DR. SILVIA GHIROTTO (Orcid ID : 0000-0003-2522-9277)

Article type : Special Issue

Distinguishing among complex evolutionary models using unphased whole-genome data through Random-Forest Approximate Bayesian Computation

Silvia Ghirotto[*§1], Maria Teresa Vizzari[*1], Francesca Tassi[2], Guido Barbujani[2] and Andrea Benazzo[§2]

[1]Department of Mathematics and Computer Science, University of Ferrara, 44121 Ferrara, Italy

[2]Department of Life Sciences and Biotechnology, University of Ferrara, 44121 Ferrara, Italy

[*] these authors contributed equally to this work

[§] correspondence should be addressed to andrea.benazzo@unife.it and silvia.ghirotto@unife.it

Abstract

Inferring past demographic histories is crucial in population genetics, and the amount of complete genomes now available should in principle facilitate this inference. In practice, however, the available inferential methods suffer from severe limitations. Although hundreds complete genomes can be simultaneously analyzed, complex demographic processes can easily exceed computational constraints, and the procedures to evaluate the reliability of the estimates contribute to increase the computational effort. Here we present an Approximate Bayesian Computation framework based on the Random Forest algorithm (ABC-RF), to infer complex past population processes using complete genomes. To this aim, we propose to summarize the data by the full genomic distribution of the four mutually exclusive categories of segregating sites (*FDSS*), a statistic fast to compute from unphased genome data and that does not require the ancestral state of alleles to be known. We constructed an efficient ABC pipeline and tested how accurately it allows one to recognize the true model among models of increasing complexity, using simulated data and taking into account different sampling strategies in terms of number of individuals analyzed, number and size of the genetic loci considered. We also compared the *FDSS* with the unfolded and folded Site Frequency Spectrum, and for these statistics we highlighted the experimental conditions

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.13263

maximizing the inferential power of the ABC-RF procedure. We finally analyzed real datasets, testing models on the dispersal of anatomically modern humans out of Africa and exploring the evolutionary relationships of the three species of Orangutan inhabiting Borneo and Sumatra.

Introduction

A faithful reconstruction of the demographic dynamics of a species is important both to improve our knowledge about the past and to disentangle the effects of demography from those of natural selection (Akey et al., 2004; Lohmueller, 2014; D. Meyer, Single, Mack, Erlich, & Thomson, 2006). In recent years, thousands of modern and ancient complete genome sequences have become available, potentially containing vast amounts of information about the evolutionary history of populations(1000 Genomes Project Consortium, 2012; Dasmahapatra et al., 2012; De Manuel et al., 2016; Mallick et al., 2016; M. Meyer et al., 2012; Moreno-Mayar et al., 2018; Prüfer et al., 2014). However, these genomes do not speak by themselves; to extract the evolutionary information they contain, appropriate inferential statistical methods are required. Some methods based on the Sequential Markovian Coalescent (SMC) model (McVean & Cardin, 2005), became popular among population geneticists due to their ability to infer population size changes through time (PSMC; Li & Durbin, 2011) and divergence times (MSMC; Schiffels & Durbin, 2014), and to scale well on whole genome sequences. Under these approaches, the local density of heterozygote sites along chromosomes is used to estimate the times of the most recent common ancestor (TMRCA) of genomic regions separated by recombination, thus providing insight into ancestral population sizes and the timing of divergence processes. These estimates are often used to indirectly support hypotheses regarding the evolution of the studied organisms. Albeit sophisticated, these methods present some limitations; the temporal resolution of the inferred demographic events seems to be strongly dependent on the number of individuals included, with poor performance in the recent past especially when analyzing single individuals. Moreover, these methods assume no gene flow among the investigated populations, which in many cases is plainly implausible. The consequences on the inferential process of violation of this assumption have been investigated using both mathematical theory (Mazet, Rodríguez, Grusea, Boitard, & Chikhi, 2016) and computer simulations (Chikhi et al., 2018).

Other methods infer demographic parameters via the diffusion approximation (Gutenkunst, Hernandez, Williamson, & Bustamante, 2010), or coalescent simulations (Beeravolu, Hickerson, Frantz, & Lohse, 2018; Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013), from the *SFS* computed on large genomic datasets. The *SFS* records the observed number of polymorphisms segregating at different frequencies in a sample of n individuals and is generally computed over a certain number of genomic regions where no influence of natural selection is assumed. The expectation of the *SFS* under different evolutionary scenarios could be approximated by the diffusion theory (as implemented e.g. in *dadi*), directly via coalescent simulations (as in *fastsimcoal* or *ABLE*), or computed analytically (Chen, 2012;

Jouganous, Long, Ragsdale, & Gravel, 2017; Kamm, Terhorst, & Song, 2017); alternative demographic histories can be compared via e.g. AIC (Akaike, 1974). Still, there are limits to the complexity of models that can be analyzed, and AIC-like approaches can only be used to understand which modifications significantly improve the model, without explicit model testing and a direct attribution of probabilities to each tested scenario. Therefore, through these approaches, model checking can be problematic (i.e. to evaluate whether and to what extent the compared models can actually be distinguished from each other, or whether the selected model can capture the observed variation), and so is quantifying the strength of the support associated to the best model (Beeravolu et al., 2018). Indeed, the only available procedure to assess the models identifiability or to test for the goodness of fit of the best scenario requires the analysis of many datasets simulated under known demographic conditions, which can be computationally prohibitive, in particular for complex evolutionary scenarios (Excoffier et al., 2013).

Recently, an inferential method that couples the ability of the SMC to deal with whole genome sequences and the population signal gathered from the *SFS* has been developed (SMC++; Terhorst, Kamm, & Song, 2017). Under this inferential framework, both the genomic and the *SFS* variation are jointly used to estimate population size trajectories through time, as well as the divergence time between pairs of populations. Although this approach seems to scale well on thousands of unphased genomes, it is based on the same assumption of classical SMC methods (with populations evolving independently), which severely limits its use whenever gene flow cannot be ruled out.

One powerful and flexible way to quantitatively compare alternative models and estimating model's parameters relies on the Approximate Bayesian Computation (ABC) methods. Under these methods, the likelihood functions need not be specified, because posterior distributions can be approximated by simulation, even under complex (and hence realistic) population models, incorporating prior information. The genetic data, both observed and simulated, are summarized by the same set of "sufficient" summary statistics, selected to be informative about the genealogic processes under investigation. The ability of the framework to distinguish among the alternative demographic models tested and the quality of the results can be evaluated with rather limited additional effort (for a review see e.g. Bertorelle, Benazzo, & Mona, 2010; Csilléry, Blum, Gaggiotti, & François, 2010).

Although ABC has the potential to deal with complex and realistic evolutionary scenarios, its application to the analysis of large genomic datasets, such as complete genomes, is still problematic. In its original formulation, indeed, the ABC procedure, depending on the complexity of the models tested (i.e. the number of parameters, and the size of the prior distributions on the parameters), may require the simulation of millions data sets of the same size of those observed. This step becomes computationally very expensive as the dataset size increases in size, or when many models need be compared. In addition, there is no accepted standard as for the choice of the summary statistics describing both observed and simulated data, as recognized since the first formal introduction of ABC (Mark A. Beaumont, Zhang, & Balding,

2002; Marjoram, Molitor, Plagnol, & Tavare, 2003). Increasing the number of summary statistics, indeed, makes it easier to choose the best model, but inevitably reduces the accuracy of the demographic inference (this problem is referred to as the "curse of dimensionality", Blum & François, 2010). Ideally, the good practice would be to select a set of summary statistics that is both low-dimensional and highly informative on the demographic parameters defining the model. In practice, however, this problem is still unsolved, despite several serious attempts (M G B Blum, Nunes, Prangle, & Sisson, 2013).

Recently, a new ABC framework has been developed based on a machine-learning tool called Random Forest (ABC-RF, Pudlo et al., 2015). Under ABC-RF, the Bayesian model selection is rephrased as a classification problem. At first, the classifier is constructed from simulations from the prior distribution via a machine learning RF algorithm. Once the classifier is constructed and applied to the observed data, the posterior probability of the resulting model can be approximated through another RF that regresses the selection error over the statistics used to summarize the data. The RF classification algorithm has been shown to be insensitive both to the correlation between the predictors (in case of ABC, the summary statistics) and to the presence of relatively large numbers of noisy variables. This means that even choosing a large collection of summary statistics, the correlation between some of them and others (which may be uninformative about the models tested), have no consequences on the RF performance, and hence on the accuracy of the inference. Moreover, compared to the standard ABC methods, the RF algorithm performs well with a radically lower number of simulations (from millions to tens of thousands per model). These properties make the new ABC-RF algorithm of particular interest for the statistical analysis of massive genetic datasets. In this light, the unfolded *SFS*, that due to the above mentioned limitations has been rarely used in a classical ABC context (Eldon, Birkner, Blath, & Freund, 2015), should be a suitable (and possibly sufficient) statistic to summarize genomic data (Lapierre, Lambert, & Achaz, 2017; Smith et al., 2017; Terhorst & Song, 2015). However, to obtain a complete representation of the frequency spectrum the ancestral state of a SNP has to be known; any uncertainty linked to the identification of the ancestral state cause indeed a bias in the reconstruction of the spectrum and, consequently, on the inference of the demographic dynamics behind it (Hernandez, Williamson, & Bustamante, 2007; Keightley & Jackson, 2018). In such cases, the folded version of the *SFS* should be used, with unavoidable loss of information (Keightley & Jackson, 2018). Moreover, since the *SFS* is based on allele frequencies, its reliability should increase as increasing the number of individuals sampled per population, that in certain condition may rather be a limiting factor (i.e. in the analysis of ancient data).

