org*

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference.*Bioinformatics, 35* (10), 1720-1728. doi:10.1093/bioinformatics/bty867

Robert, C. P., Mengersen, K., & Chen, C. (2010). Model choice versus model criticism. *Proc Natl Acad Sci U S A, 107* (3), E5; author reply E6-7. doi:10.1073/pnas.0911260107

Sisson, S. A., Fan, Y., & Beaumont, M. A. (2018). *Handbook of Approximate Bayesian Computation.* (S. A. Sisson, Y. Fan, & M. A. Beaumont Eds.). New York, USA: Chapman and Hall/CRC.

Skoglund, P., Ersmark, E., Palkopoulou, E., & Dalen, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol, 25* (11), 1515-1519. doi:10.1016/j.cub.2015.04.019

Stryjewski, K. F., & Sorenson, M. D. (2017). Mosaic genome evolution in a recent and rapid avian radiation.*Nat Ecol Evol, 1* (12), 1912-1922. doi:10.1038/s41559-017-0364-7

Tavare, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics, 145* (2), 505-518.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Thery, S., . . . Heyer, E. (2009). Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol, 19* (4), 312-318. doi:10.1016/j.cub.2008.12.049

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. *Curr Biol, 27* (16), 2529-2535 e2523. doi:10.1016/j.cub.2017.07.002

Verdu, P., & Rosenberg, N. A. (2011). A general mechanistic model for admixture histories of hybrid populations. *Genetics, 189* (4), 1413-1426. doi:10.1534/genetics.111.132787

Wakeley, J., King, L., Low, B. S., & Ramachandran, S. (2012). Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics, 190* (4), 1433-1445. doi:10.1534/genetics.111.135574

Wegmann, D., Leuenberger, C., & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics, 182* (4), 1207-1218. doi:10.1534/genetics.109.102509

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population-Structure. *Evolution, 38* (6), 1358-1370.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics, 16* (2), 97-159.

7 | DATA ACCESSIBILITY

MetHis software package is open source under the GNU General Public License v3.0, and can be downloaded with manual and example datasets from*https://github.com/romain-laurent/MetHis*

8 | CONFLICT OF INTEREST STATEMENT

Authors declare no conflict of interest for this work.

9 | AUTHOR CONTRIBUTION

CFL: Built the alpha version of the software – Conducted benchmarking and data analyses – Helped writing the article

RL: Built the beta version of the software - Conducted benchmarking and data analyses – Helped writing the article

VT: Conducted benchmarking and data analyses – Helped writing the article

BT: Helped building the beta version of the software - Conducted benchmarking and data analyses – Helped writing the article

PV: Designed and supervised the project – Conducted benchmarking and data analyses – Wrote the article

Figure, Table and Supplementary Material Content:

Main Text: 5 figures, 4 tables.

Supplementary material: 2 notes, 9 figures, 5 tables.

10 | TABLES

Table 1 . Parameter prior distributions for simulation with*MetHis* and Approximate Bayesian Computation historical inference. Parameter list correspond to the nine competing historical admixture models described in **Figure 1** and **Materials and Methods** .

