MOLECULAR ECOLOGY RESOURCES

Supplemental Tables and Figures for:

Complex genetic admixture histories reconstructed with Approximate Bayesian Computation

Cesar A. Fortes-Lima[1,2,*] | Romain Laurent[1,*] | Valentin Thouzeau[3,4] | Bruno Toupance[1] | Paul Verdu[1,#]

[1] CNRS, Muséum National d'Histoire Naturelle, Université de Paris, UMR7206 Eco-anthropologie, Paris, France
[2] Sub-department of Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
[3] CNRS, Université Paris-Dauphine, PSL University, UMR7534 Centre de Recherche en Mathématiques de la Décision, Paris, France
[4] ENS, PSL University, EHESS, CNRS, Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Etudes Cognitives, Paris, France

* Joint first authors
# Author for correspondence: Paul Verdu,
CNRS, Muséum National d'Histoire Naturelle, Université de Paris;
UMR7206 Eco-anthropologie (EA);
Address: Musée de l'Homme, 17, place du Trocadéro, 75016 Paris, France;
email: paul.verdu@mnhn.fr; tel: +33 1 44 05 73 17

KEYWORDS
Admixture; Approximate Bayesian Computation; Inference; Population Genetics; Machine Learning

MOLECULAR ECOLOGY
RESOURCES

Supplementary Table ST1. Random-Forest Approximate Bayesian Computation model-choice predictions for the ACB and ASW populations. 1,000 decision trees were considered for RF prediction for the ACB and ASW respectively. Corresponding results are plotted in Figure 3.

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

MOLECULAR ECOLOGY
RESOURCES

Supplementary Table ST2. Parameter prediction cross-validation error as a function of the number of neurons in the hidden layer and the rejection tolerance rate under the AfrDE-EurDE scenario. We considered, 1,000 random simulations in turn as pseudo-observed data to estimate posterior parameter distributions, considering 4, 5, 6, or 7 neurons in the hidden layer ("NN-HL" row), and 100,000 total simulations. Tolerance levels of 0.01, 0.05, 0.1 and 0.2 were considered ("Tolerance" row). The median values of posterior parameter distributions were used as point estimates for the error calculation.

AfrDE-EurDE NN- HL	4	4	4	4	5	5	5	5	6	6	6	6	7	7	7	7
Tolerance	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%	1%	5%	10%	20%
$s_{Afr,0}$	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
$s_{Afr,1}$	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
$s_{Afr,20}$	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
u_{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
$s_{Eur,1}$	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
$s_{Eur,20}$	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
u_{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
Average error	***0.5654***	***0.5937***	***0.6313***	***0.6393***	***0.5651***	***0.5865***	***0.6189***	***0.6359***	***0.5517***	***0.5859***	***0.6165***	***0.6325***	***0.5661***	***0.5983***	***0.6243***	***0.6413***

MOLECULAR ECOLOGY
RESOURCES

Supplementary Table ST3. Accuracy of the 95% credibility interval estimated for posterior parameters in the vicinity of the observed ACB and ASW datasets under the winning scenario AfrDE-EurDE. We provide the empirical coverage of the estimated 95% credibility interval, i.e. how many times (in percentage) the true parameter (θ_i) is found inside the estimated 95% credibility interval [2.5%quantile$(\hat{\theta}_i)$; 97.5%quantile$(\hat{\theta}_i)$], among the 1,000 posterior parameter estimations conducted using in turn the 1,000 simulations closest to our real data, separately for the ACB and ASW, as pseudo-observed datasets for four separate methods : NN estimation of the parameters taken jointly as a vector, NN estimation of the parameters taken independently, Random Forest (parameters are taken independently), and Rejection (parameters are taken independently).

	ACB				**ASW**			
AfrDE-EurDE parameters	***NN joint***	***NN indep.***	***RF indep.***	***Rejection indep.***	***NN joint***	***NN indep.***	***RF indep.***	***Rejection indep.***
$s_{Afr,0}$	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***

MOLECULAR ECOLOGY
RESOURCES

Supplementary Table ST4. Neural-Network Approximate Bayesian Computation posterior parameter errors under the loosing scenario Afr2P-Eur2P, for the ACB and ASW populations. For each target population separately, we conducted cross-validation by considering in turn 1,000 separate NN-ABC parameter inferences each using in turn one of the 1,000 closest simulations to the observed ACB (or ASW) data as the target pseudo-observed simulation. All posterior parameter estimations were conducted using 100,000 simulations under scenario Afr2P-Eur2P (**Figure 1**, **Table 1**), a 1% tolerance rate (1,000 simulations), 24 summary statistics, logit transformation of all parameters, and four neurons in the hidden layer (see **Materials and Methods**). Median was considered as the point posterior parameter estimation for all parameters. First column provides the average absolute error; second column shows the mean-squared error; third column shows the mean-squared error scaled by the parameter's observed variance (see **Materials and Methods** for error formulas).