In this paper we tested the power of the newly developed ABC-RF procedure for model selection summarizing the data through a set of summary statistics that 1- can be easily calculated from unphased genomes data, 2- do not require information about ancestral state of alleles and 3- are known to be informative about past processes of divergence and admixture (Wakeley & Hey, 1997). These statistics are the four mutually exclusive categories of segregating sites for pair of populations (i.e. private

polymorphisms in either population, shared polymorphisms and fixed differences), calculated as frequency distributions over the whole genome (hence the *FDSS*, frequency distribution of segregating sites). These statistics have already been successfully used in a standard ABC context (Robinson, Bunnefeld, Hearn, Stone, & Hickerson, 2014), but only in the form of the first four moments of the distribution across loci. Here, for the first time, and thanks to the ABC-RF procedure, we analyze the full genomic distribution of each statistic, and compare its performance with the one achievable using the unfolded and the folded pairwise joint *SFS* (calculated across all sites, including monomorphic loci).

We first performed a power analysis, to evaluate how accurately this ABC pipeline can recognize the true model among models of increasing complexity, using simulated data summarized by both the *FDSS* and the *SFS*. We also explored the performances of the presented procedure with respect to the experimental conditions, evaluating the consequences of sampling strategies involving different numbers of chromosomes, genomic loci, and locus lengths. Our results show that the ABC-RF coupled with the *FDSS* can reliably distinguish among demographic histories, in particular when few chromosomes per population are considered. In all other cases, the performances are comparable to those obtained with the *SFS*.

As a final step, we applied our method to two case studies, in all cases choosing to sample a single individual (i.e. two chromosomes) per population. First, we analyzed the demographic history of anatomically modern humans and the dynamics of migration out of the African continent, explicitly comparing two models proposed by Malaspinas et al., (2016) and by Pagani et al., (2016). Secondly, we reconstructed the past demographic history and the interaction dynamics among the three orangutan species inhabiting Borneo and Sumatra, revising the models presented by Nater et al., (2017).

Materials and Methods

The ABC-RF

In the original formulation of ABC, the most used algorithm for model selection was based on the weighted multinomial logistic regression, introduced by M. A. Beaumont (2008). Under the logistic regression method, the estimation of the coefficients for the regression between a model indicator (response) variable and the simulated summary statistics (the explanatory variables) allowed the estimation of the posterior probability for each model at the intercept condition where observed and simulated summary statistics coincide. However, this algorithm suffers from two important limitations. First, to obtain reliable estimates of the models' posterior distribution, many simulations are necessary, making it difficult to analyse massive datasets with thousands of genomic loci. The second crucial point regards the selection of the vector of summary statistics to compare simulated and observed data, that has to be, at the same time, sufficiently informative and low-dimensional (Blum & François, 2010).

These important issues related to the conventional ABC framework were recently addressed by the

introduction of a paradigm shift in the model selection procedure, based on a Machine Learning procedure called random forest (RF, Pudlo et al., 2015). Under the RF approach, the model selection stage is rephrased as a classification problem. The Machine Learning classifier is constructed from the reference table, composed by a set of simulation records made of model indices and summary statistics for the associated simulated data. The reference table serves as training database for a RF that forecasts model index based on the summary statistics. This classification method has shown to be insensitive both to the correlations among summary statistics and to the presence of uninformative variables; moreover, it accommodates large dimensional summary statistics with no consequences on the estimation performances. Once the classifier is constructed, it is applied to the real data; the posterior probability of the selected model is then approximated from a secondary RF that regresses the selection error over the available summary statistics.

The *FDSS*

To compute the FDSS we evaluated the genomic distributions of the four mutually exclusive categories of segregating sites in two populations, namely (i) segregating sites private of the first population; (ii) segregating sites private of the second populations; (iii) segregating sites that are polymorphic in both populations; and (iv) segregating sites fixed for different alleles in the two populations (Wakeley & Hey, 1997). We considered the genome as subdivided in k independent fragments of length m, and for each fragment we counted the number of sites belonging to each of the four above-mentioned categories. This way, for a locus Lj and a fixed pair of populations we have the tuple $\{Lj_i, Lj_{ii}, Lj_{iii}, Lj_{iv}\}$ of the numbers of sites in each of the four categories. The final vector of summary statistics is composed of the truncated frequency distribution of loci having from 0 to n segregating sites in each category, for each pair of populations considered. The maximum number of segregating sites in a locus of length m is fixed to n (100 in our case), and hence the last category contains all the observations higher or equal to n. Specifically, for a fixed pair of populations, the summary statistics $SS_i(z)$, $SS_{ii}(z)$, $SS_{iii}(z)$, $SS_{iv}(z)$ are:

$$SS_A(x) = \sum_{j=1}^{k} I(Lj_A = x \vee (x = n \wedge Lj_A > x)),\ where\ x \in N, x \leq n, A \in \{i, ii, iii, iv\}$$

In the one-population models, we use a single truncated frequency distribution of within-population segregating sites in a locus; in this case we thus counted the number of genomic fragments carrying from 0 to n polymorphic sites. This statistic SS(z), is hence defined as:

$$SS(x) = \sum_{j=1}^{k} I(Lj = x \vee (x = n \wedge Lj > x)),\ where\ x \in N, x \leq n$$

Power Analysis

To determine the power of both the *FDSS* and the *SFS* in distinguishing among alternative

evolutionary trajectories, we simulated genetic data considering different experimental conditions. We tested all the possible combinations of locus length (bp) {200; 500; 1,000; 2,000; 5,000}, number of loci {1,000; 5,000; 10,000} and number of chromosomes {2, 4, 10, 20}, for a total of 60 combinations of sampling conditions tested. For each combination, we generated data with intra-locus recombination (recombination rate= $1x10^{-8}$), and with a fixed mutation rate ($1x10^{-8}$ /bp/generation). We evaluated the power considering three sets of models of increasing complexity, detailed below. The *FDSS* and the two *SFS* were calculated from the *ms* (Hudson, 2002) or *msms* (Ewing & Hermisson, 2010) output of each simulation through a in-house python script (available on github https://github.com/anbena/ABC-FDSS). For each combination of experimental conditions, we compared alternative models within the three sets tested treating each simulated dataset for each model as pseudo-observed data (pods). All the ABC-RF estimates have been obtained using the function *abcrf* from the package *abcrf* and employing a forest of 500 trees, a number suggested to provide the best trade-off between computational efficiency and statistical precision (Pudlo et al., 2015). We computed the confusion matrices and we evaluated the out-of-bag classification error (CE); for each comparison we then calculated the proportion of True Positives (TP) as 1-CE. The proportion of TP is thus a measure of the power of the whole inferential procedure, considering all its features (model selection approach, alternative models compared, statistics summarizing the data, genomic parameters simulated).

One-population models

We started by considering four demographic models (Fig. 1). The first model represents a constantly evolving population with an effective population size *N1*, drawn from a uniform prior distribution (Table S1). Under the second model, the population experienced a bottleneck of intensity *i*, *T* generations ago. The intensity and the time of the bottleneck, and the ancient effective population size *Na* are drawn from uniform prior distributions, showed in Table S1. The third model represents an expanding population. The expansion (of intensity *i*) is exponential and starts *T* generations ago, with the effective population size increasing from *N1/i* to *N1* (prior distributions in Table S1). Under the last model, the population is structured in different demes, exchanging migrants at a certain rate. The actual number of demes *d*, the migration rate *m* and the effective population size *N1* are drawn from prior distributions (Table S1).

Two-populations models

We then moved to considering three demographic models with two populations (Fig. 2). The first one is a simple split model without gene flow after the divergence. Under this model, an ancestral population of size *Nanc* splits *Tsep* generation ago into two populations. These two derived populations evolve with a constant population size (*N1* and *N2*) until the present time (priors for these free parameters

are shown in Table S2). The second model also includes a continuous and bidirectional migration, all the way from the divergence moment to the present. The per generation migration rates *m12* and *m21* are drawn from priors defined in Table S2. The third and last model assumes a single pulse of bidirectional admixture at time *Tadm* after divergence. Admixture rates *adm12 adm21*, and the time of admixture are drawn from priors (Table S2).