Parameter names	**Parameter description**	**Prior distribution**	**Condition**	**Models**
$s_{\mathrm{Afr},0}$ $s_{\mathrm{Eur},0}$ = 1-$s_{\mathrm{Afr},0}$	Afr. source introgression rate at founding of H	*Uniform* [0,1]	-	all models
$t_{\mathrm{Afr,p1}}$ $t_{\mathrm{Afr,p2}}$	Times of the Afr. source introgression pulses p1 and p2	*Uniform* [0,20]	$t_{\mathrm{Afr,p1}}$ [?] $t_{\mathrm{Afr,p2}}$	Afr2P models
s_{Afr} , tAfr,p1 s_{Afr} , tAfr,p2	Afr. source introgression rates of pulses Afr,p1 and Afr,p2	*Uniform* [0,1]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	Afr2P models
$t_{\mathrm{Eur,p1}}$ $t_{\mathrm{Eur,p2}}$	Times of the Eur. source introgression pulses p1 and p2	*Uniform* [0,20]	$t_{\mathrm{Eur,p1}}$ [?] $t_{\mathrm{Eur,p2}}$	Eur2P models
s_{Eur} , tEur,p1 s_{Eur} , tEur,p2	Eur. source introgression rates of pulses Eur,p1 and Eur,p2	*Uniform* [0,1]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	Eur2P models
$s_{\mathrm{Afr},1}$	Afr. source introgression rate at the first generation after founding	*Uniform* [0,1]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	AfrDE models
$s_{\mathrm{Afr},20}$	Afr. source introgression rate in the present	*Uniform* [0, $s_{\mathrm{Afr},1}$ / 3]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	AfrDE models
u_{Afr}	Steepness of the decrease in Afr. source introgression rates	*Uniform* [0,0.5]	-	AfrDE models
$s_{\mathrm{Eur},1}$	Eur. source introgression rate at the first generation after founding	*Uniform* [0,1]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	EurDE models
$s_{\mathrm{Eur},20}$	Eur. source introgression rate in the present	*Uniform* [0, $s_{\mathrm{Eur},1}$ / 3]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	EurDE models
u_{Eur}	Steepness of the decrease in Eur. source introgression rates	*Uniform* [0,0.5]	-	EurDE models
$s_{\mathrm{Afr},1}$	Afr. source introgression rate at the first generation after founding	*Uniform* [0, $s_{\mathrm{Afr},20}$ / 3]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	AfrIN models
$s_{\mathrm{Afr},20}$	Afr. source introgression rate in the present	*Uniform* [0,1]	For all g, h_{g} = 1 - $s_{\mathrm{Afr,g}}$ - $s_{\mathrm{Eur,g}}$ in [0,1]	AfrIN models

Parameter names	**Parameter description**	**Prior distribution**	**Condition**	**Models**
u_{Afr}	Steepness of the increase in Afr. source introgression rates	*Uniform* [0,0.5]	-	AfrIN models
$s_{\text{Eur},1}$	Eur. source introgression rate at the first generation after founding	*Uniform* [0, $s_{\text{Eur},20}$ / 3]	For all g, $h_{\text{g}} = 1$ - $s_{\text{Afr,g}}$ - $s_{\text{Eur,g}}$ in [0,1]	EurIN models
$s_{\text{Eur},20}$	Eur. source introgression rate in the present	*Uniform* [0,1]	For all g, $h_{\text{g}} = 1$ - $s_{\text{Afr,g}}$ - $s_{\text{Eur,g}}$ in [0,1]	EurIN models
u_{Eur}	Steepness of the increase in Eur. source introgression rates	*Uniform* [0,0.5]	-	EurIN models

Table 2 . Neural-Network Approximate Bayesian Computation posterior parameter weighted distributions under the winning scenario AfrDE-EurDE, for the ACB and ASW populations. All posterior parameter estimations were conducted using 100,000 simulations under scenario AfrDE-EurDE (**Figure 1** , **Table 1**), a 1% tolerance rate (1,000 simulations), 24 summary statistics, logit transformation of all parameters, and 4 neurons in the hidden layer (see **Materials and Methods**).

Admixed population	AfrDE-EurDE parameters	Median	Mean	Mode	95% Credibility Interval
ACB	$s_{\text{Afr},0}$	0.3097	0.3747	0.1121	[0.0116 ; 0.9347]
	$s_{\text{Afr},1}$	0.6797	0.6769	0.6813	[0.4577 ; 0.8880]
	$s_{\text{Afr},20}$	0.2707	0.2655	0.2788	[0.1985 ; 0.2967]
	u_{Afr}	0.1409	0.1684	0.0508	[0.0041 ; 0.4507]
	$s_{\text{Eur},1}$	0.1807	0.2160	0.1158	[0.0542 ; 0.5525]
	$s_{\text{Eur},20}$	0.0100	0.0102	0.0093	[0.0018 ; 0.0200]
	u_{Eur}	0.4858	0.4627	0.4929	[0.1886 ; 0.4992]
ASW	$s_{\text{Afr},0}$	0.5258	0.5124	0.7015	[0.0262 ; 0.9758]
	$s_{\text{Afr},1}$	0.6006	0.6026	0.6081	[0.3506 ; 0.8581]
	$s_{\text{Afr},20}$	0.2352	0.2286	0.2385	[0.1222 ; 0.2714]
	u_{Afr}	0.0662	0.1105	0.0253	[0.0025 ; 0.4393]
	$s_{\text{Eur},1}$	0.2917	0.3080	0.2203	[0.1048 ; 0.5951]
	$s_{\text{Eur},20}$	0.0180	0.0189	0.0157	[0.0022 ; 0.0389]
	u_{Eur}	0.4250	0.3966	0.4567	[0.1077 ; 0.4950]

