Journal de la Société Française de Statistique
Vol. 159 No. 3 (2018)

Model choice using Approximate Bayesian Computation and Random Forests: analyses based on model grouping to make inferences about the genetic history of Pygmy human populations

Titre: Choix de modèles par calcul bayésien approximé et forêts aléatoires : analyses basées sur le groupement de modèles pour inférer l'histoire génétique des populations Pygmées

Arnaud Estoup[1,2] , Louis Raynal[3] , Paul Verdu[4] and Jean-Michel Marin[3,2]

Abstract: In evolutionary biology, simulation-based methods such as Approximate Bayesian Computation (ABC) are well adapted to make statistical inferences about complex models of natural population histories. Pudlo et al. (2016) recently developed a novel approach based on the Random Forests method (RF): the ABC-RF algorithm. Here we present the results of analyses based on ABC-RF to make inferences about the history of Pygmy human populations from Western Central Africa from a microsatellite genetic dataset. A noticeable novelty of the statistical analyses presented here is the application of ABC-RF methodology to make model choice on predefined groups of models. We formalized eight complex evolutionary scenarios which incorporate (or not) three major events: (i) whether there exists an ancestral common Pygmy population, (ii) the possibility of introgression/migration events between Pygmy and non-Pygmy populations, and (iii) the possibility of a change in size in the past in the non-Pygmy African population. We show that our grouping approach allows disentangling with strong confidence the main evolutionary events characterizing the population history of interest. The selected final scenario corresponds to a common origin of all Western Central African Pygmy groups, with the ancestral Pygmy population having diverged, with asymmetrical genetic introgression, from a demographically expanding non-Pygmy population.

Résumé : En biologie évolutive, les méthodes d'inférence fondées sur la simulation, comme le calcul bayésien approché (ABC), sont particulièrement adaptées pour traiter les modèles complexes. Pudlo et al. (2016) ont récemment développé une nouvelle approche basée sur les forêts aléatoires (RF) denommée ABC-RF. Nous présentons ici les résultats d'analyses basées sur la méthodologie ABC-RF pour inférer l'histoire des populations humaines pygmées d'Afrique centrale occidentale à partir d'un ensemble de données génétiques issues de marqueurs microsatellites. Une nouveauté notable de nos analyses statistiques concerne l'application des techniques ABC-RF pour choisir des groupes prédéfinis de modèles. Nous avons formalisé huit scénarios évolutifs complexes intégrant (ou non) trois événements majeurs : (i) l'existence d'une population pygmée ancestrale commune, (ii) la possibilité d'événements de mélange génétique / migration entre populations pygmées et non-pygmées, et (iii) la possibilité d'un changement de taille dans le passé de la population non pygmée. Nous montrons que notre approche de regroupement de scénarios permet de discerner avec une forte confiance les principaux événements évolutifs qui caractérise l'histoire populationelle d'intérêt. Le scénario sélectionné final corresponds à une origine commune de tous les groupes pygmées d'Afrique centrale occidentale, la population pygmée ancestrale ayant divergé, avec des mélanges génétiques asymétriques, d'une population non-pygmée en expansion démographique.

[1] CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France.
[2] IBC, Univ Montpellier, CNRS, Montpellier, France. E-mail: arnaud.estoup@inra.fr
[3] IMAG, Univ Montpellier, CNRS, Montpellier, France. E-mail: louis.raynal@umontpellier.fr and E-mail: jean-michel.marin@umontpellier.fr
[4] CNRS-MNHN-Univ. Paris Diderot-Sorbonne Paris Cité, UMR 7206, Ecoanthropology and Ethnobiology, Paris, France. E-mail: paul.verdu@mnhn.fr

 ISSN: 2102-6238

Keywords: Approximate Bayesian Computation, evolutionary biology, genetic variation, microsatellites, model selection, population genetics, Random Forests
Mots-clés : calcul bayésien approché, biologie évolutive, variations génétiques, microsatellites, sélection de modèle(s), génétique des populations, forêts aléatoires
AMS 2000 subject classifications: 62F15, 62P10

1. Introduction

Approximate Bayesian computation (ABC; Beaumont et al., 2002) represents 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. The method arose in population genetics (Tavaré et al., 1997; Beaumont et al., 2002) and is increasingly used in other fields, including epidemiology, system biology, ecology, agent-based modeling (reviewed in Beaumont, 2010) and population linguistics (Thouzeau et al., 2017). It undoubtedly widens the realm of models for which statistical inference can be considered. We will not detail here the general statistical features of ABC as they have been reviewed in previous publications (e.g. Bertorelle et al., 2010; Csilléry et al., 2010; Beaumont, 2010; Marin et al., 2012; Sunnåker et al., 2013). Recap on the main principles and algorithms of the methods are provided, however, in section 2.4.

Full-likelihood methods have been developed to analyse molecular data (typically DNA variation) in population genetics for some models. In a frequentist setting, they typically rely on importance sampling approximations of the likelihood: Stephens and Donnelly (2000); De Iorio and Griffiths (2004a,b); De Iorio et al. (2005); Rousset and Leblois (2007, 2012); Merle et al. (2017) and, in the Bayesian paradigm, on MCMC approximations of the posterior distributions Drummond and Rambaut (2007); Drummond and Bouckaert (2015).

All these strategies have difficulties on large datasets and are only able to consider a limited range of model structures (Beaumont, 2010). By contrast ABC is very flexible and is hence well adapted to investigate complex models of species and population history, which often involve serial or independent divergence events, changes of population sizes, and genetic admixture or migration events (e.g. Fagundes et al., 2007; Lombaert et al., 2010). In an ABC framework, such events can be easily simulated and hence incorporated into different historical and demographic evolutionary models (often called scenarios in population genetics) that can be formally tested against each other with respect to the observed data. The method can also be used to estimate the posterior distributions of demographic parameters of interest, such as divergence times, admixture rates, and effective population sizes, in a given scenario (usually the most likely one; e.g. Beaumont et al., 2002; Bertorelle et al., 2010). In practice, ABC users in the field of population and evolutionary biology can base their analyses on simulation programs that have recently been developed to provide non-specialist users with more integrated computational solutions varying in user-friendliness (see the list of ABC packages and toolboxes in Beaumont, 2010; Csilléry et al., 2010).

The formal description of a finite set of possible models lays the foundation for ABC analyses of evolutionary histories of natural populations (e.g. Estoup and Guillemaud, 2010). Choosing among models is hence a central statistical problem to be overcome when making inferences about population histories from molecular data. Both theoretical arguments and simulation experiments indicate that approximate posterior probabilities estimated from standard ABC analyses for the modeled scenarios can be inaccurate, even though the models being compared can

 ISSN: 2102-6238

still be ranked appropriately using numerical approximation (Robert et al., 2011). To overcome this problem, Pudlo et al. (2016) developed a novel approach based on a machine learning tool named Random Forests (RF; Breiman, 2001), which allows selecting among the complex models covered by ABC algorithms. This approach enables efficient discrimination among models and estimation of posterior probability of the best model, while being computationally less intensive than more standard ABC methods. Recent methodological comparisons indicate that ABC-RF and more standard ABC methods show concordant results when inferences are based on a large number of simulated datasets, but ABC-RF out-performs standard ABC methods when using a comparable and more manageable (i.e. smaller) number of simulated datasets, especially when analysing multiple complex models and large datasets (Pudlo et al., 2016; Fraimout et al., 2017).

We invite readers to consult Pudlo et al. (2016) to access to detailed statistical descriptions and testing of the ABC Random Forests (ABC-RF) method; see also section 2.4 for some mathematical and algorithmic insights about the method. Briefly and verbally, Random Forests (RF) are currently considered as one of the major state-of-art algorithm for classification or regression. It is an algorithm that learns from a database how to predict a variable from a possibly large set of covariates. In our context the database is the reference table which includes a given number of datasets that have been simulated for different models using parameter values drawn from prior distributions, each dataset being summarized with a pool of statistics (i.e. the covariates). 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 posterior probability of the best scenario). Each tree is built by using the information provided by a bootstrap sample of the database and manages to capture one part of the dependency between the output and the covariates. Based on these trees which are individually poor to predict the output, an ensemble learning technique such as RF aggregates their predictions to increase predictive performances to a high level of accuracy in favorable contexts (Breiman, 2001). The idea that underlies these methods is to learn from simulations coming from a generative model Schrider and Kern (2016); Sheehan and Song (2016). In our opinion, a relevant way to take benefits from generative model simulations is the Bayesian paradigm and then ABC type approach.

In this study, we present a set of statistical analyses using ABC-RF applied on a molecular (DNA) dataset obtained from Western Central African Pygmy and non-Pygmy populations. Central Africa and the Congo Basin are currently peopled by the largest group of forest hunter-gatherer populations worldwide, which have been historically called “Pygmies” in reference to the mythical population of short stature described by the ancient Greek poet Homer (Hewlett, 2014). Each Central African Pygmy group is in the neighborhood of several sedentary agricultural populations (hereafter called “non-Pygmies”) with whom they share complex sociocultural and economic relationships, including social rules regulating intermarriages between communities (Verdu et al., 2013; Hewlett, 2014). Due to the lack of ancient human remains in the equatorial forest, the origins of Pygmies and neighboring non-Pygmies remains largely unknown (Cavalli-Sforza et al., 1994; Cavalli-Sforza and Feldman, 2003). Moreover, Western colonizers from the 19th century somewhat arbitrarily collapsed into a single “Pygmy” group more than 20 populations that were, and still are, culturally and geographically isolated in reality, which further clouded our understanding of evolutionary relationships among these populations. Thus, the questions of (i) whether all Central African Pygmy populations have a common or an in-

 ISSN: 2102-6238

dependent origin, and (ii) whether they exchange genes through introgression/migration among one another and from neighboring non-Pygmies, were still largely debated in the anthropology and ethnology communities (Cavalli-Sforza, 1986; Hewlett, 2014; Verdu et al., 2009).