Multi-populations models

In most realistic cases, populations do interact with each other. Among the many possible scenarios, we chose to initially focus on the hypotheses proposed to explain the expansion of anatomically modern humans out of Africa. The basic alternative is between a single dispersal occurring along a Northern corridor (see e.g. Malaspinas et al., 2016) or two dispersal events, first along the so-called Southern route, and then through a Northern corridor (e.g. Pagani et al., 2016; Reyes-Centeno et al., 2014; Tassi et al., 2015). To design the models we followed the parametrization proposed by Malaspinas et al., (2016), with some minor modifications (Fig. 3). Both models share the main demographic structure: on the left the archaic groups (i.e. Neandertal, Denisova and an unknown archaic source), and on the right the anatomically modern humans (with a first separation between Africans and non-Africans and subsequent separations among population that left Africa). Given the evidence for admixture of Neandertals and Denisovans with non-African modern human populations (M. Meyer et al., 2012; Prüfer et al., 2014), we allowed for genetic exchanges from archaic to modern species, indicated in Fig. 3 by the colored arrows. The archaic populations actually sending migrants to modern humans are unknown, and hence here we used two ghost populations that diverged from the Denisovan and the Neandertal Altai samples 393 kya and 110 kya, respectively (Malaspinas et al., 2016). This way, we took into account that the archaic contributions to the modern gene pool did not necessarily come from the archaic populations that have been genotyped so far. We modeled bidirectional migration between modern populations along a stepping-stone, thus allowing for gene flow only between geographically neighboring populations. Under the Single Dispersal model (SDM) a single wave of migration outside Africa gave rise to both Eurasian and Austromelanesian populations, whereas under the Multiple Dispersal model (MDM) there are two waves of migration out of Africa, the first giving rise to Austromelanesians and the second to Eurasians. We took into account the presence of genetic structure within Africa modeling the expansion from a single unsampled "ghost" population under the SD model, and from two separated unsampled "ghost" populations for the MD model. The prior distributions for all the parameters considered in these models are in Tables S3 and S4.

We simulated both demographic models under all possible combinations of experimental parameters. We ran 50,000 simulations per model and combination of experimental parameters, using the *ms/msms* software.

Real Case: out of Africa dynamics

We explicitly compared SDM and MDM considering the high-coverage genomes of Denisova and Neandertal (M. Meyer et al., 2012; Prüfer et al., 2014), together with modern human samples from Pagani et al. (2016). A detailed description of the samples is in Table S5. All the individuals were mapped against the human reference genome hg19 build 37. To calculate the observed *FDSS* we only considered autosomal regions outside known and predicted genes +/- 10,000 bp and outside CpG islands and repeated regions (as defined on the UCSC platform, Hinrichs et al., 2016). We extracted 10,000 independent fragments of 500 bp length, separated by at least 10,000 bps in genomic regions that passed a set of minimal quality filters used for the analysis of the ancient genomes (map35_50%; M. Meyer et al., 2012; Prüfer et al., 2014). Power analysis (see *Results-Multi populations models* section), showed we could safely analyze a single individual (i.e. two chromosomes) per population. Therefore, each run of the analysis took into account the Denisova, the Neandertal, one African, one European one Asian and, in turn, either one out of six Papuans from Pagani et al. (2016) or one of 25 Papuans from Malaspinas et al. (2019) (detailed in Table S5). As for the Papuan genomes in Malaspinas et al. (2016), we downloaded the alignments in CRAM format from https://www.ebi.ac.uk/ega/datasets/EGAD00001001634. The *mpileup* and *call* commands from *samtools-1.6* (Li et al., 2009), were used to call all variants within the 10,000 neutral genomic fragments, using the *--consensus-caller* flag, without considering indels. We then filtered the initial call set according to the filters reported in Malaspinas et al. (2016) using *vcflib* and *bcftools* (Li et al., 2009). Each of the resulting 31 observed *FDSS* was separately analyzed through the ABC-RF model selection procedure. Finally, we checked whether the selected model is actually able to account for the observed variation through a Principal Component Analysis (PCA) of the simulated and observed data.

Real Case: Orangutan evolutionary history

We selected seven orangutan individuals, one from each of the populations defined by Nater et al. (2017), choosing the genomes with the highest coverage (Table S6). We downloaded the FASTQ files from https://www.ncbi.nlm.nih.gov/sra/PRJEB19688, and mapped the reads to the ponAbe2 reference genome (http://genome.wustl.edu/genomes/detail/pongo-abelii/) using the BWA-MEM v0.7.15 (Li & Durbin, 2010). We used picard-tools-1.98 (http://picard.sourceforge.net/) to add read groups and to filtered out duplicated reads from the BAM aligments. We performed local realignment around indels by the Genome Analysis Toolkit (GATK) v2.7-2 (Van der Auwera et al., 2013). To obtain genomic fragments suitable to calculate the *FDSS*, we generated a mappability mask (identified with the *GEM-mappability* module from the *GEM* library build, Derrien et al., 2012) so as to consider only genomic positions within a uniquely mappable 100-mer (up to 4 mismatches allowed). We then excluded from this mask all the exonic regions +/- 10,000 bp, repeated regions (as defined in the *Pongo abelii* Ensembl gene annotation release 78), as

well as loci on the X chromosome and in the mitochondrial genome. We then generated the final mask calculating the number of fragments separated by at least 10 kb, thus obtaining 9,000 fragments of 1,000 bp length. We called the SNPs within these fragments using the *UnifiedGenotyper* algorithm from *GATK*; the filtering step has been performed as reported in Nater et al. (2017) through *vcflib*. We finally calculated the observed *FDSS* from the quality filtered VCF file.

To investigate past population dynamics of the three Orangutan species, we designed competitive scenarios following the demographic models reported in Nater et al. (2017). We directly compared complex demographies, designing the within-species substructure as described by Nater et al. (2017), (Fig. 4A). The four competing models indeed share the same within-species features (four populations for the Bornean group, two Sumatran populations north of Lake Toba, and a single population south of Lake Toba), while differing for the tree topology, i.e. for the evolutionary relationships among the three species, as reported in Fig. 4A. We modeled bidirectional migration both among populations within a species, and between neighboring species. A detailed description of the models' parameters and of the priors are in Tables S7-S10. We ran 50,000 simulations per model using the *ms* software (Hudson, 2002), generating two chromosomes per population (4 Bornean, 1 south of Lake Toba and 2 north of Lake Toba), and 9,000 independent fragments of 1kb length per chromosome. We first assessed the power to distinguish among the four models calculating the proportion of TPs as described above, and then explicitly compared the simulated variation with the *FDSS* calculated on the observed data (Fig. 4B). Also in this case, the model checking has been performed through PCA.

Results

Power Analysis

One-population models

The four plots of Fig. 1B report the results of the power analyses obtained summarizing the data through the FDSS, whereas plots of Fig. 1C report the results obtained with the folded SFS. Being quite redundant, the results for the unfolded SFS are presented in Figure S1. In each plot, we reported the proportion of times each model was correctly recognized as the most likely one. For the FDSS, the percentage of true positives is quite high, ranging from almost 80% to 100% depending on the model generating the pod and on the combination of experimental conditions tested. The bottleneck model has the highest rate of identification, with most combinations of experimental conditions yielding nearly 100% true positives. By contrast, the least identifiable model seems the one considering a structured population, with 0.78 to 0.90 true positives. However, we observed that the decrease in the power is actually linked to the extent of gene flow among demes, and to the number of demes sampled; as rates of gene flow increase and the number of demes sampled decreases, the structured and the panmictic models converge, hence becoming harder to distinguish (Fig. S2). As expected, we observed a general increase in power with the

increase of both the locus length and the number of loci considered. By contrast, the number of sampled chromosomes does not appear to be directly linked to the increase of the proportion of true positives when the data are summarized through the FDSS. For some sampling conditions, we observed instead a decrease in the TP rate going from 2 to 20 chromosomes (see Fig. 1B). We showed that this behavior reflects the overlap of the FDSS generated by the constant and the structured models, an overlap increasing in parallel with the number of chromosomes sampled (Fig. S3). When sample size increases, indeed, the total branch length of coalescent trees is strongly influenced by the most recent part of the tree (see e.g. Wakeley & Aliacar, 2001), where the structured model behaves as a constant model because migration has not yet occurred and all lineages stay in the local deme where the data have been sampled. When the data were summarized through the SFS (both folded and unfolded) we observed, instead, significant differences in the proportion of true positives at increasing numbers of chromosomes sampled per population. When the number of chromosomes is between ten and twenty, the TP rate always ranges between 90 and 100% for all the models tested except for the structured one, which showed a slightly lower proportion of TP, between 85 and 95% (Fig. 1C, Figure S1). With only two chromosomes, and with four chromosomes for certain combination of experimental parameters, the percentage of TP only ranges between 70% and 85%. With the SFS we sometimes observed a decrease of the TP rate when considering more genetic loci, or longer locus lengths. This happened under the constant model (TP rate about 75%) and under the exponential model (TP rate about 80%).

Two-populations models

The plots in Fig. 2B, C and Figure S4 show the results for the two-populations models. When considering the *FDSS* the proportion of TP is generally quite high, with the Divergence with Migration and the Divergence with Admixture models showing the highest proportion of TP, reaching for many experimental conditions the 100%. For the Divergence model, the TP proportion is lower, ranging from 62 to 90%. Once again, the performance of the FDSS correlates with the number and the length of genetic loci, and not with the number of chromosomes. The folded and unfolded *SFS* do not show significant differences in their performance (Fig 2C and Figure S4), and we generally observed the same features emerging from the comparison of one-populations models. When only two chromosomes per population were considered the proportion of TP was between 60% and 65% for the Divergence model, between 72% and 82% for the Divergence with Migration model, and between 55% and 78% for the Divergence with Admixture model. With more chromosomes sampled we observed an increase in the TP rate, until reaching the values achieved with the *FDSS*. Both folded and unfolded SFS seem not to be sensitive to the number of loci, nor to their length.

