HAL open science

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

François-david Collin, Ghislain Durif, Louis Raynal, Eric Lombaert, Mathieu Gautier, Renaud Vitalis, Jean-Michel Marin, Arnaud Estoup

To cite this version:

François-david Collin, Ghislain Durif, Louis Raynal, Eric Lombaert, Mathieu Gautier, et al.. Extending Approximate Bayesian Computation with Supervised Machine Learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. Molecular Ecology Resources, 2021, 21 (8), pp.2598-2613. 10.1111/1755-0998.13413 . hal-03229207

HAL Id: hal-03229207

https://hal.inrae.fr/hal-03229207

Submitted on 1 Jun 2021

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.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Check for updates

Received: 10 July 2020 | Revised: 29 March 2021 | Accepted: 28 April 2021

DOI: 10.1111/1755-0998.13413

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest

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

Correspondence
Arnaud Estoup, CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France
Email: arnaud.estoup@inrae.fr

Funding information
INRAE scientific division SPE, Grant/Award Number: AAP-SPE 2016; LabEx NUMEV, Grant/Award Number: NUMEV and ANR10-LABX-20; French Agence National pour la Recherche (ANR), Grant/Award Number: SWING ANR-16-CE02-0015-01, GANDHI ANR-20-CE02-0018 and ABSint ANR-18-CE40-0034

Abstract

Simulation-based methods such as approximate Bayesian computation (ABC) are well-adapted to the analysis of complex scenarios of populations and species genetic history. In this context, supervised machine learning (SML) methods provide attractive statistical solutions to conduct efficient inferences about scenario choice and parameter estimation. The Random Forest methodology (RF) is a powerful ensemble of SML algorithms used for classification or regression problems. Random Forest allows conducting inferences at a low computational cost, without preliminary selection of the relevant components of the ABC summary statistics, and bypassing the derivation of ABC tolerance levels. We have implemented a set of RF algorithms to process inferences using simulated data sets generated from an extended version of the population genetic simulator implemented in DIYABC v2.1.0. The resulting computer package, named DIYABC Random Forest v1.0, integrates two functionalities into a user-friendly interface: the simulation under custom evolutionary scenarios of different types of molecular data (microsatellites, DNA sequences or SNPs) and RF treatments including statistical tools to evaluate the power and accuracy of inferences. We illustrate the functionalities of DIYABC Random Forest v1.0 for both scenario choice and parameter estimation through the analysis of pseudo-observed and real data sets corresponding to pool-sequencing and individual-sequencing SNP data sets. Because of the properties inherent to the implemented RF methods and the large feature vector (including various summary statistics and their linear combinations) available for SNP data, DIYABC Random Forest v1.0 can efficiently contribute to the analysis of large SNP data sets to make inferences about complex population genetic histories.

KEYWORDS
approximate Bayesian computation, demographic history, model or scenario selection, parameter estimation, pool-sequencing, population genetics, random forest, SNP, supervised machine learning

Jean-Michel Marin and Arnaud Estoup are joint senior authors of this study

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

To keep pace with a regular increase of genetic data accessible to biologists, computational methodologies for population genetic inference are constantly and rapidly being developed. Simulation-based likelihood-free methods such as approximate Bayesian computation (ABC; Beaumont et al., 2002) represent an elaborate approach to model-based inference in a Bayesian setting in which model likelihoods are difficult to calculate and must be estimated by massive simulations. Due to their great flexibility, ABC methods are well adapted to the analysis of complex models (hereafter referred to as scenarios) of populations' and species' histories, in which divergence events, population size changes, and genetic admixture or migration events are suspected (reviewed in Beaumont, 2010; Bertorelle et al., 2010; Csilléry et al., 2010). With the advent of next generation sequencing (NGS) technologies, population genetic data sets have drastically grown in size (both in terms of number of genotyped loci and number of genetically characterized populations), so that ABC users are facing two major problems: (i) the simulation of massive numbers of large data sets constituting a so-called reference table, as required for "traditional" ABC methods, becomes prohibitive without extensive computational resources, and (ii) the substantial increase in the number of nonindependent statistics used to extract information from the genetic data (an issue also valid for non-NGS data) poses various statistical problems, including the "curse of dimensionality" whereby accuracy of inferences decreases as the number of summary statistics grows (e.g., Beaumont, 2010). Although much effort has gone into dimensionality reduction and feature selection for ABC (reviewed in Blum et al., 2013; Estoup et al., 2012), reducing dimensionality might lead to loss of information if the remaining summaries fail to capture enough information from the data.

In this context, supervised machine learning (SML) methods provide attractive solutions for statistical inference. SML methods allow predicting new data points through the use of a training set of labeled simulated data examples, for which true response values are known. This data structure is reminiscent of the ABC reference table. The ability of SML methods to use simulation as a stand-in for observed data is crucial for population genetics applications, where adequately sized data sets with high-confidence labels are currently hard to obtain. Most interestingly, some SML methods are able to take advantage of high dimensional input and suffer only slightly from the curse of dimensionality (Anderson et al., 2014; Chen et al., 2013; Schrider & Kern, 2018). SML approaches are currently revolutionizing many fields (e.g., Sebastiani, 2002 in text categorization; Libbrecht & Noble, 2015 in genomics; Angermueller et al., 2016 in genomics and cellular imaging), but their use in population genetics inference is still in its infancy (see for example, Chapuis et al., 2020; Fraimout et al., 2017; Pybus et al., 2015; Schrider & Kern, 2016, 2018; Sheehan & Song, 2016; Schrider et al., 2018; Smith & Carstens, 2020; Smith et al., 2017).

The Random Forest (RF) approach proposed by Breiman (2001) is one of the major SML algorithms for classification (e.g., for scenario choice) or regression (e.g., for estimation of continuous parameters). Pudlo et al. (2016) recently developed RF algorithms to perform scenario choice from simulated data sets summarized through a large set of statistics, as typically considered in ABC, hence leading to the so-called ABC-RF approach. As compared to classical ABC methods, the ABC-RF approach enables efficient discrimination among scenarios and estimation of the posterior probability of the best scenario, with a lower computational burden. More specifically, ABC-RF and other ABC methods provide consistent results for analyses based on a large number of simulated data sets, but ABC-RF outperforms other ABC methods for analyses of multiple complex scenarios based on a smaller (hence more manageable) number of simulated data sets (Fraimout et al., 2017; Pudlo et al., 2016). Building on these results, Raynal et al. (2019) recently proposed an extension of the RF approach in a (nonparametric) regression setting to characterize the posterior distributions of parameters of interest under a given scenario. As compared to alternative ABC solutions, the RF method of Raynal et al. (2019) offers many advantages: (i) a significant improvement in robustness to the choice of summary statistics, (ii) the nonrequirement of any type of tolerance level, and (iii) a good trade-off between the precision of point estimates of parameters and the accuracy of credible intervals for a given computational burden.

The workflow for applying any SML methods to population genetic data includes several stages: (i) the simulation of data under one or several evolutionary scenarios, (ii) the encoding of both simulated and real (observed) data as feature vectors (i.e., summary statistics as in ABC), (iii) the training of the algorithm, applying it on new (observed) data point(s), and (iv) assessing its performance in term of prediction through the computation of error and accuracy measurements. Any effort to create self-contained, efficient, and user-friendly software packages capable of performing this entire workflow would streamline SML methods and make them more accessible to researchers, including nonspecialist users. To that end, we have implemented in a new computer package a set of RF algorithms to infer population' histories from genetic polymorphisms, building upon an extended version of the population genetics simulator implemented in DIYABC 2.1.0 (Cornuet et al., 2014). The data correspond to various types of genetic markers: microsatellites, DNA sequences and SNPs, including individual-sequencing and pool-sequencing SNP data (Gautier et al., 2013; Schlötterer et al., 2014). A large set of summary statistics has also been implemented to improve the extraction of genetic information from SNP data sets. The resulting package, named DIYABC Random Forest v1.0, integrates two functionalities in a user-friendly interface: the simulation under custom evolutionary scenarios of polymorphism data (summarized into a large set of descriptive statistics) and RF treatments including various statistical tools to evaluate the power and accuracy of RF-based inferences. Here we describe the main statistical features of DIYABC Random Forest v1.0 and illustrate its potentialities and functionalities for both scenario choice and parameter estimation through the analyses of pseudo-observed and real data sets corresponding to pool-sequencing and individual-sequencing SNP data.