To tackle these questions, Verdu et al. (2009) genotyped strongly variable genetic markers (namely microsatellite DNA loci; Estoup et al., 2002) in a dense sample of non-Pygmy and neighboring Pygmy populations from Western Central Africa, and used standard ABC methods (Beaumont et al., 2002; Estoup et al., 2012) to make statistical inferences. In the present study, we considered the dataset of Verdu et al. (2009) and reanalysed it using ABC-RF. A noticeable novelty of the statistical analyses presented here includes the application of ABC-RF algorithms to make scenario choice on predefined groups of models, in addition to standard analyses on the whole set of (separated) scenarios to be compared. As a matter of fact, genetic markers such as microsatellites are informative for deciphering key evolutionary events that shape genetic variation in natural populations, such as a common or an independent origin of a given set of populations, the presence or absence of genetic introgression/migration among populations, as well as major changes in effective population size, the latter feature being strongly suspected in non-Pygmy African populations (e.g. Lombaert et al., 2010; Verdu et al., 2009). Under an ABC framework, such events can be modeled explicitly hence defining different scenarios that can be grouped based on the type(s) of evolutionary events that have been incorporated into them. We show here that groups of scenarios (for instance scenarios including a common origin of Pygmy populations versus scenarios including an independent origin of those populations) can be formally and advantageously compared using ABC-RF, instead of considering all scenarios separately. Such grouping approach in scenario choice is of great interest to disentangle with strong confidence the main evolutionary events characterizing the history of natural populations.

2. Observed and simulated datasets

2.1. Observed dataset

The analysed dataset includes the genotyping at 28 microsatellite loci of 400 unrelated individuals from four Pygmy groups (i.e. the Baka, Bezan, Kola and Koya; 29 to 32 individuals per group), neighboring non-Pygmy individuals (194 individuals) from Cameroon and Gabon (Western Central Africa) (see Figure 1 and Table S1 in Verdu et al., 2009, for details about geographic location of population samples and their genetic grouping). The exact dataset used in the present study is available upon request to some authors (PV or AE), following ethical, informed consent and IRB appropriateness.

2.2. Models, groups of models, and parameters

We considered the same set of eight complex evolutionary scenarios, as in Verdu et al. (2009). These scenarios with their historical and demographic parameters are represented in Figure 1, following the notation of Verdu et al. (2009). See also Appendix A for a detailed description of the model parameters and their prior distributions.

 ISSN: 2102-6238

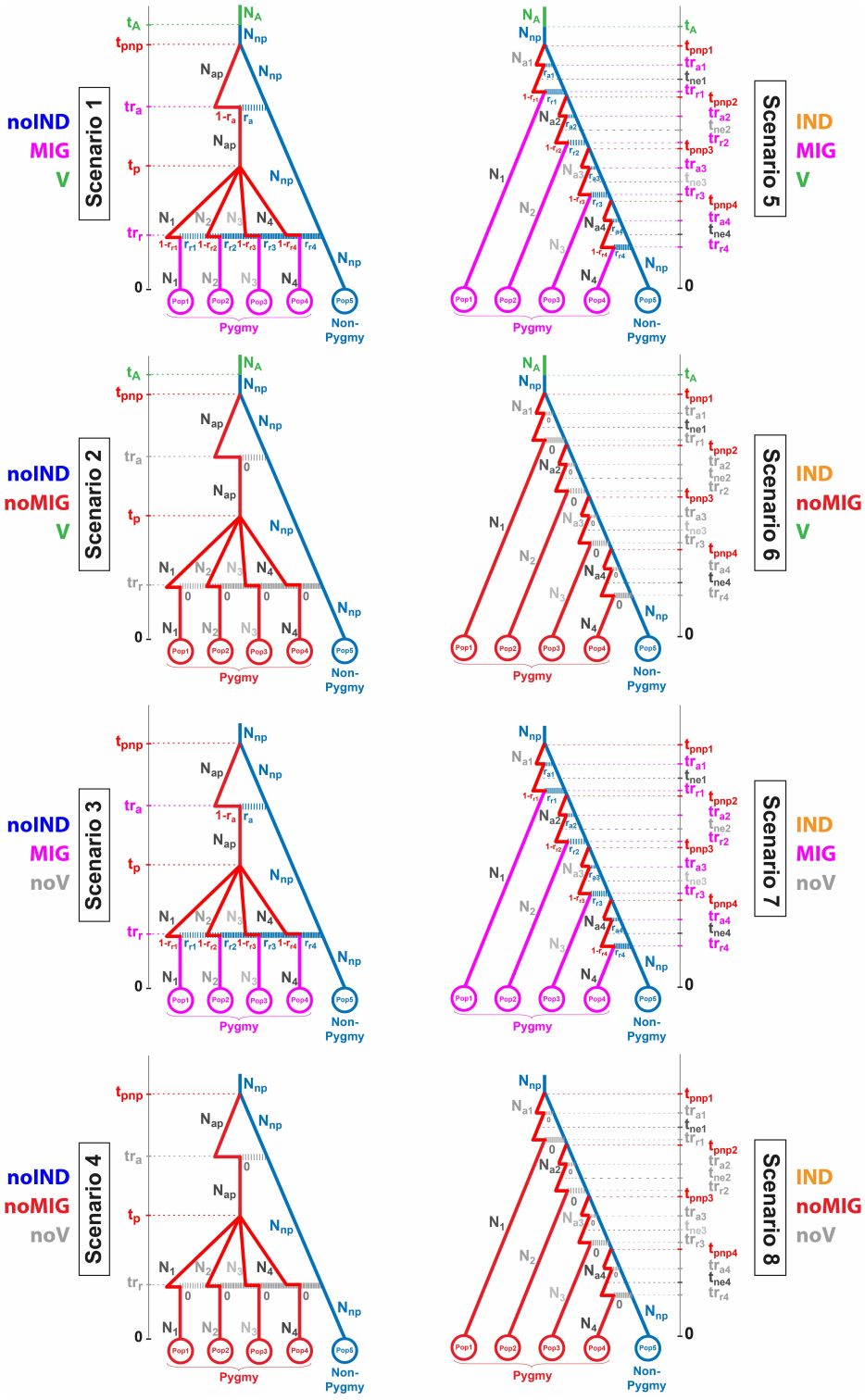


FIGURE 1. *Eight complex competing scenarios of origin and diversification of Pygmy populations from Western Central Africa*

 ISSN: 2102-6238

Note for Figure 1: The scenarios and parameters are similar to those in Verdu et al. (2009) and follow the same notation. Scenario 1 - 4 (group G1A; labelled noIND in the figure) corresponds to a common origin of Pygmy populations that diversified from a single ancestral Pygmy population at time t_p. The ancestral Pygmy population itself diverged at time t_{pnp} from the non-Pygmy African population. Scenarios 5-8 (group G1B; labelled IND) correspond to an independent origin of Pygmy groups that independently diverged from the non-Pygmy African population at times t_{pnpi}. We further incorporated in four of the eight scenarios (group G2A corresponding to scenarios 1, 3, 5 and 7; labelled MIG) both a recent and an ancient event of introgression/migration (cf. parameters tr_i and r_i) from the non-Pygmy African population into each Pygmy lineage independently. All introgression rates (r_i) were set to zero in scenarios 2, 4, 6 and 8 (group G2B; labelled noMIG). Finally, in contrast to scenarios 3, 4, 7 and 8 (group G3B; labelled noV), scenarios 1, 2, 5 and 6 (group G3A; labelled V) include a potential stepwise change of effective population size that occurred in the non-Pygmy African population at time t_A. For all scenarios, N_i correspond to the effective population size of population i. For scenarios 5-8 (group G1B), divergence times were drawn independently for each Pygmy lineage and thus, the order in which these lineages split is not pre-defined. See main text and Appendix A for details regarding prior distributions of parameters.

These eight scenarios include different combinations of three main types of evolutionary events debated in the anthropology community.

First evolutionary event: scenario group G1A (scenarios 1, 2, 3 and 4) correspond to a common evolutionary history of the four Pygmy groups, where they all originate from the same ancestral Pygmy population which initially diverged from the non-Pygmy population in a more remote past. Scenario group G1B (scenarios 5, 6, 7 and 8) correspond to an independent evolutionary history of the four Pygmy groups, where they each originate independently from the non-Pygmy African population.

Second evolutionary event: scenario group G2A (scenarios 1, 3, 5 and 7) include or exclude (group G2B corresponding to scenarios 2, 4, 6 and 8) the possibility of recent and ancient asymmetrical introgression/migration events from the African non-Pygmy gene-pool into each Pygmy population, as was already suggested by previous anthropological and genetic studies (e.g. Cavalli-Sforza, 1986; Hewlett, 1996; Destro-Bisol et al., 2004).

Third evolutionary event: scenario group G3A (scenarios 1, 2, 5 and 6) include or exclude (group G3B; scenarios 3, 4, 7 and 8) the possibility of a change of population size in the non-Pygmy African population.

It is worth stressing that there are obviously a number of other possible scenarios that might possibly fit the data just as well as if not better than the best scenario found among the present finite set of compared scenarios.