As expected, posterior estimations errors of all parameters are high under this loosing scenario, compared to the winning scenario AfrDE-EurDE (**Table 3**). This shows that there is no information in our data about the parameters of this scenario.

AfrDE-EurDE parameters	**ACB**			**ASW**		
	Av. absolute Error	Mean-square Error	Mean-square Error / Var.	Av. absolute Error	Mean-square Error	Mean-square Error / Var.
$S_{Afr,0}$	0.2477	0.0824	1.0031	0.2443	0.0809	1.0084
$S_{Afr,tAfr,p1}$	0.2337	0.0818	0.9531	0.2318	0.0752	1.0037
$S_{Afr,tAfr,p2}$	0.1263	0.0499	0.9503	0.1519	0.0481	1.0223
$t_{Afr,p1}$	4.2016	24.9695	0.9063	4.0694	22.8485	0.9168
$t_{Afr,p2}$	1.2263	3.9127	0.5173	1.0627	3.2155	0.5410
$S_{Eur,tEur,p1}$	0.2423	0.0808	0.9938	0.2316	0.0756	0.9905
$S_{Eur,tEur,p2}$	0.2385	0.0854	0.8911	0.2276	0.0814	0.9015
$t_{Eur,p1}$	3.2692	16.729	1.0001	3.3028	16.6287	1.0421
$t_{Eur,p2}$	3.2238	16.5689	0.7334	2.8862	14.6887	0.9302

MOLECULAR ECOLOGY
RESOURCES

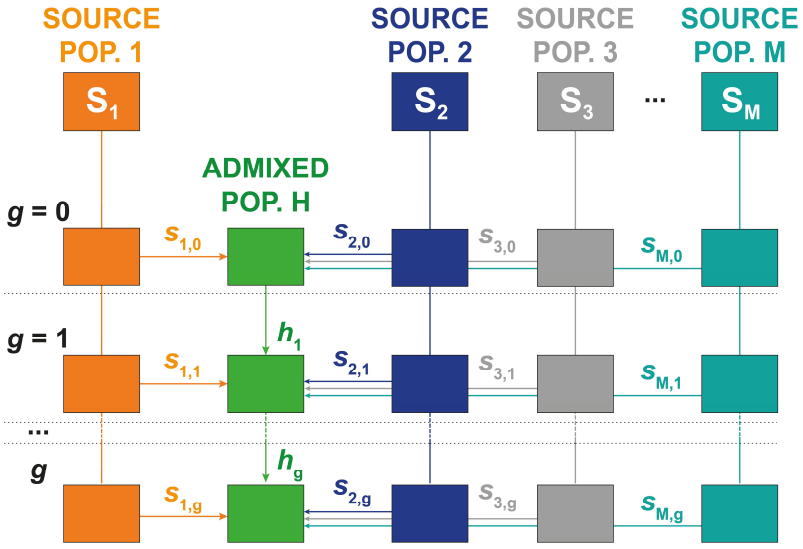
Supplementary Table ST5. Accuracy of the 95% credibility interval estimated for posterior parameters in the vicinity of the observed ACB and ASW datasets under the loosing scenario Afr2P-Eur2P. We provide the empirical coverage of the estimated 95% credibility interval, i.e. how many times (in percentage) the true parameter (θ_i) is found inside the estimated 95% credibility interval [2.5%quantile$(\hat{\theta}_i)$; 97.5%quantile$(\hat{\theta}_i)$], among the 1,000 posterior parameter estimations conducted using in turn the 1,000 simulations closest to our real data, separately for the ACB and ASW, as pseudo-observed datasets for the NN estimation of the parameters taken jointly as a vector.

As expected under the loosing scenario Afr2P-Eur2P, 95% CI are poorly estimated on average across all parameters, compared to the reasonably conservative 95% CI estimated under the winning AfrDE-EurDE scenario (**Supplementary Table ST3**).

	ACB	**ASW**
AfrDE-EurDE parameters	***NN joint***	***NN joint***
$s_{Afr,0}$	0.949	0.951
$s_{Afr,tAfr,p1}$	0.945	0.955
$s_{Afr,tAfr,p2}$	0.946	0.945
$t_{Afr,p1}$	0.893	0.905
$t_{Afr,p2}$	0.726	0.930
$s_{Eur,tEur,p1}$	0.950	0.955
$s_{Eur,tEur,p2}$	0.939	0.940
$t_{Eur,p1}$	0.846	0.851
$t_{Eur,p2}$	0.921	0.943
Average credibility interval accuracy	***0.9017***	***0.9306***

MOLECULAR ECOLOGY
RESOURCES

Supplementary Figure S1. General mechanistic model of historical admixture from Verdu and Rosenberg (2011). This general model considers, for diploid organisms, a panmictic admixture process, discrete in generations, where M source populations S_m contribute to the hybrid population H at the following generation $g + 1$ with proportions $s_{m,g}$ each in [0,1], and where the hybrid population H contributes to itself with proportion h_g in [0,1] with $h_0 = 0$, satisfying, for each value of $g \geq 0$, $\sum_{m\in[1,M]} s_{m,g} + h_g = 1$.