2 | MATERIALS AND METHODS

2.1 | ABC random forest in the realm of supervised machine learning

The guiding idea of supervised machine learning (SML) approaches is to use a set of data made of explanatory variables (input) and response values (output), in order to learn the relationship between these two, and hence emit a predicted response value for each new input of interest. More formally, SML methods learn this relationship thanks to a function, f, that predicts a response variable, y, from a feature vector, x, containing M input variables, such that $f(x) = y$. If y is a categorical variable (e.g., for scenario choice), one refers to the task as a classification problem, whereas if y is a continuous variable one refers to it as a regression problem (e.g., for parameter estimation). In supervised learning, the objective is to optimize $f{:}x \rightarrow y$ using a training set of labelled data (i.e., whose response values are known). The training set includes values of a feature vector which is a multidimensional representation of any data point made up of measurements (or features) taken from it. That is, one assumes to have a set of training data of length m of the form $\{(x_1, y_1),\ldots,(x_m, y_m)\}$, where $x \in R^M$. A variety of learning algorithms exist which can generate functions that can perform either classification or regression (reviewed in e.g., Schrider & Kern, 2018).

In our inferential framework, SML methods learn from simulations which come from one or several generative model(s) (i.e., scenario[s]). A relevant way to obtain benefits from generative scenario simulations is the Bayesian paradigm and therefore the ABC type approach (Beaumont et al., 2002). Here, the training set is equivalent to the ABC reference table, which includes a given number of data sets that have been simulated for different scenarios using parameter values drawn from prior distributions, each data set being summarized with a set of descriptive statistics. Random forest (RF; Breiman, 2001) is considered as a major SML algorithm for classification or regression. Briefly, RF aggregates the predictions of a collection of classification trees or regression trees, depending on whether the output is categorical (e.g., the identity of a finite number of compared scenarios) or quantitative (e.g., the simulated values of a parameter of interest). Each tree is built by using the information provided by a bootstrap sample of the training set and manages to capture one part of the dependency between the output and the covariates of the feature vector. Based on these random trees which are individually poor predictors of output, a random forest is built by aggregating the tree predictions in order to increase the predictive performances to a high level of accuracy, mainly due to the variance reduction of predictions compared to an individual tree (Breiman, 2001). More detail and in-depth explanation can be found in Pudlo et al. (2016), Fraimout et al. (2017), Estoup et al. (2018) and Marin et al. (2018) for scenario choice, and Raynal et al. (2019) for parameter estimation. See also Appendix S3 of Chapuis et al. (2020) for a concise overview of the RF algorithms and statistical developments used in the present study and implemented in the computer package DIYABC Random Forest v1.0.

2.2 | Simulation of the training set

Before performing RF analyses, one needs to generate a training set. The data sets composing the training set can be simulated under different scenarios and sample configurations, using parameter values drawn from prior distributions. Each resulting data set is summarized using a set of descriptive statistics. We formalized scenarios and prior distributions, and computed summary statistics using the "training set simulation" module of the main pipeline of DIYABC Random Forest v1.0, which essentially corresponds to an extended version of the population genetics simulator implemented in DIYABC v2.1.0 (Cornuet et al., 2014). As in the latter program, DIYABC Random Forest v1.0 allows consideration of complex population histories including any combination of population divergence events, symmetrical or asymmetrical admixture events (but not any continuous gene flow between populations) and changes in past population size, with population samples potentially collected at different times.

DIYABC Random Forest v1.0 accepts various types of molecular data (microsatellites, DNA sequences, and SNPs) evolving under various mutation models and located on various chromosome types (autosomal, X or Y chromosomes, and mitochondrial DNA) for diploid or haploid individuals. To simulate polymorphic data sets at a given SNP locus, we follow the algorithm proposed by Hudson (1993) – cf. *–s 1* option in the program ms associated to Hudson (2002). In DIYABC Random Forest v1.0, it is possible to impose a MAF (minimum allele frequency) criterion on both the observed and simulated data sets. For details, see the user manual of DIYABC Random Forest v1.0 (https://diyabc.github.io/doc/).

In addition to individual-sequencing SNP data (hereafter IndSeq data), DIYABC Random Forest v1.0 allows the simulation and analyses of pool-sequencing SNP data (hereafter PoolSeq data), which consist of whole-genome sequences of pools of tens to hundreds of individual DNAs (Gautier et al., 2013; Schlötterer et al., 2014). In practice, the simulation of PoolSeq data consists first in simulating individual SNP genotypes for all individuals in each population pool, and then generating pool read counts from a binomial distribution parameterized with the simulated allele counts (obtained from individual SNP genotypes) and the total pool read coverage (e.g., Hivert et al., 2018). To account for variation of the total read coverage across SNPs in the observed data set, the coverages across the pools of a given SNP are randomly drawn from the vectors of SNP coverages composing the observed data set. The "synthetic data file generation" module of the program allows the simulation of various types of pseudo-observed "raw" data sets (i.e., not summarized through statistics) without referring to any (actual) observed data set. In the case of raw PoolSeq data sets, the total coverage within each pool of each SNP is sampled from a Poisson distribution with a mean corresponding to an arbitrary coverage value (e.g., 100X) fixed by the DIYABC Random Forest v1.0 user.

It is worth noting that, in contrast to any other types of markers treated in DIYABC Random Forest v1.0 (including IndSeq SNPs), PoolSeq SNP data are considered as located on autosomal

chromosomes only. A criterion somewhat similar to the MAF was implemented for PoolSeq data: the minimum read count (MRC) which is the minimum number of sequence reads for each alleles of a SNP when pooling the reads overall population samples. For details, see the user manual of DIYABC Random Forest v1.0 (https://diyabc.github.io/doc/).

2.3 | Components of the feature vector

The feature vector includes a large number of statistics that summarize genetic variation and capture different aspects of gene genealogies and hence various features of molecular patterns generated by selectively neutral population' histories (e.g., Beaumont, 2010; Cornuet et al., 2014). For microsatellite and DNA sequence markers, DIYABC Random Forest v1.0 proposes by default the same set of summary statistics as DIYABC v2.1.0 (Cornuet et al., 2014). These summary statistics describe genetic variation within populations (e.g., numbers of alleles), between pairs (e.g., genetic distances), or per triplets (e.g., coefficients of admixture) of populations, averaged over loci.

For both IndSeq and PoolSeq SNPs, we have implemented in DIYABC Random Forest v1.0 an extended set (when compared to DIYABC v2.1.0) of summary statistics to more thoroughly describe genetic variation within populations (e.g., proportion of monomorphic loci, heterozygosity, population-specific F_{ST}) and between pair, triplet or quadruplet of populations (e.g., Nei's distance, F_{ST}-related statistics, Patterson's allele-sharing *f*-statistics, coefficients of admixture) to describe genetic variation among various population combinations. More specifically, the proportion of monomorphic loci is computed for each population, as well as for each pair and triplet of populations. Mean and variance (over loci) values are computed for all subsequent summary statistics. Heterozygosity is computed for each population and for each pair of populations as $(1-Q_1)$ and $(1-Q_2)$, where Q_1 and Q_2 are the probabilities of identity between pairs of genes (Hivert et al., 2018). F_{ST}-related statistics are computed for each population (i.e., population-specific F_{ST}; Weir & Goudet, 2017), as well as for each pair, triplet, quadruplet and overall populations (when the data set includes more than four populations), using the method-of-moments estimators described in Hivert et al. (2018). In addition, we compute Patterson's *f*-statistics for each triplet (f_3-statistics) and quadruplet (f_4-statistics) of populations as described in Patterson et al. (2012), except for the f_3-statistics for PoolSeq read count data which are computed using the unbiased estimator described in Leblois et al. (2018). Finally, distance as in Nei (1972) is computed for each pair of populations and the coefficient of admixture is computed for each triplet of populations as described in Cornuet et al. (2014). For additional details, see the user manual of DIYABC Random Forest v1.0 (https://diyabc.github.io/doc/). An illustration of the feature vector composed of all above summary statistics is given in Table S1 for the analysis of two example SNP pseudo-observed data sets.

For scenario choice, the feature vector can be expanded by values of the *d* axes of a linear discriminant analysis (LDA) processed on the above summary statistics (with *d* equal to the number of scenarios minus 1; Pudlo et al., 2016). In the same spirit, for parameter estimation, the feature vector can be completed by values of a subset of the *s* axes of a partial least squares regression analysis (PLS) also processed on the above summary statistics (with *s* equal to the number of summary statistics). The number of PLS axes added to the feature vector is determined as the number of PLS axes providing a given fraction of the maximum amount of variance explained by all PLS axes (i.e., 95% by default, but this parameter can be adjusted).