2.3. *Priors*

We chose an equiprobable prior for the compared scenarios or groups of scenarios. Similarly to Verdu et al. (2009) and references therein, we chose flat prior distributions for all demographic

 ISSN: 2102-6238

parameters specified in Figure 1 (see Appendix A for details): uniform distributions bounded between 100 and 10,000 diploid individuals for all Pygmy populations and ancestral population sizes (N_i, N_{ap}, N_{ai}, and N_A, with i between 1 and 4), between 1,000 and 100,000 for the African non-Pygmy population (N_{np}). Priors were drawn from uniform distributions between 1 and 5,000 generations for all divergence times (t_p, t_{pnp}, t_{pnpi}, with i between 1 and 4), for the population size variation times (t_{nei}, with i between 1 and 4) and for the times of "ancient" introgression of non-Pygmy genes into ancestral Pygmy lineages (tr_a, and tr_{ai}, with i between 1 and 4). For the time of change in effective population size in the non-Pygmy population (t_A), considered only in the scenario group G3A (i.e. scenarios 1, 2, 5 and 6), we drew our priors from a uniform distribution bounded between 1 and 10,000 generations. For the "recent" introgression times from non-Pygmies into the Pygmy lineages, (tr_r, and tr_{ri}, with i between 1 and 4), we chose log-Uniform prior distributions bounded between 1 and 5,000 generations. For genetic markers (i.e. microsatellite loci), the parameters and associated prior distributions of mutation models and rates were the same as in Verdu et al. (2009) and Estoup et al. (2018).

2.4. *ABC-RF: main statistical principles and analysis of groups of models*

2.4.1. *Recap on ABC parameter inference*

As previously mentioned, we concentrate on the Bayesian parametric paradigm. Let $\mathbf{y}$ denote the observed dataset. For parameter inference, the goal is to sample from the posterior distribution

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta) f(\mathbf{y}|\theta)\,.$$

We are in an intractable likelihood context. Indeed, for our complex scenarios (i.e. models), despite the simplicity of the Kingman's coalescent and of the mutation processes, we cannot expect any simplification in the calculation of the likelihood. Approximate Bayesian Computation (ABC) is a technique that only requires being able to sample from the likelihood $f(\cdot|\theta)$.

This inferential setup stemmed from population genetics models, about 15 years ago, and population geneticists still significantly contribute to methodological developments of ABC. The standard likelihood-free rejection sampler (Rubin, 1984; Pritchard et al., 1999) is introduced in Algorithm 1.

Algorithm 1: Standard likelihood-free rejection sampler

1) For $i = 1, \ldots, N$

- **a)** Generate θ' from the prior distribution $\pi(\cdot)$
- **b)** Generate a pseudo dataset $\mathbf{z}$ from the likelihood $f(\cdot|\theta')$
- **c)** If $d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \varepsilon$, set $\theta_i = \theta'$ else return to **a)**

2) Return the θ_i's

In Algorithm 1, d corresponds to a well-chosen distance and η to a projection of the observed and pseudo dataset, $\mathbf{y}$ and $\mathbf{z}$, in a space of lower dimension. The summarization process is necessary to fight against the curse of dimensionality. The obtained sample $\theta_1, \ldots, \theta_N$ is approximately distributed from the posterior distribution. ε reflects the tension between computability

 ISSN: 2102-6238

and accuracy: if $\varepsilon \to \infty$, we get simulations from the prior and if $\varepsilon \to 0$, we get simulations from the posterior. The target that corresponds to Algorithm 1 is defined by

$$\pi_\varepsilon(\theta|\mathbf{y}) = \frac{\int \pi(\theta) f(\mathbf{z}|\theta) \mathbb{I}(\mathbf{z} \in A_{\varepsilon,\mathbf{y}}) \mathrm{d}\mathbf{z}}{\int \pi(\theta) f(\mathbf{z}|\theta) \mathbb{I}(\mathbf{z} \in A_{\varepsilon,\mathbf{y}}) \mathrm{d}\mathbf{z} \mathrm{d}\theta}$$

where $A_{\varepsilon,\mathbf{y}} = \{\mathbf{z}|d(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \varepsilon\}$ is the acceptance set. There are some different views of ABC approximations. Wilkinson (2013) shows that ABC is exact but for a different model to that intended and Blum (2010) emphasizes that ABC is a kernel smoothing approximation of the likelihood function:

$$\pi_\varepsilon(\theta|\mathbf{y}) = \frac{\pi(\theta) \int f(\mathbf{z}|\theta) K(d(\eta(\mathbf{z}), \eta(\mathbf{y}))) \mathrm{d}\mathbf{z}}{\int \pi(\theta) f(\mathbf{z}|\theta) K(d(\eta(\mathbf{z}), \eta(\mathbf{y}))) \mathrm{d}\mathbf{z} \mathrm{d}\theta}.$$

In practice, it is extremely difficult to set the value of ε before running some simulations and, practitioners use instead Algorithm 2.

Algorithm 2: Nearest-Neighbors likelihood-free scheme

1) For $j = 1, \ldots, H$

- **a)** Generate θ_j from the prior $\pi(\cdot)$
- **b)** Generate $\mathbf{z}$ from the model $f(\cdot|\theta_j)$
- **c)** Calculate $d_j = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$

2) Order the distances $d_{(1)}, \ldots, d_{(H)}$

3) Return the θ_j's that correspond to the N-smallest distances

Part **1)** of Algorithm 2 corresponds to the generation of what we call a reference table with H elements. In Algorithm 2, $N = \lfloor \alpha G \rfloor$ and ε corresponds to a quantile of the distances. That corresponds to Nearest-Neighbors approximation to the posterior distribution. This fact is emphasized in Biau et al. (2015). Algorithm 2 is intuitive, simple to implement and embarrassingly parallel.

However, as for Nearest-Neighbors methods, when the number of summary statistics increases, most of the simulations are at the boundary of the space and the curse of dimensionality occurs. Also, around these standard methods, there exists a lot of developments: regression adjustments (Beaumont et al., 2002; Blum and François, 2010); MCMC and sequential algorithms (Marjoram et al., 2003; Sisson et al., 2007, 2009; Beaumont et al., 2009; Del Moral et al., 2012); strategies to select the summary statistics (Blum et al., 2013); projection approach (Fearnhead and Prangle, 2012) and more recently use of machine learning techniques, (see for instance Raynal et al., 2018).

2.4.2. Recap on ABC model choice

Let now consider the more general case where M Bayesian parametric models are in competition $f_m(\mathbf{y}|\theta_m)$, $\pi_m(\theta_m)$, $m = 1, \ldots, M$. We define prior probabilities in the model space $\mathbb{P}(\mathscr{M} = m)$ and our target are the models's posterior probabilities

$$\mathbb{P}(\mathscr{M} = m|\mathbf{y}) \propto \mathbb{P}(\mathscr{M} = m) \int f_m(\mathbf{y}|\theta_m) \pi_m(\theta_m) d\theta_m.$$

 ISSN: 2102-6238

Considering the standard 0-1 symmetric loss function, the selected model is the one with the maximum of the models's posterior probabilities

$$\text{ArgMax}_m \quad \left\{ \mathbb{P}(\mathscr{M} = m) \int f_m(\mathbf{y}|\theta_m)\pi_m(\theta_m)d\theta_m \right\}.$$

$f_m(\mathbf{y}|\theta_m)$ are not available for each model in competition, therefore it is not possible to calculate the integrated likelihood (the evidence) $\int f_m(\mathbf{y}|\theta_m)\pi_m(\theta_m)d\theta_m$. To avoid these difficulties, Grelaud et al. (2009) have introduced a model choice version of the Nearest-Neighbors likelihood-free scheme, see Algorithm 3.

Algorithm 3: Nearest-Neighbors likelihood-free model choice scheme

1) For $j = 1, \ldots, H$
- **a)** Generate m_j from the prior $\mathbb{P}(\mathscr{M} = m)$
- **b)** Generate θ'_{m_j} from the prior $\pi_{m_j}(\cdot)$
- **c)** Generate $\mathbf{z}$ from the model $f_{m_g}(\cdot|\theta'_{m_j})$
- **d)** Calculate $d_j = d(\eta(\mathbf{z}), \eta(\mathbf{y}))$

2) Order the distances $d_{(1)}, \ldots, d_{(H)}$

3) Select the model using the majority rule among the N-smallest distances index set

Part **1)** of Algorithm 3 corresponds to the generation of the reference table in a model choice context. The problem is viewed as a classification question and is solved using Nearest-Neighbors classifiers. Due to the curse of dimensionality, the methodology associated to Algorithm 3 has major difficulties. Typically, to ensure reliability of the method, the number of simulations should be large and the number of summaries statistics small.

However, exploiting a large number of summary statistics is not an issue for some machine learning methods The idea of Pudlo et al. (2016) is to learn on a huge reference table using Random Forests (Breiman, 2001). Random Forests exhibit some theoretical guarantees for sparse problems (Biau, 2012; Scornet et al., 2015). This work stands at the interface between Bayesian inference and machine learning techniques. Algorithm 4 presents the ABC-RF strategy.