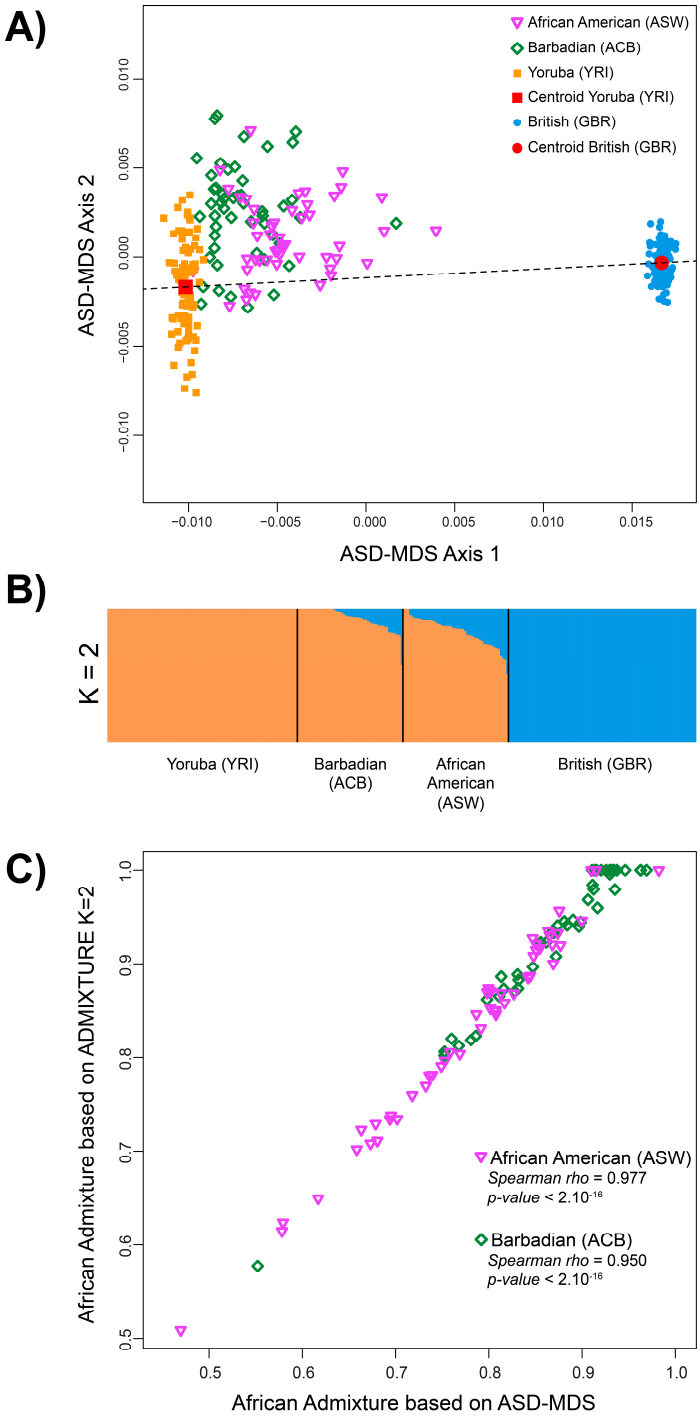


Supplementary Figure S1

MOLECULAR ECOLOGY
RESOURCES

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

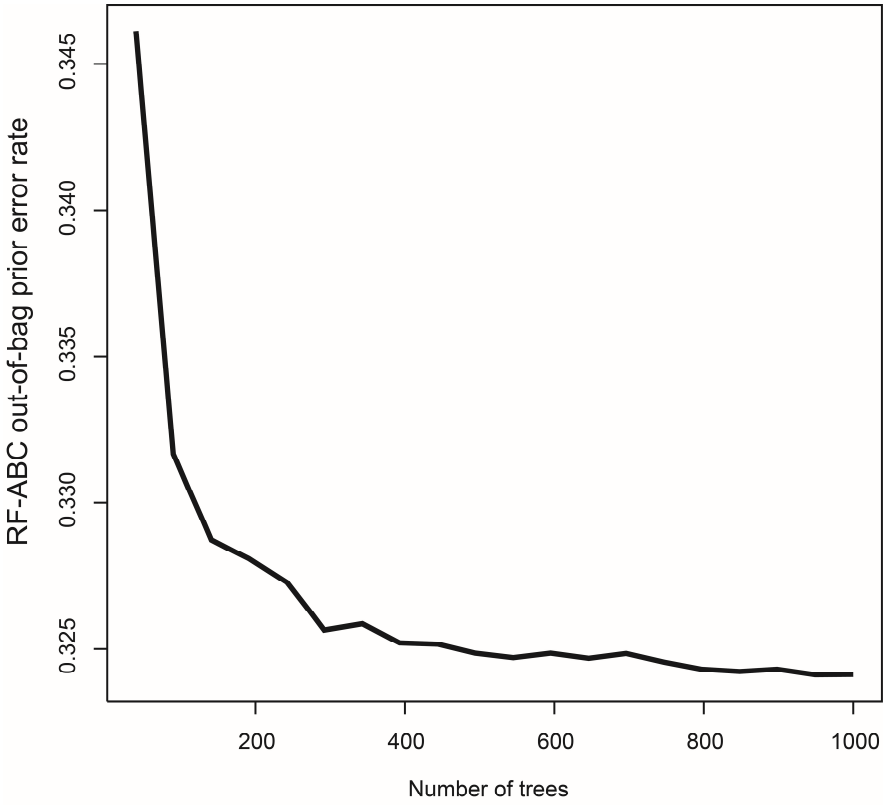
MOLECULAR ECOLOGY
RESOURCES



Supplementary Figure S2

MOLECULAR ECOLOGY
RESOURCES

Supplementary Figure S3: RF-ABC out-of-bag prior error rate as a function of the number