Università
degli Studi
di Ferrara

DOTTORATO DI RICERCA
in
Biologia Evoluzionistica ed Ecologia

In convenzione con:
Università degli Studi di Parma
Università degli Studi di Firenze

CICLO XXXIII

COORDINATORE Prof. Guido Barbujani

Supervised Machine Learning and ABC for population genetic inference

Settore Scientifico Disciplinare BIO/18

Dottoranda

Dott.ssa Maria Teresa Vizzari

Tutore

Prof.ssa Silvia Ghirotto

Cotutore

Dott. Andrea Benazzo

Anni 2017/2020

Abstract- English

In this PhD dissertation I outline the work that I did over three years, which so far has led to the publication of two papers in peer-reviewed journals. All of these studies focus on the development of a new ABC framework, based on a machine-learning tool named Random Forest, that allow the analysis of complete genome datasets to make inference about the past evolutionary processes characterizing natural populations.

Inferring past demographic histories is crucial in population genetics, and the amount of complete genomes now available should in principle facilitate this process. 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.

In this thesis I 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, I 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.

In **Chapter 4** I tested how accurately the proposed pipeline 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). Once assessed the inferential power of the ABC-RF procedure, I finally analyzed high-quality whole-genome 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.

I then extended the framework making it able to deal with low-coverage complete genomes. The low sequencing depth drastically affects the ability to reliably call genotypes, thus making low-coverage data unsuitable for inferential approaches like ABC. In **Chapter 5**, I present the results of the power analysis carried out with whole-genome datasets sequenced at different coverage levels (from 1x to 30x). I evaluated the inferential power of this procedure in distinguishing among different demographic models and in inferring model parameters. Under this approach, the *FDSS* is not directly calculated from known genotypes, but rather estimated using genotype likelihoods, so as to take into account the uncertainty linked to low-depth data in the estimation of the pattern of polymorphisms, making the simulated data directly comparable with those observed in low coverage experiments.

The inferential approaches presented in this thesis can be effectively used to analyze large panels of high- and low-coverage genomes from real populations, maximizing the information extracted from the data, in order to reconstruct complex past population dynamics.

Abstract - Italiano

Questa tesi riassume il lavoro di ricerca da me svolto durante i tre anni del dottorato, che finora ha portato alla pubblicazione di due articoli su riviste scientifiche. Questi studi sono incentrati sullo sviluppo di un nuovo framework ABC, basato su un algoritmo di machine-learning chiamato Random Forest, che consenta l'analisi di dati genomici completi per indagare i processi evolutivi passati che caratterizzano le popolazioni naturali.

L'inferenza delle dinamiche demografiche passate è cruciale negli studi di genetica delle popolazioni e la grande quantità di genomi completi ad oggi disponibile dovrebbe, in linea di principio, facilitare questo processo. In pratica, tuttavia, i metodi inferenziali disponibili soffrono di gravi limitazioni. Sebbene centinaia di genomi completi possano essere analizzati contemporaneamente, i processi demografici complessi possono facilmente superare i vincoli computazionali e le procedure per valutare l'affidabilità delle stime contribuiscono ad aumentare ulteriormente le risorse di calcolo richieste per le analisi.

In questa tesi presento un framework ABC basato sull'algoritmo di machine-learning Random Forest (ABC-RF), per inferire processi demografici passati, anche complessi, attraverso l'analisi di genomi completi. A questo scopo, propongo di riassumere i dati tramite la distribuzione genomica completa di quattro categorie di siti segreganti (*FDSS*), una statistica veloce da calcolare anche da dati genomici non fasati e che non richiede la conoscenza dello stato ancestrale degli alleli.

Nel **Capitolo 4** ho verificato con quanta accuratezza la pipeline proposta consenta di discriminare tra modelli di complessità crescente, utilizzando dati simulati e tenendo conto di diverse strategie di campionamento (in termini di numero di individui analizzati, numero e dimensione dei loci genetici considerati). Una volta valutato il potere inferenziale della procedura ABC-RF, ho analizzato diversi dataset di genomi completi di alta qualità per testare i modelli sulla dispersione degli uomini anatomicamente moderni fuori dall'Africa ed esplorare le relazioni evolutive delle tre specie di orango che abitano il Borneo e Sumatra.

Ho quindi esteso il framework rendendolo in grado di gestire anche genomi completi a bassa copertura. La bassa profondità di sequenziamento influisce drasticamente sulla capacità di identificare in modo affidabile i genotipi, rendendo così i dati a bassa copertura inadatti per approcci inferenziali come ABC. Nel **Capitolo 5**, presento i risultati dell'analisi di potenza effettuata con set di dati genomici sequenziati a diversi livelli di copertura (da 1x a 30x). Ho valutato il potere inferenziale di questa procedura nel

distinguere tra diversi modelli demografici e nell'inferire i parametri dei modelli. Con questo approccio, l'*FDSS* non viene calcolata direttamente da genotipi noti, ma piuttosto stimata utilizzando le genotype likelihoods, in modo da tenere conto dell'incertezza legata ai dati a bassa copertura nella stima del pattern dei polimorfismi, rendendo i dati simulati direttamente confrontabili con quelli osservati in esperimenti a bassa copertura.

Gli approcci inferenziali presentati in questa tesi possono essere efficacemente utilizzati per analizzare ampi dataset di genomi ad alta e bassa copertura da popolazioni reali, massimizzando le informazioni estratte dai dati, al fine di ricostruire complesse dinamiche di popolazione passate.

Table of Contents

The work presented in this thesis are published/under preparation for submission under the following titles:

Ghirotto S*, **Vizzari MT***, Tassi F, Barbujani G, Benazzo A. (2020). Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. *Mol Ecol Resour* 00:1–15. https://doi.org/10.1111/1755-0998.13263

Vizzari MT, Benazzo A, Barbujani G, Ghirotto S. (2020). A Revised Model of Anatomically Modern Human Expansions Out of Africa through a Machine Learning Approximate Bayesian Computation Approach. *Genes* 11(12):1510. https://doi.org/10.3390/genes11121510

Vizzari MT, Ghirotto S, Maisano-Delser P, Cassidy L, Manica A, Benazzo A. Robust demographic Inference from low-coverage whole-genome data through ABC. *(Manuscript in preparation)*.

Additional works in which I have took part during my PhD are:

Maisano Delser P, Krapp M, Beyer R, Jones E, Miller E, Hovhannisyan A, Parker M, **Vizzari MT**, Pearmain L, Imaz-Rosshandler I, Leonardi M, Somma GL, Hodgson J, Tysall E, Xue Z, Cassidy L, Bradley D, Eriksson A, Manica A. Climate and topography shaped human ancestral lineages. *(Manuscript in preparation)*.

Donaldson ME, Torres Vilaça S, Benazzo A, Wheeldon TJ, **Vizzari MT**, Bertorelle G, Patterson BR, Kyle CJ. Evaluation of the two versus three species model of North American wolf-like canids: genomic investigations in light of extensive patterns of contemporary hybridization. *(Manuscript in preparation)*.

1. Introduction

1.1. Population's genetic variation: from "classical markers" to whole-genome sequences

An accurate characterization of the genetic composition of a population allows to shed lights on its demographic history and on the evolutionary processes that have shaped its genetic diversity.

The first studies aimed at the analysis of the differences observed within and between species were not directly based on DNA markers, but rather on the so called "*classical markers*" such as allozymes or variation found in the human blood group system AB0. The differences in these gene products have been used as indirect evidence for the presence of variations in the DNA sequences that encode them, providing for the first time a method to empirically quantify populations' genetic variation and revolutionizing the field of population genetics and the study of evolution (Charlesworth and Charlesworth, 2017).

The introduction of these new molecular tools took place in 1966, with the publication of two papers on the genetic diversity in *Drosophila pseudoobscura* (Hubby and Lewontin, 1966) and *Homo sapiens* (Harris, 1966); these first measures were obtained analysing through gel electrophoresis several allozymes loci. The results of these seminal studies revealed substantially higher levels of genetic variation within populations than were previously predicted. Furthermore, comparing the genetic profile of different populations would allow us to investigate their past demographic dynamics. As an example, in Menozzi *et al.* (1978), one of the firsts groundbreaking studies in this context, the frequencies distribution of the alleles of some of these classic markers typed in different living population, unveiled the diffusion of Neolithic farmers from the Near Est into Europe.

As said before, *classical markers* detect molecular changes that modify the amino acidic compositions of proteins. However, these changes represent only a small fraction of all possible mutational changes occurring in DNA sequencies; due to the redundancy of the genetic code, in fact, most of the base changes that occur in the coding regions of the genome will be translated into the same amino acid and will not produce variations in the final gene products, so *classical markers* provide conservative estimates of variability because their diversity depends completely on non-synonymous mutations in gene sequences. Moreover, these markers are functional proteins potentially under selective pressure. This feature can be both a limitation or an advantage, depending on which evolutionary process we are interesting in. Non-neutral markers are not useful if we are interested in demographic reconstruction or in disentangling the evolutionary relationships between natural populations, because the diversity of these markers is strongly influenced by natural selection and can lead to biased demographic inferences. On the other hand, allozyme markers have been used to investigate adaptation processes driven by natural selection.

It was therefore clear that a direct study of DNA variation is necessary to accurately reconstruct the whole evolutionary processes; since most of the genome is non-coding and therefore possibly not under selective pressures, the molecular variation in these regions can be considered *"neutral"* and it is expected to reflect past demographic processes such as changes in populations size and admixture events.

The transition from *classical markers* to "*molecular markers"* became possible after the development of the polymerase chain reaction (PCR) method, that allowed to replicate specific regions of the genome starting from a small amount of genetic material (Mullis and Faloona, 1987), and the development of the first sequencing method, the automated Sanger sequencing (Sanger *et al.*, 1977).

One of the first categories of *molecular markers* used to surveying DNA sequence variation - still widely used in population genetics studies - is represented by microsatellites, also known as short tandem repeats (STR); these markers are repetitive genetic sequences in which the repeating unit contains from one to six bases. Microsatellites are characterized by a high mutation rate which makes them not particularly useful for inferring evolutionary events that occurred in very ancient times. However, the high level of polymorphisms of STRs make them suitable to investigate more recent demographic processes and for forensic analysis. Despite STRs have been the primary choice molecular tool for addressing evolutionary questions for nearly three decades, these markers also show several negative features such as size homoplasy, complex mutational patterns, and are prone to genotyping errors (Morin *et al.*, 2004).

The extensive use of another set of molecular markers, called single nucleotide polymorphisms (SNPs), greatly improved our power to make reliable inferences in population genetics studies. As their name suggests, SNPs consist of single base pair changes in DNA sequences and are the most abundant and widespread type of molecular variation in genomes (Brumfield *et al.*, 2003); compared to microsatellite loci, SNPs show a relatively low mutation rate and their evolution can be described by simple mutation models, such as the infinite site model. These characteristics makes SNPs the ideal type of molecular markers for analyzing past populations' dynamics (Brumfield *et al.*, 2003; Rivollat *et al.*, 2020) and for genome-wide scan to identify loci that may have been under selective pressure (Nielsen, 2005; Piras *et al.*, 2012; Mathieson *et al.*, 2015). SNPs have thus quickly become the most widespread molecular markers favoring a rapid growth in the amount of available SNP dataset and in the development of new reliable SNP genotyping technologies able to analyse several SNPs simultaneously, as the SNP arrays.

These SNP arrays provide information about the allelic state of positions in the genome that have prior evidence of variability (Nielsen, 2004); studies based on SNPs variation typed through arrays are relatively low cost and have been performed on massive numbers of different species. Nevertheless, SNP array returns a partial representation of the entire genome and presents limitations regarding how the loci included in these panels were discovered and typed. SNP array data were originally identified through a SNP discovery process that tends to select loci with particular allelic distribution from a small number of individuals, thus introducing an ascertainment bias which will affect parameter estimates and lead to false demographic inferences (Nielsen, 2004). Although recently developed SNP arrays include panels with genotypic positions from different populations, which should in principle reduce the bias (Patterson *et al.*, 2012), almost all available population genetic methods assume that the genetic variation under investigation have been randomly sampled among the pool of all the genomic variants that are present in a population, a condition that can be only achieved with whole-genome sequencing data (Nielsen, 2004).

The development of next-generation sequencing (NGS) technologies allowed to overcome many of the limitations of the previous methods by generating high-throughput sequence data from entire genomes, resulting in a reduction in ascertainment biases and an increase in the ability to detect evolutionary processes (van Dijk *et al.*, 2014). NGS technologies allowed parallel sequencing of millions of DNA fragments, reducing the cost of sequencing, and increasing the amount of data generated. Additionally, these technologies produce large numbers of short sequencing reads, feature that make them particularly useful for analyzing short DNA fragments, such as those normally found in ancient DNA samples (Haber *et al.*, 2016).

1.1.1 The Next Generation Sequencing (NGS) revolution

In the late '70s the first method that allowed the sequencing of DNA molecules was developed by Frederick Sanger. Sanger's sequencing technology takes advantage of the

DNA molecule synthesis process and it involves the use of use of fluorescently labelled nucleotides as irreversible DNA chain terminators (Sanger *et al.*, 1977). The final output is given by a chromatogram when a laser excites the label on the nucleotide at the end of each sequence. This technology, defined as first-generation sequencing, was the primary sequencing method used to reconstruct the first reference sequence for the human genome, released in 2004 after 14 years of work and almost 3 billion dollars spent (International Human Genome Sequencing Consortium, 2004). The publication of the human genome sequence was a monumental achievement that paved the way for the analysis of whole genome data, from both humans and other species, to investigate evolutionary dynamics at an unprecedented resolution. However, the Human Genome Project required a great deal of time and resources and highlighted the need to develop faster and cheaper methodologies to obtain genomic data. Since the release of the first human complete genome new methods that allows to overcome the limits of Sanger sequencing have been developed and are represented by the next generation sequencing (NGS) technologies (van Dijk *et al.*, 2014).

The NGS technologies, also known as second-generation sequencing, provides cheaper, faster, and reliable large scale DNA sequencing data and have become the standard tool for many applications in differ field of biology. For example, NGS data can be used to analyze the genomic variation of economically important species (e.g., Elsik *et al.*, 2009; The Potato Genome Sequencing Consortium, 2011); to perform population genetic studies that aim to understand the effect of evolutionary forces (such as mutation, natural selection, genetic drift (Fu *et al.*, 2016; Mathieson, 2020) in shaping the observed genetic variation in modern and ancient populations; to evaluate the genetic composition of different endangered organisms in order to define more effective conservation' actions (Supple and Shapiro, 2018). Different NGS methods have been developed, all sharing three main features: (i) they do not rely on the cloning of DNA fragments through a bacterial vector, (ii) the sequencing of the DNA library is done in thousands parallel reactions and (iii) the sequencing output is obtained directly without the use of the electrophoresis (van Dijk *et al.*, 2014).

Although these technologies are extremely powerful, they still have some drawbacks. Being characterized by an output of short reads (36-300bp reads length), NGS heavily relies on bioinformatics tools to obtain the complete sequence of a genome and to identify all its variable sites. The reconstruction of a genome sequence is done through the alignment of the generated reads to a reference genome sequence; this reference must belong to the same species of the sample sequenced but, if it does not exist, the choice should fall on the closest phylogenetic species whose genome sequence is already available. Despite using a reference, the reconstruction of a genomic sequence may be difficult due to the presence of highly repeated and particularly complex regions (i.e., telomeres, centromeres, regions containing short tandem repeats) extending for several base pairs; the length of the NGS reads is therefore not sufficient to accurately resolve the complexity of all the above-mentioned genomic regions (van Dijk *et al.*, 2018).

In the last few years, the introduction of the third-generation sequencing technologies allowed to overcome some of the limitations of previous sequencing technologies (van Dijk *et al.*, 2018). Unlike the second-generation sequencing, the third-generation technologies generate longer reads (5-30kb length) but at an higher cost. These long reads are typically used to resolve the complex regions of the genome and to reconstruct high-quality genomic sequence even if a reference genome is not available; in the latter case, we refer to a bioinformatics procedure called "*de novo*" assembly.

The main weak point of genomic data produced through NGS technologies is the higher error rates compared to the more reliable Sanger sequencing. These error rates vary across different sequencing platforms and arise as a consequence of base-calling and alignment errors (Nielsen *et al.*, 2011). Typically, short-reads sequencing machines present an average miscall error rate of ~1%, whereas for long-reads platforms this value reaches 10-15% (van Dijk *et al.*, 2018). These errors should be taken into account or even corrected, to avoid biased and inaccurate demographic inference and genetic analyses; a possibility to do that is to rely on high-coverage sequencing. Sequencing coverage, or depth, represent the number of times every base pair of the underlying unknown genome is read during the sequencing process (Sims *et al.*, 2014) and it is usually referred as an average throughout the whole genome; usually the higher is the coverage the more accurate the characterization of polymorphisms and individuals' genotypes will be. Having multiple evidence of which nucleotides align to each genomic position would facilitate the identification of sequencing errors and will make the variant calling output more accurate. Furthermore, since reads are not evenly distributed over the genome, some genomic regions will be covered by fewer reads than the average depth and this is something that can affect the analyses or that could potentially introduce some degree of bias. Also, when it comes to detect low frequency variants, high coverage sequencing (>20x) is much more informative and reduce the rate of false-positive SNP detection (Xu *et al.*, 2017). It has been shown indeed, that at very low-coverage (2x) level the number of false-positive called singletons is quite high (Han *et al.*, 2014). However, the huge amount of information derived through high coverage sequencing comes with a significant economical effort: given a limited financial budget the compromise is whether sequencing few high coverage samples (>20x) or sequencing more individuals at low (<5x) to medium (~10x) coverage. In any case, it is important to keep in mind that sequencing few samples at higher depths certainly increases the confidence in the called genotypes, but also restricts the analysis to a small sample of individuals which may not be representative of the genetic variations of the entire population. Likewise, choosing to sequence a larger sample of individuals at lower depths may be a viable strategy to improve the accuracy of population genetic analysis, especially those based on alleles frequencies; in fact, the uncertainty of the called genotypes can be compensated by the large number of individuals typed in the population, as shown in Fumagalli *et al.* (2013). So ultimately there is always a trade-off between the sample size and the sequencing depth.

In some circumstances, however, this choice cannot be made, especially when dealing with ancient DNA (aDNA) because of the lower availability of suitable samples. After the death of an organism, indeed, DNA molecules begins to degrade accumulating chemical damages that lead to highly fragmented aDNA samples; typically, average fragments' length can vary between 60 and 150bp (Prüfer *et al.*, 2010). Moreover, postmortem damages often cause misincorporations in the nucleotide's composition of the sample, leading to additional biases that can affect downstream analyses. One of the main challenges of aDNA studies concerns the presence of contamination in the ancient samples. Unlike modern samples, the genetic material extracted from an archaeological sample or a museum specimen tends to have a non-endogenous origin, mostly represented by microbial DNA, together with the DNA of those who handled the sample (Prüfer *et al.*, 2010). For these reasons, endogenous DNA recovery from ancient samples can be difficult, causing further reduction in sequencing depth and sample size than those achieved with modern data.

NGS technologies have revolutionized the field of population genetics, allowing the sequencing of hundred thousand complete genome sequencing from both living populations and ancient samples that are tens of thousands of years old, making it possible to explicitly study evolutionary processes over time and space. The main challenge now is how to deal with such a huge amount of information in order to make reliable inference about these past population' dynamics, exploiting the information contained in high- and low-coverage complete genomes.

1.2. Inferring past demographic dynamics from NGS data

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; Meyer *et al.*, 2006; Lohmueller, 2014). 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; Meyer *et al.*, 2012; Prüfer *et al.*, 2014; De Manuel *et al.*, 2016; Mallick *et al.*, 2016; Moreno-Mayar *et al.*, 2018). 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 and Cardin, 2005), became popular among population geneticists due to their ability to infer population size changes through time (PSMC; Li and Durbin, 2011) and divergence times (MSMC; Schiffels and 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 *et al.*, 2016) and computer simulations (Chikhi *et al.*, 2018).

Other methods infer demographic parameters via the diffusion approximation (Gutenkunst *et al.*, 2010), or coalescent simulations (Excoffier *et al.*, 2013; Beeravolu *et al.*, 2018), 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 *et al.*, 2017; Kamm *et al.*, 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 model's 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 *et al.*, 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 *et al.*, 2010; Csilléry *et al.*, 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 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 (Beaumont *et al.*, 2002; Marjoram *et al.*, 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. 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.

Recently, a new ABC framework has been developed based on a machine-learning tool called Random Forest (ABC-RF, Pudlo *et al.*, 2016; Raynal *et al.*, 2019). 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 *et al.*, 2015), should be a suitable (and possibly sufficient) statistic to summarize genomic data (Terhorst and Song, 2015; Lapierre *et al.*, 2017; Smith *et al.*, 2017). 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 *et al.*, 2007; Keightley and Jackson, 2018). In such cases, the folded version of the *SFS* should be used, with unavoidable loss of information (Keightley and 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).

The above-mentioned methods assume that the complete genome sequences analyzed are characterized at a sufficient sequencing depth and quality to obtain a faithful representation of the genetic variation that is present in the individuals under investigation. With NGS experiments, indeed, there is a level of uncertainty associated to the genotype calling that increases with decreasing coverage levels (Nielsen *et al.*, 2012). When the coverage is low it is for instance more probable to not recognize heterozygous sites or NGS errors, thus introducing a bias in the reconstructed sequence (Nielsen *et al.*, 2012).

As said before, methods based on SMC models (Li and Durbin, 2011; Schiffels and Durbin, 2014) use the distributions of the heterozygous sites across the genome to infer the coalescent times' distribution, and changes in the effective population size (N_e) over time. The uncertainty linked to the identification of polymorphisms and to the called genotypes, typical of low-coverage data, can alter these estimates and consequently, can lead to false demographic inferences. It has been estimated indeed that the minimum coverage level required to perform these kinds of analysis should be 18x (Nadachowska-Brzyska *et al.*, 2016). Methods relying on the *SFS*, being based on population's allele frequencies estimation and not on individual's genotypes, present, in principle, more flexible data-quality requirements, because the low quality of the individual data should be compensated by the high number of samples analyzed (Beichman *et al.*, 2018). In some circumstances it is however impossible to obtain enough sampled individuals to reliable estimates allele frequencies (i.e. for ancient populations or for elusive species), and the resulting *SFS* can be seriously affected by a bias that should be taken into account (Han *et al.*, 2014). As for model-based methods such as ABC, since the simulated genetic variation produced can be considered as highly accurate as that retrieved from high-quality genomes, the entire ABC procedure is only effective when compared with high-quality observed data (i.e., genomes sequenced at high-coverage levels).

Given the continuous production of whole-genome data, many of which at low coverage, it has become necessary to develop methods able to deal with the uncertainty resulting from genotype calling. These methods rely on a probabilistic approach in which polymorphisms and genotypes are not directly characterized from the data, but rather estimated from the so-called *genotype likelihoods* (GLs). These methods account for the uncertainty linked to low-coverage data and integrate the error rate of the sequencing machines (the quality scores) to generate GLs (Nielsen *et al.*, 2011). The genotype likelihoods are directly computed from the aligned reads and express the probability of the sequencing data given a certain genotype, at a particular site, for a particular individual. It is either possible, using a bayesian framework, to incorporate prior information to the inference (as the population allele frequencies) in order to produce genotypes posterior probabilities (Nielsen *et al.*, 2011). Under these approaches, the genotype calling phase could be avoided and the GLs could be used to estimate the *SFS* (Korneliussen *et al.*, 2014), to characterize population's structure (Meisner and Albrechtsen, 2018), to test for introgression (Soraggi *et al.*, 2018) and to reliably use ancient DNA by explicitly incorporating post-mortem DNA damage in the estimation of GLs (Kousathanas *et al.*, 2017).

Several studies showed that using GLs instead of calling genotypes produce a higher number of true positive polymorphic sites, thus reducing the bias that would have affected downstream population analyses (Han *et al.*, 2014; Korneliussen *et al.*, 2014; Kousathanas *et al.*, 2017). Despite these probabilistic approaches allows to deal with the uncertainty of low depth whole-genome sequences, they are not used so far within model-based inferential framework such as ABC.

2. Aim of the study

The developing of sequencing technologies and the rapidly declining production costs have made it possible to collect genome-scale data from numerous populations sampled from a wide range of species. This massive availability of genomic data requires the developing of new statistical frameworks capable to effectively exploiting whole genome sequencing datasets to make reliable inferences about the underling evolutionary processes of natural populations.

In this thesis I present a new approach for inferring demographic history from whole-genome data. The idea is to summarize the genomic data through the full genomic distributions of segregating sites (*FDSS*) within an Approximate Bayesian Computation framework, using a Machine Learning (Random Forest) approach.

I tested the performance of the framework through an extensive power analysis simulating data under different sampling strategies in terms of number of individuals analysed, number and size of the genetic loci considered, for sets of models of increasing complexity. I also applied the procedure to the analysis of real data, comparing complex alternative models on human dispersal out of Africa, and assessing the evolutionary history of the three species of Orangutan inhabiting Borneo and Sumatra islands (see **Chapter 4**).

Furthermore, I extended the framework to the analysis of genomic data sequenced at different coverage levels, developing a procedure where the *FDSS* is not directly calculated from known genotypes, but rather estimated using genotype likelihoods, so as to take into account the uncertainty linked to low-coverage data in the estimation of the pattern of polymorphisms. The so-generated simulated data are hence directly comparable with those observed even in low-coverage experiments. I evaluated the inferential power of this procedure in distinguishing among different demographic models and in inferring model parameters under different experimental conditions, assessing the effect of coverage (from 1x to 30x), number of individuals, number, and size of the simulated genetic loci (see **Chapter 5**).

3. Method

The pattern of genetic variation observed in the genome reflects the unique and complex evolutionary processes of a species that cannot be analytically predicted. Powerful tools to make inference about past demographic events that have shaped the species genetic variation rely on simulation-based methods. Simulations are widely used in population genetics. Through simulations it is possible to explicitly model the genetic variation expected under specific demographic histories using *stochastic models* (i.e. mathematical description of random evolution trough time, such as the *coalescent* (Kingman, 1982). The general principle is to generate *in-silico* datasets of genetic variation according to specific evolutionary scenarios and to compare them with the real observed genetic variation, in order to infer historical processes or to evaluate the effects of different evolutionary forces, such as natural selection and genetic drift, and to understand the interactions between them. Simulations can also be used to validate the properties and the inferential power of newly developed statistical methods (Hoban *et al.*, 2012).

The population genetics simulation algorithms can be classified in two main categories: *forward-in-time (*or *individual-based simulations)* and *backward-in-time (*or *coalescent simulations)*:

- *Forward-in-time* simulations are based on the life history modelling of each single individual in the population under investigation. This approach allows to keep track of the evolution in the genetic composition of a population starting from the present generation (t=0) and moving towards the subsequent generations (t+1, t+2, t+3 etc..); in this way the population properties can be observed at any generation.
- *Backward-in-time* simulations starts from the present generation and works backward along the lineages of a sample of a population. This approach first reconstructs the genealogy of the samples up to a single ancestor, namely *most recent common ancestor* (MRCA), and then works forward up to the current generation, introducing mutations into the generated genealogy accordingly to a specific mutation model.

Although in principle the forward-in-time approach is flexible in being able to simulate any evolutionary and demographic scenario, computational time and memory usage are still a crucial issue, especially when the number of generations simulated, or the size of the population generated, is large.

On the other hand, coalescent backward simulations represent an excellent framework for population genetics aims, because this approach is computationally efficient since it only traces the history of the observed sample backward in time and could address complex, and hence realistic, demographic scenarios at a large sequence level.

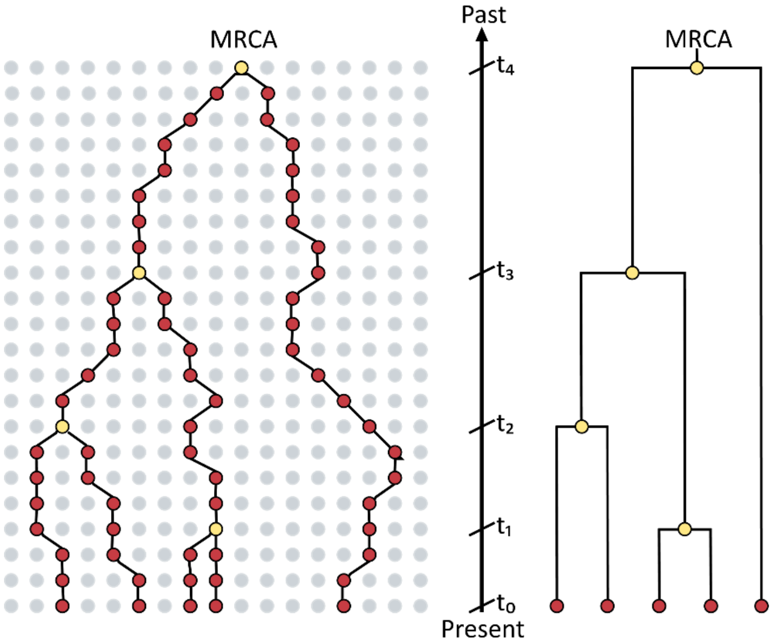
The *coalescent*, described formally in mathematical terms by John Kingman in the 1982 (Kingman, 1982), provides a description of the genealogical relationships among a sample of DNA sequences (or *loci*) drawn from a population. The coalescent models the evolutionary processes proceeding backward in time in order to identify all the points in which a pair of loci finds a common ancestor (i.e., when two lineages *coalesce*). This process continues until the common ancestor of the entire sample, namely *most recent common ancestor (MRCA),* is reached.

The genealogies reconstructed contain information about past demographic events and about the processes that have shaped the diversity of a population. In fact, the probability that two lineages share a common ancestor in the previous generation depends on the size of the population under consideration. A sample coming from a small population will share a common ancestor only few generations in the past and the coalescent rate will be higher than a sample coming from a bigger population, where the common ancestor will be located many generations back in past.

Figure 3.1 illustrate the idea that underlies the coalescent using a sample evolving accordingly to a Wright-Fisher population model (Fisher, 1930; Wright, 1931).

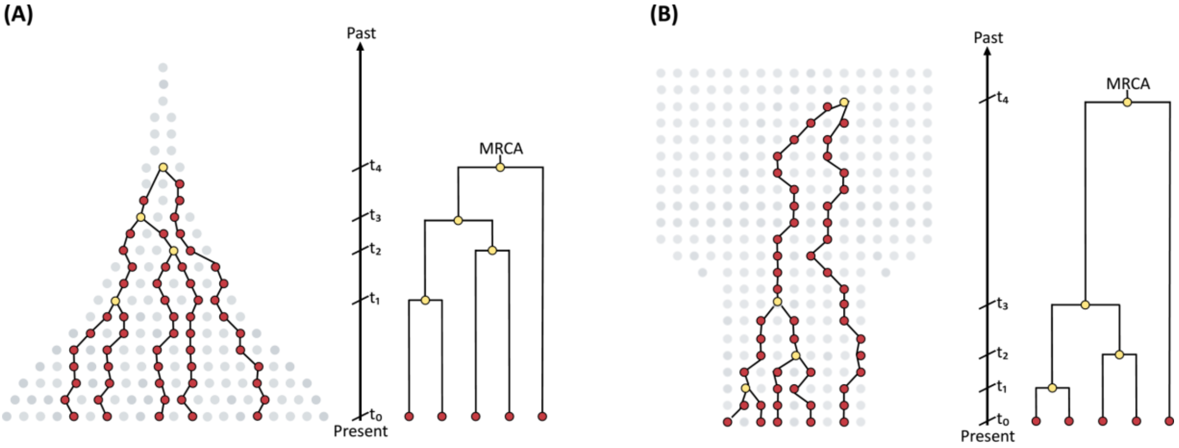
In this model, we consider a panmictic constant population composed by *N* individuals. Each row represents a single generation where only the offspring of the preceding generation survives. There are not selecting forces acting on the population, and all individuals have an equal chance to reproduce. We can sample *n* individuals from present times and tracing their ancestry using the coalescent theory. As we move to the past generations the lines of ancestry decrease from *n* to *n-1*, then from *n-1* to *n-2* and so on until a single line remains. The last coalescent event is called the time of the most recent common ancestor (TMRCA) and the last lineage represents the MRCA of the sample.

Figure 3.1. An example of genealogy of a sample of *N* individuals.



The expected distribution of coalescent times is greatly affected by changes in population size. A population that has experienced a demographic decline will be characterized by a high frequency of coalescence events in recent times, due to the limited size of the modern population (Figure 3.2 B). Similarly, in a growing population, the population size at present is greater than the population size in the past. In this case, the first coalescence events occur slowly but, as we move backwards the population size decrease and so the coalescence rate increase (Figure 3.2 A).

Figure 3.2. Genealogy of a group of *N* individuals sampled from an expanding population (A) and from a declining population (B).



In the years following its introduction, the coalescent was extended to include population dynamics and genetic processes (such as natural selection, recombination, and migration) that have increased its capability to generate the genealogical history of samples according to more realistic evolutionary dynamics. Once the genealogy has been reconstructed it is possible to simulate the genetic variability of the sample using any mutational model by inserting the mutations proportionally to the length of the branches. In this way one can observe the level of genetic variability that can be generated by a specific demographic model.

3.1. Approximate Bayesian Computation

The Approximate Bayesian Computation (ABC) approach is a powerful simulation-based framework developed to compare alternative models of evolution and to infer their parameters. Its flexibility is due to the likelihood-free inference allowing to analyse complex and realistic demographic models for which the likelihood function cannot be analytically derived (Bertorelle *et al.*, 2010; Csilléry *et al.*, 2010).

The ABC algorithm was formally defined and introduced for the first time by Beaumont in 2002 (Beaumont *et al.*, 2002). The general method includes the following steps: once the demography of the populations under investigation are defined using a model of evolution with specific demographic parameters, a large number of coalescent simulations are produced accordingly. The values of the parameters defining the model (such as population sizes and demographic event's times) are extracted from broad prior distributions, i.e. the probability distributions of parameter's values before any data are examined. Both observed and simulated datasets are summarized using the same set of summary statistics. Finally, observed data sets are then compared to the simulated ones, in order to identify the most supported model among those tested (*Model Selection*) and to estimate its parameters (*Parameters Estimation*).

Since the whole ABC method is based on the comparison between simulated and observed data, the correct choice of the statistics used to summarize them is one of the most important steps of the entire procedure (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003). The selected vector of statistics has to be able to capture the relevant information contained in the data about the processes under investigation, in other words the chosen summary statistics should be *sufficient*. Unfortunately, the sufficiency of the statistics is difficult to define, and it is strictly dependent on the model, the parameters, and the data analysed. Intuitively, including a limited number of summary statistics leads to a very rough representation of the information contained in the data, producing biases in the ABC estimates. Conversely, calculating a large number of summary statistics increase the amount of information considered on the data, but at the same time introduce stochastic noise that increases the errors in the posterior estimates. This problem, known as *course of dimensionality* (Blum and François, 2010), is not yet solved despite several serious attempts has been done (Blum *et al.*, 2013).

3.1.1. Model Selection

Trough ABC it is possible to compare alternative hypotheses about a process and assign a probability to each of them, thus allowing to identify the model, among those compared, that best explains the observed variation. These posterior probabilities, in the original formulation of ABC, can be calculated following two different approaches.

The first procedure relies on a simple acceptance-rejection algorithm (AR) (Beaumont *et al.*, 2002): for each simulated dataset, a Euclidean distance δ between the observed and simulated summary statistic is calculated and an arbitrary distance threshold is chosen; the value of the threshold is defined such that only a small fraction of the simulations, corresponding to the simulations shows the shortest δ between observed and simulated data, are retained. The posterior probabilities are then computed as the proportion of the accepted simulations for each model tested.

The second approach, proposed by Beaumont in 2008 (Beaumont, 2008), relies on the use of a weighted multinomial logistic regression procedure (LR) to compute the posterior probability: in this case the summary statistics are the predictive variable, and the model parameters are the response variable; the dependent categorical variables are represented by the models. The posterior probability of the models is evaluated in the point corresponding to the observed summary statistics.

3.1.2. Parameter's Estimation

In the original formulation of ABC, only a subset of simulations is retained to perform parameter estimations (in general the 1% closest to the observed data). At this point the estimation step can be performed in two different ways.

The first approach, also known as the "direct approach", consist in using the parameters values of the retained simulations as a sample of their posterior distribution. However, the posterior distributions of the retained simulations strictly depend on the threshold value used to filter the simulated datasets; if the threshold is too permissive and a large number of simulations are retained, the posterior distribution obtained could completely overlap the prior. The direct approach works well when the threshold is very stringent, but in this case a huge number of simulations are needed to obtain a reasonable number of retained datasets for the estimates.