Algorithm 4: ABC-RF scheme

Input a reference table used as learning set, made of H elements, each one composed of a model index $m^{(h)}$ and d summary statistics. A possibly large collection of summary statistics can be used, from scientific theory to machine-learning alternatives.

$$\begin{bmatrix} m^{(1)} & \eta_1(\mathbf{z}^{(1)}) & \eta_2(\mathbf{z}^{(1)}) & \ldots & \eta_d(\mathbf{z}^{(1)}) \\ m^{(2)} & \eta_1(\mathbf{z}^{(2)}) & \eta_2(\mathbf{z}^{(2)}) & \ldots & \eta_d(\mathbf{z}^{(2)}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m^{(H)} & \eta_1(\mathbf{z}^{(H)}) & \eta_2(\mathbf{z}^{(H)}) & \ldots & \eta_d(\mathbf{z}^{(H)}) \end{bmatrix}$$

Learning construct a classification Random Forest $\hat{m}(\cdot)$ to infer model indexes

Output apply the Random Forest classifier to the observed dataset $\mathbf{y}$

For the observed dataset $\mathbf{y}$, the Random Forest classifier predicts the MAP model index. The predictor is good enough to select the most likely model but not to derive directly the associated

 ISSN: 2102-6238

posterior probabilities. Indeed, the frequency of trees associated with the majority model is not a proper substitute to the true posterior probability. We have

$$\mathbb{P}[\mathscr{M} = \hat{m}(\eta(\mathbf{y}))|\eta(\mathbf{y})] = 1 - \mathbb{E}\left[\mathbb{I}(\mathscr{M} \neq \hat{m}(\eta(\mathbf{y})))|\eta(\mathbf{y})\right] .$$

Therefore, as explained in Pudlo et al. (2016), this justifies using a second Random Forest in regression to estimate the posterior probability of the selected model; see Algorithm 5.

Algorithm 5: Scheme to estimate the posterior probability of the selected model

Input the value of $\mathbb{I}(\mathscr{M} \neq \hat{m}(\eta(\mathbf{z}))$ for the trained Random Forest and for all terms in the reference table using the out-of-bag classifiers

Learning construct a regression Random Forest $\hat{\mathbb{E}}(\cdot)$ to infer $\mathbb{E}\left[\mathbb{I}(\mathscr{M} \neq \hat{m}(\eta(\mathbf{z})))|\eta(\mathbf{z})\right]$

Output an estimate of the posterior probability of the selected model $m(\eta(\mathbf{y}))$

$$\hat{\mathbb{P}}[\mathscr{M} = \hat{m}(\eta(\mathbf{y}))|\eta(\mathbf{y})] = 1 - \hat{\mathbb{E}}\left[\mathbb{I}(\mathscr{M} \neq \hat{m}(\eta(\mathbf{y})))|\eta(\mathbf{y})\right]$$

Taking benefit of the out-of-bag trick, we can use the same reference table for the two Random Forests and avoid overfitting.

The R package abcrf implements Algorithms 4 and 5. Version 1.7.1 introduces a new very interesting feature which is at the core of our paper: the study of groups of models (scenarios). Using pre-defined groups, we create a new variable indexing the groups and use it to train the forests. We partition the set of initial models or a subset of it. Groups have empty intersections and the recorded datasets of the reference table that corresponds to unused models are not considered. For groups of models, we propose to use Algorithms 4 and 5 on the reference table altered in this way.

2.5. ABC-RF: analyses conducted on the observed dataset

Following Pudlo et al. (2016), ABC-RF treatments were processed on a reference table including 100,000 simulated datasets (i.e. 12,500 per scenario). Datasets were summarized using the whole set of summary statistics proposed by DIYABC (Cornuet et al., 2014) for microsatellite markers, describing genetic variation per population (e.g. number of alleles), per pair (e.g. genetic distance), or per triplet (e.g. coefficient of admixture) of populations, averaged over the 26 loci (see Appendix B for details about such statistics), plus the linear discriminant analysis (LDA) axes as additional summary statistics. The total number of summary statistics was 130 plus a single discriminant (LDA) axis when analysing pairwise groups of scenarios or seven linear discriminant analysis axes when analysing the eight models considered separately, as additional summary statistics. We checked that the number of simulated datasets of the reference table was sufficient by evaluating the stability of prior error rates (i.e. the probability to choose a wrong model when drawing model index and parameter values into priors) and posterior probabilities estimations on 80,000, 90,000 and 100,000 simulated datasets (results not shown). The number of trees in the constructed Random Forests was fixed to n = 1,000; see Appendix C for a justification of choosing such a number of trees per forest.

For each ABC-RF analysis, we predicted the best group of scenarios or individual scenario (based on the number of votes), estimated its posterior probabilities, but also the prior error rate

 ISSN: 2102-6238

as well as a proximal measure of the posterior predictive error rate. Both types of error rates were computed from 10,000 simulated pseudo-observed datasets (pods), for which the true scenario identity (ID) is known. The proximal measure of the posterior predictive error rate was computed conditionally to the observed dataset by selecting the ID model and the evolutionary parameter values within the 100 best simulations (i.e. those closest to the observed dataset as deduced by computing standardized Euclidean distances between the vectors of observed and simulated summary statistics) among a total of 800,000 simulated datasets generated from priors. It is worth stressing, that when pods are drawn randomly into prior distributions for both the scenario ID and the parameter values, one estimates global error levels computed over the whole (and usually huge) data space defined by the prior distributions. The levels of error may be substantially different depending on the location of an observed or pseudo-observed dataset in the prior data space. Indeed, some peculiar combination of parameter values may correspond to situations of strong (weak) discrimination among the compared scenarios. Aside from their use to select the best classifier and set of summary statistics, prior-based indicators are hence relatively poorly relevant since, for a given dataset, the only point of importance in the data space is the observed dataset itself. Computing error indicators conditionally to the observed dataset (i.e., focusing around the observed dataset by using a "posterior distribution") is hence clearly more relevant than blindly computing indicators over the whole prior data space.

2.6. Computer programs and computer times

For the simulation of data following the above model-prior design, we used the package DIYABC v2.1.0 (Cornuet et al., 2014), freely available with a detailed user-manual and example projects for academic and teaching purposes at http://www1.montpellier.inra.fr/CBGP/diyabc. Briefly, Cornuet et al. (2008, 2010, 2014) developed DIYABC to provide a user-friendly interface, which allows biologists with little background in programming to perform inferences via ABC. DIYABC is a coalescent-based program (Nordborg, 2001) which can consider complex population histories, including any number of divergence (without migration), admixture (i.e. punctual genetic introgression), and population size variation events, for population samples that may have been collected at different times. The package accepts various types of molecular data (microsatellites, DNA sequences, and SNP) evolving under various mutation models and located on various chromosome types (autosomal, X or Y chromosomes, and mitochondrial DNA). Regarding ABC-RF treatments which follow the generation of the reference table using DIYABC, computations were performed with the R package abcrf (version 1.7.1) available on the CRAN.

In the present study, all analyses were processed on a 16 cores Intel Xeon E5-2650 computer (Linux Debian platform, 64 bits system, with a maximum of 20 Gb of RAM used for the heaviest treatments). The production of a reference table including 100,000 simulated datasets (and summary statistics) took 40 minutes with 30% of the running time devoted to the computation of the 130 summary statistics for each simulated dataset. Optimizing computer code procedures to efficiently compute summary statistics is important especially in the case of high-dimensional ABC analyses which may include several thousand summary statistics. New efforts in this direction on a future version of the DIYABC program allowed us to reduce to only 5% the fraction of the running time devoted to the computation of 8,800 summary statistics associated to a high dimensional genetics dataset including 15,000 SNP loci (unpublished results). Such technical

 ISSN: 2102-6238

optimizations open new perspectives for the analysis of (very) high-dimensional datasets in population genetics. RF treatments, following the generation of the reference table and based on the R package abcrf, took four and eight minutes for scenarios grouping and individual scenarios configurations, respectively.

3. Results

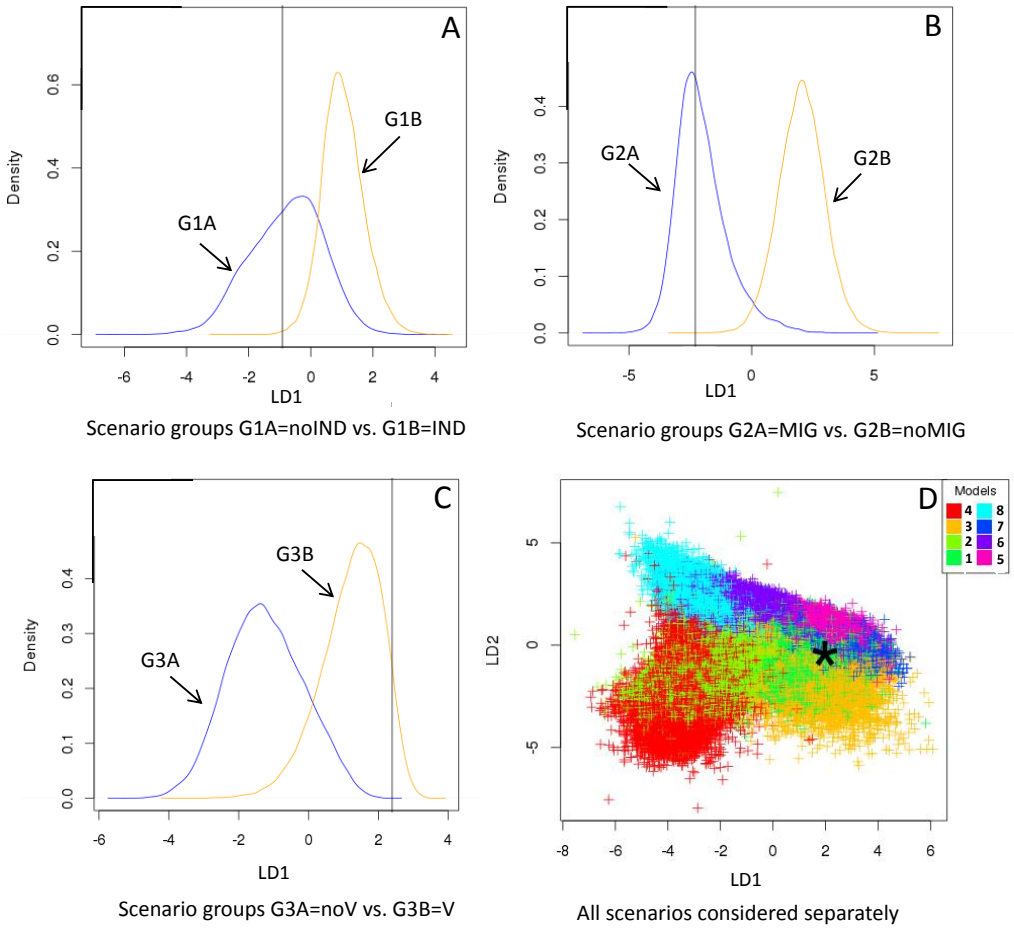
We first conducted ABC-RF treatments to make model choice on predefine groups of scenarios (group G1A vs group G1B; group G2A vs. group G2B; and group G3A vs. group G3B). We then carried out ABC-RF treatments on the eight scenarios considered separately.