2.4 | Prediction using random forest

We used the "random forest analyses" module of the main pipeline of the software DIYABC Random Forest v1.0 to perform RF analyses (i.e. predictions) on a given target data set. For scenario choice, the outcome of the first step of RF computation is a classification vote for each scenario which represents the number of times a scenario is selected in a forest of *n* trees. The scenario with the highest classification vote corresponds to the scenario best suited to the target data set among the set of compared scenarios. This first RF predictor is good enough to select the most likely scenario but not to derive directly the associated posterior probabilities. A second analytical step based on a second random forest in regression is necessary to provide an estimation of the posterior probability of the best-supported scenario (Pudlo et al., 2016). Raynal et al. (2019) extended the RF approach to estimate the posterior distributions of parameters of interest in a given scenario. Their approach requires the derivation of a new RF for each component of interest of the parameter vector. Practitioners of Bayesian inference often report the posterior mean, posterior variance or posterior quantiles, rather than the full posterior distribution, since the former are easier to interpret than the latter. We implemented the methodologies detailed in Raynal et al. (2019) to provide estimations of the posterior mean, variance, median (i.e., 50% quantile) as well as 5% and 95% quantiles (and hence 90% credibility interval) of each parameter of interest. The posterior distribution of each parameter of interest was obtained using importance weights following the work by Meinshausen (2006) on quantile regression forests.

2.5 | Assessing the quality of predictions

For scenario choice and parameter estimation, DIYABC Random Forest v1.0 allows evaluating the robustness of inferences. Because the level of errors on scenario choice and accuracy of parameter estimation may substantially differ depending on the location of an observed data set in the prior data space, prior-based indicators are poorly relevant, aside from their use to select the best classification method and possibly a set of highly informative components of the feature vector. Therefore, in addition to

global (i.e., prior) error/accuracy corresponding to prediction quality measures computed over the entire data space, it is crucial to compute local (i.e., posterior) error/accuracy conditionally on the observed data set, corresponding to prediction quality exactly at the position of the observed data set. For scenario choice, the global prior errors, including the confusion matrix (i.e., the contingency table of the true and predicted classes for each example in the training set) and the mean misclassification error rate, were computed using the out-of-bag (a.k.a. out-of-bootstrap) training data as a free test data set. The out-of-bag data set corresponds to the data of the training set that were not selected when creating the different tree bootstrap samples and is hence equivalent to using an independent test data set (Breiman, 2001; Pudlo et al., 2016; Raynal et al., 2019). Using the out-of-bag prediction method for estimating global and local error/accuracy measures is computationally efficient as this approach makes use of the data sets already present in the training set and hence avoids the computationally costly simulations (especially for large SNP data sets) of additional test data sets. Chapuis et al. (2020) highlighted that the local (posterior) error for scenario choice can be computed as 1 minus the posterior probability of the selected scenario.

For parameter estimation, we also relied on out-of-bag predictions to compute both global (i.e., prior) and local (i.e., posterior) accuracy measures, as detailed in the Appendix S3 of Chapuis et al., 2020. Accuracy measures include: (i) both the global and local NMAE (i.e., the normalized mean absolute error which is the average absolute difference between the point estimate and the true simulated value divided by the true simulated value) with the mean or the median taken as point estimate; (ii) both the global and local MSE and NMSE (i.e., the mean square error which is the average squared difference between the point estimate and the true simulated value for MSE, divided by the true simulated value for NMSE), again with the mean or the median taken as point estimate; and (iii) several confidence interval measures, computed only at the global scale, including the 90% coverage (i.e., the proportion of true simulated values located between the estimated 5% and 95% quantiles), and the mean or the median of the 90% amplitude and relative 90% amplitude (i.e., the mean or median of the difference between the estimated 5% and 95% quantiles for the 90% amplitude, divided by the true simulated value for the relative 90% amplitude).

2.6 | Main technical features of the package DIYABC Random Forest v1.0

The package DIYABC Random Forest v1.0 is composed of three parts: the data set 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 main technical features of the package (implementation, interface, outputs, memory space and computing time) are described in Appendix S1.

2.7 | Illustration using pseudo-observed SNP data sets

2.7.1 | Compared scenarios and prior distributions

We considered a case study where one wants to make inferences about the genetic origin of a population of interest (for example a recent invasive population) among a set of possible source populations (for which the topology is known; see Figure 1). The target population (pop 4) has three possible single population sources (pop1, 2 or 3) and three possible admixed pairwise population sources (i.e., admixture between pop1 & 2, pop1 & 3 and pop2 & 3). We hence formalized six competing scenarios that constitute two groups of scenarios when referring to the presence or absence of an admixture event when founding the target population 4: group 1 includes three scenarios including an admixture event (scenarios 1, 2 and 3) and group 2 three scenarios without any admixture event (scenarios 4, 5 and 6). Such grouping approach in scenario choice is relevant to disentangle in our analysis the level of confidence to make inferences about a given (or several) specific evolutionary event of interest, here the presence or absence of an admixed origin of population 4 (Chapuis et al., 2020; Estoup et al., 2018).

Demographic and historical parameters include four effective population sizes N_1, N_2, N_3 and N_4 (for pop 1, 2, 3, and 4, respectively) and three divergence or admixture time events (t_1, t_2 and t_3), with t_1 the divergence or admixture time of pop4, t_2 the divergence time of pop3 from pop2, and t_3 the divergence time of pop2 from pop1 (Figure 1). For the three scenarios with admixture, the parameter r_a corresponds to the proportion of genes of a given source population entering into the admixed pop4. Prior values for time events (t_1, t_2, and t_3) were drawn from uniform distributions bounded between 10 and 1,000 generations, with $t_3>t_2>t_1$. We used uniform prior distributions bounded between 1×10^2 and 1×10^4 diploid individuals for each effective population sizes N_1, N_2, N_3 and N_4. The admixture rate r_a was drawn from a uniform prior distribution bounded between 0.05 and 0.95.

2.7.2 | Pseudo-observed data sets

Our prediction targets correspond to four pseudo-observed data sets that were generated using the "Synthetic data file generation" module of DIYABC Random Forest v1.0 under the (admixed) scenario 3 or the (nonadmixed) scenario 6 using the following parameter values: N_1 = 7,000, N_2 = 2,000, N_3 = 4,000, N_4 = 3,000, t_1 = 200, t_2 = 300, t_3 = 500, and r_a = 0.3 for scenario 3. The short divergence times and large effective population sizes values correspond to a situation of low level of genetic differentiation among populations (cf. parwise F_{ST} values ranging from 3% to 7%) and hence to a difficult case study. The four pseudo-observed data sets correspond to a PoolSeq read count data set and an IndSeq allele count data set generated under scenario 3 and under scenario 6, each with 30,000 SNPs. They represent similar sequencing efforts: a 100X coverage

