



HAL
open science

Complex genetic admixture histories reconstructed with Approximate Bayesian Computations

Cesar A Fortes-Lima, Romain Laurent, Valentin Thouzeau, Bruno Toupance,
Paul Verdu

► **To cite this version:**

Cesar A Fortes-Lima, Romain Laurent, Valentin Thouzeau, Bruno Toupance, Paul Verdu. Complex genetic admixture histories reconstructed with Approximate Bayesian Computations. 2020. hal-03065543

HAL Id: hal-03065543

<https://hal.science/hal-03065543>

Preprint submitted on 14 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Title:**

2 **Complex genetic admixture histories reconstructed with Approximate Bayesian**
3 **Computations**

4
5 **Running Title:**

6 **Admixture history reconstructed with ABC**

7
8 **Authors:**

9 Cesar A. Fortes-Lima^{†,‡,*}, Romain Laurent^{†,*}, Valentin Thouzeau^{§,&}, Bruno Toupance[†] and Paul
10 Verdu^{†,#}

11
12 **Affiliation:**

13 [†] CNRS, Muséum National d'Histoire Naturelle, Université de Paris, Unité Eco-anthropologie
14 (EA), UMR7206, Paris, France

15 [‡] Sub-department of Human Evolution, Department of Organismal Biology, Evolutionary
16 Biology Centre, Uppsala University, Uppsala, Sweden

17 [§] CNRS, Université Paris-Dauphine, PSL University, UMR 7534 Centre de Recherche en
18 Mathématiques de la Décision, Paris, France

19 [&] ENS, PSL University, EHESS, CNRS, Laboratoire de Sciences Cognitives et
20 Psycholinguistique, Département d'Etudes Cognitives, Paris, France

21
22 * These authors contributed equally to this work

23
24 # Corresponding author: Paul Verdu,

25 Institution: CNRS, Muséum National d'Histoire Naturelle, Université de Paris;

26 Lab: Unité Eco-anthropologie (EA) UMR7206;

27 Address: Musée de l'Homme, 17, place du Trocadéro, 75016 Paris, France;

28 email: paul.verdu@mnhn.fr;

29 tel: +33 1 44 05 73 17

30
31 **Keywords:** Admixture; Approximate Bayesian Computation; Inference; Population Genetics;
32 Machine Learning

33

34 **Author Contributions:**

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

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

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

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

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

44

45 **Acknowledgements:**

46 We warmly thank Frédéric Austerlitz, Erkan O. Buzbas, Antoine Cools, Flora Jay, Evelyne
47 Heyer, Margueritte Lapierre, Guillaume Laval, Nina Marchi, Etienne Patin, Noah A.
48 Rosenberg, and Zachary A. Szpiech for useful comments and discussions. This project was
49 funded in part by the French Agence Nationale de la Recherche project METHIS (ANR 15-
50 CE32-0009-01). CFL was funded in part by the Sven and Lilly Lawski's Foundation (N2019-
51 0040).

52 Authors declare no conflict of interest for this work.

53

54

55 **Figure and Table Content:**

56 Main Text: 5 figures, 4 tables.

57 Supplementary material: 1 note, 6 figures, 3 tables.

58

59 **Novel online resources:**

60 *MetHis* software package can be downloaded with manual and example dataset from

61 <https://github.com/romain-laurent/MetHis>

62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84

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 dense genome-wide datasets mostly in human populations. Nevertheless, complex admixture histories, beyond one or two pulses of admixture, remain methodologically challenging to reconstruct, especially when large datasets are unavailable. We develop an Approximate Bayesian Computations (ABC) framework to reconstruct complex admixture histories from independent genetic markers. We built the software package *MetHis* to simulate independent SNPs in a two-way admixed population for scenarios with multiple admixture pulses, or monotonically decreasing or increasing admixture at each generation; drawing model-parameter values from prior distributions set by the user. For each simulated dataset, we calculate 24 summary statistics describing genetic diversity and moments of individual admixture fraction. We coupled *MetHis* with existing ABC algorithms and investigate the admixture history of an African American and a Barbadian population. Results show that Random-Forest ABC scenario-choice, followed by Neural-Network ABC posterior parameter estimation, can distinguish most complex admixture scenarios and provide accurate model-parameter estimations. For both admixed populations, we find that monotonically decreasing contributions over time, from the European and African sources, explain the observed data more accurately than multiple admixture pulses. Furthermore, we find contrasted trajectories of introgression decay from the European and African sources between the two admixed populations. This approach will allow for reconstructing detailed admixture histories in numerous populations and species, particularly when maximum-likelihood methods are intractable.

INTRODUCTION

85

86

87 Hybridization between species and admixture between populations are powerful mechanisms
88 influencing biological evolution. Genetic admixture patterns have thus been extensively studied
89 to understand migrations and admixture-related adaptation (HELICONIUS GENOME CONSORTIUM
90 2012; HELLENTHAL *et al.* 2014; SKOGLUND *et al.* 2015; BRANDENBURG *et al.* 2017). The
91 increasing availability of genome-wide data in numerous species, and particularly humans (e.g.
92 1000 GENOMES PROJECT CONSORTIUM 2015), further provides unprecedented opportunities to
93 understand the genomic architecture of admixture, characterize the contribution of admixture
94 to adaptive evolution, and infer demographic histories of admixture from genetic data.

95 Based on a long history of statistical developments aimed at investigating admixture patterns
96 from genetic data (BERNSTEIN 1931; CAVALLI-SFORZA and BODMER 1971; CHAKRABORTY and
97 WEISS 1988; LONG 1991; FALUSH *et al.* 2003; PATTERSON *et al.* 2012), population geneticists
98 recently developed methods to reconstruct the genomic architecture of admixed segments
99 deriving from each source population, and to describe admixture linkage-disequilibrium (LD)
100 patterns (SANKARARAMAN *et al.* 2008; PRICE *et al.* 2009; LAWSON *et al.* 2012; MAPLES *et al.*
101 2013; GUAN 2014; SALTER-TOWNSHEND and MYERS 2019). In *Homo sapiens*, these methods
102 have been extensively used to infer populations' ancestral genetic origins and map local
103 ancestry along individual genomes, often for disease-mapping purposes (e.g. SHRINER *et al.*
104 2011). Furthermore, by coupling admixture mapping approaches with natural selection scans,
105 sometimes accounting for ancient and recent demographic history, it is possible to identify
106 signatures of adaptive introgression or post-admixture selection having influenced genomic
107 diversity patterns in human populations (JEONG *et al.* 2014; RACIMO *et al.* 2015; PATIN *et al.*
108 2017).

109 In this context, several maximum-likelihood approaches have been developed to estimate the
110 parameters of admixture models (time of admixture events and their associated intensities) that
111 vastly improved our understanding of detailed admixture histories in particular for human
112 populations (e.g. PICKRELL and PRITCHARD 2012; HELLENTHAL *et al.* 2014). The two classes
113 of methods most extensively deployed in the past rely, respectively, on the moments of allelic
114 frequency spectrum divergences among populations (REICH *et al.* 2009; PATTERSON *et al.* 2012;
115 PICKRELL and PRITCHARD 2012; LIPSON *et al.* 2013), and on admixture LD patterns (POOL and
116 NIELSEN 2009; MOORJANI *et al.* 2011; GRAVEL 2012; LOH *et al.* 2013; HELLENTHAL *et al.* 2014;
117 CHIMUSA *et al.* 2018). They allow for identifying admixture events in a given set of populations,

118 estimating admixture fractions, and inferring time since each pulse of admixture. Notably,
119 Gravel (GRAVEL 2012) developed an approach to fit the observed curves of admixture LD decay
120 to those theoretically expected under admixture models involving one or two possible pulses of
121 admixture from multiple source populations. This major advance significantly improved our
122 ability to reconstruct detailed admixture histories using genetic data, for instance among several
123 populations descending from the Transatlantic Slave Trade (TAST) across the Americas (e.g.
124 MORENO-ESTRADA *et al.* 2013; BAHARIAN *et al.* 2016; FORTES-LIMA *et al.* 2017).

125 Despite the unquestionable importance of these previous developments, existing admixture
126 history inference methods somewhat suffer from inherent limitations acknowledged by the
127 authors (GRAVEL 2012; LIPSON *et al.* 2013; HELLENTHAL *et al.* 2014). First, most likelihood
128 approaches can only consider one or two pulses of admixture in the history of the hybrid
129 population. Nevertheless, admixture processes in numerous species are known to be often much
130 more complex, involving multiple admixture-pulses or periods of recurring admixture over time
131 from each source population separately. It is not yet clear how these methods might behave
132 when they can consider only simplified versions of the true admixture history underlying the
133 observed data (GRAVEL 2012; LIPSON *et al.* 2013; LOH *et al.* 2013; HELLENTHAL *et al.* 2014;
134 MEDINA *et al.* 2018; NI *et al.* 2019). Second, while it is possible to compare maximum-
135 likelihood values obtained from fitting one or two admixture pulses to the observed data as a
136 guideline to find the “best” scenario, formal statistical comparison of model posterior
137 probabilities is often out of reach of these approaches (GRAVEL 2012; FOLL *et al.* 2015; NI *et*
138 *al.* 2019). Finally, admixture-LD methods, in particular, rely on fine mapping of local ancestry
139 segments in individual genomes and thus require substantial amounts of genomic data (typically
140 several hundred thousand to several millions of SNPs), and, sometimes, accurate phasing.
141 These still represent major challenges for most species, including humans.

142 To overcome these limitations, Approximate Bayesian Computation (ABC) approaches
143 (TAVARÉ *et al.* 1997; PRITCHARD *et al.* 1999; BEAUMONT *et al.* 2002) represent a promising
144 class of methods to infer complex admixture histories from observed genetic data. Indeed, ABC
145 has been successfully used previously in different species (including humans), and using
146 different types of genetic data, to formally test alternative demographic scenarios hypothesized
147 to be underlying observed genetic patterns, and to estimate, a posteriori, the parameters of the
148 winning models (VERDU *et al.* 2009; BOITARD *et al.* 2016; FRAIMOUT *et al.* 2017).

149 ABC model-choice and posterior parameter inference rely on comparing observed summary
150 statistics to the same set of statistics, calculated from a usually large number of genetic
151 simulations explicitly parametrized by the user, and produced under competing demographic
152 scenarios (BEAUMONT *et al.* 2002; WEGMANN *et al.* 2009; BLUM and FRANÇOIS 2010; CSILLÉRY
153 *et al.* 2012; PUDLO *et al.* 2016; SISSON *et al.* 2018). Each simulation, and corresponding vector
154 of summary statistics, is produced using model-parameters drawn randomly from prior
155 distributions informed adequately by the user. Therefore, the flexibility of ABC relies mostly
156 on explicit genetic data simulations set by the user. This makes ABC *a priori* particularly well
157 suited to investigate highly complex historical admixture scenarios for which likelihood
158 functions are very often intractable, but for which simulation of genetic data is feasible
159 (PRITCHARD *et al.* 1999; VERDU and ROSENBERG 2011; GRAVEL 2012). However, ABC has
160 until now seldom been used to investigate admixture processes beyond a single admixture pulse
161 or constant migrations (BUZBAS and ROSENBERG 2015; BUZBAS and VERDU 2018).

162 In this paper, we show how ABC can be successfully applied to reconstruct, from genetic data,
163 highly complex admixture histories beyond exploring models with a single or two pulses of
164 admixture. In particular, we focus on evaluating how a relatively limited number of independent
165 SNPs can be used for accurately distinguishing major classes of historical admixture models,
166 such as multiple admixture-pulses versus recurring increasing or decreasing admixture over
167 time, and for conservative posterior parameter inference under the winning model.
168 Furthermore, we show that the quantiles and higher moments of the distribution of admixture
169 fractions in the admixed population are highly informative summary-statistics for ABC model-
170 choice and posterior-parameter estimation, as expected analytically (VERDU and ROSENBERG
171 2011; GRAVEL 2012; BUZBAS and VERDU 2018).

172 In order to do so, and since genetic data simulation under highly complex admixture models is
173 not trivial using existing coalescent approaches (WAKELEY *et al.* 2012), we propose a novel *ad*
174 *hoc* forward-in-time genetic data simulator and a set of parameter-generator and summary-
175 statistics calculation tools embedded in an open source C software package called *MetHis*. It is
176 adapted to conduct primarily ABC inferences with existing ABC tools implemented in the R
177 (R DEVELOPMENT CORE TEAM 2017) packages *abc* (CSILLÉRY *et al.* 2012) and *abcrf* (PUDLO *et*
178 *al.* 2016; RAYNAL *et al.* 2019).

179 We exemplify our approach by reconstructing the complex admixture histories underlying
180 observed genetic patterns separately for the African American (ASW) and Barbadian (ACB)

181 populations from the 1000 Genomes Project Phase 3 (1000 GENOMES PROJECT CONSORTIUM
182 2015). Both populations are known to be admixed populations of European and African descent
183 in the context of the TAST (e.g. GRAVEL 2012; BAHARIAN *et al.* 2016; MARTIN *et al.* 2017).
184 We find admixture histories much more complex than previously inferred for these populations
185 and further reveal the diversity of admixture histories undergone by populations descending
186 from the TAST in the Americas.

187

188

189

MATERIAL AND METHODS

190 We aimed at evaluating how ABC model-choice and posterior parameter estimation could
191 allow reconstructing highly complex historical admixture processes using independent genome-
192 wide SNPs. To do so, we chose to focus on the recent admixture history of populations of
193 African and European ancestry, descending from European colonization and the TAST in the
194 Americas. This case-study represents an appropriate setting for empirically testing our ABC
195 approach, since this period of history starting in the late 15th century has been extensively
196 studied in population genetics based on the same publicly available datasets.

197 First, we describe the targeted case-study population and genetic datasets. Second, we present
198 in detail the complex admixture processes here investigated and the associated demographic
199 parameters. Third, we describe the novel simulation and summary statistics calculation software
200 package called *MetHis*, here proposed to investigate these admixture processes. Fourth, we
201 detail the Random-Forest ABC procedure used for scenario-choice inference and the
202 performance of this approach both in general for the tested models and specifically for the real
203 data here investigated. Finally, we detail the Neural Network ABC procedure deployed to
204 estimate posterior parameter distributions, its parameterization, and the cross-validation
205 procedures conducted to evaluate its power and accuracy.

206 Population Genetics Dataset

207 We considered the admixture histories of the African American (ASW) and Barbadian (ACB)
208 population samples from the 1000 Genomes Project Phase 3 (1000 GENOMES PROJECT
209 CONSORTIUM 2015). Previous studies identified, within the same database, the West European
210 Great-Britain (GBR) and the West African Yoruba (YRI) population samples as reasonable
211 proxies for the genetic sources of the admixture of both ACB and ASW populations,
212 consistently with the macro-history of the TAST in the former British colonial empire in Africa
213 and the Americas (BAHARIAN *et al.* 2016; MARTIN *et al.* 2017; VERDU *et al.* 2017).

214 We excluded from our sample set, individuals previously identified to be more closely related
215 than first-degree cousins in the four populations separately (VERDU *et al.* 2017). We also
216 excluded the three ASW individuals showing traces of Native American or East-Asian
217 admixture beyond that from Europe and Africa, as reported in previous studies (MARTIN *et al.*
218 2017). This allows us to consider only two source populations for the admixture history of both
219 admixed populations investigated here. Among the remaining individuals we randomly drew

220 50 individuals in the targeted admixed ACB and ASW populations, respectively, and included
221 the remaining 90 YRI individuals and 89 GBR individuals.

222 We extracted biallelic polymorphic sites (SNPs as defined by the 1000 Genomes Project Phase
223 3) from the merged ACB+ASW+GBR+YRI data set, excluding singletons. Furthermore, we
224 focused only on independent SNPs by LD pruning the data set using the PLINK (PURCELL *et*
225 *al.* 2007) --indep-pairwise option with a sliding window of 100 SNPs, moving in increments of
226 10 SNPs, and r^2 threshold of 0.1 (ALEXANDER *et al.* 2009). Finally, we randomly drew 100,000
227 SNPs from the remaining SNP set.

228 **Competing complex admixture scenarios**

229 We aimed at investigating comprehensive admixture histories with, after the original
230 foundation of the admixed population, possibly multiple pulses (>1) of admixture, or recurring
231 monotonically increasing or decreasing admixture, from each source population separately. To
232 do so, we chose to work under the general mechanistic model presented in Verdu and Rosenberg
233 (VERDU and ROSENBERG 2011), henceforth called the VR2011 model, derived from Ewens and
234 Spielman (EWENS and SPIELMAN 1995). Briefly (**Supplementary Figure S1**), the VR2011
235 general model considers, for diploid organisms, a panmictic admixture process, discrete in
236 generations, where M source populations S_m contribute to the hybrid population H at the
237 following generation $g + 1$ with proportions $s_{m,g}$ each in $[0,1]$, and where the hybrid population
238 H contributes to itself with proportion h_g in $[0,1]$ with $h_0 = 0$, satisfying, for each value of $g \geq$
239 0 , $\sum_{m \in [1,M]} s_{m,g} + h_g = 1$.