The second method, introduced by Beaumont *et al*. (2008), is based on the computation of a local weighted linear regression between each parameter and the summary statistics of the retained simulations. In this case, a weight is assigned to each simulated dataset, that increases as the distance between the observed and simulated data sets decreases. The regression slope is then used to adjust each parameter value from the retained simulations towards the value expected in correspondence to the observed summary statistics vector. The distribution of the values obtained represents the posterior distribution of the demographic parameter. Usually, the mode, the mean and the median value of the obtained posterior distributions are used as points estimates of the parameters.

3.2. Supervised Machine Learning and ABC: Random Forest

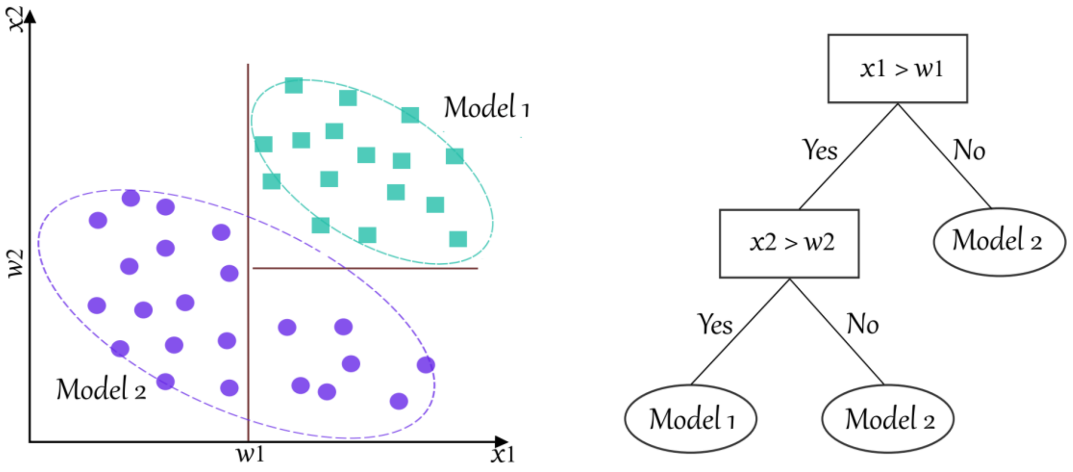
In the original formulation of ABC, the most used algorithm for model selection was based on the weighted multinomial logistic regression, introduced by Beaumont (2008). However, this algorithm suffers from two important limitations. First, to obtain reliable estimates of the models' posterior distribution, many simulations are necessary, making it difficult to analyze massive datasets with thousands of genomic loci, or to generate data from complex demographic histories. 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 and 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 and parameters' estimation procedures, based on a Machine Learning algorithm called *Random Forest* (RF, Pudlo *et al.*, 2016).

Machine Learning (ML) is one of the modern approaches being adapted for population genetic inferences. Generally, ML algorithms are divided in two categories: *unsupervised learning* and *supervised learning*. The main difference between them is that unsupervised algorithms can automatically identify structures within a dataset without prior knowledge of how the data are organized and are commonly used for exploratory analysis like principal component analysis (PCA) (Schrider and Kern, 2018). Supervised algorithms, on the other hand, is based on the exploitation of prior knowledge about a known dataset to make prediction about a new set of data and are usually used for classification and regression problems.

Random Forest (RF) is a popular ML algorithm that belongs to the supervised learning technique. RF is based on the concept of *ensemble learning*, which is a process of combining multiple classifiers to solve a complex problem (Breiman, 2001). The classifiers at the core of the RF procedure are represented by individual *decision trees*. In this type of tree-structured classifiers, the internal nodes represent the features of a dataset, the branches represent a decision rules, and each terminal node represents an outcome.

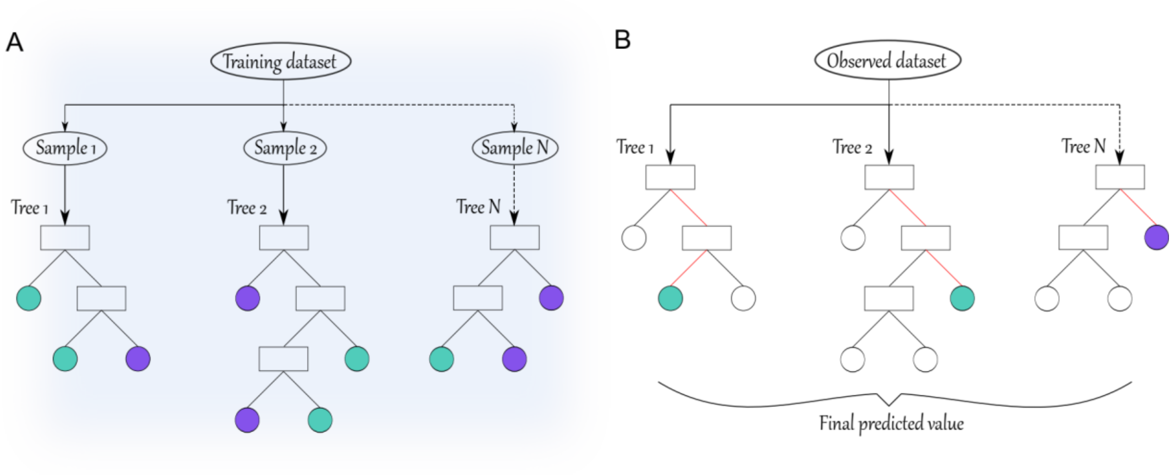
In a decision tree, there are two kind of nodes: the *decision node* and the *leaf node*. Decision nodes are used to make any decision and have multiple branches, whereas leaf nodes are the output of those decisions and do not contain any further branches. The decisions are performed based on the features of the given dataset. A decision tree simply asks a question and based on the answer (Yes/No), it further splits the tree into subtrees until an outcome is reached. Figure 3.3 explains the general structure of a decision tree.

Figure 3.3. Structure of a decision tree. Left panel, plot of datasets produced under two different model, as a function of two features (*x1* and *x2*). Right panel, the resulting decision tree that describe the structure of the data in function of the features' values (*w1* and *w2*).



As said before, the aim of a supervised learning algorithm such as RF is to learn a function (*f*) that, given a training set of labelled data, best approximates the relationship between a response variable (*y*) and a vector of features (*x*), such that *f(x)=y*. If the response variable is categorical, we are constructing a classification forest, whereas if y is a continuous variable the forest is defined as regression forest (Schrider and Kern, 2018).

Figure 3.4 General structure of a random forest classifier. A) Training forest. The RF algorithm learns the relationships between the response variable (y) and a vector of characteristics (x) by constructing *N* classification/regression trees by subsampling the reference table. B) Prediction on a target dataset. Once the classifier is constructed, a prediction on the observed data is evaluated with the obtained random forest. The final outcome is assigned following a majority vote rule (classification forest) or by averaging the predictions produced by the trees in the forest (regression forest).



In an ABC context, the training set consist of the simulated dataset of genomic variation produced under the specified evolutionary models tested (called Reference Table). The response variable y can be represented by the model indices (classification forest) or by the parameters' values (regression forest); the vector of features x, is represented by the vector of summary statistics chosen to summarize the data.

Under the RF approach, the model selection stage is so rephrased as a classification problem. The Machine Learning classifier is constructed from the reference table, composed by a set of simulation records made of models' indices and summary statistics for the associated simulated data. The reference table serves as training database for RF that forecasts model index based on the summary statistics. 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 (Pudlo *et al.*, 2016).

Similarly, the parameters estimation stage is treated as a regression problem. In this case the reference table is composed by the simulated summary statistics and the parameters' values sampled from the prior distribution and used to simulate the data. The outcome of the regression forest is an estimated value for each demographic parameter (Raynal *et al.*, 2019).

Random Forest has shown to be insensitive both to the correlations among summary statistics and to the presence of uninformative variables, and it accommodates large dimensional summary statistics with no consequences on the estimation performances. Moreover, the number of simulations necessary to obtain reliable estimates passed from a few millions (needed for the classic ABC approach) to few thousands (Pudlo *et al.*, 2016).

All these features make it now possible to apply ABC-RF to the study of complex evolutionary models through the analysis of complete genomes without incurring in computational constrains and in *curse of dimensionality* issues.

3.3. Assessing the quality of the ABC procedure

One of the most interesting features of ABC is its high flexibility in assessing the quality of the estimates inferred from real data. This is mainly achieved through the analysis of pseudo-observed data (*pods*), i.e. simulated datasets generated under known conditions.

To assess the reliability of the model selection procedures, the proportion of True Positives (TP) can be evaluated. To do this, a set of pseudo-observed data is generated using each of the models considered in the model selection analysis; these *pods* are then treated as observed datasets. The TP rate can be calculated as the proportion of cases in which the model selection procedures is able to recover the right model. High values of TP mean that the genetic data used in the analysis allow one to distinguish between the demographic models tested. I applied this procedure to evaluate the power of the inferential framework proposed in this thesis, calculating the proportion of True Positives using 1,000 *pods* generated from each of the models under investigation (results detailed in **Chapter 4**).

Similarly, to assess the quality of the parameters estimate one can calculate different indices like the coefficient of determination (R^2), the bias and the root mean square error (RMSE). For this purpose, to assess the quality of the parameter estimation performed in the study presented in this thesis, I exploited 1,000 *pods* generated from the models tested (results detailed in **Chapter 5**) and calculated the following indices:

- The coefficient of determination (R^2). R^2 is the fraction of variance of the parameters explained by the summary statistics used to build the regression model. In the absence of an established threshold value, there is a general agreement that when $R^2 < 0.10$, the summary statistics do not convey enough information about the parameter estimate (Neuenschwander *et al.*, 2008).

- The relative bias. To calculate the relative bias, I estimated the parameters for each pod with the same approach used for the observed data. The bias depends on the sum of differences between the 1,000 estimates of each parameter thus obtained and the known (true) value, and it is calculated as:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\theta_i - \theta}{\theta}$$

where θ_i is the estimator of the parameter θ (true value), and *n* is the number of pods used (1,000 in our case). Because bias is relative, a value of 1 corresponds to a bias equal to 100% of the true value.

- The root mean square error (RMSE). To calculate the RMSE I re-estimated parameters using pods. The RMSE depends on the sum of squared differences between the 1,000 estimates of each parameter thus obtained and the true value and it is calculated as:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \theta)^2}$$

- The factor *2*, representing the proportion of the 1,000 estimated median values lying between 50% and 200% of the true value.
- The 50% coverage, defined as the proportion of times that the known value lies within the 50% credible interval of the 1,000 estimates.

3.4. Summarize whole-genome variation: the *FDSS.*

The introduction of the Random Forest algorithm in the ABC inferential framework allowed to overcome the issues related to the dimensionality of summary statistics chosen to summarize the genomic data. Unfortunately, the *sufficiency* of the summary statistics is still an assumption for a proper inference. No general rules are available about which and how many statistics should be used, except for simple evolutionary models such as a constant population size through time. The choice of summary statistics is often completely arbitrary, and it needs to be carefully evaluated case by case, performing some preliminary analysis evaluating the ability of the chosen summary statistics to recover the relevant aspects of the genomic data.

In this thesis I 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 for any pair of populations, 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 and 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 *et al.*, 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, I analyze the full genomic distribution of each statistic.

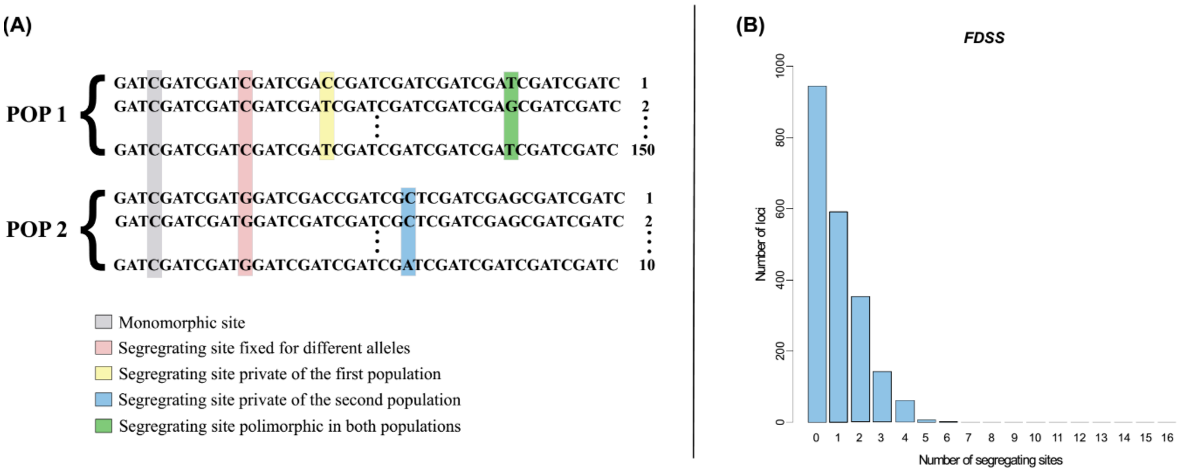
To compute the *FDSS,* I evaluated the genomic distributions of the four mutually exclusive categories of segregating sites in two populations, namely (i) segregating sites private of the first population; (ii) segregating sites private of the second populations; (iii) segregating sites that are polymorphic in both populations; and (iv) segregating sites fixed for different alleles in the two populations. I considered the genome as subdivided in k independent fragments of length m, and for each fragment I 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}, $\mathrm{Lj}_{\mathrm{ii}}$, $\mathrm{Lj}_{\mathrm{iii}}$, $\mathrm{Lj}_{\mathrm{iv}}$} 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:

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

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

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

Figure 3.5. Summary statistics used for the ABC-RF analysis. (A) The four categories of segregating sites. (B) An example of frequency distribution of segregating sites computed for a single category of sites. This plot shows the number of loci along the genome (y-axis) carrying a certain number of segregating sites (x-axis).


(A)
POP 1
GATCGATCGATCGATCGACCGATCGATCGATCGATCGATCGATC 1
GATCGATCGATCGATCGATCGATCGATCGATCGAGCGATCGATC 2
GATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC 150
POP 2
GATCGATCGATGGATCGACCGATCGCTCGATCGAGCGATCGATC 1
GATCGATCGATGGATCGATCGATCGCTCGATCGAGCGATCGATC 2
GATCGATCGATGGATCGATCGATCGATCGATCGATCGATCGATC 10
Monomorphic site
Segregrating site fixed for different alleles
Segregrating site private of the first population
Segregrating site private of the second population
Segregrating site polimorphic in both populations
(B)
FDSS
Number of loci
0
200
400
600
800
1000
Number of segregating sites
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

4. Inference using High-Coverage data

In this work I developed and tested a new ABC pipeline combining the Random Forest procedure of model selection with the *FDSS* statistic to summarize the data. I analyze the full genomic distribution of the *FDSS* and compare its performance under a wide range of experimental conditions with the one achievable using a statistic that is commonly used to summarize the genetic variation: the Site Frequency Spectrum (*SFS* hereafter, calculated across all sites, including monomorphic loci). I calculated both the *folded SFS*, in which the ancestral/derived state of the alleles is unknown, and the *unfolded SFS*, which assumes knowledge of whether the alleles are ancestral or derived.

I 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*.

As a final step, I applied the new ABC procedure to two case studies, in all cases choosing to sample a single individual (i.e., two chromosomes) per population. First, I 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, I 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).

4.1. Power Analysis

To determine the power of both the *FDSS* and the *SFS* in distinguishing among alternative evolutionary trajectories, I performed a power analysis simulating and testing genetic data according to different experimental conditions. I 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 sampled for each population {2, 4, 10, 20}, for a total of 60 combinations of sampling conditions tested. For each combination, I generated data with intra-locus recombination (recombination rate=1×10^{-8}), and with a fixed mutation rate (1×10^{-8} /bp/generation). I 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 and Hermisson, 2010) output of each simulation through a in-house python script (available on github https://github.com/anbena/ABC-FDSS). When analyzing demographic models assuming more than one population, I calculated the paiwise *SFS*. For each combination of experimental conditions, I 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.*, 2016). I computed the confusion matrices and I evaluated the out-of-bag classification error (CE); for each comparison I 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).

4.1.1. One-population models

I started by considering four demographic models depicting the evolutionary dynamics of a single population (Figure 4.1). The first model (*Constant*) represents a constantly evolving population with a certain effective population size. Under the second model *(Bottleneck*), the population experienced an instantaneous 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. The third model (*Exponential Growth*) 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*. Under the last model (*Structure*), 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 uniform prior distributions (Table 4.1).

Table 4.1. Demographic parameters and prior distributions of One-Population models. Mutation and Recombination rates are expressed per nucleotide per generation.

Demographic Parameters	**Prior Distributions**
Effective population size *(N1)*	Uniform {500:50,000}
Intensity bottleneck *(i)*	Uniform {10:100}
Intensity exponential growth *(i)*	Uniform {10:100}
Time bottleneck *(T)*	Uniform {100:20,000}
Time exponential growth *(T)*	Uniform {100:20,000}
Number of demes *(d)*	Uniform {2:10}
Migration rate *(m)*	Exponential {0.1}
Mutation rate	1×10^{-8} {Fixed}
Recombination rate	1×10^{-8} {Fixed}

4.1.2. Two-populations models

I then moved to considering three demographic models with two populations (Figure 4.2). The first one (*Divergence*) 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. The second model (*Divergence with Migration*) 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 exponential priors with mean 0.1. The third and last model (*Divergence with Admixture*) assumes a single pulse of bidirectional admixture at time *Tadm* after divergence. Admixture rates *adm12 adm21*, and the time of admixture are drawn from uniform priors (Table 4.2).

Table 4.2. Demographic parameters and prior distributions of Two-Populations models. Mutation and Recombination rates are expressed per nucleotide per generation. Time is in generations. In the simulation step I considered a *Tadm* value only if (*Tsep*-*Tadm*)/*Tsep* was between 0.2 and 0.8.

Demographic Parameters	**Prior Distributions**
Effective population size *(Nanc, N1, N2)*	Uniform {500:50,000}
Time split *(Tsep)*	Uniform {300:20,000}
Migration rate *(m12, m21)*	Exponential {0.1}
Time admixture *(Tadm)*	Uniform {50:2,500}
Admixture rate *(adm12, adm21)*	Uniform {0.05:0.20}
Mutation rate	$1\text{x}10^{-8}$ {Fixed}
Recombination rate	$1\text{x}10^{-8}$ {Fixed}

4.1.3. Multi-populations models

In most realistic cases, populations do interact with each other, and it is also of deep interest to test how the ABC procedure that it is presented in this thesis would behave when more realistic dynamics are taken into account. Among the many possible scenarios, I chose to initially focus and compare 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, I followed the parametrization proposed by Malaspinas *et al.* (2016), with some minor modifications (Figure 4.3, Tables 4.3-4.4). 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 (Meyer *et al.*, 2012; Prüfer *et al.*, 2014), I allowed for genetic exchanges from archaic to modern species, indicated in Figure 4.3 by the colored arrows. The archaic populations actually sending migrants to modern humans are unknown, and hence here I 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, I 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. I 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. I 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.

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

Table 4.3. Demographic parameters and prior distributions of multi-populations models: Single Dispersal model. Migration and admixture rates are expressed per generation, times in years. I considered a generation time of 29 years as in Malaspinas *et al.* (2016). Per nucleotide per generation mutation and recombination rates are fixed as in Malaspinas *et al.* (2016).

Demographic Parameters	Prior Distributions
Effective population size *(Ne)*	Uniform {500:50,000}
Migration rate *(ModernPop)*	Uniform {10^{-6}: 10^{-3}}
Time split Africa-Ghost	Uniform{40,000:145,000}yrs
Duration time bottleneck	2,900yrs
Intensity bottleneck	Uniform {2:100}
Time split Eurasia/Papua-Ghost*(OOA)*	Uniform {35,000:EndBottlGhost}yrs
Time split Europe-Asia	Uniform {20,000:30,000}yrs
Time admixture Nea-Asia	Uniform {20,000:Time split Europe-Asia}yrs

Time admixture Nea-Eurasia	Uniform {Time split Europe-Asia:EndbottlOOA}yrs
Time admixture Den-Papua	Uniform {30,000:EndBottlOOA}yrs
Time admixture Arc-Papua	Uniform {TimeAdmix.Den-Papua: EndBottl.OOA}yrs
Time admixture Nea-Ghost	Uniform{Time OOA:EndBottl.Ghost}yrs
Admixture rate	Uniform {10^{-3}:10^{-1}}
Time split Nea-NeaR	110,000yrs {Fixed}
Time split Den-DenR	393,000yrs {Fixed}
Time split Den-Nea	495,000yrs {Fixed}
Time split Arc-Nea/Den	580,000yrs {Fixed}
Time split Ancient-Modern	638,000yrs {Fixed}
Sample time Neanderthal	85,735yrs {Fixed}
Sample time Denisova	67,570yrs {Fixed}
Mutation rate	1.25x10^{-8}{Fixed}
Recombination rate	1.12x10^{-8}{Fixed}

Table 4.4. Demographic parameters and prior distributions of multi-populations models: Multiple Dispersal model. Migration and admixture rates are expressed per generation, times in years. I considered a generation time of 29 years as in Malaspinas *et al.* (2016). Per nucleotide per generation mutation and recombination rates are fixed as in Malaspinas *et al.* (2016).

Demographic Parameters	**Prior Distributions**
Effective population size *(Ne)*	Uniform {500:50,000}
Migration rate *(ModerPop)*	Uniform {10^{-6}: 10^{-3}}
Time split Africa-Ghosts	Uniform {40,000:145,000}yrs
Duration time bottleneck	2,900yrs
Intensity bottleneck	Uniform {2:100}
Time split Papua-Ghost1*(OOA1)*	Uniform {40,000:Time split. Africa-Ghost1}yrs
Time split Eurasia-Ghost2*(OOA2)*	Uniform {35,000:EndBott.Papua}yrs
Time split Europe-Asia	Uniform {20,000:EndBott.Eurasia}yrs
Time admixture Nea-Asia	Uniform {20,000:Time split Europe-Asia}yrs
Time admixture Nea-Eurasia	Uniform {Time split Europe-Asia:EndBott.Eurasia}yrs
Time admixture Den-Papua	Uniform {30,000: EndBott.Papua}yrs
Time admixture Arc-Papua	Uniform {Time admix. Den-Papua:EndBott.Papua}yrs
Time admixture Nea-Ghost2	Uniform {Time split Euras-Ghost2:Time split Africa-Ghost2}yrs
Admixture rate	Uniform {10^{-3}:10^{-1}}
Time split Nea-NeaR	110,000yrs {Fixed}
Time split Den-DenR	393,000yrs {Fixed}
Time split Den-Nea	495,000yrs {Fixed}
Time split Arc-Nea/Den	580,000yrs {Fixed}
Time split Ancient-Modern	638,000yrs {Fixed}
Sample time Neanderthal	85,735yrs {Fixed}
Sample time Denisova	67,570yrs {Fixed}
Mutation rate	1.25x10^{-8} {Fixed}
Recombination rate	1.12x10^{-8} {Fixed}

4.2. Real Case: out of Africa dynamics

I explicitly compared SDM and MDM considering the high-coverage genomes of Denisova and Neandertal (Meyer *et al.*, 2012; Prüfer *et al.*, 2014), together with modern human samples from Pagani *et al.* (2016). All the individuals were mapped against the human reference genome hg19 build 37. To calculate the observed *FDSS* I 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). I 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%; Meyer *et al.*, 2012; Prüfer *et al.*, 2014). Power analysis (see *Results-Multi populations models* section) showed I 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.* (2016). As for the Papuan genomes in Malaspinas *et al.* (2016), I 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. I 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.

4.3. Real Case: Orangutan evolutionary history

I selected seven orangutan individuals, one from each of the populations defined by Nater *et al.* (2017), choosing the genomes with the highest coverage. I 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 and Durbin, 2010). I used picard-tools-1.98 (http://picard.sourceforge.net/) to add read groups and to filtered out duplicated reads from the BAM aligments. I 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*, I 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). I 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. I 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. I 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*. I finally calculated the observed *FDSS* from the quality filtered VCF file.

To investigate past population dynamics of the three Orangutan species, I designed competitive scenarios following the demographic models reported in Nater *et al.* (2017). I directly compared complex demographies, designing the within-species substructure as described by Nater *et al.* (2017), (Figure 4.5 A). 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 Figure 4.5 B. Under the first model (*1a*) both the North Toba (first) and Borneo (later) populations separated from *Pongo tapanuliensis*, located south of Lake Toba. The second model (*2a*), assumes a first separation of South Toba from Nord Toba, followed by the divergence of the Borneo Orangutan from South Toba. Under the third model (*1b*) both the Borneo (first) and North Toba (later) populations separated from South Toba. The fourth and last model (*2b*) describe a first separation of South Toba from Borneo Orangutan, followed by the divergence of North Toba from South Toba. I modeled bidirectional migration both among populations within a species, and between neighboring species. I 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. I 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 (Figure 4.5 B).

Table 4.7. Demographic parameters and prior distributions for Model 1a. Migration rates are expressed per generation, times in years. I used a generation time of 25 years as in Nater *et al.* (2017). The per nucleotide per generation mutation rate is fixed as in Nater *et al.* (2017).

Demographic Parameters	**Prior Distributions**
Effective population size *(Ne-ModernPop)*	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate *(Intra BO)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(Intra NT)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(ST-strucNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancBO)*	Loguniform {10^{-6}: 10^{-2}}
Time sep.. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. BO-ST	Uniform {400,000:1,500,000}yrs
Time stop migration *(ST-ancBO)*	Uniform {TimeBottlBO:Time sep. BO-ST}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep. ancNT-ST	Uniform {1,500,000:4,000,000}yrs
Mutation rate	$1.5\text{x}10^{-8}$ {Fixed}

Table 4.8. Demographic parameters and prior distributions for Model 2a. Migration rates are expressed per generation, times in years. I used a generation time of 25 years as in Nater *et al.* (2017). The per nucleotide per generation mutation rate is fixed as in Nater *et al.* (2017).

Demographic Parameters	**Prior Distributions**
Effective population size *(Ne-ModernPop)*	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate *(Intra BO)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(Intra NT)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(ST-strucNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancBO)*	Loguniform {10^{-6}: 10^{-2}}
Time sep. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. BO-ST	Uniform {1,500,000:4,000,000}yrs
Time stop migration *(ST-ancBO)*	Uniform {TimeBottlBO:Time sep. BO-ST}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep. ancNT-ST	Uniform {TimeStrucNT:Time sep. BO-ST}yrs
Mutation rate	$1.5\text{x}10^{-8}$ {Fixed}

Table 4.9. Demographic parameters and prior distributions for Model 1b. Migration rates are expressed per generation, times in years. I used a generation time of 25 years as in Nater *et al.* (2017). The per nucleotide per generation mutation rate is fixed as in Nater *et al.* (2017).

Demographic Parameters	**Prior Distributions**
Effective population size *(Ne-ModernPop)*	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate *(Intra BO)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(Intra NT)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(ST-strucNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancBO)*	Loguniform {10^{-6}: 10^{-2}}
Time sep. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. BO-ST	Uniform {400,000:1,500,000}yrs
Time stop migration *(ST-ancBO)*	Uniform {TimeBottlBO:Time sep. BO-ST}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep. ST-ancNT	Uniform {1,500,000:4,000,000}yrs
Mutation rate	$1.5\text{x}10^{-8}$ {Fixed}

Table 4.10. Demographic parameters and prior distributions for Model 2b. Migration rates are expressed per generation, times in years. I used a generation time of 25 years as in Nater *et al.* (2017). The per nucleotide per generation mutation rate is fixed as in Nater *et al.* (2017).

Demographic Parameters	**Prior Distributions**
Effective population size *(Ne-ModernPop)*	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate *(Intra BO)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(Intra NT)*	Loguniform {10^{-4}: 10^{-1}}
Migration rate *(ST-strucNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancNT)*	Loguniform {10^{-5}: 10^{-1}}
Migration rate *(ST-ancBO)*	Loguniform {10^{-6}: 10^{-2}}
Time sep. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. ST-BO	Uniform {1,500,000:4,000,000}yrs
Time stop migration *(ST-ancBO)*	Uniform {TimeBottlBO:Time sep. ST-BO}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep. ST-ancNT	Uniform {TimeStrucNT:Time sep. ST-BO}yrs
Mutation rate	$1.5\text{x}10^{-8}$ {Fixed}

4.5. Results

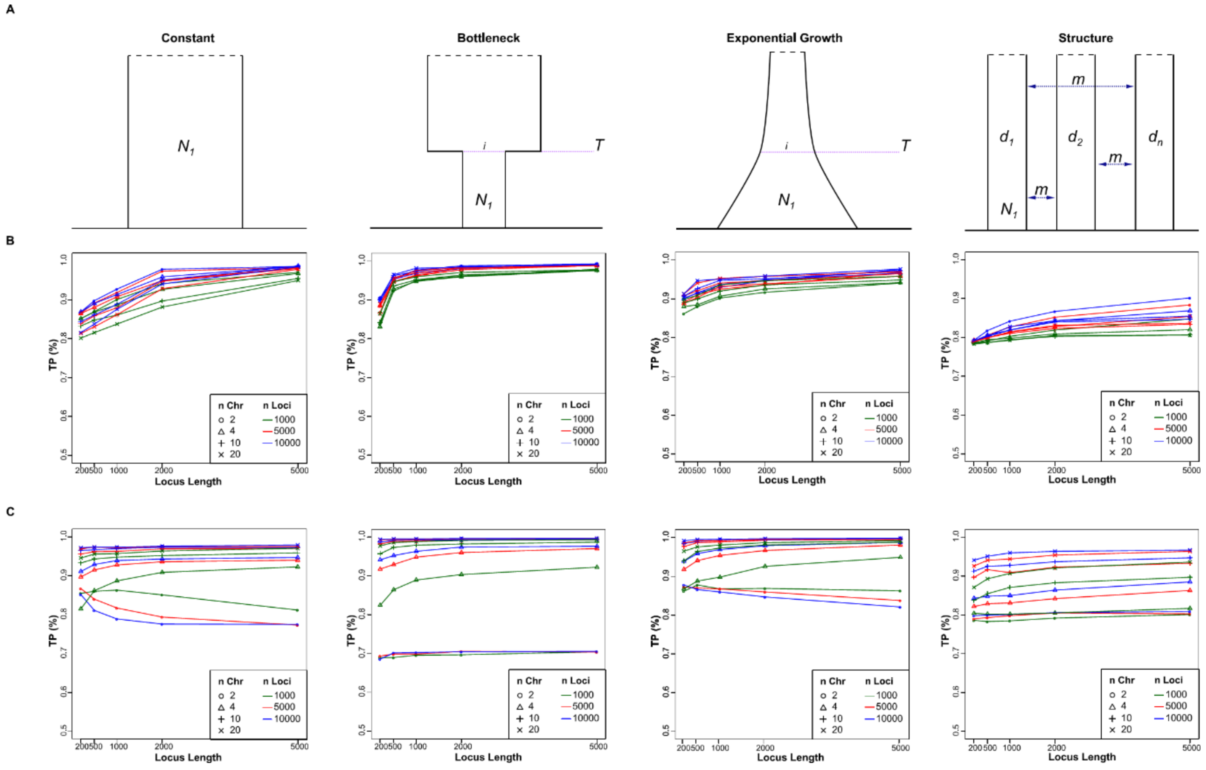
4.5.1. Power Analysis

4.5.1.1. One-population models

The four plots of Figure 4.1 B report the results of the power analyses obtained summarizing the data through the *FDSS*, whereas plots of Figure 4.1 C report the results obtained with the folded *SFS*. Being consistent, the results for the unfolded *SFS* are reported in the *Supplementary Materials* section (Supplementary Figure 4.1). In each plot, I 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, I 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 (Supplementary Figure 4.2). As expected, I 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, I observed instead a decrease in the TP rate going from 2 to 20 chromosomes (see Figure 4.1 B). I 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. 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 and 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) I 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% (Figure 4.1 C, Supplementary Figure 4.1 A).

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* I 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%).

Figure 4.1. 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.

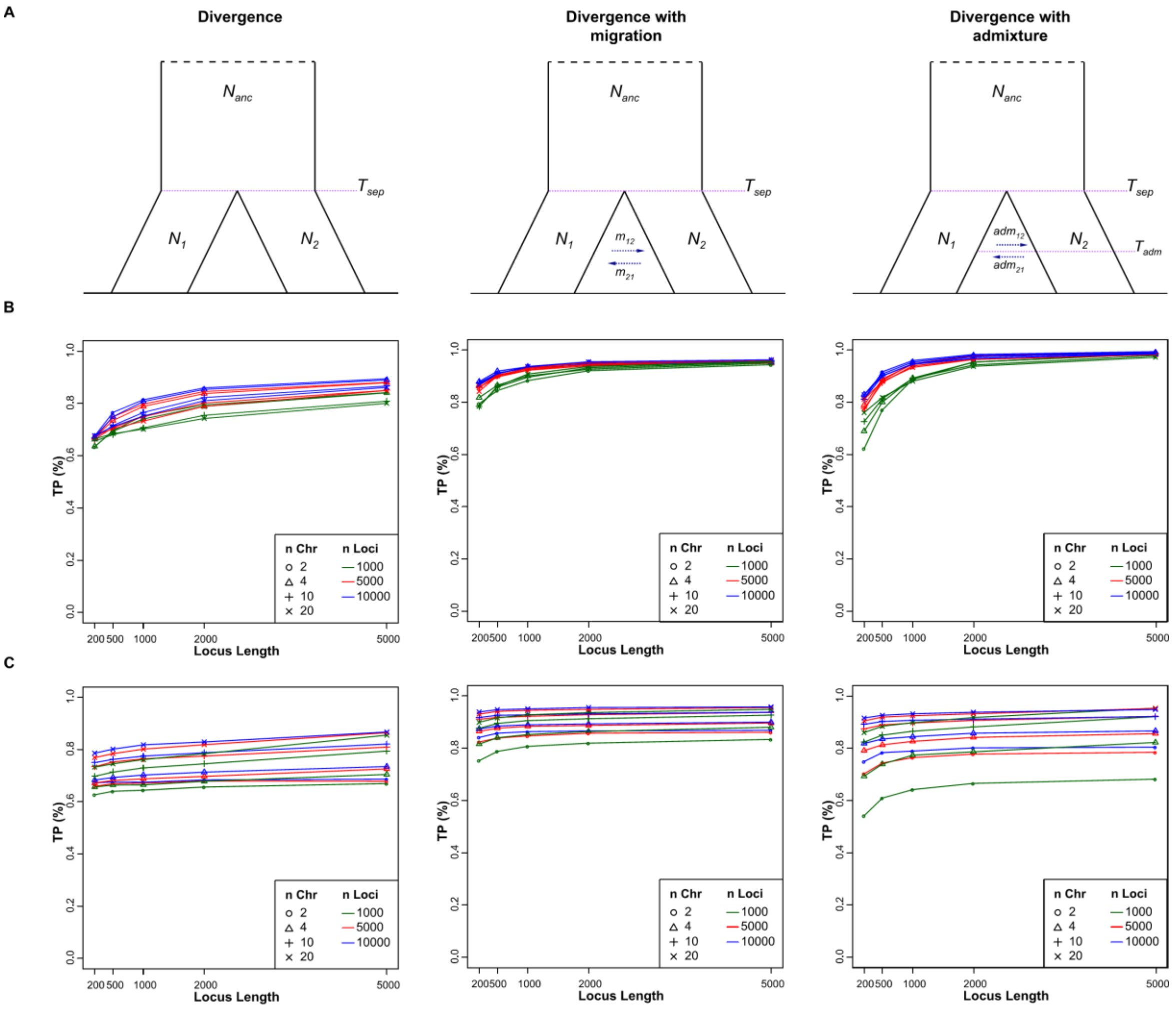


4.5.1.2. Two-populations models

The plots in Figure 4.2 B, C and Supplementary Figure 4.1 B 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 (Figure 4.2 C and Supplementary Figure 4.1 B), and I 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 I 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.

Figure 4.2. 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 Figure 4.1.

