Supporting/Supplemental Information

Invited paper - Special Issue for *Molecular Ecology Resources* on Machine Learning techniques in Evolution and Ecology

Extending Approximate Bayesian Computation with Supervised Machine Learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest

Short title: DIYABC Random Forest to infer population history

François-David Collin [1], Ghislain Durif [1], Louis Raynal [1], Eric Lombaert [2], Mathieu Gautier [3], Renaud Vitalis [3], Jean-Michel Marin [1,*], Arnaud Estoup [3,*]

[1] IMAG, Univ Montpellier, CNRS, UMR 5149, Montpellier, France

[2] ISA, INRAE, CNRS, Univ Côte d'Azur, Sophia Antipolis, France

[3] CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

[*] These authors are joint senior authors on this work

Corresponding author: Arnaud Estoup. E-mail: arnaud.estoup@.inrae.fr
CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

TABLE S1. Summary statistics provided by DIYABC Random Forest v1.0 with corresponding values computed from the pseudo-observed PoolSeq and IndSed SNP datasets generated under the (admixed) scenario 3.

The two pseudo observed datasets were simulated under scenario 3 (see Figure 1) using the following parameter values: N_1=7,000, N_2=2,000, N_3=4,000, N_4=3,000, t_1=200 r_a=0.3, t_2=300 and t_3=500. The names of summary statistics are those given in the program DIYABC Random Forest v1.0 (see also the associated manual report available at https://diyabc.github.io/doc). The two pseudo observed SNP datasets were summarized using a total of 130 statistics. ML1p, ML2p and ML3p: proportion of monomorphic loci computed for each population and each pair or triplet of populations, respectively. Mean (cf. suffix m) and variance (cf. suffix v) were computed over loci for all subsequent summary statistics. HW: heterozygosity within each population. HB: heterozygosity for each pair of populations. FST: F_{ST} estimates for each population (FST1), for each pair (FST2), triplet (FST3), and quadruplet (FST4) of populations. NEI: Nei's (1972) distance for each pair of populations. AML: Cornuet et al. (2014)'s coefficient of admixture for each triplet of populations. F: allele-shared Patterson et al. (2012)'s *f*-statistics computed for each triplet (F3) and quadruplet (F4) of populations. The population index(s) are indicated at the end of each statistics and correspond to those in Figure 1. For instance ML1p_1 corresponds to the proportion of monomorphic loci in population 1 and F3m_4.1.3 to the mean F3 statistics with population 4 as target and populations 1 and 3 as external populations. The feature vector was enriched with one LDA axis or five LDA axes for scenario choice when comparing the two groups of scenarios or individual scenarios, respectively, and by 4 to 24 PLS axes for parameter estimation, depending on the estimated parameter and the analyzed training set. Five "noise variables", randomly drawn into uniform distributions bounded between 0 and 1, were also added to the feature vector in order to evaluate which summary statistics were informative in our different inferential settings, when conducting scenario choice or parameter estimation.

Statistics	Population(s)	Values for the PoolSeq pseudo-observed dataset	Values for the IndSeq pseudo-observed dataset
ML1p	1 2 3 4	0.1826 0.2678 0.2208 0.1664	0.1498 0.1746 0.1244 0.1102
ML2p	1.2 1.3 1.4 2.3 2.4 3.4	0.0746 0.0624 0.0680 0.1440 0.0866 0.1010	0.0362 0.0262 0.0332 0.0478 0.0318 0.0310
ML3p	1.2.3 1.2.4 1.3.4 2.3.4	0.0214 0.0180 0.0330 0.0592	0.0048 0.0048 0.0078 0.0104
HWm	1 2 3 4	0.20954404 0.19838350 0.20488395 0.21055354	0.28478421 0.28119158 0.28998420 0.29305474
HWv	1 2 3 4	0.03017888 0.03269557 0.03209324 0.03047639	0.03190730 0.03378459 0.02993610 0.02933304
HBm	1.2 1.3 1.4 2.3 2.4 3.4	0.22103712 0.22007464 0.21984261 0.21242519 0.21683138 0.21476721	0.30713500 0.30632800 0.30218300 0.30215000 0.30505300 0.30241900
HBv	1.2 1.3 1.4 2.3 2.4 3.4	0.03050988 0.02952776 0.02889917 0.03152973 0.03053319 0.02997418	0.02691466 0.02543388 0.02534241 0.02699442 0.02585686 0.02462178
FST1m	1 2 3 4	0.03656301 0.08787669 0.05798906 0.03192155	0.06386061 0.07567027 0.04676723 0.03667384
FST1v	1 2 3 4	0.63796837 0.69117010 0.67843706 0.64425761	0.34477777 0.36506305 0.32347805 0.31696137
FST2m	1.2 1.3 1.4 2.3 2.4 3.4	0.07728727 0.05828701 0.04442737 0.05080998 0.05682634 0.03279332	0.07862049 0.06184152 0.04389236 0.05481418 0.05877615 0.03604114

FST2v	1.2 1.3	0.00662912 0.00482556	0.01430816 0.01205242
	1.4 2.3	0.00347683 0.00415286	0.00947003 0.01095105
	2.4 3.4	0.00479387 0.00255583	0.01132813 0.00827309
NEIm	1.2 1.3	0.02526759 0.01989388	0.04891985 0.04303372
	1.4 2.3	0.01634083 0.01766116	0.03607558 0.04038162
	2.4 3.4	0.01942893 0.01285703	0.04173982 0.03331893
NEIv	1.2 1.3	0.00304633 0.00178349	0.00709198 0.00580780
	1.4 2.3	0.00123277 0.00156038	0.00414014 0.00527857
	2.4 3.4	0.00200541 0.00082593	0.00508674 0.00369040
AMLm	1.2.3 2.1.3	0.46204517 0.40770154	0.4613780 0.444626580
	3.1.2 1.2.4	0.44138604 0.41696963	0.48438619 0.42576090
	2.1.4 4.1.2	0.44406000 0.52419587	0.45916842 0.52907402
	1.3.4 3.1.4	0.43788739 0.40023811	0.45458238 0.43777103
	4.1.3 2.3.4	0.45678690 0.54884198	0.48674400 0.52262473.
	3.2.4 4.2.3	0.47752730 0.42632776	0.46453923 0.44485723
AMLv	1.2.3 2.1.3	0.20487761 0.19859487	0.20011036 0.20000220
	3.1.2 1.2.4	0.18815848 0.19462928	0.19081113 0.19336935
	2.1.4 4.1.2	0.21034224 0.18424422	0.20604561 0.18747876
	1.3.4 3.1.4	0.20335850 0.19214446	0.20465516 0.19706727
	4.1.3 2.3.4	0.18189880 0.20830083	0.18931158 0.20938216
	3.2.4 4.2.3	0.19153337 0.18909750	0.19133678 0.19324775
FST3m	1.2.3 1.2.4	0.06232238 0.05957347	0.06515089 0.06052300
	1.3.4 2.3.4	0.04527343 0.04681239	0.04732179 0.04990147
FST3v	1.2.3 1.2.4	0.00338161 0.00320667	0.00855638 0.00803107
	1.3.4 2.3.4	0.00233752 0.00255658	0.00667524 0.00679905
FST4m	1.2.3.4	0.05353053	0.05574299
FST4v	1.2.3.4	0.00222549	0.00587558
F3m	1.2.3 2.1.3	0.00957127 0.00750209	0.01326439 0.01088271
	3.1.2 1.2.4	0.00328938 0.00725216	0.00567939 0.00974039
	2.1.4 4.1.2	0.00982119 0.00254166	0.01440671 0.00352313
	1.3.4 3.1.4	0.00780301 0.00505764	0.01065389 0.00828989
	4.1.3 2.3.4	0.00199082 0.00805293	0.00260963 0.01179621
	3.2.4 4.2.3	0.00273854 0.00430993	0.00476589 0.00613363
F3v	1.2.3 2.1.3	0.00067502 0.00059355	0.00200864 0.00177915
	3.1.2 1.2.4	0.00038770 0.00054309	0.00150415 0.00152669
	2.1.4 4.1.2	0.00074646 0.00033338	0.00206510 0.00117159
	1.3.4 3.1.4	0.00044435 0.00030498	0.00150024 0.00129820
	4.1.3 2.3.4	0.00022943 0.00052647	0.00098230 0.00170687
	3.2.4 4.2.3	0.00025187 0.00032306	0.00108649 0.00114095
F4m	1.2.3.4 1.3.2.4	-0.0023191 -0.00176826	0.00352400 -0.0026105
	1.4.2.3	0.0005508	0.0009135
F4v	1.2.3.4 1.3.2.4	0.00031890 0.00036603	0.00110473 0.00131977
	1.4.2.3	0.0002816	0.0010138

TABLE S2. Results for scenario choice under the (non-admixed) scenario 6.