240 Here, we adapted the two source-populations version of the general VR2011 ($M = 2$), and
241 define, next, the nine competing complex admixture scenarios considered to reconstruct the
242 history of introgression from Africa and Europe into the gene-pool of the ACB and ASW
243 admixed populations (see above), separately (**Figure 1**).

244 **Foundation of the admixed population H**

245 For all scenarios (**Figure 1, Table 1**) we chose a fixed time for the foundation (generation 0,
246 forward-in-time) of population H occurring 21 generations before present, with admixture
247 proportions $s_{Afr,0}$ and $s_{Eur,0}$ from the African and the European sources respectively, with $s_{Afr,0}$
248 $+ s_{Eur,0} = 1$, and $s_{Afr,0}$ in $[0,1]$. This corresponds to the first arrival of European permanent settlers
249 in the Americas and Caribbean in the late 15th and early 16th centuries, considering 20 or 25

250 years per generation and the sampled generation born in the 1980s. Note that simulations
251 considering a parameter $s_{Afr,0}$ close to 0, or alternatively 1, correspond to foundations of the
252 population H from either one source population, therefore delaying the first “real” genetic
253 admixture event to the next, more recent, demographic event. Following foundation, we
254 consider three alternative scenarios for the admixture contribution of each source population S,
255 African or European in our case, separately.

256 Admixture-pulse(s) scenarios

257 For a given source population S, African (Afr) or European (Eur), scenarios *S-2P* consider two
258 possible pulses of admixture into population H occurring respectively at time $t_{S,p1}$ and $t_{S,p2}$
259 distributed in $[1,20]$ with $t_{S,p1} \neq t_{S,p2}$, with associated admixture proportion $s_{S,tS,p1}$ and $s_{S,tS,p2}$ in
260 $[0,1]$ satisfying, at all times t , $\sum_{S \in (Afr, Eur)} s_{S,t} \leq 1$ (**Figure 1, Table 1**). Note that for one of
261 either $s_{S,t}$ parameter values close to 0, the two-pulse scenarios are equivalent to single pulse
262 scenarios after the foundation of H. Furthermore, for both $s_{S,t}$ values close to 0, scenarios *S-2P*
263 are nested with scenarios where only the founding admixture pulse 21 generations ago is the
264 source of genetic admixture in population H. Alternatively, $s_{S,t}$ parameter values close to 1
265 consider a virtual complete genetic replacement of population H by source population S at that
266 time. Finally, certain *S-2P* scenarios with two consecutive pulses from a given source S ($t_{S,p1} =$
267 $t_{S,p2} - 1$), may be strongly resembling single-pulse scenarios (after foundation).

268 Recurring decreasing admixture scenarios

269 For a given source population S, scenarios *S-DE* consider a recurring monotonically decreasing
270 admixture from source population S at each generation between generation 1 (after foundation
271 at generation 0) and generation 20 (sampled population) (**Figure 1, Table 1**). In these scenario,
272 $s_{S,g}$, with g in $[1..20]$, are the discrete numerical solutions of a rectangular hyperbola function
273 over the 20 generations of the admixture process until present as described in **Supplementary**
274 **Note S1**. In brief, this function is determined by parameter u_S , the “steepness” of the curvature
275 of the decrease, in $[0,1/2]$, $s_{S,1}$, the admixture proportion from source population S at generation
276 1 (after foundation), in $[0,1]$, and $s_{S,20}$, the last admixture proportion in the present, in $[0, s_{S,1}/3]$.
277 Note that we chose the boundaries for $s_{S,20}$ in order to reduce the parameter space and nestedness
278 among competing scenarios, and explicitly force scenarios *S-DE* into a substantially decreasing
279 admixture process. Indeed, defining $s_{S,20}$ in $[0, s_{S,1}]$ instead would have also allowed for both
280 decreasing admixture processes and relatively constant recurring admixture processes.
281 Furthermore, note that parameter u_S values close to 0 create pulse-like scenarios occurring

282 immediately after foundation of intensity $s_{S,1}$, followed by constant recurring admixture at each
283 generation until present of intensity $s_{S,20}$. Alternatively, parameter u_S values close to 1/2 create
284 scenarios with a linearly decreasing admixture between $s_{S,1}$ and $s_{S,20}$ from source population S
285 at each generation after the foundation of population H.

286 Recurring increasing admixture scenarios

287 Finally, for a given source population S, scenarios *S-IN* mirrors the *S-DE* scenarios by
288 considering instead a recurring monotonically increasing admixture from source population S
289 (**Figure 1, Table 1**). Here, $s_{S,g}$, with g in [1..20], are the discrete numerical solutions of the
290 same function as in the S-DE decreasing scenarios (see above), flipped over time between
291 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
292 $[0, 1/2]$ parametrizes the “steepness” of the curvature of the increase. Note that in this case,
293 parameter u values close to 0 create pulse-like scenarios occurring in the present of intensity
294 $s_{S,20}$, preceded by constant recurring admixture of intensity $s_{S,1}$ at each generation since
295 foundation. Alternatively, parameter u_S values close to 1/2 create scenarios with a linearly
296 increasing admixture between $s_{S,1}$ and $s_{S,20}$ from source population S at each generation after
297 the foundation of population H.

298 Combining admixture scenarios from either source populations

299 We combine these three scenarios to obtain nine alternative scenarios with two source
300 populations, African (Afr) and European (Eur) respectively, for the admixture history of
301 population H (**Figure 1, Table 1**), the ASW or ACB alternatively, with the only condition that,
302 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
303 the remaining contribution of the admixed population H to itself at the generation g . Four
304 scenarios (Afr2P-EurDE, Afr2P-EurIN, AfrDE-Eur2P, and AfrIN-Eur2P) consider a mixture
305 of pulse-like and recurring admixture from each source. Three scenarios (Afr2P-Eur2P, AfrDE-
306 EurDE, and AfrIN-EurIN), consider symmetrical classes of admixture scenarios from either
307 source. Two scenarios (AfrIN-EurDE and AfrDE-EurIN) consider mirroring recurring
308 admixture processes. Importantly, this scenario design considers nested historical scenarios in
309 specific parts of the parameter space, as exemplified above.

310 Forward-in-time simulations with *MetHis*

311 Simulation of genome-wide independent SNPs under highly complex admixture histories is
312 often not trivial under the coalescent and using classical existing software (WAKELEY *et al.*

313 2012). In this context, we developed *MetHis*, a C open-source software package available at
314 <https://github.com/romain-laurent/MetHis>, to simulate large amounts of genetic data under the
315 two-source populations VR2011 model and calculate summary statistics of interest to
316 population geneticists interested in complex admixture processes. *MetHis*, in its current form,
317 can be used to simulate any number of independent SNPs in the admixed population H.
318 However, *MetHis* does not allow simulating the source populations for the admixture process.
319 Instead, this can be done efficiently using coalescent-based simulations with existing software
320 such as *fastsimcoal2* (EXCOFFIER and FOLL 2011; EXCOFFIER *et al.* 2013), or other forward-in-
321 time genetic data simulators such as *SLIM v3* (HALLER and MESSER 2019).

322 *Simulating source populations*

323 Here, we wanted to focus our investigation specifically on the admixture process undergone by
324 the admixed population descending from the TAST. Therefore, we made several *ad hoc*
325 simplification choices for simulating source population genetic data under the nine competing
326 models described next.

327 We consider that the African and European populations at the source of the admixture processes
328 are very large populations at the drift-mutation equilibrium, accurately represented by the
329 Yoruban YRI and British GBR datasets here investigated. Therefore, we first build two separate
330 datasets each comprising 20,000 haploid genomes of 100,000 independent SNPs, each SNP
331 being randomly drawn in the site frequency spectrum (SFS) observed for the YRI and GBR
332 datasets respectively. These two datasets are used as fixed gamete reservoirs for the African
333 and European source population datasets separately, at each generation of the forward-in-time
334 admixture process. From these reservoirs, at each generation separately, we build an effective
335 individual gene-pool of diploid size N_g (see below), by randomly pairing gametes avoiding
336 selfing. These virtual source populations provide the parental pool for simulating individuals in
337 the admixed population H, at each generation separately. Thus, while our gamete reservoirs are
338 fixed over the 21 generations of the admixture processes here considered, the parental genetic
339 pools are randomly built anew at each generation of the admixture process.

340 *Simulating the admixed population*

341 At each generation, *MetHis* performs simple Wright-Fisher (FISHER 1922; WRIGHT 1931)
342 forward-in-time simulations, individual-centered, in a panmictic admixed population H of
343 diploid effective size N_g . For a given individual in the hybrid population at the following

344 generation ($g + 1$), *MetHis* independently draws each parent from the source populations with
345 probability $s_{S,g}$ (**Figure 1, Table 1**), or from the hybrid population with probability h_g ,
346 randomly builds a haploid gamete of 100,000 independent SNPs for each parent, and pairs the
347 two constructed gametes to create the new individual. Here, we decided to neglect mutation
348 over the 21 generations of admixture considered. This is reasonable when studying relatively
349 recent admixture histories. Nevertheless, this will be improved in future versions of the
350 software, in particular to allow studying much more ancient admixture histories. Finally, while
351 we chose explicitly to simulate only the individuals in the admixed population H here, note that
352 future developments of *MetHis* will allow to also simulate individual genetic data in the source
353 populations in the same way.

354 Effective population size in the source and the admixed populations

355 To focus on the admixture process itself without excessively increasing the parameter space,
356 we consider, for each nine-competing model, both source populations and the admixed
357 population H with constant effective population size $N_g = 1000$ diploid individuals at each
358 generation. Nevertheless, note that *MetHis* software readily allows the user to easily
359 parameterize changes in the effective size of population H at each generation.

360 Sampling simulated unrelated individuals

361 After each simulation, we randomly draw individual samples matching sample sizes in our
362 observed dataset: 90 and 89 individuals respectively from the African and European sources,
363 and 50 individuals in the admixed population H. We sample individuals until our sample set
364 contains no individuals related at the 1st degree cousin within each population and between the
365 admixed population and either source populations, based on explicit parental flagging during
366 the last 2 generations of the simulations.

367 Simulating by randomly drawing parameter values from prior distributions

368 With this implementation of *MetHis*, we performed 10,000 independent simulations under each
369 nine competing scenarios described above and in **Figure 1**, drawing the corresponding model-
370 parameters (pulse-times and associated admixture intensities, “steepness” of the recurring
371 admixture-increases or decreases and associated initial and final admixture intensities), in prior-
372 distributions detailed in **Table 1**. Although the user can perform *MetHis* simulations with an
373 external parameter list, we readily provide *ad hoc* scripts in *MetHis*, which allow to easily
374 generate parameter lists for a large number of complex admixture scenarios set by the user.

375 For the best models identified using Random-Forest ABC model-choice approach (PUDLO *et*
376 *al.* 2016) for the ACB and ASW admixed populations respectively (see **Results**), we conducted
377 an additional 90,000 independent simulations with the same parameter priors as in the 10,000
378 simulations already conducted. Thus, we considered 100,000 simulations for the best scenarios
379 for the ACB and ASW respectively, to be used for ABC posterior parameter inference (see
380 below).

381 Summary Statistics

382 We considered 24 summary statistics for ABC model-choice and posterior parameter inference,
383 computed on each simulated dataset with *MetHis*. Four statistics were strictly within-
384 populations; four statistics were strictly between-populations; and 16 statistics were specifically
385 calculated to describe the distribution of admixture among individuals within the admixed
386 population H. Indeed, previous theoretical works have shown that this distribution and all its
387 moments carried signatures of the underlying complex historical process (VERDU and
388 ROSENBERG 2011; GRAVEL 2012). Numerous descriptive statistical approaches have been
389 successfully developed to estimate admixture fractions from genetic data in admixed
390 populations (e.g. ALEXANDER *et al.* 2009; PATTERSON *et al.* 2012; PICKRELL and PRITCHARD
391 2012). However, most methods remain computationally costly when iterated for large to very
392 large sets of simulated genetic data. Therefore, only a few previous ABC historical inference
393 approaches have considered the distribution of admixture fraction as a summary statistics
394 (BUZBAS and ROSENBERG 2015; BUZBAS and VERDU 2018), although some admixture-related
395 statistics have been embedded in ABC software packages (CORNUET *et al.* 2014).

396 Distribution of admixture fractions as a summary statistic

397 We estimated individual admixture distribution based on allele-sharing-dissimilarity (ASD)
398 (BOWCOCK *et al.* 1994) and multidimensional scaling (MDS) (PASCHOU *et al.* 2007; PRICE *et*
399 *al.* 2009). For each simulated dataset, we first calculated a pairwise inter-individual ASD matrix
400 using *asd* software (<https://github.com/szpiech/asd>) on all pairs of sampled individuals and all
401 100,000 independent SNPs. Then we projected in two dimensions this pairwise ASD matrix
402 with classical unsupervised metric MDS using the *cmdscale* function R (R DEVELOPMENT CORE
403 TEAM 2017). We expect individuals in population H to be dispersed along an axis joining the
404 centroids of the two proxy source populations on the two-dimensional MDS plot. We projected
405 individuals orthogonally on this axis, and calculate individual's relative distance to each
406 centroid. We considered this measure to be an estimate of individual average admixture level

407 from either source population. Note that by doing so, some individuals might show “admixture
408 fractions” higher than one, or lower than zero, as they might be projected on the other side of
409 the centroid when being genetically close to 100% from one source population or the other.
410 Under an ABC framework, this is not a difficulty since this may happen also on the real data *a*
411 *priori*, and our goal is to use summary statistics that mimic the observed ones. This individual
412 admixture estimation method has been shown to be highly concordant with cluster membership
413 fractions as estimated with ADMIXTURE (ALEXANDER *et al.* 2009) in real data analyses (e.g.
414 VERDU *et al.* 2017). Considering the real data here investigated, we confirm these previous
415 findings since we obtain a Spearman correlation (calculated using the *cor.test* function in *R*), of
416 $\rho = 0.950$ (p-value $< 2.10^{-16}$) and $\rho = 0.977$ (p-value $< 2.10^{-16}$) between admixture estimates
417 based on ASD-MDS and on ADMIXTURE, for the ACB and ASW respectively
418 **(Supplementary Figure S2).**

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

422 Within population summary statistics

423 We calculated SNP by SNP heterozygosities (NEI 1978) using *vcftools* (DANECEK *et al.* 2011),
424 and considered the mean and variance of this quantity across SNPs in the admixed population
425 as two separate summary statistics for ABC inference. Note that, these quantities are fixed for
426 each source population, respectively, and thus uninformative in our case study, since source
427 populations are simulated only once and used for all subsequent simulations under the nine
428 competing models (see above).

429 In addition, as we computed the individual pairwise ASD matrix for calculating the distribution
430 of admixture fraction (see above), we also considered the mean and variance of ASD values
431 across pairs of individuals within the admixed population H, as two within-population summary
432 statistics.

433 Between populations summary statistics

434 In addition to previous summary statistics, we considered multilocus pairwise F_{ST} (WEIR and
435 COCKERHAM 1984) between population H and each source population respectively, calculated
436 using *vcftools* (DANECEK *et al.* 2011). Note that the F_{ST} between the source populations is fixed,
437 since simulated source populations are themselves fixed (see above), and thus uninformative in

438 our case study. Furthermore, we calculated the mean ASD between individuals in population H
439 and, separately, individuals in either source population. Finally, we computed anew from
440 Patterson (PATTERSON *et al.* 2012) the f_3 statistics based on allelic frequencies obtained with
441 *vcftools* (DANECEK *et al.* 2011). In the two-source population case, this statistic is extensively
442 employed to test the original source of the admixture of a target admixed population, infer the
443 time since admixture, and estimate admixture intensities using maximum-likelihood
444 approaches.

445 Approximate Bayesian Computations

446 *MetHis* has been designed to operate under an ABC framework for model choice and parameter
447 inference. Thus, it allows simulating genetic data under numerous possible models by drawing
448 parameter values in a priori distributions set by the user in a flexible way. In addition, *MetHis*
449 allows for calculating numerous summary statistics a priori of interest to admixture processes,
450 and provides, as outputs, scenarios-parameter vectors and corresponding summary-statistics
451 vectors in reference tables ready to be used with the machine-learning ABC *abc* (CSILLÉRY *et*
452 *al.* 2012), and *abcrf* (PUDLO *et al.* 2016; RAYNAL *et al.* 2019) *R* packages (R DEVELOPMENT
453 CORE TEAM 2017).

454 Prior-checking

455 We evaluated, a priori, if the above simulation design and novel tools can simulate genetic data
456 for which summary statistics are coherent with those observed for the ACB and ASW as the
457 targeted admixed population. To do so, we first plotted each prior summary statistics
458 distributions and visually verified that the observed summary statistics for the ACB and ASW
459 respectively fell within the simulated distributions (**Supplementary Figure S3**). Second, we
460 explored the first four PCA axes computed with the *princomp* function in *R*, based on the 24
461 summary statistics and all 90,000 total simulations preformed for the nine competing scenarios,
462 and visually checked that observed summary statistics were within the cloud of simulated
463 statistics (**Supplementary Figure S4**). Finally, we performed a goodness-of-fit approach using
464 the *gfit* function from the *abc* package in *R*, with 1,000 replicates and tolerance level set to 0.01
465 (**Supplementary Figure S5**).