of trees considered to build the forest for the model-choice procedure considering nine-competing scenarios (Figure 1).



Supplementary Figure S3

MOLECULAR ECOLOGY
RESOURCES

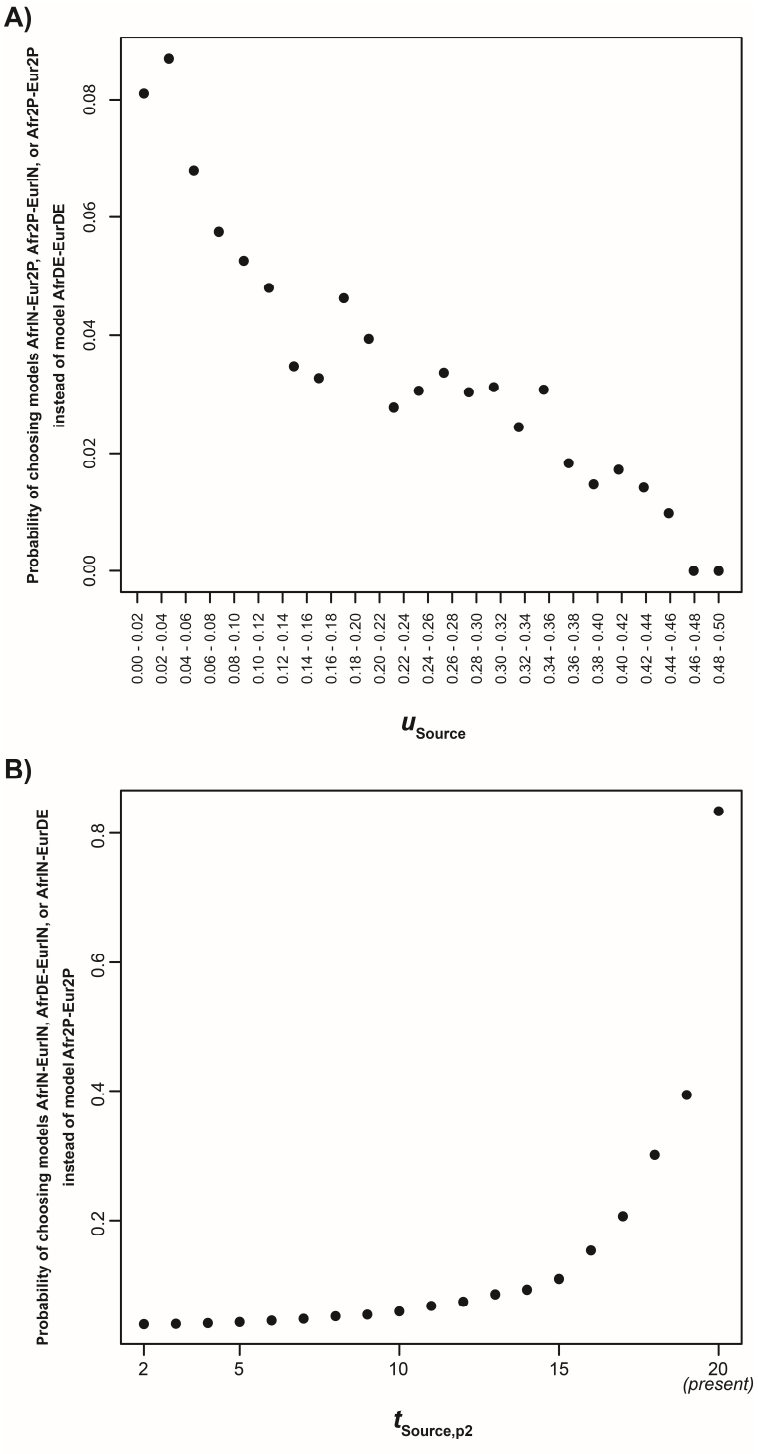
Supplementary Figure S4: Examples of Random-Forest Approximate Bayesian Computation model-choice error as a function of scenario-parameters.