The six compared scenarios and the two groups of scenarios are detailed in Figure 1. Results are given for the two example pseudo-observed datasets (PoolSeq and IndSeq) which were simulated under the scenario 6 using the following parameter values: N_1=7,000, N_2=2,000, N_3=4,000, N_4=3000, t_1=200, t_2=300 and t_3=500. RF analyses used a training set including feature vector values from 12,000 simulated datasets (2,000 per scenario) and the number of trees was 1,000. Global (prior) and local (posterior) error rates were estimated using out-of-bag estimators from a sample of 10,000 data randomly chosen in the training set. Standard deviations over the ten replicate analyses are given between brackets for each metrics, in addition to the means. In the "RF with LDA" treatments, five LDA axes were added to the set of 130 summary statistics composing the feature vector. ABC rejection or ABC mnlog: inference methods based on a simple rejection or a multinomial regression algorithm (using the R package abc v2.1; Csilléry, François, & Blum 2012). NC: not computable. Similar results were obtained for the pseudo-observed datasets generated under the (non-admixed) scenario 6 than for those generated under the (admixed) scenario 3 (Table 2). The only discrepancy is that, in contrast to pseudo-observed datasets generated under scenario 6, the posterior probabilities of the selected scenario were higher (and hence local error rate higher) when excluding LDA axes. Note that the true/expected posterior probabilities value are unknown in these case studies.

Type of dataset	Type of treatment		Global error rate	Local error rate	Vote scen. 1	Vote scen. 2	Vote scen. 6	Vote scen. 4	Vote scen. 5	Vote scen. 6	Posterior probability
PoolSeq	Groups of scenarios: with vs. without admixture	RF with LDA	0.176 (0.005)	0.224 (0.022)		218.2 (14.920)			781.8 (14.920)		0.776 [group 2] (0.022)
		RF without LDA	0.187 (0.004)	0.172 (0.008)		244.0 (14.008)			756.0 (14.008)		0.828 [group 2] (0.009)
		ABC rejection	0.266	NC		NC			NC		0.524 [group 2]
		ABC mnlog	0.202	NC		NC			NC		0.999 [group 2]
	All scenarios considered separately	RF with LDA	0.191 (0.004)	0.117 (0.009)	135.6 (17.037)	1.6 (1.265)	21.6 (6.415)	0.1 (0.316)	3.8 (2.394)	837.3 (20.844)	0.883 [scen. 6] (0.009)
		RF without LDA	0.216 (0.005)	0.112 (0.005)	137.2 (8.456)	2.6 (1.506)	33.5 (8.708)	0.3 (0.675)	6.4 (2.675)	820.0 (11.832)	0.888 [scen. 6] (0.005)
		ABC rejection	0.372	NC	NC	NC	NC	NC	NC	NC	0.441 [scen. 6]
		ABC mnlog	0.261	NC	NC	NC	NC	NC	NC	NC	0.999 [scen. 6]
IndSeq	Groups of scenarios: with	RF with LDA	0.206 (0.004)	0.193 (0.016)		346.7 (11.196)			653.3 (11.196)		0.807 [group 2] (0.018)

vs. without admixture	RF without LDA	0.215 (0.002)	0.194 (0.016)		276.2 (16.956)			723.8 (16.956)		0.806 [group 2] (0.016)
	ABC rejection	0.344	NC		NC			NC		0.455 [group 2]
	ABC mnlog	0.261	NC		NC			NC		1.000 [scen. 6]
All scenarios considered separately	RF with LDA	0.240 (0.005)	0.169 (15.515)	180.5 (1.792)	3.1 (1.792)	105.8 (7.757)	0.4 (0.516)	11.2 (2.658)	699 (15.677)	0.831 [scen. 6] (0.023)
	RF without LDA	0.252 (0.001)	0.155 (0.0197)	165.7 (16.351)	1.5 (1.179)	56.1 (9.386)	0.1 (0.316)	11.5 (2.224)	765.1 (20.311)	0.845 [scen. 6] (0.020)
	ABC rejection	0.473	NC	NC	NC	NC	NC	NA	NC	0.362 [scen. 6]
	ABC mnlog	0.332	NC	NC	NC	NC	NC	NA	NC	0.995 [scen. 6]

TABLE S3. Results for estimation of parameters of interest under the (non-admixed) scenario 6.

Results are given for two example pseudo-observed datasets (PoolSeq and IndSeq) which were simulated under the scenario 6 using the following parameter values: t_1 = 200, N_4 = 3,000 and t_1/N_4 = 0.067. RF analyses used a training set including feature vector values from 10,000 simulated datasets and the number of trees was 1,000. Global (prior) and local (posterior) NMAE values were estimated using out-of-bag estimators from a sample of 10,000 data randomly chosen in the training set. Standard deviations over the ten replicate analyses are given between brackets for each metrics, in addition to the means. In the "RF with PLS" treatments, the number of PLS axes which were added to the set of 130 summary statistics of the feature vector for the PoolSeq (IndSeq) datasets was equal to 13 (18), 17 (16), and 4 (4) for t_1, N_4 and t_1/N_4, respectively. CI: credibility interval. 90% coverage: proportion of test parameter values comprise between the estimated 5% and the 95% quantiles. ABC rejection or ABC logRidge: inference method based on a simple rejection or a regression with a Ridge regulation algorithm (using the R package abc v2.1; Csilléry, François, & Blum 2012). NC: not computable. Note the particularly narrow 90% coverage values obtained when using ABC logRidge.

Type of dataset	Type of treatment	Parameter	Posterior point estimates of			Global (prior) NMAE computed from		Local (posterior) NMAE computed from		90% Coverage
			Mean	Median	90% CI	Mean	Median	Mean	Median	
PoolSeq	RF with PLS	t_1	289.8 (2.344)	282.0 (1.826)	191.0 – 412.2 (5.925) - (4.454)	0.223 (0.0002)	0.211 (0.0001)	0.138 (0.0045)	0.137 (0.0043)	0.962 (0.0007)
		N_4	5101 (35.82)	4914 (64.71)	2759 - 8190 (79.22) - (52.60)	0.283 (0.0003)	0.262 (0.0003)	0.258 (0.0099)	0.247 (0.0090)	0.943 (0.0008)
		t_1/N_4	0.060 (0.0001)	0.060 (0.0002)	0.052 - 0.068 (0.0001) (0.0005)	0.114 (0.0003)	0.109 (0.0002)	0.053 (0.0012)	0.054 (0.0012)	0.969 (0.0004)
	RF without PLS	t_1	295.1 (2.092)	291.0 (4.807)	186.4 - 419.5 (5.621) - (6.972)	0.226 (0.0003)	0.215 0.0002)	0.140 (0.0032)	0.140 (0.0037)	0.962 (0.0008)
		N_4	5437 (82.42)	5320 (108.82)	2840 - 8408 (59.46) (91.88)	0.286 (0.0003)	0.267 (0.0004)	0.258 (0.0107))	0.251 (0.0096)	0.942 (0.0009)
		t_1/N_4	0.060 (0.0004)	0.060 (0.0006)	0.053 - 0.068 (0.0008) (0.0006)	0.115 (0.0009)	0.110 (0.0003)	0.053 (0.0021)	0.054 (0.0017)	0.969 (0.0005)
	ABC	t_1	292.4	279.5	102.8- 534.3	0.848	0.707	NC	NC	0.956

	rejection	N_4	6166	6317	2038 - 9576	0.766	0.699	NC	NC	0.939
		t_1/N_4	0.051	0.049	0.019 - 0.093	0.959	0.717	NC	NC	0.975
	ABC logRidge	t_1	295.4	295.7	208.5 – 371.6	0.414	0.252	NC	NC	0.686
		N_4	5127	5125	3816 - 6531	0.325	0.320	NC	NC	0.659
		t_1/N_4	0.065	0.065	0.062 - 0.068	0.121	0.119	NC	NC	0.674
IndSeq	RF with PLS	t_1	370.8 (3.219)	384.7 (3.093)	215.7 – 491.2 (5.805) (5.922)	0.283 (0.0003)	0.266 (0.0003)	0.185 (0.0076)	0.184 (0.0078)	0.951 (0.0006)
		N_4	6835 (64.87)	6884 (41.56)	2258 - 9301 (260.7) - (73.7)	0.325 (0.0004)	0.307 (0.0002)	0.199 (0.0077)	0.201 (0.0079)	0.932 (0.0013)
		t_1/N_4	0.062 (0.0003)	0.062 (0.0003)	0.052 - 0.070 (0.0002) (0.0002)	0.160 (0.0002)	0.152 (0.0002)	0.063 (0.0014)	0.0631 (0.0015)	0.946 (0.0007)
	RF without PLS	t_1	383.0 (2.702)	394.6 (2.989)	223.9 – 503.4 (4.677) - (9.919)	0.285 (0.0002)	0.269 (0.0003)	0.179 (0.0078)	0.179 (0.0081)	0.950 (0.0009)
		N_4	6824 (64.87)	6881 (41.56)	2182 - 9318 (260.7) - (73.7)	0.325 (0.0004)	0.308 (0.0002)	0.199 (0.0077)	0.201 (0.0079)	0.932 (0.0007)
		t_1/N_4	0.062 (0.0002)	0.063 (0.0004)	0.052 - 0.070 (0.0003) (0.0002)	0.161 (0.0002)	0.153 (0.0001)	0.063 (0.0017)	0.063 (0.0016)	0.946 (0.0007)
	ABC rejection	t_1	321.5	302.0	102.9 – 610.1	1.016	0.851	NC	NC	0.932
		N_4	6208	6274	2417 - 9550	0.874	0.782	NC	NC	0.891
		t_1/N_4	0.055	0.053	0.021 – 0.094	1.128	0.863	NC	NC	0.964
	ABC logRidge	t_1	295.3	295.4	234 – 348.6	0.356	0.356	NC	NC	0.673
		N_4	5147	5297	2287 - 7256	0.467	0.467	NC	NC	0.684
		t_1/N_4	0.064	0.064	0.058 – 0.071	0.199	0.199	NC	NC	0.700