Table 3. Neural-Network Approximate Bayesian Computation posterior parameter errors under the winning scenario AfrDE-EurDE, for the ACB and ASW populations. For each target population separately, we conducted cross-validation by considering in turn 1,000 separate NN-ABC parameter inferences each using in turn one of the 1,000 closest simulations to the observed ACB (or ASW) data as the target pseudo-observed simulation. All posterior parameter estimations were conducted using 100,000 simulations under scenario AfrDE-EurDE (**Figure 1** , **Table 1**), a 1% tolerance rate (1,000 simulations), 24 summary statistics, logit transformation of all parameters, and four neurons in the hidden layer (see **Materials and Methods**

). Median was considered as the point posterior parameter estimation for all parameters. First column provides the average absolute error; second column shows the mean-squared error; third column shows the mean-squared error scaled by the parameter's observed variance (see **Materials and Methods** for error formulas).

AfrDE-EurDE parameters	**ACB**	**ACB**	**ACB**	**ASW**	**A**
	Av. absolute Error	Mean-square Error	Mean-square Error / Var.	Av. absolute Error	M
$s_{\mathbf{Afr,0}}$	0.2530	0.0857	1.0070	0.2444	0.
$s_{\mathbf{Afr,1}}$	0.1206	0.0216	0.8533	0.1158	0.
$s_{\mathbf{Afr,20}}$	0.02744	0.0012	0.4162	0.0219	0.
$u_{\mathbf{Afr}}$	0.1166	0.0198	0.9974	0.1254	0.
$s_{\mathbf{Eur,1}}$	0.0952	0.0164	1.0526	0.1001	0.
$s_{\mathbf{Eur,20}}$	0.0044	0.0001	0.6452	0.0069	0.
$u_{\mathbf{Eur}}$	0.1084	0.0174	0.9431	0.1021	0.

Table 4. Approximate Bayesian Computation mean posterior parameter errors under the winning Scenario AfrDE-EurDE, for the ACB and ASW populations separately, using four different methods: NN estimation of the parameters taken jointly as a vector, NN estimation of the parameters taken separately, Random Forest (parameters taken separately), and Rejection (parameters taken separately). For each target population separately and for each method, we conducted an out-of-bag cross validation by considering in turn 1,000 separate parameter inferences each using one of the 1,000 closest simulation to the observed ACB (or ASW) data as the target pseudo-observed dataset. All posterior parameter estimations were conducted using the other 99,999 simulations under the AfrDE-EurDE scenario (**Figure 1** ,**Table 1**), a 1% tolerance rate (i.e. 1,000 simulations), 24 summary statistics, logit transformation of all parameters, four neurons in the hidden layer per neural network and 500 trees per random forest. Median was considered as the point posterior parameter estimation for all parameters. First column provides the average absolute error; second column shows the mean-squared error; third column shows the mean-squared error scaled by the parameter's observed variance (see **Materials and Methods** for error formulas).

Posterior parameter estimation ABC method	**ACB**	**ACB**	**ACB**	**A**
	Av. absolute Error	Mean-squared Error	Mean-squared Error / Var.	A
NN joint	0.1037	0.0232	0.8450	0.
NN independent	0.1032	0.0236	0.8294	0.
RF independent	0.1042	0.0246	0.8534	0.
Rejection independent	0.1071	0.0238	0.9299	0.

11 | FIGURES

Figure 1 . Nine competing scenarios for reconstructing the admixture history of African American ASW or Barbadian ACB populations descending from West European and West sub-Saharan African source populations during the Transatlantic Slave Trade. "EUR" represents the Western European and "AFR" represents the West Sub-Saharan African source populations for the admixed population H. See **Table 1**and **Materials and Methods** for model parameter descriptions.