466 Model-choice with Random-Forest Approximate Bayesian Computation

467 We used Random-Forest ABC (RF-ABC) for model-choice implemented in the *abcrf* function
468 of the *abcrf* *R* package to obtain the cross-validation table and associated prior error rate using

469 an out-of-bag approach (**Figure 2**). We considered the same prior probability for the nine
470 competing models each represented by 10,000 simulations in the reference table. For the ACB
471 and ASW observed data separately, we performed model-choice prediction and estimation of
472 posterior probabilities of the winning model using the *predict.abcrf* function in the same *R*
473 package, using the complete simulated reference table for training the Random-Forest
474 algorithm (**Figure 3, Supplementary Table S1**). Both sets of analyses were performed
475 considering 1,000 decision trees in the forest after visually checking that error-rates converged
476 appropriately (**Supplementary Figure S6**), using the *err.abcrf* function in the *R* package *abcrf*.
477 Each summary statistics relative importance to the model-choice cross-validation was
478 computed using the *abcrf* function (**Figure 2**). RF-ABC cross-validation procedures using
479 groups of scenarios were conducted using the group definition option in the *abcrf* function
480 (ESTOUP *et al.* 2018).

481 Posterior parameter estimation with Neural-Network Approximate Bayesian Computation

482 It is difficult to estimate jointly the posterior distribution of all model parameters with RF-ABC
483 (RAYNAL *et al.* 2019). Furthermore, although RF-ABC performs satisfactorily well with an
484 overall limited number of simulations under each model (PUDLO *et al.* 2016), posterior
485 parameter estimation with other ABC approaches, such as simple rejection (PRITCHARD *et al.*
486 1999), regression (BEAUMONT *et al.* 2002; BLUM and FRANÇOIS 2010) or Neural-Network (NN)
487 (CSILLÉRY *et al.* 2012), require substantially more simulations a priori. Therefore, we
488 performed 90,000 additional simulations, for a total of 100,000 simulations for the best
489 scenarios identified with RF-ABC among the nine competing models for the ACB and ASW
490 separately.

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

492 For each parameter estimation analysis, we determined empirically the NN tolerance level (i.e.
493 the number of simulations to be included in the NN training), and number of neurons in the
494 hidden layer. Indeed, while the NN needs a substantial amount of simulations for training, there
495 is also a risk of overfitting posterior parameter estimations when considering too large a number
496 of neurons in the hidden layer. However, there are no absolute rules for choosing both numbers
497 (CSILLÉRY *et al.* 2012; JAY *et al.* 2019).

498 Therefore, using the 100,000 simulations for the winning scenarios identified with RF-ABC
499 (see above), we tested four different tolerance levels to train the NN (0.01, 0.05, 0.1, and 0.2),

500 and a number of neurons ranging between four and seven (the number of free parameters in the
501 winning scenarios, see **Results**). For each pair of tolerance level and number of neurons values,
502 we conducted cross-validation checking of posterior parameter estimations with 1,000
503 randomly chosen simulated datasets in turn used as pseudo-observed data with the “*cv4abc*”
504 function of the *R* package *abc*. We considered the median point-estimate of each posterior
505 parameter ($\hat{\theta}_i$) to be compared with the true parameter value used for simulation (θ_i). The
506 cross-validation parameter prediction error was then calculated across the 1,000 separate
507 posterior estimations for pseudo-observed datasets for each pair of tolerance level and number
508 of neurons, and for each parameter θ_i , as $\sum_1^{1000}(\hat{\theta}_i - \theta_i)^2 / (1000 \times \text{Variance}(\theta_i))$, allowing
509 to compare errors for scenarios-parameters across NN tolerance-levels and numbers of hidden
510 neurons, using the *summary.cv4abc* function in the *R* package *abc* (CSILLÉRY *et al.* 2012).
511 Results showed that, *a priori*, all numbers of neurons considered performed very similarly for
512 a given tolerance level (**Supplementary Table S2**). Furthermore, results showed that
513 considering 1% closest simulations to the pseudo-observed ones, to train the NN for parameter
514 estimation, reduces the average error for each tested number of neurons. Thus, we decided to
515 opt for four neurons in the hidden layer and a 1% tolerance level for training the NN in all
516 subsequent NN-ABC analyses, in order to avoid overfitting in parameter posterior estimations.

517 Estimation of model-parameters posterior distributions for ACB and ASW

518 We jointly estimated model-parameters posterior distributions for the ACB and ASW admixed
519 population separately, using 100,000 simulations for the best scenarios identified for each
520 admixed population separately, using NN-ABC (“*neuralnet*” methods’ option in the *R* package
521 *abc*) based on the logit-transformed (“*logit*” transformation option in the *R* package *abc*)
522 summary statistics using a 1% tolerance level to train the NN (i.e. considering only the 1,000
523 closest simulations to the observed data), fitted using a single-hidden-layer neural network with
524 four hidden neurons (**Figure 4, Table 2**).

525 Posterior parameter estimation error and credibility interval accuracy

526 For the ACB and ASW admixed populations separately, we wanted to evaluate the posterior
527 error performed by our NN-ABC approach on the median point estimate of each parameter, in
528 the vicinity of our observed data rather than randomly on the entire parameter space. To do so,
529 we first identified the 1,000 simulations closest to the real data with a tolerance level of 1%, for
530 the ACB and ASW respectively. Then, separately for the ACB and ASW set of closest

531 simulations, we performed, similarly as above for the real data parameter estimation procedure,
532 1,000 separate NN-ABC parameter estimations using the “neural” method in the *abc* function
533 with a NN trained with 1% tolerance level and four neurons in the hidden layer, using in turn
534 the other 99,999 simulations as reference table, and recorded the median point estimate for each
535 parameter. We then compared these estimates with the true parameter used for each 1,000
536 pseudo-observed target in the vicinity of our observed data and provide three types of error
537 measurements in **Table 3**. The mean-squared error scaled by the variance of the true parameter
538 $\sum_1^{1000}(\hat{\theta}_i - \theta_i)^2 / (1000 \times \text{Variance}(\theta_i))$ as previously (Csilléry et al. 2012); the mean-
539 squared error $\sum_1^{1000}(\hat{\theta}_i - \theta_i)^2 / 1000$, allowing to compare estimation errors for a given
540 scenario-parameter between the ACB and ASW analyses; and the mean absolute error
541 $\sum_1^{1000}|\hat{\theta}_i - \theta_i| / 1000$, which provides a more intuitive parameter estimation error.

542 Finally, based on these cross-validation procedures, we evaluated *a posteriori* if, in the vicinity
543 of the ACB and ASW observed datasets respectively, the lengths of the estimated 95%
544 credibility intervals for each parameter was accurately estimated or not (JAY et al. 2019). To do
545 so, we calculated how many times the true parameter (θ_i) was found inside the estimated 95%
546 credibility interval [2.5%quantile($\hat{\theta}_i$) ; 97.5%quantile($\hat{\theta}_i$)], among the 1,000 out-of-bag NN-
547 ABC posterior parameter estimation, separately for the ACB and ASW (**Supplementary Table**
548 **S3**). For each parameter, if less than 95% of the true parameter values are found inside the 95%
549 credibility interval estimated for the observed data, we consider the length of this credibility
550 interval as underestimated indicative of a non-conservative behavior of the parameter
551 estimation. Alternatively, if more than 95% of the true parameter-values are found inside the
552 estimated 95% credibility interval, we consider its length as overestimated, indicative of an
553 excessively conservative behavior of this parameter estimation.

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

555 With the above procedure, we aimed at estimating the posterior parameter distributions jointly
556 for all parameters, and their errors for the scenario most likely explaining observed genetic data
557 for the ACB and ASW respectively. Nevertheless, NN-ABC and RF-ABC parameter inference
558 procedures also allow estimating each parameter posterior distribution in turn and separately
559 rather than jointly. This can further provide insights into how both ABC parameter inference
560 approaches perform in the parameter space of the winning scenarios. To do so, we performed

561 several out-of-bag cross-validation parameter estimation analyses for the ACB and ASW
562 separately.

563 We compared four methods: NN estimation of the parameters taken jointly as a vector (similarly
564 as in the above procedure), NN estimation of the parameters taken in turn separately, RF
565 estimation of the parameters which also considers parameters in turn and separately (RAYNAL
566 *et al.* 2019), and simple Rejection estimation for each parameter separately (PRITCHARD *et al.*
567 1999). For each method, we used in turn the 1,000 simulations closest to the real data as pseudo-
568 observed data, and set a tolerance level of 1% of the 99,999 remaining simulations. We consider
569 four neurons in the hidden-layer per neural network, and we considered 500 decision trees per
570 random forest to limit the computational cost of these analyses at little accuracy cost *a priori*
571 (**Supplementary Figure S6**). We then computed the mean-squared errors scaled by the
572 variance of the true parameters, the mean-squared errors, and the mean absolute errors similarly
573 as previously. Finally, we estimated the accuracy of the 95% credibility intervals for each
574 method and for each parameter similarly as previously.

575

576

577

RESULTS

578 First, we present results about the ability of *MetHis* to simulate data close to the observed ones.
579 Second, we evaluate the ability of RF-ABC to distinguish, in the entire parameter space, the
580 nine complex admixture scenarios in competition, and evaluate how each one of the 24
581 summary statistics contribute to distinguish among scenarios. Third, we use Random-Forest
582 ABC to specifically predict the best fitting scenario for the history of admixture of two recently
583 admixed populations descending from the Transatlantic Slave Trade in the Americas (African
584 American ASW and Barbadian ACB). Fourth, we use Neural-Network ABC to estimate
585 posterior parameter distributions under the winning scenario for the ACB and the ASW
586 separately. Fifth, we evaluate in detail the accuracy of our posterior parameter estimation, and
587 compare with other ABC posterior parameter inference approaches. Finally, we synthesize the
588 complex admixture history thus reconstructed for the ASW and ACB populations.

589 *Simulating the observed data with MetHis*

590 With *MetHis*, we conducted 10,000 simulations for each one of the nine competing scenarios
591 for the admixture history of the ASW or the ACB populations, described in detail in **Figure 1**
592 and **Material and Methods**, with corresponding model parameters drawn in *a priori*
593 distributions described in **Table 1**.

594 We produced 90,000 vectors of 24 summary statistics each, overall highly consistent with the
595 observed ones for the ACB and the ASW populations respectively. First, we found that each
596 observed statistic is visually reasonably well simulated under the nine competing scenarios here
597 considered (**Supplementary Figure S3**). Second, the observed data each fell into the simulated
598 sets of summary statistics projected in the first four PCA dimensions (**Supplementary Figure**
599 **S4**) considering all 24 summary statistics in the analysis. Finally, the observed vectors of 24
600 summary statistics computed for the ACB and ASW, respectively, were not significantly
601 different (p-value = 0.468 and 0.710 respectively) from the 90,000 simulated sets of statistics
602 using a goodness-of-fit approach (**Supplementary Figure S5**). Therefore, we successfully
603 simulated datasets producing sets of summary statistics reasonably close to the observed ones,
604 despite considering constant effective population sizes, fixed virtual source population genetic
605 pool-sets, and neglecting mutation during the 21 generations of forward-in-time simulations
606 performed using *MetHis*.

607 *Complex admixture scenarios cross-validation with RF-ABC*

608 We trained the RF-ABC model-choice algorithm using 1,000 trees, which guaranteed the
609 convergence of the model-choice prior error rates (**Supplementary Figure S6**). Based on this
610 training, the complete out-of-bag cross-validation matrix showed that the nine competing
611 scenarios of complex historical admixture could be relatively reasonably distinguished using
612 our set of 24 summary statistics and 10,000 simulations under each competing scenario, despite
613 the high level of nestedness of the scenarios here considered (see **Material and Methods**).
614 Indeed, we calculated an out-of-bag prior error rate of 32.41%, considering each 90,000
615 simulation, in turn, as out-of-bag pseudo-observed target dataset and the rest of simulations
616 (89,999) as the training dataset for RF-ABC scenario-choice. Furthermore, we found the
617 posterior probabilities of identifying the correct scenario ranging from 55.17% (prior
618 probability = 11.11% for each competing scenario), for the two-pulses scenarios from both the
619 African and European sources (Afr2P-Eur2P), to 77.71% for the scenarios considering
620 monotonically decreasing recurring admixture from both sources (AfrDE-EurDE) (**Figure 2A**).
621 Importantly, the average probability, for a given admixture scenario, of choosing any one
622 alternative (wrong) scenario were on average 4.05% across the eight alternative scenarios,
623 ranging from 2.79% for the AfrDE-EurDE scenario, to 5.60% for the Afr2P-Eur2P scenario
624 (**Figure 2A**). This shows that our approach did not systematically favor one or the other
625 competing scenario when wrongly choosing a scenario instead of the true one, despite high
626 levels of nestedness among scenarios.

627 We find that the six summary statistics most contributing to the observed cross validation results
628 for RF-ABC model-choice among the 24 statistics here tested were statistics describing
629 specifically the admixture-fraction distribution: minimum and maximum admixture fraction
630 values, variance, skewness, as well as the 10% and 90% quantiles of the distribution (**Figure**
631 **2B**). Interestingly, within and between populations summary-statistics often used in population
632 genetics (including F_{ST} , mean heterozygosity, and f_3 statistics), contributed to distinguishing
633 the competing complex admixture scenarios to a lesser extent.

634 Finally, note that scenarios considering monotonically recurring admixture from each source
635 populations (AfrDE-EurDE, AfrDE-EurIN, AfrIN-EurDE, AfrIN-EurIN) can be relatively well
636 distinguished, using our RF-ABC framework, from scenarios with at least one source
637 population contributing to the admixed population with two possible pulses after the foundation
638 event (Afr2P-Eur2P, Afr2P-EurDE, Afr2P-EurIN, AfrDE-Eur2P, AfrIN-Eur2P). Indeed, we
639 found an out-of-bag prior error rate of 13.85%, and posterior cross-validation probabilities of

640 identifying the correct group of scenarios of 86.08% and 86.23% respectively for the two groups
641 (ESTOUP *et al.* 2018).

642 *Complex admixture histories for the Barbadian and African American populations*

643 *Random-Forest ABC scenario-choice*

644 We performed RF-ABC model-choice with 1,000 decision trees and 10,000 simulations per
645 each nine competing scenarios (**Figure 1** and **Table 1, Material and Methods**), separately for
646 the admixture history of the Barbadian (ACB) and the African American (ASW) populations.
647 For the ACB, **Figure 3** shows that the majority of votes (53.1%) went to an admixture scenario
648 AfrDE-EurDE with a posterior probability of the winning scenario of 60.28%. This scenario
649 encompassed monotonically decreasing recurring contributions from both the African and
650 European source populations over the last 20 generations before present. The second most
651 chosen scenario considered a monotonically decreasing recurring contribution from the African
652 source population over the last 20 generations, while the European source population
653 contributed two admixture pulses to this admixed population after the founding pulse (scenario
654 AfrDE-Eur2P). However, this scenario is voted for 3.5 times less often than the winning
655 scenario AfrDE-EurDE, gathering 15.1% of the 1,000 votes, only slightly above the 11.11%
656 prior probability for each nine-competing scenario (**Figure 3; Supplementary Table S1**).

657 Concerning the admixture history of the ASW, RF-ABC scenario-choice results were less
658 segregating. **Figure 3** shows that the AfrDE-EurDE scenario also gathered the majority of votes
659 for the admixture history of the ASW, albeit with lower posterior probability than for the ACB
660 (33.5% of 1,000 votes, with posterior probability = 48.0% for the ASW). The second most
661 chosen scenario, AfrDE-Eur2P, was only slightly less chosen with 31.7% of the votes (**Figure**
662 **3, Supplementary Table S1**). For the ASW, considering only the two best scenarios (AfrDE-
663 EurDE and AfrDE-Eur2P) to train the Random Forest, and re-conducting the RF-ABC
664 scenario-choice, improved the scenario discrimination in favor of the AfrDE-EurDE scenario.
665 While we found only a slight majority of votes (51.8%) also in favor of the AfrDE-EurDE
666 scenario, we found a substantially increased posterior probability for this model equal to 57.9%.
667 This increased posterior probability of the AfrDE-EurDE scenario compared to the previous
668 RF-ABC scenario-choice considering the nine competing scenarios (48.0%), indicated that this
669 scenario best explains the ASW observed genetic patterns, despite overall limited
670 discriminatory power of our approach in the part of the summary-statistics space occupied by
671 the ASW.

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

673 We performed 100,000 simulations using *MetHis* for the AfrDE-EurDE scenarios, in order to
674 estimate, using Neural-Network ABC, posterior parameter distributions and the corresponding
675 parameter prediction cross-validation errors, considering in turn the ACB and the ASW
676 populations (**Figure 4** and **5**, **Table 2**, **Table 3**, and **Supplementary Table S3**).