The projection of the microsatellite population datasets from the reference table on a single (when analysing pairwise groups of scenarios) or on the first two LDA axes (when analysing the eight scenarios considered separately) provides a first visual indication about our capacity to discriminate among the compared scenarios (Figure 2). Simulations under the different pairwise groups of scenarios weakly overlapped indicating a strong power to discriminate among the pairwise groups of scenarios of interest. When considering the whole set of eight scenarios individually, the projected points substantially overlapped for at least some of the scenarios suggesting an overall lower power to discriminate among scenarios considered separately than when considering pairwise groups of scenarios. As a first inferential clue, one can note that the location of the observed dataset (indicated by vertical line or a star symbol in Figure 2) suggests, albeit without any formal quantification, a marked association with the scenario groups G1A, G2A and G3A, and, to a lower extent with the scenario 1.

A quantitative measure of the power to discriminate among groups of scenarios (scenarios) was obtained by estimating the probability to choose a wrong group of scenarios (scenario) when drawing index and parameter values of group of scenarios (scenario) into priors (i.e. prior error rates). Table 1 indicates substantially lower prior error rates when discriminating among groups of scenarios (i.e. 8.85% for G1A vs. G1B, 2.65% for G2A vs. G2B, and 10.54% for G3A vs. G3B) than among scenarios considered individually (prior error rate = 20.67%). Because for a given dataset, the only point of importance in the data space is the observed dataset, we conducted a second quantitative estimation of error rates corresponding to a proximal measure of the posterior predictive error rate computed conditionally to the observed dataset (Table 1). We found that posterior predictive error rates were substantially lower than prior error rates (i.e. posterior predictive error rate = 4.99 % for G1A vs. G1B, 0.91 % for G2A vs. G2B, 0.20 % for G3A vs. G3B, and 6.17 % for the scenarios considered separately), indicating that the observed dataset belongs to a region of the data space where the power to discriminate among groups of scenarios (individual scenarios) is higher than the global power computed over the whole prior data space. The conclusion of a higher power to discriminate among groups of scenarios than among scenarios considered separately still holds.

Figure 3 shows that RF analysis is able to automatically determine the (most) relevant statistics for model comparison. A typical feature of RF analysis is that LDA axes always correspond to the most informative statistics, which makes sense knowing their intrinsic construction structure. Interestingly, many of the most informative population genetics summary statistics were not selected by the experts in Verdu et al. (2009), especially some crude estimates of admixture rates based on population triplets (i.e. AML statistics; see Appendix B). A possible explanation is that

 ISSN: 2102-6238



Note: The location of the additional (observed) dataset is indicated by a vertical line in panels A, B and C, and a large black star in panel C. Panel A: in scenario group G1A the four Pygmy populations originate non-independently from the same ancestral Pygmy population (noIND) or independently from a non-Pygmy African population (group G1B=IND). Panel B: scenario group G2A include (MIG) or exclude (group G2B = noMIG) the possibility of recent and ancient asymmetrical admixture (i.e. migration) events between each Pygmy population and the non-Pygmy African one. Panel C: scenario group G3A include (V) or exclude (group G3B = noV) the possibility a change of population size in the non-Pygmy African population. Curves in A-C are estimated kernel (or equivalent) densities.

FIGURE 2. *Projection of the microsatellite population datasets from the reference table on a single (when analysing pairwise groups of scenarios) or on the first two LDA axes (when analysing the eight scenarios considered separately)*

 ISSN: 2102-6238

TABLE 1. *Error rates on scenario group (individual scenarios) choice and posterior probabilities of the selected scenario groups (scenario) when discriminating among evolutionary scenario groups (scenarios) of Pygmy human populations using Algorithms 4 and 5*

	Groups of scenarios			Individual scenarios
	G1A vs. G1B	G2A vs. G2B	G3A vs. G3B	
Prior error rate	8.85%	2.65%	10.54%	20.67%
Posterior predictive error rate	4.99%	0.91%	0.20%	6.17%
Posterior probability of the selected group of scenarios or scenario	0.923 (G1A)	0.987 (G2A)	0.955 (G3A)	0.851 (Scenario 1)

Note: Scenarios (see Figure 1) are grouped according to three types of major evolutionary events and then compared (column 2-4), or the eight compared scenarios are considered separately (column 5). The three evolutionary events included (or not) in the scenarios are (i) an independent (group G1A) vs. non-independent (group G1B) divergence from the non-pygmy African population, (ii) the presence (group G2A) vs. absence (group G2B) of asymmetrical introgression/migration from the non-Pygmy African population into the Pygmy populations, and (iii) the presence (group 3A) vs absence (group 3B) of variation in the effective population size (N_A) of the non-Pygmy African population (i.e. demographic expansion). Posterior probabilities of the selected scenario group or scenario are given in the last line for each type of analyses.

experts in population genetics are biased towards choosing summary statistics that are informative for parameter estimation under a given model. However, according to our own experience on this issue, the most informative statistics for model choice are often different than those that are informative for parameter estimates (Raynal et al., 2018; Robert et al., 2011). It is worth stressing that the most informative statistics differ depending on the model choice design. AML, FST and LIK statistics are among the most informative when discriminating among groups of scenarios dealing with independence/dependence of divergence events (G1A-B and individual scenarios) and introgresion/migration events (G2A-B and individual scenarios), whereas intra-population statistics such as VAR, V2P and MWG; see Appendix B) are the most informative ones when discriminating among groups of scenarios dealing with population size variation events (G3A-B). These differences are easy to interpret intuitively as divergence and introgression/migration events strongly impact the branching pattern of the tree topology summarizing the relationships among populations, which are informed by two and three sample statistics measuring the amount of genetic variation shared between populations (e.g. AML and FST), whereas population size variation events mainly impact the level of genetic variation within populations, which corresponds to the type of variation targeted by single sample statistics (e.g. VAR, V2P and MWG).

The outcome of the first step of the ABC-RF statistical treatment applied to a given target dataset is a classification vote for each scenario groups (or individual scenarios) which represents the number of times a given scenario group (or single scenario) is selected in a forest of n trees. The group of scenarios (or single scenario) with the highest classification vote corresponds to the (set of) model(s) best suited to the target dataset among the set of compared groups of scenarios or individual scenarios. In our case study, the classification vote estimated for the observed human microsatellite dataset was by far the highest for the scenario groups G1A (i.e. ensemble of scenarios in which the four Pygmy populations originate non-independently from a common ancestral Pygmy population; cf. 940 of the n = 1,000 RF-trees selected the scenario group G1A),

 ISSN: 2102-6238

the scenario groups G2A (i.e. scenarios including asymmetrical introgression/migration events between each Pygmy population and the non-Pygmy one; cf. 984 of the 1,000 trees), and the scenario groups G3A (i.e. scenarios including a change of population size in the non-Pygmy African population; cf. 959 of the 1,000 trees). When considering the eight scenarios separately, the classification vote estimated for the observed human microsatellite dataset was the highest for the scenario 1 which congruently includes all three above-selected evolutionary events (877 of the 1,000 trees).

It is worth stressing that there is no direct connection between the frequencies of the allocation of the data of the groups of scenarios (or individual scenarios) among the tree classifiers (i.e. the classification vote) and the posterior probabilities of the competing groups of scenarios (individual scenarios) (see Figure S2 in Pudlo et al., 2016). We therefore conducted the second RF analytical step corresponding to the algorithm 3 in Pudlo et al. (2016) to obtain a reliable estimation of posterior probability of the best group of scenarios (or individual scenario) (Table 1). The high posterior probability value provides a strong confidence in selecting the scenario groups G1A, G2A and G3A ($p = 0.923$, 0.987 and 0.955, respectively). When considering all scenarios separately, we found that the selected scenario 1 was associated to a moderately high posterior value (at least lower compared to those for groups of scenarios) of 0.851.