Multi-populations models

Fig. 3B, C and Figure S5 summarize the power analysis comparing SDM and MDM. For the *FDSS* the proportion of true positives ranges between 0.65 and 0.70 for the SDM, and between 0.65 and 0.8 for the MDM, in this case with a slight increase of the power with the size of the fragments simulated and the number of loci simulated. Because the SDM and the MDM share several features, in particular when under MD the time interval between the first and second exit is short, we also evaluated the ability of the *FDSS* to be informative about the correct model as a function of this interval. To do this, we considered 10,000 pods from the MDM. We then subdivided these 10,000 pods in 6 bins of increasing interval between these two events (up to 60,000 years), measuring, within each bin, the proportion of times in which the MDM is correctly recognized by the ABC-RF procedure. As might be expected, the proportion of true positives increases with increasing time intervals (Fig. S6), reaching values of 90% for some combinations of experimental parameters. When the data are summarized through the *SFS* the proportion of TP reach 75% for the SDM and 0.8 for the MDM. In this case the highest proportions of TP are observed for twenty chromosomes, with negligible or null impact of the number of genetic loci or locus length.

Real Case: out of Africa dynamics

Simulations in the previous section show that alternative models can be distinguished using the *FDSS* to summarize the data, except when the difference between them becomes so small that the models overlap. Interestingly, the success of *FDSS* in distinguishing models does not seem to depend on the number of chromosomes analyzed; a single individual sampled per population shows a comparable discrimination power as twenty chromosomes. Thus, it seems that ABC models comparison through *FDSS* is particularly suited for small sample sizes, e.g. in studies of ancient DNA. To further explore this feature we applied the *FDSS* to estimate posterior probabilities of alternative models about early human expansion from Africa. Whether human demographic history is better understood assuming one (Malaspinas et al., 2016; Mallick et al., 2016) or two (Pagani et al., 2016; Reyes-Centeno et al., 2014; Tassi et al., 2015) major episodes of African dispersal is still an open question. While concluding that indigenous Australians and Papuans seem to derive their ancestry from the same African wave of dispersal as most Eurasians, Mallick et al. (2016) admitted that these inferences change depending on the computational method used for phasing haplotypes. Therefore, it made sense to compare the SDM and the MDM through our ABC approach. The proportion of true positives for the combination of experimental parameters here considered (i.e. 10,000 loci of 500 bp length and 2 chromosomes per population) was 0.68 for the SDM, and 0.74 for the MDM (Fig. 3A).

Regardless of the Papuan individual considered in each run of 31 replicated experiments, the results always supported the MDM, with posterior probabilities ranging from 0.74 to 0.76 for the Pagani et

al. (2016) genomes, and from 0.69 to 0.74 for the Malaspinas et al. (2016) genomes (Fig. 5 and Tables S11-S12), The PCA of the simulated and observed data shown in Figure S7 confirms that the MDM is able to reproduce the genetic variation found in real data.

Real Case: Orangutan evolutionary history

As a second application, we investigated the past demographic and evolutionary dynamics of the orangutan. In addition to the two species previously recognized in Borneo (*Pongo pygmeus*) and in Sumatra, North of Lake Toba (*Pongo abelii*), Nater et al. (2017) described a new species of Sumatran orangutan, *Pongo tapanuliensis*, South of Lake Toba. To reduce the otherwise excessive computational effort in their ABC analysis, Nater et al. (2017) had to resort to an ad-hoc procedure, incorporating factors such as bottlenecks and population structure only after comparing simplified versions of their models; this raises questions on the robustness of the conclusions thus reached. As we saw, the ABC-RF approach can handle complex model comparisons, and the analysis of a single individual per population further accelerates the simulation step. We first assessed the ability to correctly recognize the four models through a power analysis (Fig. 4A). The most identifiable model (TP=0.802) appeared to be the model 2b, under which there is a first separation of South Toba from Borneo Orangutan, followed by the divergence of North Toba from South Toba. The model assuming an early separation of South Toba form North Toba, followed by the separation of Borneo from South Toba, actually showed the lowest proportion of true positives (0.480). The application to real data favored the model 1a, (also associated with the highest posterior probability in Nater et al., 2017), with a posterior probability of 0.49. Under the most supported model both the North Toba (first) and Borneo (later) separated from *Pongo tapanuliensis* (Fig. 4B). Model 1a also proven to be able to account for real variation, as it is shown in Figure S8.

Discussion

The cost of genotyping has dramatically dropped lately, making population-scale genomic data available for a large set of organisms (1000 Genomes Project Consortium, 2012; Benazzo et al., 2017; Dasmahapatra et al., 2012; De Manuel et al., 2016; Miller et al., 2012). The main challenge now is how to extract as much information as possible from these data, developing flexible and robust statistical methods of analysis (Excoffier et al., 2013; Li & Durbin, 2011; Schiffels & Durbin, 2014). Approximate Bayesian Computation, explicitly comparing alternative demographic models and estimating the models' probabilities, represents a powerful inferential tool about past demographic events (Mark A. Beaumont, 2010). One of the main advantages of such a simulation-based approach is the possibility to easily check whether the models being compared are actually distinguishable, hence quantifying the reliability of the estimates produced (Csilléry et al., 2010). Nevertheless, despite few successful attempts (Boitard, Rodríguez, Jay, Mona, & Austerlitz, 2016), only recently, with the development of the Random Forest

procedure for ABC model selection (Pudlo et al., 2015), it has become possible to definitely overcome the issues linked to the use of uninformative/correlated summary statistics, and to significantly reduce the computational effort of the simulation step. With this work, we took advantage of this newly proposed algorithm to test the flexibility of an ABC-based framework in comparing different demographic models. As customary, we summarized the data through the folded and unfolded version of the *SFS*, but the novelty of this work lies in the use of the *FDSS*, namely the complete genomic distribution of the four mutually exclusive categories of segregating sites for pairs of populations (Wakeley & Hey, 1997).

Power Analysis

Initially, we analyzed sets of models with increasing levels of complexity, simulating genetic data under a broad spectrum of experimental conditions. This extensive power analysis showed that both the *SFS* and the *FDSS* allow one to often recognize the model under which the data were generated, with some uncertainties only when two models are just marginally different. This was the case for both simple (one or two-population scenarios, Figs 1 and 2) and complex (multi-populations scenarios, Fig. 3) demographies. When we compared one-population scenarios, the *FDSS* is necessarily composed only by a single distribution, representing the frequency of genomic fragments carrying a certain number of polymorphic sites. Nonetheless the model identifiability, calculated as the proportion of TPs over 50,000 pods, reached values between 80% and 100%, with slightly lower values only for the structured model. This reduction in power was always due to the levels of gene flow among demes; when it is high, the structured model tends to panmixia (Fig. S2), as has already been known since Wright's times (Wright, 1931). We also showed that the power depends on the number of demes; indeed, the proportion of TPs increases in parallel with the number of demes considered in the structured model (Fig. S2).

Among the two-populations demographies, the models with bi-directional migration at a constant rate and with pulse of admixture proved easiest to identify, with almost 100% TPs, regardless of the combination of experimental parameters tested. With the *FDSS* we obtained lower TP rates (about 70-80%) only when using 1,000 short loci, whereas with the *SFS* the proportion of TP correlates with the number of chromosomes used.

Even when rather complicated scenarios were compared (e.g. the multi-populations models), the rate of accurate results is close to 70% TPs. As expected, when processes occur at short time distances, they are difficult to discriminate. When, under MDM, the two expansions from Africa are simulated at very close times, the SDM and the MDM models become extremely similar. Accordingly, we observed an increase in the power of the test at increasing intervals between the African divergence and the second exit (Fig. S6), reaching values close to 90%.

We also tested whether using the complete frequency distribution of the four categories of

segregating sites actually entails an advantage respect to the use of its summary (as e.g. in Robinson et al., 2014), comparing one, two and multi-populations models through the first two moments of the four distributions. The results, reported in Figs S9-S11, are significantly in favor of the use of the full distribution, and increasingly so with the complexity of the models, in particular when few chromosomes (two or four) or short locus lengths are analyzed.

Comparison between *SFS* and *FDSS*

In general, our results showed that both the (folded and unfolded) *SFS* and the *FDSS* obtained good discrimination power, regardless of the complexity of the models being compared. Going into detail, the *FDSS* shows a better performance with respect to the *SFS* when few chromosomes per population (i.e. two or four) are available, as emerged in particular from the analysis of one- and two-populations models. Under these models the dimensionality of the folded *SFS* for two or four chromosomes is often lower than the number of models' parameters, possibly making it difficult to discriminate among the demographic scenarios tested. On the other hand, when tens of chromosomes may be analyzed, the *SFS* seem to be the better choice to summarize the data. Considering the *FDSS*, the accuracy of the model selection seems to be more dependent on the number of loci considered and on the locus length rather than on the number of individuals sampled per population. As opposed to the *SFS*, the *FDSS* is then a suitable summary of whole genome data for ABC-RF analysis of even suboptimal datasets, such as those coming from the study of ancient DNA data, or of elusive species. Moreover, when dealing with highly complex models, the simulation of a small number of chromosomes also reduces the computational costs of the simulation step.