Random Forest trained on 9000 simulations per nine scenarios (**Figure 1**) as a reference table. Each simulation considers 100,000 SNPs and 50 individuals sampled in population H and 90 and 89 in the African and European source respectively. 1000 additional simulations per scenario, considered in turn as pseudo-observations. RF-ABC model-choice performed in turn for each pseudo-observed target simulation using 1,000 decision trees and 24 summary-statistics (see **Materials and Methods**).

A) Probability of wrongly choosing scenarios Afrf2P-Eur2P, Afr2P-EurIN, or AfrIN-Eur2P instead of the true simulated target model AfrDE-EurDE, as a function of u_{Afr} and u_{Eur} increasing values in 2% bins. Both u_{Source} values need to be in the same bin for calculations.

B) Probability of wrongly choosing scenarios AfrIN-EurIN, AfrDE-EurIN, or AfrIN-EurDE instead of the true simulated target model Afr2P-Eur2P, as a function of the time for the second admixture pulse from either source, $t_{Afr,p2}$ and $t_{Eur,p2}$, increasing values.

MOLECULAR ECOLOGY
RESOURCES



Supplementary Figure S4

MOLECULAR ECOLOGY
RESOURCES

Supplementary Figure S5: Three Random-Forest Approximate Bayesian Computation model-choice cross-validation.

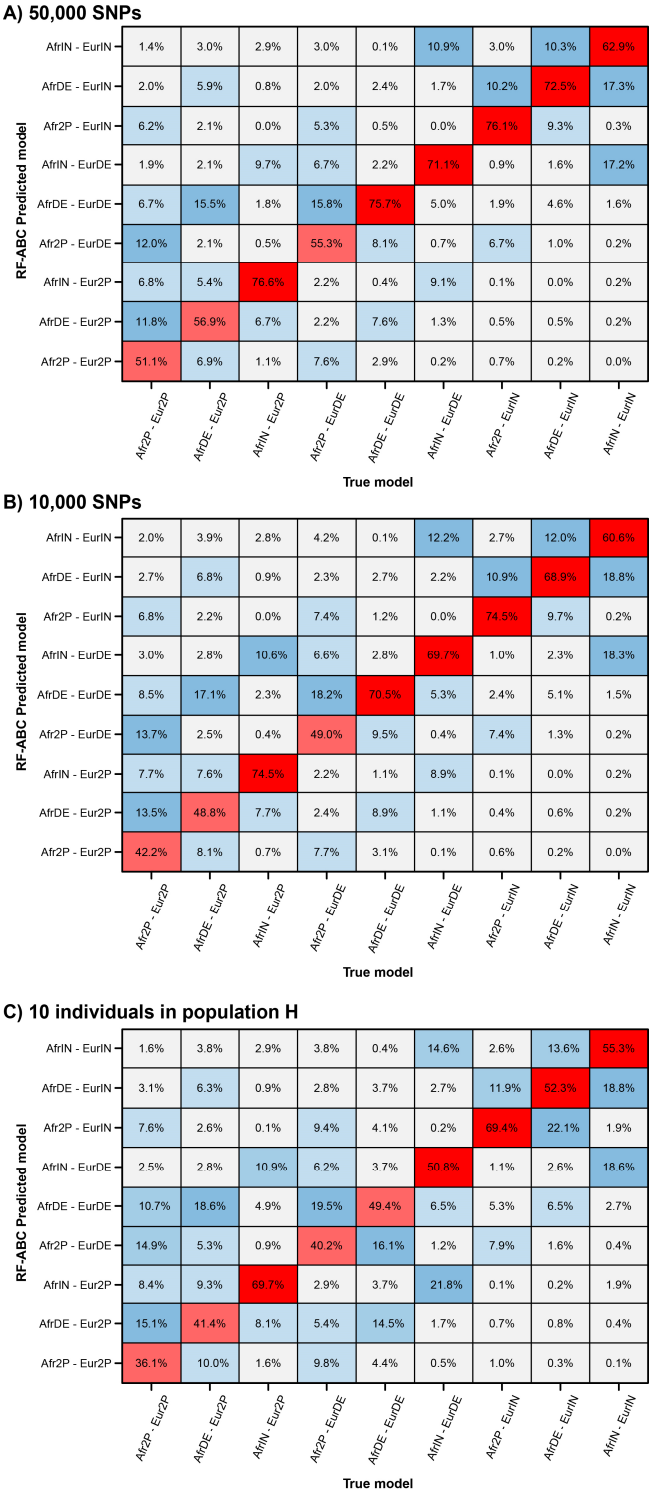
Heat map of the out-of-bag cross-validation results considering each 10,000 simulations per each nine competing models (**Figure 1**, **Table 1**) in turn as pseudo-observed target for RF-ABC model-choice. Prior probability of correctly choosing a given scenario is 11%. RF-ABC model-choice performed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

A) 50,000 SNPs simulated in the hybrid and source populations; 50 individuals sampled in population H, 90 and 89 respectively in the African and European source populations. Out-of-bag prior error rate is 33.53%.

B) 10,000 SNPs simulated in the hybrid and source populations; 50 individuals sampled in population H, 90 and 89 respectively in the African and European source populations. Out-of-bag prior error rate is 37.93%.