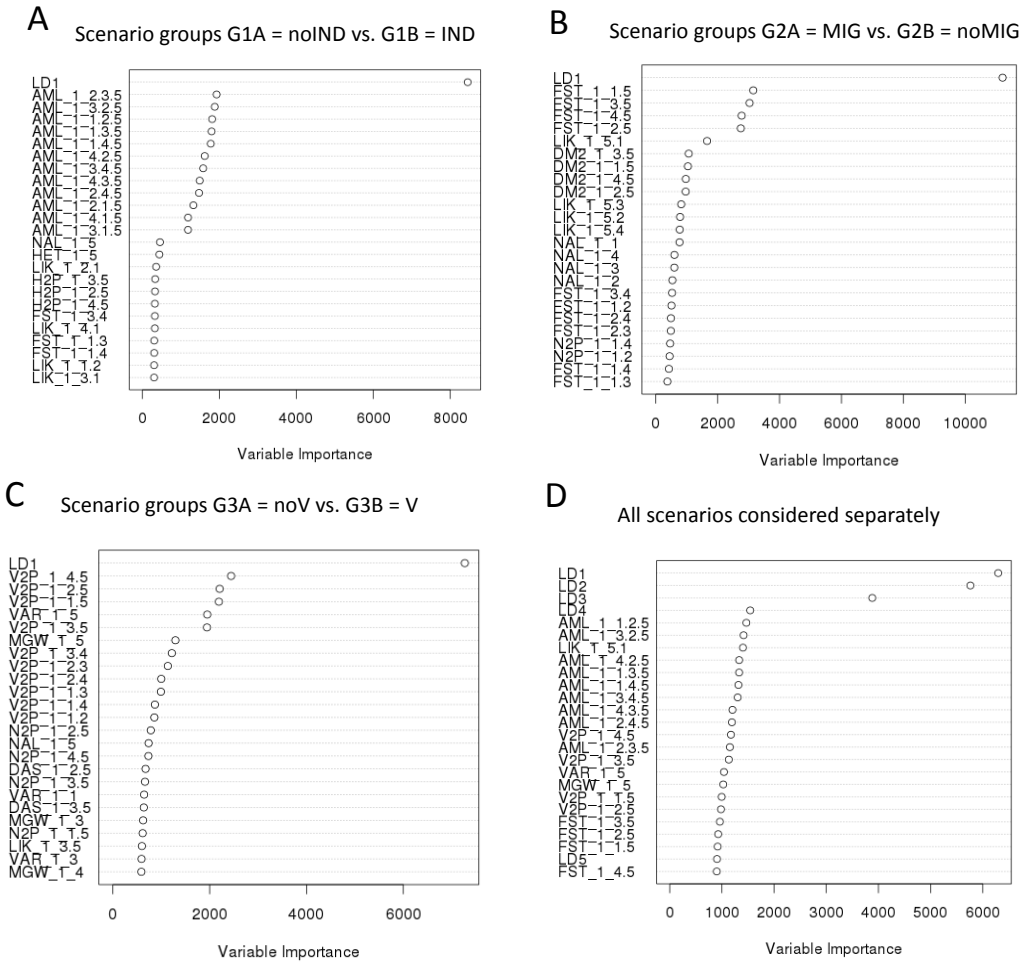
As for any Bayesian inference, the shape of the priors used for dataset simulations may affect both the posterior probabilities of scenarios and the posterior parameter estimation under ABC inference (e.g. Sunnåker et al., 2013). To empirically evaluate the influence of prior shape on our inferences, we conducted all ABC-RF analyses assuming a set of alternative non-flat priors for the simulations corresponding to the prior set 2 in Verdu et al. (2009). We found that such alternative statistical treatment did not change model choice results and that error rates and posterior probabilities were only moderately affected (results not shown).

4. Practical recommendations regarding the implementation of the Random Forests algorithms

We develop here several points, formalized as questions, which should help users seeking to apply our methodology on their dataset for statistical model choice.

Are my models and/or associated priors compatible with the observed dataset? This question is of prime interest and applies to any type of ABC treatment, including both standard ABC treatments and treatments based on ABC Random Forests. This issue is particularly crucial knowing that with complex models and high dimensional datasets (i.e. big an hence very informative datasets), as more and more encountered in population genomics, "all models are wrong...". Basically, if none of the proposed model - prior combinations produces some simulated datasets in a reasonable vicinity of the observed dataset, this is a signal of incompatibility, and we consider that is then useless to attempt model choice inference. In such situations, we strongly advise reformulating the compared models and/or the associated prior distributions in order to achieve some compatibility in the above sense. We propose here a visual way to address this issue, namely through the simultaneous projection of the simulated reference table datasets and of the observed dataset on the first LDA axes. Such a graphical assessment can be achieved using our R package abcrf version 1.7.1. In the LDA projection, the observed dataset has to be

 ISSN: 2102-6238



Note: The contribution of each statistics is evaluated as the total amount of decrease of the residual sum of squares, divided by the number of trees, for each of the 130 used population genetics summary statistics provided by DIYABC plus the LDA axes also used as statistics. The higher the variable importance the more informative is the statistic. The population index(s) are indicated at the end of each statistics (1 - 4 = Pygmy populations and 5 = non-Pygmy African population). The common number after the acronym of the summary statistics (i.e. "_1_") stands for the group of microstatellite markers used for inferences. For instance AML_1_3.2.5 corresponds to the Maximum likelihood coefficient of admixture considering that population 5 is an admixted (i.e. genetically introgressed) population with parental populations 3 and 2, computed using microsatellite marker group 1. More details about statistics can be found in Appendix B: See legend of Figure 1 for details about scenario groups.

FIGURE 3. *Contributions of the 25 most informative statistics to the Random Forest*

located reasonably within the clouds of simulated datasets (see Figure 2 as an illustration). Note that visual representations of a similar type (although based on PCA) as well as computation for each summary statistics and for each model of the probabilities of the observed values in the prior distributions have been proposed by Cornuet et al. (2010) and are already automatically provided by the DIYABC software.

Frazier et al. (2018) very recently analysed the behavior of approximate Bayesian computation (ABC) when the model generating the simulated data differs from the actual data generating process; i.e., when the data simulator in ABC is misspecified. They demonstrate that when the model is misspecified different versions of ABC can lead to substantially different results and they suggest approaches to diagnose model misspecification in ABC.

Did I simulate enough datasets for my reference table? A rule of thumb is to simulate between 5,000 and 20,000 datasets per model among those compared. In the present example we simulated 12,500 datasets for each of the eight compared scenarios (total = 100,000 datasets in the reference table). To evaluate whether or not this number is sufficient for Random Forests analysis, we recommend to compute global prior error rates from both the entire reference table and a subset of the reference table (for instance from a subset of 80,000 simulated datasets if the reference table includes a total of 100,000 simulated datasets). If the prior error rate value obtained from the subset of the reference table is similar, or only slightly higher, than the value obtained from the entire reference table, one can consider that the reference table contains enough simulated datasets. If a substantial difference is observed between both values, then we recommend an increase in the number of datasets in the reference table.

Did my forest grow enough trees? According to our experience, a forest made of 500 trees often constitutes an interesting trade-off between computation efficiency and statistical precision (Breiman, 2001). To evaluate whether or not this number is sufficient, we recommend plotting the estimated values of the prior error rate and/or the posterior probability of the best model as a function of the number of trees in the forest. The shapes of the curves provide a visual diagnostic of whether such key quantities stabilize when the number of trees tends to a given value (1,000 trees in the present study).We provide illustrations of such procedure and visual representations in the case of inferences about Human Pygmy population history (see Appendix C in which graphical representation have been produced by our R package abcrf version 1.7.1).

5. Conclusions and perspectives

Choosing among a group of models (individual scenarios) is a crucial inferential issue as it allows the identification of major historical and evolutionary events formalized into a set of compared scenarios formalized as a combination of such evolutionary events. We illustrate this issue through ABC-RF analyses to make inferences about the genetic history of Pygmy human populations. The eight formalized complex scenarios incorporate (or not) three main evolutionary events: (i) whether there is an independent or non-independent origin of Pygmy groups, (ii) the possibility of introgression/migration events between Pygmy and non-Pygmy African populations, and (iii) the possibility of a change in effective size in the past in the non-Pygmy African population. We found that our scenario grouping approach allows disentangling with strong con-

 ISSN: 2102-6238

fidence (i.e. low error rates and high posterior probabilities) the main events that compose the evolutionary history of interest. The final selected scenario (when comparing all eight scenarios separately) corresponds to a common origin of all Western Central African Pygmy groups considered, with the ancestral Pygmy populations having diverged from the non-Pygmy African population in a more remote past. Furthermore, it encompasses both recent and ancient asymmetrical introgression events from the non-Pygmy African gene-pool into each Pygmy population considered, and a change of population size in the non-Pygmy African population. Our ABC-RF analyses confirm and strengthen the initial historical interpretation of Verdu et al. (2009). We inferred a probable common origin of all Western Central African populations categorized as Pygmies by Western explorers, despite the vast cultural, morphological, and genetic diversity observed today among these populations (Hewlett, 2014). We also confirmed recent asymmetrical and heterogeneous genetic introgressions from non-Pygmies into each Pygmy population. Altogether, these results are in agreement with the ethno-historical scenario proposed by Verdu et al. (2009) in which the relatively recent expansion of non-Pygmy agriculturalist populations in Western Central Africa which occurred 2000-5000 YBP may have modified the pre-existing social relationships in the ancestral Pygmy population, in turn resulting in its fragmentation into isolated groups. Since then, enhanced genetic drift in isolated populations with small effective sizes, and different levels of genetic introgression from non-Pygmies into each Pygmy population led to the rapid genetic diversification of the various Western Central African Pygmy populations observed today.