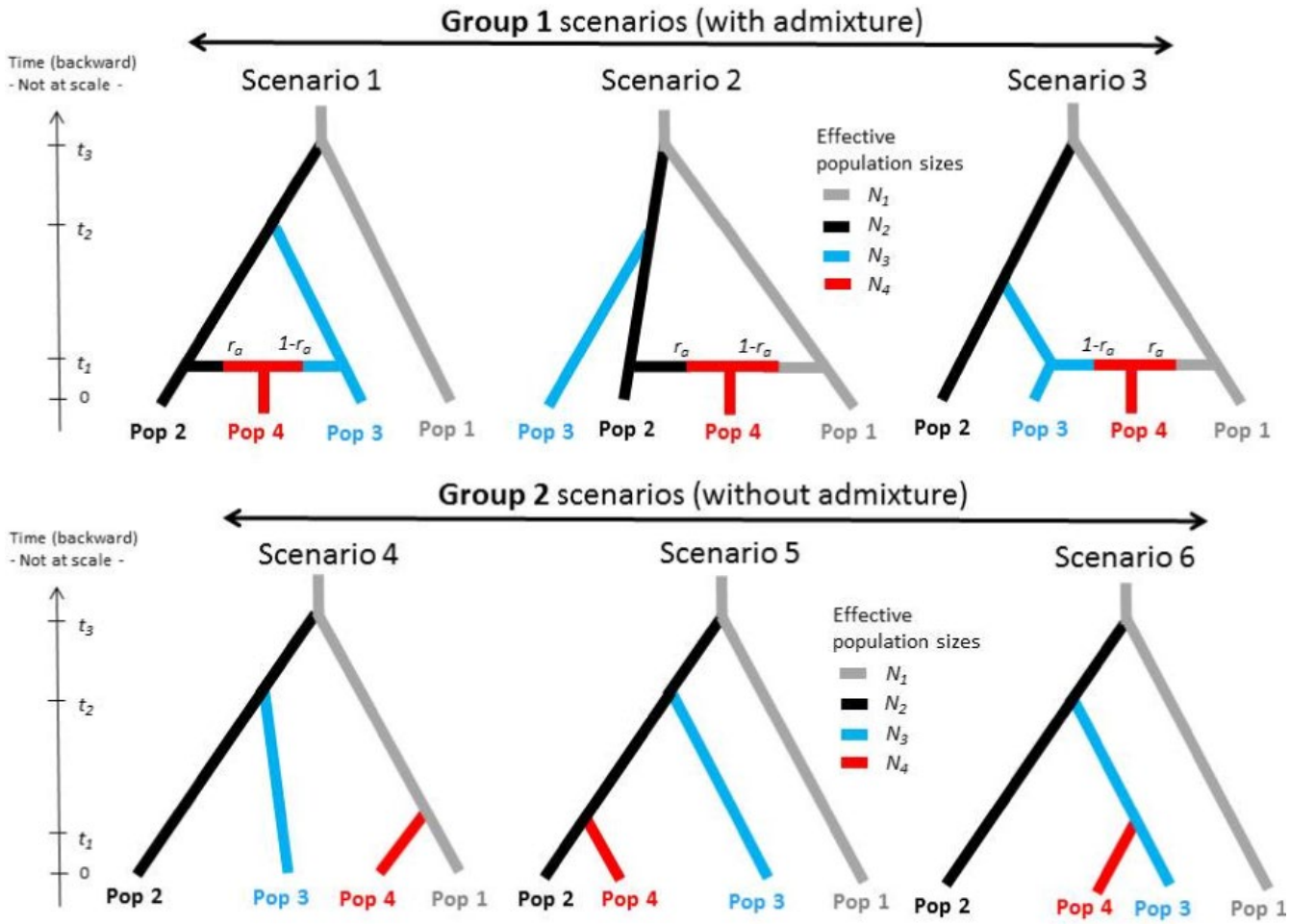

FIGURE 1 Evolutionary scenarios compared. The target population (pop 4) has three possible single (i.e., nonadmixed) population sources (pop 1, pop 2 or pop 3) composing a group of three scenarios without admixture (group 2 in the figure) and three possible admixed pairwise population sources (i.e., admixture between pop1& pop2, pop 1& pop3 and pop 2 & pop3) composing a group of three scenarios with admixture (group 1 in the figure). Demographic and historical parameters include four effective population sizes N_1, N_2, N_3 and N_4 (for populations 1, 2, 3, and 4, respectively) and three divergence or admixture time events (t_1, t_2 and t_3), For the scenarios with admixture, the parameter r_a corresponds to the proportion of genes of a given source population entering into the admixed population 4. See text for details about prior distribution of parameters

for each population of the PoolSeq data sets (with 100 individuals per population pool) and 10 individuals sequenced per population for the IndSeq data sets with a 10X coverage for each sequenced individual (the latter parameter being not explicitly indicated in the program as individual SNP genotypes are considered to be inferred without errors). Analyses were processed on a subset of 5,000 SNPs with a MRC = 5 for the PoolSeq data sets and a MAF = 5% for the IndSeq data sets.

2.7.3 | Scenario choice

We processed scenario choice analyses grouping scenarios based on the presence or absence of an admixed origin of population 4, and then considered all six scenarios separately. The training sets which included a total of 12,000 simulated data sets (i.e., 2,000 per scenario) were generated using the "Training set simulation" module of DIYABC Random Forest v1.0, drawing parameter values into the prior distributions described above and summarizing SNP data using 130 statistics (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 sets. Following Pudlo et al. (2016), we checked that 12,000 data sets in the training set was sufficient by evaluating the stability of prior error rates and posterior probabilities estimations of the best scenario on subsets of 10,000, 11,000 and 12,000 data of the training set (results not shown). The number of trees in the constructed Random Forest was fixed to 1,000, as this number was large enough to ensure a stable estimation of the global error rate (Figure S1). 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 on the same data sets using two traditional ABC methods: the ABC rejection method and the ABC mnlog method based on a simple rejection and a multinomial regression algorithm, respectively (Blum, 2018; Csilléry et al., 2012). For all analyses, we used a tolerance rate of 5% and hence the 600 simulated data sets closest to the observed data set. 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 data sets.

2.7.4 | Parameter estimation

Following Raynal et al. (2019), we conducted independent RF treatments for each parameter of interest. For the sake of concision, we focused our estimations on four parameters involved in the admixture event in scenario 3 (i.e., the selected scenario after processing scenario choice for the pseudo-observed data sets generated under the admixed scenario 3): the founding/admixture time for the target pop 4 (t_1), the admixture rate (r_a corresponding to the proportion of genes originating from pop 1), the effective population size of pop 4 (N_4), and the compound parameter corresponding to the ratio t_1/N_4. The same parameters, except r_a, were estimated for the pseudo-observed data sets generated under the nonadmixed scenario 6. Considering ratios (or products) of parameters - here the admixture time scaled by the effective population size as drift parameter - allows reducing parameter identifiability issues of some scenarios (e.g., Beaumont, 2010). The training sets included 10,000 data sets simulated under scenario 3 or scenario 6, and summarized using the same 130 statistics (Table S1) plus 4 to 24 PLS axes depending on the parameter estimated and the training set analysed. For each parameter, we inferred point estimates and computed global and local accuracy metrics 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. We checked that 10,000 data sets in the training set were sufficient by evaluating the stability of the global accuracy metrics (i.e., NMAE using the mean as point estimates) on subsets of 8,000, 9,000 and 10,000 data of the training set (results not shown). The number of trees in the constructed random forest was fixed to 1,000, as this number wase large enough to ensure a stable estimation of the global accuracy metrics (Figure S1). 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 on the same data sets using the ABC rejection method and the ABC logRidge method based on a simple rejection and a regression with a Ridge regulation algorithm, respectively (Blum, 2018; Csilléry et al., 2012). For all analyses, we used a tolerance rate of 5% and hence the 500 simulated data sets closest to the observed data set. We used an independent test data set including 1,000 data sets obtained from prior distributions to compute the global NMAE (with the mean and the median as point estimate) and the 90% coverage as accuracy metrics.

2.8 | Illustration using a real IndSeq SNP data set of human populations

We analysed an IndSeq real data set including 5,000 SNP markers genotyped in four human populations by The 1000 Genomes Project Consortium (2012). The four populations include Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African ancestry. Our intention is not to bring new insights into human population history, but to illustrate the potential of DIYABC Random Forest in this context. We compared six scenarios of evolution of the four human populations and focused on the estimation of the admixture rate associated to American individuals of African ancestry. The scenarios and prior distribution, the real and simulated IndSeq data sets, and the statistical methods used for inferences, including Random Forest and traditional ABC methods, were similar to those described for the analyses of the pseudo-observed data sets in section 2.7 (Figures S3 and S4). See Appendix S2 for details.

3 | RESULTS

For both scenario choice and parameter estimation, we illustrate the inferential power and functionalities of DIYABC Random Forest v1.0 through the analysis of four pseudo-observed SNP data sets corresponding to PoolSeq and IndSeq data. We first processed RF analyses grouping scenarios based on the presence or absence of an admixed origin of the target population 4, and then considered all six compared scenarios separately. We then estimated parameters of interests under the selected (best) scenario. We contrasted our inferential results with and without adding LDA axes (for scenario choice) or PLS axes (for parameter estimation) to the RF feature vector initially composed of 130 summary statistics. Finally, for comparative purposes, we present results obtained using two traditional ABC methods. In the following sections 3.1, 3.2 and 3.3, we detail results for the two pseudo-observed data sets generated under the (admixed) scenario 3. Similar results were indeed obtained for the two pseudo-observed data sets generated under the (nonadmixed) scenario 6 (see Tables S2 and S3).

3.1 | Scenario choice

The projection of the data sets of the training set on a single (when analysing the two groups of scenarios) or on the first two LDA axes (when analysing the six scenarios considered separately) provides a first visual indication about our capacity to discriminate among the compared scenarios (Figure 2). Simulations under the two groups of scenarios moderately overlapped suggesting a substantial power to discriminate among them. When considering the six scenarios individually, the projected points overlapped in a more marked way, at least for some of the scenarios, suggesting an overall lower power to discriminate among scenarios considered separately than when considering the two groups of scenarios. As a first inferential clue, the