The performances of the folded and unfolded SFS are comparable, with a slight increase in the power of the unfolded spectrum for some specific conditions (usually when considering four chromosomes) or demographic model analyzed (as one-populations models or MDM). However, we should remind that we generated the unfolded SFS through simulations, thus assuming that the ancestral state of alleles is known with certainty. When analyzing real data the spectrum instead needs to be polarized, meaning that the ancestral and derived alleles have to be defined using an outgroup, where the outgroup allele is typically taken as ancestral under parsimony assumption. Parallel changes or peculiar features of the demographic structure of the outgroup population (i.e. structured population) could introduce a bias in the definition of ancestral states, leading to a skew toward sites with a high frequency of the derived state and, therefore, potentially generating inaccurate demographic signals (Baudry & Depaulis, 2003; Hernandez et al., 2007; Morton, Dar, & Wright, 2009). It is anyway worth noting that this is not the case for the *FDSS*, which may be calculated from the number of polymorphic sites across populations, without further assumptions on the state of alleles.

Applications to real datasets

We finally analyzed two demographic models about the anatomically modern human expansion out of Africa, combining ancient and modern genome data. The former (Neandertal and Denisova, in our case) are characterized by highly fragmented DNA, and so, we restricted the analysis to short DNA stretches (500 bp) to maximize the number of independent loci retrievable. Despite this limitation, even with 2 chromosomes per population we obtained a good ability to tell models apart (Fig. 3). Thirty-one replicated experiments, differing for the Papuan genome being considered, consistently supported the MDM over the SDM (Fig. 5), i.e. a first expansion from Africa of the ancestors of the current Austro-Melanesians, followed by a second expansion leading to the peopling of Eurasia. Considering different modern individuals from African, European and Asian populations did not change the support for the MDM. These results raise several questions; indeed, it was the SDM that showed the best fit in Malaspinas et al. (2016), whereas the MDM appeared to account for the data only when the analysis was restricted to modern populations. However, our findings are in agreement with those by Pagani et al. (2016), who estimated that at least 2% of the Papuan genomes derive from an earlier, and distinct, dispersal out of Africa. Other genomic studies (Tassi et al., 2015), but not all (Mallick et al., 2016), and phenotypic analyses (Reyes-Centeno et al., 2014) appear in closer agreement with the MDM, which calls for further research in this area. Note that Malaspinas and collaborators argued that apparent support for multiple dispersal events really came from the confounding effect of Denisovan admixture in the Australian-Papuans' ancestors; however, both in this and in a previous (Tassi et al., 2015) study, we found statistically-significant support for the MDM after correcting for possible Denisovan admixture. Be that as it may, in no other study besides the present one (i) the alternative hypotheses are explicitly compared analyzing complete genomes; (ii) posterior probabilities are estimated for each model, and (iii) the accuracy of the estimates is assessed by power analysis.

We then moved to investigating the evolutionary history of the three extant Orangutan species. We basically improved the ABC analysis performed by Nater et al. (2017) summarizing the data through *FDSS*, sampling a single individual per population, and applying the ABC-RF model selection framework. Nater and colleagues (2017) started comparing simplified evolutionary scenarios, and considered population substructure and gene flow only when estimating parameters, but not in the phase of model choice. ABC-RF allowed us to avoid this uncertain procedure, confirming Nater et al.'s (2017) conclusion that the first split separated the North Toba and the newly identified South Toba species (Fig. 4B). The main difference was about the strength of the support associated to this model. While Nater and colleagues (2017) estimated high posterior probabilities for the best-fitting model (73% when comparing the 4 models and 98% when comparing the two best scenarios), our procedure associated to the same model a posterior probability of 49% (Fig. 4B). Moreover, the power analysis that we conducted (absent in Nater et al., 2017), revealed that the ability to correctly distinguish among the four tested models is between 48% and 80%, with the selected model that can be erroneously recognized as the most probable one in the 38% of cases. Although model 1a

has been selected as the most supported scenario, the uncertainty emerged from the classification error suggests that the true evolutionary history of Orangutan species is still largely unknown. These results emphasize (i) the importance of including complex demographic histories in the model selection step, so as to evaluate the real posterior probability associated to the best model, on which the parameter estimation will be performed and (ii) the importance of performing a power analysis of the models tested, so as to be aware of the level of uncertainty about the conclusions of the study.

Conclusions

In this paper we showed that ABC-RF can often reconstruct a complex series of demographic processes, based both on the *SFS* and on the *FDSS*. The *FDSS* generally exhibited better performance when few chromosomes per populations were analyzed; this feature, together with the ease of estimation from whole genome data without further assumptions, makes this statistic particularly suitable for demographic inference through an ABC approach. It is also worth noting that the power to correctly identify the true model was quite good when we simulated short fragments, even in the comparison of complex demographies (Fig. 3). This finding means that the ABC-RF model selection procedure through *FDSS* or *SFS* is suitable for the analysis of ancient data (M. Meyer et al., 2012) and of RAD sequencing data (Rowe, Renaut, & Guggisberg, 2011), where short DNA fragments are more the rule than the exception.

In all our analyses we considered the *FDSS* or the *SFS* as calculated from known genotypes, meaning that the presented procedure is currently optimized for high-coverage data (De Manuel et al., 2016; Mallick et al., 2016; Miller et al., 2012). A natural extension of this work will thus be to implement the use of low coverage data, developing an approach able to retrieve the *FDSS* taking into account the genotype uncertainty and sequencing errors, for instance through the use of the genotype likelihoods (as, e.g., in ANGSD, Korneliussen, Albrechtsen, & Nielsen, 2014).

The flexibility of the ABC-RF model selection approach, combined with the inferential power proven by the summary statistics that we proposed to calculate on genomic data, may contribute to a detailed and comprehensive study of complex demographic dynamics for any species for which few high coverage genomes are available.

Acknowledgments

We would like to thank the DFG Center for Advanced Studies, ‘‘Words, Bones, Genes, Tools,’’ at University of Tübingen, that hosted AB and SG during the first phase of the project, and Igor Yanovich for his help with the revision. This study has been supported by “Fondo per l’Incentivazione alla Ricerca” (FIR) 2018 of the University of Ferrara. FT was supported by ERC Advanced Grant 295733, ‘LanGeLin’.

References

1000 Genomes Project Consortium. (2012). An integrated map of genetic variation. *Nature*, *492*, 56–65. doi: 10.1038/nature11632

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., & Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*, *2*(10), e286. doi: 10.1371/journal.pbio.0020286

Baudry, E., & Depaulis, F. (2003). Effect of Misoriented Sites on Neutrality Tests with Outgroup. *Genetics*, *165*(3), 1619–1622.

Beaumont, M. A. (2008). Joint determination of topology, divergence time, and immigration in population trees. In *Simulations, Genetics and Human Prehistory* (pp. 135–154).

Beaumont, Mark A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 379–406. doi: 10.1146/annurev-ecolsys-102209-144621

Beaumont, Mark A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035. doi: Genetics December 1, 2002 vol. 162 no. 4 2025-2035

Beeravolu, C. R., Hickerson, M. J., Frantz, L. A. F., & Lohse, K. (2018). ABLE: blockwise site frequency spectra for inferring complex population histories and recombination. *Genome Biology*, *19*(1), 145. doi: 10.1186/s13059-018-1517-y

Benazzo, A., Trucchi, E., Cahill, J. A., Maisano Delser, P., Mona, S., Fumagalli, M., … Bertorelle, G. (2017). Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. *Proceedings of the National Academy of Sciences*, *114*(45), E9589–E9597. doi: 10.1073/pnas.1707279114

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. doi: 10.1111/j.1365-294X.2010.04690.x

Blum, M G B, Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Statist. Sci.*, *28*, 189–208. doi: 10.1214/12-STS406

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

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLoS Genetics*, *12*(3), e1005877. doi: 10.1371/journal.pgen.1005877

Chen, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology*, *81*(2), 179–195. doi: 10.1016/j.tpb.2011.11.004

Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., & Mazet, O. (2018). The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: Insights into demographic inference and model choice. *Heredity*, *120*, 13–24. doi: 10.1038/s41437-017-0005-6

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

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., … Jiggins, C. D. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*(7405), 94–98. doi: 10.1038/nature11041

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. doi: 10.1126/science.aag2602

Derrien, T., Estellé, J., Sola, S. M., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability. *PLoS ONE*, *7*(1), e30377. doi: 10.1371/journal.pone.0030377

Eldon, B., Birkner, M., Blath, J., & Freund, F. (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger Coalescents? *Genetics*, *199*(3), 841–856. doi: 10.1534/genetics.114.173807

Ewing, G., & Hermisson, J. (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, *26*(16), 2064–2065. doi: 10.1093/bioinformatics/btq322

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, *9*(10), e1003905. doi: 10.1371/journal.pgen.1003905

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2010). Diffusion Approximations for Demographic Inference : DaDi. *Nature Precedings*. doi: 10.1038/NPRE.2010.4594.1

Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2007). Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Molecular Biology and Evolution*, *24*(8), 1792–1800. doi: 10.1093/molbev/msm108