ABC-RF makes two major advances regarding the use of summary statistics, and the identification of the most probable model. For the summary statistics, given a pool of different metrics available, ABC-RF extracts the maximum of information from the entire set of proposed statistics. This avoids the arbitrary choice of a subset of statistics, which is often applied in ABC analyses, and also avoids what in the statistical sciences is called "the curse of dimensionality" (see Blum et al., 2013, for a comparative review of dimension reduction methods in ABC). With respect to identifying the most probable model, ABC-RF uses a classification vote system rather than the posterior probabilities traditionally used in ABC analysis. Pudlo et al. (2016) shows that, as compared to previous ABC methods, ABC-RF offers at least four advantages (i) it significantly reduces the model classification error as measured by the prior error rate (see also Fraimout et al., 2017, for a detailed study on this issue); (ii) it is robust to the number and choice of summary statistics, as RF can handle many superfluous and/or strongly correlated statistics with no impact on the performance of the method (Marin et al., 2018; Raynal et al., 2018); (iii) the computing effort is considerably reduced as RF requires a much smaller reference table compared with alternative methods (i.e., a few thousands of simulated datasets versus hundreds of thousands to millions of simulations per compared model for more standard ABC approaches). For instance, our ABC-RF treatments on Pygmy populations required a ca. 40 times shorter computational duration than when using more standard ABC approaches which required a reference table of four million simulated datasets instead of 100,000 for ABC-RF (Verdu et al., 2009; Estoup et al., 2018); and (iv) ABC-RF provides more reliable estimation of posterior probability of the selected model (i.e., the model that best fit the observed dataset). Such major computational and inference improvements broadens the possibility of using simulation-based methods in computationally expensive case studies.

It is worth stressing that ABC-RF approaches has been recently developed by our group to

 ISSN: 2102-6238

estimate the posterior distributions of model parameters of interest (Raynal et al., 2018), for instance divergence times, introgression/admixture rates, and effective population sizes in the present Pygmy population history. Raynal et al. (2018) propose an approach which relies on the Random Forests methodology of Breiman (2001) applied in a (non-parametric) regression setting. They advocate the derivation of a new Random Forest for each component of the parameter vector of interest. When compared with earlier ABC solutions, this new RF method offers significant gains in terms of robustness to the choice of the summary statistics, does not depend on any type of tolerance level, and is a good trade-off in term of quality of point estimator precision of parameters and credible interval estimations for a given computation time.

The present study highlights the major potentialities of ABC-RF methods for inference on complex population origins and demographic histories using molecular datasets. More generally, we believe that ABC methods, especially ABC-RF algorithms, will be of considerable interest for the statistical processing of massive datasets whose availability rapidly increases in various fields of research including, but not limited to, population genetics.

References

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

Beaumont, M. A., Zhang, W., and Balding, D. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Bertorelle, G., Benazzo, A., and 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.

Biau, G. (2012). Analysis of a random forest model. *Journal of Machine Learning Research*, 13:1063–1095.

Biau, G., Cérou, F., and Guyader, A. (2015). New insights into Approximate Bayesian Computation. *Annales de l'Institut Henri Poincaré B, Probability and Statistics*, 51(1):376–403.

Blum, M. (2010). Approximate Bayesian Computation: A Nonparametric Perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.

Blum, M. and François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20:63–73.

Blum, M., Nunes, M., Prangle, D., and Sisson, S. (2013). A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Statistical Science*, 28(2):189–208.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Cavalli-Sforza, L. (1986). African pygmies: an evaluation of the state of research. In Cavalli-Sforza, L., editor, *African pygmies*, pages 361–426. Orlando Academic Press.

Cavalli-Sforza, L. and Feldman, M. (2003). The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, 33:266–275.

Cavalli-Sforza, L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press.

Choisy, M., Franck, P., and Cornuet, J.-M. (2004). Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology*, 13:955–968.

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J.-M., and 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.

Cornuet, J.-M., Ravigné, V., and Estoup, A. (2010). Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, 11(1):401.

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 24(23):2713–2719.

 ISSN: 2102-6238

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418.

De Iorio, M. and Griffiths, R. (2004a). Importance sampling on coalescent histories. i. *Advances in Applied Probability*, 36(2):417–433.

De Iorio, M. and Griffiths, R. (2004b). Importance sampling on coalescent histories. ii: Subdivided population models. *Advances in Applied Probability*, 36(2):434–454.

De Iorio, M., Griffiths, R., Leblois, R., and Rousset, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology*, 68(1):41–53.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020.

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglià, A., Tofanelli, S., Spedini, G., and Capelli, C. (2004). Variation of Female and Male Lineages in Sub-Saharan Populations: the Importance of Sociocultural Factors. *Molecular Biology and Evolution*, 21(9):1673–1682.

Drummond, A. and Bouckaert, R. (2015). Bayesian evolutionary analysis by sampling trees. In *Bayesian Evolutionary Analysis with BEAST*, pages 79–96. Cambridge University Press.

Drummond, A. and Rambaut, A. (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214.

Estoup, A. and Guillemaud, T. (2010). Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology*, 19(19):4113–4130.

Estoup, A., Jarne, P., and Cornuet, J.-M. (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, 11(9):1591–1604.

Estoup, A., Lombaert, E., Marin, J.-M., Robert, C., Guillemaud, T., Pudlo, P., and Cornuet, J.-M. (2012). Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Molecular Ecology Ressources*, 12(5):846–855.

Estoup, A., Verdu, Marin, Robert, Dehne-Garcia, Cornuet, and Pudlo (2018). Application of approximate Bayesian computation to infer the genetic history of Pygmy hunter-gatherers populations from Western Central Africa. In Sisson, S., Fan, Y., and Beaumont, M., editors, *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.

Excoffier, L., Estoup, A., and Cornuet, J.-M. (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, 169:1727–1738.

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, 104(45):17614–17619.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for Approximate Bayesian Computation: semi-automatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Fraimout, A., Debat, V., Fellous, S., Hufbauer, R. A., Foucaud, J., Pudlo, P., Marin, J.-M., Price, D. K., Cattel, J., Chen, X., et al. (2017). Deciphering the Routes of invasion of Drosophila suzukii by Means of ABC Random Forest. *Molecular biology and evolution*, 34(4):980–996.

Frazier, D. T., Robert, C. P., and Rousseau, J. (2018). Model Misspecification in ABC: Consequences and Diagnostics. *ArXiv e-prints*, (1708.01974v2).

Garza, J. and Williamson, E. (2001). Detection of reduction in population size using data from microsatellite DNA. *Molecular Ecology*, 10:305–318.

Goldstein, D., Linares, A., Cavalli-Sforza, L., and Feldman, N. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139:463–471.

Grelaud, A., Marin, J.-M., Robert, C., Rodolphe, F., and Tally, F. (2009). Likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 3(2):427–442.

Hewlett, B. S. (1996). Cultural diversity among african pygmies. In Kent, S., editor, *Cultural Diveristy among Twentieth-Century Foragers: An African Perspective*, pages 361–426. Cambridge: Cambridge University Press.

Hewlett, B. S. (2014). *Hunter-gatherers of the Congo Basin: cultures, histories, and biology of African pygmies*. New Brunswick: Transactions Publishers.

Jin, L. and Chakraborty, R. (1994). Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data. *Molecular Biology and Evolution*, 11(1):120–127.

Lombaert, E., Guillemaud, T., Cornuet, J.-M., Malausa, T., Facon, B., and Estoup, A. (2010). Bridgehead effect in

 ISSN: 2102-6238

the worldwide invasion of the biocontrol harlequin ladybird. *PloS one*, 5(3):e9743.
Marin, J.-M., Pudlo, P., Estoup, A., and Robert, C. P. (2018). Likelihood-free model choice. in handbook of approximate bayesian computation. In Sisson, S., Fan, Y., and Beaumont, M., editors, *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.
Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, pages 1–14.
Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.
Merle, C., Leblois, R., Rousset, F., and Pudlo, P. (2017). Resampling: An improvement of importance sampling in varying population size models. *Theoretical Population Biology*, 114:70–87.
Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York, USA.
Nordborg, M. (2001). Coalescent theory. *Handbook of statistical genetics*, pages 179–212.
Pascual, M., Chapuis, M., Mestres, F., Balanyà, J., Huey, R., Gilchrist, G., and Estoup, A. (2007). Introduction history of *drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Molecular Ecology*, 19:3069–3083.
Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16:1791–1798.
Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866.
Rannala, B. and J., M. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences, USA*, 94:9197–9201.
Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2018). ABC random forests for Bayesian parameter inference. *Bioinformatics*. bty867.
Robert, C., Cornuet, J.-M., Marin, J.-M., and Pillai, N. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117.
Rousset, F. and Leblois, R. (2007). Likelihood and Approximate Likelihood Analyses of Genetic Structure in a Linear Habitat: Performance and Robustness to Model Mis-Specification. *Molecular Biology and Evolution*, 24(12):2730–2745.
Rousset, F. and Leblois, R. (2012). Likelihood-based inferences under isolation by distance: Two-dimensional habitats and confidence intervals. *Molecular Biology and Evolution*, 29(3):957–973.
Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12:1151–1172.
Schrider, D. and Kern, A. (2016). S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLOS Genetics*, 12(3):1–31.
Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741.
Sheehan, S. and Song, Y. (2016). Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12:1–28.
Sisson, S., Fan, Y., and Tanaka, M. (2009). Sequential Monte Carlo without likelihoods: Errata. *Proceedings of the National Academy of Sciences*, 106(39):16889.
Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.
Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B*, 62(4):605–655. With discussion and a reply by the authors.
Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS Computational Biology*, 9(1):e1002803.
Tavaré, S., Balding, D., Griffiths, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.
Thouzeau, V., Mennecier, P., Verdu, P., , and Austerlitz, F. (2017). Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations. *Proceedings of the Royal Society of London B: Biological Sciences*, 284(1861).
Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J.-M., et al. (2009). Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, 19(4):312–318.
Verdu, P., Becker, N. S., Froment, A., Georges, M., Grugni, V., Quintana-Murci, L., Hombert, J.-M., Van der Veen, L.,

 ISSN: 2102-6238

Le Bomin, S., Bahuchet, S., et al. (2013). Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Molecular Biology and Evolution*, 30(4):918–937.