TABLE S4. Results for scenario choice: ABC Random Forest (RF) versus traditional ABC methods

The six compared scenarios and the two groups of scenarios are detailed in Figure 1. Results are given for the two example pseudo-observed datasets (PoolSeq and IndSeq) which were simulated under the (admixed) scenario 3 using the following parameter values: N_1=7,000, N_2=2,000, N_3=4,000, N_4=3000, t_1=200, r_a=0.3, t_2=300 and t_3=500. In the "RF with LDA" treatments, five LDA axes were added to the set of 130 summary statistics composing the feature vector. Standard deviations over the ten replicate analyses are given between brackets for each metrics, in addition to the means. Traditional ABC methods are ABC rejection or ABC mnlog and correspond to inference based on a simple rejection or a multinomial regression algorithm (using the R package abc v2.1; Csilléry, François, & Blum 2012). NC: not computable. The global error rate of the selected admixture group of scenarios was notably high (and the posterior probabilities low) with the ABC rejection method. For the ABC multinomial logistic method, global prior error rates were higher than for ABC Random Forest and the posterior probabilities of the best scenario were always equal to 1.000 for the pseudo-observed datasets.

Type of dataset	Type of treatment		Global error rate	Local error rate	Posterior probability
PoolSeq	Groups of scenarios: with vs. without admixture	RF with LDA	0.172 (0.001)	0.085 (0.009)	0.915 [group 1] (0.009)
		ABC rejection	0.342	NC	0.616 [group 1]
		ABC mnlog	0.212	NC	1.000 [group 1]
	All scenarios considered separately	RF with LDA	0.196 (0.0008)	0.135 (0.011)	0.865 [scen. 3] (0.011)
		ABC rejection	0.457	NC	0.333 [scen 3]
		ABC mnlog	0.271	NC	1.000 [scen 3]
IndSeq	Groups of scenarios: with vs. without admixture	RF with LDA	0.212 (0.001)	0.177 (0.016)	0.823 [group 1] (0.016)
		ABC rejection	0.351	NC	0.633 [group 1]
		ABC mnlog	0.263	NC	1.000 [group 1]
	All scenarios considered separately	RF with LDA	0.248 (0.001)	0.268 (0.018)	0.732 [scen. 3] (0 .018)
		ABC rejection	0.473	NC	0.371 [scen 3]
		ABC mnlog	0.330	NC	1.000 [scen 3]

TABLE S5. Results for estimation of parameters of interest: ABC Random Forest (RF) versus traditional ABC methods

Results are given for the two example pseudo-observed datasets (PoolSeq and IndSeq) which were simulated under the (admixed) scenario 3 using the following parameter values: r_a = 0.3, t_1 = 200, N_4 = 3,000 and t_1/N_4 = 0.067. In the "RF with PLS" treatments, the number of PLS axes which were added to the set of 130 summary statistics of the feature vector for the PoolSeq (IndSeq) datasets was equal to 12 (12), 18 (21), 23 (24), and 4 (4) for r_a, t_1, N_4 and t_1/N_4, respectively. 90% coverage: proportion of test parameter values comprise between the estimated 5% and the 95% quantile. CI: credibility interval. Standard deviations over the ten replicate analyses are given between brackets for each metrics, in addition to the means. Traditional ABC methods are ABC rejection or ABC mnlog and correspond to inference based on a simple rejection or a regression with a Ridge regulation algorithm (using the R package abc v2.1; Csilléry, François, & Blum 2012). NC: not computable. NMAE values with the ABC logRidge method were low (i.e. similar to those for Random Forest), but ABC logRidge was characterized by particularly narrow 90% coverage values (i.e. around 0.70), indicating that confidence intervals are prejudicially too narrow (i.e. the true parameter values is often outside the limits of the confidence interval) with this method, a feature previously noted by Raynal et al. 2019.

Type of dataset	Type of treatment	Parameter	Posterior point estimates of			Global (prior) NMAE computed from		Local (posterior) NMAE computed from		90% Coverage
			Mean	Median	90% CI	Mean	Median	Mean	Median	
PoolSeq	RF with PLS	r_a	0.346 (0.0018)	0.352 (0.0030)	0.248 - 0.422 (0.0041) (0.0040)	0.133 (0.0002)	0.123 (0.0002)	0.089 (0.0028)	0.089 (0.0024)	0.974 (0.0008)
		t_1	291.4 (3.366)	300.5 (2.273)	147.6 - 441.0 (3.777) - (3.887)	0.312 (0.0003)	0.290 (0.0003)	0.202 (0.0047)	0.200 (0.0045)	0.960 (0.0009)
		N_4	4040 (37.16)	3658 (58.55)	1861 - 7399 (90.42) - (161.6)	0.416 (0.0005)	0.380 (0.0006)	0.317 (0.0094)	0.285 (0.0093)	0.939 (0.0007)
		t_1/N_4	0.067 (0.0004)	0.068 (0.0005)	0.049 - 0.084 (0.0010) (0.0006)	0.217 (0.0008)	0.178 (0.0002)	0.079 (0.0020)	0.077 (0.0016)	0.979 (0.0004)
	ABC rejection	r_a	0.449	0.439	0.130 - 0.822	0.572	0.524	NC	NC	0.947
		t_1	304.3	290.0	111.9 – 543.0	1.102	0.918	NC	NC	0.934
		N_4	5940	6100	1805 - 9701	0.890	0.793	NC	NC	0.907
		t_1/N_4	0.058	0.055	0.051 – 0.104	1.450	0.994	NC	NC	0.961

	ABC logRidge	r_a	0.269	0.269	0.265 - 0.273	0.163	0.159	NC	NC	0.676
		t_1	298.1	299.4	250.6 – 334.2	0.294	0.273	NC	NC	0.670
		N_4	4612	4703	3155 - 5726	0.383	0.383	NC	NC	0.694
		t_1/N_4	0.073	0.073	0.069 – 0.075	0.203	0.205	NC	NC	0.702
IndSeq	RF with PLS	r_a	0.402 (0.0041)	0.391 (0.0040)	0.275 - 0.611 (0.0041) (0.0096	0.172 (0.0003)	0.154 (0.0003)	0.161 (0.0021)	0.150 (0.0020)	0.963 (0.0011)
		t_1	400.5 (3.133)	395.6 (2.875)	231.5 - 574.1 (4.478) (11.083)	0.398 (0.0006)	0.357 (0.0006)	0.179 (0.0056)	0.179 (0.0051)	0.957 (0.0008)
		N_4	6608 (53.15)	6796 (55.61)	2861 - 9513 (111.6) (148.7)	0.476 (0.0006)	0.442 (0.0007)	0.249 (0.0117)	0.249 (0.0105)	0.927 (0.0008)
		t_1/N_4	0.061 (0.0004)	0.061 (0.0004)	0.044 - 0.077 (0.0006) (0.0009)	0.262 (0.0009)	0.220 (0.0006)	0.091 (0.0025)	0.090 (0.0025)	0.975 (0.0007)
	ABC rejection	r_a	0.450	0.442	0.126 - 0.802	0.513	0.472	NC	NC	0.932
		t_1	321.7	303.0	96.90 - 625.0	1.148	0.968	NC	NC	0.942
		N_4	6175	6459	2138 - 9587	0.929	0.856	NC	NC	0.907
		t_1/N_4	0.0555	0.051	0.022 - 0.097	1.319	0.993	NC	NC	0.962
	ABC logRidge	r_a	0.374	0.345	0.071 - 0.884	0.190	0.184	NC	NC	0.690
		t_1	336.6	334.1	247.5 - 424.2	0.428	0.425	NC	NC	0.656
		N_4	5283	5276	4105 - 6509	0.464	0.462	NC	NC	0.703
		t_1/N_4	0.067	0.067	0.061 - 0.074	0.241	0.238	NC	NC	0.697

FIGURE S1. Evolution of prediction power relatively to the number of trees in the forest when analyzing PoolSeq data.