677 For the ACB under the AfrDE-EurDE scenario (**Figure 4A**, **Table 2**), we found that the two
678 recent admixture intensities from Africa and Europe ($s_{\text{Afr},20}$ and $s_{\text{Eur},20}$, respectively) and the
679 steepness of the European decrease in contribution over time (u_{Eur}) had sharp posterior densities
680 clearly distinct from their respective priors. Note that the cross-validation error on these
681 parameters in the vicinity of our real data were low (average absolute error 0.02744, 0.0044,
682 and 0.1084, respectively for $s_{\text{Afr},20}$, $s_{\text{Eur},20}$, and u_{Eur}) (**Table 3**), and lengths of 95% credibility
683 intervals reasonably accurate (96.4%, 94.4%, 94.1% of 1,000 cross-validation true parameter
684 values fell into estimated 95% credibility intervals, **Supplementary Table S3**). This shows the
685 reliability of our method to accurately infer the three parameters in the part of the space of
686 summary statistics occupied by the ACB observed data.

687 Furthermore, the two ancient admixture intensities from Africa and Europe at generation 1
688 immediately following the initial foundation of the admixed population H ($s_{\text{Afr},1}$ and $s_{\text{Eur},1}$,
689 respectively), also had posterior densities apparently distinguished from their prior
690 distributions, but both had much wider 95% credibility intervals (**Figure 4A**, **Table 2**).
691 Consistently, we found a slightly increased posterior parameter error in this part of the
692 parameter space for both these parameters, with average absolute error 0.121 and 0.095
693 respectively for $s_{\text{Afr},1}$ and $s_{\text{Eur},1}$ (**Table 3**). Nevertheless, note that 95.8% and 94.7% of 1,000
694 cross-validation true values for those two parameters fell into the estimated 95% credibility
695 intervals (**Supplementary Table S3**). This shows a reasonably conservative behavior of our
696 method for these estimations, further indicating that information is lacking in our data or set of
697 summary statistics for a more accurate estimation of these parameters, rather than an inaccuracy
698 of our approach.

699 Interestingly (**Figure 4A**, **Table 2**), we found that accurate posterior estimation of the steepness
700 of the African decrease in admixture over time (u_{Afr}) is difficult. Indeed, the posterior density
701 of this parameter only showed a tendency towards small values slightly departing from the
702 prior, indicative of a limit of our method to estimate this parameter (**Figure 4A**, **Table 2**).
703 Finally (**Figure 4A**, **Table 2**), we found that we had virtually no information to estimate the

704 founding admixture proportions from Africa and Europe at generation 0, as our posterior
705 estimates barely departed from the prior and associated mean absolute error was high (0.2530,
706 **Table 3**). Nevertheless, our method seemed to be performing reasonably conservatively for
707 these two latter parameters (95.6% and 95.3% of 1,000 cross-validation true parameter values
708 fell into estimated 95% credibility intervals, **Supplementary Table S3**). This indicates that
709 information is strongly lacking in our data or summary statistics for successfully capturing these
710 parameters, rather than inherent inaccuracy of our ABC method.

711 For the African American ASW under the AfrDE-EurDE model, our posterior parameter
712 estimation accuracy results were overall quantitatively slightly less accurately estimated
713 compared to those obtained for the ACB population, as indicated by overall larger credibility
714 intervals and cross-validation errors (**Figure 4B, Table 2, Table 3, Supplementary Table S3**).
715 This was consistent with the more ambiguous RF-ABC model-choice results obtained for this
716 population (**Figure 3**).

717 Comparing NN, RF, and Rejection ABC posterior parameter estimation accuracy

718 For posterior parameter estimations considering the ACB or the ASW population, the means of
719 the three types of errors (scaled mean-square error, mean-square error, absolute error, see
720 **Material and Methods**) were systematically lower for the two NN methods (joint or
721 independent posterior parameter estimation) than for the RF and Rejection independent
722 posterior parameter estimation methods (**Table 4**). Furthermore, we found that the means of the
723 three types of errors were qualitatively comparable between the NN estimation of the
724 parameters taken as a joint vector and the NN estimation of the parameters taken separately.
725 Altogether, these results showed that considering the NN estimation for parameters taken
726 jointly as a vector is overall preferable for the ACB and ASW populations, since it further
727 allowed the joint interpretation of parameter values estimated *a posteriori*, with little difference
728 in accuracy between the two methods.

729 Finally, results showed that the lengths of 95% credibility intervals estimated with NN joint
730 parameter estimation was, across all parameters, more accurate than all other methods with, on
731 average, 95.1% and 95.2% of true parameter values falling within the estimated 95% credibility
732 intervals, for the ACB and ASW respectively (**Supplementary Table S3**). Furthermore, we
733 found that lengths of 95% credibility intervals estimated with NN and RF independent posterior
734 parameter estimations were systematically under-estimated, with less than 94% of the true
735 parameter values falling into the 95% credibility intervals estimated. Finally, we found that

736 lengths of 95% credibility intervals estimated with the Rejection method were also rather
737 accurately estimated although on average slightly over-estimated compared to the NN joint
738 parameter estimation with, on average, 95.5% of the 1,000 cross-validation true parameter
739 values within the estimated 95% credibility intervals for the ACB, and 95.8% for the ASW.

740 Admixture histories of the African American ASW and Barbadian ACB

741 **Figure 5** visually synthesized the estimated posterior parameters of the complex admixture
742 scenarios reconstructed with our novel *MetHis* – machine-learning ABC framework, and
743 associated 95% credibility intervals (**Table 2**).

744 We found a virtual complete replacement of the ACB and ASW populations at generation 1
745 after foundation, thus consistent with our inability to accurately estimate the founding
746 proportions from the African and European sources at generation 0. Furthermore, we found an
747 increasingly precise posterior estimation of African and European contributions to the gene-
748 pool of the ACB and ASW populations forward in time, with most recent estimations exhibiting
749 narrow credibility intervals. This is also consistent with the nature of recurrent admixture
750 processes, where older information is often lost or replaced when more recent admixture events
751 occur.

752 Most interestingly, we found that the recurring contribution of the European gene pool to the
753 admixed populations rapidly decreases after generation 1 for both the ACB and ASW albeit
754 with substantial differences (**Figure 5**). Indeed, we found that the recurring contribution from
755 the European source to the ACB gene pool falls below 10% at generation 9 until no more than
756 1% in the present (generation 20). Comparatively, we found that the European contribution
757 diminished substantially less rapidly for the ASW, going below 10% only after generation 12
758 until roughly 2% in the present. This indicates that the European contribution to the African
759 American gene pool was more sustained over time than for the Barbadian.

760 Finally, we found substantial recurring contributions from the African source population to the
761 gene pool of both admixed populations (**Figure 5**). For the ACB population, we found a
762 progressive decrease of the African recurring introgression until a virtually constant recurring
763 admixture close to 28% from generation 10 and onward. For the ASW, our results showed a
764 sharper decrease of the African contribution after foundation until a virtually constant recurring
765 admixture process close to 24% from generation 5 until present (generation 20). The high
766 overall African recurring introgression into the admixed-populations gene pools captures the

767 importance of recurring admixture in explaining the observed patterns for both populations
768 descending from the TAST.

769

DISCUSSION

770 We evaluated how machine-learning Approximate Bayesian Computation methods can bring
771 new insights to the reconstruction of highly complex admixture histories using genetic data. To
772 illustrate our proof of concept and thoroughly investigate the power and accuracy of our
773 approach using real data, we aimed at reconstructing the recent complex admixture history for
774 the African American (ASW) and Barbadian (ACB) population samples from the 1000
775 Genomes project (Phase 3).

776 Our results demonstrated that our novel *MetHis* forward-in-time simulator and summary
777 statistics calculator coupled with RF-ABC scenario-choice can often clearly infer the best class
778 of highly complex admixture histories underlying independent SNP data diversity, in a
779 reasonable-size sample and genetic dataset. In the two source-populations admixture models
780 here investigated, we distinguished scenarios encompassing two pulses of admixture from each
781 source, after the founding admixture event, monotonically increasing or decreasing admixture
782 intensities over time, or a combination of these three scenarios. Furthermore, we found that
783 NN-ABC provide accurate posterior parameter inference of most demographic parameters of
784 recurring monotonically decreasing admixture processes, compared to other classes of ABC
785 posterior parameter inference methods. Finally, we empirically demonstrated that the moments
786 of the distribution of admixture fractions within the admixed population estimated using
787 independent SNPs were highly informative for reconstructing the admixture history using an
788 ABC approach, as expected theoretically (VERDU and ROSENBERG 2011; GRAVEL 2012).

789 While we found that distinguishing among competing models is more difficult in certain parts
790 of the parameter space due to scenario-nestedness (ROBERT *et al.* 2010), our *MetHis* – ABC
791 method already vastly extends the array of complex admixture models explored with most,
792 classically used, maximum-likelihood inference approaches (ROBERT *et al.* 2010; GRAVEL
793 2012; LOH *et al.* 2013; HELLENTHAL *et al.* 2014). It is challenging to analytically predict
794 genomic diversity patterns expected under realistic complex admixture histories, as likelihood
795 calculations under such models are very often intractable (VERDU and ROSENBERG 2011;
796 GRAVEL 2012; MEDINA *et al.* 2018; NI *et al.* 2019). In turn, this makes it difficult to understand
797 how most existing efficient maximum-likelihood admixture inference methods, which often
798 only consider one or two pulses of admixture, behave when the observed genetic data in fact
799 results from much more complex admixture processes (GRAVEL 2012; HELLENTHAL *et al.*
800 2014).

801 In this context, the proof of concept here presented more generally shows that ABC can be
802 fruitfully attempted to explore, virtually, any other admixture model beyond the case studies
803 here-conducted, provided that, *a priori*, simulation and summary statistics calculation are
804 feasible. To these ends, other recent efficient forward-in-time genetic data simulators can also
805 be successfully used in an ABC framework instead of *MetHis* (HALLER and MESSER 2019; NI
806 *et al.* 2019). In reality, studies investigating ABC approaches for admixture reconstruction,
807 while allowing for exploring scenarios out of reach of other methods, will inevitably face the
808 same difficulties as any ABC inference; such as high dimensional parameter and summary-
809 statistics spaces, lack of information from summary statistics, and scenario nestedness
810 (CSILLÉRY *et al.* 2010; ROBERT *et al.* 2010; SISSON *et al.* 2018).

811 Importantly, the current *MetHis* – ABC approach does not make use of admixture linkage-
812 disequilibrium patterns in the admixed population, and only relies on independent genetic
813 markers. Nevertheless, admixture LD has consistently proved to bring massive information
814 about the complex admixture history of numerous populations worldwide (GRAVEL 2012;
815 HELLENTHAL *et al.* 2014; MEDINA *et al.* 2018; NI *et al.* 2019). However, existing methods to
816 calculate admixture LD patterns remain computationally intensive and require numerous
817 markers and accurate phasing, which is difficult under ABC where such statistics have to be
818 calculated for each one of the numerous simulated datasets. In this context, RF-ABC (PUDLO
819 *et al.* 2016; RAYNAL *et al.* 2019) or AABC (BUZBAS and ROSENBERG 2015) methods allow
820 substantially diminishing the number of simulations required for satisfactory scenario-choice
821 and posterior parameter inference, which makes both approaches promising tools for using, in
822 the future, admixture LD patterns to reconstruct complex admixture processes from genomic
823 data.

824 Sex-biased admixture processes are known to have influenced admixed populations, and in
825 particular populations descending from the TAST (MORENO-ESTRADA *et al.* 2013; FORTES-
826 LIMA *et al.* 2018). Future version of our *MetHis* – ABC framework will explicitly implement
827 sex-specific admixture processes with, in addition to autosomal data, the possibility to
828 investigate sex-related genetic data (X-chromosome, Y-chromosome, and mitochondrial DNA)
829 (GOLDBERG *et al.* 2014; GOLDBERG and ROSENBERG 2015).

830 Finally, although *MetHis* readily allows considering changes of effective population size in the
831 admixed population at each generation as a parameter of interest to ABC inference, we did not,
832 for simplicity, investigate here how such changes affected our results for the African American

833 and Barbadian admixed population. Future work using *MetHis* will allow specifically
834 investigating how effective size changes may influence genetic patterns in the admixed
835 population, a question of major interest as numerous admixed populations are expected to have
836 experienced founding events and/or bottlenecks during their history (e.g. BROWNING *et al.*
837 2018).

838 For all these reasons, it is crucial, in general and in the future, to further develop novel
839 methodological tools and evaluate how genetic patterns evolve over time as a function of each
840 parameter of complex historical admixture models separately (BUZBAS and VERDU 2018;
841 MEDINA *et al.* 2018; NI *et al.* 2019). *MetHis* can help to this task since it allows the users to
842 investigate how parameters of the complex admixture process can influence, over time, a large
843 number of population genetics summary-statistics calculated in the simulated admixed
844 population at each generation.

845 Concerning the specific admixture history of the two admixed populations descending from the
846 TAST here reconstructed, note that several competing scenarios can clearly be discarded for
847 explaining the observed genetic patterns. In particular, the Afr2P – Eur2P scenario considering
848 two possible pulses of introgression after the founding event, separately from the African and
849 European source, does not significantly exceeds the prior probability of choosing any nine-
850 competing scenario (4.6% and 11.2% of the 1,000 votes for the Afr2P – Eur2P, respectively for
851 the ACB and the ASW, **Figure 3**). Note that this scenario embeds models analogous to the most
852 complex admixture scenarios that have been previously tested for these populations with
853 maximum-likelihood approaches based on extensive genome-wide data and admixture-LD
854 based statistics (GRAVEL 2012; BAHARIAN *et al.* 2016). Interestingly, very recent migrations
855 from either Africa or Europe to the Americas are known to have been intense demographically
856 in the 19th and 20th century (BERLIN 2010). However, the recent increased demographic
857 migrations do not seem to have left the equivalent signature in the genetic admixture process
858 of both the ACB and ASW populations, as monotonically recurring increasing admixture
859 scenarios can here be rejected confidently.

860 Nevertheless, we found that genetic admixture of African origin in both admixed populations,
861 although decreasing since foundation, retained high levels in the present day (between 20% and
862 30%). These results could stem from the known importance of African recurring forced
863 migrations during the TAST into the Americas; further prompts the influence of African slave
864 descendants forced migrations within the Americas after the initial crossing of the ocean (often

865 called the Middle Passage); and highlights the major importance of post-slavery migrations of
866 TAST descendant populations within the Americas (BERLIN 2010; ELTIS and RICHARDSON
867 2010; BAHARIAN *et al.* 2016). For instance, intense migrations from Haitian slave-descendants
868 in the 19th century have already been shown to possibly have contributed to the admixture
869 patterns of other populations in the Caribbean and continental America (MORENO-ESTRADA *et*
870 *al.* 2013; FORTES-LIMA *et al.* 2018).

871 Finally, we found that the genetic contribution from Europe rapidly decreases, after the
872 foundation of both admixed populations, to marginal amounts during the 20th century.
873 Therefore, it seems that neither sustained European migrations, nor the relaxation of social and
874 legal constraints on admixture between descendant communities subsequent to the abolition of
875 slavery and the end of segregation, have translated into increased European genetic contribution
876 to the gene-pool of admixed populations descending from European and African forced or
877 voluntary migrations into the Americas after the TAST.

878 Altogether, our results for the two recently-admixed human populations illustrated how our
879 *MetHis* – ABC framework can bring fundamental new insights into the complex demographic
880 history of admixed populations; a framework that can easily be adapted for investigating
881 admixture history in numerous populations and species, particularly when maximum-likelihood
882 methods are intractable.

883

884 **Figures Legends**

885

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

892

893 **Figure 2:** Random-Forest Approximate Bayesian Computation model-choice cross-validation.
894 (A) Heat map of the out-of-bag cross-validation results considering each 10,000 simulations
895 per each nine competing models (**Figure 1, Table 1**) in turn as pseudo-observed target for RF-
896 ABC model-choice. Out-of-bag prior error rate is 32.41%. RF-ABC model-choice performed
897 using 1,000 decision trees and 24 summary-statistics (see **Material and Methods**). (B)
898 Summary statistics’ respective importance in the RF-ABC model-choice out-of-bag cross-
899 validation (Pudlo et al. 2016).

900

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

905

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

914 kernel, and are constrained to the prior limits. Posterior parameter densities are indicated by a
915 solid line; prior parameter densities are indicated by black dotted lines.

916

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

924

925

926 **Supplementary Figure S1.** General mechanistic model of historical admixture from Verdu and
927 Rosenberg (2011).

928

929 **Supplementary Figure S2:** Comparison of individual admixture estimates using ASD-MDS
930 and ADMIXTURE for the Barbadian (ACB) and the African American (ASW). 100,000
931 independent SNPs were considered from the 1000 Genome Project Phase 3 for 279 unrelated
932 individuals (90 Yoruba (YRI), 89 British (GBR), 50 Barbadian (ACB), 50 African American
933 (ASW)). (A) Allele Sharing Dissimilarity was computed between all pairs of individuals and
934 the resulting matrix projected on the first two dimensions of a metric MDS. The two-
935 dimensional centroid of the Yoruba (YRI) and, respectively, the British (GBR) are indicated in
936 red and connected by a black dotted line. ACB and ASW individuals are projected orthogonally
937 onto this line and their relative distance to the Yoruba centroid is calculated to obtain ASD-
938 MDS based individual admixture estimates. (B) A single run of unsupervised ADMIXTURE
939 (Alexander et al. 2009) has been computed using the 279 individuals and 100,000 SNPs and
940 results were plotted using DISTRUCT (Rosenberg 2004). Individual membership proportions
941 to the “orange” cluster mostly represented by Yoruba (YRI) genotypes was considered as an
942 estimate of African admixture for the ACB and ASW respectively. (C) Spearman correlation

943 between ASD-MDS and ADMIXTURE-based estimates of African admixture for the ACB and
944 ASW individuals separately.