Figure 1

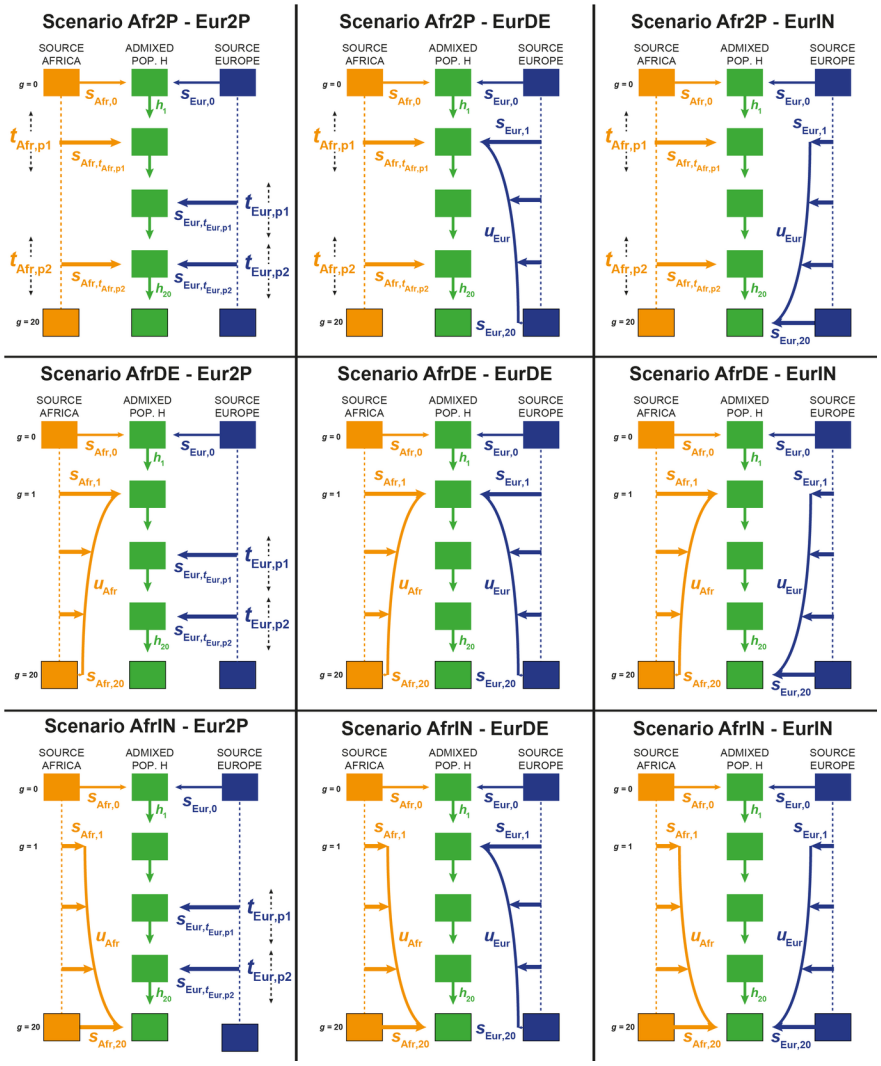

Figure 2 : Random-Forest Approximate Bayesian Computation model-choice cross-validation. Heat map of the out-of-bag cross-validation results considering each 10,000 simulations per each nine competing models (**Figure 1** , **Table 1**) in turn as pseudo-observed target for RF-ABC model-choice. Prior probability of correctly choosing a given scenario is 11%. Out-of-bag prior error rate is 32.41%. RF-ABC model-choice performed using 1,000 decision trees and 24 summary-statistics (see **Materials and Methods**).