Prediction power was evaluated by computing the global (prior) error rate for scenario choice and the global (prior) mean squared error with the mean taken as point estimate, for parameter estimation. The feature vectors included LDA or PLS axes, and five noise variables. For the presented analyses (and all others), the gain of increasing the number of trees becomes limited for a number of trees > 900; hence our final choice of building forests from 1,000 trees. Similar results were obtained for analyses of the IndSeq dataset (results not shown).

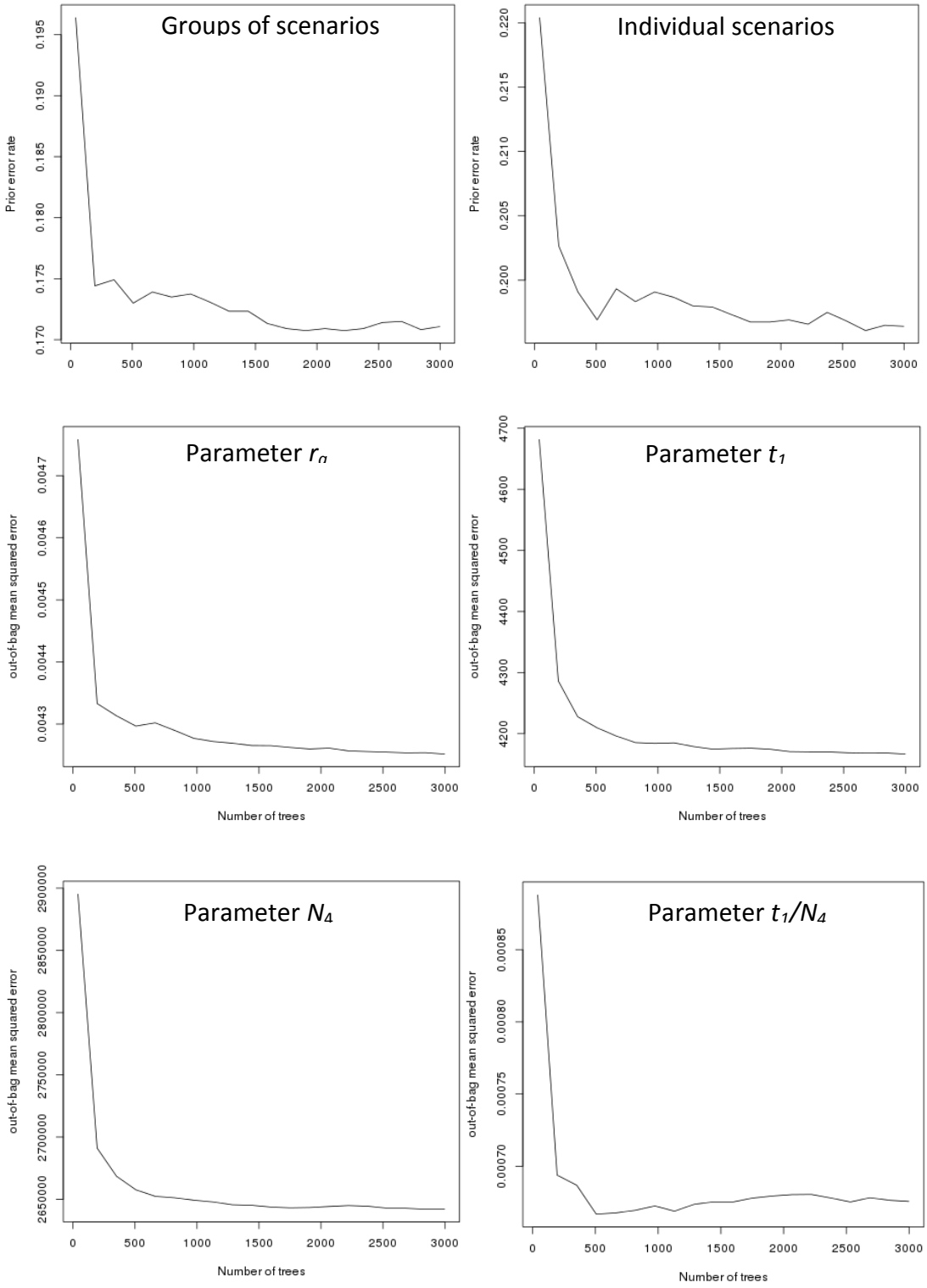
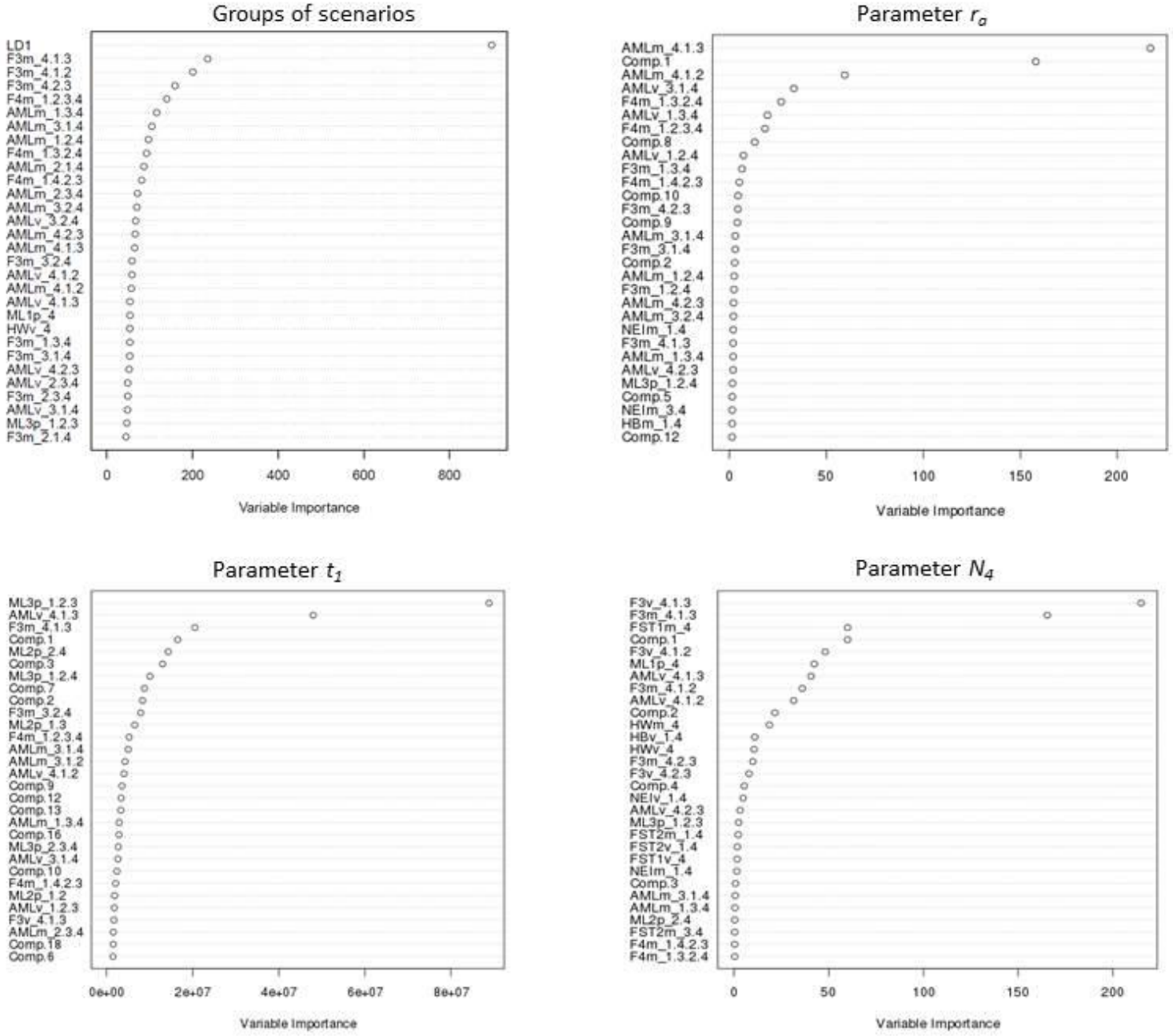


FIGURE S2. Contributions for the PoolSeq data analyses of the 30 most informative statistics of the feature vector to the Random Forest when choosing among the two groups of scenarios and when estimating the parameters r_a, t_1 and N_4 under scenario 3.

The variable importance of each statistics is computed as the mean decrease of impurity across the trees, where the impurity measure is the Gini index, and the residual sum of squares for scenario choice and parameter inference, respectively. It was computed for each of the 130 summary statistics provided by DIYABC, plus the LDA axes for scenario choice (denoted LD) or the PLS axes for parameter estimation (denoted Comp.) that were added to the feature vector. The higher the variable importance the more informative is the statistic. Population index(s) are indicated at the end of each statistics and correspond to those in Figure 1. More details about summary statistics can be found in Table S1. See Figure 3 for an illustration of the contributions of the most informative statistics when choosing among the scenarios separately and when estimating the parameter t_1/N_4.