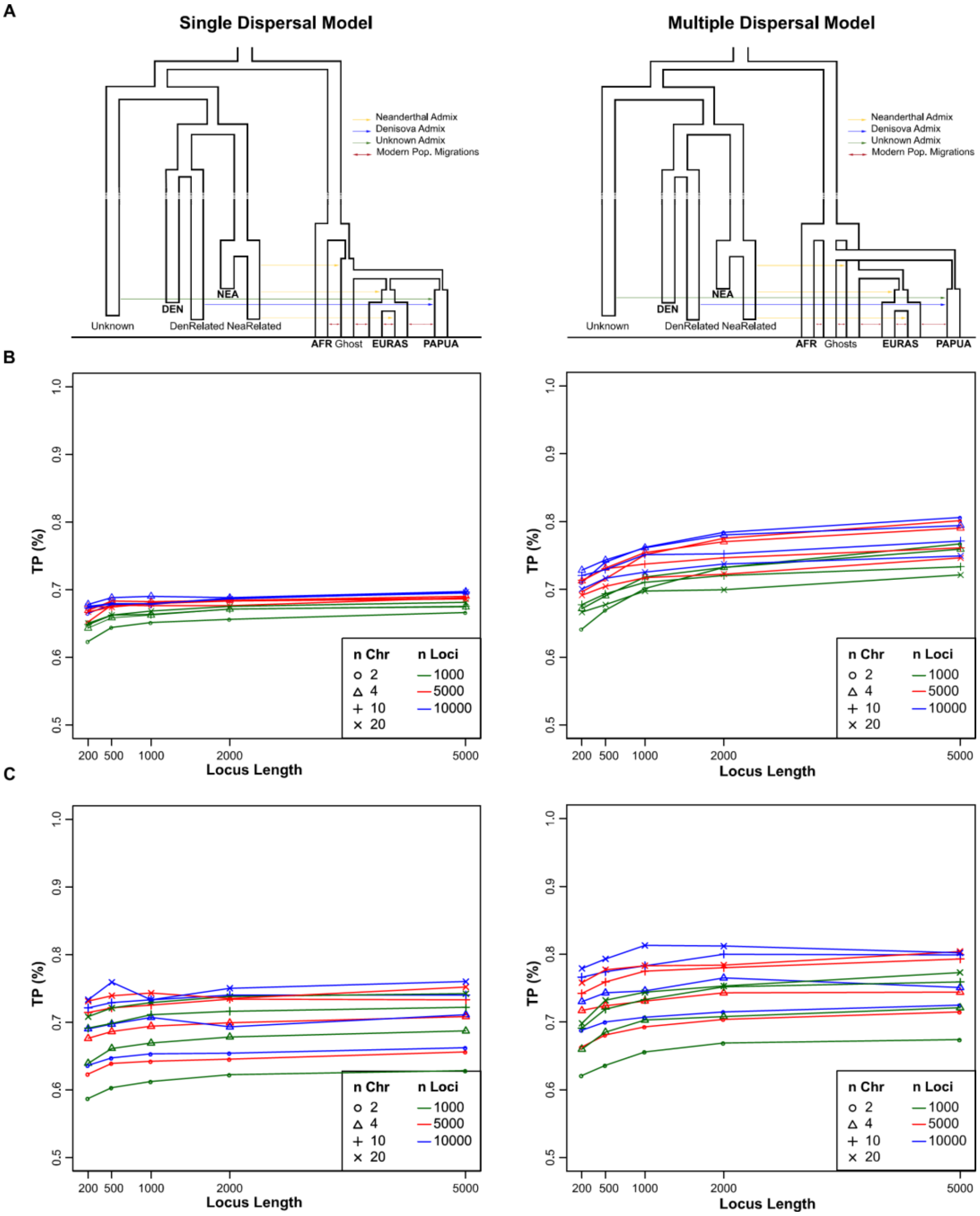


4.5.1.3. Multi-populations models

Figure 4.3 B, C and Supplementary Figure 4.1 C 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, I also evaluated the ability of the *FDSS* to be informative about the correct model as a function of this interval. To do this, I considered 10,000 pods from the MDM. I 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, reaching values of 90% for some combinations of experimental parameters (details in **Paper I**). 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.

Figure 4.3. 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 Figure 4.1.



4.5.2. 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 length of the fragments considered, or 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 model comparison through *FDSS* is particularly suited for fragmented genomes and small sample sizes, as in case of ancient DNA studies. To further explore this feature, I 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) indeed admitted that these inferences change depending on the computational method used for phasing haplotypes. Therefore, it made sense to explicitly 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 (Figure 4.3 A).

Regardless of the Papuan individual considered in each run of 31 replicated experiments, the results were always consistent in supporting 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 (Figure 4.4).

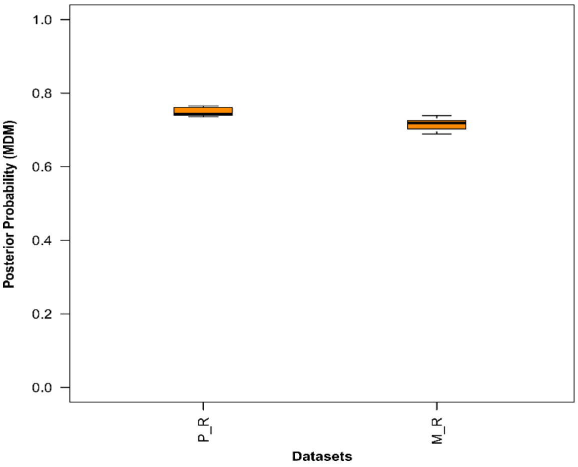
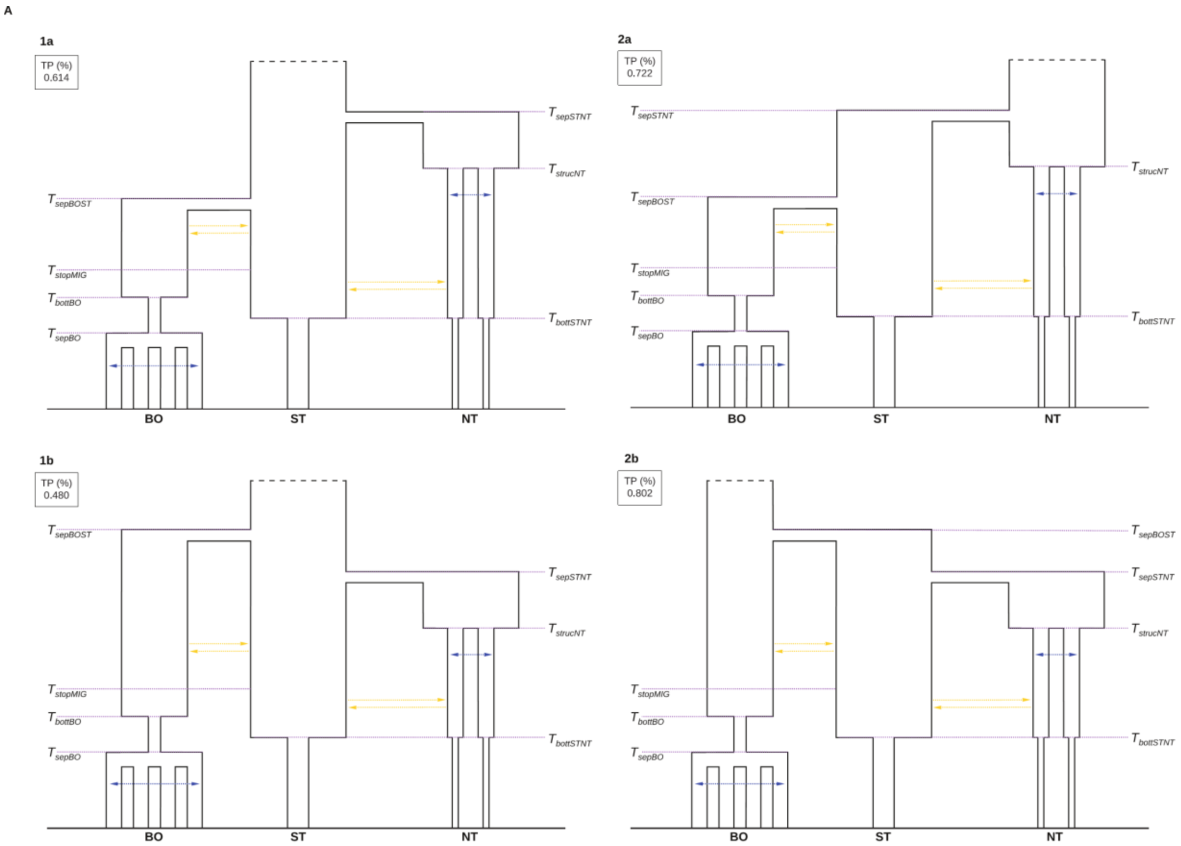


Figure 4.4. Posterior Probabilities for the MDM. Left panel: posterior probabilities obtained analyzing 6 Papuan individuals from Pagani *et al.* (2016) (PR). Right panel: posterior probabilities obtained analyzing 25 Papuan individuals from Malaspinas *et al.* (2016) (MR).

4.5.3. Real Case: Orangutan evolutionary history

As a second application, I 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. I first assessed the ability to correctly recognize the four models through a power analysis (Figure 4.5 A). 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* (Figure 4.5 B).

Figure 4.5. 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).



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

4.6. 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; Dasmahapatra *et al.*, 2012; Miller *et al.*, 2012; De Manuel *et al.*, 2016; Benazzo *et al.*, 2017). 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 and Durbin, 2011; Schiffels and Durbin, 2014). Approximate Bayesian Computation, explicitly comparing alternative demographic models and estimating the models' probabilities, represents a powerful inferential tool about past demographic events (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 *et al.*, 2016), only recently, with the development of the Random Forest procedure for ABC model selection (Pudlo *et al.*, 2016), 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. In this thesis, I took advantage of this newly proposed algorithm to test the flexibility of a new ABC-based procedure in comparing different demographic models. To ensure sufficiency in the summary of the data, I proposed the use of the *FDSS*, namely the complete genomic distribution of the four mutually exclusive categories of segregating sites for pairs of populations (Wakeley and Hey, 1997). I tested the ability and the efficiency of the whole framework in distinguishing among models of increasing complexity while generating data under a broad spectrum of experimental conditions. I also compared the power obtained summarizing the data through the *FDSS* with that reachable through the folded and unfolded version of the *SFS*.

4.6.1. Power Analysis

Initially, I 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, Figures 4.1 and 4.2) and complex (multi-populations scenarios, Figure 4.3) evolutionary models. When I 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 (Supplementary Figure 4.2 A); when it is high, the structured model tends to panmixia, as has already been known since Wright's times (Wright, 1931). I 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 (Supplementary Figure 4.2 B).

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* I 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, I observed an increase in the power of the test at increasing intervals between the African divergence and the second exit, reaching values close to 90%.

4.6.2. Comparison between SFS and FDSS

In general, the results presented in this thesis show 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 I 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 and Depaulis, 2003; Hernandez *et al.*, 2007; Morton *et al.*, 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.

4.6.3. Applications to real datasets

I 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, I 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 I obtained a good ability to tell models apart (Figure 4.3). Thirty-one replicated experiments, differing for the Papuan genome being considered, consistently supported the MDM over the SDM (Figure 4.4), 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, a statistically-significant support for the MDM was found after correcting for possible Denisovan admixture. I then moved to investigating the evolutionary history of the three extant Orangutan species. I 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 (Figure 4.5 B). 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), the procedure here presented assigned to the same model a posterior probability of 49% (Figure 4.5 B). Moreover, the power analysis that I conducted, and that was absent in the Nater *et al.*, 2017 work, 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. Both these features can be easily addressed through the ABC pipeline presented in this thesis.

The results of this study led to the two publications listed in the Papers section, p. 113.

PAPER I: Ghirotto S*****, **Vizzari MT***, Tassi F, Barbujani G, Benazzo A. (2020). Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. *Mol Ecol Resour* 00:1–15. https://doi.org/10.1111/1755-0998.13263

PAPER II: Vizzari MT, Benazzo A, Barbujani G, Ghirotto S. (2020). A Revised Model of Anatomically Modern Human Expansions Out of Africa through a Machine Learning Approximate Bayesian Computation Approach. *Genes* 11(12):1510. https://doi.org/10.3390/genes11121510

5. Inference using Low-Coverage data

Sequencing depth is an important feature of NGS data because it is strictly related to the accuracy in the identification of genetic variants within whole genomes. The correct characterization of variable sites leads to more accurate genotype calling and hence to more reliable demographic inferences. Polymorphic sites called from high coverage data are more accurate than those detected from data covered at lower sequencing depth (Fumagalli *et al.*, 2013). In an ABC context, observed genomic variation is compared to the genomic variation simulated under different evolutionary scenarios in order to identify the model, among those tested, that produced datasets closer to the observed ones. The simulated data must have the same features of the observed data, in terms of number and length of available loci and number of individuals sampled per population. Furthermore, since simulated genotypes are considered “true genotypes”, they should be compared only with observation coming from high quality and high coverage sequenced data. For this reason, poor quality (i.e., low-coverage) data should not be directly used to perform inferential analysis based on genetic simulations; to date, no ABC studies exploiting low-coverage genomes are available. On the other hand, the amount of low-coverage data available is significant, especially concerning genomic data coming from ancient remains, and it would be of great interest to have the possibility of exploiting this information in a model-based inferential framework.

For this reason, in the second part of my PhD I concentrated my efforts in developing and implementing an ABC framework able to efficiently deal with low-coverage whole-genome data, providing unbiased parameter estimates and model selection. I also estimated the impact of coverage level in generating accurate results, both for model selection and parameters estimation procedures.

The idea behind the development of this new framework is to integrate the uncertainty typical of low-coverage data in the simulations step through the generation of genotype likelihoods for specific coverage level, instead of genotypes. The *FDSS* it is then estimated directly from the genotype likelihoods (GLs) in both observed and simulated data. Through the so simulated GLs we are able to explicitly account for differences in the coverage level and to consider the sequencing error rate. In this way, the simulated datasets will show the same feature of the low-coverage observed datasets, thus allowing us in principle to perform a safe comparison and an unbiased inferential procedure.

I calculated the *FDSS* based on GLs from the output of the *ms* coalescent simulator

(Hudson, 2002) using an in-house python script, following the steps detailed below:

1. Simulate a certain number of independent loci of length w base pairs, using the ms coalescent simulator.
2. For each locus, for each one of the w sites (both polymorphic and monomorphic), I sample the number of "reads" covering that site for each diploid individual according to a Poisson distribution with a mean equal to a user-defined mean coverage. A specific mean coverage for each simulated individual could be optionally set.
3. Sampled "reads" are then assigned to the two chromosomes (assuming diploidy) of each individual accordingly to a binomial distribution with a probability of success of 0.5.

By sampling the number of available reads from a Poisson distribution given a mean coverage, I take into account the possibility that some nucleotide positions may be not covered, thus making our method able to deal with the presence of missing data (typically observed in low-coverage genomes).

4. At this point, a Phred quality score is assigned to each sampled read. The Phred quality score is a measure of the quality of the identification of the nucleobases generated by DNA sequencing, and it represents the chances that a base is incorrectly called and is defined as $Q = -10\log_{10} P$, were P represent the base-colling error probability. In the simulations, I assumed that every base has the same quality score of 30. A Phred quality score Q30 means that the probability of an incorrect base call is 1 in 1,000.
5. A certain amount of errors (wrong nucleotides) are introduced according to a binomial distribution $B(n,p)$, where n is the number of reads covering each site and p is the NGS error rate (that I fixed to 1% as observed for the Illumina platform).
6. Given the error rate, the Phred quality scores and the list of nucleotides (sampled "reads", D), we can finally calculate for each position the genotype likelihoods, $P(D|G)$, for all the 10 possible diploid genotypes following the method implemented in *GATK* (Van der Auwera *et al.*, 2013):

$$P(D|G=\{A_1,A_2\}) = \prod_{i=1}^{M} P(b_i|G=\{A_1,A_2\}) = \prod_{i=1}^{M} (\frac{1}{2}P(b_i|A_1) + \frac{1}{2}P(b_i|A_2))$$

$$P(b|A) = \begin{cases} \dfrac{e}{3} & : b \neq A \\ 1-e & : b \neq A \end{cases}$$

where *M* is the sequencing depth, b_i is the observed base in read *i* and *e* is the probability of error calculated from the Phred quality score.

7. The conditional posterior probability of the genotype G given the observed data *D*, formally *P(G|D)*, is finally computed according to the Bayes' Theorem:

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}$$

where *D* is the list of nucleotides observed at each site. This probability depends on the prior probability of the genotype, *P*(*G*), and on the conditional probability of the data given the genotype, *P*(*D*|*G*), computed in 6.

The prior probability of a genotype, *P*(*G*), represents how probably we expect to see a certain genotype in the population according to the evolutionary model. In this framework the prior of the genotype frequency is computed using the allele frequency counts from reads under the assumption of Hardy-Weinberg Equilibrium.

8. In the case of a single population, the genotype posterior probabilities across all individuals are used to estimate the probability of the site to be segregating in the population, $P(POPx_{poly})$, as follow:

$$P(POPx_{mono}) = \sum_{G}^{AA,TT,CC,GG} \left(\prod_{k=1}^{n_x} P(G|D) \right)$$

$$P\big(POPx_{poly}\big) = 1 - POPx_{mono}$$

where *P*(*G*|*D*) represents the posterior probabilities of the homozygous genotypes and n_x indicates the number of individuals sampled from the population *x*.

Single-site probabilities were combined to determine the expected number of segregating sites in a single locus (of length *w*) and the *FDSS* was finally computed across all the simulated loci.

In case of two populations, I computed the probability of the site to be:

a) Monomorphic in *Pop1* and *Pop2*:

$$P(mm) = P(POP1_{mono}) + P(POP2_{mono})$$

b) Segregating in *Pop1* but monomorphic in *Pop2*:

$$P(pm) = P(POP1_{poly}) + P(POP2_{mono})$$

c) Monomorphic in *Pop1* but segregating in *Pop2*:

$$P(mp) = P(POP1_{mono}) + P\left(POP2_{poly}\right)$$

d) Segregating in *Pop1* and *Pop2*:

$$P(pp) = P(POP1_{poly}) + P\left(POP2_{poly}\right)$$

e) Fixed for different alleles in *Pop1* and *Pop2*:

$$P(fixed) = 1 - (P(mm) + P(pm) + P(mp) + P(pp))$$

The *FDSS* was then computed across loci for each segregating sites category as previously described.

The approach described above works by position, generating 10 genotype posterior probabilities at both polymorphic and monomorphic sites. Computing the genotype posterior probabilities for a short locus is very fast but the entire process may easily become computationally intensive and time consuming when analyzing thousands of independent loci of several Kb in length for each coalescent simulation. To partially overcome this issue, I decided to approximate the estimation of the GLs for the monomorphic fraction of each simulated locus, that is the most demanding phase of the framework, as described below.

Since the genotype posterior probabilities for a monomorphic site only depend on the error rate and the mean coverage, I repeated the steps 2-7 in a large sample of monomorphic sites (10,000 sites). Then, I used the formulas in 8 to compute the expected probability for a monomorphic site to be correctly identified as monomorphic (a) or polymorphic in every segregating sites category (b, c, d, e) due to the error rate and mean coverage. These expected probabilities were generated once and then used to model the monomorphic part of each locus in all the simulated dataset, independently for each of the combination of experimental condition tested.

5.1. Power Analysis

To evaluate the robustness of our procedure, I carried out an extensive simulation study conducting a power analysis on different coverage levels. I also explored the inferential power of the presented approach with respect to different experimental conditions, evaluating the consequences of sampling strategies involving different numbers of chromosomes, different numbers of loci, and different locus lengths. I tested all the possible combinations of locus length (bp) {200; 1,000}, number of loci {1,000; 5,000}, number of chromosomes sampled per population {10, 20, 50} and four different coverage levels {1x, 2x, 5x, 30x} for a total of 48 combinations of sampling strategies tested. For each combination, I generated 100,000 simulated datasets with a fixed intra-locus recombination rate ($1x10^{-8}$/bp/generation), and with a fixed mutation rate ($1x10^{-8}$ /bp/generation). I evaluated the power considering two sets of models that are detailed below. The *FDSS* were estimated from the genotype likelihood calculated from the *ms* (Hudson, 2002) output of each simulation through a in-house python script. For each combination of experimental conditions, I compared alternative one- and two-population models treating each simulated dataset as pseudo-observed data (*pods*); similarly, to evaluate the power of our framework in estimating demographic parameters, I generated 1,000 pods for each model and combination of experimental condition tested. All the ABC-RF estimates have been obtained using the functions *abcrf*, for model selection, and *regAbcrf*, for parameters estimation, and employing forests of 2,000 trees; both functions are integrated in the R-package *abcrf* (Pudlo *et al.*, 2016; Raynal *et al.*, 2019). I evaluated the out-of-bag classification error (CE) and the proportion of True Positives (1-CE) as a measure of the power of the model selection procedure and, to determine the power of our procedure in estimating the demographic parameters of a true model, I calculated all the indices detailed in **Chapter 3** (*Method*, section *3.3*).

5.1.1. One-population models

I first analyzed a set of models involving a single population evolving under three different scenarios (Figure 5.1). The first model (*Constant*) represents a population with a constant effective population size through time (*N1*). The second model (*Bottleneck*) describe a population that *T* generations ago has undergone an instantaneous bottleneck event*,* with the effective population size decreasing from *NaBott* to *N1Bott*. The third model (*Exponential Growth*) represents an exponentially growing population. The expansion starts *T* generations ago, with the effective population size increasing from *NaExp* to *N1Exp*. The demographic parameters associated to each demographic model are drawn from uniform prior distributions (Table 5.1).

Figure 5.1. One-population models. Demographic models compared: Constant, Bottleneck, Expansion.

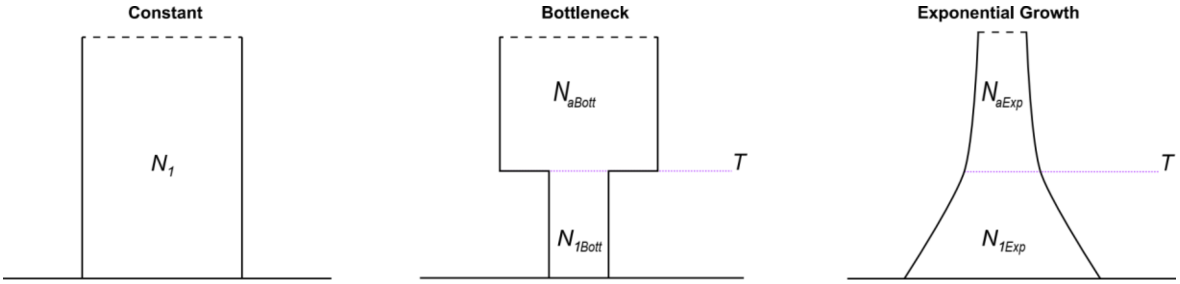


Table 5.1. Demographic parameters and prior distributions of One-Population models. Mutation and Recombination rates are expressed per nucleotide per generation.

Demographic Parameters	**Prior Distributions**
Effective population size *(N1)*	Uniform {500:50,000}
Effective population size *(NaBott)*	Uniform {25,000:100,000}
Effective population size *(N1Bott)*	Uniform {500:5,000}
Effective population size *(NaExp)*	Uniform {500:5,000}
Effective population size *(N1Exp)*	Uniform {25,000:100,000}
Time bottleneck *(T)*	Uniform {100:20,000}
Time exponential growth *(T)*	Uniform {100:20,000}
Mutation rate	$1\text{x}10^{-8}$ {Fixed}
Recombination rate	$1\text{x}10^{-8}$ {Fixed}

5.1.2. Two-populations models

I then moved to considering three demographic models with two populations (Figure 5.2). This set of models is parametrized as detailed in **Chapter 4** (*Application to High-coverage data,* section *4.1.2*). The first model (*Divergence*) describes an ancestral population of size

Nanc that splits *Tsep* generation ago into two different populations. These two derived populations evolve with a constant population size (*N1* and *N2*) until present time. The second model (*Divergence with Migration*) also includes a continuous and bidirectional migration event between the two derived populations, from the divergence to the present. Under the third model (*Divergence with Admixture*), a single pulse of bidirectional admixture occurred at time *Tadm* after the divergence. Admixture rates (*adm12, adm21*), and event' times are drawn from uniform priors; migration rates (*m12*, *m21*) are drawn from exponential priors with mean 0.1 (Table 5.2).

Figure 5.2. Two-populations models. Demographic models compared: Divergence with isolation, Divergence with migration, Divergence with a single pulse of admixture.

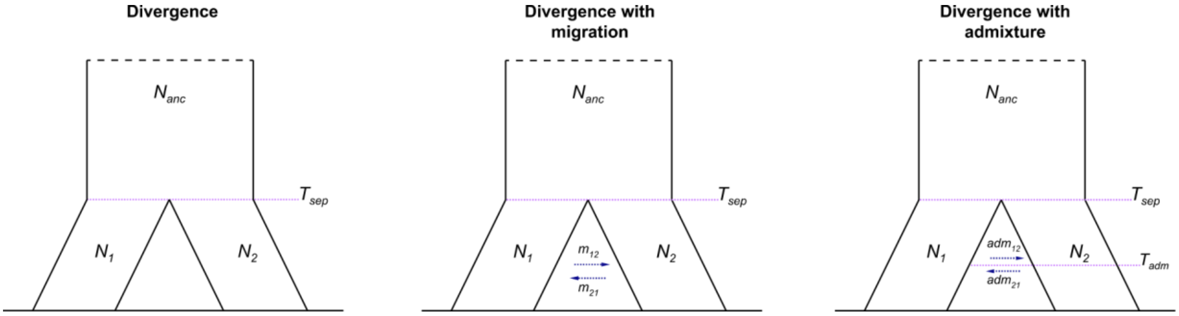


Table 5.2. Demographic parameters and prior distributions of Two-Populations models. Mutation and Recombination rates are expressed per nucleotide per generation. Time is in generations. In the simulation step I considered a *Tadm* value only if (*Tsep-Tadm*)/*Tsep* was between 0.2 and 0.8.

Demographic Parameters	**Prior Distributions**
Effective population size *(Nanc, N1, N2)*	Uniform {500:50,000}
Time split *(Tsep)*	Uniform {300:20,000}
Migration rate *(m12, m21)*	Exponential {0.1}
Time admixture *(Tadm)*	Uniform {50:2,500}
Admixture rate *(adm12, adm21)*	Uniform {0.05:0.20}
Mutation rate	1×10^{-8} {Fixed}
Recombination rate	1×10^{-8} {Fixed}

5.2. Results

5.2.1. Model Selection

5.2.1.1. One-population models

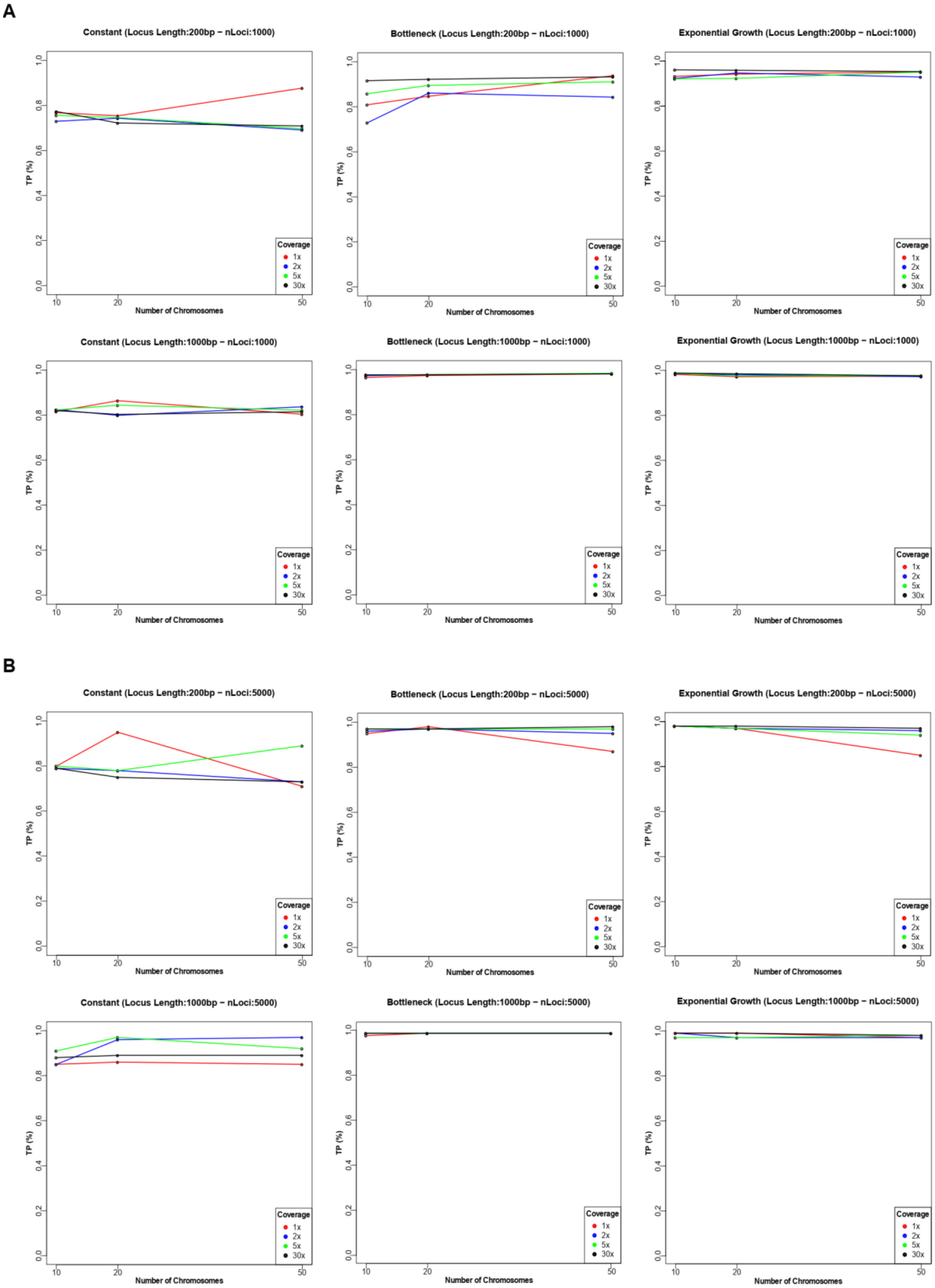
The plots of Figure 5.3 report the results of the power analyses obtained summarizing the data through the *FDSS* estimated from genotype likelihoods (GLs). In each plot, I reported the True Positive (TP) rates for each demographic model and the combination of experimental parameters tested; plots in Panel A show the TP percentages obtained simulating 1,000 loci, whereas in Panel B those obtained simulating 5,000 loci.

In general, the true positives rate is quite high, ranging from almost 80% to 100% depending on the model and on the combination of parameters generating the data. The exponential growth model is the most clearly identifiable ones, with all combinations of experimental conditions yielding nearly 100% of true positives. On the contrary, the least identifiable model is the one describing a constant population, with TP rate ranging from 75% to 90%.

The proportion of true positives generally increase with the increase of both locus length and number of loci considered; the number of chromosomes does not seem to affect the true positives rate. These observations are consistent with the results obtained for the High-Coverage data, presented in **Chapter 4** and detailed in **Paper I** (Ghirotto *et al.,* 2020).

As expected, the proportion of TP increase also with the increase of the coverage level. For both bottleneck and exponential growth models the proportions of true positives obtained simulating the data at low coverage (1x, 2x and 5x) are comparable with those obtained for the high coverage (30x), ranging from 95% to 100% for almost all the combinations of parameters tested; differences are observed only for the bottleneck model when considering 1,000 short loci. In this specific case the proportion of TP varies between 0.78 and 0.95. For the constant model, the true positives percentage considering data with the lower coverage levels (about ~80%-90% TP) is higher than that obtained with a coverage level of 30x (~80% TP); this difference is more pronounced when analyzing short loci.

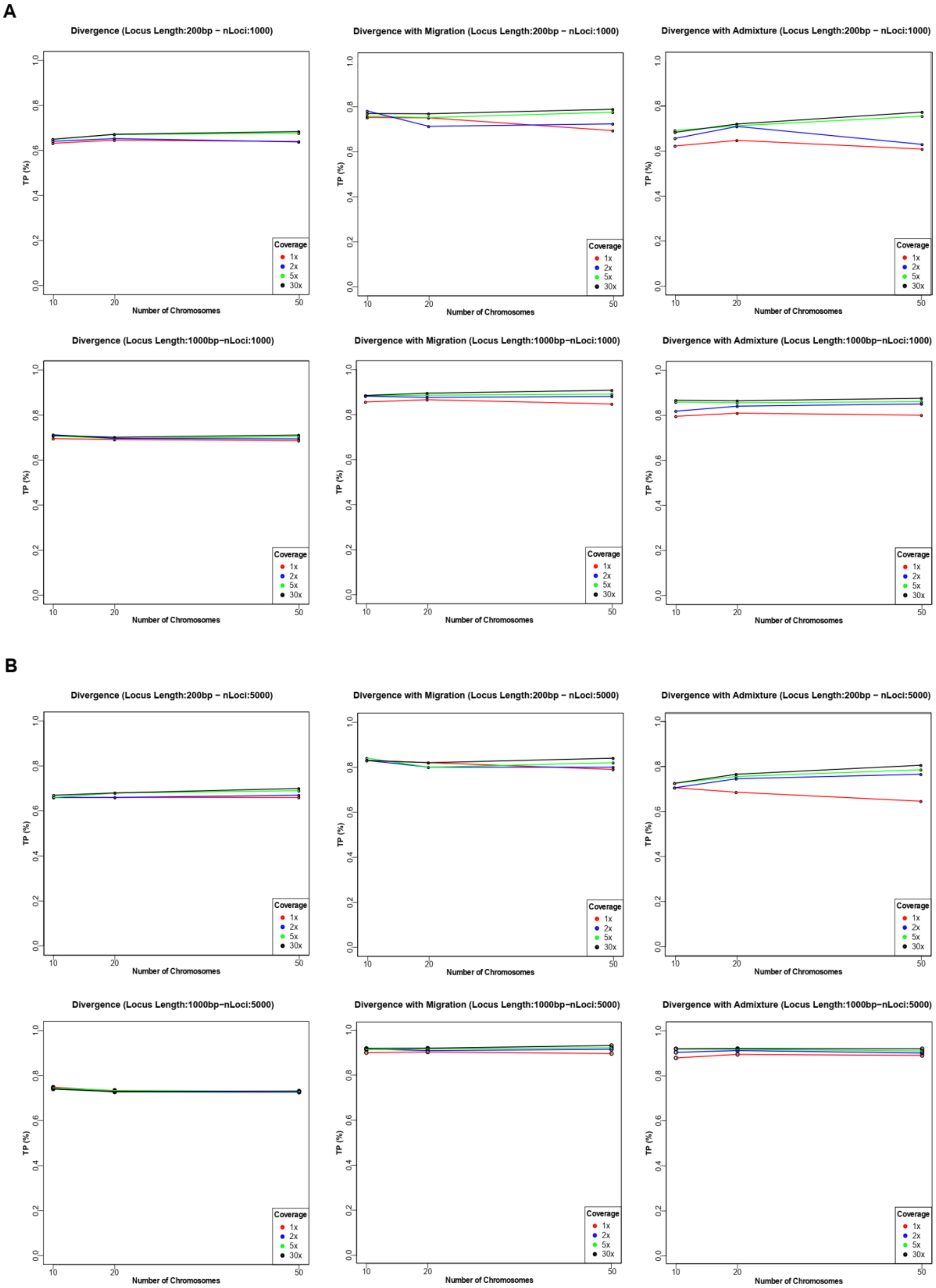
Figure 5.3. Proportion of True Positives for the one-population models. The plot below represents the proportion of TPs obtained analyzing pods coming from the three models under 48 combinations of experimental parameters. A) Combinations considering 1,000 loci. B) Combinations considering 5,000 loci. Number of chromosomes is in the x-axis, coverage levels are represented by different colors.



5.2.1.2. Two-population models

The plots in Figure 5.4 show the results for the two-populations models. 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, ranging from 62% to 90%. The true positives rate for the divergence model is lower, ranging from 60% to 78%. Once again, the proportion of TP increase with the increasing of locus length and number of loci; the number of chromosomes does not affect the true positive rate. The results do not show significant differences in the percentage of true positives when simulating the data at different coverage levels. As observed for the one-population models, the TP rate obtained when analysing low-coverage levels is comparable with that obtained with high coverage data.

Figure 5.4. Proportion of True Positives for the two-populations models. A) Combinations considering 1,000 loci. B) Combinations considering 5,000 loci. The plots have the same features of Figure 5.3.