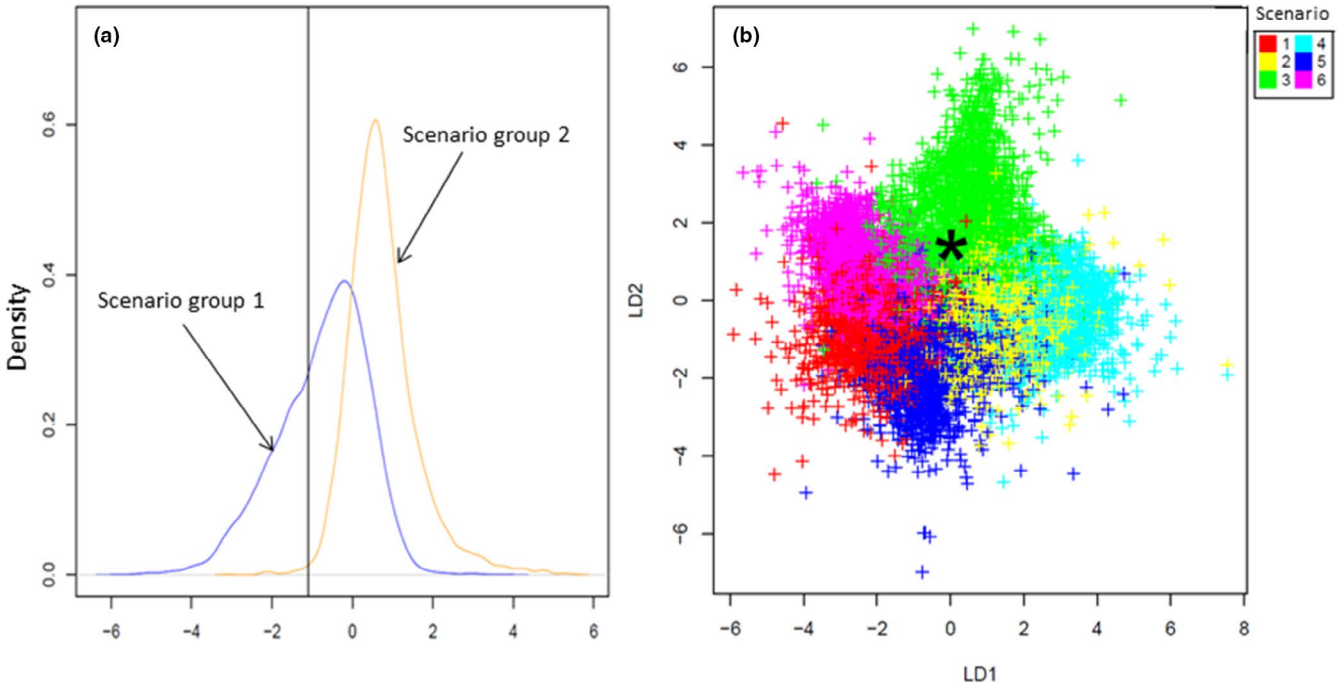

FIGURE 2 Projection of the PoolSeq data sets from the training set on a single LDA axis when analysing the two groups of scenarios (a) or on the first two LDA axes when analysing the six scenarios separately (b). The six compared scenarios and the two groups of scenarios are detailed in Figure 1. The location of the PoolSeq pseudo-observed data set in the LDA projection is indicated by a vertical line and a star symbol in panels a and b, respectively. The pseudo-observed data sets was simulated under the (admixed) scenario 3 (belonging to the group 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

location of the observed data set (indicated by a vertical line and a star symbol in Figure 2a,b, respectively) suggests, albeit without any formal quantification, a marked association with the scenario group 1 and with the scenario 3.

The RF classification votes and posterior probabilities estimated for both the PoolSeq and IndSeq pseudo-observed data sets (with or without adding LDA axes to the feature vector) were the highest for the scenario group 1, which includes an admixture event (Table 1). When considering the six scenarios separately, the highest classification votes and posterior probabilities were for scenario 3, which congruently includes an admixture event between the pop 1 & 3 as sources of the target pop 4. The posterior probabilities of scenario group 1 and scenario 3 were relatively high (from 0.657 to 0.891), which is satisfactory when considering the difficulty of the example case study (cf. low level of genetic differentiation among populations). We found that including LDA axes in the RF vector feature substantially improved scenario choice predictions (e.g., global prior error rates were 3% to 12% lower when including LDA axes; Table 1). The levels of errors were considerably different at the global and local scales, with lower levels at the local scale for analyses of the PoolSeq data set, and a trend for higher levels at the local scale for analyses of the IndSeq data set.

Finally, we obtained better prediction levels (with or without LDA axes) for the PoolSeq data set than the IndSeq data set (e.g., global prior error rates were 14% to 27% lower for the PoolSeq data set; Table 1). This indicates that, for a similar sequencing effort, a PoolSeq strategy is preferable to an IndSeq strategy, at least when a substantially large number of individual samples are available. This result, which might basically stem from a more accurate estimation of allele frequency when using PoolSeq data, echoes theoretical results in the comparative study by Gautier et al. (2013).

Traditional ABC methods provide qualitatively similar results, but precision metrics were poorer compared to those obtained using ABC Random Forest (Table S4).

3.2 | Parameter estimation

NMAE values show that estimations were substantially more accurate both at the global and local scales for the admixture rate r_a and the compound parameter t_1/N_4 (cf. the low NMAE values for these parameters) than for t_1 and N_4 (Table 2). This result is also illustrated for the two pseudo-observed data sets by point estimates close to the true values and narrow 90% CI for r_a and t_1/N_4. NMAE values computed from median point estimates were systematically smaller (albeit sometimes only to a small extent) than those computed from mean point estimates, indicating that the median is globally a better point estimate of the parameter than the mean. As expected when considering point estimates for the two pseudo-observed data sets, this trend did not translate for all parameters.

We found that including PLS axes in the RF feature vector improved parameter estimation in a heterogeneous way. The accuracy

TABLE 1 Results for scenario choice

Type of data set	Type of treatment	Global error rate	Local error rate	Vote scen. 1	Vote scen. 2	Vote scen. 3	Vote scen. 4	Vote scen. 5	Vote scen. 6	Posterior probability
PoolSeq	Groups of scenarios: with vs. without admixture									
	RF with LDA	0.172 (0.001)	0.085 (0.009)	925.1 (10.754)			74.9 (10.754)			.915 (group 1) (.009)
	RF without LDA	0.192 (0.001)	0.162 (0.015)	891.9 (13.585)			108.1 (13.585)			.838 [group 1] (.015)
	All scenarios considered separately									
	RF with LDA	0.196 (0.0008)	0.135 (0.011)	4.1 (1.297)	65.3 (7.364)	897.0 (8.056)	8.5 (2.121)	15.7 (4.808)	9.4 (2.011)	.865 [scen. 3] (.011)
	RF without LDA	0.220 (0.0009)	0.202 (0.013)	9.3 (3.020)	98.5 (17.264)	829.4 (20.304)	8.3 (2.669)	32.9 (3.381)	21.6 (4.671)	.798 [scen. 3] (.013)
IndSeq	Groups of scenarios: with vs. without admixture									
	RF with LDA	0.212 (0.001)	0.177 (0.016)	840.6 (12.816)			159.4 (12.816)			.823 [group 1] (.016)
	RF without LDA	0.220 (0.002)	0.270 (0.016)	805.5 (18.940)			194.5 (18.940)			.730 [group 1] (.016)
	All scenarios considered separately									
	RF with LDA	0.248 (0.001)	0.268 (0.018)	6.9 (3.107)	105.9 (10.692)	817.0 (13.021)	12.7 (3.057)	41.2 (5.159)	16.3 (3.653)	.732 [scen. 3] (.018)
	RF without LDA	0.262 (0.001)	0.343 (0.0197)	9.0 (2.981)	123.5 (7.322)	769.6 (12.366)	15.8 (4.049)	60.7 (8.982)	21.4 (4.274)	.657 [scen. 3] (.020)

Note: The six compared scenarios and the two groups of scenarios are detailed in Figure 1 Results are given for the two example pseudo-observed data sets (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 = 3,000, 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 10 replicate analyses are given between brackets for each metrics, in addition to the means. See Table S2 for a comparison with traditional ABC methods.

TABLE 2 Results for estimation of parameters of interest

Type of data set	Type of treatment	Parameter	Posterior point estimates of			Global (prior) NMAE computed from		Local (posterior) NMAE computed from		
			Mean	Median	90% CI	Mean	Median	Mean	Median	90% coverage
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)
	RF without PLS	r_a	0.364 (0.0028)	0.368 (0.0032)	0.245–0.483 (0.0072) (0.0055)	0.143 (0.00026)	0.130 (0.00028)	0.010 (0.0032)	0.098 (0.0025)	0.973 (0.0004)
		t_1	288.2 (3.752)	301.7 (2.540)	134.4–443.0 (8.044) (4.546)	0.322 (0.00042)	0.301 (0.00046)	0.235 (0.0095)	0.231 (0.0091)	0.959 (0.0007)
		N_4	5517 39.60	5539 94.16	2319–8662 (104.3) (133.8)	0.456 0.0006	0.421 0.0006	0.324 (0.0116)	0.297 (0.0091)	0.936 (0.0011)
		t_1/N_4	0.068 (0.0004)	0.068 (0.0006)	0.050–0.085 (0.0008) (0.0006)	0.218 (0.0009)	0.179 (0.0003)	0.079 (0.0021)	0.078 (0.0017)	0.980 (00006)
		t_1/N_4	0.073	0.073	0.069–0.075	0.203	0.205	NC	NC	0.702
IndSeq		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)
	RF without PLS	r_a	0.417 (0.0052)	0.410 (0.0041)	0.284–0.612 (0.0050) (0.0143)	0.173 (0.0003)	0.155 (0.0003)	0.162 (0.0055)	0.153 (0.0048)	0.963 (0.0007)
		t_1	399.0 (3.184)	395.2 (3.370)	227.9–591.9 (4.653) (4.758)	0.407 (0.0005)	0.366 (0.0005)	0.191 (0.0042)	0.190 (0.0041)	0.959 (0.0007)
		N_4	5837 (50.70)	5978 (93.02)	2524–9210 (134.6) (166.8)	0.499 (0.0006)	0.467 (0.0006)	0.295 (0.0053)	0.292 (0.0051)	0.926 (0.0007)
		t_1/N_4	0.061 (0.0003)	0.061 (0.0003)	0.045–0.078 (0.0005) (0.0008)	0.263 (0.0008)	0.221 (0.0005)	0.092 (0.0025)	0.092 (0.0024)	0.976 (0.0008)