Appendix S1: Supplementary information about the main technical features of the package DIYABC Random Forest v1.0

Implementation

The package DIYABC Random Forest v1.0 is composed of three parts: the dataset simulator, the Random Forest inference engine and the graphical user interface. The whole is packaged as a standalone and user-friendly application available at https://diyabc.github.io. The different developer and user manuals for each component of the package are available on the same site. DIYABC Random Forest v1.0 is a multithreaded program which runs on three operating systems: GNU/Linux, Microsoft Windows and MacOS. Computational procedures of the simulator and the Random Forest inference engine are written in C++.

For the Random Forest part of the package, we used our own version of the core RF (written in C++) from the package ranger (Wright & Ziegler 2017). In this new version, that we named abcranger, the Random Forest computations are optimized in order to grow a limited batch of trees in memory (but still computed in parallel to leverage multicore architectures) in sequential – i.e. batch-wise order. Tree growing and predictions are computed in a single pass, predictions are stored or accumulated and each tree is then discarded. Although we still need the entire training set at once, processing in this way avoids the in-memory storage of the whole forest at zero performance cost. The abcranger package hence opens new perspective to efficiently compute RF from training sets of (very) large size. For instance, a training set including > 100,000 particles of a feature vectors composed of > 10,000 summary statistics could be treated without any memory overflow (results not shown). It is worth stressing that abcranger is not limited to population genetics applications as the program can be used as an inference engine independently from the DIYABC simulator. However, for the moment, the binary standalone used by the DIYABC interface handles only outputs produced by the DIYABC simulator. A python wrapper (and example notebooks) is available at https://github.com/diyabc/abcranger and an R wrapper will be soon provided at the same site.

Interface

DIYABC Random Forest v1.0 can be used through a modern and user-friendly graphical interface designed as an R shiny application (Chang, Cheng Allaire, Xie, & McPherson, 2019). For a fluid and simplified user experience, this interface is available through a standalone application, which does not require installing R or any dependencies and hence can be used independently. The application is also implemented in an R package providing a standard shiny web application (with the same graphical interface) that can be run locally as any shiny application, or hosted as a web service to provide a DIYABC Random Forest v1.0 server for multiple users.

The main pipeline of the interface is divided into two modules corresponding to the two phases of a statistical treatment based on DIYABC Random Forest v1.0: module 1 = "Training set simulation" and module 2 = "Random Forest analyses". In module 1, users specify what type and how simulated data will be generated under the ABC framework to produce a training set. Module 2 guides users through scenario choice and parameter inference by providing a simple interface for the supervised learning framework based on Random Forest methodologies. An additional module named "Synthetic data file generation" (based on the DIYABC simulation engine) is also available in the application. It can be used to easily generate datafile(s) for various types of genetic markers corresponding to synthetic "ground truth" raw data (not summarized through statistics) under a given historical scenario and a set of fixed parameter values. The formats of the generated datafiles are similar to those of the observed input datafiles read by DIYABC Random Forest v1.0 (for details see user manual at https://diyabc.github.io/doc/).

Outputs

The integration of various graphical outputs (historical scenario representation, error or accuracy metrics, posterior curves, contribution to inferences of components of the feature vector, etc.) is managed with the ggplot2 R package (Wickham 2016), allowing users to create and export high-quality graphics related to the analyses. We encourage users to consult the user manual of the program available at

https://diyabc.github.io/doc/ for details regarding the various numerical and graphical outputs provided by DIYABC Random Forest v1.0. It is worth noting that a number of such outputs have been used in the present paper to illustrate the results obtained when analyzing SNP pseudo-observed or real datasets.

Memory space and computing time

All analyses carried out in the present paper were processed on a 16 cores Intel Xeon E5-2650 computer (Linux Debian platform, 64 bits system), with a maximum of 26 Gb and 1.8 Gb of RAM used for the heaviest treatments regarding the simulation of the training set (with a loop-size of 50 datasets corresponding to the number of simulated datasets distributed over all computer threads) as well as for RF analyses. Optimizing computer code procedures to efficiently compute summary statistics is important especially in the case of high-dimensional analyses which may include several thousand summary statistics. Substantial efforts in this direction on DIYABC Random Forest v1.0 allowed to considerably reduce (compared to the simulation module of DIYABC v2.1.0) both the fraction of the running time and the memory space devoted to the computation of summary statistics. Such optimizations open new perspectives for the analysis of (very) high-dimensional datasets in population genetics. Regarding the pseudo-observed datasets used as illustration, the production of a training set including 10,000 simulated datasets took 13 min (respectively 26 h) with only 4% (respectively 10%) of the running time devoted to the computation of the 130 summary statistics for the IndSeq (respectively PoolSeq) data. Note that the computation time difference between IndSeq and PoolSeq reflects the ten time larger number of individuals included in the PoolSeq simulation setting. RF treatments following the generation of the training set took less than 30 sec for scenario choice and 1 min for each parameter estimation, with 37% of the time used to compute local NMAE accuracy measures estimated using out-of-bag estimators from a sample of 10,000 data randomly chosen in the training set.

References cited

Chang, W., Cheng, J., Allaire, JJ, Xie, Y., & McPherson J. (2019). Shiny: Web Application Framework for R. R package version 1.4.0. https://CRAN.R-project.org/package=shiny

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-0-387-98141-3

Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77(1)*, 1-17. https ://doi.10.18637/jss.v077.i01

Appendix S2: Supplementary information about the illustration using a real IndSeq SNP dataset of Human populations

Motivation and background

We analyzed an IndSeq SNP dataset obtained from individuals originating from four Human populations (30 unrelated individuals per population) using the freely accessible public 1000 Genome databases (i.e. the vcf format files including variant calls available at http://www.1000genomes.org/data; The 1000 Genome Project Consortium, 2012). The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied by sequencing many individuals lightly (i.e. at a 4X coverage). A major interest of using SNP data from this source is that they do not suffer from any ascertainment bias (i.e. the deviations from expected theoretical results due to the SNP discovery process in which a small number of individuals from selected populations are used as discovery panel), which is a prerequisite when using the population genetic simulator implemented in DIYABC Random Forest v1.0. The four Human populations included the Yoruba population (Nigeria) as representative of Africa (encoded YRI in the 1000 genome database), the Han Chinese population (China) as representative of the East Asia (encoded CHB), the British population (England and Scotland) as representative of Europe (encoded GBR), and the population composed of Americans of African Ancestry in SW-USA (encoded ASW). The SNP loci were selected from the 22 autosomal chromosomes using the following criteria: (i) all 30x4 analyzed individuals have a genotype characterized by a quality score (GQ)>10 (on a PHRED scale), (ii) polymorphism is present in at least one of the 30x4 individuals in order to fit the SNP simulation algorithm used in DIYABC Random Forest v1.0, (iii) the minimum distance between two consecutive SNPs is 1 kb in order to minimize linkage disequilibrium between SNP, and (iv) SNP loci showing significant deviation from Hardy-Weinberg equilibrium at a 1% threshold (Wigginton, Cutler & Abecasis 2005) in at least one of the four populations has been removed (35 SNP loci concerned). After applying the above criteria, we obtained a dataset including 51,250 SNP loci scattered over the 22 autosomes (with a median distance between two consecutive SNPs equal to 7 kb) among which a subset of 5,000 SNP loci with a MAF > 1% were randomly chosen for applying our ABC random forest algorithms.

In this application, we compared six scenarios (i.e. models) of evolution of the four Human populations which differ from each other by one ancient and one recent historical event: (i) A single out-of-Africa colonization event giving an ancestral out-of-Africa population which secondarily split into one European and one East Asian populational lineage, versus two independent out-of-Africa colonization events, one giving the European lineage and the other one giving the East Asian lineage. The possibility of a second ancient (i.e. >100,000 years) out of Africa colonization event through the Arabian peninsula toward Southern Asia has been suggested by archaeological studies (e.g. Rose et al. 2011). (ii) **The possibility (or not) of a recent genetic admixture of the Americans of African Ancestry in SW-USA between their African ancestors and individuals of European or East Asia origins.**

The six different scenarios as well as the prior distributions of the time event and effective population size parameters used to simulate SNP datasets using the software DIYABC Random Forest v1.0 are detailed in Figure S3. We stress here that our intention is not to bring new insights into Human population history, which has been and is still studied in greater details in a number of studies using genetic data, but to illustrate the potential of DIYABC Random Forest v1.0 for the statistical processing of a real IndSeq SNP dataset in the context of a complex evolutionary histories.