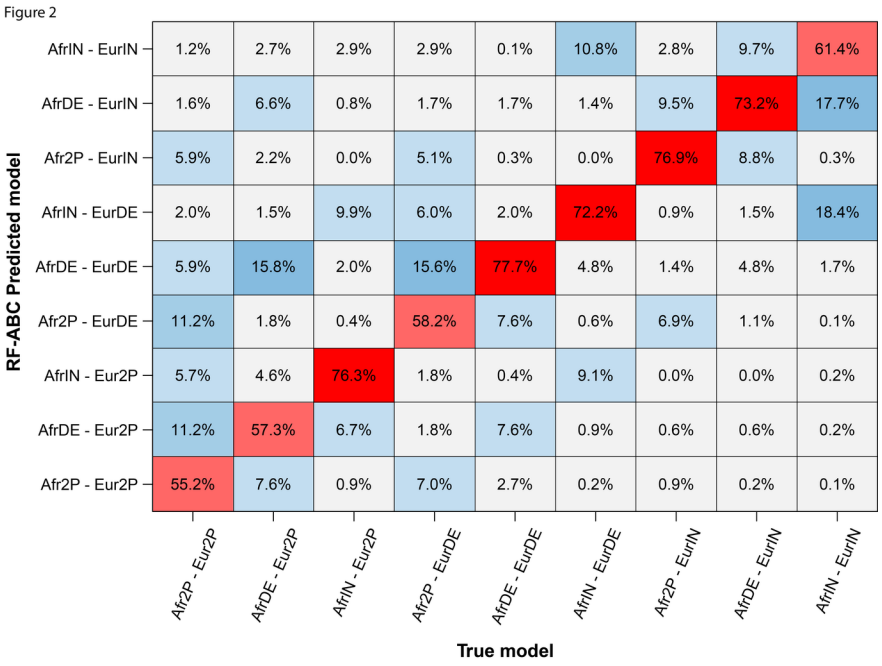

Figure 3 : Random-Forest Approximate Bayesian Computation model-choice predictions for the ACB (left panel) and ASW (right panel) populations. Nine competing models were compared, each with 10,000 simulations (**Figure 1** , **Table 1**). 1,000 decision trees were considered in the model-choice prediction, respectively for each population.

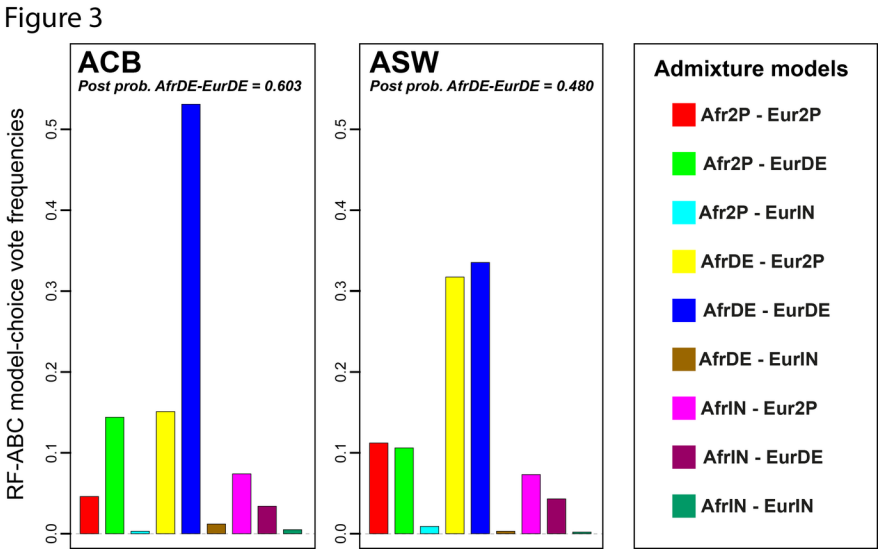

Figure 4 : Neural-Network Approximate Bayesian Computation posterior parameters estimated densities under the winning scenario AfrDE-EurDE, for (A) the ACB and (B) the ASW populations. Median posterior point estimates are indicated by the red vertical line, 95% credibility intervals are indicated by the colored area under the posterior curve (**Table 2**). All posterior parameter estimations were conducted using 100,000 simulations under scenario AfrDE-EurDE, a 1% tolerance rate (1,000 simulations), 24 summary statistics, logit transformation of all parameters, and four neurons in the hidden layer (see **Materials and Methods**).

For all parameters separately, densities are plotted with 1,000 points, a Gaussian kernel, and are constrained to the prior limits. Posterior parameter densities are indicated by a solid line; prior parameter densities are indicated by black dotted lines.