945

946 **Supplementary Figure S3:** Summary statistics prior-distribution densities for each nine
947 competing models considered (**Figure 1**). 10,000 simulations were performed for each nine-
948 competing scenario and prior densities plotted with a different color indicated for each scenario.
949 Corresponding statistics observed from the ACB and ASW population separately are
950 represented, on each plot, by vertical dotted-lines (red and blue respectively for ACB and ASW).
951 The 24 separate summary statistics considered are described in **Material and Methods**.

952

953 **Supplementary Figure S4:** Four first axes of the principal component analysis for the 90,000
954 sets of 24 summary statistics computed on simulated data under each nine-competing scenario
955 (**Figure 1**). The 24 same statistics calculated for the observed ACB and ASW population
956 samples, respectively, are then projected on the PCA and represented by, respectively, a red
957 and blue star. All two-dimensional projections are orthonormal.

958

959 **Supplementary Figure S5:** Histogram of the goodness-of-fit for the observed set of 24
960 summary statistics computed for (A) the ACB population, and (B) the ASW population, in turn
961 serving as the observed admixed population H considering the YRI population sample as the
962 African source and the GBR population sample as the European source (see **Material and**
963 **Methods**). Goodness-of-fit statistics were calculated as the mean distance between observed
964 and accepted summary statistics. Observed statistics are fitted to the full 90,000 sets of the same
965 statistics calculated from 10,000 simulations performed under each nine-competing models
966 (**Figure 1**). Goodness-of-fit was obtained considering 1,000 repetitions and a tolerance value
967 of 0.01.

968

969 **Supplementary Figure S6:** RF-ABC out-of-bag prior error rate as a function of the number of
970 trees considered to build the forest for the model-choice procedure considering nine-competing
971 scenarios (**Figure 1**).

972 **Tables Legends**

973

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

977

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

984

985 **Table 3.** Neural-Network Approximate Bayesian Computation posterior parameter errors under
986 the winning scenario AfrDE-EurDE, for the ACB and ASW populations. For each target
987 population separately, we conducted cross-validation by considering in turn 1,000 separate NN-
988 ABC parameter inferences each using in turn one of the 1,000 closest simulations to the
989 observed ACB (or ASW) data as the target pseudo-observed simulation. All posterior parameter
990 estimations were conducted using 100,000 simulations under scenario AfrDE-EurDE (**Figure**
991 **1, Table 1**), a 1% tolerance rate (1,000 simulations), 24 summary statistics, logit transformation
992 of all parameters, and four neurons in the hidden layer (see **Material and Methods**). Median
993 was considered as the point posterior parameter estimation for all parameters. First column
994 provides the average absolute error; second column shows the mean-squared error; third
995 column shows the mean-squared error scaled by the parameter's observed variance (see
996 **Material and Methods** for error formulas).

997

998 **Table 4.** Approximate Bayesian Computation mean posterior parameter errors under the
999 winning Scenario AfrDE-EurDE, for the ACB and ASW populations separately, using four
1000 different methods: NN estimation of the parameters taken jointly as a vector, NN estimation of
1001 the parameters taken separately, Random Forest (parameters taken separately), and Rejection

1002 (parameters taken separately). For each target population separately and for each method, we
1003 conducted an out-of-bag cross validation by considering in turn 1,000 separate parameter
1004 inferences each using one of the 1,000 closest simulation to the observed ACB (or ASW) data
1005 as the target pseudo-observed dataset. All posterior parameter estimations were conducted
1006 using the other 99,999 simulations under the AfrDE-EurDE scenario (**Figure 1, Table 1**), a 1%
1007 tolerance rate (i.e. 1,000 simulations), 24 summary statistics, logit transformation of all
1008 parameters, four neurons in the hidden layer per neural network and 500 trees per random forest.
1009 Median was considered as the point posterior parameter estimation for all parameters. First
1010 column provides the average absolute error; second column shows the mean-squared error;
1011 third column shows the mean-squared error scaled by the parameter's observed variance (see
1012 **Material and Methods** for error formulas).

1013

1014

1015 **Supplementary Table S1.** Random-Forest Approximate Bayesian Computation model-choice
1016 predictions for the ACB and ASW populations. 1,000 decision trees were considered for RF
1017 prediction for the ACB and ASW respectively. Corresponding results are plotted in **Figure 3**.

1018

1019 **Supplementary Table S2.** Parameter prediction cross-validation error as a function of the
1020 number of neurons in the hidden layer and the rejection tolerance rate under the AfrDE-EurDE
1021 scenario. We considered, 1,000 random simulations in turn as pseudo-observed data to estimate
1022 posterior parameter distributions, considering 4, 5, 6, or 7 neurons in the hidden layer (“NN-
1023 HL” row), and 100,000 total simulations. Tolerance levels of 0.01, 0.05, 0.1 and 0.2 were
1024 considered (“Tolerance” row). The median values of posterior parameter distributions were
1025 used as point estimates for the error calculation.

1026

1027 **Supplementary Table S3.** Accuracy of the 95% credibility interval estimated for posterior
1028 parameters in the vicinity of the observed ACB and ASW datasets. We provide the empirical
1029 coverage of the estimated 95% credibility interval, i.e. how many times (in percentage) the true
1030 parameter (θ_i) is found inside the estimated 95% credibility interval [$2.5\% \text{quantile}(\hat{\theta}_i)$;
1031 $97.5\% \text{quantile}(\hat{\theta}_i)$], among the 1,000 posterior parameter estimations conducted using in turn

1032 the 1,000 simulations closest to our real data, separately for the ACB and ASW, as pseudo-
1033 observed datasets for four separate methods : NN estimation of the parameters taken jointly as
1034 a vector, NN estimation of the parameters taken independently, Random Forest (parameters are
1035 taken independently), and Rejection (parameters are taken independently).

1036

1037 **Table 1.**

Parameter Names	Prior distribution	Condition	Models
$s_{Afr,0}$	<i>Uniform</i> [0,1]	-	all models
$t_{Afr,p1}$ $t_{Afr,p2}$	<i>Uniform</i> [0,20]	$t_{Afr,p1} \neq t_{Afr,p2}$	Afr2P models
$s_{Afr,1}, t_{Afr,p1}$ $s_{Afr,1}, t_{Afr,p2}$	<i>Uniform</i> [0,1]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	Afr2P models
$t_{Eur,p1}$ $t_{Eur,p2}$	<i>Uniform</i> [0,20]	$t_{Eur,p1} \neq t_{Eur,p2}$	Eur2P models
$s_{Eur,1}, t_{Eur,p1}$ $s_{Eur,1}, t_{Eur,p2}$	<i>Uniform</i> [0,1]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	Eur2P models
$s_{Afr,1}$	<i>Uniform</i> [0,1]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrDE models
$s_{Afr,20}$	<i>Uniform</i> [0, $s_{Afr,1} / 3$]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrDE models
u_{Afr}	<i>Uniform</i> [0,0.5]	-	AfrDE models
$s_{Eur,1}$	<i>Uniform</i> [0,1]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurDE models
$s_{Eur,20}$	<i>Uniform</i> [0, $s_{Eur,1} / 3$]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurDE models
u_{Eur}	<i>Uniform</i> [0,0.5]	-	EurDE models
$s_{Afr,1}$	<i>Uniform</i> [0, $s_{Afr,20} / 3$]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrIN models
$s_{Afr,20}$	<i>Uniform</i> [0,1]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrIN models
u_{Afr}	<i>Uniform</i> [0,0.5]	-	AfrIN models
$s_{Eur,1}$	<i>Uniform</i> [0, $s_{Eur,20} / 3$]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurIN models
$s_{Eur,20}$	<i>Uniform</i> [0,1]	For all g , $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurIN models
u_{Eur}	<i>Uniform</i> [0,0.5]	-	EurIN models

1038

1039 **Table 2.**

1040

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

1041

1042

1043 **Table 3.**

1044

<i>AfrDE-EurDE</i> <i>parameters</i>	ACB			ASW		
	Av. absolute Error	Mean-square Error	Mean-square Error / Var.	Av. absolute Error	Mean-square Error	Mean-square Error / Var.
$s_{Afr,0}$	0.2530	0.0857	1.0070	0.2444	0.0805	1.0081
$s_{Afr,1}$	0.1206	0.0216	0.8533	0.1158	0.0197	0.9259
$s_{Afr,20}$	0.02744	0.0012	0.4162	0.0219	0.0007	0.4773
u_{Afr}	0.1166	0.0198	0.9974	0.1254	0.0216	0.9757
$s_{Eur,1}$	0.0952	0.0164	1.0526	0.1001	0.0157	1.0152
$s_{Eur,20}$	0.0044	0.0001	0.6452	0.0069	0.0001	0.6623
u_{Eur}	0.1084	0.0174	0.9431	0.1021	0.0153	0.8036

1045

1046

1047 **Table 4.**

1048

<i>Posterior parameter estimation ABC method</i>	ACB			ASW		
	Av. absolute Error	Mean-squared Error	Mean-squared Error / Var.	Av. absolute Error	Mean-squared Error	Mean-squared Error / Var.
<i>NN joint</i>	0.1037	0.0232	0.8450	0.1024	0.0219	0.8383
<i>NN independent</i>	0.1032	0.0236	0.8294	0.1025	0.0225	0.8344
<i>RF independent</i>	0.1042	0.0246	0.8534	0.1036	0.0233	0.8697
<i>Rejection independent</i>	0.1071	0.0238	0.9299	0.1050	0.0223	0.8951

1049

1050

1051

1052 **Supplementary Table S1.**

1053

Competing Model Target population	Afr2P- Eur2P	Afr2P- EurDE	Afr2P- EurIN	AfrDE- Eur2P	AfrDE- EurDE	AfrDE- EurIN	AfrIN- Eur2P	AfrIN- EurDE	AfrIN- EurIN
<i>ACB</i>	46	144	3	151	531	12	74	34	5
<i>ASW</i>	112	106	9	317	335	3	73	43	2

1054

1055

1056

1057 **Supplementary Table S2.**

1058

AfrDE-EurDE	4	4	4	4	5	5	5	5	6	6	6	6	7	7	7	7
NN- HL																
Tolerance	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%
<i>S</i> _{Afr,0}	1.0161	0.9980	1.0003	1.0014	1.0037	1.0017	0.9987	0.9980	1.0018	0.9957	1.0015	0.9987	1.0063	0.9957	0.9981	0.9985
<i>S</i> _{Afr,1}	0.4588	0.4968	0.4924	0.4972	0.4877	0.4674	0.4841	0.4929	0.4763	0.4330	0.4702	0.5025	0.4837	0.4965	0.4613	0.4812
<i>S</i> _{Afr,20}	0.1420	0.2160	0.2976	0.3018	0.1468	0.2178	0.2678	0.3264	0.1455	0.2071	0.2738	0.3090	0.1312	0.2209	0.2765	0.3279
<i>u</i> _{Afr}	0.8800	0.8844	0.9355	0.9482	0.8759	0.8969	0.9040	0.9080	0.8309	0.8752	0.9017	0.9347	0.8621	0.9029	0.9344	0.9130
<i>S</i> _{Eur,1}	0.4445	0.4955	0.4822	0.5057	0.4804	0.4444	0.5097	0.4962	0.4596	0.4827	0.4693	0.4819	0.4836	0.4938	0.4673	0.5363
<i>S</i> _{Eur,20}	0.1589	0.2346	0.3071	0.3127	0.1272	0.2117	0.2522	0.3239	0.1173	0.2167	0.2923	0.2923	0.1552	0.2186	0.3164	0.3012
<i>u</i> _{Eur}	0.8574	0.8304	0.9038	0.9078	0.8340	0.8658	0.9161	0.9056	0.8305	0.8907	0.9069	0.9085	0.8403	0.8594	0.9159	0.9312
<i>Average error</i>	<i>0.5654</i>	<i>0.5937</i>	<i>0.6313</i>	<i>0.6393</i>	<i>0.5651</i>	<i>0.5865</i>	<i>0.6189</i>	<i>0.6359</i>	<i>0.5517</i>	<i>0.5859</i>	<i>0.6165</i>	<i>0.6325</i>	<i>0.5661</i>	<i>0.5983</i>	<i>0.6243</i>	<i>0.6413</i>

1059

1060

1061

1062 **Supplementary Table S3.**

1063

1064

<i>AfrDE-EurDE</i> <i>parameters</i>	ACB				ASW			
	<i>NN joint</i>	<i>NN indep.</i>	<i>RF indep.</i>	<i>Rejection indep.</i>	<i>NN joint</i>	<i>NN indep.</i>	<i>RF indep.</i>	<i>Rejection indep.</i>
$s_{Afr,0}$	0.956	0.934	0.929	0.952	0.952	0.931	0.937	0.950
$s_{Afr,1}$	0.958	0.929	0.942	0.968	0.958	0.914	0.942	0.963
$s_{Afr,20}$	0.964	0.926	0.956	0.971	0.963	0.928	0.960	0.978
u_{Afr}	0.953	0.932	0.930	0.950	0.944	0.914	0.925	0.945
$s_{Eur,1}$	0.947	0.909	0.939	0.949	0.950	0.912	0.930	0.955
$s_{Eur,20}$	0.944	0.908	0.930	0.957	0.952	0.919	0.929	0.968
u_{Eur}	0.941	0.919	0.927	0.943	0.947	0.928	0.936	0.952
Average credibility interval accuracy	0.951	0.922	0.936	0.955	0.952	0.920	0.937	0.958

1065

1066

1067

1068 **REFERENCES**

- 1069 1000 GENOMES PROJECT CONSORTIUM, 2015 A global reference for human genetic variation. *Nature* **526**:
1070 68-74.
- 1071 ALEXANDER, D. H., J. NOVEMBRE and K. LANGE, 2009 Fast model-based estimation of ancestry in
1072 unrelated individuals. *Genome Res* **19**: 1655-1664.
- 1073 BAHARIAN, S., M. BARAKATT, C. R. GIGNOUX, S. SHRINGARPURE, J. ERRINGTON *et al.*, 2016 The Great
1074 Migration and African-American Genomic Diversity. *PLoS Genet* **12**: e1006059.
- 1075 BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate Bayesian computation in population
1076 genetics. *Genetics*. **162**: 2025-2035.
- 1077 BERLIN, I., 2010 *The making of African America : the four great migrations*. Viking, New York.
- 1078 BERNSTEIN, F., 1931 Die geographische Verteilung der Blutgruppen und ihre anthropologische
1079 Bedeutung, pp. 227-243 in *Comitato Italiano per o studio dei problemi della popolazione*.
1080 Istituto Poligraphico dello Stato, Roma.
- 1081 BLUM, M. G. B., and O. FRANÇOIS, 2010 Non-linear regression models for Approximate Bayesian
1082 Computation. *Statistics and Computing* **20**: 63-67.
- 1083 BOITARD, S., W. RODRIGUEZ, F. JAY, S. MONA and F. AUSTERLITZ, 2016 Inferring Population Size History
1084 from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian
1085 Computation Approach. *PLoS Genet* **12**: e1005877.
- 1086 BOWCOCK, A. M., A. RUIZ-LINARES, J. TOMFOHRDE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of
1087 human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455-457.
- 1088 BRANDENBURG, J. T., T. MARY-HUARD, G. RIGAILL, S. J. HEARNE, H. CORTI *et al.*, 2017 Independent
1089 introductions and admixtures have contributed to adaptation of European maize and its
1090 American counterparts. *PLoS Genet* **13**: e1006666.
- 1091 BROWNING, S. R., B. L. BROWNING, M. L. DAVIGLUS, R. A. DURAZO-ARVIZU, N. SCHNEIDERMAN *et al.*, 2018
1092 Ancestry-specific recent effective population size in the Americas. *PLoS Genet* **14**: e1007385.
- 1093 BUZBAS, E. O., and N. A. ROSENBERG, 2015 AABC: approximate approximate Bayesian computation for
1094 inference in population-genetic models. *Theor Popul Biol* **99**: 31-42.
- 1095 BUZBAS, E. O., and P. VERDU, 2018 Inference on admixture fractions in a mechanistic model of
1096 recurrent admixture. *Theor Popul Biol* **122**: 149-157.
- 1097 CAVALLI-SFORZA, L. L., and W. F. BODMER, 1971 *The genetics of human populations*. W. H. Freeman, San
1098 Francisco,.
- 1099 CHAKRABORTY, R., and K. M. WEISS, 1988 Admixture as a tool for finding linked genes and detecting that
1100 difference from allelic association between loci. *Proc Natl Acad Sci U S A* **85**: 9119-9123.
- 1101 CHIMUSA, E. R., J. DEFO, P. K. THAMI, D. AWANY, D. D. MULISA *et al.*, 2018 Dating admixture events is
1102 unsolved problem in multi-way admixed populations. *Brief Bioinform*.
- 1103 CORNUET, J. M., P. PUDLO, J. VEYSSIER, A. DEHNE-GARCIA, M. GAUTIER *et al.*, 2014 DIYABC v2.0: a software
1104 to make approximate Bayesian computation inferences about population history using single
1105 nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* **30**: 1187-
1106 1189.
- 1107 CSILLÉRY, K., M. G. BLUM, O. E. GAGGIOTTI and O. FRANCOIS, 2010 Approximate Bayesian Computation
1108 (ABC) in practice. *Trends Ecol Evol* **25**: 410-418.
- 1109 CSILLÉRY, K., O. FRANÇOIS and M. G. B. BLUM, 2012 abc: an R package for approximate Bayesian
1110 computation (ABC). *Methods in Ecology and Evolution* **3**: 475-479.
- 1111 DANECEK, P., A. AUTON, G. ABECASIS, C. A. ALBERS, E. BANKS *et al.*, 2011 The variant call format and
1112 VCFtools. *Bioinformatics* **27**: 2156-2158.
- 1113 ELTIS, D., and D. RICHARDSON, 2010 Atlas of the transatlantic slave trade, pp. in *The Lewis Walpole*
1114 *series in eighteenth-century culture and history*. Yale University Press,, New Haven.
- 1115 ESTOUP, A., L. RAYNAL, P. VERDU and J. M. MARIN, 2018 Model choice using Approximate Bayesian
1116 Computation and Random Forests: analyses based on model grouping to make inferences
1117 about the genetic history of Pygmy human populations. *Journal of the Sfds* **159**: 167-190.