Hinrichs, A. S., Raney, B. J., Speir, M. L., Rhead, B., Casper, J., Karolchik, D., … Kent, W. J. (2016). UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics*, *32*(9), 1430–1432. doi: 10.1093/bioinformatics/btv766

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

Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, *206*, 1549–1567. doi: 10.1534/genetics.117.200493

Kamm, J. A., Terhorst, J., & Song, Y. S. (2017). Efficient Computation of the Joint Sample Frequency Spectra for Multiple Populations. *Journal of Computational and Graphical Statistics*, *26*(1), 182–194. doi: 10.1080/10618600.2016.1159212

Keightley, P. D., & Jackson, B. C. (2018). Inferring the probability of the derived vs. The ancestral allelic state at a polymorphic site. *Genetics*, *209*(3), 897–906. doi: 10.1534/genetics.118.301120

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. doi: 10.1186/s12859-014-0356-4

Lapierre, M., Lambert, A., & Achaz, G. (2017). Accuracy of demographic inferences from the site frequency spectrum: The case of the yoruba population. *Genetics*, *206*(1), 439–449. doi: 10.1534/genetics.116.192708

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. doi: 10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi:

10.1093/bioinformatics/btp352

Lohmueller, K. E. (2014). The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLoS Genetics*, *10*(5), e1004379. doi: 10.1371/journal.pgen.1004379

Malaspinas, A. S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., … Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214. doi: 10.1038/nature18299

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. doi: 10.1038/nature18964

Marjoram, P., Molitor, J., Plagnol, V., & Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0306899100

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: Instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity*, *116*(4), 362–371. doi: 10.1038/hdy.2015.104

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1387–1393. doi: 10.1098/rstb.2005.1673

Meyer, D., Single, R. M., Mack, S. J., Erlich, H. A., & Thomson, G. (2006). Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics*, *173*(4), 2121–2142. doi: 10.1534/genetics.105.052837

Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., … Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222–226. doi: 10.1126/science.1224344

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., … Lindqvist, C. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences*, *109*(36), E2382-90. doi: 10.1073/pnas.1210506109

Moreno-Mayar, J. V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Chan, J., Spence, J. P., … Willerslev, E. (2018). Early human dispersals within the Americas. *Science*, *362*(6419), eaav2621. doi: 10.1126/science.aav2621

Morton, B. R., Dar, V.-N., & Wright, S. I. (2009). Analysis of Site Frequency Spectra from Arabidopsis

with Context-Dependent Corrections for Ancestral Misinference. *Plant Physiology*, *149*(2), 616–624. doi: 10.1104/pp.108.127787

Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., de Manuel, M., Desai, T., … Krützen, M. (2017). Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Current Biology*, *27*(22), 3576–3577. doi: 10.1016/j.cub.2017.09.047

Pagani, L., Lawson, D. J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., … Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, *538*(7624), 238–242. doi: 10.1038/nature19792

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. doi: 10.1038/nature12886

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2015). Reliable ABC model choice via random forests. *Bioinformatics*, *32*(6), 859–866. doi: 10.1093/bioinformatics/btv684

Reyes-Centeno, H., Ghirotto, S., Detroit, F., Grimaud-Herve, D., Barbujani, G., & Harvati, K. (2014). Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proceedings of the National Academy of Sciences*, *111*(20), 7248–7253. doi: 10.1073/pnas.1323666111

Robinson, J. D., Bunnefeld, L., Hearn, J., Stone, G. N., & Hickerson, M. J. (2014). ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology*, *23*(18), 4458–4471. doi: 10.1111/mec.12881

Rowe, H. C., Renaut, S., & Guggisberg, A. (2011). RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, *20*(17), 3499–3502. doi: 10.1111/j.1365-294X.2011.05197.x

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. doi: 10.1038/ng.3015

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. doi: 10.1111/mec.14223

Tassi, F., Ghirotto, S., Mezzavilla, M., Vilaça, S. T., De Santi, L., & Barbujani, G. (2015). Early modern human dispersal from Africa: Genomic evidence for multiple waves of migration. *Investigative Genetics*, *6*, 6–13. doi: 10.1186/s13323-015-0030-2

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309. doi: 10.1038/ng.3748

Terhorst, J., & Song, Y. S. (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, *112*(25), 7677–7682. doi: 10.1073/pnas.1503717112

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43

Wakeley, J., & Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics*, *159*(2), 893–905.

Wakeley, J., & Hey, J. (1997). Estimating ancestral population parameters. *Genetics*, *145*(3), 847–855. doi: 10.1111/j.1365-294X.2011.05413.x

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97–159.

Figure Legends

Fig 1. One-population models and proportion of True Positives. A) Demographic models compared: Constant, Bottleneck, Expansion, Structured population. N_1 is the effective population size, i the intensity of the bottleneck or of the expansion, T the time of the bottleneck or of the start of the expansion, m is the migration rate. B) True Positives rates for the *FDSS*. C) True Positives rates for the *folded SFS*.
The plot below each of the four models represents the proportion of TPs obtained analyzing pods coming from the above model under 60 combinations of experimental parameters. Different locus lengths are in the x-axes, number of loci is represented by different colors and the number of chromosomes is represented by different symbols.

Fig 2. Two-populations models and proportion of True Positives. A) Demographic models compared: Divergence with isolation, Divergence with migration, Divergence with a single pulse of admixture. N_{anc} is the effective population size of the ancestral population, N_1 and N_2 are the effective population sizes of the diverged populations, T_{sep} is the time of the split, m_{12} and m_{21} the migration rates, T_{adm} is the time of the single pulse of admixture, adm_{12} and adm_{21} the proportions of admixture. B) True Positives rates for the *FDSS*. C) True Positives rates for the *folded SFS*. The plots have the same features of Fig 1.

Fig 3. Multi-populations models and proportion of True Positives. A) Demographic models compared: Single Dispersal and Multiple Dispersals. The populations sampled are indicated in bold. B) True Positives rates for the *FDSS*. C) True Positives rates for the *folded SFS*. The plots have the same features of Fig 1.

Fig 4. Demographic models tested to study the evolutionary history of Orangutan species. A) Four demographic models compared. The numbers in the black boxes indicate the proportion of TP calculated analyzing 50,000 pods coming from that demographic model. NT, Sumatran populations north of Lake Toba; ST, the Sumatran population south of Lake Toba; BO, Bornean populations. B) Number of votes associated to each model by ABC-RF and posterior probability of the most supported model (model 1a).

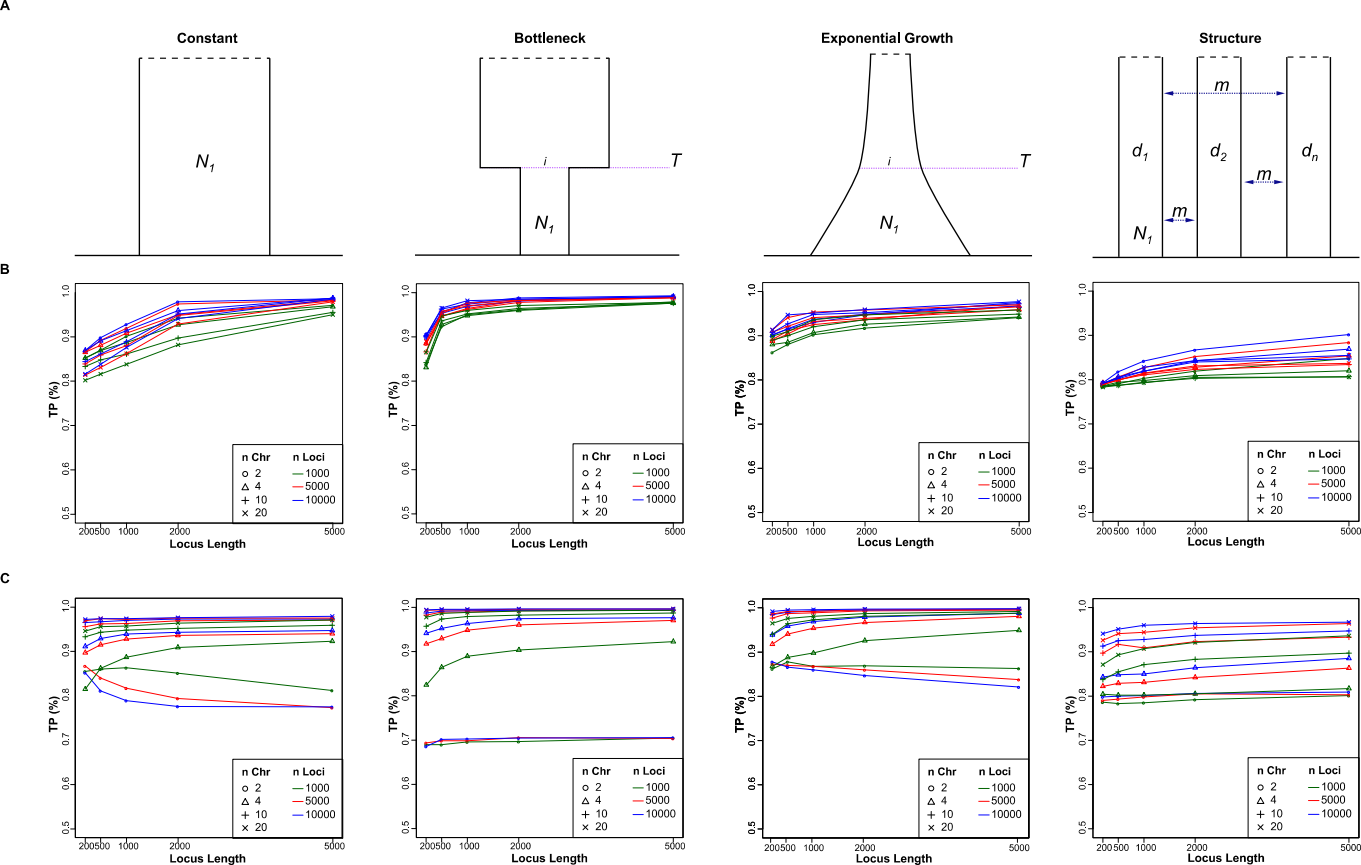
Fig 5. Posterior Probabilities for the MDM. Left panel: posterior probabilities obtained analyzing 6 Papuan individuals from Pagani et al. (2016) (PR). Right panel: posterior probabilities obtained analyzing 25 Papuan individuals from Malaspinas et al. (2016) (MR).