Figure 4

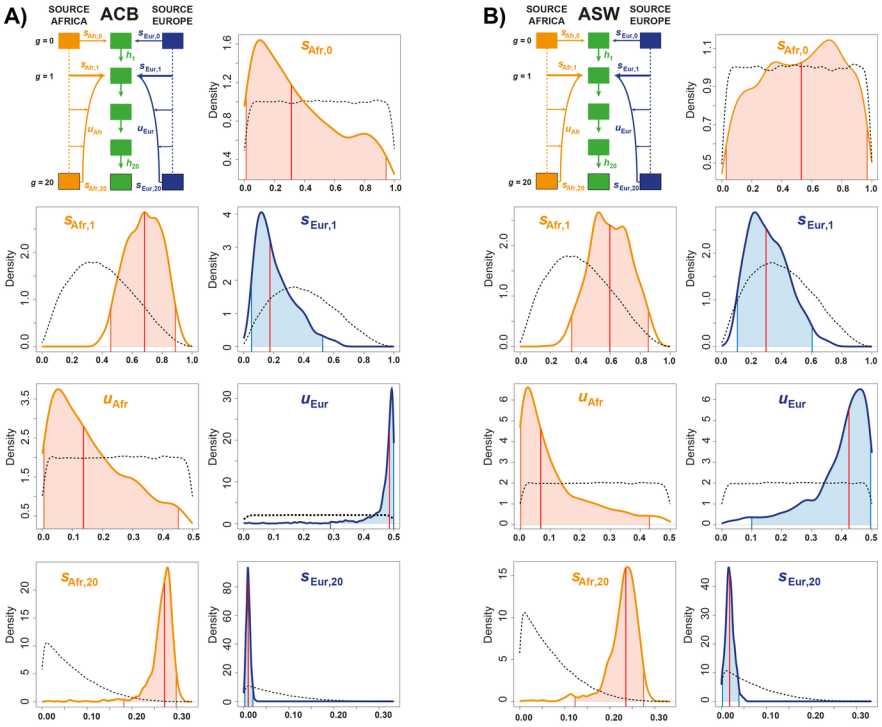

Figure 5 : Approximate Bayesian Computation inference of the admixture history of the ACB and ASW populations respectively. Top panels are based on median point-estimates of intensity parameters at each generation. Bottom panels show 95% credibility intervals for each inferred parameter around the median point-estimates. The African introgression is plotted in orange, the European introgression in blue, and in green the remaining contribution of the admixed population to itself at the following generation. (A) Results for the ACB under the AfrDE-EurDE winning scenario; (B) Results for the ASW under the AfrDE-EurDE winning scenario.