- 1118 EWENS, W. J., and R. S. SPIELMAN, 1995 The transmission/disequilibrium test: history, subdivision, and
1119 admixture. *Am J Hum Genet* **57**: 455-464.
- 1120 EXCOFFIER, L., I. DUPANLOUP, E. HUERTA-SANCHEZ, V. C. SOUSA and M. FOLL, 2013 Robust demographic
1121 inference from genomic and SNP data. *PLoS Genet* **9**: e1003905.
- 1122 EXCOFFIER, L., and M. FOLL, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic
1123 diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**: 1332-1334.
- 1124 FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2003 Inference of population structure using multilocus
1125 genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567-1587.
- 1126 FISHER, R. A., 1922 Darwinian evolution of mutations. *Eugen Rev* **14**: 31-34.
- 1127 FOLL, M., H. SHIM and J. D. JENSEN, 2015 WFABC: a Wright-Fisher ABC-based approach for inferring
1128 effective population sizes and selection coefficients from time-sampled data. *Mol Ecol*
1129 *Resour* **15**: 87-98.
- 1130 FORTES-LIMA, C., J. BYBJERG-GRAUHM, L. C. MARIN-PADRON, E. J. GOMEZ-CABEZAS, M. BAEKVAD-HANSEN *et al.*,
1131 2018 Exploring Cuba's population structure and demographic history using genome-wide
1132 data. *Sci Rep* **8**: 11422.
- 1133 FORTES-LIMA, C., A. GESSAIN, A. RUIZ-LINARES, M. C. BORTOLINI, F. MIGOT-NABIAS *et al.*, 2017 Genome-wide
1134 Ancestry and Demographic History of African-Descendant Maroon Communities from French
1135 Guiana and Suriname. *Am J Hum Genet* **101**: 725-736.
- 1136 FRAIMOUT, A., V. DEBAT, S. FELLOUS, R. A. HUFBAUER, J. FOUCAUD *et al.*, 2017 Deciphering the Routes of
1137 invasion of *Drosophila suzukii* by Means of ABC Random Forest. *Mol Biol Evol* **34**: 980-996.
- 1138 GOLDBERG, A., and N. A. ROSENBERG, 2015 Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased
1139 Admixture on the X Chromosome. *Genetics* **201**: 263-279.
- 1140 GOLDBERG, A., P. VERDU and N. A. ROSENBERG, 2014 Autosomal admixture levels are informative about
1141 sex bias in admixed populations. *Genetics* **198**: 1209-1229.
- 1142 GRAVEL, S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607-619.
- 1143 GUAN, Y., 2014 Detecting structure of haplotypes and local ancestry. *Genetics* **196**: 625-642.
- 1144 HALLER, B. C., and P. W. MESSER, 2019 SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher
1145 Model. *Mol Biol Evol* **36**: 632-637.
- 1146 HELICONIUS GENOME CONSORTIUM, 2012 Butterfly genome reveals promiscuous exchange of mimicry
1147 adaptations among species. *Nature* **487**: 94-98.
- 1148 HELLENTHAL, G., G. B. J. BUSBY, G. BAND, J. F. WILSON, C. CAPELLI *et al.*, 2014 A genetic atlas of human
1149 admixture history. *Science* **343**: 747-751.
- 1150 JAY, F., S. BOITARD and F. AUSTRERLITZ, 2019 An ABC Method for Whole-Genome Sequence Data:
1151 Inferring Paleolithic and Neolithic Human Expansions. *Mol Biol Evol* **36**: 1565-1579.
- 1152 JEONG, C., G. ALKORTA-ARANBURU, B. BASNYAT, M. NEUPANE, D. B. WITONSKY *et al.*, 2014 Admixture
1153 facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* **5**: 3281.
- 1154 LAWSON, D. J., G. HELLENTHAL, S. MYERS and D. FALUSH, 2012 Inference of population structure using
1155 dense haplotype data. *PLoS Genet* **8**: e1002453.
- 1156 LIPSON, M., P. R. LOH, A. LEVIN, D. REICH, N. PATTERSON *et al.*, 2013 Efficient moment-based inference of
1157 admixture parameters and sources of gene flow. *Mol Biol Evol* **30**: 1788-1802.
- 1158 LOH, P. R., M. LIPSON, N. PATTERSON, P. MOORJANI, J. K. PICKRELL *et al.*, 2013 Inferring admixture histories
1159 of human populations using linkage disequilibrium. *Genetics* **193**: 1233-1254.
- 1160 LONG, J. C., 1991 The genetic structure of admixed populations. *Genetics* **127**: 417-428.
- 1161 MAPLES, B. K., S. GRAVEL, E. E. KENNY and C. D. BUSTAMANTE, 2013 RFMix: a discriminative modeling
1162 approach for rapid and robust local-ancestry inference. *Am J Hum Genet* **93**: 278-288.
- 1163 MARTIN, A. R., C. R. GIGNOUX, R. K. WALTERS, G. L. WOJCIK, B. M. NEALE *et al.*, 2017 Human Demographic
1164 History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**:
1165 635-649.
- 1166 MEDINA, P., B. THORNLOW, R. NIELSEN and R. CORBETT-DETIG, 2018 Estimating the Timing of Multiple
1167 Admixture Pulses During Local Ancestry Inference. *Genetics* **210**: 1089-1107.
- 1168 MOORJANI, P., N. PATTERSON, J. N. HIRSCHHORN, A. KEINAN, L. HAO *et al.*, 2011 The history of African gene
1169 flow into Southern Europeans, Levantines, and Jews. *PLoS Genet* **7**: e1001373.

- 1170 MORENO-ESTRADA, A., S. GRAVEL, F. ZAKHARIA, J. L. MCCAULEY, J. K. BYRNES *et al.*, 2013 Reconstructing the
1171 population genetic history of the Caribbean. *PLoS Genet* **9**: e1003925.
- 1172 NEI, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of
1173 individuals. *Genetics* **89**: 583-590.
- 1174 NI, X., K. YUAN, C. LIU, Q. FENG, L. TIAN *et al.*, 2019 MultiWaver 2.0: modeling discrete and continuous
1175 gene flow to reconstruct complex population admixtures. *Eur J Hum Genet* **27**: 133-139.
- 1176 PASCHOU, P., E. ZIV, E. G. BURCHARD, S. CHOUDHRY, W. RODRIGUEZ-CINTRON *et al.*, 2007 PCA-correlated
1177 SNPs for structure identification in worldwide human populations. *PLoS Genet* **3**: 1672-1686.
- 1178 PATIN, E., M. LOPEZ, R. GROLLEMUND, P. VERDU, C. HARMANT *et al.*, 2017 Dispersals and genetic
1179 adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**: 543-
1180 546.
- 1181 PATTERSON, N., P. MOORJANI, Y. LUO, S. MALLICK, N. ROHLAND *et al.*, 2012 Ancient admixture in human
1182 history. *Genetics* **192**: 1065-1093.
- 1183 PICKRELL, J. K., and J. K. PRITCHARD, 2012 Inference of population splits and mixtures from genome-wide
1184 allele frequency data. *PLoS Genet* **8**: e1002967.
- 1185 POOL, J. E., and R. NIELSEN, 2009 Inference of historical changes in migration rate from the lengths of
1186 migrant tracts. *Genetics* **181**: 711-719.
- 1187 PRICE, A. L., A. TANDON, N. PATTERSON, K. C. BARNES, N. RAFAELS *et al.*, 2009 Sensitive detection of
1188 chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**:
1189 e1000519.
- 1190 PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN and M. W. FELDMAN, 1999 Population growth of
1191 human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**: 1791-
1192 1798.
- 1193 PUDLO, P., J. M. MARIN, A. ESTOUP, J. M. CORNUET, M. GAUTIER *et al.*, 2016 Reliable ABC model choice via
1194 random forests. *Bioinformatics* **32**: 859-866.
- 1195 PURCELL, S., B. NEALE, K. TODD-BROWN, L. THOMAS, M. A. FERREIRA *et al.*, 2007 PLINK: a tool set for whole-
1196 genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559-575.
- 1197 R DEVELOPMENT CORE TEAM, 2017 R: A language and environment for statistical computing, pp. R
1198 Foundation for Statistical Computing, Vienna, Austria.
- 1199 RACIMO, F., S. SANKARARAMAN, R. NIELSEN and E. HUERTA-SANCHEZ, 2015 Evidence for archaic adaptive
1200 introgression in humans. *Nat Rev Genet* **16**: 359-371.
- 1201 RAYNAL, L., J. M. MARIN, P. PUDLO, M. RIBATET, C. P. ROBERT *et al.*, 2019 ABC random forests for Bayesian
1202 parameter inference. *Bioinformatics* **35**: 1720-1728.
- 1203 REICH, D., K. THANGARAJ, N. PATTERSON, A. L. PRICE and L. SINGH, 2009 Reconstructing Indian population
1204 history. *Nature* **461**: 489-494.
- 1205 ROBERT, C. P., K. MENGERSEN and C. CHEN, 2010 Model choice versus model criticism. *Proc Natl Acad Sci*
1206 *U S A* **107**: E5; author reply E6-7.
- 1207 SALTER-TOWNSHEND, M., and S. MYERS, 2019 Fine-Scale Inference of Ancestry Segments Without Prior
1208 Knowledge of Admixing Groups. *Genetics* **212**: 869-889.
- 1209 SANKARARAMAN, S., S. SRIDHAR, G. KIMMEL and E. HALPERIN, 2008 Estimating local ancestry in admixed
1210 populations. *Am J Hum Genet* **82**: 290-303.
- 1211 SHRINER, D., A. ADEYEMO, E. RAMOS, G. CHEN and C. N. ROTIMI, 2011 Mapping of disease-associated
1212 variants in admixed populations. *Genome Biol* **12**: 223.
- 1213 SISSON, S. A., Y. FAN and M. A. BEAUMONT, 2018 *Handbook of Approximate Bayesian Computation*. .
1214 Chapman and Hall/CRC, New York, USA.
- 1215 SKOGLUND, P., E. ERSMARK, E. PALKOPOULOU and L. DALEN, 2015 Ancient wolf genome reveals an early
1216 divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol* **25**:
1217 1515-1519.
- 1218 TAVARÉ, S., D. J. BALDING, R. C. GRIFFITHS and P. DONNELLY, 1997 Inferring coalescence times from DNA
1219 sequence data. *Genetics* **145**: 505-518.
- 1220 VERDU, P., F. AUSTERLITZ, A. ESTOUP, R. VITALIS, M. GEORGES *et al.*, 2009 Origins and genetic diversity of
1221 pygmy hunter-gatherers from Western Central Africa. *Curr Biol* **19**: 312-318.

- 1222 VERDU, P., E. M. JEWETT, T. J. PEMBERTON, N. A. ROSENBERG and M. BAPTISTA, 2017 Parallel Trajectories of
1223 Genetic and Linguistic Admixture in a Genetically Admixed Creole Population. *Curr Biol* **27**:
1224 2529-2535 e2523.
- 1225 VERDU, P., and N. A. ROSENBERG, 2011 A general mechanistic model for admixture histories of hybrid
1226 populations. *Genetics* **189**: 1413-1426.
- 1227 WAKELEY, J., L. KING, B. S. LOW and S. RAMACHANDRAN, 2012 Gene genealogies within a fixed pedigree,
1228 and the robustness of Kingman's coalescent. *Genetics* **190**: 1433-1445.
- 1229 WEGMANN, D., C. LEUENBERGER and L. EXCOFFIER, 2009 Efficient approximate Bayesian computation
1230 coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**: 1207-1218.
- 1231 WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-Statistics for the Analysis of Population-Structure.
1232 *Evolution* **38**: 1358-1370.
- 1233 WRIGHT, S., 1931 Evolution in Mendelian Populations. *Genetics* **16**: 97-159.

1234

1235

Supplementary note S1

We used the rectangular hyperbola class of functions to obtain increasing/decreasing patterns using only one shape parameter. We give here the derivation of the equations used, giving the example of a decreasing pattern.

A decreasing hyperbola is given by the function:

$$f(x) = \frac{a(1-x)}{a+x} \quad (\text{S1.1})$$

with $x \in [0; 1]$, $f(x) \in [0; 1]$ and $a \in [0; +\infty[$. Parameter a controls the shape (“steepness”) of the curve obtained (figure S1.1).

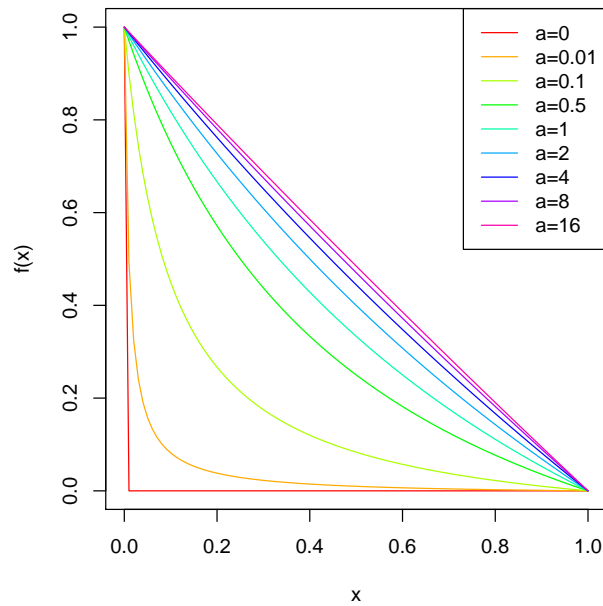


Figure S1.1: Influence of a on equation S1.1

The intersection between the hyperbola and $y = x$ is given by

$$x = y = -a + \sqrt{a + a^2}$$

thus, we can sample an uniform deviate $u \in [0; \frac{1}{2}]$ and set parameter a :

$$a = \frac{u^2}{1 - 2u}$$

to obtain all hyperbola shapes.

We then transformed equation S1.1 to rescale the ranges of x and $f(x)$:

$$f(x) = \frac{a(y_{max} - y_{min})\left(1 - \frac{x - x_{min}}{x_{max} - x_{min}}\right)}{a + \frac{x - x_{min}}{x_{max} - x_{min}}} + y_{min} \quad (\text{S1.2})$$

with $x \in [x_{min}; x_{max}]$ and $f(x) \in [y_{min}; y_{max}]$ (figure S1.2).