Scenario choice

Following the new approach proposed by Estoup et al. (2018), we used DIYABC Random Forest v1.0 to process RF analyses grouping scenarios based on the presence or absence of an admixed origin of the ASW population, and then considered all six scenarios separately. The training sets were generated using the "Training set simulation" module of DIYABC Random Forest v1.0, drawing parameter values into the prior distributions described in the legend of Figure S3 and summarizing SNP data using the same 130 statistics as those used for the pseudo-observed dataset examples in the main text (see Table S1) plus one LDA axis or five LDA axes (i.e., the number of scenarios minus 1; see Pudlo et al. 2016) computed when comparing the

two groups of scenarios or individual scenarios, respectively. We then used the "Random Forest analyses" module of DIYABC Random Forest v1.0 to process RF treatments on the training set which included a total of 12,000 simulated datasets (i.e., 2,000 per scenario). The number of trees in the constructed Random Forest was fixed to 1,000, as this number turned out to be large enough to ensure a stable estimation of the global error rate (Figure S4). We predicted the best scenario and estimated its posterior probability, as well as the global and local error rates, over ten replicate RF analyses based on the same training set.

For comparative purposes, we used the R package abc v2.1 to process scenario choice inferences on the same datasets using two standard ABC methods: the ABC rejection method and the ABC mnlog method based on a simple rejection and a multinomial regression algorithm, respectively (Csilléry, François, & Blum 2012; Blum 2018). For all analyses, we used a tolerance rate of 5% and hence the 600 simulated datasets closest to the observed dataset. The leave-one-out cross-validation method implemented in abc v2.1 was used to compute global error rates from a sample of 10,000 datasets.

Parameter estimation

We focused our estimations on the admixture rate associated to American individuals of African ancestry (i.e. the parameter ra). The training set included 10,000 datasets simulated under scenario 2 (i.e. the selected scenario after processing scenario choice with DIYABC Random Forest v1.0) and summarized using the same 130 statistics plus 2 PLS axes. We inferred point estimates and computed global and local accuracy indices corresponding to global and local NMAE (with the mean and the median as point estimates), as well as the 90% coverage, using out-of-bag estimators from a sample of 10,000 data randomly chosen in the training set (Raynal et al., 2019; Chapuis et al., 2020). The number of trees in the constructed Random Forest was fixed to 1,000, as this number turned out to be large enough to ensure a stable estimation of the global accuracy metrics (Figure S4). For each parameter, we conducted ten replicate RF analyses based on the same training set.

For comparative purposes, we used the R package abc v2.1 to process parameter estimation inference on the same datasets using the ABC rejection method and the ABC logRidge method based on a simple rejection and a regression with a Ridge regulation algorithm, respectively (Csilléry, François, & Blum 2012; Blum 2018). For all analyses, we used a tolerance rate of 5% and hence the 500 simulated datasets closest to the observed dataset. We used an independent test dataset including 1,000 datasets obtained from prior distributions to compute the global NMAE (with the mean and the median as point estimate) and the 90% coverage as accuracy indices.

References cited

Blum, M.G.B. (2018). Regression approaches for ABC. In Sisson, S., Fan, Y., & Beaumont, M., editors, Handbook of Approximate Bayesian Computation. Chapman and Hall/CRC. https://doi.org/10.1201/9781315117195

Chapuis, M-P R., Raynal, L., Plantamp, C., Meynard, CN., Blondin, L., Marin, J-M., & Estoup, A. (2020). A young age of subspecific divergence in the desert locust *Schistocerca gregaria*, inferred by ABC Random Forest. *Molecular Ecology 29(23),* 4542-4558. https//doi:10.1111/mec.15663. Previous version reviewed and recommended by *Peer Community in Evolutionary Biology*, bioRxiv, 671867, 10.24072/pci.evolbiol.100091

Csilléry, K., François, O., & Blum, M. G. (2012). abc: an r package for approximate bayesiancomputation (abc). *Methods in Ecology and Evolution, 3(3),* 475-479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

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

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

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

Rose JI, Usik VI, Marks AE, Hilbert YH, Galletti CS, et al. (2011) The Nubian Complex of Dhofar, Oman: An African Middle Stone Age Industry in Southern Arabia. *PLoS ONE 6(11)*: e28239. doi:10.1371/journal.pone.0028239

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491(7422)*, 56–65. https://doi: 10.1038/nature11632

Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, *76(5)*: 887–893. https:// doi: 10.1086/429864

FIGURE S3. Six scenarios of evolution of four Human populations.
The genotyped populations are YRI = Yoruba (Nigeria, Africa), CHB = Han (China, East Asia), GBR = British (England and Scotland, Europe), and ASW = Americans of African Ancestry (SW USA). The six scenarios differ from each other by one ancient and one recent historical event: (i) a single out-of-Africa colonization event giving an ancestral out-of-Africa population which secondarily split into one European and one East Asian population lineage (scenarios 1, 2 and 3), versus two independent out-of-Africa colonization events, one giving the European lineage and the other one giving the East Asian lineage (scenarios 4, 5 and 6). (ii) The possibility (or not; scenarios 1 and 4) of a recent genetic admixture of ASW individuals with their African ancestors and individuals of European (scenarios 2 and 5) or East Asia (scenarios 3 and 6) origins. The prior distributions of the parameters used to simulate SNP datasets are as followed: Uniform[100; 10000] for the split times t2 and t3 (in number of generations), Uniform[1; 30] for the admixture (or split) time t1, Uniform[0.05; 0.95] for the admixture rate ra (proportion of genes with a non-African origin; only for scenarios with admixture), Uniform[1000; 100000] for the stable effective population sizes N1, N2, N4, N4 and N34 (in number of diploid individuals), Uniform[5; 500] for the bottleneck effective population sizes Nbn3, Nbn4, and Nbn34, Uniform[5; 500] for the bottleneck durations d3, d4, and d34, Uniform[100; 10000] for both the ancestral effective population size Na and the time of change to Na. Conditions on time events were t4>t3>t2 for scenarios 1, 2 and 3, and t4>t3 and t4>t2 for scenarios 4, 5 and 6.