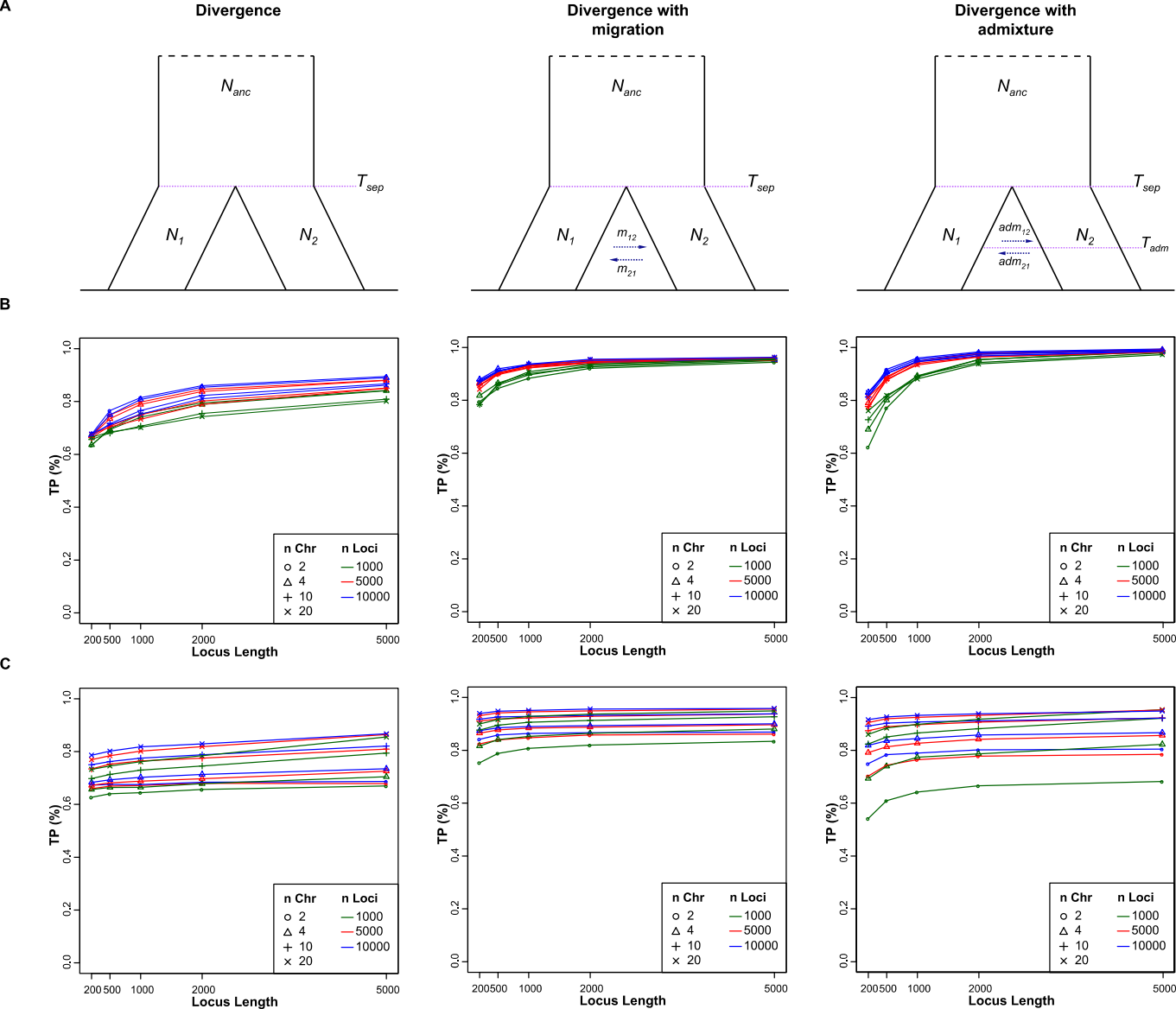
Data Accessibility

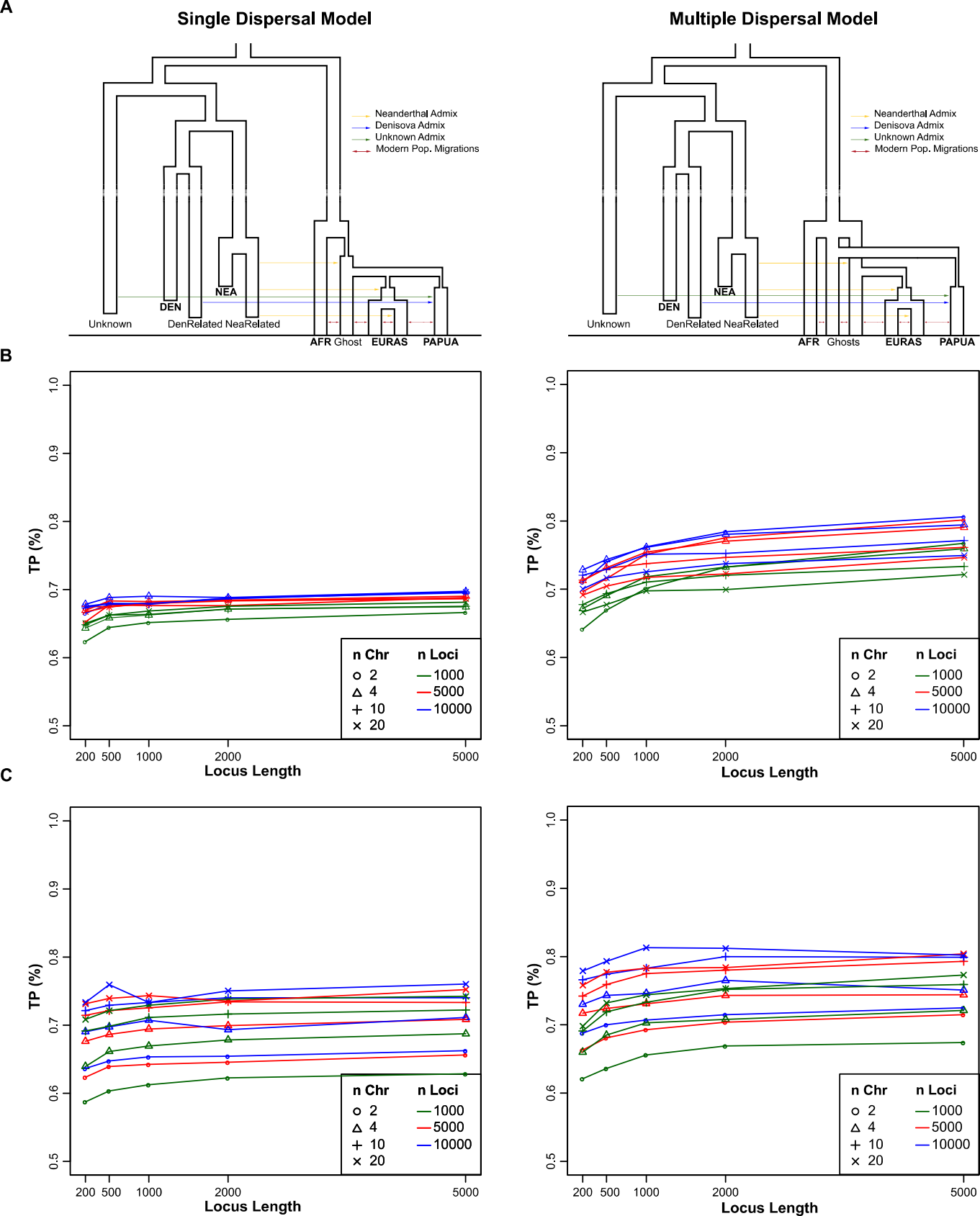
All the scripts used or produced by the authors can be found at https://github.com/anbena/ABC-FDSS.