C) 100,000 SNPs simulated in the hybrid and source populations; 10 individuals sampled in population H, 18 and 18 respectively in the African and European source populations. Out-of-bag prior error rate is 48.39%.

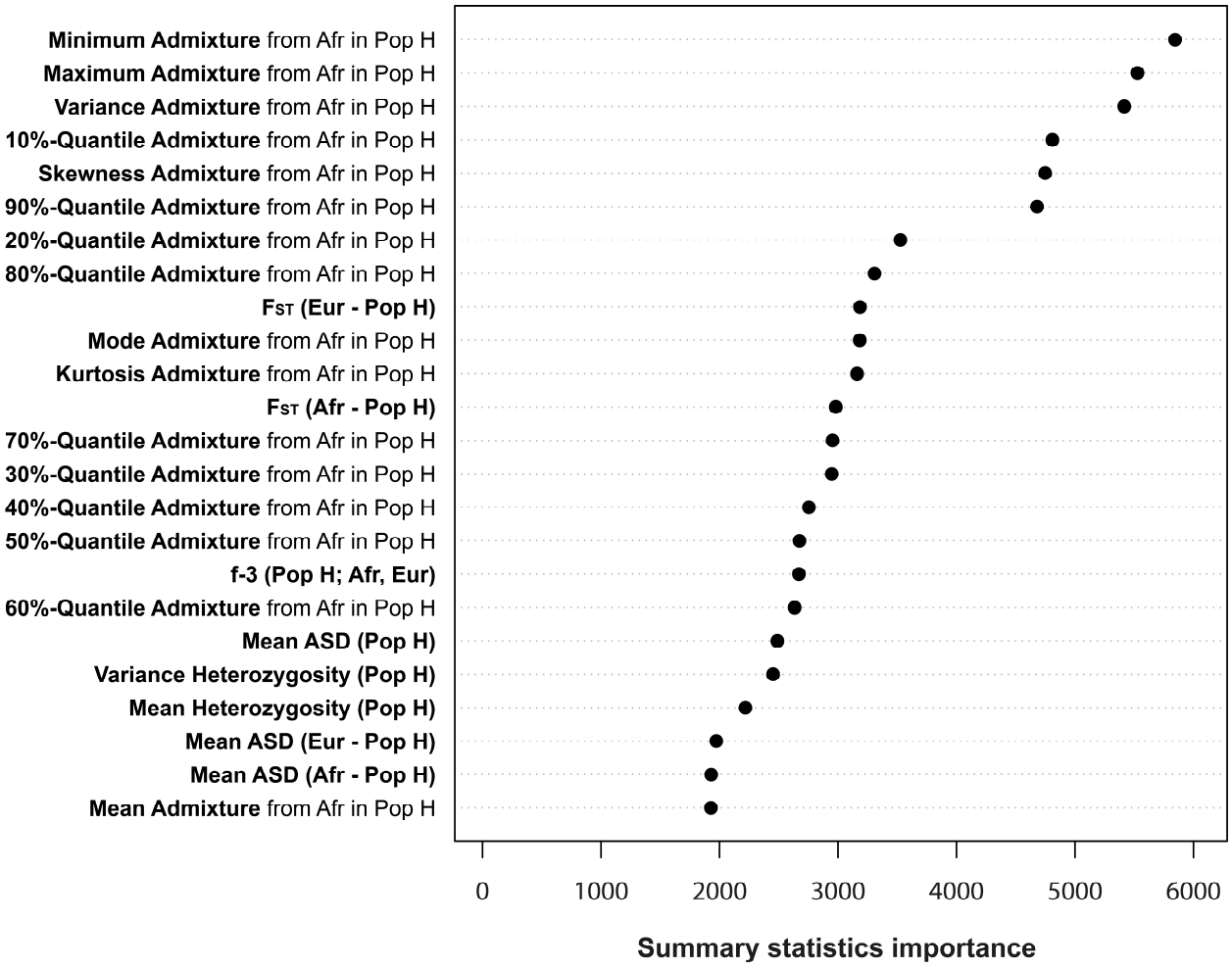
MOLECULAR ECOLOGY
RESOURCES



Supplementary Figure S5

MOLECULAR ECOLOGY
RESOURCES

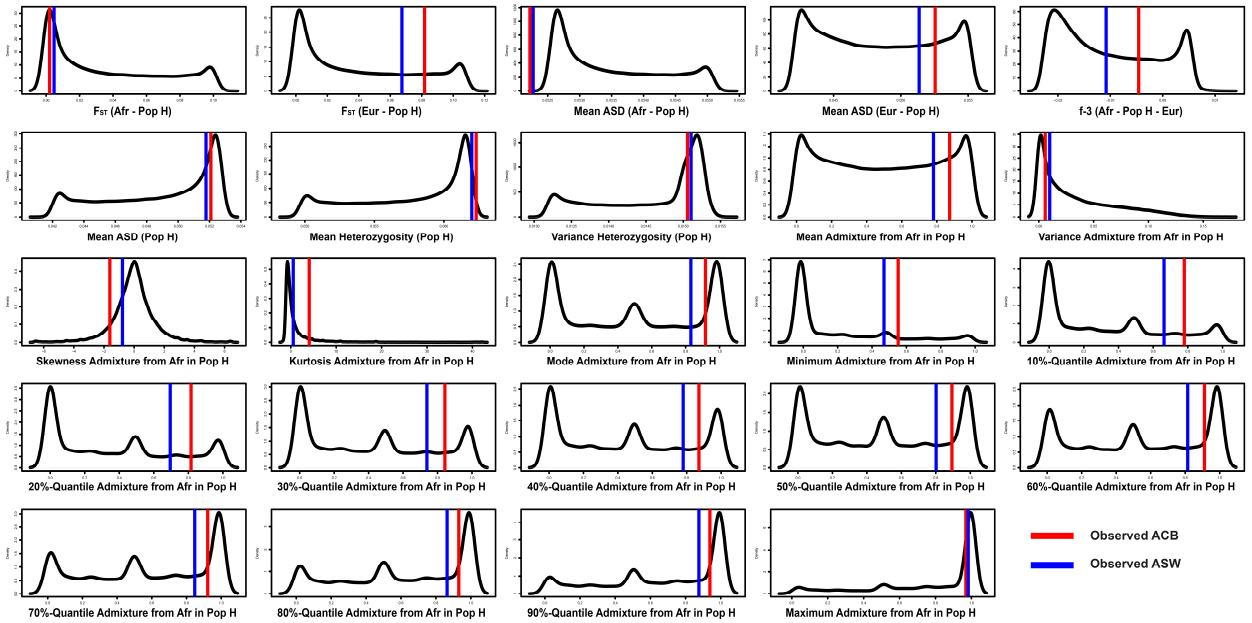
Supplementary Figure S6: Summary statistics' respective importance in the RF-ABC model-choice out-of-bag cross-validation presented in **Figure 2**.



Supplementary Figure S6

MOLECULAR ECOLOGY