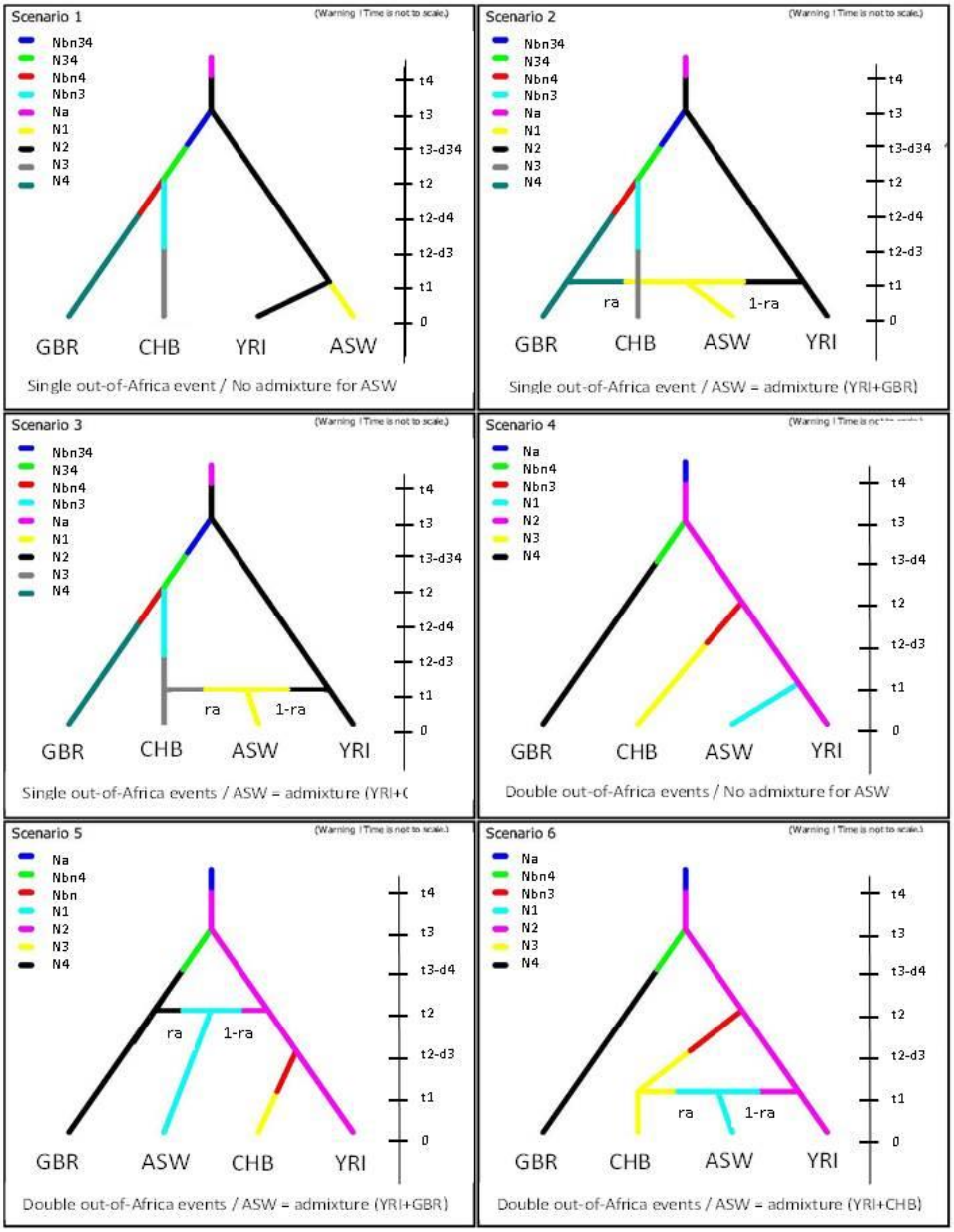
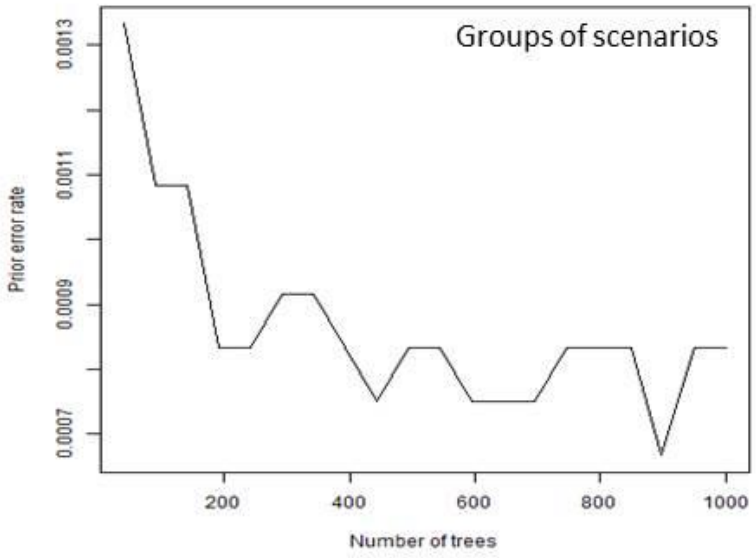
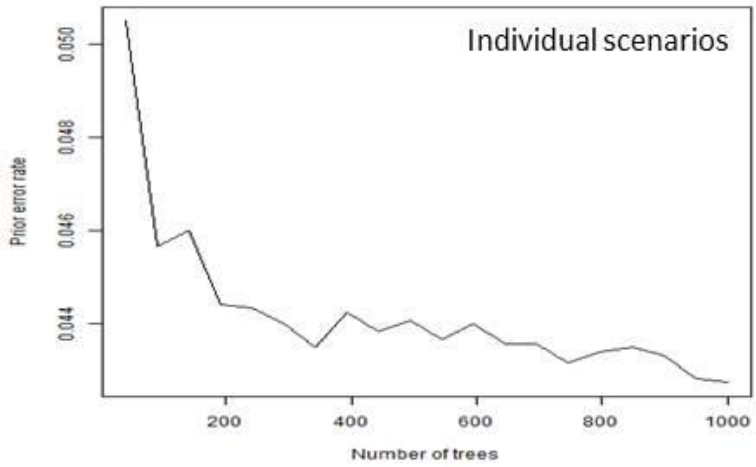


FIGURE S4. Evolution of prediction power relatively to the number of trees in the forest when analyzing the Human population IndSeq dataset.
Prediction power was evaluated by computing the global (prior) error rate for scenario choice and the global (prior) mean squared error with the mean taken as point estimate, for estimation of the parameter ra. The feature vectors included five LDA (for scenario choice) or two PLS (for parameter estimation) axes, and five noise variables. Groups of scenarios = with (scenarios 2, 3, 5 and 6) and without (scenarios 1 and 4) a recent genetic admixture of ASW individuals.





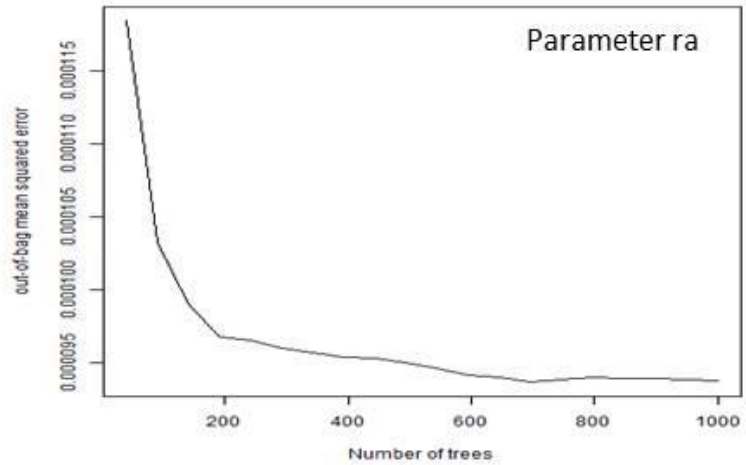


FIGURE S5. Projection of the real Human population IndSeq datasets from the training set on a single LDA axis when analyzing two groups of scenarios (A) or on the first two LDA axes when analyzing the six scenarios of figure S3 separately (B).
The location of the Human population IndSeq observed dataset in the LDA projection is indicated by a vertical line and a star symbol in panels A and B, respectively. Scenario group 1 = with a recent genetic admixture of ASW individuals. Scenario group 2 = without a recent genetic admixture of ASW individuals.