Note: Results are given for the two example pseudo-observed data sets (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) data sets 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. CI: credibility interval. 90% coverage: proportion of test parameter values comprise between the estimated 5% and the 95% quantile. Standard deviations over the 10 replicate analyses are given between brackets for each metrics, in addition to the means. See Table S3 for a comparison with traditional ABC methods.

gain of including PLS axes ranged from negligible (e.g., IndSeq global NMAE for t_1/N_4 based on median of 0.220 and 0.221 with and without PLS, respectively) to substantial (e.g., PoolSeq global NMAE for N_4 based on median of 0.380 and 0.421 with and without PLS, respectively). The accuracy levels were always lower at the global than local scale, sometimes to a large extent (Table 2). In the present case study, the pseudo-observed data sets are hence located in a favourable part of the prior space. Finally, like scenario choice analyses, we obtained considerably higher accuracy (i.e., lower NMAE values with or without PLS axes) for the PoolSeq data set than the IndSeq data set. In accordance with this, point estimates for all parameters of the two pseudo-observed data sets were closer to the true values with narrower ranges of 90% CI for PoolSeq than IndSeq data sets. This reinforces our previous conclusion that, for a similar sequencing effort, it is preferable to use a PoolSeq strategy than an IndSeq strategy when a large number of individual samples are available.

Similar trends were observed when using traditional ABC methods, but the later type of methods were generally characterized by poorer accuracy metrics when compared to those obtained using ABC random forest (Table S5).

3.3 | Contribution to random forest inferences of components of the feature vector

Learning more about how various summary statistics relate to scenarios or parameters would be useful for population genetics going forward. In the realm of traditional ABC methods, it is not clear which summary statistics are responsible for a signal. By contrast, many SML methods including RF allow direct measurement of the contribution of each component included in the feature vector. RF hence offer direct ways to assess which features of the input are driving inferences, information which can yield insights about the underlying processes. Figure 3 illustrates how RF automatically ranks the components of the feature vector according to their level of information when building trees of the forest. Figure 3 and Figure S2 show that informative statistics are different depending on the comparisons (individual scenarios or groups of scenarios) and the analysed parameter in a given scenario. Four- and three-sample *f*-statistics, as well as the related three-sample coefficients of admixture (i.e., AML statistics), were among the most informative to discriminate scenarios (Figure 3a). In accordance with this, such statistics are by construction highly sensitive to the topology connecting populations and including or not an admixture event (Estoup et al., 2018; Patterson et al., 2012). A typical feature of RF scenario choice is that one or several LDA axes always correspond to the best informative statistics.

For parameter estimation, the most informative summary statistics were different depending on the parameter of interest (Figure 3b and Figure S2). Figure 3b shows that for the (well-estimated) compound parameter t_1/N_4, the most informative statistics included three-sample *f*-statistics and AML statistics with the pop 4 as target, the population-specific F_{ST}, ML1p (proportion of monomorphic loci) and heterozygosity - all for pop 4 -, and pairwise-population statistics (F_{ST} and Nei's distance) that included pop 4. For other parameter values, the set of informative statistics differed among parameters, but always included a large number of four-sample and three-sample *f*-statistics, as well as three-sample AML statistics (Figure S2). In contrast to LDA axes (used for scenario choice), only a subset of PLS components were ranked among the 30 most informative statistics and they were never ranked at first position

We added five noise variables (corresponding to values randomly drawn into uniform distributions bounded between 0 and 1) to the feature vector processed by RF in order to evaluate the threshold of variable importance metrics below which components of the vector were not informative anymore. We found that for both scenario choice and parameter estimation, a substantial proportion of summary statistics was not informative. We found that 28% to 38% and 20% to 65% of the summary statistics were informative for scenario choice and parameter estimation, respectively. It is worth stressing that noninformative components of the feature vector are simply not or seldom chosen when constructing each individual trees of the forest, and hence do not alter RF inferences (Breiman, 2001; Marin et al., 2018; Raynal et al., 2019). In agreement with this, removing noise variables from the feature vector did not impact the levels of errors in scenario choice and of accuracy in parameter estimation in the present case study (results not shown).

3.4 | Illustration using a real IndSeq SNP data set of human populations

We analysed an IndSeq real data set including 5,000 SNP markers genotyped in four human populations, including Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African ancestry. We compared six scenarios of evolution of these populations and focused on the estimation of the admixture rate associated with American individuals of African ancestry (Figure S3). The scenarios and prior distributions, the real and simulated IndSeq data sets, and the statistical methods used for inferences, including Random Forest and two standard ABC methods, are detailed in Appendix S2.

Regarding scenario choice, ABC Random Forest using the LDA axes provides the best results. The RF algorithm selects (according to the number of votes) the group of scenarios including an admixture event and more specifically scenario 2 as the forecasted scenario, an answer suggested visually on the LDA projections of Figure S5 in Appendix S2. The posterior probability of the selected group of scenarios was 1.000 with LDA axes in the feature vector (global prior error rate = 0.0008) and 1.000 without LDA axes (global prior error rate = 0.0011). The posterior probability of the (admixed) scenario 2 was 0.997 with LDA axes (global prior error rate = 0.042) and 0.995 without LDA axes (global prior error rate = 0.061). Considering previous population genetics studies

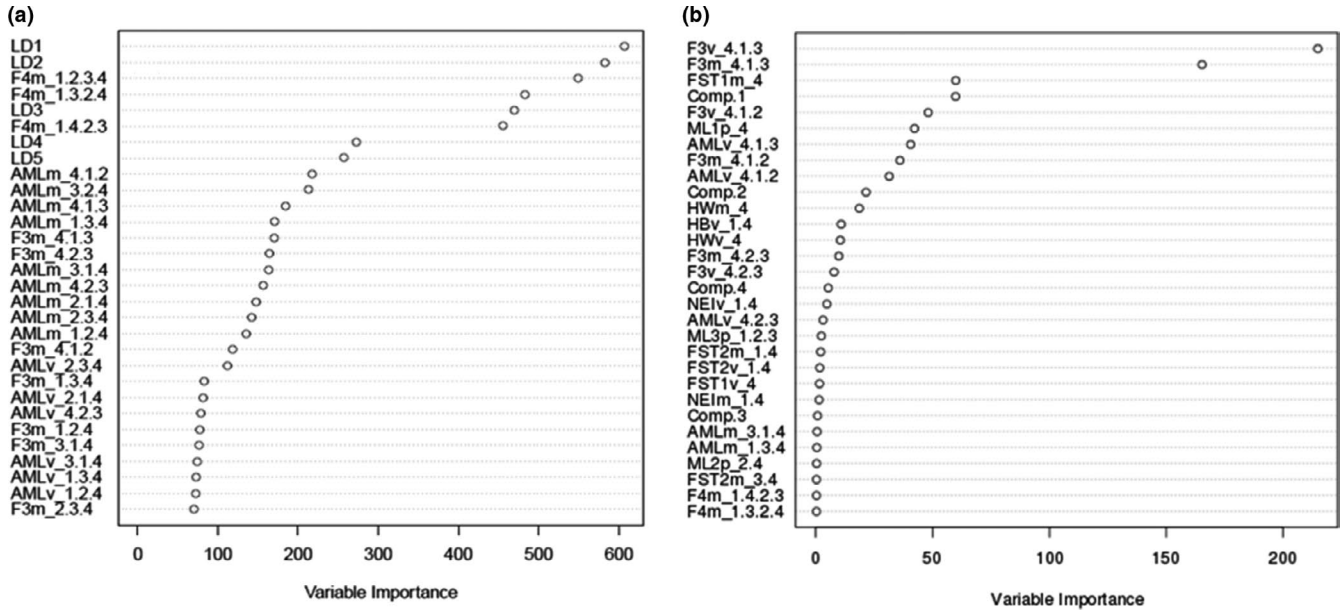

FIGURE 3 Contributions for the PoolSeq data analyses of the 30 most informative statistics to the random forest when choosing among scenarios considered separately (a) and when estimating the parameter t_1/N_4 under scenario 3 (b). The variable importance of each statistics is computed as the mean decrease 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. For each variable, the sum of the impurity decrease across every tree of the forest is accumulated every time that variable is chosen to split a node. The sum is divided by the number of trees in the forest to give an average. The scale is irrelevant: only the relative values matter. The variable importance was computed for each of the 130 summary statistics provided by DIYABC Random Forest, plus the LDA axes for scenario choice (denoted LD) or the PLS components 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 S3 for an illustration of the contributions of the most informative statistics when choosing among the two groups of scenarios and when estimating the parameters r_a, t_1 and N_4.