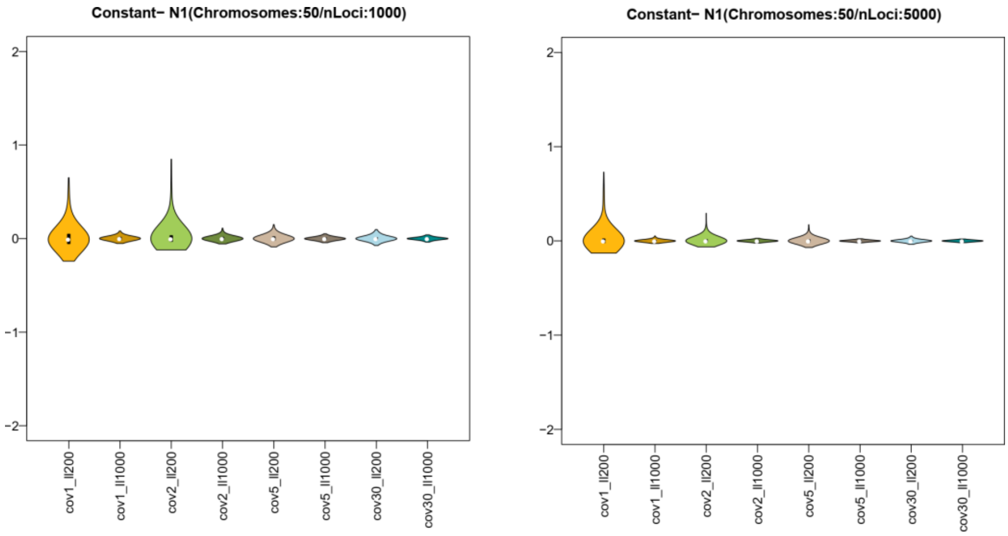
5.2.2. Parameters Estimation

5.2.2.1. One-population models

Supplementary Tables 5.1-5.6 report the results of the quality assessment of the parameters estimation procedure for the one-populations models. The R^2, Bias, RMSE, Factor2 and 50% Coverage associated to each demographic parameter were estimated exploiting 1,000 pseudo-observed datasets (pods) and reference tables of 100,000 simulated datasets. In each box I listed the indices calculated for each coverage level tested (1x, 2x, 5x and 30x) together with the number of chromosomes and the locus lengths; in Supplementary Tables 5.1, 5.3 and 5.5 I reported all the combinations considering 1,000 loci, whereas in Supplementary Tables 5.2, 5.4 and 5.6 those considering 5,000 loci. The quality of the estimates is generally consistent varying the number of chromosomes sampled thus, for the sake of clarity, only the combinations of experimental parameters considering 50 chromosomes per population are reported in the main text. Figures 5.5, 5.6 and 5.7 report the distribution of the relative Bias values over the 1,000 pods as a general index quantifying the goodness of the estimates.

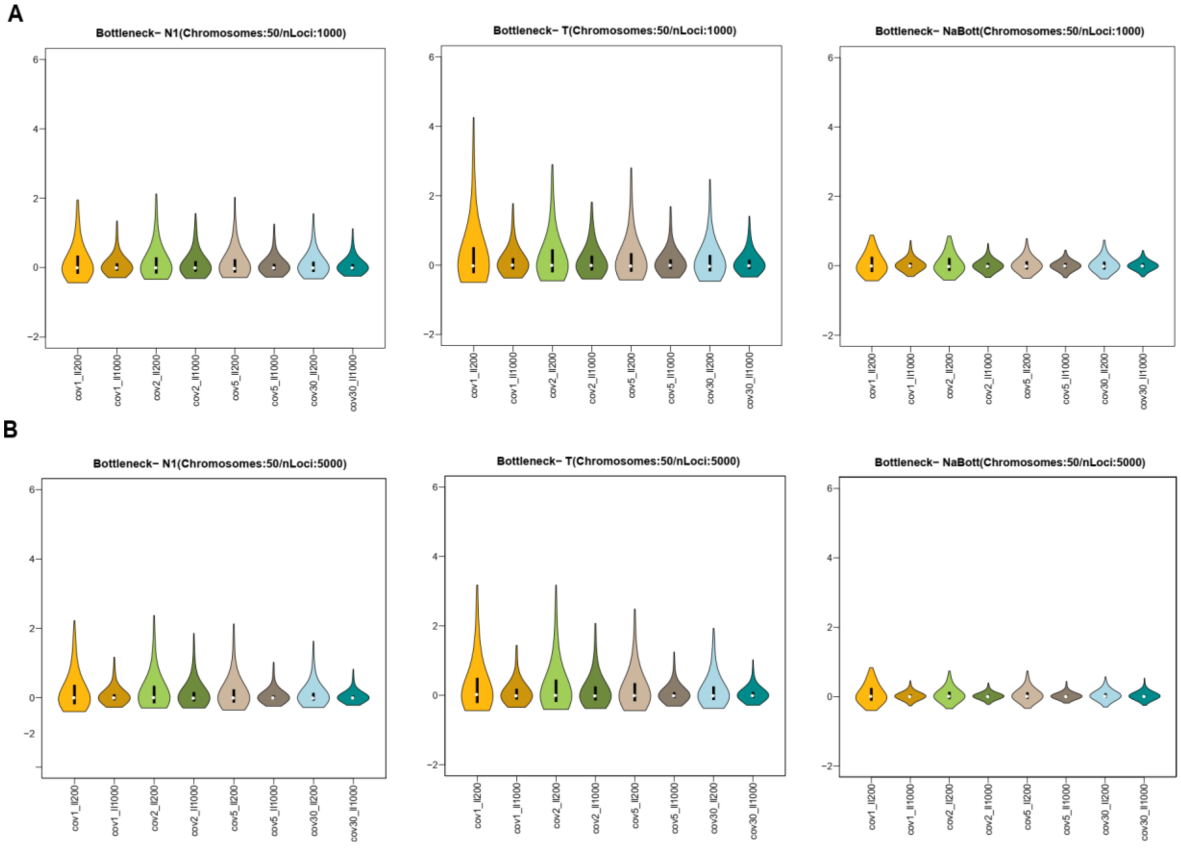
The constant model is defined by a single demographic parameter, i.e. the effective population size (*N1*); the quality assessment procedure indicates a good ability to estimates *N1*, with R^2 values reaching 100% and median Bias values always close to 0 for all the combinations of experimental conditions tested. The variance of the estimates decreases when using more loci (Figure 5.5 left vs right panel) or when increasing their length. In general, the coverage level seems to not affect the median quality estimate measures, indeed the results obtained considering low-coverage levels (1x, 2x and 5x) are comparable with those obtained with a 30x coverage (Supplementary Tables 5.1, 5.2 and Figure 5.5). However, the dispersion of the relative bias is considerably reduced for shorter loci (200bp in length) for coverage levels higher than 2x. The same pattern is not observed for longer loci (1,000bp) where a small dispersion was already observed at 1x coverage.

Figure 5.5. Relative Bias distributions for the Constant model's demographic parameters. Coverage levels are represented by different colors. Lighter colors indicate short loci (200bp), darker colors indicate longer loci (1,000bp). Left panel: combinations considering 1,000 loci. Right panel: combinations considering 5,000 loci.



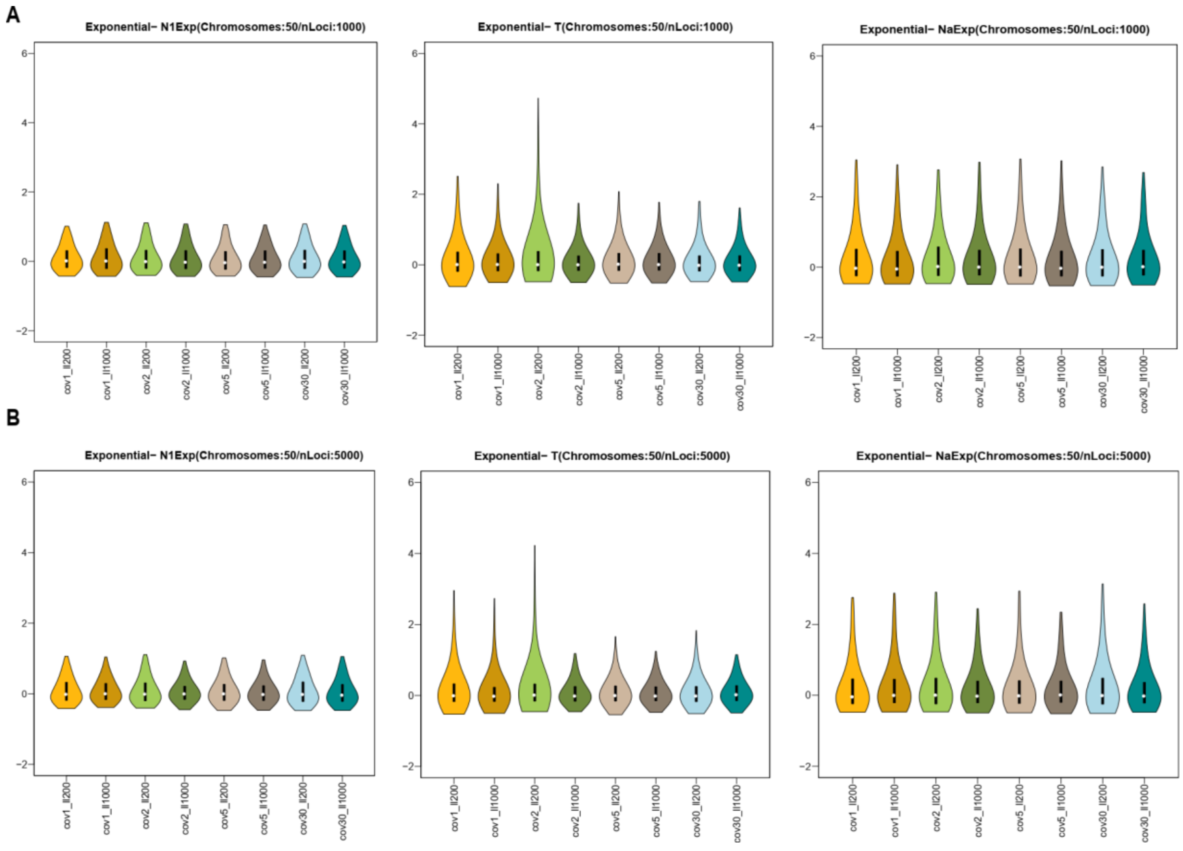
The quality estimates of the three demographic parameters defining the bottleneck model (*N1*, *NaBott* and *T*) shows similar results with R^2 values >50% and median Bias values ranging from 0.01 to 0.4. All quality indices improve with the increase of the number and the length of loci and the number of chromosomes. As regarding to the coverage levels, I observed a slight improvement of the estimates with the increase of the sequencing depth (median Bias 30x: 0.01-0.2; median Bias 1x: 0.04-0.4). Even in this case, the dispersion of the relative Bias decrease with the increase of the number and length of loci. However, the observed reductions in the dispersion of the relative bias seems to be very limited across almost all the experimental condition analyzed and among demographic parameters (Supplementary Tables 5.3, 5.4 and Figure 5.6).

Figure 5.6. Relative Bias distributions for the Bottleneck model's demographic parameters. These plots have the same features of Figure 5.5. A) Combinations considering 1,000 loci. B) Combinations considering 5,000 loci**.**



For the exponential growth model, I observed the same general results with R^2 >10% and median Bias that varies between 0.06 and 0.5 for most of the combinations of experimental condition tested. The quality of the estimates slightly increases with the increase of locus length, number of loci and number of chromosomes. The demographic parameter that shows lower quality indices is the current effective population size (*N1*) with R^2 values ranging from 2% to 30% depending on the experimental conditions; in particular, the R^2 is lower than 10% when I considered only 1,000 loci and 10 sampled chromosomes (Supplementary Tables 5.5, 5.6). I did not observe significant differences in the quality of the estimates with respect to the coverage levels analysed. Even in this case the results obtained with 1x are indeed comparable with those obtained with 30x (Supplementary Tables 5.5, 5.6 and Figure 5.7).

Figure 5.7. Relative Bias distributions for the Exponential Growth model's demographic parameters. These plots have the same features of Figure 5.5. A) Combinations considering 1,000 loci. B) Combinations considering 5,000 loci**.**



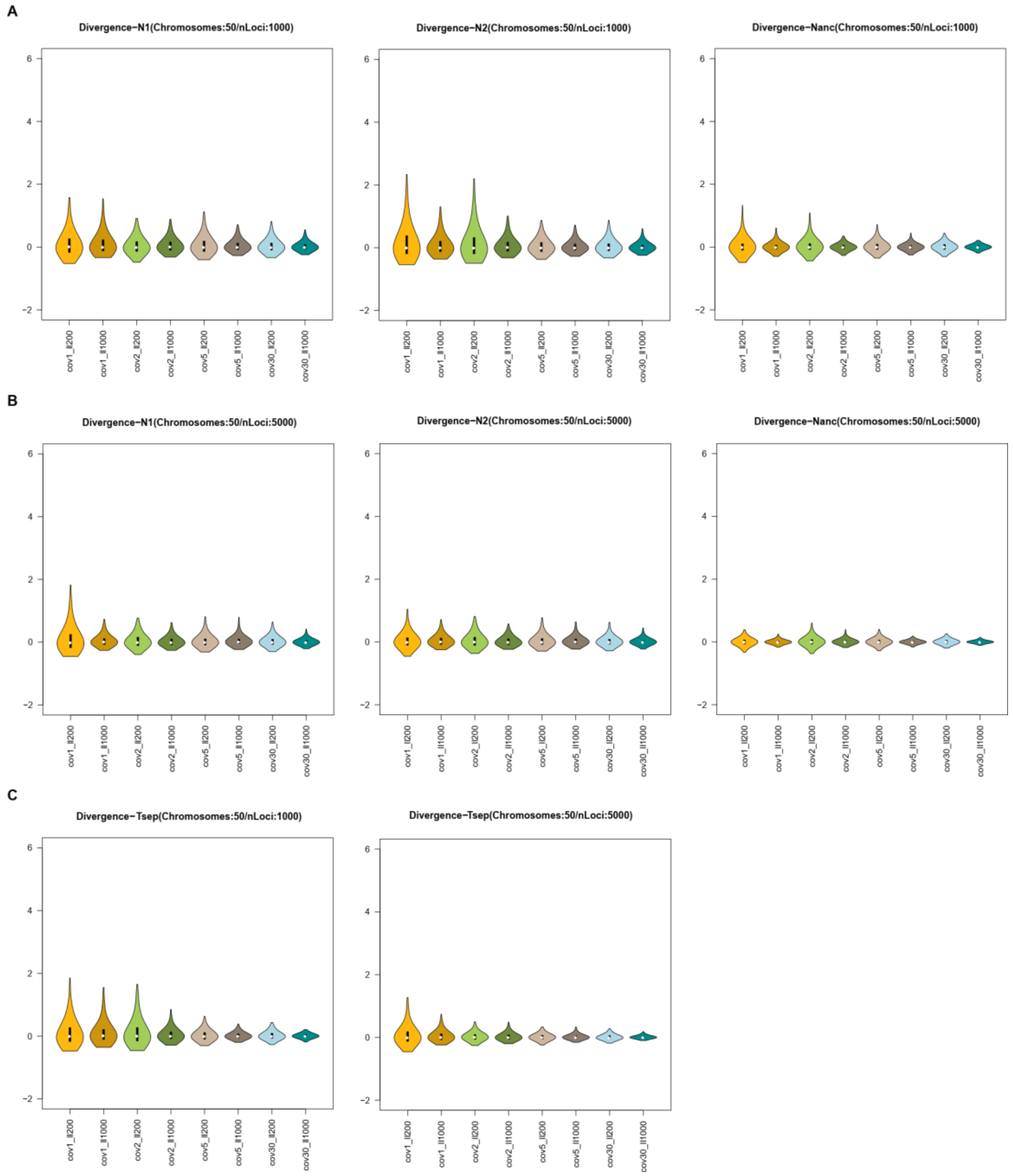
5.2.2.2. Two-population models

In the following tables are reported the results of the quality assessment of the parameters estimation procedure for the two-populations models (Supplementary Tables 5.7-5.12). These tables are structured in the same way of the tables presented for the one-population models. Figures 5.8, 5.9, 5.10 report the distribution of the relative Bias values over the 1,000 pods, for the combinations of experimental parameters considering 50 chromosomes.

The general results of the quality of the estimates are similar to those obtained for the one-population models, with $R^2 > 50\%$ and median Bias values ranging from 0.003 to 0.4 for most of the demographic parameters estimated (Supplementary Tables 5.7-5.12). The estimates improve with the increase of the number of chromosomes, number and length of loci considered in the analysis (Figures 5.8-5.10). As for the one-population models, the effect of the coverage level on the quality of the estimates is negligible, the results remain indeed consistent regardless the sequencing depth (Figures 5.8-5.10).

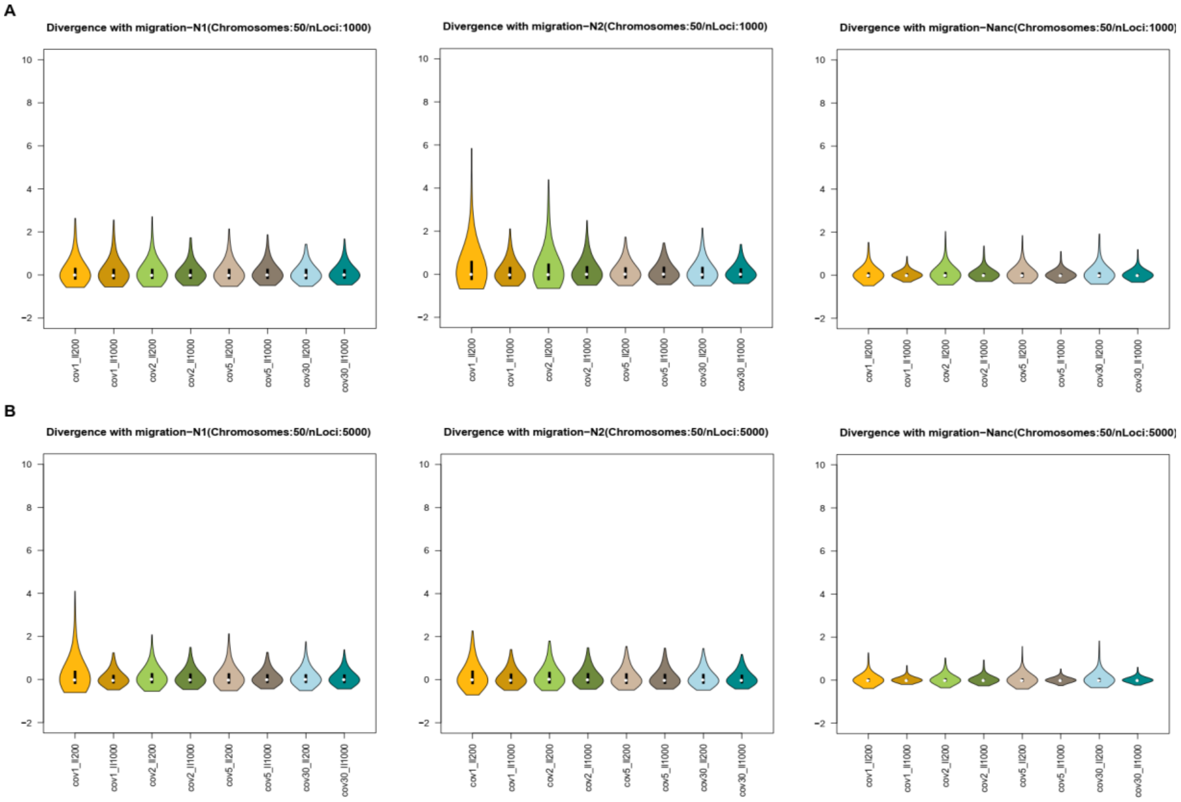
The model that shows the best quality indices is the divergence model. In this scenario the demography is defined by four demographic parameters: the effective population sizes of the ancestral population (*Nanc*) and of the two derived population (*N1* and *N2*) and the divergence time (*Tsep*). All these parameters show R^2 values ranging from 65% to almost 100% and low median Bias (Supplementary Tables 5.7-5.8). The variance of the relative Bias decrease, uniformly for all demographic parameters, with the increase of the number and length of loci analysed and of the sequencing depth (Figure 5.8).

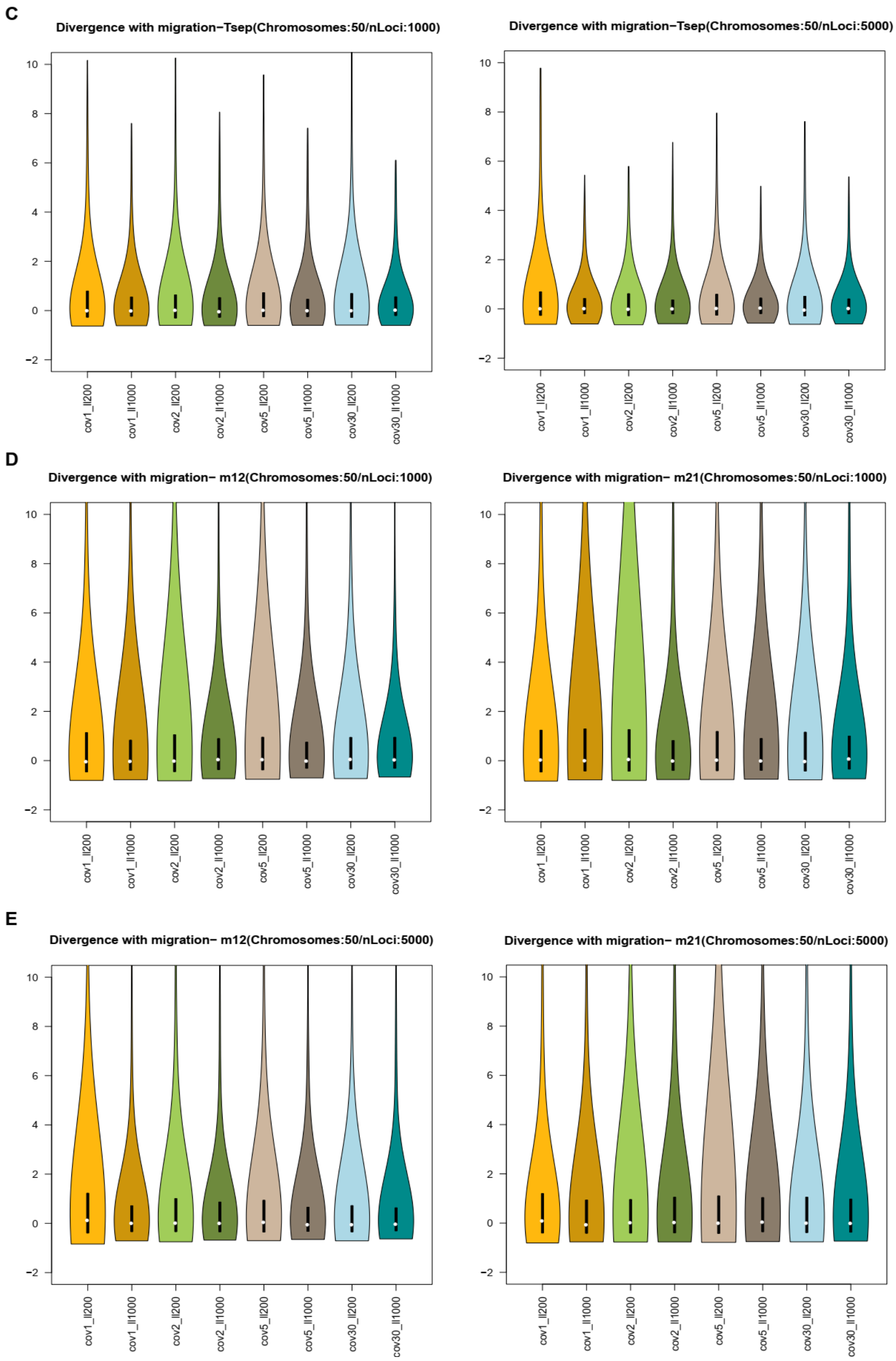
Figure 5.8. Relative Bias distributions for the Divergence model's demographic parameters. These plots have the same features of Figure 5.5. Panel A and left of Panel C: combinations considering 1,000 loci. Panel B and right of Panel C: combinations considering 5,000 loci.



The divergence with migration model shows lower quality indices compared to the other two-population models. The R^2 of the three effective population sizes varies between 40% and almost 100%, whether the median Bias's values vary between 0.06 and 0.4. The worst estimated parameters are the two migration rates (*m12* and *m21*) with $R^2 < 40\%$ and higher levels of Bias with values ranging from 1 to 34 (Supplementary Tables 5.9-5.10). As regard of the dispersion of the relative Bias, I observed a general decrease with the increase of the length and the number of loci and of the coverage level. This pattern is more evident for the three effective population sizes (*N1*, *N2* and *Nanc*; Figure 5.9 A and B); The two migration rates and the separation time (*Tsep*), show instead greater levels of variance that slightly improves with the increase of the sequencing depth, the number and length of loci (Figure 5.9 C, D and E).

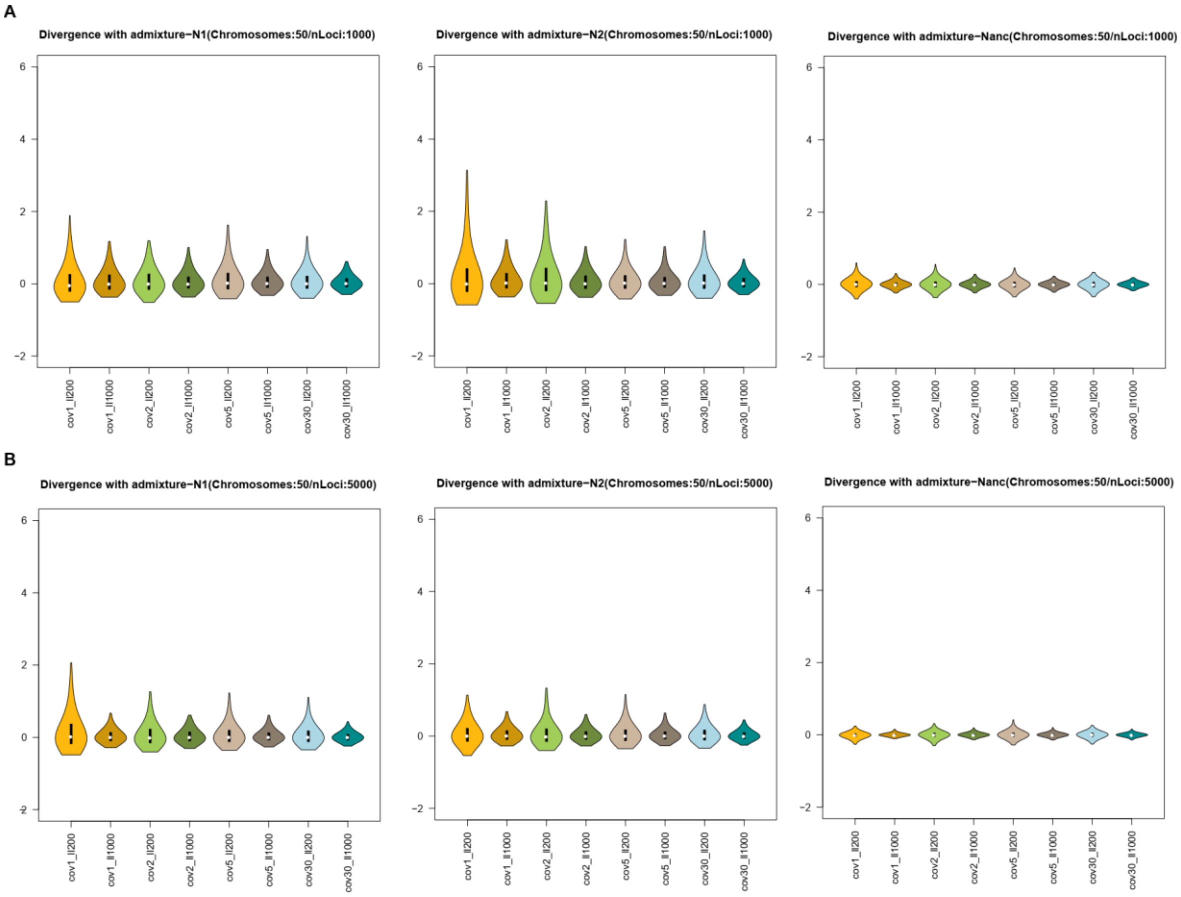
Figure 5.9. Relative Bias distributions for the Divergence with migration model's demographic parameters. These plots have the same features of Figure 5.5. Panels A, C and D: combinations considering 1,000 loci. Panels B, C and E: combinations considering 5,000 loci.

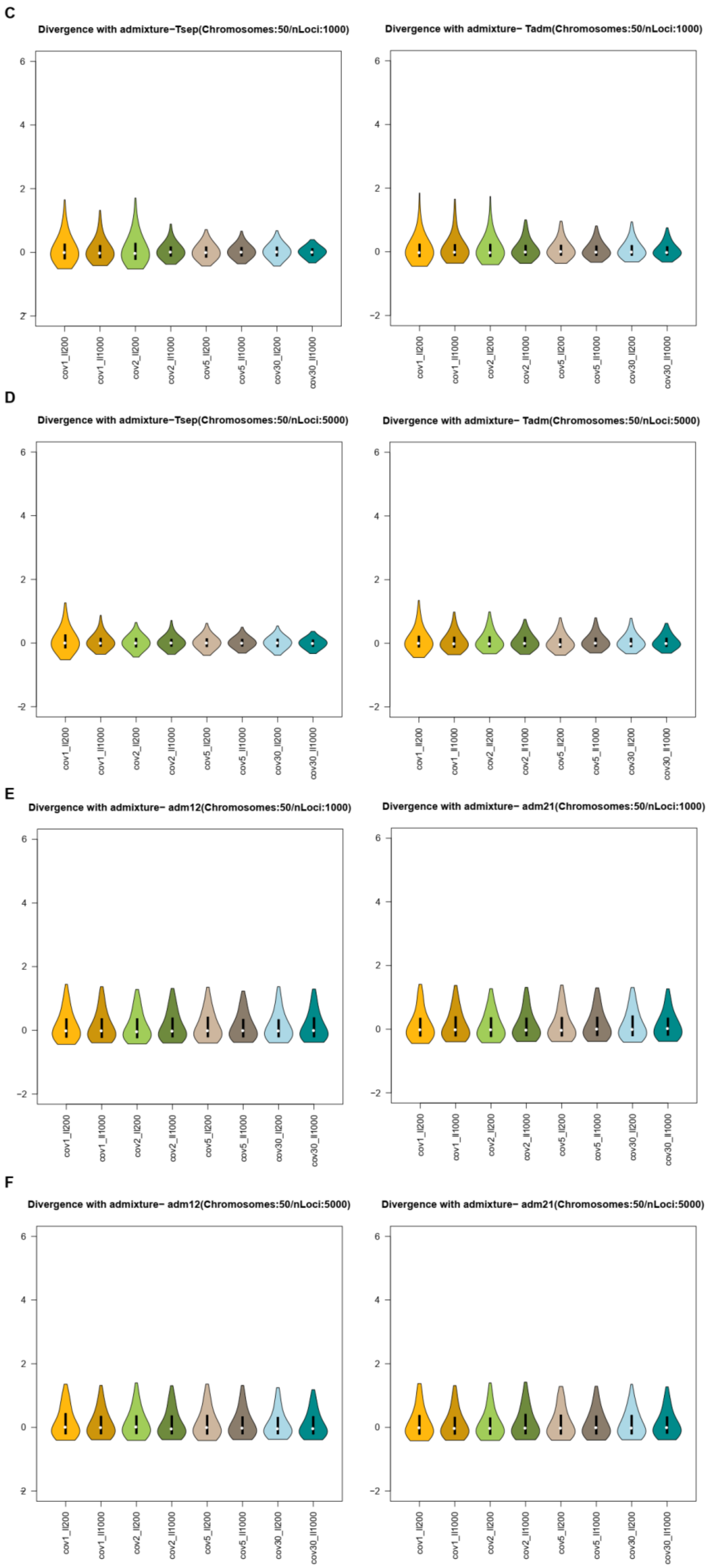



C
Divergence with migration−Tsep(Chromosomes:50/nLoci:1000)
10
8
6
4
2
0
−2
cov1_ll200
cov1_ll1000
cov2_ll200
cov2_ll1000
cov5_ll200
cov5_ll1000
cov30_ll200
cov30_ll1000
Divergence with migration−Tsep(Chromosomes:50/nLoci:5000)
10
8
6
4
2
0
−2
cov1_ll200
cov1_ll1000
cov2_ll200
cov2_ll1000
cov5_ll200
cov5_ll1000
cov30_ll200
cov30_ll1000
D
Divergence with migration− m12(Chromosomes:50/nLoci:1000)
10
8
6
4
2
0
−2
cov1_ll200
cov1_ll1000
cov2_ll200
cov2_ll1000
cov5_ll200
cov5_ll1000
cov30_ll200
cov30_ll1000
Divergence with migration− m21(Chromosomes:50/nLoci:1000)
10
8
6
4
2
0
−2
cov1_ll200
cov1_ll1000
cov2_ll200
cov2_ll1000
cov5_ll200
cov5_ll1000
cov30_ll200
cov30_ll1000
E
Divergence with migration− m12(Chromosomes:50/nLoci:5000)
10
8
6
4
2
0
−2
cov1_ll200
cov1_ll1000
cov2_ll200
cov2_ll1000
cov5_ll200
cov5_ll1000
cov30_ll200
cov30_ll1000
Divergence with migration− m21(Chromosomes:50/nLoci:5000)
10
8
6
4
2
0
−2
cov1_ll200
cov1_ll1000
cov2_ll200
cov2_ll1000
cov5_ll200
cov5_ll1000
cov30_ll200
cov30_ll1000

Finally, for the divergence with a single pulse of admixture model I observed good quality indices for the three population sizes (*N1*, *N2* and *Nanc*), the time of the admixture event (*Tadm*) and the divergence time (*Tsep*) with $R^2 > 50\%$ and low level of median Bias with values ranging from 0.005 to 0.1. The worst estimated parameters are the two admixture rates (*adm12* and *adm21*) with $R^2 < 10\%$ for most of the combination of experimental parameters tested (Supplementary Tables 5.11-5.12). Even in this case, the variance of the relatives Bias decrease with the increase of locus length, number of loci and coverage levels (Figure 5.10). This pattern is more evident for the three effective population sizes (Figure 5.10 A and B), than for the divergence time (*Tsep*), the admixture time (*Tadm*) and the two admixture rates (Figure 5.10 C, D, E and F).

Figure 5.10. Relative Bias distributions for the Divergence with admixture model's demographic parameters. These plots have the same features of Figure 5.5. Panels A, C and E: combinations considering 1,000 loci. Panels B, D and F: combination considering 5,000 loci.




C
Divergence with admixture−Tsep(Chromosomes:50/nLoci:1000)
Divergence with admixture− Tadm(Chromosomes:50/nLoci:1000)
D
Divergence with admixture−Tsep(Chromosomes:50/nLoci:5000)
Divergence with admixture− Tadm(Chromosomes:50/nLoci:5000)
E
Divergence with admixture− adm12(Chromosomes:50/nLoci:1000)
Divergence with admixture− adm21(Chromosomes:50/nLoci:1000)
F
Divergence with admixture− adm12(Chromosomes:50/nLoci:5000)
Divergence with admixture− adm21(Chromosomes:50/nLoci:5000)
6
4
2
0
−2
cov1_ll200
cov1_ll1000
cov2_ll200
cov2_ll1000
cov5_ll200
cov5_ll1000
cov30_ll200
cov30_ll1000

5.3. Discussion

In the past years, Next-Generation Sequencing (NGS) technologies have revolutionized population genetics studies, and now high-quality whole-genome data can be safely used to investigate the past demographic dynamics of many species (Prüfer *et al.*, 2014; Mallick *et al.*, 2016; Benazzo *et al.*, 2017; Nater *et al.*, 2017, Ghirotto *et al.*, 2020). However, despite the large availability of complete genomes data, the number of well covered genomes (i.e., that have been sequenced at high coverage level), for which the genotypes are known with certainty, are still limited. This is true in particular for ancient DNA data (Haber *et al.*, 2016) or for non-model species (Beichman *et al.*, 2018). Furthermore, for large-scale population studies, where many samples must be sequenced to obtain a broad picture of the genetic variation of the population, producing high-coverage genomes may not be economically feasible. These scenarios pushed researchers to lean on low-coverage sequencing strategies, sacrificing confidence in genotype-calling in return for much greater sample sizes (Fumagalli *et al.*, 2013).