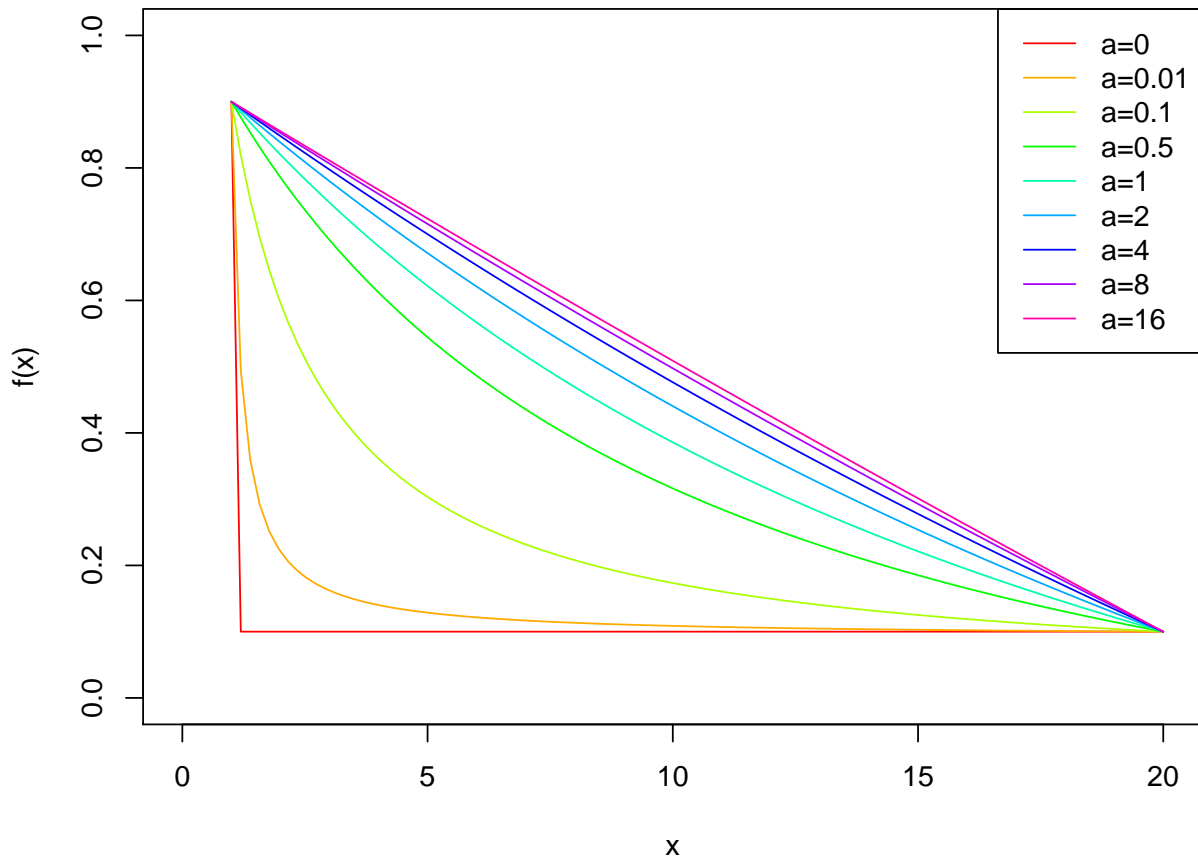


Figure S1.2: Influence of a on equation S1.2 with $x_{min} = 1$, $x_{max} = 20$, $y_{min} = 0.1$ and $y_{max} = 0.9$

With the notation used in the main text for contributions, and considering 20 generations of admixture, we obtain:

$$s_{S,g} = \frac{a(s_{S,1} - s_{S,20})\left(1 - \frac{g-1}{20-1}\right)}{a + \frac{g-1}{20-1}} + s_{S,20} \quad (\text{S1.3})$$

and an example of the patterns obtained for different u values is given in figure S1.3.

Literature Cited in Supplementary Materials

Rosenberg, N. A., 2004 DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* 4: 137–138.

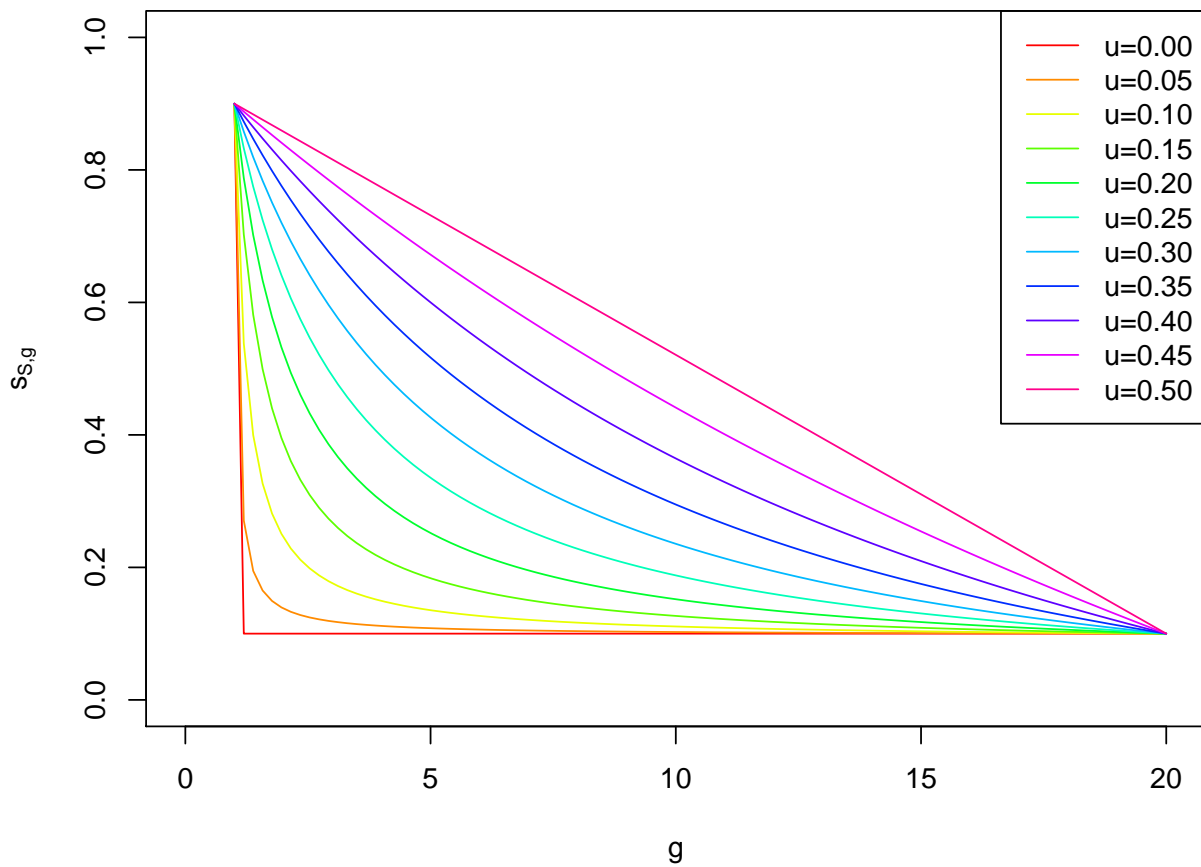


Figure S1.3: Influence of u on equation S1.3 with $s_{S,20} = 0.1$ and $s_{S,1} = 0.9$

Figure 1

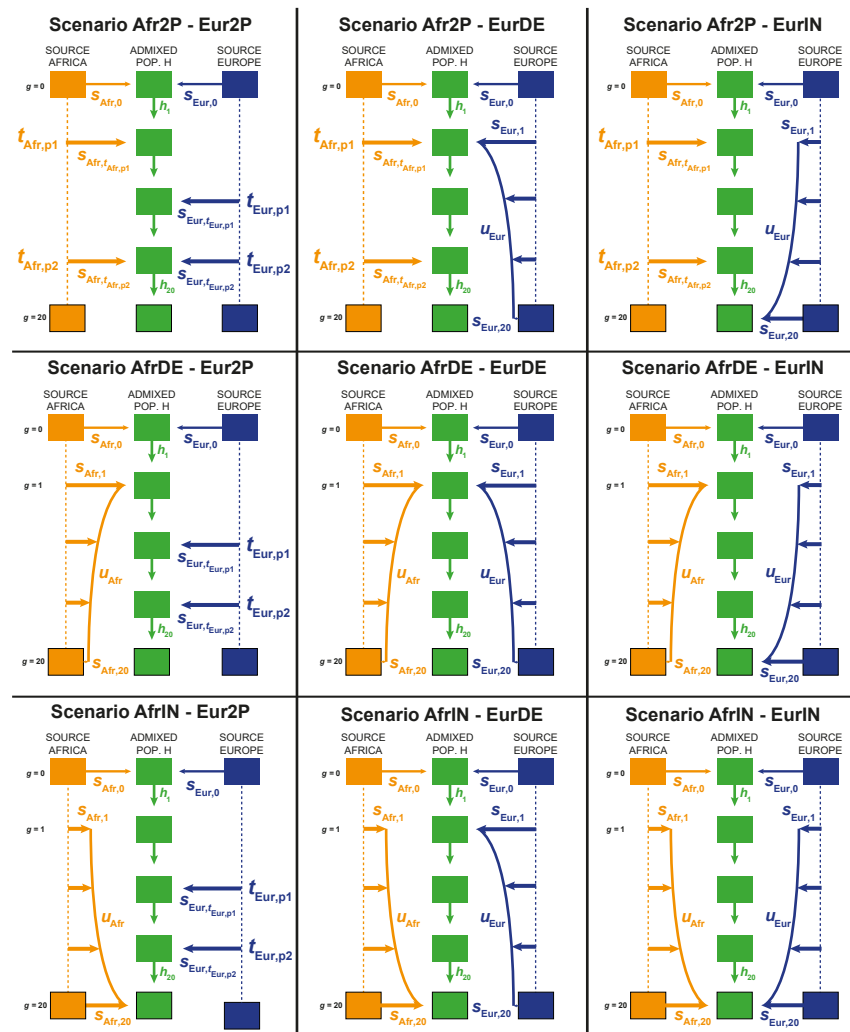


Figure 2

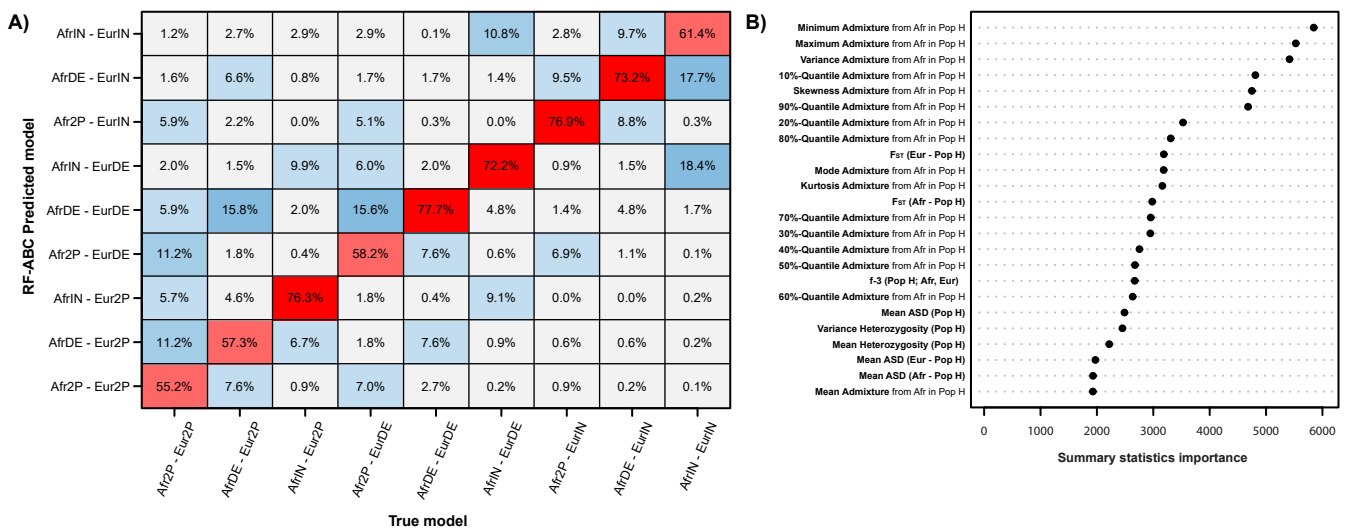


Figure 3

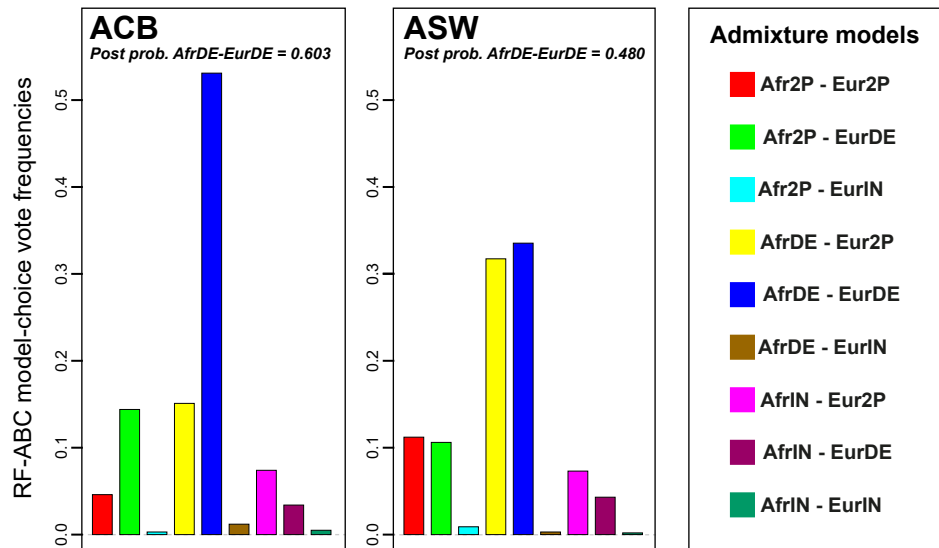


Figure 4

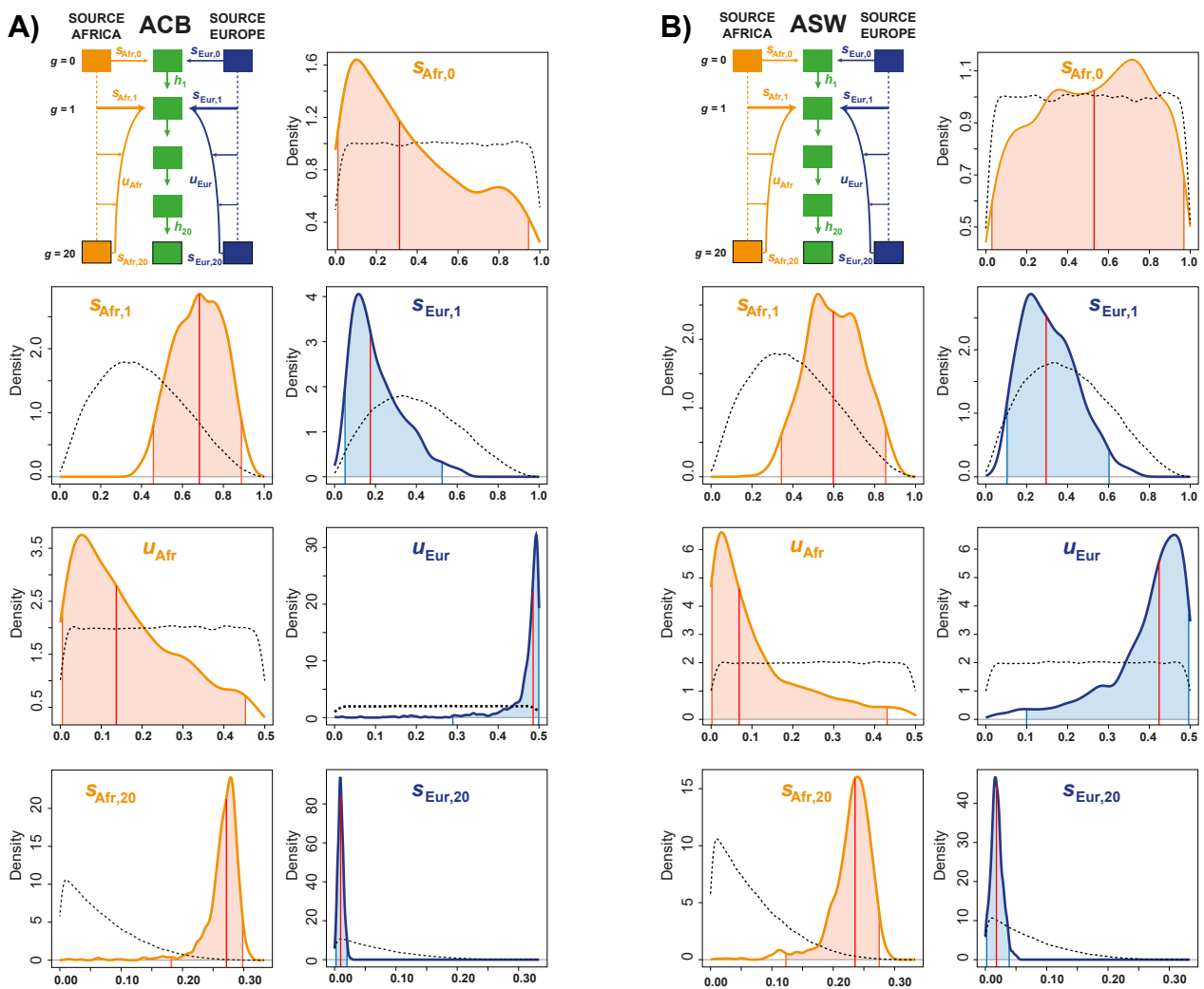
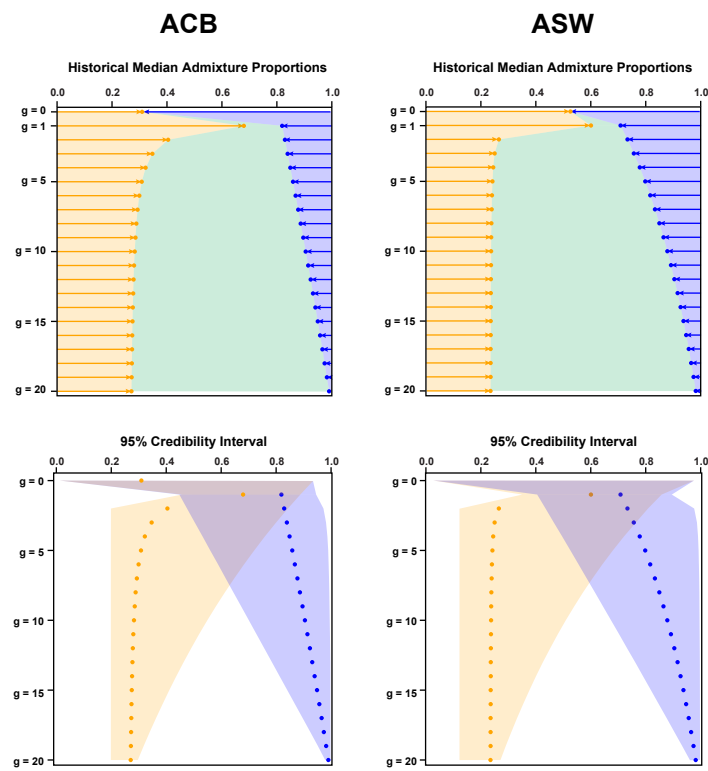
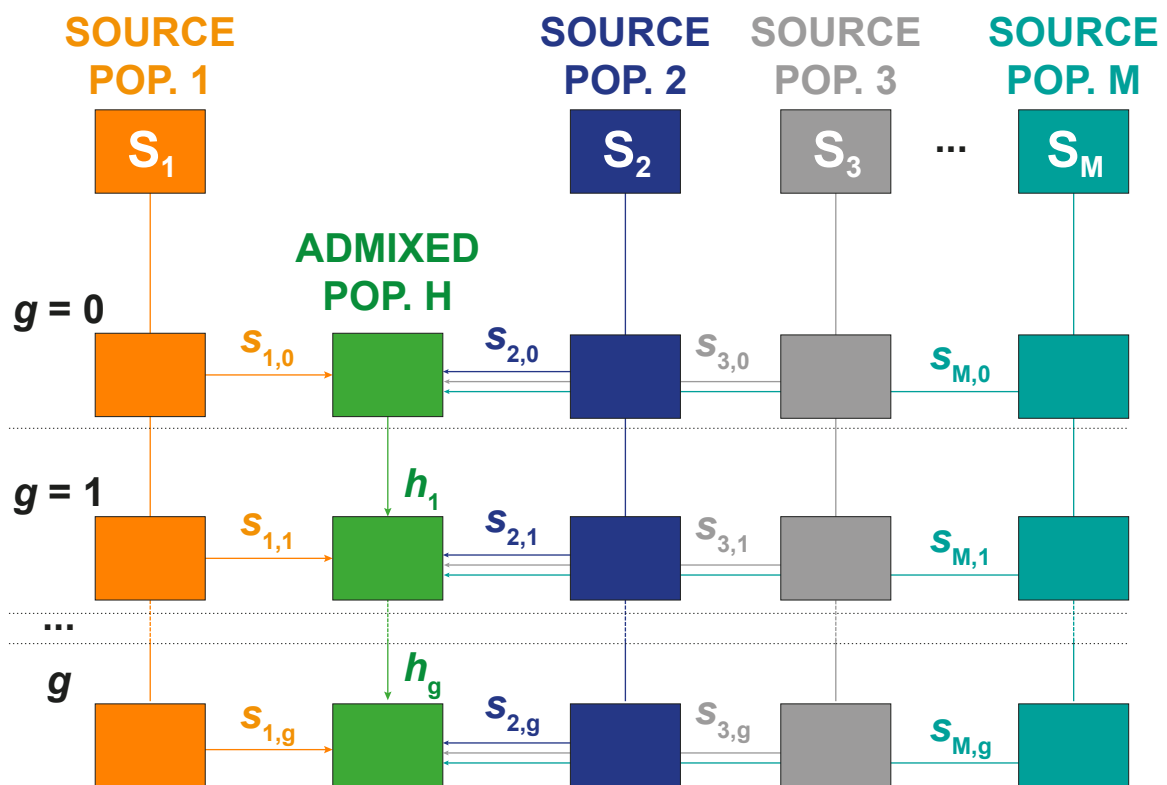