in the field, it is not surprising that scenario 2, which includes a single out-of-Africa colonization event giving an ancestral out-of-Africa population with a secondary split into one European and one East Asian population lineage and a recent genetic admixture of Americans of African origin with their African ancestors and European individuals, was selected (e.g., Bryc et al., 2015). LDA axes, four-sample and three-sample *f*-statistics, and three-sample coefficients of admixture (i.e., AML statistics) were among the most informative statistics of the feature vector to discriminate scenarios (Figure S6). Traditional ABC methods provided qualitatively similar results, but precision metrics were poorer compared to those obtained using ABC Random Forest. The posterior probability of the selected admixture group of scenarios was 0.663 with the ABC rejection method (global prior error rate = 0.162) and 1.000 with the ABC multinomial logistic method (global prior error rate = 0.016). The posterior probability of scenario 2 was 0.369 with the rejection method (global prior error rate = 0.321) and 1.000 with the multinomial logistic method (global prior error rate = 0.125). We observed substantial instability of the posterior probability of the best scenario as we found that, when using different threshold for selecting the closest simulated data sets, the posterior probability was always equal to 1.000 but was sometimes associated to a different scenario than scenario 2 (results not shown).

We then focused on scenario 2 under which we estimated the admixture rate (r_a) associated to American individuals of African ancestry. Using DIYABC Random Forest and including two PLS axes, the estimations for r_a were equal to 0.230 (median) and 0.229 (mean), with 95% CIs of 0.201 and 0.261. Without PLS axes, similar estimations were obtained (median = 0.230, mean = 0.231, and 95% CIs [0.203, 0.264]). The latter estimates lay well within previous estimates of the mean proportion of genes of European ancestry within African American individuals, which typically ranged from 0.070 to 0.270 (with most estimates around 0.200), depending on individual exclusions, the population samples and sets of genetic markers considered, as well as the evolutionary models assumed and inferential methods used (reviewed in Bryc et al., 2015). Global (prior) NMAE values were equal to 0.025 for all types of ABC-RF computation (i.e., with or without PLS axes and when computed on both median and mean). Local (posterior) NMAE were slightly smaller with PLS axes (0.030 and 0.032 with and without PLS axes, respectively). The 90% coverages were equal to 0.994 with and without PLS axes. The most informative statistics included the first PLS component, three-sample AML statistics with the population ASW as target, the pairwise-population statistics (F_{ST} and Nei's distance) including the population ASW, and MLp (proportion of monomorphic loci) statistics (Figure S6).

Traditional ABC methods provided estimations of r_a close to those obtained with ABC random forest: (i) median = 0.240, mean = 0.253 and a large 95% CIs (0.078, 0.445) for ABC rejection, and (ii) median = mean = 0.241 and a very narrow 95% CIs = (0.240, 0.243) for ABC logRidge. NMAE values were large for ABC rejection (0.276 and 0.299 for NMAE on median and mean, respectively) and small for ABC logRidge (0.023 for both NMAE on median and mean). The 90% coverage was equal to 0.96 for ABC rejection and was particularly narrow (i.e., 0.71) for ABC logRidge.

4 | DISCUSSION

Population genetics is now poised for an explosion in the use of SML approaches (Schrider & Kern, 2018). In this context, any effort to create self-contained, efficient, and user-friendly software packages capable of performing the entire workflow associated to SML methods would streamline such methods and make them more accessible to researchers, especially for nonspecialist users. For this purpose, we developed the package DIYABC Random Forest v1.0 which integrates, within a user-friendly interface, a set of methods to simulate training sets for various types of molecular data under custom evolutionary scenarios, encode both the simulated and observed (target) data as large size feature vectors (summary statistics), train RF algorithms, apply them on observed data point(s), and assess their performance in term of prediction (using various metrics to evaluate error and accuracy). We illustrate the main potentialities and functionalities of DIYABC Random Forest v1.0 through the treatments of pseudo-observed and real data sets corresponding to PoolSeq and IndSeq SNP data sets. Our results indicate that SML methods such as RF show great promise in scenario selection and demographic estimation using genetic data and we argue that they may soon be the preferred choice over alternative methods based on traditional ABC.

The first advantage of RF is that, given a pool of different metrics available (here various nonindependent summary statistics and their linear combinations), the method extracts the maximum information from the entire set of the proposed component of the feature vector. This avoids the arbitrary choice of a subset of components, which is often applied in ABC analyses. It also minimizes the curse of dimensionality whereby accuracy of inferences decreases as the number of summary statistics grows. As a matter of fact, SML methods such as RF can handle many statistics, even if they are strongly correlated and/or unnecessary (i.e., virtually noninformative), with a limited impact on the performance of the method (Marin et al., 2018; Raynal et al., 2019). In practice, and in contrast to traditional ABC methods, SML methods perform better when the input data have a large number of features, in what is commonly called the "blessing of dimensionality" (e.g., Anderson et al., 2014; Breiman, 2001). In agreement with this, inputs that consist of thousands of variables have been used with great success (e.g., Amit & Geman, 1997; Chen et al., 2013; and unpublished results obtained using feature vectors of >10,000 summary statistics to treat SNP data sets under complex evolutionary scenarios with DIYABC Random Forest v1.0).

Regarding the composition of the feature vector, defining informative statistics to be included in this vector remains an important issue of any SML method. We have implemented a new set of summary statistics to better extract the genetic information contained in the selectively neutral and independent SNP markers simulated in DIYABC Random Forest v1.0. For both scenario choice and parameter estimation, our results show, at least in the evolutionary contexts we explored, the high level of information content of four-populations and three-populations *f*-statistics (Patterson et al., (2012), as well as the related three-sample AML statistics (Cornuet et al., 2014). We found that inferences were more accurate with this new set of SNP summary statistics than with the one previously proposed in DYABC v2.1.0 (Cornuet et al., 2014). For instance, comparative treatments based on the pseudo-observed IndSeq data set generated under scenario 3, show that error levels were substantially lower and accuracy higher with the new set of SNP summary statistics (results not shown). The addition into the feature vector of linear combinations of statistics (LDA and PLS axes for scenario choice and parameter estimation, respectively) also globally improved our statistical inferences. While the inferential gain was systematic and substantial for LDA axes, we found that including PLS axes in the RF vector feature improved parameter estimation in a heterogeneous way, with a negligible gain in some cases.

The second advantage of SML methods such as RF is that they naturally use all of the simulations to learn the mapping of data to scenarios and/or parameters. This contrasts to the rejection step of ABC methods which precludes an optimal use of the data sets that are not retained. This advantage remains although work has been done to retain more of the simulations in ABC, for instance by weighing their influence on parameter estimation according to their similarity to the observed data (e.g., Blum & François, 2010). Consequently, the computing effort is considerably reduced for RF, as the method requires a substantially smaller training set compared to ABC methods (e.g., a few thousand simulated data sets versus hundreds of thousands of simulations per scenario for most ABC approaches; Blum & François, 2010; Fraimout et al., 2017; Pudlo et al., 2016; Raynal et al., 2019). Given the ever-increasing dimensionality of modern genetic data generated using NGS technologies, this is a particularly appealing property of SML methods. Moreover, it is worth noting that DIYABC Random Forest v1.0 relies on out-of-bag prediction to evaluate the error and accuracy of inferences, so that no additional potentially costly simulations of test data sets are necessary for this purpose.

RF is often considered as a "tuning-free" method in the sense that it does not require meticulous calibrations. This represents an important advantage of this method, especially for nonexpert users. On the opposite, ABC methods require calibration to optimize their use, such calibration being time consuming when different levels of tolerance are tested and/or used. In practice, we nevertheless advise users to consider several check points, before finalizing inferential treatments using DIYABC Random Forest v1.0. These are detailed in Appendix S3.

Various SML methods have been recently developed (e.g., Schrider & Kern, 2018; Wang et al., 2021). In particular, neural networks are machine learning methods which are used increasingly in population genetics, often under the term "deep learning" (Sheehan & Song, 2016), and sometimes using an ABC framework (Mondal et al., 2019). Deep learning, with its incredibly flexible input and output structure, is expected to be an important area of future research in many different fields including population genetics (e.g., Angermueller et al., 2016; Flagel et al., 2018; Schrider & Kern, 2018; Sheehan & Song, 2016). In contrast to RF, deep learning methods are not tuning-free and often require meticulous calibrations, including the specification of the number of layers composing the neural network, as well as thorough investigation of the regularization parameter of the cost function. Moreover, deep learning methods require data sets of larger size and substantially larger computing resources than RF. We hence believe that RF remains one of the most competitive SML methods when no tuning of parameters is desired. The RF method remains particularly attractive for nonexpert machine-learning users, especially when it is embedded in an integrative user-friendly interfaced program such as DIYABC Random Forest 1.0.