When analysing low-coverage genomes in a population genetic context we should take into account the genotype uncertainty and sequencing errors associated to low quality data. In the recent past, some methods have been developed to this aim that exploit the genotype likelihoods to provide estimates of genetic variation (Korneliussen *et al.*, 2014; Kousathanas *et al.*, 2017; Meisner and Albrechtsen, 2018). Unfortunately, there is still a lack of methods to reconstruct past events with this type of data, so the main challenge now is to develop more flexible and robust statistical frameworks to be able to exploit low-depth genomes in an inferential context with the aim of making accurate demographic inference.

The reconstruction of the past demographic histories relies on the pattern of genetic variation shown by the sampled populations; this means that an accurate estimation of genotypes is crucial for a reliable inference of past processes. One of the approaches to reconstruct complex evolutionary dynamics is represented by an Approximate Bayesian Computation (ABC) framework. It exploits coalescent simulations to generate the expected level of variation, represented by known genotypes, under different evolutionary scenarios. Demographic inference is then performed by comparing the simulated data with the genotypes called in the sampled individuals. The low sequencing depth drastically affects the ability to reliably call genotypes, thus making low-coverage data unsuitable for such powerful inferential approaches.

A possible strategy to integrate the use of low coverage data in an ABC context relies on correcting the uncertainty linked to the coverage level in the calling of the genotypes when performing the calculation of the observed Summary Statistics; a way to do this is to work with genotypes likelihoods. ANGSD (Korneliussen *et al.*, 2014) is one of the most widely used tools to calculate population genetics indices from low-coverage data through the GLs, and can be in principle exploited to make the observed statistics comparable with those generated through simulations. Correcting the observed statistics for the coverage level would represent a fast and flexible option to integrate low-coverage genomic data within an ABC approach; however, our simulation experiments show that the correction performed by ANGSD is not always effective, as in case of low sequencing depth. Supplementary Figure 5.1 reports the results of the power analysis I performed, to test the effectiveness of the ANGSD correction on the observed data at different coverage levels in identify the true demographic history. This power analysis has been performed under two different sets of models, namely the one- and two-populations models described in **Chapter 5** (Sections *5.1.1* and *5.1.2*). The pods generated with a high coverage level (30x) were almost always assigned to the true demographic history. Irrespective to the model's features, indeed, the TP proportion was about 80-90%, that is comparable with that expected when analyzing genotypes directly generated through simulations (Ghirotto *et al.*, 2020).

When comparing the one-population models considering a coverage of 5x, I observed a significant reduction in the power for the exponential and the constant model, with a proportion of TP that decreases from 80-90% to 0-20%. The proportion of TP further decreases with lower coverage levels -1x and 2x- where any pod was correctly assigned (TP= 0%). The bottleneck model showed, instead, high proportion of TP even at the lower coverage levels; however, I verified that this result is actually an artefact resulting from the analysis of low coverage sequencing data. There is indeed a negative correlation between the impact of the sequencing error and the level of coverage, with a higher number of sites erroneously identified as heterozygous in low coverage data. This biased prediction of genomic variation causes a distortion in the summary statistic distribution towards pattern of variation only generable from our bottleneck model; consequently, the lower is the coverage of a pod, the higher is the probability of assignment to the bottleneck model. All the low coverage pods generated through the bottleneck model were hence correctly assigned, thus explaining the unreliable high TP proportion observed.

When comparing the two-population models, with a coverage of 5x I observed a reduction in the power for the divergence and the divergence with migration model, with a proportion of TP that decreases from 80-90% to 20-80%. The TP rate further decreases with lower coverage levels, 1x and 2x (0-40%). The divergence with admixture model, showed good TP proportions, with values ranging from 60% to 90%. Even in this case, this behaviour could be explained by the bias in the correct identification of the pattern of polymorphisms due to the low-depth sequencing. With very low-coverage levels, 1-2x, the uncertainty linked to the identification of the single true genotype prevent indeed the classification of private polymorphic sites. This results in an artificially higher proportion of loci containing shared polymorphic sites with respect to sites belonging to the other three categories, that is what is naturally generated by the divergence with admixture model. Such distortion results in an higher proportion of pods assigned to the divergence with admixture model, and consequently in an high TP rate when pods actually come from the same model.

Taken together, these results highlight that the simple correction of low coverage observed data through the method embedded in ANGSD to perform ABC model choice, does not produce reliable results when the sequencing depth is low (<= 5x). This drawback would severely limit the possibility to exploit the present framework to study past demographic processes trough the analysis of genomic data from high degraded samples, as those extracted from ancient remains (whose achievable coverage is often lower than 5x) or through non-invasive sampling.

To make possible and effective the inclusion of these kind of genomic data in the ABC inferential process, I developed a new inferential framework, in which, rather than correcting the low-covered observed information, genomic data are generated through simulations according to a specific coverage level. Observed and simulated data are hence directly comparable, and statistics are calculated in both cases through genotype likelihoods. I summarized the data using the full genomic distribution of the four mutually exclusive categories of segregating sites (*FDSS*), a powerful and easy to compute statistics already successfully used to summarized whole-genome data (details in **Chapter 4** and **Paper I**).

Under this framework, the *FDSS* is not directly calculated from known genotypes, but rather computed using genotype likelihoods in both simulated and observed data, so as to take into account the uncertainty linked to low-coverage data in the estimation of model's features.

I evaluated the inferential power of ABC, coupled with *FDSS* using genotype likelihoods, in distinguishing among different demographic models and in inferring model parameters under different experimental conditions. I evaluated the effect of different levels of coverage (1x to 30x), number of individuals, number and size of the simulated genetic loci on the generation of the *FDSS*.

I defined two different set of demographic models describing a single or two populations evolving under three different scenarios.

When I compared one-population scenarios the model identifiability, calculated as the proportion of TPs over 100,000 pods, reached values between 80% and 100%, regardless of the coverage level considered. The true positives rates obtained simulating the data at a low sequencing depth (1x to 5x) are comparable with those obtained simulating the data at 30x coverage. The proportion of true positives for the constant model, considering data with the lower coverage levels, is slightly higher than that obtained with a coverage level of 30x (~80%-90% vs ~80% TP). This result may be related to the skewed level of polymorphism typical of low-coverage conditions (Nielsen *et al.*, 2012; Fumagalli *et al.*, 2013) that may amplify differences in the polymorphism levels generated by the models, especially in some regions of the parameter's space. This behavior is not observed in the bottleneck and the exponential growth models, suggesting that the performance of the inferential framework at lower coverage levels could be model dependent and hence a careful analysis of the model choice performance should be always performed before analyzing real datasets.

I then tested the power of our framework in estimating the models' parameters; the general quality of the estimates was quite good, with high R^2 values, low Bias and RMSE, and consequently high values of factor2 and 50% coverage. Once again, the quality of the estimates is not influenced by the coverage, indeed, the performances of the estimation process are comparable regardless of whether the coverage was 1x or 30x. The only case in which I observed lower quality indices was when I estimated the current effective population size (*N1*) under the Exponential Growth model. These results were somehow expected: the ability to accurately estimates the present effective population size of an exponentially growing population strictly depends on the time of the beginning of the growth; if the growth start in recent times, we need larger samples to characterize the effective population size before and after the expansion (Boitard *et al.*, 2016).

Among the two-population models the proportion of TPs ranged from 60% to 90% and the estimated demographic parameters showed high quality indices, except for both migration and admixture rates. As for the one-population models, the coverage levels considered in the analysis does not affect the power of our procedure. The results obtained for both model selection and parameters' estimation procedures seemed not to be affected by sequencing depth, suggesting that integrating the genotype likelihood in the estimation process is an effective way to deal with sequencing conditions characterized by high genotype uncertainty.

These results demonstrate that, for the first time, low-coverage sequencing data can be safely integrated into an ABC-based inferential procedure.

The results of this study highlight that the proposed framework, based on the simulation of datasets at a certain coverage level in which the *FDSS* is estimated from genotype likelihoods, is able to produce, although for very simple demographic scenarios, reliable identification of the true model and unbiased estimates of its demographic parameters.

The current study is limited to the analysis of relatively simple demographic models, but these promising preliminary results pave the way for a successful comparison of more complex models.

6. General conclusions

In recent years, thanks to the continuous development of new NGS technologies, we have witnessed an explosion in the production of whole-genome data; not only from modern samples, but also from historical samples, or directly from the environment, thus giving the possibility to investigate evolutionary processes and dynamics hitherto unexplored.

As genomic datasets grew in size, it became essential to develop new methods of analysis allowing us to deal with this kind of data and making sense of this vastness of information. In this regard, an increasing number of population genetics studies are now exploiting the flexibility and analytical power of machine learning tools (ML). One interesting advantage of ML algorithms is that they are well suited to process high-dimensional input data, being able to identify which features, among the thousand defining the data, are the most informative about the processes under investigation (Schrider and Kern, 2018).

ML techniques are currently highly exploited in bioinformatics to make prediction about the regulatory regions of the genome, to identify variants that are potentially linked to serious diseases (Zou *et al.*, 2019), and, in population genetics to identify patterns of natural selection (Schrider and Kern, 2016; Torada *et al.,* 2019) and to facilitate demographic inferences, trough ABC procedures, using large genomic datasets (Pudlo *et al.*, 2016; Mondal *et al.*, 2019). In the latter case, the introduction of ML algorithms made it possible to overcome two of the main ABC's limitations: the dimensionality of the summary statistics and the number of simulations required for a proper analysis. This would facilitate the application of ABC also to the study of complex and more realistic demographic models but does not solve the limitations related to the choice of informative statistics to efficiently summarize the genomic variation under investigation. In this PhD thesis I demonstrated that *FDSS*, coupled with ABC-RF, is an efficient inferential tool to reconstruct past complex demographic dynamics using high- and low-coverage genomes.

The results of the extensive power analysis, presented in **Chapter 4** and detailed in **Paper I**, show indeed how the *FDSS* is therefore an appropriate summary of the whole genome data for ABC-RF analysis, even for non-optimal datasets, i.e. where we cannot access a large sample of individuals or when dealing with very fragmented DNA samples.

The preliminary results presented in **Chapter 5** show the ability of the framework in distinguishing among different evolutionary scenarios and in making reliable estimate of their demographic parameters analysing low-coverage data. To this purpose, I integrated the uncertainty associated to the identification of genotypes, directly calculating the *FDSS* from the genotype likelihoods generated in the simulation step.

As future perspectives I am going to assess the power of this framework analysing more complex evolutionary scenarios. If these promising preliminary results will be confirmed, this would facilitate the integration of the information contained in low-coverage genomes (for example those coming from ancient samples) in the analysis of past population processes and will improve our ability in shedding light on evolutionary and demographic processes.

7. Bibliography

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

Akaike H (1974). A New Look at the Statistical Model Identification. *IEEE Trans Automat Contr* **19**: 716–723.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, *et al.* (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* **2**: e286.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, *et al.* (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinforma* **43**: 11.10.1-11.10.33.

Baudry E, Depaulis F (2003). Effect of Misoriented Sites on Neutrality Tests with Outgroup. *Genetics* **165**: 1619–22.

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

Beaumont MA (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annu Rev Ecol Evol Syst* **41**: 379–406.

Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–35.

Beeravolu CR, Hickerson MJ, Frantz LAF, Lohse K (2018). ABLE: blockwise site frequency spectra for inferring complex population histories and recombination. *Genome Biol* **19**: 145.

Beichman AC, Huerta-Sanchez E, Lohmueller KE (2018). Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annu Rev Ecol Evol Syst* **49**: 433–456.

Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, *et al.* (2017). Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. *Proc Natl Acad Sci* **114**: E9589–E9597.

Bertorelle G, Benazzo A, Mona S (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Mol Ecol* **19**: 2609–25.

Blum MGB, François O (2010). Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* **20**: 63–73.

Blum MGB, Nunes MA, Prangle D, Sisson SA (2013). A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Stat Sci* **28**: 189–208.

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 Genet* **12**: e1005877.

Breiman L (2001). Random forests. *Mach Learn* **45**: 5–32.

Brumfield RT, Beerli P, Nickerson DA, Edwards S V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol*.

Charlesworth B, Charlesworth D (2017). Population genetics from 1966 to 2016. *Heredity (Edinb)* **118**: 2–9.

Chen H (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theor Popul Biol* **81**: 179–195.

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 (Edinb)* **120**: 13–24.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* **25**: 410–8.

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, *et al.* (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–8.

Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, *et al.* (2012). Fast computation and applications of genome mappability. *PLoS One* **7**: e30377.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014). Ten years of next-generation sequencing technology. *Trends Genet* **30**: 418–426.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018). The Third Revolution in Sequencing Technology. *Trends Genet* **34**: 666–681.

Eldon B, Birkner M, Blath J, Freund F (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger Coalescents? *Genetics* **199**: 841–56.

Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, *et al.* (2009). The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* **324**: 522–528.

Ewing G, Hermisson J (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**: 2064–5.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet* **9**: e1003905.

Fisher F (1930). *The Genetical Theory of Natural Selection*. Clarendon Press.

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, *et al.* (2016). The genetic history of Ice Age Europe. *Nature* **534**: 200–205.

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, *et al.* (2013). Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. *Genetics* **195**: 979–992.

Ghirotto S, Vizzari MT, Tassi F, Barbujani G, Benazzo A. (2020). Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. *Mol Ecol Resour* **00**:1–15.

Gutenkunst R, Hernandez R, Williamson S, Bustamante C (2010). Diffusion Approximations for Demographic Inference: DaDi. *Nat Preced*.

Haber M, Mezzavilla M, Xue Y, Tyler-Smith C (2016). Ancient DNA and the rewriting of human history: be sparing with Occam's razor. *Genome Biol* **17**: 1.

Han E, Sinsheimer JS, Novembre J (2014). Characterizing Bias in Population Genetic Inferences from Low-Coverage Sequencing Data. *Mol Biol Evol* **31**: 723–735.

Harris H (1966). C. Genetics of Man Enzyme polymorphisms in man. *Proc R Soc London Ser B Biol Sci* **164**: 298–310.

Hernandez RD, Williamson SH, Bustamante CD (2007). Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol* **24**: 1792–800.

Hinrichs AS, Raney BJ, Speir ML, Rhead B, Casper J, Karolchik D, *et al.* (2016). UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics* **32**: 1430–2.

Hoban S, Bertorelle G, Gaggiotti OE (2012). Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet* **13**: 110–122.

Hubby JL, Lewontin RC (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. *Genetics*.

Hudson RR (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.

International Human Genome Sequencing Consortium (2004). International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*.

Jouganous J, Long W, Ragsdale AP, Gravel S (2017). Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics* **206**: 1549–1567.

Kamm JA, Terhorst J, Song YS (2017). Efficient Computation of the Joint Sample Frequency Spectra for Multiple Populations. *J Comput Graph Stat* **26**: 182–194.

Keightley PD, Jackson BC (2018). Inferring the probability of the derived vs. The ancestral allelic state at a polymorphic site. *Genetics* **209**: 897–906.

Kingman JFC (1982). The coalescent. *Stoch Process their Appl* **13**: 235–248.

Korneliussen TS, Albrechtsen A, Nielsen R (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Kousathanas A, Leuenberger C, Link V, Sell C, Burger J, Wegmann D (2017). Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics* **205**: 317–332.

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

Li H, Durbin R (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–95.

Li H, Durbin R (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–6.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9.

Lohmueller KE (2014). The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLoS Genet* **10**: e1004379.

Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, *et al.* (2016). A genomic history of Aboriginal Australia. *Nature* **538**: 207–214.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, *et al.* (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**: 201–206.

De Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, *et al.* (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**: 477–481.

Marjoram P, Molitor J, Plagnol V, Tavare S (2003). Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci* **100**: 15324–15328.

Mathieson I (2020). Human adaptation over the past 40,000 years. *Curr Opin Genet Dev* **62**: 97–104.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, *et al.* (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**: 499–503.

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 (Edinb)* **116**: 362–71.

McVean GAT, Cardin NJ (2005). Approximating the coalescent with recombination. *Philos Trans R Soc B Biol Sci* **360**: 1387–1393.

Meisner J, Albrechtsen A (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics* **210**: 719–731.

Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**: 786–792.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, *et al.* (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**:222-6.

Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G (2006). Signatures of demographic history and natural selection in the human major histocompatibility complex loci.

Genetics **173**: 2121–42.

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, *et al.* (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc Natl Acad Sci* **109**: E2382-90.

Mondal M, Bertrampetit J, Lao O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nat Commun* **10**(1): 246.

Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, *et al.* (2018). Early human dispersals within the Americas. *Science* **362**: eaav2621.

Morin PA, Luikart G, Wayne RK, the SNP workshop group (2004). SNPs in ecology, evolution and conservation. *Trends Ecol Evol* **19**: 208–216.

Morton BR, Dar V-N, Wright SI (2009). Analysis of Site Frequency Spectra from Arabidopsis with Context-Dependent Corrections for Ancestral Misinference. *Plant Physiol* **149**: 616–24.

Mullis KB, Faloona FA (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. In: *Methods in Enzymology*,, pp 335–350.

Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol Ecol* **25**: 1058–1072.

Nater A, Mattle-Greminger MP, Nurcahyo A, Nowak MG, de Manuel M, Desai T, *et al.* (2017). Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Curr Biol* **27**: 3576–3577.

Neuenschwander S, Largiadèr CR, Ray N, Currat M, Vonlanthen P, Excoffier L (2008). Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): Inference under a Bayesian spatially explicit framework. *Mol Ecol* **17**: 757–72.

Nielsen R (2004). Population genetic analysis of ascertained SNP data. *Hum Genomics* **1**: 218.

Nielsen R (2005). Genomic scans for selective sweeps using SNP data. *Genome Res* **15**: 1566–1575.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS One* **7**: e37558.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443–451.

Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, *et al.* (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**: 238–242.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, *et al.* (2012). Ancient Admixture in Human History. *Genetics* **192**: 1065–1093.

Piras IS, De Montis A, Calò CM, Marini M, Atzori M, Corrias L, *et al.* (2012). Genome-wide scan with nearly 700 000 SNPs in two Sardinian sub-populations suggests some regions as candidate targets for positive selection. *Eur J Hum Genet* **20**: 1155–1161.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, *et al.* (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**: 43–9.

Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE (2010). Computational challenges in the analysis of ancient DNA. *Genome Biol* **11**: R47.

Pudlo P, Marin J-MM, Estoup A, Cornuet J-MM, Gautier M, Robert CP (2016). Reliable ABC model choice via random forests. *Bioinformatics* **32**: 859–866.

Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics* **35**: 1720–1728.

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. *Proc Natl Acad Sci* **111**: 7248–53.

Rivollat M, Jeong C, Schiffels S, Küçükkalıpçı İ, Pemonge M-H, Rohrlach AB, *et al.* (2020). Ancient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers. *Sci Adv* **6**: eaaz5344.

Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ (2014). ABC inference of multi-population divergence with admixture from unphased population genomic data. *Mol Ecol* **23**: 4458–71.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* **74**: 5463–5467.

Schiffels S, Durbin R (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**: 919–925.

Schrider DR, Kern AD (2016). Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLoS Genet* **12**(3): e1005928.

Schrider DR, Kern AD (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet* **34**: 301–312.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**: 121–132.

Smith ML, Ruffley M, Espíndola A, Tank DC, Sullivan J, Carstens BC (2017). Demographic model selection using random forests and the site frequency spectrum. *Mol Ecol* **26**: 4562–4573.

Soraggi S, Wiuf C, Albrechtsen A (2018). Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. *G3; Genes|Genomes|Genetics* **8**: 551–566.

Supple MA, Shapiro B (2018). Conservation of biodiversity in the genomics era. *Genome*

Biol **19**: 131.

Tassi F, Ghirotto S, Mezzavilla M, Vilaça ST, De Santi L, Barbujani G (2015). Early modern human dispersal from Africa: Genomic evidence for multiple waves of migration. *Investig Genet* **6**: 6–13.

Terhorst J, Kamm JA, Song YS (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**: 303–309.

Terhorst J, Song YS (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc Natl Acad Sci* **112**: 7677–82.

The Potato Genome Sequencing Consortium T (2011). Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189–195.

Torada L, Lorenzon L, Beddis A, Isildak U, Pattini L, Mathieson S, Fumagalli M (2019). ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics* **20**: 337.

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

Wakeley J, Hey J (1997). Estimating ancestral population parameters. *Genetics* **145**: 847–855.

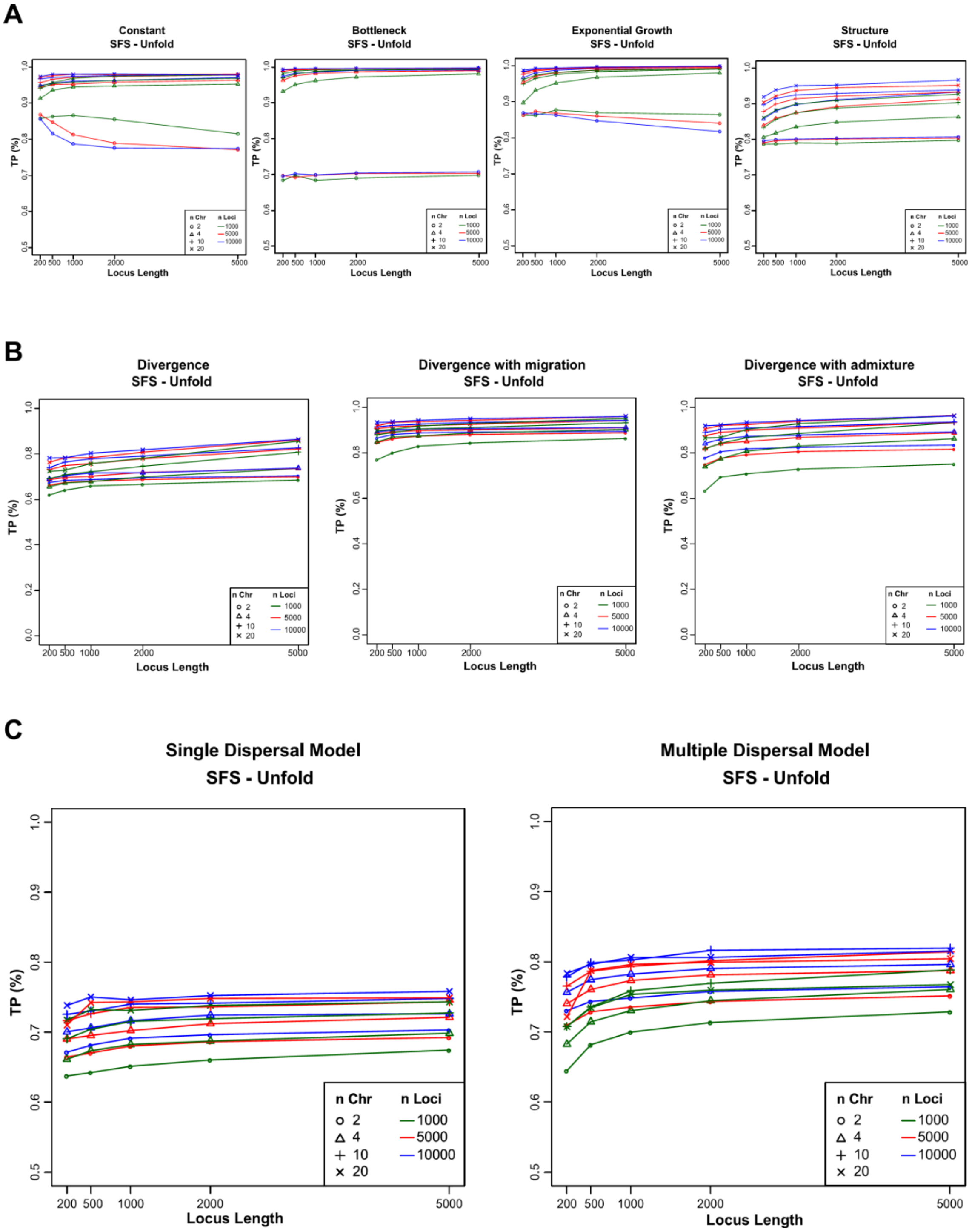
Wright S (1931). Evolution in Mendelian Populations. *Genetics* **16**: 97–159.

Xu C, Wu K, Zhang J-G, Shen H, Deng H-W (2017). Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study. *Genet Epidemiol* **41**: 187–197.

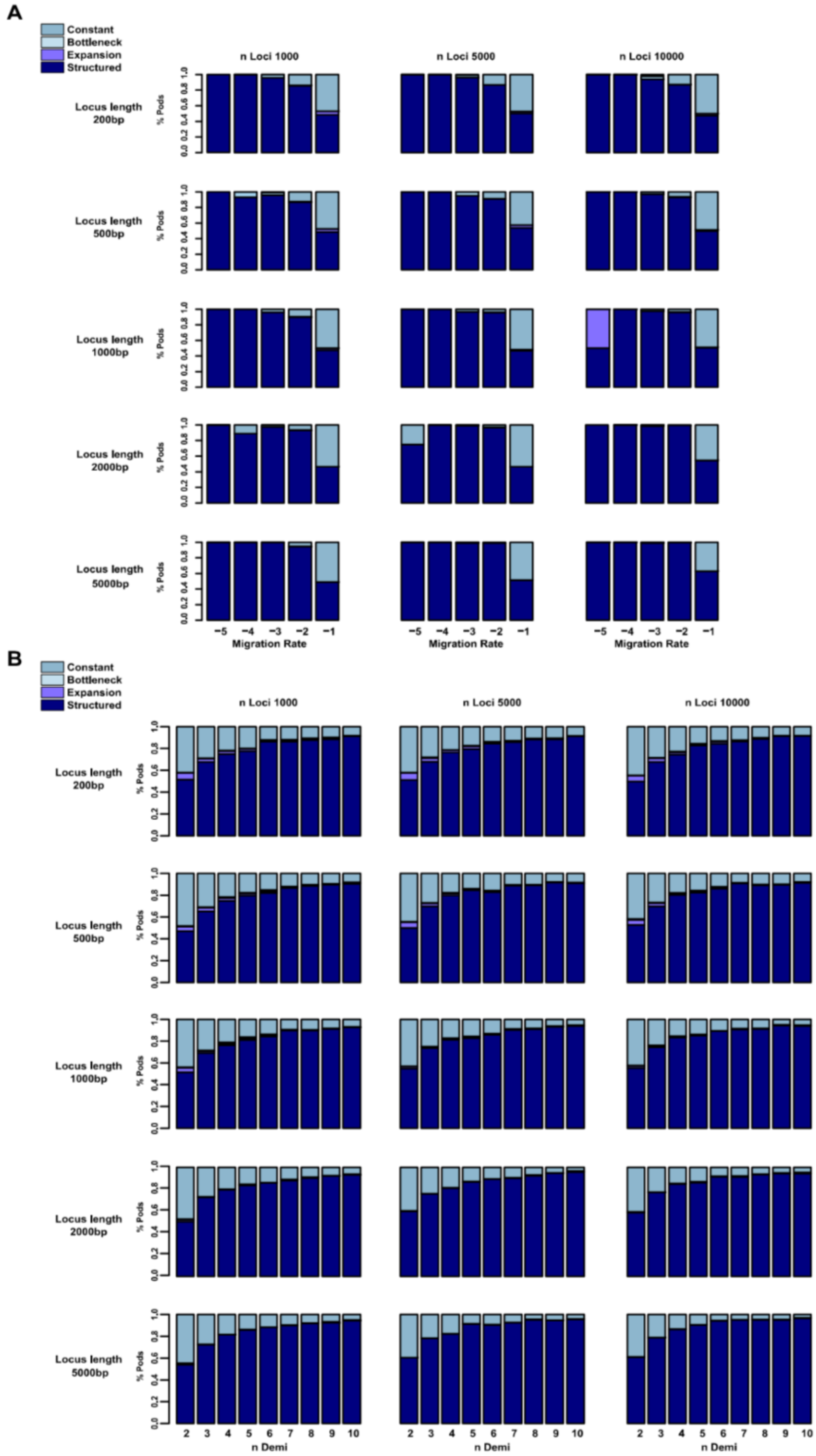
Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Talenti A (2019). A primer on deep learning in genomics. *Nat Genet* **51**: 12–18.

8. Supplementary Materials

Supplementary Figure 4.1. Proportion of True Positives for (A) the one-population models, (B) the two-population models and (C) the multi-population models summarized through the *unfolded SFS*. The plots have the same features of Figure 4.1.



Supplementary Figure 4.2. Proportion of True positives for the one-population structured model as a function of the migration rate (A) and the number of demes considered (B). (A) Each plot represents the proportion of pods from the structured model assigned to each of the four one-population models with the migration rates among demes in the structured model constrained at ranges of increasing values (from 1*10-5 to 1*10-1). All the plots consider two chromosomes and a specific combination of locus length and number of loci; the number of demes in the structured model is fixed to four. In general, the TP rate (in dark blue) decreases as increasing the migration rate among demes, with the constant model erroneously recognize as the true model for higher migration rates. (B) Proportion of pods from the structured model assigned to each of the four one-population models as a function of the number of demes (from 2 to 10). The TP rate increase with the number of demes, regardless of the level of migration among demes.



Supplementary Table 5.1. Accuracy of the estimated parameters of the Constant model assessed by 1,000 pods. Combinations of experimental parameters considering 1,000 loci. The number of chromosomes is indicated with *nc*, whereas *ll* indicates the locus length.

		Coverage 1x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.955	0.033	1545.833	0.989	0.644
	ll1000		0.994	0.002	766.837	1.000	0.570
nc20	ll200		0.984	0.016	1308.498	0.995	0.603
	ll1000		0.999	0.004	609.009	1.000	0.554
nc50	ll200		0.990	0.036	1713.427	0.983	0.589
	ll1000		1.000	0.005	503.096	1.000	0.615

		Coverage 2x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.997	0.033	1499.297	0.985	0.573
	ll1000		0.996	0.002	617.687	1.000	0.594
nc20	ll200		0.992	0.017	1247.473	0.992	0.609
	ll1000		0.998	0.001	511.822	1.000	0.600
nc50	ll200		0.966	0.079	1175.160	0.979	0.755
	ll1000		0.999	0.007	472.530	0.999	0.618

		Coverage 5x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	1.000	0.013	1211.403	0.996	0.533
	ll1000		1.000	0.000	518.848	1.000	0.663
nc20	ll200		1.006	0.011	1167.460	0.994	0.531
	ll1000		0.996	0.002	548.235	1.000	0.609
nc50	ll200		1.000	0.014	775.294	0.995	0.667
	ll1000		0.999	0.000	440.270	1.000	0.619

		Coverage 30x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.998	0.001	1188.937	1.000	0.508
	ll1000		0.995	0.002	549.645	1.000	0.558
nc20	ll200		0.994	0.001	920.206	1.000	0.561
	ll1000		0.999	0.001	499.945	1.000	0.577
nc50	ll200		1.000	0.004	700.686	1.000	0.620
	ll1000		1.000	0.000	382.375	1.000	0.591

Supplementary Table 5.2. Accuracy of the estimated parameters of the Constant model assessed by 1,000 pods. Combinations of experimental parameters considering 5,000 loci.

		Coverage 1x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	1.000	0.007	590.230	1.000	0.565
	ll1000		0.999	0.001	334.277	1.000	0.570
nc20	ll200		0.996	0.005	622.665	0.999	0.590
	ll1000		0.998	0.002	285.148	1.000	0.554
nc50	ll200		0.989	0.053	805.546	0.979	0.571
	ll1000		0.998	0.010	246.237	0.997	0.669

		Coverage 2x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.996	0.008	524.767	0.999	0.605
	ll1000		1.000	0.001	307.994	1.000	0.610
nc20	ll200		0.999	0.012	482.244	0.998	0.647
	ll1000		0.999	0.000	268.772	1.000	0.594
nc50	ll200		0.995	0.020	455.324	0.994	0.670
	ll1000		1.000	0.002	229.755	1.000	0.575

		Coverage 5x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	1.000	0.003	462.882	1.000	0.595
	ll1000		0.998	0.000	253.277	1.000	0.580
nc20	ll200		1.001	0.003	417.779	1.000	0.592
	ll1000		1.000	0.001	229.886	1.000	0.638
nc50	ll200		0.998	0.012	450.187	0.996	0.596
	ll1000		1.000	0.000	208.397	1.000	0.597

		Coverage 30x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.999	0.002	408.336	1.000	0.643
	ll1000		0.999	0.001	267.408	1.000	0.606
nc20	ll200		1.000	0.001	379.749	1.000	0.555
	ll1000		0.999	0.000	228.323	1.000	0.646
nc50	ll200		0.999	0.002	316.463	1.000	0.555
	ll1000		0.998	0.000	215.947	1.000	0.625

Supplementary Table 5.3. Accuracy of the estimated parameters of the Bottleneck model assessed by 1,000 pods. Combinations of experimental parameters considering 1,000 loci.

		Coverage 1x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.653	0.168	868.519	0.932	0.552
		T	0.655	0.277	3695.004	0.875	0.538
		NaBott	0.597	0.059	14885.458	0.976	0.515
	ll1000	*N1*	0.795	0.100	728.413	0.968	0.495
		T	0.861	0.149	2522.540	0.939	0.521
		NaBott	0.819	0.021	9948.504	0.994	0.626
nc20	ll200	*N1*	0.660	0.156	853.297	0.941	0.527
		T	0.703	0.279	3707.116	0.875	0.485
		NaBott	0.622	0.062	14135.565	0.984	0.537
	ll1000	*N1*	0.683	0.175	902.776	0.926	0.513
		T	0.777	0.216	3153.159	0.903	0.522
		NaBott	0.769	0.027	10732.898	0.984	0.561
nc50	ll200	*N1*	0.565	0.218	980.206	0.894	0.529
		T	0.524	0.470	4187.524	0.810	0.503
		NaBott	0.559	0.070	17791.336	0.983	0.531
	ll1000	*N1*	0.781	0.118	740.648	0.959	0.493
		T	0.911	0.175	2353.816	0.934	0.547
		NaBott	0.800	0.041	10184.577	0.987	0.545

		Coverage 2x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.528	0.213	1024.312	0.893	0.491
		T	0.632	0.381	4088.446	0.822	0.464
		NaBott	0.562	0.073	15641.259	0.969	0.533
	ll1000	*N1*	0.765	0.126	749.295	0.954	0.506
		T	0.846	0.159	2152.658	0.941	0.510
		NaBott	0.780	0.019	9889.567	0.991	0.599
nc20	ll200	*N1*	0.630	0.187	931.103	0.929	0.469
		T	0.733	0.277	3697.900	0.872	0.492
		NaBott	0.719	0.043	13001.665	0.987	0.559
	ll1000	*N1*	0.771	0.106	709.390	0.964	0.531
		T	0.869	0.135	2089.638	0.947	0.567
		NaBott	0.831	0.030	9436.385	0.988	0.601
nc50	ll200	*N1*	0.606	0.202	950.933	0.915	0.542
		T	0.624	0.318	3898.129	0.856	0.513
		NaBott	0.531	0.064	16497.826	0.979	0.526
	ll1000	*N1*	0.740	0.152	823.798	0.940	0.463
		T	0.772	0.189	2941.760	0.923	0.471
		NaBott	0.822	0.022	10573.725	0.992	0.537

		Coverage 5x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.670	0.209	975.256	0.905	0.521
		T	0.688	0.298	3861.590	0.842	0.486
		NaBott	0.582	0.073	15124.538	0.976	0.533
	ll1000	*N1*	0.808	0.101	688.596	0.959	0.514
		T	0.892	0.135	2047.077	0.942	0.520
		NaBott	0.801	0.026	9778.200	0.991	0.602
nc20	ll200	*N1*	0.611	0.200	964.997	0.915	0.472
		T	0.694	0.276	3786.642	0.872	0.505
		NaBott	0.596	0.053	15257.369	0.983	0.493
	ll1000	*N1*	0.821	0.084	626.388	0.973	0.532
		T	0.907	0.109	1935.173	0.958	0.546
		NaBott	0.818	0.021	9603.793	0.988	0.625
nc50	ll200	*N1*	0.643	0.187	918.610	0.931	0.487
		T	0.668	0.282	3617.178	0.878	0.500
		NaBott	0.681	0.040	12577.801	0.992	0.562
	ll1000	*N1*	0.805	0.098	702.097	0.963	0.509
		T	0.878	0.140	1981.857	0.939	0.503
		NaBott	0.832	0.010	10151.471	0.994	0.566

		Coverage 30x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.709	0.177	866.735	0.924	0.540
		T	0.717	0.260	3334.580	0.887	0.536
		NaBott	0.709	0.048	12218.474	0.984	0.542
	ll1000	*N1*	0.801	0.072	613.803	0.974	0.572
		T	0.929	0.099	1884.530	0.961	0.584
		NaBott	0.793	0.018	9409.010	0.995	0.661
nc20	ll200	*N1*	0.676	0.162	847.922	0.936	0.525
		T	0.745	0.257	3329.713	0.894	0.514
		NaBott	0.688	0.043	12887.452	0.990	0.545
	ll1000	*N1*	0.815	0.083	656.112	0.969	0.522
		T	0.909	0.096	1783.618	0.958	0.555
		NaBott	0.870	0.015	8601.397	0.992	0.640
nc50	ll200	*N1*	0.732	0.139	817.584	0.948	0.521
		T	0.710	0.215	3320.334	0.894	0.534
		NaBott	0.690	0.037	13136.178	0.984	0.546
	ll1000	*N1*	0.793	0.097	644.400	0.968	0.536
		T	0.877	0.132	1907.834	0.956	0.546
		NaBott	0.774	0.015	9593.672	0.994	0.617

Supplementary Table 5.4. Accuracy of the estimated parameters of the Bottleneck model assessed by 1,000 pods. Combinations of experimental parameters considering 5,000 loci.