Figure 5

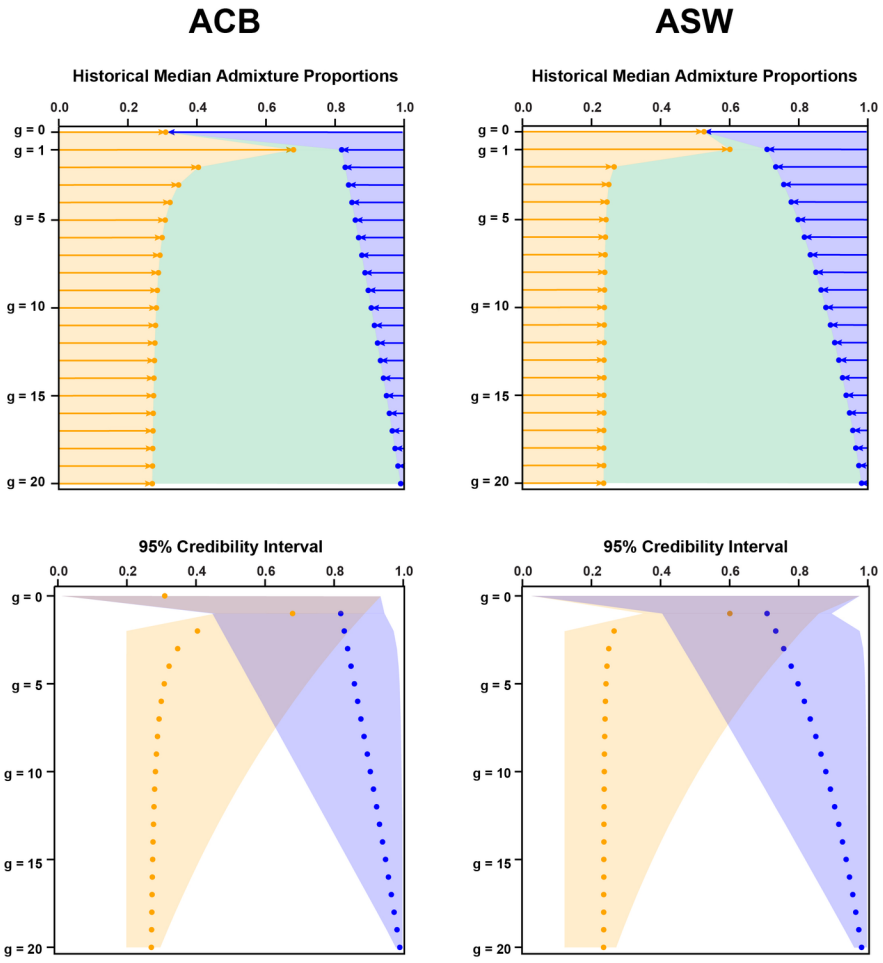

Figure 1

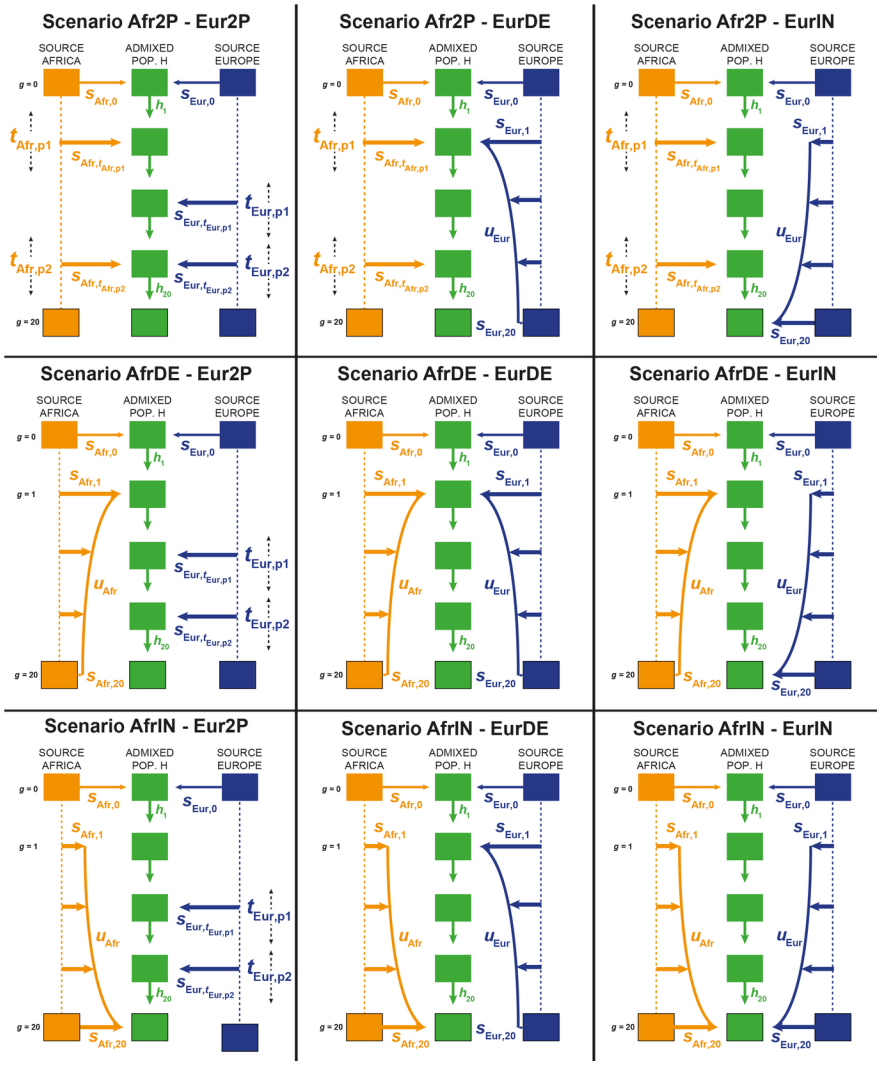

Figure 2