Author Contributions

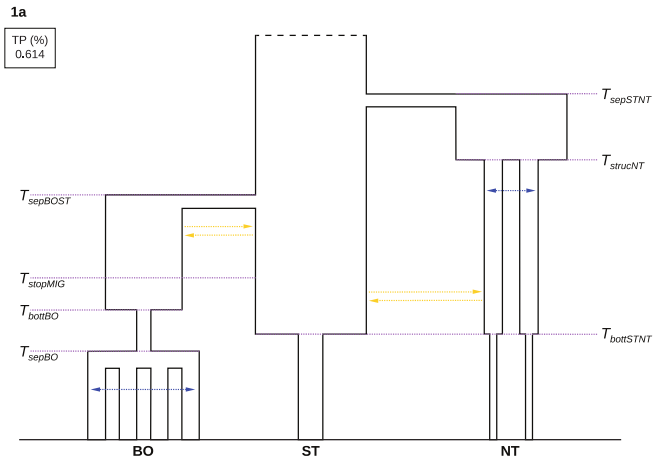
AB conceived the study; AB and SG designed the experiments; MTV, AB, SG and FT analyzed the data; SG, MTV, FT, GB and AB discussed the results; SG, GB and AB wrote the paper with inputs from all coauthors.

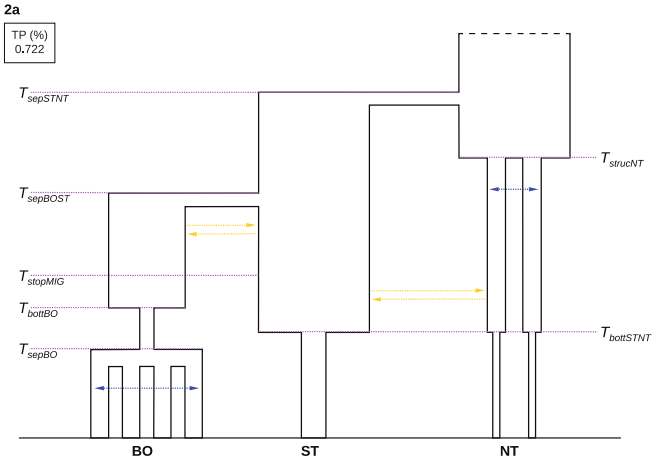

A
Constant
Bottleneck
Exponential Growth
Structure
N_1
i
T
N_1
i
T
N_1
m
d_1
d_2
d_n
m
m
B
C
TP (%)
Locus Length
n Chr
n Loci
2
4
10
20
1000
5000
10000
200
500
1000
2000
5000

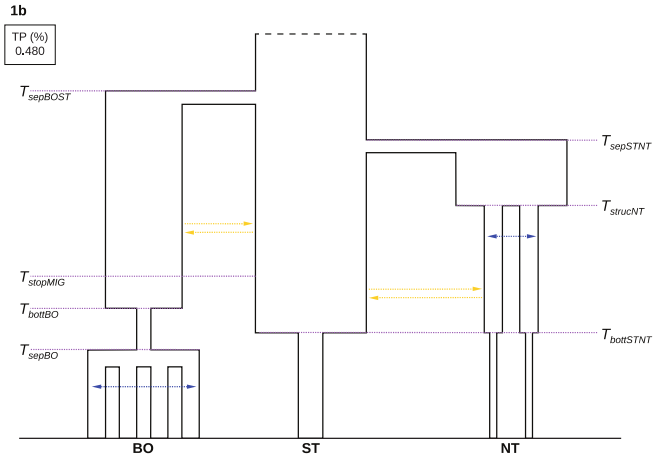

A
Divergence
Divergence with migration
Divergence with admixture
N_{anc}
T_{sep}
N_1
N_2
m_{12}
m_{21}
adm_{12}
adm_{21}
T_{adm}
B
C
TP (%)
Locus Length
n Chr
n Loci
2
4
10
20
1000
5000
10000
200
500
1000
2000
5000

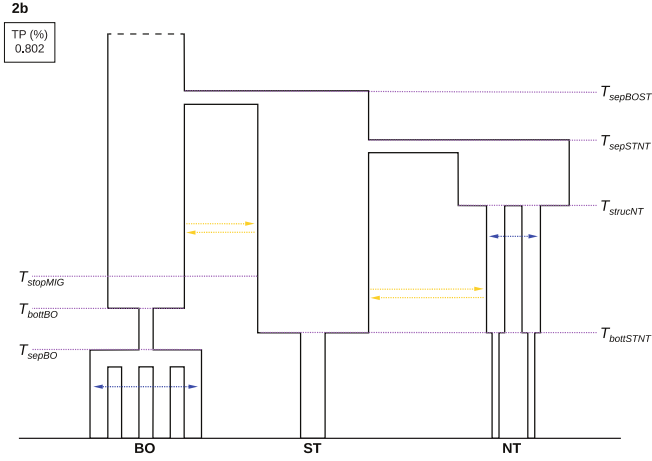

A
Single Dispersal Model
Multiple Dispersal Model
Neanderthal Admix
Denisova Admix
Unknown Admix
Modern Pop. Migrations
Unknown
DEN
NEA
DenRelated
NeaRelated
AFR
Ghost
Ghosts
EURAS
PAPUA
B
C
TP (%)
Locus Length
n Chr
n Loci
2
4
10
20
1000
5000
10000
200
500
1000
2000
5000
0.5
0.6
0.7
0.8
0.9
1.0

A



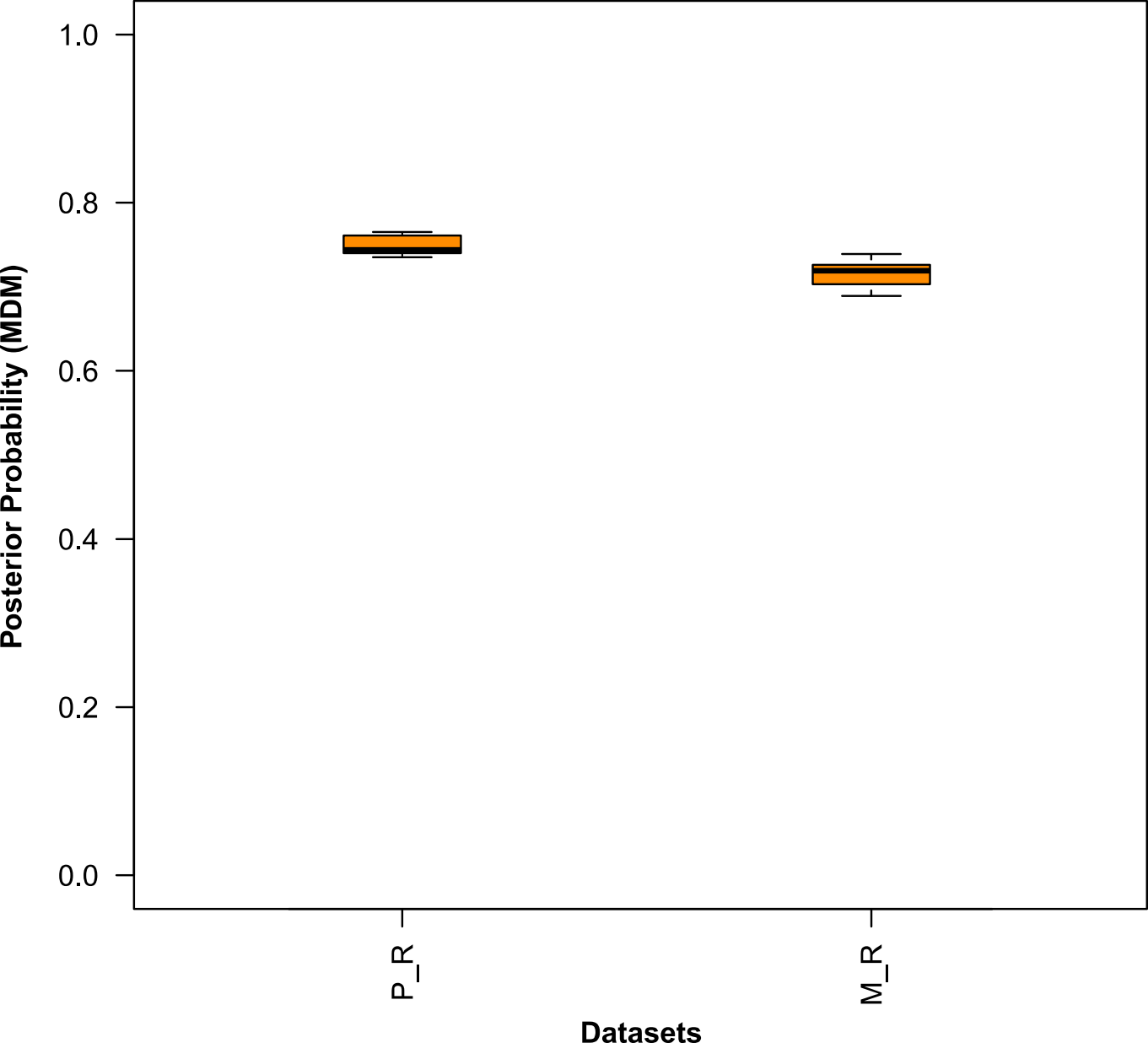






B

Selected Model	Votes model 1A	Votes model 2A	Votes model 1B	Votes model 2B	PP
1A	0.398	0.190	0.292	0.120	0.489


1.0
0.8
0.6
0.4
0.2
0.0
Posterior Probability (MDM)
P_R
M_R
Datasets