		Coverage 1x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.689	0.200	886.946	0.922	0.539
		T	0.759	0.265	3207.655	0.887	0.516
		NaBott	0.790	0.034	10771.029	0.990	0.562
	ll1000	*N1*	0.799	0.112	708.054	0.965	0.556
		T	0.882	0.136	1873.723	0.949	0.567
		NaBott	0.860	0.016	7959.279	0.991	0.633
nc20	ll200	*N1*	0.723	0.135	809.505	0.948	0.501
		T	0.761	0.198	3042.307	0.911	0.517
		NaBott	0.722	0.030	11906.801	0.988	0.578
	ll1000	*N1*	0.757	0.132	770.193	0.946	0.535
		T	0.854	0.161	2400.323	0.927	0.532
		NaBott	0.843	0.033	8244.129	0.989	0.633
nc50	ll200	*N1*	0.456	0.232	1017.675	0.887	0.515
		T	0.534	0.375	4169.887	0.830	0.523
		NaBott	0.424	0.071	15794.643	0.992	0.551
	ll1000	*N1*	0.778	0.095	690.049	0.966	0.543
		T	0.875	0.120	2090.666	0.948	0.553
		NaBott	0.872	0.014	8264.823	0.993	0.577

		Coverage 2x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.682	0.185	892.240	0.929	0.487
		T	0.771	0.252	3204.015	0.888	0.501
		NaBott	0.779	0.031	10773.951	0.989	0.533
	ll1000	*N1*	0.799	0.082	640.848	0.964	0.541
		T	0.895	0.098	1517.488	0.956	0.565
		NaBott	0.900	0.011	7402.099	0.995	0.663
nc20	ll200	*N1*	0.717	0.157	836.943	0.940	0.524
		T	0.756	0.198	2866.225	0.914	0.525
		NaBott	0.791	0.017	10213.974	0.993	0.582
	ll1000	*N1*	0.742	0.122	724.023	0.960	0.560
		T	0.882	0.147	2010.084	0.951	0.567
		NaBott	0.867	0.021	7348.372	0.992	0.640
nc50	ll200	*N1*	0.629	0.247	977.828	0.909	0.502
		T	0.650	0.329	3812.149	0.861	0.490
		NaBott	0.665	0.047	11944.613	0.989	0.580
	ll1000	*N1*	0.728	0.151	811.795	0.942	0.534
		T	0.810	0.191	2696.072	0.926	0.531
		NaBott	0.889	0.012	7586.827	0.996	0.582

		Coverage 5x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.747	0.103	774.891	0.959	0.510

		T	0.833	0.164	2884.581	0.928	0.508
		NaBott	0.779	0.035	11222.058	0.987	0.554
	ll1000	*N1*	0.767	0.149	746.970	0.951	0.580
		T	0.847	0.165	1962.224	0.934	0.589
		NaBott	0.905	0.016	6977.608	0.994	0.700
nc20	ll200	*N1*	0.746	0.163	818.798	0.943	0.526
		T	0.759	0.220	2789.918	0.908	0.504
		NaBott	0.794	0.026	10171.610	0.988	0.591
	ll1000	*N1*	0.767	0.107	688.589	0.953	0.575
		T	0.877	0.121	1809.353	0.949	0.569
		NaBott	0.882	0.019	7305.038	0.994	0.684
nc50	ll200	*N1*	0.643	0.179	901.868	0.926	0.514
		T	0.735	0.244	3610.010	0.888	0.500
		NaBott	0.730	0.047	11336.572	0.985	0.548
	ll1000	*N1*	0.826	0.079	584.968	0.973	0.538
		T	0.938	0.089	1359.596	0.965	0.553
		NaBott	0.890	0.020	7389.880	0.988	0.666

		Coverage 30x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.698	0.151	809.926	0.942	0.510
		T	0.801	0.191	2776.364	0.921	0.520
		NaBott	0.782	0.022	9809.908	0.994	0.571
	ll1000	*N1*	0.842	0.073	583.787	0.972	0.571
		T	0.951	0.085	1212.466	0.968	0.600
		NaBott	0.895	0.023	6657.717	0.993	0.710
nc20	ll200	*N1*	0.751	0.153	832.845	0.943	0.518
		T	0.857	0.205	2786.045	0.913	0.531
		NaBott	0.796	0.034	9999.243	0.990	0.572
	ll1000	*N1*	0.812	0.052	549.215	0.976	0.593
		T	0.945	0.066	1167.277	0.971	0.614
		NaBott	0.892	0.011	7109.018	0.994	0.691
nc50	ll200	*N1*	0.757	0.138	783.094	0.949	0.535
		T	0.818	0.202	2733.306	0.916	0.526
		NaBott	0.811	0.031	9703.053	0.994	0.594
	ll1000	*N1*	0.871	0.068	537.977	0.979	0.578
		T	0.926	0.085	1259.342	0.973	0.578
		NaBott	0.867	0.026	8417.561	0.986	0.696

Supplementary Table 5.5. Accuracy of the estimated parameters of the Exponential Growth model assessed by 1,000 pods. Combinations of experimental parameters considering 1,000 loci.

		Coverage 1x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.025	0.156	21116.612	0.927	0.510
		T	0.469	0.935	3795.697	0.785	0.559
		NaExp	0.238	0.251	1112.305	0.856	0.560
	ll1000	*N1*	0.063	0.152	21461.613	0.928	0.466
		T	0.733	0.307	3254.462	0.872	0.515
		NaExp	0.448	0.276	1005.609	0.860	0.512
nc20	ll200	*N1*	0.064	0.124	21139.963	0.935	0.511
		T	0.715	0.442	3284.099	0.859	0.550
		NaExp	0.322	0.263	1135.775	0.855	0.521
	ll1000	*N1*	0.237	0.107	21450.154	0.946	0.466
		T	0.804	0.313	3074.880	0.889	0.504
		NaExp	0.507	0.194	1019.320	0.875	0.496
nc50	ll200	*N1*	0.244	0.107	19227.428	0.970	0.501
		T	0.646	0.268	3729.226	0.835	0.558
		NaExp	0.137	0.326	1189.785	0.812	0.514
	ll1000	*N1*	0.182	0.125	21122.796	0.952	0.497
		T	0.799	0.260	2944.336	0.896	0.475
		NaExp	0.298	0.280	1154.632	0.828	0.488

		Coverage 2x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.047	0.139	21084.219	0.926	0.506
		T	0.643	0.483	3944.169	0.818	0.509
		NaExp	0.312	0.280	1117.932	0.826	0.509
	ll1000	*N1*	0.124	0.148	22106.251	0.929	0.484
		T	0.804	0.303	3144.395	0.900	0.496
		NaExp	0.497	0.210	1025.683	0.886	0.523
nc20	ll200	*N1*	0.079	0.140	20489.851	0.939	0.517
		T	0.743	0.323	3321.596	0.880	0.524
		NaExp	0.295	0.300	1118.425	0.839	0.514
	ll1000	*N1*	0.126	0.130	21538.449	0.945	0.480
		T	0.771	0.279	3031.834	0.901	0.489
		NaExp	0.329	0.284	1106.004	0.831	0.499
nc50	ll200	*N1*	0.102	0.119	20480.930	0.950	0.511
		T	0.587	0.504	3130.020	0.856	0.615
		NaExp	0.123	0.337	1175.402	0.826	0.524
	ll1000	*N1*	0.221	0.093	20576.524	0.958	0.489
		T	0.793	0.192	2897.855	0.911	0.549
		NaExp	0.236	0.315	1145.096	0.828	0.533

		Coverage 5x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.086	0.151	21265.463	0.939	0.491

		T	0.727	0.444	3688.859	0.823	0.494
		NaExp	0.300	0.323	1156.656	0.816	0.488
	ll1000	*N1*	0.147	0.140	21139.987	0.925	0.516
		T	0.765	0.369	2946.913	0.879	0.511
		NaExp	0.510	0.189	999.537	0.877	0.531
nc20	ll200	*N1*	0.184	0.088	20573.404	0.959	0.497
		T	0.775	0.292	3324.238	0.862	0.481
		NaExp	0.311	0.287	1156.554	0.823	0.494
	ll1000	*N1*	0.210	0.108	21900.028	0.936	0.495
		T	0.803	0.265	2960.026	0.888	0.494
		NaExp	0.396	0.277	1142.955	0.836	0.476
nc50	ll200	*N1*	0.199	0.081	20230.472	0.963	0.520
		T	0.793	0.247	3004.159	0.894	0.515
		NaExp	0.197	0.334	1219.434	0.817	0.494
	ll1000	*N1*	0.319	0.085	19967.123	0.962	0.503
		T	0.795	0.177	2998.871	0.910	0.487
		NaExp	0.339	0.285	1214.465	0.803	0.491

		Coverage 30x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.091	0.129	21239.587	0.940	0.501
		T	0.710	0.332	3222.802	0.873	0.540
		NaExp	0.298	0.307	1155.671	0.834	0.494
	ll1000	*N1*	0.136	0.117	21817.818	0.938	0.481
		T	0.782	0.306	2973.979	0.891	0.503
		NaExp	0.441	0.261	1067.447	0.847	0.470
nc20	ll200	*N1*	0.141	0.110	20669.920	0.959	0.491
		T	0.757	0.323	3062.532	0.880	0.517
		NaExp	0.257	0.288	1179.409	0.833	0.508
	ll1000	*N1*	0.176	0.124	21391.027	0.947	0.493
		T	0.789	0.166	2961.960	0.914	0.503
		NaExp	0.332	0.304	1182.575	0.829	0.502
nc50	ll200	*N1*	0.300	0.103	21308.067	0.951	0.488
		T	0.805	0.161	3066.184	0.907	0.491
		NaExp	0.270	0.311	1262.277	0.792	0.489
	ll1000	*N1*	0.272	0.090	20182.885	0.962	0.514
		T	0.779	0.142	2911.160	0.917	0.505
		NaExp	0.280	0.310	1168.234	0.822	0.538

Supplementary Table 5.6. Accuracy of the estimated parameters of the Exponential Growth model assessed by 1,000 pods. Combinations of experimental parameters considering 5,000 loci.

		Coverage 1x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.069	0.127	21227.819	0.934	0.505
		T	0.743	0.499	3390.393	0.843	0.524
		NaExp	0.399	0.271	1075.252	0.857	0.477
	ll1000	*N1*	0.085	0.140	21437.573	0.936	0.506
		T	0.816	0.251	2908.127	0.901	0.520
		NaExp	0.577	0.210	911.310	0.888	0.537
nc20	ll200	*N1*	0.193	0.108	18917.859	0.970	0.550
		T	0.695	0.209	2865.130	0.915	0.581
		NaExp	0.241	0.264	1038.725	0.864	0.549
	ll1000	*N1*	0.226	0.126	20288.809	0.944	0.508
		T	0.817	0.366	2940.792	0.876	0.520
		NaExp	0.573	0.164	971.513	0.878	0.509
nc50	ll200	*N1*	0.187	0.107	19408.133	0.964	0.525
		T	0.677	0.346	3207.315	0.859	0.558
		NaExp	0.193	0.301	1171.286	0.825	0.530
	ll1000	*N1*	0.318	0.100	17965.959	0.967	0.504
		T	0.812	0.254	2823.050	0.889	0.520
		NaExp	0.323	0.306	1105.199	0.851	0.512

		Coverage 2x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.075	0.112	21122.708	0.943	0.500
		T	0.730	0.454	3245.578	0.869	0.520
		NaExp	0.335	0.316	1084.128	0.846	0.505
	ll1000	*N1*	0.125	0.121	21941.546	0.946	0.482
		T	0.819	0.247	2904.739	0.889	0.483
		NaExp	0.630	0.211	938.403	0.884	0.502
nc20	ll200	*N1*	0.129	0.142	21867.864	0.926	0.496
		T	0.785	0.307	2910.053	0.889	0.488
		NaExp	0.323	0.288	1125.399	0.839	0.482
	ll1000	*N1*	0.454	0.083	19287.059	0.953	0.487
		T	0.854	0.091	2736.813	0.939	0.487
		NaExp	0.685	0.154	873.470	0.904	0.512
nc50	ll200	*N1*	0.161	0.106	19833.435	0.959	0.521
		T	0.657	0.424	2949.957	0.885	0.608
		NaExp	0.183	0.322	1163.026	0.834	0.532
	ll1000	*N1*	0.542	0.066	17767.134	0.967	0.488
		T	0.836	0.115	2823.295	0.939	0.485
		NaExp	0.465	0.255	1104.797	0.845	0.464

		Coverage 5x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.128	0.139	21602.752	0.926	0.489

		T	0.764	0.336	3208.567	0.872	0.511
		NaExp	0.371	0.306	1123.043	0.833	0.492
	ll1000	*N1*	0.213	0.131	22100.502	0.928	0.500
		T	0.844	0.190	2744.433	0.895	0.494
		NaExp	0.651	0.168	894.740	0.893	0.505
nc20	ll200	*N1*	0.140	0.122	20460.829	0.936	0.519
		T	0.789	0.202	2970.885	0.913	0.508
		NaExp	0.290	0.283	1131.855	0.851	0.521
	ll1000	*N1*	0.391	0.082	20204.232	0.945	0.483
		T	0.829	0.091	2752.506	0.940	0.468
		NaExp	0.616	0.203	936.682	0.880	0.493
nc50	ll200	*N1*	0.357	0.085	20097.029	0.954	0.476
		T	0.825	0.137	2872.786	0.916	0.513
		NaExp	0.309	0.278	1137.105	0.839	0.517
	ll1000	*N1*	0.488	0.060	18362.734	0.965	0.486
		T	0.835	0.118	2823.749	0.942	0.486
		NaExp	0.506	0.266	1055.913	0.827	0.509

		Coverage 30x					
		Parameter	**R^2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.102	0.120	21027.171	0.954	0.490
		T	0.770	0.298	3008.685	0.893	0.525
		NaExp	0.315	0.269	1113.259	0.855	0.517
	ll1000	*N1*	0.184	0.123	21545.419	0.934	0.507
		T	0.881	0.155	2418.797	0.935	0.531
		NaExp	0.653	0.133	819.388	0.914	0.511
nc20	ll200	*N1*	0.148	0.135	20712.768	0.939	0.506
		T	0.815	0.213	2844.496	0.904	0.514
		NaExp	0.273	0.287	1142.299	0.843	0.522
	ll1000	*N1*	0.212	0.119	21445.524	0.936	0.497
		T	0.820	0.144	2709.066	0.927	0.510
		NaExp	0.490	0.249	1025.747	0.851	0.495
nc50	ll200	*N1*	0.304	0.092	21360.327	0.948	0.461
		T	0.822	0.157	2926.798	0.912	0.490
		NaExp	0.287	0.322	1248.717	0.811	0.501
	ll1000	*N1*	0.294	0.069	20772.366	0.958	0.493
		T	0.821	0.116	2823.599	0.942	0.513
		NaExp	0.378	0.255	1097.335	0.857	0.511

Supplementary Table 5.7. Accuracy of the estimated parameters of the Divergence model assessed by 1,000 pods. Combinations of experimental parameters considering 1,000 loci.

		Coverage 1x					
		Parameter	R2	Bias	RMSE	Factor2	Coverage50%
nc10	ll200	*N1*	0.654	0.161	9155.313	0.904	0.514
		N2	0.726	0.132	8607.355	0.932	0.499
		Nanc	0.953	0.029	2190.376	0.985	0.542
		Tsep	0.874	0.150	2342.224	0.930	0.554
	ll1000	*N1*	0.774	0.110	6909.512	0.962	0.594
		N2	0.783	0.091	6874.929	0.967	0.550
		Nanc	0.982	0.011	1368.092	0.998	0.653
		Tsep	0.888	0.079	1569.978	0.972	0.629
nc20	ll200	*N1*	0.750	0.120	8345.493	0.942	0.502
		N2	0.789	0.123	8075.368	0.947	0.516
		Nanc	0.987	0.028	2400.162	0.984	0.539
		Tsep	0.892	0.070	1940.873	0.972	0.564
	ll1000	*N1*	0.819	0.082	6234.466	0.979	0.592
		N2	0.822	0.077	6398.797	0.969	0.573
		Nanc	0.980	0.028	1451.580	0.993	0.627
		Tsep	0.900	0.065	1414.270	0.977	0.659
nc50	ll200	*N1*	0.748	0.163	9159.231	0.908	0.502
		N2	0.677	0.252	10046.658	0.869	0.493
		Nanc	0.935	0.076	4559.102	0.942	0.495
		Tsep	0.797	0.154	3274.603	0.918	0.477
	ll1000	*N1*	0.765	0.146	7193.454	0.946	0.531
		N2	0.755	0.118	7284.723	0.952	0.535
		Nanc	0.949	0.033	2430.447	0.983	0.522
		Tsep	0.816	0.168	2378.697	0.937	0.608

		Coverage 2x					
		Parameter	R2	Bias	RMSE	Factor2	Coverage50%
nc10	ll200	*N1*	0.771	0.139	8006.414	0.950	0.537
		N2	0.722	0.164	8227.542	0.943	0.504
		Nanc	0.970	0.024	2102.069	0.982	0.556
		Tsep	0.921	0.079	1842.791	0.970	0.551
	ll1000	*N1*	0.848	0.077	6051.145	0.974	0.586
		N2	0.792	0.076	5915.952	0.979	0.586
		Nanc	0.978	0.015	1326.036	0.997	0.686
		Tsep	0.938	0.050	1183.651	0.987	0.634
nc20	ll200	*N1*	0.692	0.181	8445.646	0.928	0.539
		N2	0.708	0.145	8163.903	0.929	0.537
		Nanc	0.986	0.034	2648.142	0.981	0.526
		Tsep	0.905	0.050	1770.296	0.980	0.533
	ll1000	*N1*	0.881	0.074	5561.477	0.985	0.588
		N2	0.813	0.075	6166.650	0.975	0.568
		Nanc	0.979	0.008	1550.086	0.996	0.651
		Tsep	0.933	0.036	1181.370	0.994	0.661

nc50	ll200	*N1*	0.793	0.061	7811.865	0.956	0.537
		N2	0.676	0.213	9380.944	0.888	0.510
		Nanc	0.938	0.081	4182.347	0.959	0.525
		Tsep	0.737	0.170	3203.063	0.906	0.541
	ll1000	*N1*	0.809	0.086	6216.943	0.974	0.574
		N2	0.795	0.105	6435.257	0.971	0.543
		Nanc	0.961	0.014	2313.210	0.992	0.586
		Tsep	0.875	0.068	1819.484	0.980	0.641

		Coverage 5x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.737	0.142	7468.780	0.946	0.524
		N2	0.785	0.085	7241.530	0.957	0.543
		Nanc	0.968	0.013	2432.867	0.992	0.569
		Tsep	0.881	0.073	1695.713	0.969	0.601
	ll1000	*N1*	0.867	0.067	5827.342	0.984	0.583
		N2	0.822	0.058	5438.979	0.984	0.605
		Nanc	0.976	0.002	1318.983	1.000	0.701
		Tsep	0.936	0.023	922.973	0.996	0.734
nc20	ll200	*N1*	0.794	0.087	7304.609	0.961	0.516
		N2	0.801	0.107	7280.575	0.958	0.527
		Nanc	0.950	0.036	2776.566	0.979	0.543
		Tsep	0.922	0.044	1527.070	0.984	0.585
	ll1000	*N1*	0.850	0.061	5486.603	0.982	0.591
		N2	0.843	0.061	5345.512	0.985	0.626
		Nanc	0.974	0.020	1469.626	0.994	0.676
		Tsep	0.944	0.018	880.199	0.997	0.729
nc50	ll200	*N1*	0.761	0.092	7296.399	0.958	0.532
		N2	0.822	0.072	6637.222	0.978	0.525
		Nanc	0.943	0.051	3306.031	0.976	0.549
		Tsep	0.907	0.036	1731.411	0.992	0.557
	ll1000	*N1*	0.849	0.035	4603.966	0.989	0.630
		N2	0.872	0.034	4497.882	0.993	0.626
		Nanc	0.974	0.006	1760.389	0.998	0.680
		Tsep	0.969	0.012	857.029	0.998	0.719

		Coverage 30x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.739	0.158	7842.213	0.945	0.554
		N2	0.765	0.096	7788.829	0.954	0.543
		Nanc	0.970	0.025	2148.120	0.993	0.568
		Tsep	0.911	0.057	1686.612	0.972	0.578
	ll1000	*N1*	0.835	0.039	5170.745	0.990	0.617
		N2	0.852	0.059	5825.852	0.982	0.601
		Nanc	0.975	0.004	1339.137	0.998	0.709
		Tsep	0.939	0.030	881.715	0.996	0.742
nc20	ll200	*N1*	0.812	0.065	7073.019	0.972	0.543
		N2	0.775	0.113	7329.004	0.958	0.543

		Nanc	0.976	0.035	2514.890	0.988	0.553
		Tsep	0.922	0.050	1340.213	0.986	0.612
	ll1000	*N1*	0.843	0.044	4985.146	0.989	0.610
		N2	0.832	0.047	5365.946	0.988	0.622
		Nanc	0.984	0.007	1505.727	0.998	0.693
		Tsep	0.962	0.014	769.479	1.000	0.790
nc50	ll200	*N1*	0.813	0.062	6301.582	0.976	0.549
		N2	0.827	0.064	6040.040	0.975	0.584
		Nanc	0.954	0.018	2938.266	0.990	0.558
		Tsep	0.922	0.027	1359.340	0.995	0.630
	ll1000	*N1*	0.873	0.066	5139.122	0.984	0.622
		N2	0.842	0.060	5303.823	0.982	0.626
		Nanc	0.967	0.025	1981.573	0.988	0.635
		Tsep	0.923	0.025	1207.228	0.995	0.696

Supplementary Table 5.8. Accuracy of the estimated parameters of the Divergence model assessed by 1,000 pods. Combinations of experimental parameters considering 5,000 loci.

		Coverage 1x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.807	0.071	6178.998	0.968	0.541
		N2	0.753	0.119	7303.880	0.943	0.549
		Nanc	0.993	0.007	1287.755	0.997	0.571
		Tsep	0.901	0.072	1762.754	0.973	0.594
	ll1000	*N1*	0.857	0.045	5122.714	0.992	0.656
		N2	0.854	0.047	4828.987	0.991	0.621
		Nanc	0.991	0.004	826.891	1.000	0.706
		Tsep	0.921	0.048	1071.065	0.993	0.724
nc20	ll200	*N1*	0.792	0.138	6912.684	0.954	0.572
		N2	0.837	0.057	5927.945	0.980	0.551
		Nanc	0.992	0.017	1503.813	0.993	0.626
		Tsep	0.897	0.062	1634.868	0.979	0.591
	ll1000	*N1*	0.889	0.042	4605.538	0.989	0.650
		N2	0.847	0.048	4801.478	0.990	0.642
		Nanc	0.993	0.006	926.783	0.998	0.691
		Tsep	0.946	0.024	781.951	0.997	0.695
nc50	ll200	*N1*	0.702	0.172	8230.381	0.919	0.517
		N2	0.826	0.088	6455.792	0.953	0.564
		Nanc	0.981	0.015	2692.224	0.977	0.541
		Tsep	0.811	0.097	2580.004	0.941	0.541
	ll1000	*N1*	0.849	0.062	4687.746	0.983	0.617
		N2	0.877	0.062	4378.261	0.984	0.643
		Nanc	0.977	0.009	1335.554	0.995	0.671
		Tsep	0.911	0.060	1522.001	0.986	0.668

Coverage 2x					
Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**

nc10	ll200	*N1*	0.840	0.076	5928.333	0.973	0.576
		N2	0.783	0.097	6285.699	0.965	0.582
		Nanc	0.987	0.014	1306.460	0.997	0.603
		Tsep	0.918	0.068	1395.310	0.980	0.647
	ll1000	*N1*	0.896	0.052	4385.383	0.981	0.646
		N2	0.890	0.049	4507.875	0.987	0.670
		Nanc	0.987	0.001	745.340	0.999	0.766
		Tsep	0.960	0.037	716.737	0.992	0.787
nc20	ll200	*N1*	0.809	0.068	6364.359	0.974	0.567
		N2	0.803	0.078	6314.701	0.971	0.572
		Nanc	0.982	0.006	1668.263	0.997	0.582
		Tsep	0.919	0.046	1223.175	0.989	0.653
	ll1000	*N1*	0.866	0.042	4630.029	0.987	0.670
		N2	0.884	0.050	4982.631	0.987	0.646
		Nanc	0.989	0.008	988.891	0.997	0.703
		Tsep	0.968	0.026	766.731	0.998	0.749
nc50	ll200	*N1*	0.850	0.061	6320.253	0.980	0.560
		N2	0.821	0.067	6513.861	0.979	0.546
		Nanc	0.966	0.034	2686.499	0.978	0.539
		Tsep	0.938	0.032	1555.236	0.994	0.603
	ll1000	*N1*	0.837	0.042	4607.084	0.988	0.642
		N2	0.854	0.046	4621.131	0.989	0.606
		Nanc	0.976	0.027	1427.035	0.993	0.654
		Tsep	0.927	0.035	1199.911	0.992	0.690

		Coverage 5x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.867	0.054	5591.553	0.981	0.600
		N2	0.803	0.096	6650.584	0.954	0.577
		Nanc	0.988	0.001	1289.186	0.999	0.648
		Tsep	0.943	0.043	1039.058	0.989	0.682
	ll1000	*N1*	0.875	0.034	4574.602	0.993	0.639
		N2	0.910	0.029	3978.111	0.996	0.665
		Nanc	0.991	0.002	789.572	0.998	0.742
		Tsep	0.964	0.015	569.830	1.000	0.818
nc20	ll200	*N1*	0.856	0.068	5841.968	0.984	0.537
		N2	0.828	0.071	5930.085	0.979	0.552
		Nanc	0.972	0.004	1723.490	1.000	0.592
		Tsep	0.959	0.030	981.257	0.996	0.633
	ll1000	*N1*	0.896	0.029	4100.869	0.993	0.671
		N2	0.883	0.035	4308.574	0.990	0.642
		Nanc	0.987	0.008	988.304	0.998	0.717
		Tsep	0.959	0.013	615.090	0.999	0.807
nc50	ll200	*N1*	0.857	0.054	5796.990	0.979	0.595
		N2	0.859	0.059	5646.839	0.984	0.566
		Nanc	0.973	0.021	2170.771	0.993	0.584
		Tsep	0.951	0.017	1162.794	0.999	0.655
	ll1000	*N1*	0.881	0.059	4261.002	0.984	0.653

		N2	0.890	0.056	4507.761	0.983	0.641
		Nanc	0.977	0.007	1280.497	0.994	0.690
		Tsep	0.956	0.023	883.002	0.997	0.729

		Coverage 30x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.825	0.090	6416.640	0.966	0.610
		N2	0.857	0.056	5657.955	0.982	0.577
		Nanc	0.981	0.004	1416.384	0.998	0.620
		Tsep	0.951	0.039	1075.387	0.991	0.705
	ll1000	*N1*	0.873	0.045	4352.897	0.988	0.683
		N2	0.885	0.032	4208.989	0.993	0.652
		Nanc	0.987	0.003	824.537	0.998	0.722
		Tsep	0.963	0.029	606.403	0.997	0.837
nc20	ll200	*N1*	0.872	0.045	5595.820	0.987	0.561
		N2	0.862	0.063	5662.925	0.980	0.590
		Nanc	0.977	0.004	1576.824	0.998	0.603
		Tsep	0.945	0.022	908.414	0.994	0.714
	ll1000	*N1*	0.913	0.035	3782.402	0.995	0.668
		N2	0.908	0.034	4001.358	0.992	0.663
		Nanc	0.991	0.004	913.430	0.998	0.736
		Tsep	0.976	0.019	536.396	0.999	0.849
nc50	ll200	*N1*	0.863	0.044	5264.049	0.985	0.604
		N2	0.877	0.040	5228.546	0.990	0.601
		Nanc	0.981	0.006	1856.134	0.997	0.617
		Tsep	0.958	0.013	980.051	0.997	0.703
	ll1000	*N1*	0.895	0.018	3594.999	0.996	0.691
		N2	0.898	0.020	3546.578	0.998	0.681
		Nanc	0.978	0.003	1020.853	1.000	0.734
		Tsep	0.968	0.013	582.188	1.000	0.818

Supplementary Table 5.9. Accuracy of the estimated parameters of the Divergence with migration model assessed by 1,000 pods. Combinations of experimental parameters considering 1,000 loci.

		Coverage 1x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.500	0.461	11412.411	0.794	0.505
		N2	0.461	0.445	11673.160	0.781	0.509
		Nanc	0.932	0.079	4415.267	0.964	0.504
		Tsep	0.274	0.627	5309.609	0.731	0.507
		m12	0.125	7.595	40.622	0.516	0.510
		m21	0.123	3.611	40.730	0.465	0.494
	ll1000	*N1*	0.622	0.356	9862.770	0.829	0.525
		N2	0.538	0.340	10646.689	0.826	0.513
		Nanc	0.979	0.029	2451.871	0.973	0.577
		Tsep	0.369	0.583	4917.960	0.748	0.538

		m12	0.184	1.410	35.288	0.562	0.534
		m21	0.132	3.127	41.252	0.483	0.502
nc20	ll200	*N1*	0.605	0.342	10495.142	0.826	0.484
		N2	0.576	0.256	10938.138	0.823	0.527
		Nanc	0.955	0.062	3516.476	0.969	0.546
		Tsep	0.355	0.695	5147.629	0.725	0.492
		m12	0.180	15.968	38.718	0.510	0.503
		m21	0.160	4.655	39.264	0.475	0.492
	ll1000	*N1*	0.622	0.249	9548.361	0.855	0.509
		N2	0.592	0.375	10354.273	0.842	0.487
		Nanc	0.967	0.068	2019.813	0.975	0.597
		Tsep	0.396	0.672	4684.152	0.769	0.513
		m12	0.208	4.353	38.277	0.582	0.529
		m21	0.153	2.616	37.795	0.521	0.516
nc50	ll200	*N1*	0.714	0.293	9933.899	0.859	0.475
		N2	0.490	0.555	12289.071	0.743	0.488
		Nanc	0.916	0.139	5337.672	0.927	0.488
		Tsep	0.274	0.986	5662.100	0.685	0.477
		m12	0.187	2.639	40.058	0.485	0.471
		m21	0.165	3.192	40.944	0.477	0.482
	ll1000	*N1*	0.675	0.269	9051.272	0.853	0.535
		N2	0.639	0.234	9598.678	0.882	0.523
		Nanc	0.966	0.066	2526.095	0.971	0.587
		Tsep	0.355	0.669	4966.753	0.731	0.520
		m12	0.205	2.606	39.451	0.561	0.522
		m21	0.153	3.609	39.414	0.481	0.491

		Coverage 2x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.618	0.357	10659.104	0.814	0.490
		N2	0.513	0.426	11703.306	0.793	0.518
		Nanc	0.946	0.088	3409.459	0.970	0.521
		Tsep	0.329	0.762	5262.514	0.732	0.501
		m12	0.172	3.183	38.482	0.550	0.536
		m21	0.128	5.198	39.217	0.485	0.505
	ll1000	*N1*	0.647	0.314	9762.591	0.851	0.518
		N2	0.586	0.247	10268.196	0.853	0.520
		Nanc	0.962	0.049	1987.866	0.980	0.601
		Tsep	0.433	0.482	4938.470	0.754	0.513
		m12	0.207	1.432	40.742	0.590	0.532
		m21	0.134	4.083	38.691	0.524	0.523
nc20	ll200	*N1*	0.637	0.317	10269.919	0.832	0.522
		N2	0.540	0.327	11170.528	0.803	0.511
		Nanc	0.932	0.046	3930.110	0.969	0.519
		Tsep	0.304	0.855	5449.880	0.698	0.480
		m12	0.176	2.745	38.719	0.504	0.485
		m21	0.141	1.577	40.170	0.511	0.516
	ll1000	*N1*	0.690	0.250	9128.129	0.861	0.507

		N2	0.562	0.291	10281.309	0.842	0.489
		Nanc	0.956	0.056	1997.829	0.981	0.619
		Tsep	0.402	0.587	4685.042	0.751	0.542
		m12	0.242	2.782	32.390	0.603	0.540
		m21	0.143	3.635	41.568	0.513	0.504
nc50	ll200	*N1*	0.752	0.248	8814.155	0.887	0.502
		N2	0.438	0.437	12448.803	0.765	0.507
		Nanc	0.904	0.180	4931.489	0.934	0.509
		Tsep	0.246	0.934	5691.736	0.715	0.477
		m12	0.166	15.790	42.416	0.498	0.499
		m21	0.162	6.794	38.286	0.494	0.510
	ll1000	*N1*	0.777	0.172	8194.833	0.903	0.538
		N2	0.633	0.278	9615.627	0.870	0.493
		Nanc	0.945	0.136	2660.518	0.960	0.599
		Tsep	0.340	0.712	5013.077	0.759	0.490
		m12	0.234	1.685	35.069	0.596	0.521
		m21	0.175	1.799	38.409	0.571	0.548

		Coverage 5x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.608	0.305	10754.244	0.808	0.503
		N2	0.554	0.361	11440.532	0.787	0.491
		Nanc	0.956	0.056	3325.220	0.973	0.575
		Tsep	0.326	0.877	5322.455	0.697	0.489
		m12	0.157	2.909	36.520	0.516	0.507
		m21	0.143	2.542	39.800	0.511	0.513
	ll1000	*N1*	0.656	0.270	9010.178	0.873	0.488
		N2	0.581	0.246	9780.536	0.890	0.516
		Nanc	0.954	0.086	2648.879	0.971	0.592
		Tsep	0.467	0.575	4864.997	0.749	0.487
		m12	0.230	2.382	33.928	0.604	0.535
		m21	0.156	76.879	39.386	0.502	0.484
nc20	ll200	*N1*	0.719	0.238	9298.476	0.868	0.511
		N2	0.626	0.227	10084.465	0.848	0.508
		Nanc	0.934	0.104	4007.075	0.956	0.525
		Tsep	0.328	0.743	5295.169	0.705	0.468
		m12	0.245	3.023	32.680	0.584	0.542
		m21	0.160	2.974	40.577	0.498	0.498
	ll1000	*N1*	0.721	0.192	8568.671	0.901	0.503
		N2	0.671	0.191	9186.787	0.888	0.498
		Nanc	0.958	0.052	2187.812	0.976	0.595
		Tsep	0.444	0.512	4539.489	0.785	0.552
		m12	0.228	2.535	35.677	0.596	0.511
		m21	0.183	195.083	41.262	0.533	0.526
nc50	ll200	*N1*	0.740	0.214	8796.621	0.891	0.499
		N2	0.729	0.159	9062.182	0.900	0.488
		Nanc	0.912	0.135	3990.508	0.944	0.520
		Tsep	0.281	0.956	5234.892	0.711	0.479

		m12	0.203	2.608	37.965	0.564	0.511
		m21	0.180	7.142	36.108	0.508	0.501
	ll1000	*N1*	0.781	0.163	7593.648	0.910	0.489
		N2	0.714	0.172	8653.066	0.922	0.510
		Nanc	0.936	0.078	2855.482	0.961	0.584
		Tsep	0.346	0.696	4810.749	0.779	0.523
		m12	0.260	2.745	33.679	0.641	0.560
		m21	0.202	2.342	35.991	0.560	0.521

		Coverage 30x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.597	0.368	10510.496	0.817	0.522
		N2	0.586	0.309	10604.448	0.813	0.511
		Nanc	0.953	0.066	3713.520	0.969	0.522
		Tsep	0.296	0.897	5190.186	0.716	0.513
		m12	0.166	2.962	38.768	0.514	0.494
		m21	0.137	14.923	41.897	0.475	0.493
	ll1000	*N1*	0.678	0.220	9135.007	0.882	0.497
		N2	0.618	0.214	9430.574	0.869	0.527
		Nanc	0.969	0.049	1965.322	0.977	0.630
		Tsep	0.409	0.544	4644.812	0.783	0.500
		m12	0.222	2.115	33.930	0.617	0.543
		m21	0.157	9.580	36.890	0.529	0.531
nc20	ll200	*N1*	0.736	0.228	8727.574	0.884	0.512
		N2	0.573	0.289	10578.938	0.848	0.517
		Nanc	0.923	0.114	3606.423	0.947	0.538
		Tsep	0.263	0.950	5299.382	0.717	0.477
		m12	0.220	2.912	33.331	0.580	0.519
		m21	0.165	2.322	37.555	0.506	0.507
	ll1000	*N1*	0.700	0.202	8604.772	0.880	0.496
		N2	0.691	0.171	8584.153	0.916	0.521
		Nanc	0.960	0.053	2110.011	0.983	0.652
		Tsep	0.407	0.574	4585.663	0.760	0.536
		m12	0.235	2.053	34.106	0.619	0.530
		m21	0.181	5.190	38.455	0.531	0.494
nc50	ll200	*N1*	0.765	0.138	8063.151	0.912	0.482
		N2	0.664	0.182	9667.601	0.880	0.487
		Nanc	0.897	0.147	4698.857	0.930	0.545
		Tsep	0.276	1.000	5231.801	0.733	0.505
		m12	0.245	3.198	33.545	0.583	0.534
		m21	0.188	5.971	38.807	0.499	0.492
	ll1000	*N1*	0.794	0.151	7100.099	0.929	0.475
		N2	0.741	0.132	8051.725	0.934	0.519
		Nanc	0.955	0.084	2756.833	0.965	0.613
		Tsep	0.365	0.631	4760.645	0.766	0.528
		m12	0.249	2.100	31.714	0.628	0.539
		m21	0.210	2.429	34.605	0.579	0.544

Supplementary Table 5.10. Accuracy of the estimated parameters of the Divergence with migration model assessed by 1,000 pods. Combinations of experimental parameters considering 5,000 loci.