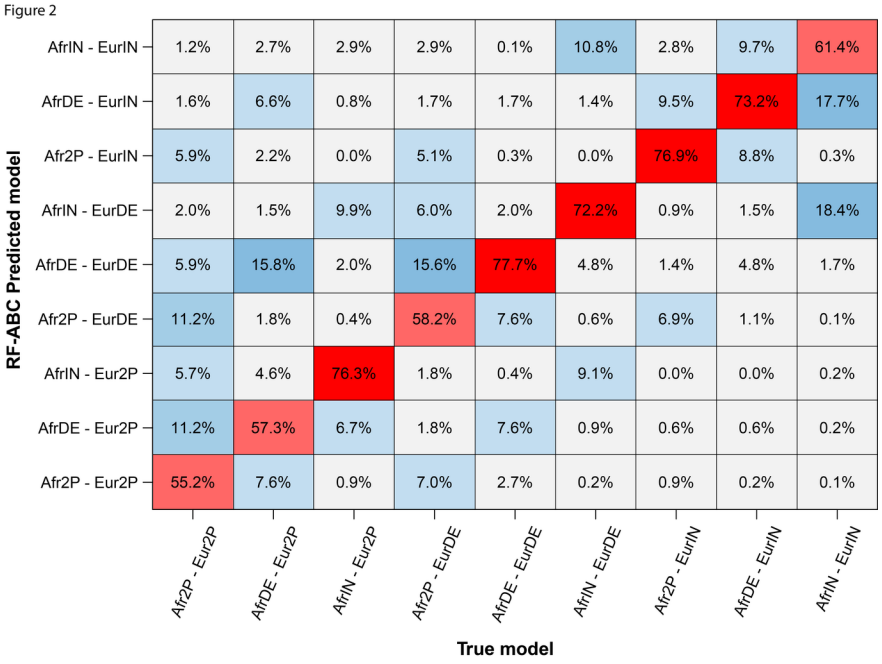

Figure 3

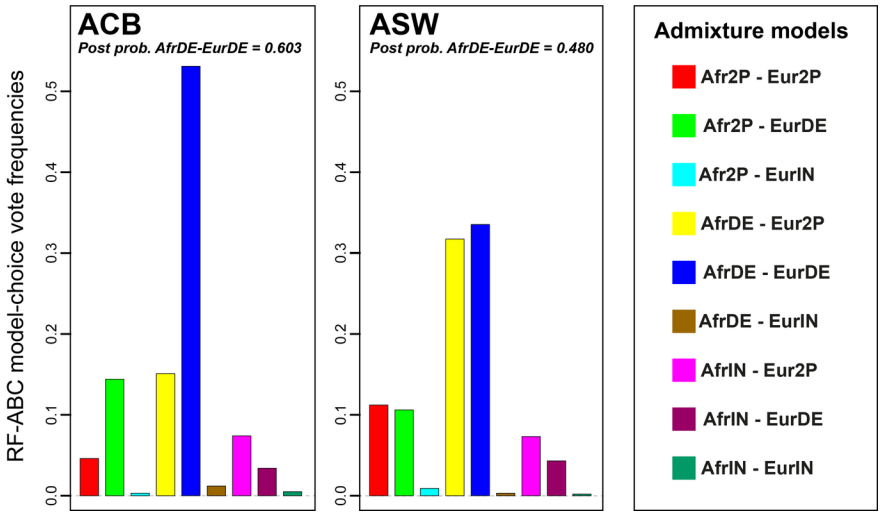

Figure 4

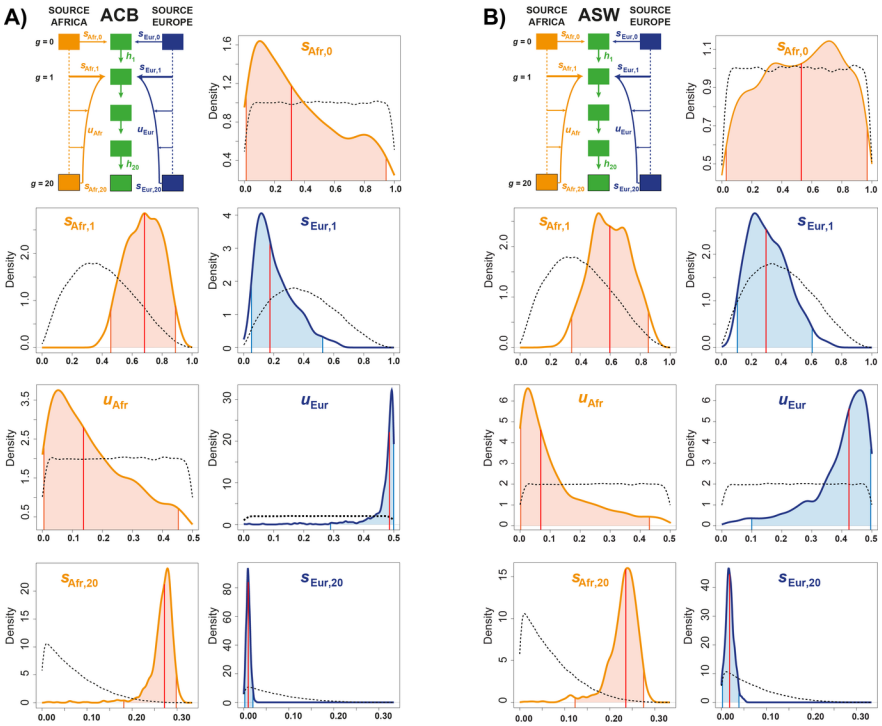

Figure 5