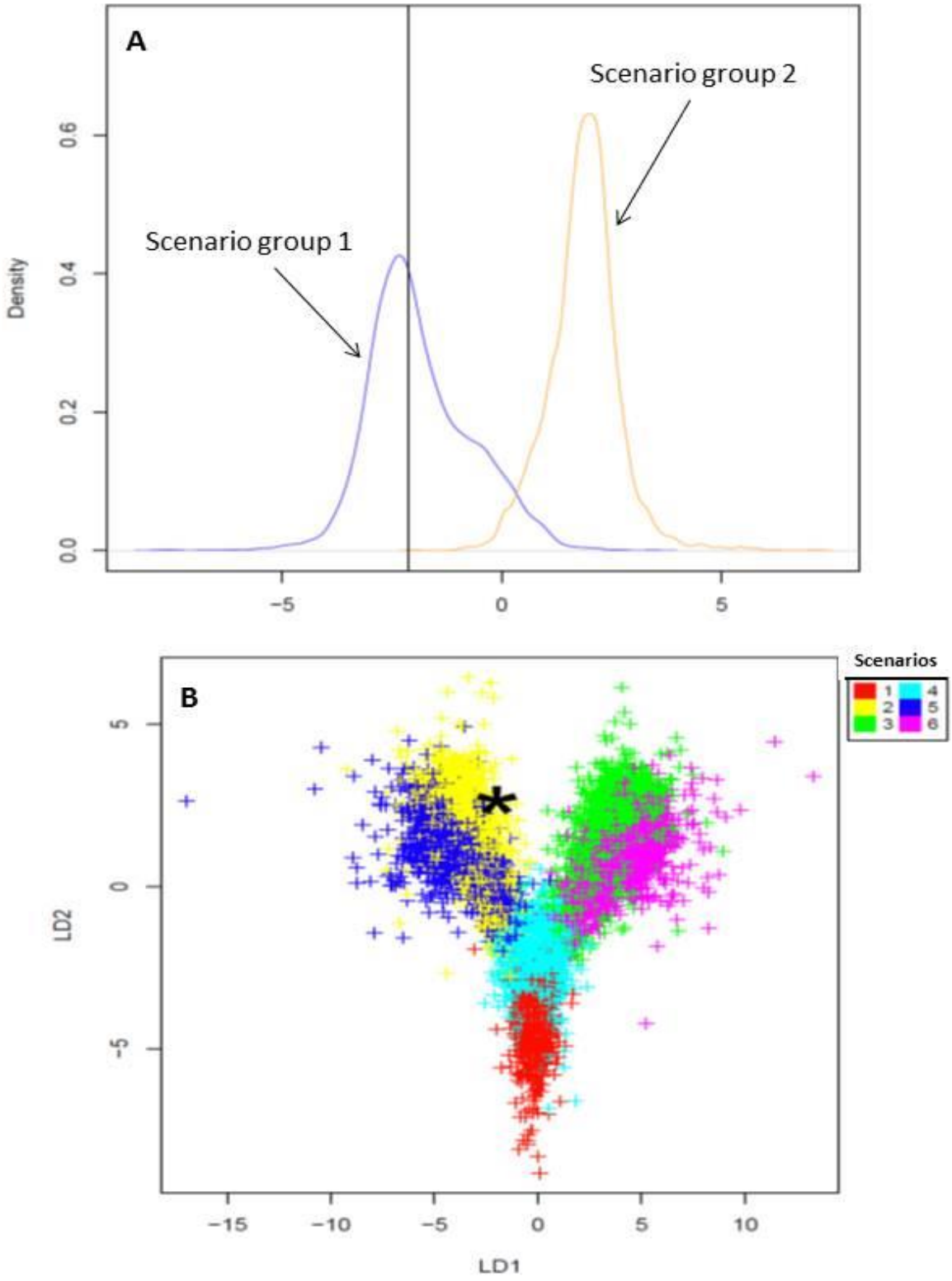
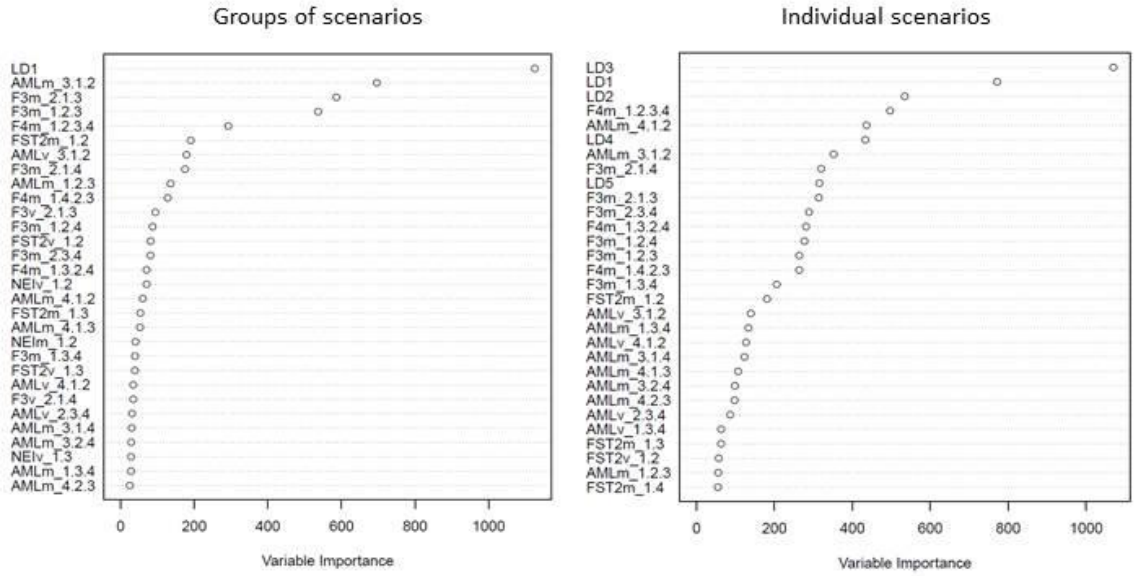
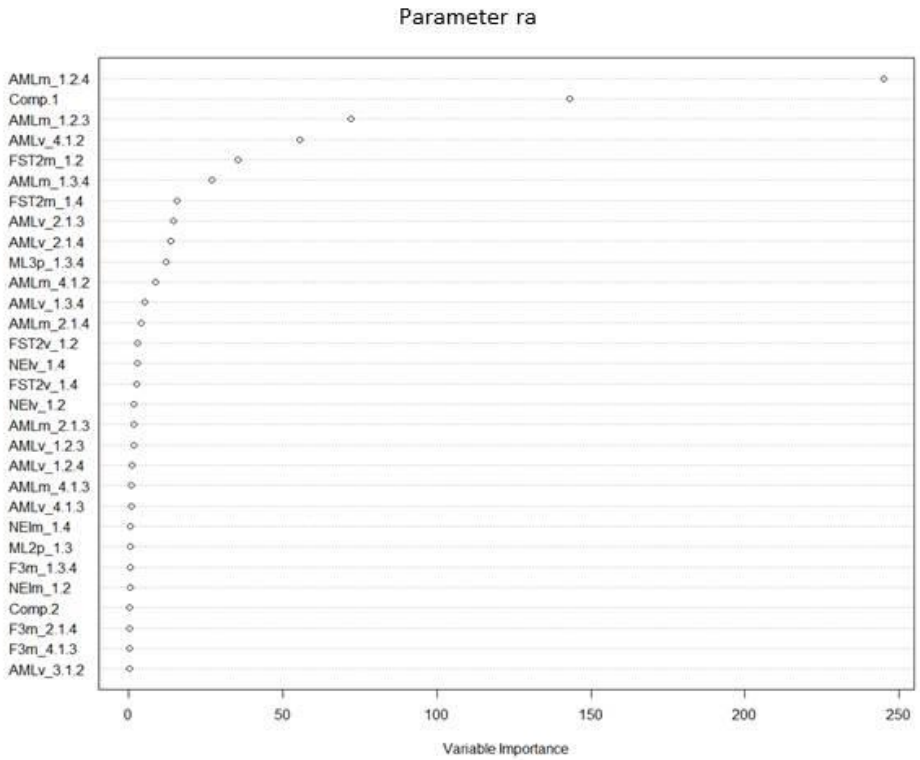


FIGURE S6. Contributions for the real Human population IndSeq dataset analyses of the 30 most informative statistics of the feature vector to the Random Forest when choosing among two groups of scenarios (with and without a recent genetic admixture of ASW individuals), among the six scenarios separately, and when estimating the admixture parameter *ra* under the scenario 2 of figure S3.
The variable importance of each statistics is computed as the mean decrease of impurity across the trees, where the impurity measure is the Gini index, and the residual sum of squares for scenario choice and parameter inference, respectively. It was computed for each of the 130 summary statistics provided by DIYABC, plus the LDA axes for scenario choice (denoted LD) or the PLS axes for parameter estimation (denoted Comp.) that were added to the feature vector. The higher the variable importance the more informative is the statistic. Population index(s) are 1, 2, 3 and 4 for populations ASW, YRI, CHB and GBR, respectively (see figure S3).





Appendix S3: Checking points - thereafter formalized as questions - before finalizing inferential treatments using DIYABC Random Forest v1.0.

1/ Are my scenarios and/or associated priors compatible with the observed dataset?

This question is of prime interest and applies to ABC Random Forest as well as to any alternative ABC treatments. This issue is particularly crucial, given that complex scenarios and high dimensional datasets (i.e., large and hence very informative datasets) are becoming the norm in population genomics. Basically, if none of the proposed scenario / prior combinations produces some simulated datasets in a reasonable vicinity of the observed dataset, this is a signal of incompatibility and it is not recommended to attempt any inferences. In such situations, we strongly advise reformulating the compared scenarios and/or the associated prior distributions in order to achieve some compatibility in the above sense. DIYABC Random Forest v1.0 proposes a visual way to address this issue through the simultaneous projection of datasets of the training set and of the observed dataset on the first LDA axes (e.g., Figure 2 of main text and Figure S5); see also other dedicated diagnostic tools in the notice of the software. In the LDA projection, the observed dataset has to be reasonably located within the clouds of simulated datasets.

2/ Did I simulate enough datasets for my training set?

A rule of thumb is, for scenario choice to simulate between 2,000 and 20,000 datasets per scenario among those compared (Pudlo et al., 2016; Estoup et al., 2018), and for parameter estimation to simulate between 10,000 and 100,000 datasets under a given scenario (Raynal et al., 2019; Chapuis et al., 2020). To evaluate whether or not this number is sufficient for RF analysis, we recommend to compute error/accuracy metrics such as those proposed by DIYABC Random Forest v1.0 from both the entire training set and a subset of the latter (for instance from a subset of 80,000 simulated datasets if the training set includes a total of 100,000 simulated datasets). If error (accuracy) metrics from the subset are similar, or only slightly higher (lower) than the value obtained from the entire database, one can consider that the training set contains enough simulated datasets. If a substantial difference is observed between both values, then we recommend increasing the number of simulated datasets in the training set.

3/ Did my forest grow enough trees?

According to our experience, a forest made of 500 to 2,000 trees often constitutes an interesting trade-off between computation efficiency and statistical precision (Breiman, 2001; Chapuis et al., 2020; Pudlo et al., 2016, Raynal et al., 2019). To evaluate whether or not this number is sufficient, we recommend plotting error/accuracy metrics as a function of the number of trees in the forest. The shapes of the curves provide a visual diagnostic of whether such key metrics stabilize when the number of trees tends to a given value. DIYABC Random Forest v1.0 provides such a plot-figure as output (e.g. Figures S1 and S4).

References cited

Breiman, L. (2001). Random forests. *Machine Learning, 45(1),* 5–32. https://doi.org/10.1023/A:1010933404324

Chapuis, M-P R., Raynal, L., Plantamp, C., Meynard, CN., Blondin, L., Marin, J-M., & Estoup, A. (2020). A young age of subspecific divergence in the desert locust *Schistocerca gregaria*, inferred by ABC Random Forest. *Molecular Ecology 29(23),* 4542-4558. https//doi:10.1111/mec.15663. Previous version reviewed and recommended by *Peer Community in Evolutionary Biology,* bioRxiv, 671867, 10.24072/pci.evolbiol.100091

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

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

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