		Coverage 1x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.659	0.292	9835.659	0.849	0.534
		N2	0.489	0.383	11685.146	0.782	0.486
		Nanc	0.973	0.041	2800.031	0.976	0.556
		Tsep	0.402	0.670	4996.280	0.738	0.516
		m12	0.204	2.162	36.645	0.597	0.554
		m21	0.153	5.423	39.303	0.503	0.505
	ll1000	*N1*	0.652	0.222	9375.775	0.877	0.515
		N2	0.663	0.258	9396.072	0.871	0.527
		Nanc	0.979	0.047	1383.401	0.988	0.659
		Tsep	0.513	0.575	4245.307	0.791	0.518
		m12	0.239	2.926	35.944	0.640	0.540
		m21	0.165	1.956	39.744	0.517	0.507
nc20	ll200	*N1*	0.588	0.327	10518.363	0.838	0.533
		N2	0.641	0.204	9987.678	0.869	0.522
		Nanc	0.971	0.048	2666.986	0.978	0.524
		Tsep	0.436	0.722	5029.804	0.738	0.496
		m12	0.180	2.541	39.894	0.535	0.491
		m21	0.132	3.519	38.910	0.508	0.517
	ll1000	*N1*	0.724	0.193	8262.431	0.917	0.525
		N2	0.635	0.208	8673.650	0.888	0.545
		Nanc	0.985	0.019	1259.140	0.990	0.598
		Tsep	0.551	0.514	4210.267	0.796	0.552
		m12	0.294	2.388	29.948	0.653	0.541
		m21	0.192	2.356	36.484	0.557	0.539
nc50	ll200	*N1*	0.634	0.372	10247.254	0.808	0.532
		N2	0.699	0.258	10416.350	0.819	0.489
		Nanc	0.942	0.081	3315.686	0.954	0.512
		Tsep	0.334	1.043	5415.652	0.704	0.485
		m12	0.169	3.378	40.292	0.498	0.516
		m21	0.153	2.111	38.613	0.492	0.507
	ll1000	*N1*	0.746	0.103	7380.514	0.936	0.533
		N2	0.724	0.127	8399.388	0.922	0.533
		Nanc	0.966	0.048	1614.420	0.981	0.631
		Tsep	0.488	0.467	4369.081	0.809	0.534
		m12	0.260	1.748	33.814	0.643	0.535
		m21	0.204	3.665	35.868	0.551	0.531

		Coverage 2x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.668	0.285	8982.826	0.873	0.538
		N2	0.548	0.332	10507.895	0.839	0.525
		Nanc	0.963	0.040	2720.449	0.975	0.568
		Tsep	0.383	0.701	4951.037	0.722	0.488

		m12	0.198	2.869	34.815	0.603	0.541
		m21	0.143	2.483	40.548	0.505	0.520
	ll1000	*N1*	0.672	0.257	8977.575	0.879	0.502
		N2	0.632	0.203	9747.851	0.872	0.502
		Nanc	0.975	0.017	1114.137	0.994	0.661
		Tsep	0.516	0.470	4179.326	0.798	0.539
		m12	0.267	2.364	31.909	0.654	0.536
		m21	0.163	4.465	37.771	0.531	0.520
nc20	ll200	*N1*	0.691	0.212	9139.610	0.885	0.501
		N2	0.685	0.183	9300.829	0.878	0.537
		Nanc	0.958	0.063	2348.220	0.977	0.544
		Tsep	0.397	0.657	4771.680	0.768	0.515
		m12	0.250	3.360	34.448	0.589	0.523
		m21	0.152	4.472	37.814	0.489	0.491
	ll1000	*N1*	0.703	0.212	8657.876	0.882	0.509
		N2	0.644	0.181	8841.140	0.895	0.528
		Nanc	0.985	0.021	1330.530	0.990	0.641
		Tsep	0.534	0.497	4325.052	0.780	0.524
		m12	0.264	2.785	33.939	0.627	0.503
		m21	0.194	2.251	36.925	0.529	0.489
nc50	ll200	*N1*	0.825	0.236	8125.193	0.895	0.505
		N2	0.701	0.233	9345.243	0.887	0.502
		Nanc	0.936	0.084	3350.919	0.966	0.547
		Tsep	0.338	0.621	5292.303	0.733	0.490
		m12	0.239	2.691	34.784	0.579	0.517
		m21	0.179	13.599	37.181	0.548	0.524
	ll1000	*N1*	0.733	0.175	7535.241	0.923	0.523
		N2	0.731	0.158	7988.443	0.921	0.515
		Nanc	0.969	0.057	1753.408	0.974	0.648
		Tsep	0.472	0.551	4234.751	0.807	0.563
		m12	0.303	2.304	33.166	0.634	0.493
		m21	0.210	24.453	37.708	0.543	0.514

		Coverage 5x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.689	0.277	9196.270	0.876	0.509
		N2	0.579	0.311	10645.089	0.831	0.503
		Nanc	0.974	0.045	2500.958	0.975	0.566
		Tsep	0.421	0.625	4911.774	0.756	0.519
		m12	0.227	2.306	36.768	0.566	0.508
		m21	0.159	3.653	38.352	0.500	0.500
	ll1000	*N1*	0.701	0.253	8582.010	0.877	0.486
		N2	0.662	0.224	8724.177	0.895	0.511
		Nanc	0.980	0.034	1429.570	0.984	0.637
		Tsep	0.537	0.454	4143.882	0.801	0.537
		m12	0.270	13.950	30.847	0.661	0.546
		m21	0.177	4.090	38.814	0.542	0.502
nc20	ll200	*N1*	0.744	0.198	8410.893	0.900	0.508

		N2	0.711	0.166	8911.047	0.907	0.522
		Nanc	0.949	0.061	2721.146	0.966	0.564
		Tsep	0.405	0.616	5004.993	0.750	0.520
		m12	0.248	2.999	36.054	0.619	0.511
		m21	0.197	1.780	36.265	0.541	0.526
	ll1000	*N1*	0.740	0.186	7615.388	0.909	0.556
		N2	0.669	0.175	8939.756	0.918	0.494
		Nanc	0.992	0.030	1442.174	0.985	0.650
		Tsep	0.504	0.413	4394.375	0.810	0.530
		m12	0.281	71.587	33.364	0.697	0.558
		m21	0.157	3.236	43.540	0.575	0.529
nc50	ll200	*N1*	0.767	0.192	8068.101	0.882	0.490
		N2	0.711	0.141	8757.809	0.916	0.516
		Nanc	0.957	0.080	3161.402	0.958	0.557
		Tsep	0.340	0.784	5189.352	0.734	0.498
		m12	0.262	8.020	35.131	0.599	0.494
		m21	0.190	11.098	36.162	0.539	0.496
	ll1000	*N1*	0.799	0.123	6963.919	0.940	0.532
		N2	0.709	0.146	8016.151	0.927	0.498
		Nanc	0.986	0.037	1568.912	0.986	0.647
		Tsep	0.496	0.494	4183.320	0.832	0.539
		m12	0.306	7.787	32.411	0.666	0.525
		m21	0.226	21.369	33.942	0.578	0.521

		Coverage 30x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.655	0.252	9444.570	0.864	0.520
		N2	0.640	0.269	9837.096	0.858	0.524
		Nanc	0.973	0.048	2304.978	0.972	0.585
		Tsep	0.417	0.559	4886.560	0.740	0.507
		m12	0.227	3.177	35.416	0.578	0.496
		m21	0.167	4.275	38.643	0.506	0.503
	ll1000	*N1*	0.703	0.244	7829.534	0.907	0.524
		N2	0.686	0.170	8311.963	0.912	0.535
		Nanc	0.975	0.040	1161.914	0.984	0.670
		Tsep	0.527	0.331	4058.775	0.820	0.558
		m12	0.264	4.496	32.820	0.690	0.566
		m21	0.167	5.145	31.451	0.606	0.503
nc20	ll200	*N1*	0.740	0.216	8358.351	0.886	0.481
		N2	0.655	0.183	9419.081	0.883	0.511
		Nanc	0.943	0.145	2908.572	0.958	0.570
		Tsep	0.378	0.687	4804.651	0.748	0.510
		m12	0.255	4.623	35.286	0.604	0.508
		m21	0.197	2.054	34.490	0.547	0.521
	ll1000	*N1*	0.770	0.217	7404.215	0.910	0.521
		N2	0.688	0.179	8458.851	0.920	0.500
		Nanc	0.983	0.038	1638.021	0.981	0.662
		Tsep	0.498	0.457	4181.855	0.809	0.562

		m12	0.336	1.612	27.841	0.697	0.552
		m21	0.205	9.499	35.989	0.556	0.503
nc50	ll200	*N1*	0.809	0.179	7504.063	0.921	0.530
		N2	0.764	0.142	8448.758	0.924	0.496
		Nanc	0.932	0.153	3263.440	0.948	0.574
		Tsep	0.341	0.659	5193.381	0.752	0.497
		m12	0.233	1.497	34.498	0.646	0.542
		m21	0.200	4.439	37.357	0.546	0.525
	ll1000	*N1*	0.813	0.122	6706.854	0.947	0.529
		N2	0.725	0.104	7704.226	0.945	0.525
		Nanc	0.970	0.041	1678.281	0.984	0.669
		Tsep	0.466	0.525	4308.982	0.821	0.553
		m12	0.331	2.110	27.455	0.703	0.564
		m21	0.220	34.586	34.958	0.584	0.525

Supplementary Table 5.11. Accuracy of the estimated parameters of the Divergence with pulse of admixture model assessed by 1,000 pods. Combinations of experimental parameters considering 1,000 loci.

		Coverage 1x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.546	0.262	11130.775	0.848	0.487
		N2	0.599	0.214	10091.491	0.876	0.509
		Nanc	0.974	0.017	1825.056	0.994	0.540
		Tadm	0.475	0.170	466.348	0.938	0.511
		Tsep	0.605	0.155	1702.689	0.929	0.538
		adm12	0.076	0.146	0.043	0.930	0.526
		adm21	0.077	0.151	0.045	0.924	0.502
	ll1000	*N1*	0.672	0.177	8194.508	0.923	0.567
		N2	0.716	0.124	8229.567	0.948	0.554
		Nanc	0.985	0.005	1271.255	0.997	0.634
		Tadm	0.525	0.134	424.338	0.952	0.510
		Tsep	0.658	0.122	1401.779	0.948	0.587
		adm12	0.067	0.149	0.043	0.922	0.499
		adm21	0.065	0.163	0.044	0.915	0.493
nc20	ll200	*N1*	0.644	0.180	10111.944	0.901	0.492
		N2	0.658	0.158	10211.257	0.893	0.480
		Nanc	0.969	0.006	2005.487	0.997	0.535
		Tadm	0.552	0.137	441.278	0.940	0.504
		Tsep	0.724	0.087	1457.427	0.952	0.530
		adm12	0.098	0.161	0.045	0.905	0.480
		adm21	0.108	0.144	0.045	0.914	0.491
	ll1000	*N1*	0.721	0.105	8001.344	0.945	0.548
		N2	0.717	0.100	8057.902	0.944	0.533
		Nanc	0.992	0.002	1254.815	1.000	0.604
		Tadm	0.559	0.113	398.900	0.965	0.512
		Tsep	0.722	0.056	1223.826	0.986	0.559

		adm12	0.073	0.140	0.043	0.927	0.511
		adm21	0.082	0.160	0.043	0.922	0.527
nc50	ll200	*N1*	0.654	0.154	9928.273	0.898	0.491
		N2	0.556	0.357	11223.367	0.816	0.494
		Nanc	0.990	0.022	3650.681	0.973	0.478
		Tadm	0.521	0.150	495.954	0.922	0.475
		Tsep	0.555	0.130	2015.829	0.905	0.511
		adm12	0.134	0.154	0.046	0.912	0.493
		adm21	0.137	0.147	0.046	0.901	0.496
	ll1000	*N1*	0.731	0.115	7573.528	0.951	0.544
		N2	0.787	0.131	7474.297	0.952	0.551
		Nanc	0.972	0.013	1765.633	0.990	0.568
		Tadm	0.439	0.156	440.803	0.930	0.525
		Tsep	0.522	0.098	1650.991	0.948	0.583
		adm12	0.067	0.146	0.045	0.914	0.484
		adm21	0.081	0.153	0.044	0.918	0.498

		Coverage 2x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.605	0.173	10167.407	0.889	0.475
		N2	0.566	0.193	10406.735	0.905	0.496
		Nanc	0.992	0.021	1813.034	0.993	0.527
		Tadm	0.538	0.130	429.489	0.949	0.519
		Tsep	0.697	0.100	1484.007	0.948	0.527
		adm12	0.087	0.139	0.044	0.933	0.505
		adm21	0.097	0.124	0.045	0.935	0.464
	ll1000	*N1*	0.737	0.100	8002.048	0.963	0.584
		N2	0.742	0.099	7963.528	0.952	0.556
		Nanc	0.984	0.005	1166.919	1.000	0.644
		Tadm	0.574	0.101	383.733	0.965	0.527
		Tsep	0.733	0.056	1126.227	0.987	0.560
		adm12	0.074	0.144	0.043	0.931	0.533
		adm21	0.074	0.156	0.044	0.924	0.480
nc20	ll200	*N1*	0.612	0.212	10169.393	0.870	0.533
		N2	0.599	0.193	9914.493	0.889	0.525
		Nanc	0.979	0.013	2083.050	0.993	0.547
		Tadm	0.587	0.089	428.780	0.960	0.494
		Tsep	0.728	0.082	1372.957	0.972	0.527
		adm12	0.096	0.146	0.044	0.917	0.509
		adm21	0.094	0.143	0.044	0.922	0.518
	ll1000	*N1*	0.735	0.087	7272.509	0.968	0.556
		N2	0.691	0.085	7633.071	0.962	0.554
		Nanc	0.982	0.002	1364.099	1.000	0.657
		Tadm	0.608	0.050	385.242	0.975	0.494
		Tsep	0.809	0.036	1028.959	0.992	0.594
		adm12	0.076	0.162	0.043	0.915	0.511
		adm21	0.089	0.160	0.043	0.918	0.525
nc50	ll200	*N1*	0.749	0.117	8569.528	0.925	0.494

		N2	0.584	0.264	10501.004	0.851	0.521
		Nanc	0.957	0.020	3187.190	0.974	0.517
		Tadm	0.425	0.148	486.768	0.932	0.497
		Tsep	0.443	0.128	2105.536	0.902	0.526
		adm12	0.106	0.129	0.046	0.926	0.473
		adm21	0.111	0.127	0.045	0.923	0.483
	ll1000	*N1*	0.752	0.084	6634.115	0.973	0.586
		N2	0.769	0.085	7223.169	0.968	0.565
		Nanc	0.980	0.011	1809.860	0.993	0.586
		Tadm	0.557	0.103	398.087	0.968	0.522
		Tsep	0.678	0.063	1336.505	0.979	0.608
		adm12	0.066	0.153	0.044	0.924	0.484
		adm21	0.070	0.136	0.043	0.926	0.517

		Coverage 5x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.606	0.189	9771.302	0.890	0.560
		N2	0.661	0.154	9756.026	0.902	0.513
		Nanc	0.983	0.009	1899.363	0.998	0.559
		Tadm	0.566	0.110	426.074	0.959	0.497
		Tsep	0.695	0.073	1392.094	0.972	0.525
		adm12	0.092	0.147	0.042	0.927	0.538
		adm21	0.099	0.130	0.043	0.932	0.505
	ll1000	*N1*	0.757	0.071	7316.149	0.970	0.566
		N2	0.695	0.069	7507.254	0.973	0.556
		Nanc	0.983	0.004	1182.485	0.999	0.686
		Tadm	0.652	0.084	359.958	0.971	0.519
		Tsep	0.782	0.042	991.977	0.989	0.603
		adm12	0.080	0.157	0.043	0.919	0.504
		adm21	0.076	0.147	0.043	0.927	0.500
nc20	ll200	*N1*	0.708	0.111	8539.791	0.947	0.518
		N2	0.652	0.151	9316.438	0.917	0.505
		Nanc	0.976	0.019	2351.998	0.990	0.571
		Tadm	0.631	0.077	400.676	0.964	0.484
		Tsep	0.753	0.060	1297.350	0.986	0.520
		adm12	0.103	0.144	0.044	0.923	0.508
		adm21	0.094	0.200	0.045	0.885	0.488
	ll1000	*N1*	0.792	0.102	6593.486	0.973	0.571
		N2	0.764	0.109	7028.718	0.966	0.551
		Nanc	0.981	0.008	1390.129	0.996	0.644
		Tadm	0.675	0.080	349.530	0.980	0.539
		Tsep	0.806	0.024	1049.131	0.991	0.568
		adm12	0.077	0.165	0.044	0.914	0.482
		adm21	0.101	0.173	0.042	0.922	0.506
nc50	ll200	*N1*	0.710	0.160	8195.573	0.928	0.540
		N2	0.775	0.107	7717.886	0.949	0.553
		Nanc	0.956	0.011	2912.344	0.985	0.545
		Tadm	0.617	0.095	398.707	0.975	0.493

		Tsep	0.661	0.033	1504.706	0.975	0.521
		adm12	0.089	0.156	0.044	0.919	0.500
		adm21	0.085	0.144	0.045	0.915	0.494
	ll1000	*N1*	0.818	0.089	6160.797	0.976	0.596
		N2	0.786	0.088	6650.572	0.968	0.591
		Nanc	0.968	0.007	1720.445	0.991	0.626
		Tadm	0.675	0.068	351.189	0.990	0.490
		Tsep	0.739	0.043	1237.254	0.988	0.548
		adm12	0.087	0.137	0.042	0.940	0.511
		adm21	0.092	0.158	0.043	0.920	0.522

		Coverage 30x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.606	0.183	9823.368	0.899	0.516
		N2	0.708	0.153	9161.712	0.909	0.510
		Nanc	0.976	0.006	2026.205	0.993	0.540
		Tadm	0.548	0.122	434.992	0.950	0.491
		Tsep	0.663	0.087	1484.241	0.966	0.536
		adm12	0.095	0.164	0.043	0.925	0.516
		adm21	0.088	0.167	0.045	0.916	0.492
	ll1000	*N1*	0.739	0.080	7097.446	0.973	0.558
		N2	0.763	0.097	6819.645	0.975	0.568
		Nanc	0.984	0.001	1315.735	1.000	0.675
		Tadm	0.600	0.083	376.392	0.978	0.504
		Tsep	0.759	0.048	936.737	0.990	0.645
		adm12	0.073	0.146	0.043	0.929	0.509
		adm21	0.072	0.147	0.043	0.934	0.497
nc20	ll200	*N1*	0.729	0.105	8367.586	0.950	0.533
		N2	0.637	0.128	8624.309	0.935	0.519
		Nanc	0.963	0.022	2316.546	0.991	0.554
		Tadm	0.628	0.103	381.771	0.968	0.551
		Tsep	0.738	0.064	1242.293	0.988	0.552
		adm12	0.088	0.190	0.045	0.901	0.500
		adm21	0.089	0.162	0.043	0.914	0.510
	ll1000	*N1*	0.802	0.067	6205.341	0.986	0.574
		N2	0.805	0.077	6223.132	0.982	0.567
		Nanc	0.970	0.003	1474.738	0.998	0.669
		Tadm	0.654	0.057	352.295	0.988	0.509
		Tsep	0.787	0.022	945.127	0.996	0.623
		adm12	0.085	0.135	0.042	0.930	0.510
		adm21	0.081	0.131	0.042	0.928	0.519
nc50	ll200	*N1*	0.742	0.107	7564.815	0.959	0.509
		N2	0.745	0.144	7976.194	0.943	0.515
		Nanc	0.949	0.000	3116.077	0.990	0.551
		Tadm	0.663	0.087	370.928	0.976	0.495
		Tsep	0.712	0.041	1354.106	0.984	0.557
		adm12	0.097	0.145	0.043	0.918	0.525
		adm21	0.084	0.158	0.044	0.916	0.478

	ll1000	*N1*	0.817	0.056	5614.259	0.993	0.572
		N2	0.839	0.057	5651.357	0.991	0.579
		Nanc	0.974	0.002	1559.145	0.999	0.667
		Tadm	0.717	0.059	335.390	0.991	0.533
		Tsep	0.789	0.015	1093.603	0.993	0.572
		adm12	0.089	0.152	0.043	0.920	0.498
		adm21	0.093	0.146	0.042	0.931	0.531

Supplementary Table 5.12. Accuracy of the estimated parameters of the Divergence with pulse of admixture model assessed by 1,000 pods. Combinations of experimental parameters considering 5,000 loci.

		Coverage 1x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.756	0.116	8291.115	0.953	0.531
		N2	0.646	0.174	9134.845	0.911	0.541
		Nanc	0.986	0.004	1148.603	0.998	0.557
		Tadm	0.533	0.110	434.016	0.952	0.505
		Tsep	0.701	0.099	1391.240	0.953	0.529
		adm12	0.081	0.172	0.044	0.918	0.479
		adm21	0.080	0.150	0.044	0.914	0.507
	ll1000	*N1*	0.771	0.072	6647.890	0.978	0.565
		N2	0.826	0.066	5883.734	0.985	0.573
		Nanc	0.994	0.002	695.711	1.000	0.681
		Tadm	0.547	0.104	403.042	0.975	0.467
		Tsep	0.720	0.059	978.607	0.988	0.640
		adm12	0.086	0.137	0.043	0.920	0.500
		adm21	0.083	0.137	0.043	0.935	0.492
nc20	ll200	*N1*	0.670	0.163	8440.250	0.927	0.523
		N2	0.756	0.078	8226.346	0.949	0.498
		Nanc	0.997	0.005	1349.302	0.993	0.535
		Tadm	0.577	0.124	419.847	0.948	0.516
		Tsep	0.704	0.080	1377.175	0.966	0.551
		adm12	0.104	0.181	0.045	0.914	0.479
		adm21	0.121	0.151	0.044	0.913	0.509
	ll1000	*N1*	0.829	0.058	5823.267	0.995	0.603
		N2	0.814	0.062	6025.304	0.988	0.617
		Nanc	0.993	0.002	745.688	0.998	0.668
		Tadm	0.660	0.064	372.707	0.984	0.487
		Tsep	0.857	0.037	858.195	0.993	0.602
		adm12	0.100	0.177	0.044	0.913	0.486
		adm21	0.090	0.168	0.044	0.927	0.488
nc50	ll200	*N1*	0.618	0.219	9066.509	0.887	0.519
		N2	0.760	0.090	8247.314	0.936	0.497
		Nanc	0.985	0.005	2025.551	0.992	0.552
		Tadm	0.578	0.130	446.197	0.939	0.512
		Tsep	0.630	0.106	1736.961	0.918	0.528
		adm12	0.107	0.173	0.045	0.907	0.475

		adm21	0.120	0.158	0.046	0.901	0.503
	ll1000	*N1*	0.844	0.053	5236.700	0.988	0.600
		N2	0.855	0.063	5225.656	0.988	0.599
		Nanc	0.980	0.002	935.789	1.000	0.640
		Tadm	0.650	0.081	389.137	0.972	0.504
		Tsep	0.707	0.064	1199.189	0.987	0.596
		adm12	0.089	0.139	0.043	0.934	0.516
		adm21	0.098	0.123	0.044	0.922	0.517

		Coverage 2x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.773	0.088	7339.078	0.967	0.561
		N2	0.779	0.111	7043.112	0.961	0.605
		Nanc	0.988	0.008	1088.319	0.997	0.602
		Tadm	0.568	0.095	390.686	0.965	0.501
		Tsep	0.751	0.077	1177.906	0.973	0.585
		adm12	0.085	0.133	0.043	0.923	0.515
		adm21	0.078	0.148	0.045	0.920	0.488
	ll1000	*N1*	0.799	0.037	5861.431	0.989	0.598
		N2	0.813	0.057	6125.758	0.986	0.580
		Nanc	0.997	0.002	678.179	1.000	0.712
		Tadm	0.672	0.077	363.650	0.983	0.481
		Tsep	0.830	0.034	839.707	0.993	0.620
		adm12	0.103	0.152	0.042	0.923	0.500
		adm21	0.099	0.161	0.043	0.925	0.513
nc20	ll200	*N1*	0.776	0.102	7845.828	0.952	0.551
		N2	0.730	0.126	8120.566	0.941	0.554
		Nanc	0.991	0.002	1403.866	1.000	0.577
		Tadm	0.638	0.087	388.444	0.978	0.490
		Tsep	0.778	0.034	1074.476	0.990	0.577
		adm12	0.113	0.136	0.044	0.924	0.503
		adm21	0.107	0.149	0.044	0.922	0.500
	ll1000	*N1*	0.815	0.055	5906.411	0.987	0.583
		N2	0.798	0.059	5881.622	0.985	0.591
		Nanc	0.994	0.001	842.357	0.999	0.675
		Tadm	0.678	0.081	371.951	0.978	0.470
		Tsep	0.865	0.045	860.378	0.993	0.580
		adm12	0.093	0.172	0.042	0.925	0.519
		adm21	0.096	0.163	0.043	0.924	0.502
nc50	ll200	*N1*	0.715	0.114	7655.604	0.950	0.516
		N2	0.717	0.102	7897.550	0.948	0.527
		Nanc	0.985	0.007	2314.283	0.990	0.554
		Tadm	0.630	0.094	386.857	0.975	0.493
		Tsep	0.727	0.026	1351.056	0.984	0.518
		adm12	0.119	0.158	0.044	0.926	0.508
		adm21	0.119	0.125	0.044	0.922	0.509
	ll1000	*N1*	0.811	0.051	5546.903	0.992	0.594
		N2	0.853	0.044	4628.994	0.997	0.618

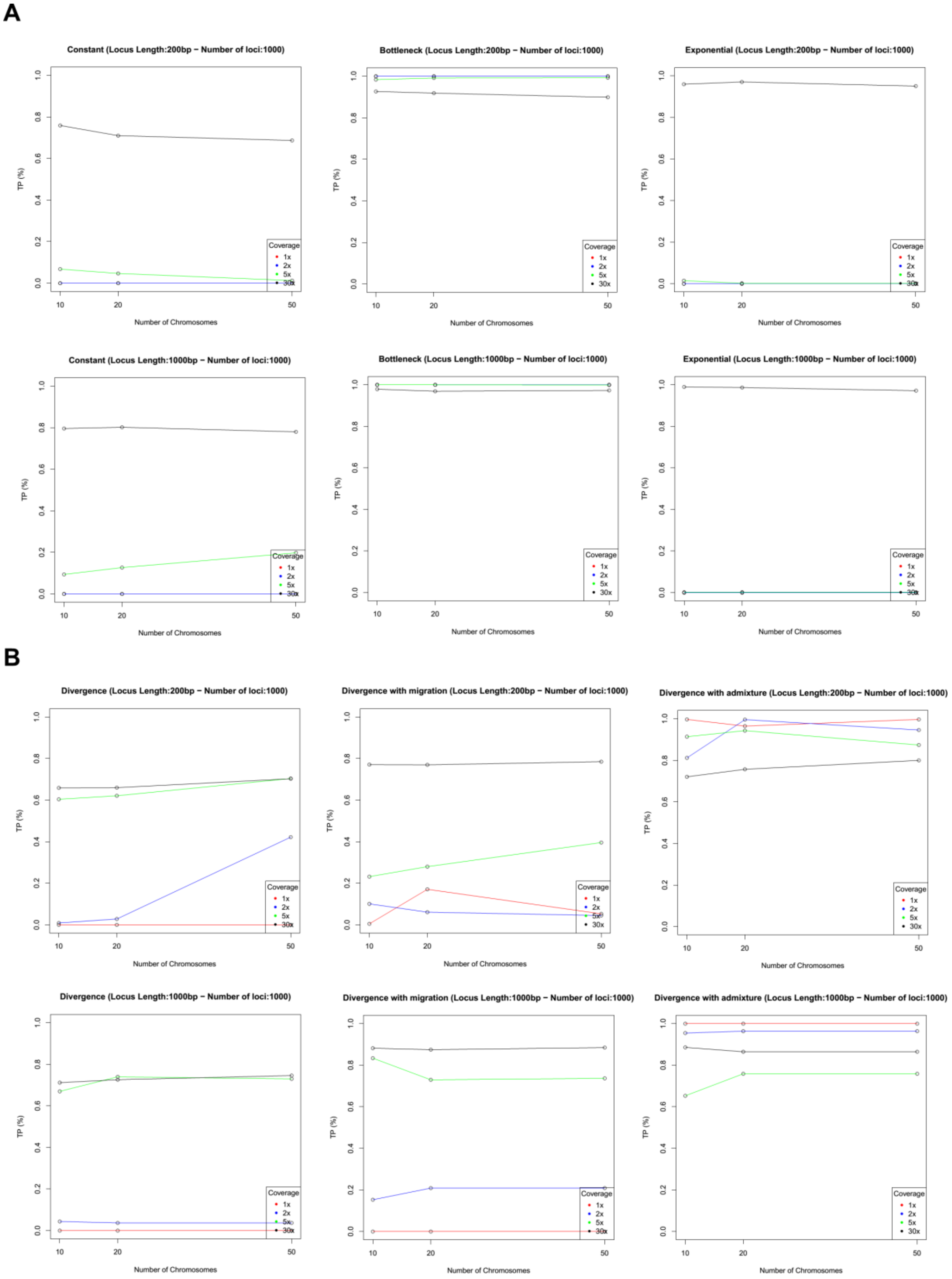
		Nanc	0.985	0.006	1098.199	0.999	0.641
		Tadm	0.670	0.068	361.052	0.989	0.526
		Tsep	0.734	0.039	1181.380	0.986	0.610
		adm12	0.103	0.135	0.043	0.927	0.505
		adm21	0.101	0.163	0.044	0.912	0.490

		Coverage 5x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.753	0.106	8268.709	0.948	0.514
		N2	0.710	0.151	8307.573	0.936	0.552
		Nanc	1.000	0.002	1128.252	0.998	0.581
		Tadm	0.672	0.085	373.503	0.969	0.535
		Tsep	0.789	0.050	1026.727	0.987	0.568
		adm12	0.107	0.154	0.044	0.920	0.499
		adm21	0.092	0.142	0.045	0.920	0.469
	ll1000	*N1*	0.796	0.033	5726.459	0.992	0.602
		N2	0.844	0.041	5831.839	0.995	0.578
		Nanc	0.991	0.000	681.811	1.000	0.714
		Tadm	0.668	0.066	366.881	0.987	0.466
		Tsep	0.844	0.033	777.561	0.996	0.608
		adm12	0.085	0.142	0.042	0.936	0.502
		adm21	0.097	0.157	0.044	0.912	0.482
nc20	ll200	*N1*	0.812	0.085	7019.189	0.973	0.537
		N2	0.741	0.096	7089.348	0.973	0.550
		Nanc	0.977	0.012	1559.038	0.996	0.582
		Tadm	0.669	0.069	364.814	0.991	0.480
		Tsep	0.834	0.039	976.436	0.996	0.562
		adm12	0.097	0.137	0.044	0.926	0.498
		adm21	0.103	0.144	0.043	0.934	0.510
	ll1000	*N1*	0.828	0.051	5122.222	0.988	0.612
		N2	0.815	0.056	5475.032	0.985	0.614
		Nanc	0.989	0.002	856.310	1.000	0.666
		Tadm	0.680	0.064	357.623	0.985	0.479
		Tsep	0.827	0.019	878.846	0.998	0.550
		adm12	0.099	0.165	0.045	0.909	0.463
		adm21	0.105	0.137	0.043	0.927	0.502
nc50	ll200	*N1*	0.793	0.109	7061.329	0.959	0.547
		N2	0.766	0.092	7210.604	0.962	0.513
		Nanc	0.964	0.013	2187.226	0.994	0.566
		Tadm	0.704	0.044	371.245	0.983	0.498
		Tsep	0.746	0.024	1275.213	0.989	0.534
		adm12	0.102	0.156	0.045	0.917	0.481
		adm21	0.099	0.153	0.045	0.920	0.475
	ll1000	*N1*	0.876	0.054	4784.508	0.993	0.647
		N2	0.861	0.049	4751.239	0.994	0.647
		Nanc	0.985	0.008	1052.834	0.997	0.671
		Tadm	0.678	0.072	337.184	0.989	0.524
		Tsep	0.791	0.021	1094.193	0.994	0.559

		adm12	0.116	0.136	0.042	0.930	0.507
		adm21	0.114	0.137	0.043	0.931	0.499

		Coverage 30x					
		Parameter	**R2**	**Bias**	**RMSE**	**Factor2**	**Coverage50%**
nc10	ll200	*N1*	0.724	0.108	7912.677	0.959	0.545
		N2	0.754	0.083	7804.380	0.969	0.527
		Nanc	0.984	0.000	1318.414	0.999	0.575
		Tadm	0.617	0.086	387.329	0.968	0.488
		Tsep	0.742	0.049	1097.765	0.985	0.589
		adm12	0.105	0.146	0.044	0.929	0.490
		adm21	0.102	0.136	0.044	0.930	0.492
	ll1000	*N1*	0.815	0.055	5372.516	0.993	0.589
		N2	0.822	0.050	5364.411	0.992	0.585
		Nanc	0.987	0.001	721.031	1.000	0.711
		Tadm	0.643	0.086	359.445	0.983	0.505
		Tsep	0.844	0.034	781.658	0.991	0.646
		adm12	0.092	0.161	0.042	0.925	0.521
		adm21	0.100	0.163	0.041	0.924	0.517
nc20	ll200	*N1*	0.819	0.084	6709.087	0.984	0.552
		N2	0.780	0.105	7503.843	0.962	0.557
		Nanc	0.974	0.003	1611.378	0.999	0.589
		Tadm	0.651	0.080	361.369	0.972	0.499
		Tsep	0.769	0.026	1039.546	0.992	0.572
		adm12	0.109	0.123	0.044	0.919	0.500
		adm21	0.105	0.138	0.044	0.942	0.484
	ll1000	*N1*	0.849	0.054	5179.492	0.990	0.591
		N2	0.841	0.056	5077.558	0.993	0.609
		Nanc	0.991	0.003	800.275	0.999	0.712
		Tadm	0.718	0.062	350.865	0.990	0.504
		Tsep	0.863	0.023	882.932	0.996	0.578
		adm12	0.111	0.137	0.042	0.935	0.505
		adm21	0.096	0.166	0.044	0.918	0.465
nc50	ll200	*N1*	0.739	0.092	6888.367	0.967	0.536
		N2	0.840	0.085	6545.927	0.976	0.529
		Nanc	0.972	0.006	1966.477	0.996	0.576
		Tadm	0.722	0.059	340.551	0.988	0.521
		Tsep	0.791	0.016	1175.963	0.989	0.543
		adm12	0.122	0.119	0.042	0.933	0.523
		adm21	0.103	0.139	0.043	0.937	0.500
	ll1000	*N1*	0.876	0.024	4378.537	0.996	0.620
		N2	0.862	0.030	4576.962	0.994	0.618
		Nanc	0.990	0.003	968.492	0.997	0.688
		Tadm	0.719	0.054	324.168	0.992	0.518
		Tsep	0.817	0.001	1027.107	0.998	0.526
		adm12	0.127	0.117	0.042	0.944	0.500
		adm21	0.115	0.135	0.042	0.938	0.529

Supplementary Figure 5.1. Proportion of True Positives for (A) the one-population models, (B) the two-population models, obtained correcting the pods through the genotype likelihood computed by ANGSD. The plots have the same features of Figure 5.3.