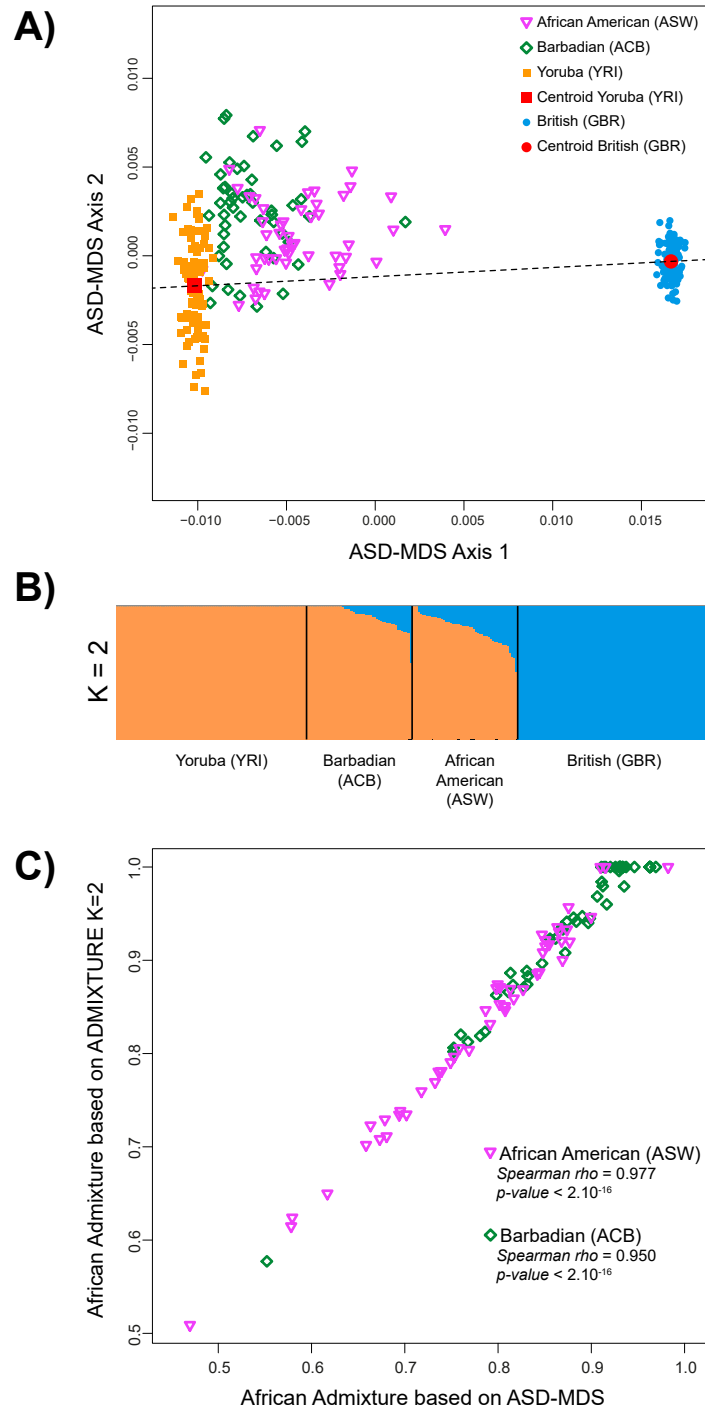
Figure 5



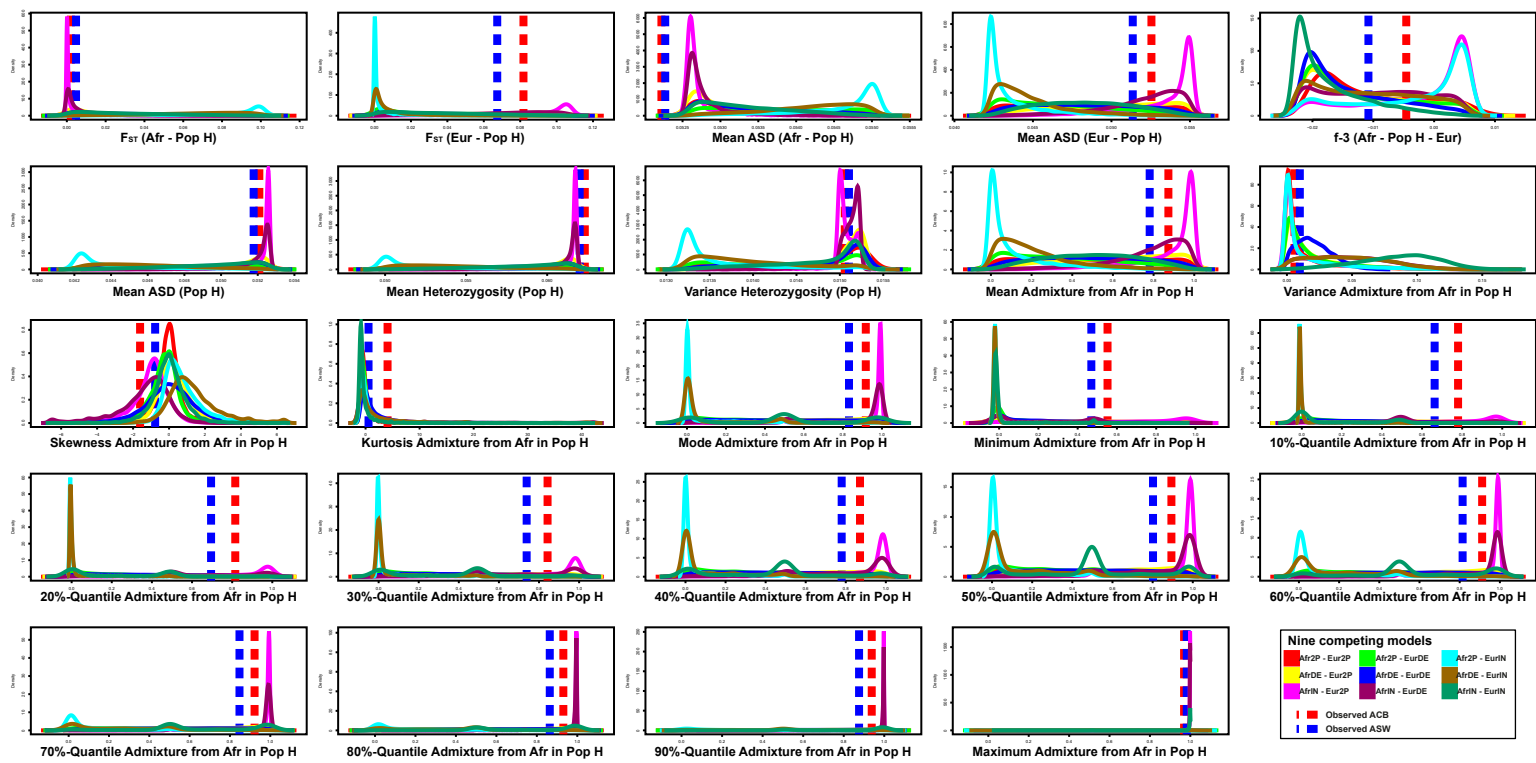
Supplementary Figure S1



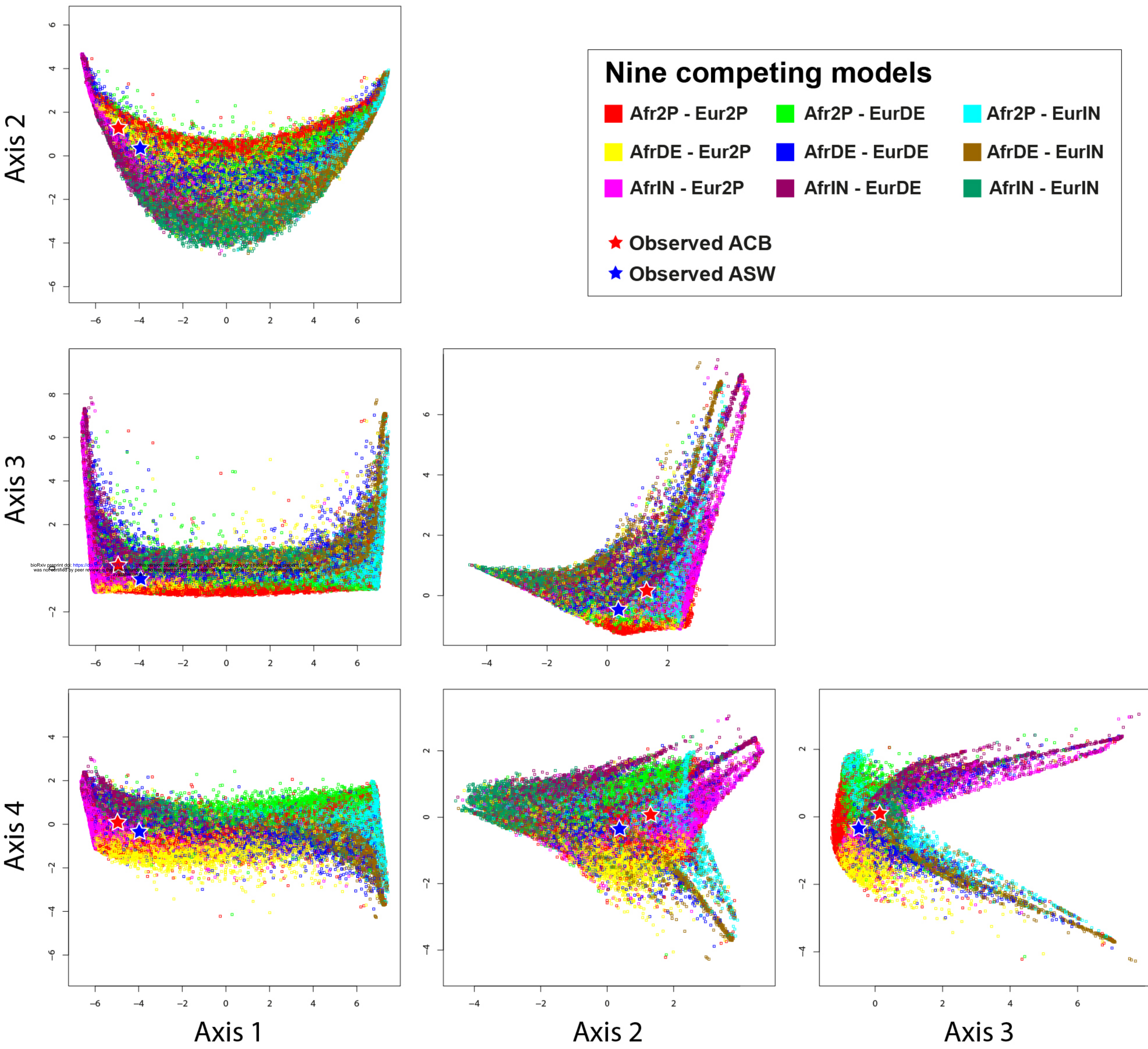
Supplementary Figure S2



Supplementary Figure S3

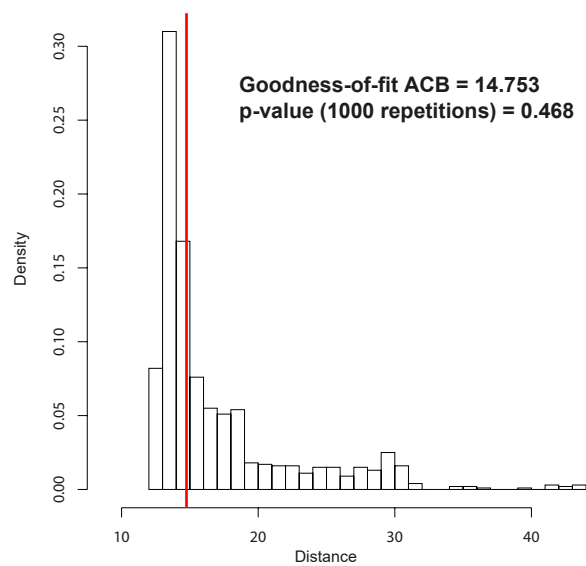


Supplementary Figure S4

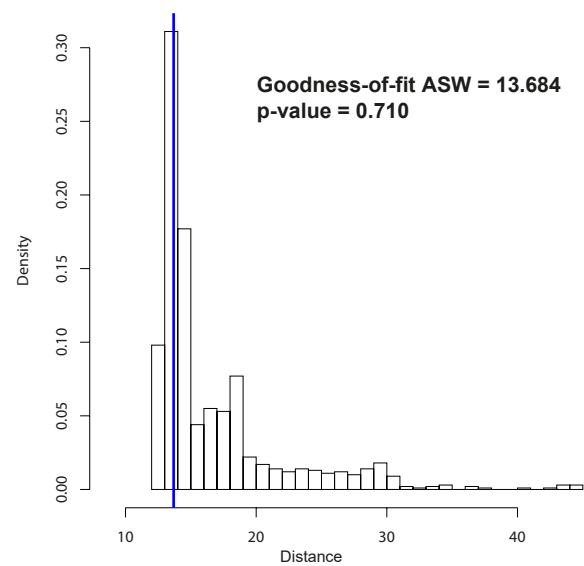


Supplementary Figure S5

A)



B)



Supplementary Figure S6