Weir, B. and Cockerham, C. (1984). Estimating F-statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370.

Wilkinson, R. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141.

Acknowledgements

A.E. acknowledges financial support by the National Research Fund ANR (France) through the project ANR-16-CE02-0015-01 (SWING), and the INRA scientific department SPE (AAP-SPE 2016). This work has also been funded by Labex NUMEV (NUMEV, ANR10-LABX-20).

Appendix A: Description of the historical and demographic model-parameters and their prior distributions, for the eight competing scenarios considered for the origin and diversification of Pygmy populations from Western Africa

The eight scenarios with their historical and demographic parameters are represented in Figure 1 of the main text. The column "Scenarios" indicates in which scenario each model parameter appears. The column "Group" indicates in which group of scenarios each model parameter appears when processing a model-grouping approach. The index i indicated for some parameters corresponds to population index (1 - 4 = Pygmy populations and 5 = non-Pygmy African population). Scenarios and associated model parameters follow the same notation as in Verdu et al. (2009).

Parameter type	Parameter name	Prior distribution	Scenarios	Group
Divergence times [(a)]	t_{pnp}; t_p	U[1,5000]	1, 2, 3, 4	G1A
	t_{pnpi} , with i in $\{1,\ldots,4\}$	U[1,5000]	5, 6, 7, 8	G1B
Admixture times [(a)]	tr_a	U[1,5000]	1, 2, 3, 4	G1A
	tr_r	Log-U[1,5000]	1, 2, 3, 4	G1A
	tr_{ai} , with i in $\{1,\ldots,4\}$	U[1,5000]	5, 6, 7, 8	G1B
	tr_{ri} , with i in $\{1,\ldots,4\}$	Log-U[1,5000]	5, 6, 7, 8	G1B
Times of effective population size changes [(a)]	t_A	U[1,10000]	1, 2, 5, 6	G3A
	no t_A		3, 4, 7, 8	G3B
	t_{nei}, with i in $\{1,\ldots,4\}$	U[1,5000]	5, 6, 7, 8	G1B
Admixture rates	r_a; r_{ai}; r_{ri}, with i in $\{1,\ldots,4\}$	U[0,1]	1, 3, 5, 7	G2A
	r_a; r_{ai}; r_{ri}, with i in $\{1,\ldots,4\}$	0 (no admixture)	2, 4, 6, 8	G2B
Effective population sizes [(b)]	N_A	U[100,10000]	1, 2, 5, 6	G3A
	no N_A		3, 4, 7, 8	G3B
	N_{ap}	U[100,10000]	1, 2, 3, 4	G1A
	N_{ai} , with i in $\{1,\ldots,4\}$	U[100,10000]	5, 6, 7, 8	G1B
	N_{np}	U[1000,100000]	1, 2, 3, 4, 5, 6, 7, 8	G1A, G1B, G2A, G2B, G3A, G3B
	N_i , with i in $\{1,\ldots,4\}$	U[100,10000]	1, 2, 3, 4, 5, 6, 7, 8	G1A, G1B, G2A, G2B, G3A, G3B

[(a)] In number of generations (assuming a generation time of 25 year; Verdu et al. (2009))

[(b)] In number of (reproductively effective) diploid individuals

 ISSN: 2102-6238

Appendix B: Summary statistics

For microsatellite markers, the program DIYABC v.2.1.0 proposes a series of summary statistics among those used by population geneticists. These summary statistics are mean values over loci and characterize a single, a pair or a trio of population samples. [(a)] Number of statistics computed in the present dataset which includes five population samples. [(b)] In contrast to other two sample statistics, the mean index of classification for a pair of population samples (LIK) is asymmetrical as it corresponds to the mean individual likelihoods of individuals sampled in population i being assigned to population j (LIK $i.j$) and of individuals sampled in population j being assigned to population i (LIK $j.i$); hence a total of 20 LIK statistics for a dataset including five population samples. [(c)] AML $i.j.k$ triplets refer to a target population sample i (the admixed population) and two parental population samples (j and k); hence a total of 30 AML statistics for a dataset including five population samples. Computational details about statistics can be found in the references given in the last column of the table.

Type of statistics	Summary statistics	Acronym	Number of statistics computed [(a)]	Reference
Single sample statistics across loci	Mean number of alleles	NAL	5	
	Mean gene diversity	HET	5	Nei (1987)
	Mean allele size variance	VAR	5	
	Mean M index	MWG	5	Garza and Williamson (2001) and Excoffier et al. (2005)
Two sample statistics across loci pooling two samples	Mean number of alleles	N2P	10	
	Mean gene diversity	H2P	10	
	Mean allele size variance	V2P	10	
Two sample statistics	FST between two samples	FST	10	Weir and Cockerham (1984)
	Mean index of classification	LIK	20 [(b)]	Rannala and J. (1997) and Pascual et al. (2007)
	Shared allele distance	DAS	10	Jin and Chakraborty (1994)
	$(\delta\mu)^2$ distance	DM2	10	Goldstein et al. (1995)
Three sample statistics	ML coefficient of admixture	AML	30 [(c)]	Choisy et al. (2004)
Total number of statistics			130	

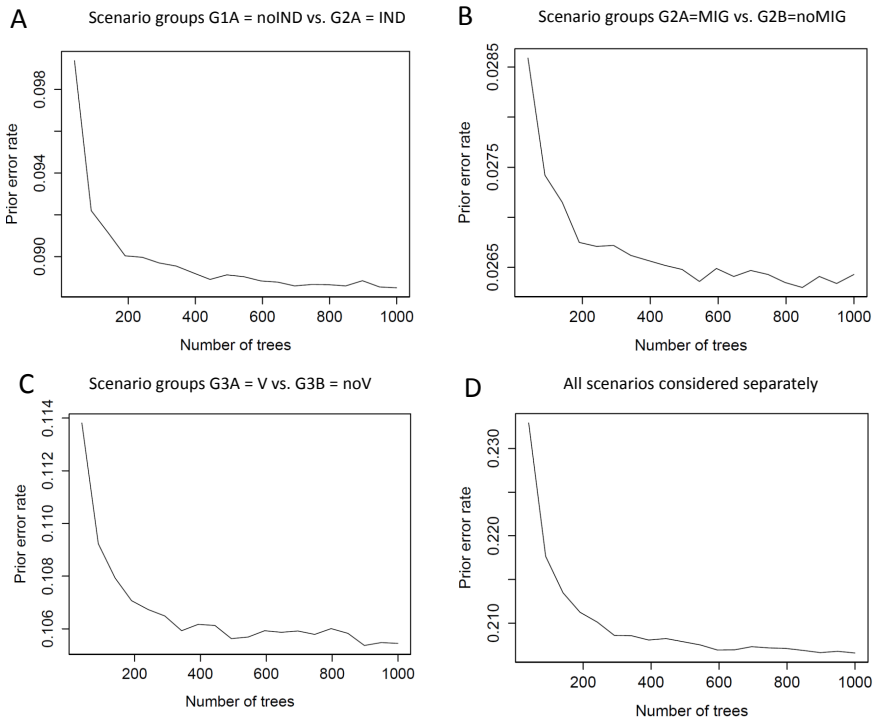
All 130 statistics computed in the present study are listed below. The population sample index(s) are indicated at the end of each statistics (1 - 4 = Pygmy populations and 5 = non-Pygmy African population). For instance `AML - 1.2.3` corresponds to the Maximum likelihood coefficient of admixture considering that population 1 is an admixted population with parental populations 2 and 3.

 ISSN: 2102-6238

```
NAL - 1 2 3 4 5
HET - 1 2 3 4 5
VAR - 1 2 3 4 5
MGW - 1 2 3 4 5
N2P - 1.2 1.3 1.4 1.5 2.3 2.4 2.5 3.4 3.5 4.5
H2P - 1.2 1.3 1.4 1.5 2.3 2.4 2.5 3.4 3.5 4.5
V2P - 1.2 1.3 1.4 1.5 2.3 2.4 2.5 3.4 3.5 4.5
FST - 1.2 1.3 1.4 1.5 2.3 2.4 2.5 3.4 3.5 4.5
LIK - 1.2 1.3 1.4 1.5 2.1 2.3 2.4 2.5 3.1 3.2
      3.4 3.5 4.1 4.2 4.3 4.5 5.1 5.2 5.3 5.4
DAS - 1.2 1.3 1.4 1.5 2.3 2.4 2.5 3.4 3.5 4.5
DM2 - 1.2 1.3 1.4 1.5 2.3 2.4 2.5 3.4 3.5 4.5
AML - 1.2.3 1.2.4 1.2.5 1.3.4 1.3.5 1.4.5 2.1.3
      2.1.4 2.1.5 2.3.4 2.3.5 2.4.5 3.1.2 3.1.4
      3.1.5 3.2.4 3.2.5 3.4.5 4.1.2 4.1.3 4.1.5
      4.2.3 4.2.5 4.3.5 5.1.2 5.1.3 5.1.4 5.2.3
      5.2.4 5.3.4
```

Appendix C: Evolution of ABC-RF prior error rates with respect to the number of trees in the forest

The following graphs represent the decrease of the ABC-RF prior error rate with the number of trees in the forest for the four RF analyses conducted using a reference table including 100,000 simulated datasets. For all analyses the gain of increasing the number of trees becomes limited for a number of trees > 800; hence our final choice of building forests from 1,000 trees.



 ISSN: 2102-6238