RESOURCES

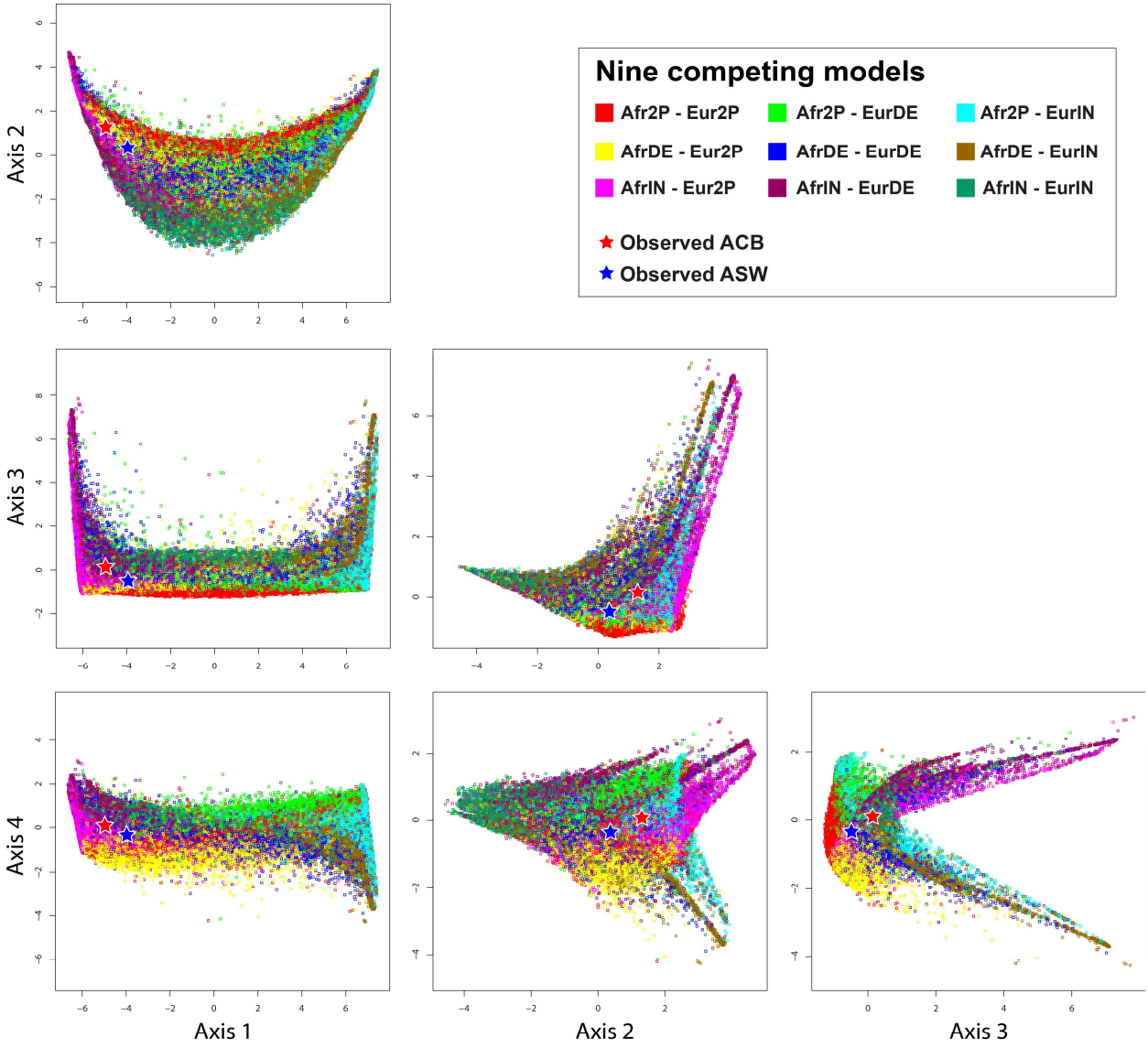
Supplementary Figure S7: Summary statistics prior-distribution densities for each nine competing models considered (Figure 1). 10,000 simulations were performed for each nine-competing scenario. Prior densities are plotted for the nine scenarios altogether. Corresponding statistics observed from the ACB and ASW population separately are represented, on each plot, by vertical lines (red and blue respectively for ACB and ASW). The 24 separate summary statistics considered are described in **Materials and Methods**.