In conclusion, although SML approaches are revolutionizing many fields, their use in population genetics inference is still in its infancy (Schrider & Kern, 2018). However, the recent successes of SML approaches in the latter scientific field demonstrate that they have the potential to revolutionize the practice of population genetic data analysis. In particular, SML methods such as RF may soon be the preferred choice over ABC method in scenario selection and demographic estimation, especially when analysing multiple complex scenarios and large-size data sets. In this context, DIYABC Random Forest v1.0 provides an integrative operational solution streamlining the entire workflow to applying RF methods to various types of population genetic data. We believe that because of the general properties of the implemented RF methods and the large set of summary statistics available for SNP data, DIYABC Random Forest v1.0 represents a useful resource to make efficient inferences about population genetic history from high dimensional genetic data sets, as typically obtained from NGS technologies.

ACKNOWLEDGEMENTS

This work was supported by funds from the French Agence National pour la Recherche (projects SWING ANR-16-CE02-0015-01, GANDHI ANR-20-CE02-0018 and ABSint ANR-18-CE40-0034), and the LabEx NUMEV (NUMEV, ANR10-LABX-20). We thank Pierre Pudlo for useful discussions, and Jean-Marie Cornuet and Alex Dehne Garcia for computer code expertise at the onset of the ABC Random Forest project.

AUTHOR CONTRIBUTIONS

François-David Collin, Louis Raynal, Jean-Michel Marin and Arnaud Estoup were responsible for the conceptualization, F--D.C., L.R., J-M.M. and A.E.; François-David Collin was responsible for the core program coding and Ghislain Durif for the interface coding; Mathieu Gautier, Renaud Vitalis and Arnaud Estoup provided the new SNP summary statistics: François-David Collin, Eric Lombaert and Arnaud Estoup were responsible for program debugging and testing: Arnaud Estoup was responsible for the example data set analysis, Arnaud Estoup wrote the original draft of the manuscript; François-David Collin, Louis Raynal, Ghislain Durif, Mathieu Gautier, Renaud Vitalis, Eric Lombaert, Jean-Michel Marin and Arnaud Estoup wrote, reviewed and edited the manuscript, Jean-Michel Marin and Arnaud Estoup were responsible for funding acquisition.

DATA AVAILABILITY STATEMENT

For the pseudo-observed and real PoolSeq and IndSeq data sets used as examples in this paper, we provide the corresponding pseudo-observed data sets (read numbers or genotype data and summary statistics: i.e., <file_name.snp>and statobsRF.txt), the headerRF files (headerRF.txt) and the training set files (reftableRF.bin) at https://github.com/diyabc/MER_publication_materials/tree/main/MER_2021_DATASET_EXAMPLES.

ORCID

Arnaud Estoup https://orcid.org/0000-0002-4357-6144

REFERENCES

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, *9*, 1545–1588.

Anderson, J., Belkin, M., Goyal, N., Rademacher, L., & Voss, J. (2014). The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. *Proceedings of The 27th Conference on Learning Theory* (pp. 1135–1164). PMLR 35.

Angermueller, C., Parnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651

Beaumont, M. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.PMID: 12524368.

Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology*, *19*(13), 2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x

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

Blum, M. G. B., & François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, *20*(1), 63–73. https://doi.org/10.1007/s11222-009-9116-0

Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, *28*(2), 189–208. https://doi.org/10.1214/12-STS406

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

Bryc, K., Durand, E. Y., Macpherson, M., Reich, D., & Mountain, J. L. (2015). The genetic ancestry of African Americans, Latinos, and

European Americans across the United States. *American Journal of Human Genetics*, *96*(1), 37–53. https://doi.org/10.1016/j.ajhg.2014.11.010

Chapuis, M.-P.- R., Raynal, L., Plantamp, C., Meynard, C. N., 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.org/10.1111/mec.15663. Previous version reviewed and recommended by Peer Community in Evolutionary Biology, bioRxiv, 671867. https://doi.org/10.24072/pci.evolbiol.100091

Chen, D., Cao, X., Wen, F., & Sun, J. (2013). Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3025–3032). IEEE.

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J. M., & Estoup, A. (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, *30*(8), 1187–1189. https://doi.org/10.1093/bioinformatics/btt763

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

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., Lombaert, E., Marin, J.-M., Robert, C., Guillemaud, T., Pudlo, P., & Cornuet, J.-M. (2012). Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Molecular Ecology Resources*, *12*(5), 846–855. https://doi.org/10.1111/j.1755-0998.2012.03153.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.

Flagel, L., Brandvain, Y., & Schrider, D. R. (2018). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, *36*(2), 220–238. https://doi.org/10.1093/molbev/msy224

Fraimout, A., Debat, V., Fellous, S., Hufbauer, R. A., Foucaud, J., Pudlo, P., Marin, J. M., Price, D. K., Cattel, J., Chen, X., Depra, M., Duyck, P. F., Pudlo, P., Guedot, C., Kenis, M., Pudlo, P., Kimura, M. T., Loeb, G., Loiseau, A., & Estoup, A. (2017). Deciphering the routes of invasion of *Drosophila suzukii* by means of ABC Random Forest. *Molecular Biology and Evolution*, *34*(4), 980–996. https://doi.org/10.1093/molbev/msx050

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., & Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766–3779. https://doi.org/10.1111/mec.12360

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, *210*(1), 31–330. https://doi.org/10.1534/genetics.118.300900

Hudson, R. R. (1993). The how and why of generating gene genealogies. In N. Takahata, & A. G. Clark (Eds.), *Mechanisms of Molecular Evolution* (pp. 23–36). Sinauer Associates.

Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Leblois, R., Gautier, M., Rohfritsch, A., Foucaud, J., Burban, C., Galan, M., & Kerdelhué, C. (2018). Deciphering the demographic history of allochronic differentiation in the pine processionary moth *Thaumetopoea pityocampa*. *Molecular Ecology*, *27*(1), 264–278. https://doi.org/10.1111/mec.14411

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332. https://doi.org/10.1038/nrg3920

Marin, J.-M., Pudlo, P., Estoup, A., & Robert, C. P. (2018). Likelihood-free model choice. In handbook of approximate Bayesian computation. In S. Sisson, Y. Fan & M. Beaumont (Eds), *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315117195

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999. https://doi.org/10.5555/1248547.1248582

Mondal, M., Bertranpetit, J., & Lao, O. (2019). Approximate Bayesian Computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, *10*, 246. https://doi.org/10.1038/s41467-018-08089-7

Nei, M. (1972). Genetic distance between populations. *American Naturalist*, *106*, 283–292. https://doi.org/10.1086/282771

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

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.org/10.1093/bioinformatics/btv684

Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., Bertranpetit, J., & Engelken, J. (2015). Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, *31*(24), 3946–3952. https://doi.org/10.1093/bioinformatics/btv493

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.org/10.1093/bioinformatics/bty867

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749–763. https://doi.org/10.1038/nrg3803

Schrider, D. R., Ayroles, J., Matute, D. R., & Kern, A. D. (2018). Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genetics*, *14*(4), e1007341. https://doi.org/10.1371/journal.pgen.1007341

Schrider, D. R., & Kern, A. D. (2016). S/HIC: robust Identification of soft and hard sweeps using machine learning. *PLoS Genetics*, *12*(3), e1005928. https://doi.org/10.1371/journal.pgen.1005928.

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computional Biology*, *12*(3), e1004845. https://doi.org/10.1371/journal.pcbi.1004845

Smith, M. L., & Carstens, B. C. (2020). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, *74*(2), 216–229. https://doi.org/10.1111/evo.13878

Smith, M. L., Ruffley, M., Espíndola, A., Tank, D. C., Sullivan, J., & Carstens, B. C. (2017). Demographic model selection using random forests and the site frequency spectrum. *Molecular Ecology*, *26*(17), 4562–4573. https://doi.org/10.1111/mec.14223

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

Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H. H., Mathieson, I., & Mathieson, S. (2021). Automatic inference of demographic parameters using generative adversarial networks. *Mol Ecol Resour.*, 1–17. https://doi.org/10.1111/1755-0998.13386

Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, *206*(4), 2085–s2103. https://doi.org/10.1534/genetics.116.198424

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Collin F-D, Durif G, Raynal L, et al. Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. *Mol Ecol Resour.* 2021;00:1–16. https://doi.org/10.1111/1755-0998.13413