Supplementary Figure S7

MOLECULAR ECOLOGY
RESOURCES

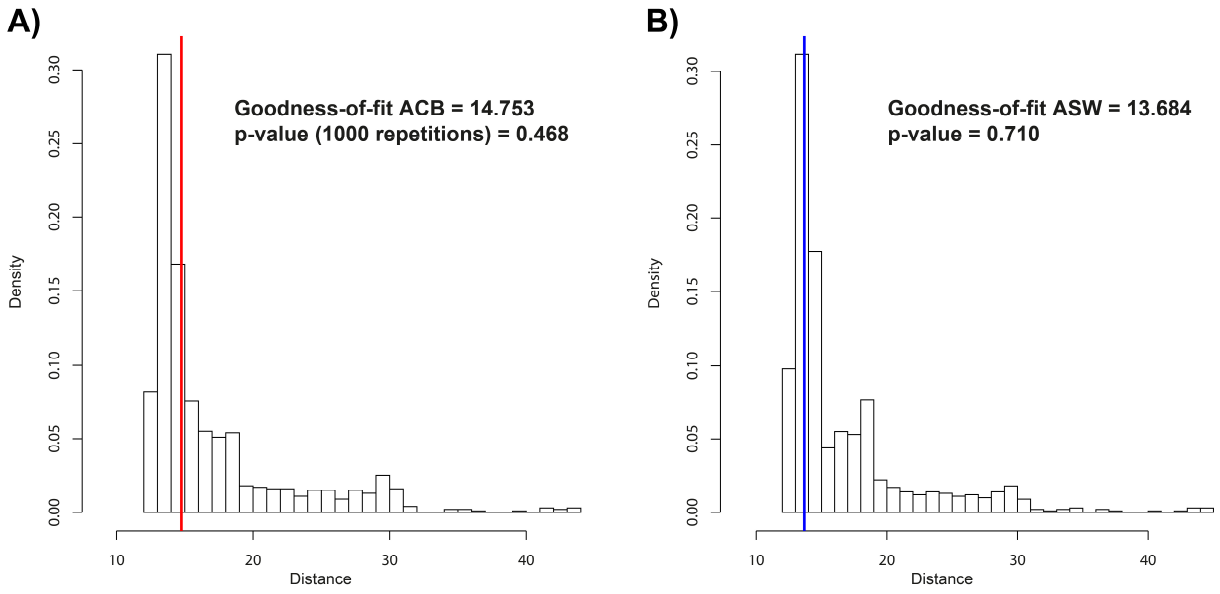
Supplementary Figure S8: Four first axes of the principal component analysis for the 90,000 sets of 24 summary statistics computed on simulated data under each nine-competing scenario (Figure 1). The 24 same statistics calculated for the observed ACB and ASW population samples, respectively, are then projected on the PCA and represented by, respectively, a red and blue star. All two-dimensional projections are orthonormal.



Supplementary Figure S8

MOLECULAR ECOLOGY
RESOURCES

Supplementary Figure S9: Histogram of the goodness-of-fit for the observed set of 24 summary statistics computed for (A) the ACB population, and (B) the ASW population, in turn serving as the observed admixed population H considering the YRI population sample as the African source and the GBR population sample as the European source (see Materials and Methods). Goodness-of-fit statistics were calculated as the mean distance between observed and accepted summary statistics. Observed statistics are fitted to the full 90,000 sets of the same statistics calculated from 10,000 simulations performed under each nine-competing models (Figure 1). Goodness-of-fit was obtained considering 1,000 repetitions and a tolerance value of 0.01.



Supplementary Figure S9