Paper I

Received: 13 November 2019 | Revised: 28 August 2020 | Accepted: 7 September 2020

DOI: 10.1111/1755-0998.13263

SPECIAL ISSUE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

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] | Andrea Benazzo[2]

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

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

Correspondence
Andrea Benazzo, Department of Life Sciences and Biotechnology, University of Ferrara, 44121 Ferrara, Italy.
Email: andrea.benazzo@unife.it

Silvia Ghirotto, Department of Mathematics and Computer Science, University of Ferrara, 44121 Ferrara, Italy.
Email: silvia.ghirotto@unife.it

Funding information
Fondo per l'Incentivazione alla Ricerca 2018 Università di Ferrara; H2020 European Research Council, Grant/Award Number: 295733

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 analysed, 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 analysed, number and size of the genetic loci considered. We also compared the *FDSS* with the unfolded and folded site frequency spectrum (*SFS*), and for these statistics we highlighted the experimental conditions maximizing the inferential power of the ABC-RF procedure. We finally analysed real data sets, 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.

1 | 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; Meyer et al., 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 (1,000 Dasmahapatra et al., 2012; De Manuel et al., 2016; Genomes Project Consortium, 2012; Mallick et al., 2016; 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

Ghirotto and Vizzari authors contributed equally to this work.

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 analysing 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 et al., 2016) and computer simulations (Chikhi et al., 2018).

Other methods infer demographic parameters via the diffusion approximation (Gutenkunst et al., 2010), or coalescent simulations (Beeravolu et al., 2018; Excoffier et al., 2013), from the *SFS* computed on large genomic data sets. 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 et al., 2017; Kamm et al., 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 analysed, 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 data sets 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 et al., 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 et al., 2010; Csilléry et al., 2010).

Although ABC has the potential to deal with complex and realistic evolutionary scenarios, its application to the analysis of large genomic data sets, 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 data set 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 (Beaumont et al., 2002; Marjoram et al., 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 Blum et al., 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 data sets. In this light, the unfolded *SFS*, that due to the above mentioned limitations has been rarely used in a classical ABC context (Eldon et al., 2015), should be a suitable (and possibly sufficient) statistic to summarize genomic data (Lapierre et al., 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 et al., 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 (a) can be easily calculated from unphased genomes data; (b) do not require information about ancestral state of alleles; and (c) 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 et al., 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 analyse 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 analysed 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).

2 | MATERIALS AND METHODS

2.1 | 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 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 data sets 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.

2.2 | The *FDSS*

To compute the *FDSS* we evaluated the genomic distributions of the four mutually exclusive categories of segregating sites in two populations, namely (a) segregating sites private of the first population; (b) segregating sites private of the second populations; (c) segregating sites that are polymorphic in both populations; and (d) 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}, $\mathrm{Lj}_{\mathrm{ii}}$, $\mathrm{Lj}_{\mathrm{iii}}$, $\mathrm{Lj}_{\mathrm{iv}}$} 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), $\mathrm{SS}_{\mathrm{ii}}$(z), $\mathrm{SS}_{\mathrm{iii}}$(z), $\mathrm{SS}_{\mathrm{iv}}$(z) are:

$$\mathrm{SS}_A(x)=\sum_{j=1}^{k} I\left(\mathrm{Lj}_A=x\vee(x=n\wedge \mathrm{Lj}_A>x)\right), \quad \text{where } x\in N, x\leq n, A\in\{\mathrm{i}, \mathrm{ii}, \mathrm{iii}, \mathrm{iv}\}$$

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:

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

2.3 | 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 intralocus recombination (recombination rate = 1 x 10^{-8}), and with a fixed mutation rate (1 × 10^{-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 an 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 data set 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).

2.3.1 | One-population models

We started by considering four demographic models (Figure 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 (a) 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).

2.3.2 | Two-populations models

We then moved to considering three demographic models with two populations (Figure 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).

2.3.3 | Multipopulations 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

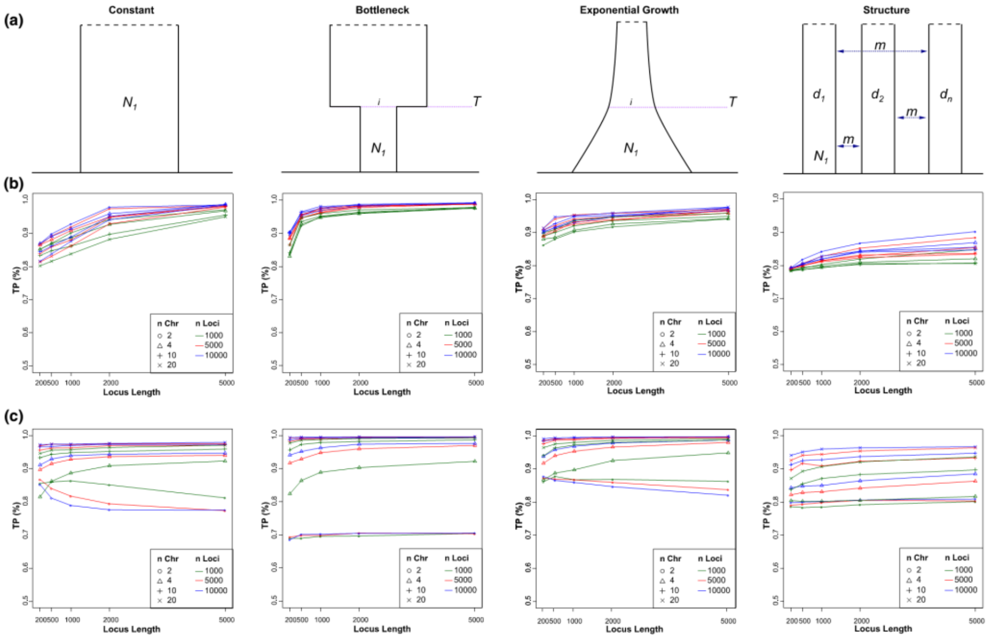


FIGURE 1 One-population models and proportion of true positives. (a) Demographic models compared: Constant, Bottleneck, Expansion, Structured population. N_1, the effective population size; *I*, intensity of the bottleneck or of the expansion; *T*, the time of the bottleneck or of the start of the expansion; *m*, the migration rate. (b) True positive rates for the *FDSS*. (c) True positive rates for the folded *SFS*. The plot below each of the four models represents the proportion of TPs obtained analysing 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 colours and the number of chromosomes is represented by different symbols

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 (Figure 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 (Meyer et al., 2012; Prüfer et al., 2014), we allowed for genetic exchanges from archaic to modern species, indicated in Figure 3 by the coloured 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 modelled bidirectional migration between modern populations along a stepping-stone, thus allowing for gene flow only between geographically neighbouring 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 modelling 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.

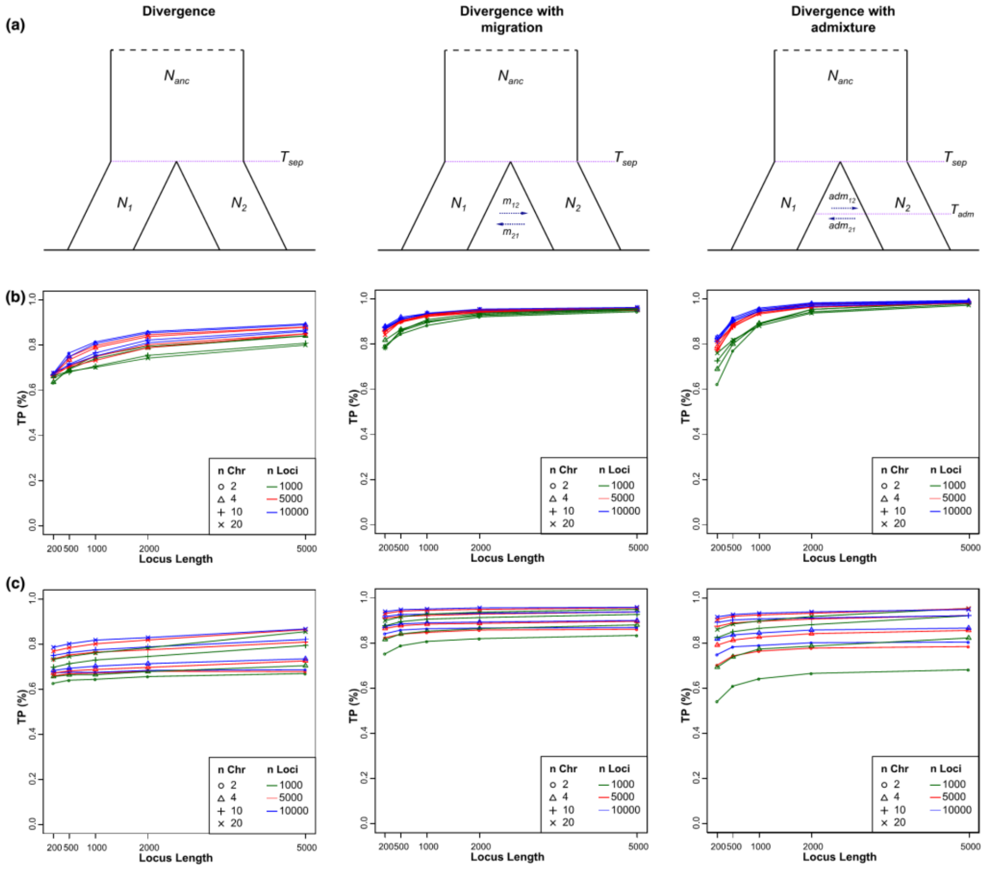


FIGURE 2 Two-populations models and proportion of true positives. (a) Demographic models compared: Divergence with isolation, Divergence with migration, Divergence with a single pulse of admixture. N_{anc} is the effective population size of the ancestral population, N_1 and N_2 are the effective population 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 positive rates for the *FDSS*. (c) True positive rates for the folded *SFS*. The plots have the same features of Figure 1

2.4 | Real case: Out of Africa dynamics

We explicitly compared SDM and MDM considering the high-coverage genomes of Denisova and Neandertal (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%; Meyer et al., 2012; Prüfer et al., 2014). Power analysis (see Results – Multipopulations models section), showed we could safely analyse 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

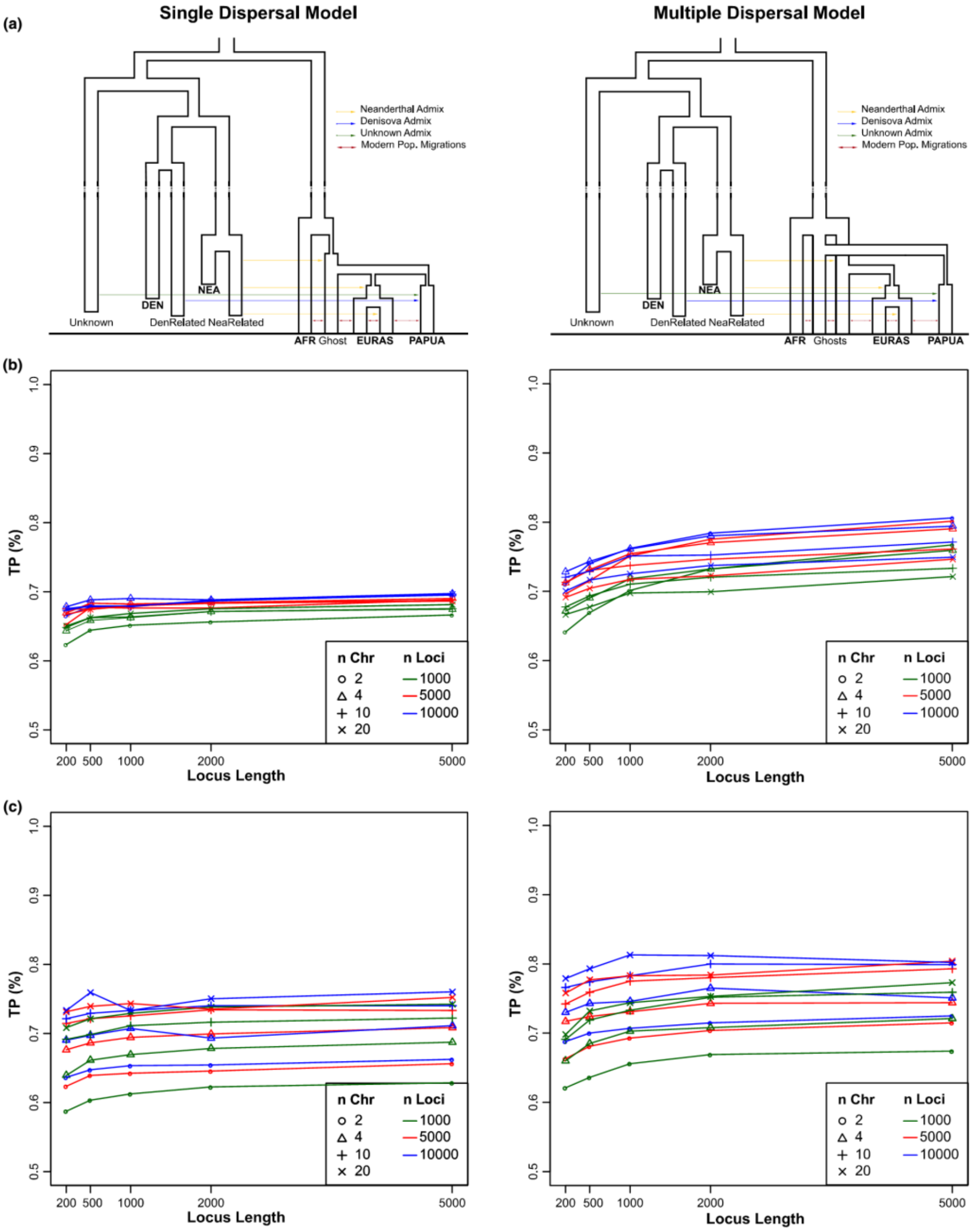
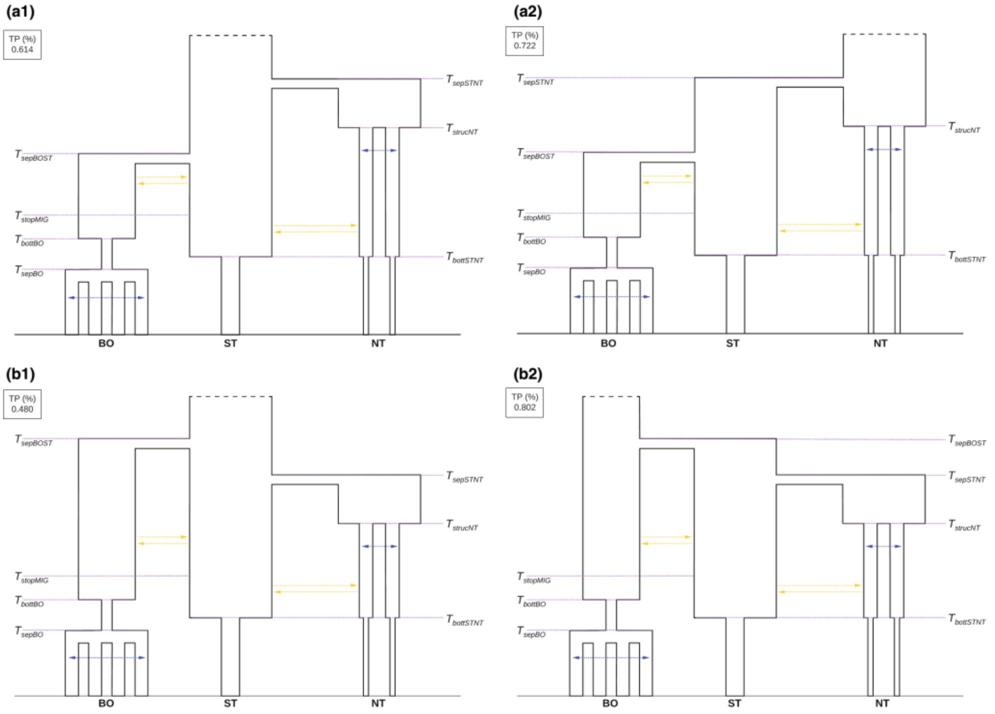


FIGURE 3 Multipopulation models and proportion of true positives. (a) Demographic models compared: Single dispersal and multiple dispersals. The populations sampled are indicated in bold. (b) True positive rates for the *FDSS*. (c) True ositive rates for the folded *SFS*. The plots have the same features of Figure 1



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

FIGURE 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 analysing 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)

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 analysed 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.

2.5 | 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 four 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), (Figure 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 Figure 4a. We modelled bidirectional migration both among populations within a species, and between neighbouring 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 (four Bornean, one south of Lake Toba and two north of Lake Toba), and 9,000 independent fragments of 1 kb 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 (Figure 4b). Also in this case, the model checking has been performed through PCA.

3 | RESULTS

3.1 | Power analysis

3.1.1 | One-population models

The four plots of Figure 1b report the results of the power analyses obtained summarizing the data through the FDSS, whereas plots of Figure 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 (Figure 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 Figure 1b). We showed that this behaviour 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 (Figure 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 10 and 20, 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% (Figure 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%).

3.1.2 | Two-populations models

The plots in Figure 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 (Figure 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.

3.1.3 | Multipopulations models

Figure 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 six 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 (Figure 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.

3.2 | 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 analysed; 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 two chromosomes per population) was 0.68 for the SDM, and 0.74 for the MDM (Figure 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 (Figure 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.

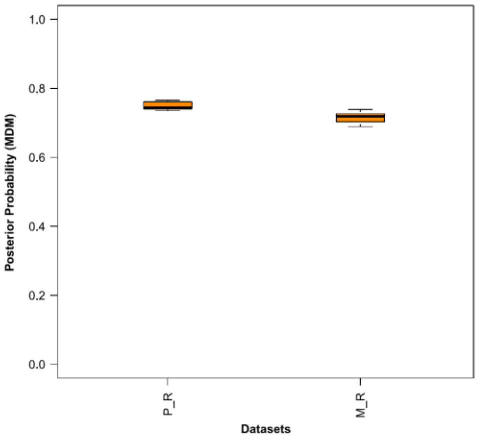


FIGURE 5 Posterior probabilities for the MDM. Left panel: posterior probabilities obtained analysing six Papuan individuals from Pagani et al. (2016) (PR). Right panel: posterior probabilities obtained analysing 25 Papuan individuals from Malaspinas et al. (2016) (MR)

3.3 | 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 (Figure 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 favoured 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* (Figure 4b). Model 1a also proven to be able to account for real variation, as it is shown in Figure S8.

4 | DISCUSSION

The cost of genotyping has dramatically dropped lately, making population-scale genomic data available for a large set of organisms (1,000 Benazzo et al., 2017; Dasmahapatra et al., 2012; De Manuel et al., 2016; Genomes Project Consortium, 2012; 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 (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 et al., 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).

4.1 | Power analysis

Initially, we analysed 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, Figures 1 and 2) and complex (multipopulations scenarios, Figure 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 (Figure 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 (Figure S2).

Among the two-populations demographies, the models with bidirectional 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 multipopulations 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 (Figure 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 multipopulations models through the first two moments of the four distributions. The results, reported in Figures S9–S11, are significantly in favour 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 analysed.

4.2 | 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 analysed, 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 data sets, 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.

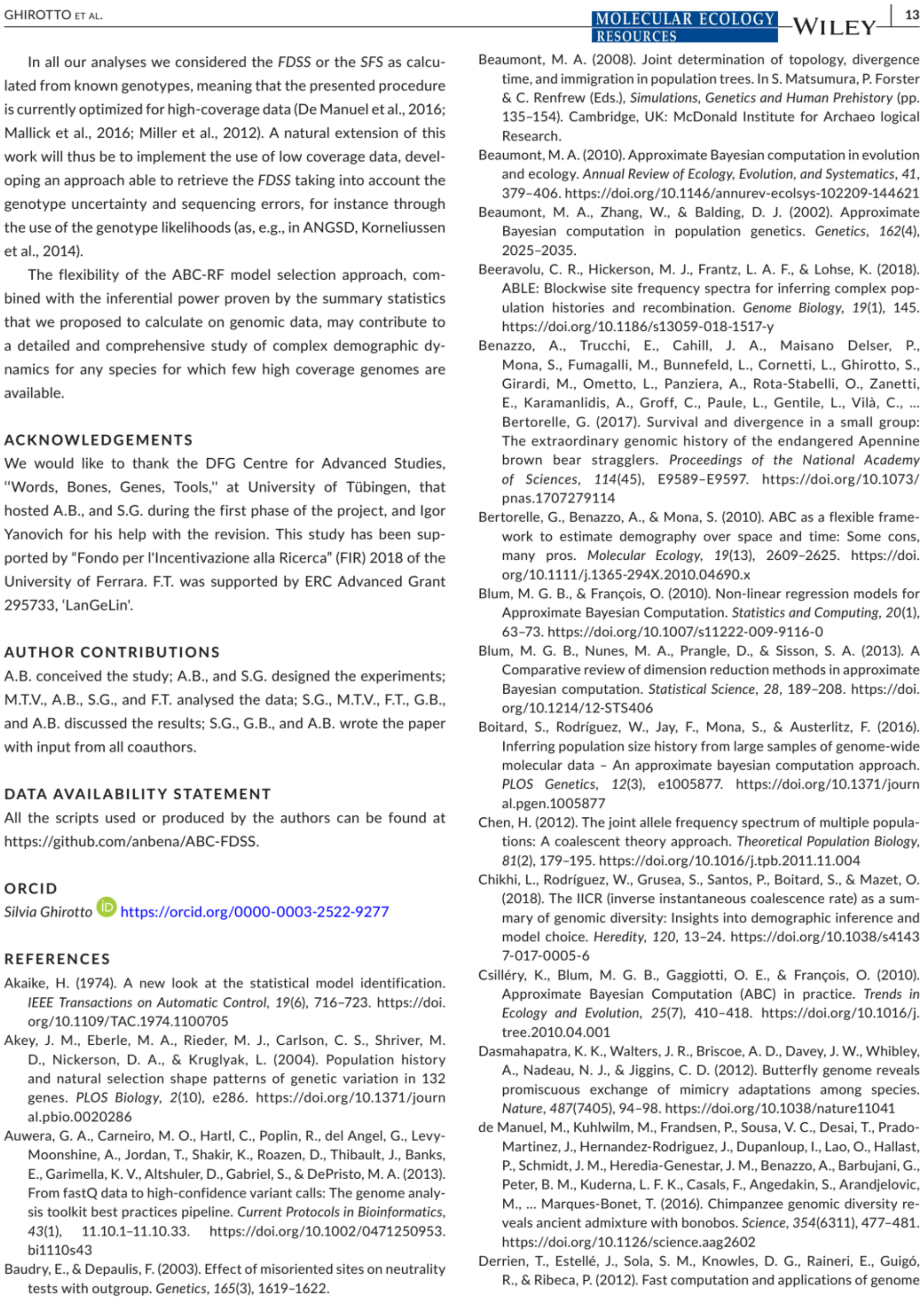
The performances of the folded and unfolded SFS are comparable, with a slight increase in the power of the unfolded spectrum for some specific conditions (usually when considering four chromosomes) or demographic model analysed (as one-populations models or MDM). However, we should remind that we generated the unfolded *SFS* through simulations, thus assuming that the ancestral state of alleles is known with certainty. When analysing 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 et al., 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.

4.3 | Applications to real data sets

We finally analysed 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 two chromosomes per population we obtained a good ability to tell models apart (Figure 3). Thirty-one replicated experiments, differing for the Papuan genome being considered, consistently supported the MDM over the SDM (Figure 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 study (Tassi et al., 2015), 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 (a) the alternative hypotheses are explicitly compared analysing complete genomes; (b) posterior probabilities are estimated for each model; and (c) 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 et al. (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 the conclusion of Nater et al. (2017) that the first split separated the North Toba and the newly identified South Toba species (Figure 4b). The main difference was about the strength of the support associated to this model. While Nater et al. (2017) estimated high posterior probabilities for the best-fitting model (73% when comparing the four models and 98% when comparing the two best scenarios), our procedure associated to the same model a posterior probability of 49% (Figure 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 (a) 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 (b) 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.

In conclusion, 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 analysed; 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 (Figure 3). This finding means that the ABC-RF model selection procedure through *FDSS* or *SFS* is suitable for the analysis of ancient data (Meyer et al., 2012) and of RAD sequencing data (Rowe et al., 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 et al., 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.

ACKNOWLEDGEMENTS

We would like to thank the DFG Centre for Advanced Studies, ''Words, Bones, Genes, Tools,'' at University of Tübingen, that hosted A.B., and S.G. 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. F.T. was supported by ERC Advanced Grant 295733, 'LanGeLin'.

AUTHOR CONTRIBUTIONS

A.B. conceived the study; A.B., and S.G. designed the experiments; M.T.V., A.B., S.G., and F.T. analysed the data; S.G., M.T.V., F.T., G.B., and A.B. discussed the results; S.G., G.B., and A.B. wrote the paper with input from all coauthors.

DATA AVAILABILITY STATEMENT

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

ORCID

Silvia Ghirotto https://orcid.org/0000-0003-2522-9277

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/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. https://doi.org/10.1371/journal.pbio.0020286

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & 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. https://doi.org/10.1002/0471250953.bi1110s43

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 S. Matsumura, P. Forster & C. Renfrew (Eds.), *Simulations, Genetics and Human Prehistory* (pp. 135–154). Cambridge, UK: McDonald Institute for Archaeo logical Research.

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

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

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. https://doi.org/10.1186/s13059-018-1517-y

Benazzo, A., Trucchi, E., Cahill, J. A., Maisano Delser, P., Mona, S., Fumagalli, M., Bunnefeld, L., Cornetti, L., Ghirotto, S., Girardi, M., Ometto, L., Panziera, A., Rota-Stabelli, O., Zanetti, E., Karamanlidis, A., Groff, C., Paule, L., Gentile, L., Vilà, C., ... 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. https://doi.org/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. https://doi.org/10.1111/j.1365-294X.2010.04690.x

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

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

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1038/nature11041

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., ... Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1371/journal.pgen.1003905

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

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2010). Diffusion approximations for demographic inference: DaDi. *Nature Precedings*, https://doi.org/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. https://doi.org/10.1093/molbev/msm108

Hinrichs, A. S., Raney, B. J., Speir, M. L., Rhead, B., Casper, J., Karolchik, D., Kuhn, R. M., Rosenbloom, K. R., Zweig, A. S., Haussler, D., & Kent, W. J. (2016). UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics*, *32*(9), 1430–1432. https://doi.org/10.1093/bioinformatics/btv766

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*, 337–338. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1093/bioinformatics/btp698

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/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. https://doi.org/10.1371/journal.pgen.1004379

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., … Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214. https://doi.org/10.1038/nature18299

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/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*, *100*(26), 15324–15328.

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1534/genetics.105.052837

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prufer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Paabo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., Kim, H. L., Burhans, R. C., Drautz, D. I., Wittekindt, N. E., Tomsho, L. P., Ibarra-Laclette, E., Herrera-Estrella, L., Peacock, E., Farley, S., Sage, G. K., Rode, K., Obbard, M., Montiel, R., … 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–E2390. https://doi.org/10.1073/pnas.1210506109

Moreno-Mayar, J. V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Chan, J., Spence, J. P., Allentoft, M. E., Vimala, T., Racimo, F., Pinotti, T., Rasmussen, S., Margaryan, A., Iraeta Orbegozo, M., Mylopotamitaki, D., Wooller, M., Bataille, C., Becerra-Valdivia, L., Chivall, D., Comeskey, D., … Willerslev, E. (2018). Early human dispersals within the Americas. *Science*, *362*(6419), eaav2621. https://doi.org/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. https://doi.org/10.1104/pp.108.127787

Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., de Manuel, M., Desai, T., Groves, C., Pybus, M., Sonay, T. B., Roos, C., Lameira, A. R., Wich, S. A., Askew, J., Davila-Ross, M., Fredriksson, G., de Valles, G., Casals, F., Prado-Martinez, J., Goossens, B., … Krützen, M. (2017). Morphometric, behavioral, and genomic evidence for a new orangutan species. *Current Biology*, *27*(22), 3576–3577. https://doi.org/10.1016/j.cub.2017.09.047

Pagani, L., Lawson, D. J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L., Wall, J. D., Cardona, A., Mägi, R., Sayres, M. A. W., Kaewert, S., Inchley, C., Scheib, C. L., Järve, M., Karmin, M., … Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, *538*(7624), 238–242. https://doi.org/10.1038/nature19792

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann,

M., Meyer, M., Ongyerth, M., ... Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1073/pnas.1503717112

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. https://doi.org/10.1111/j.1365-294X.2011.05413.x

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

SUPPORTING INFORMATION

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

How to cite this article: Ghirotto S, Vizzari MT, Tassi F, Barbujani G, Benazzo A. Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. *Mol Ecol Resour.* 2020;00:1–15. https://doi.org/10.1111/1755-0998.13263

MOLECULAR ECOLOGY
RESOURCES

Supplemental Information for:

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

Fig S1. Proportion of True Positives for the one-population models summarized through the *unfolded SFS*. The plots have the same features of Fig 1.

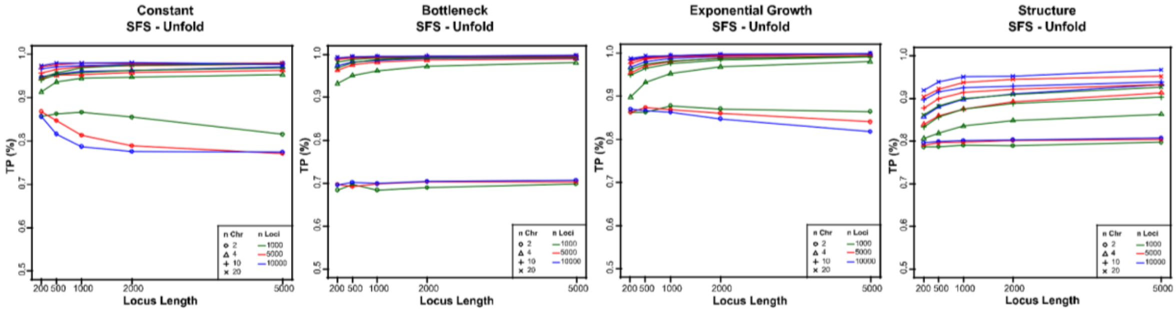


Fig S2. Proportion of True positives for the one-population structured model as a function of the migration rate (A) and the number of demes considered (B). (A) Each plot represents the proportion of pods from the structured model assigned to each of the four one-population models with the migration rates among demes in the structured model constrained at ranges of increasing values (from 1*10-5 to 1*10-1). All the plots consider two chromosomes and a specific combination of locus length and number of loci; the number of demes in the structured model is fixed to four. In general, the TP rate (in dark blue) decreases as increasing the migration rate among demes, with the constant model erroneously recognize as the true model for higher migration rates. (B) Proportion of pods from the structured model assigned to each of the four one population models as a function of the number of demes (from 2 to 10). The TP rate increase with the number of demes, regardless of the level of migration among demes.

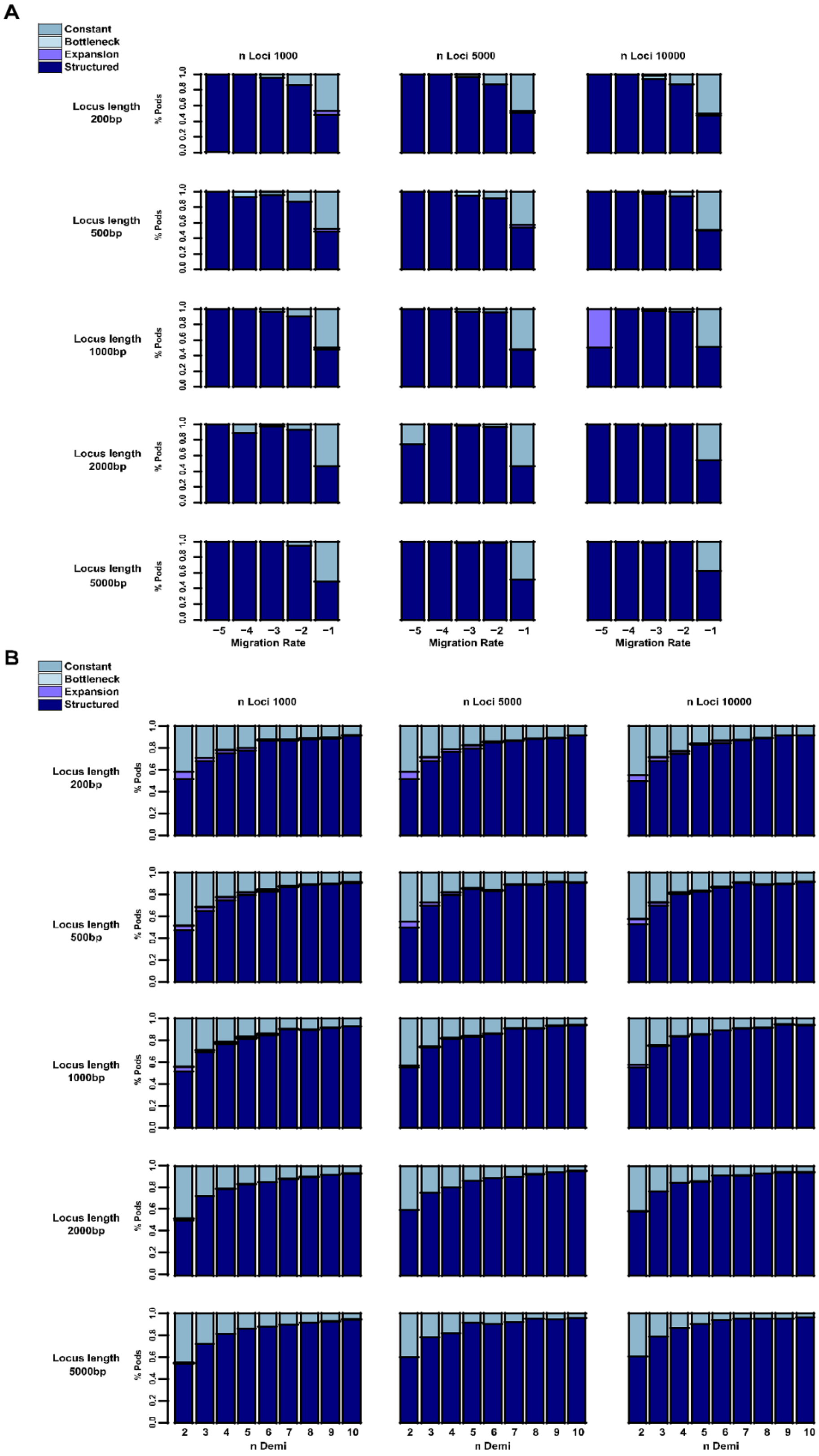

A
Constant
Bottleneck
Expansion
Structured
n Loci 1000
n Loci 5000
n Loci 10000
Locus length 200bp
Locus length 500bp
Locus length 1000bp
Locus length 2000bp
Locus length 5000bp
% Pods
0.0 0.2 0.4 0.6 0.8 1.0
−5 −4 −3 −2 −1
Migration Rate
B
Constant
Bottleneck
Expansion
Structured
n Loci 1000
n Loci 5000
n Loci 10000
Locus length 200bp
Locus length 500bp
Locus length 1000bp
Locus length 2000bp
Locus length 5000bp
% Pods
0.0 0.2 0.4 0.6 0.8 1.0
2 3 4 5 6 7 8 9 10
n Demi

Fig S3. *FDSS* generated under the one-population models for each number of chromosomes tested. Each plot represents the *FDSS* simulated under each of the four one-population models considering 1,000 fragments of 1,000 base pair length, for a specific number of chromosomes sampled. Going from two to twenty chromosomes, we observe an increase of the overlapping between the *FDSS* generated under the Constant and the Structured model, thus possibly explaining the decrease in the models' identifiability as increasing the number of chromosomes considered.

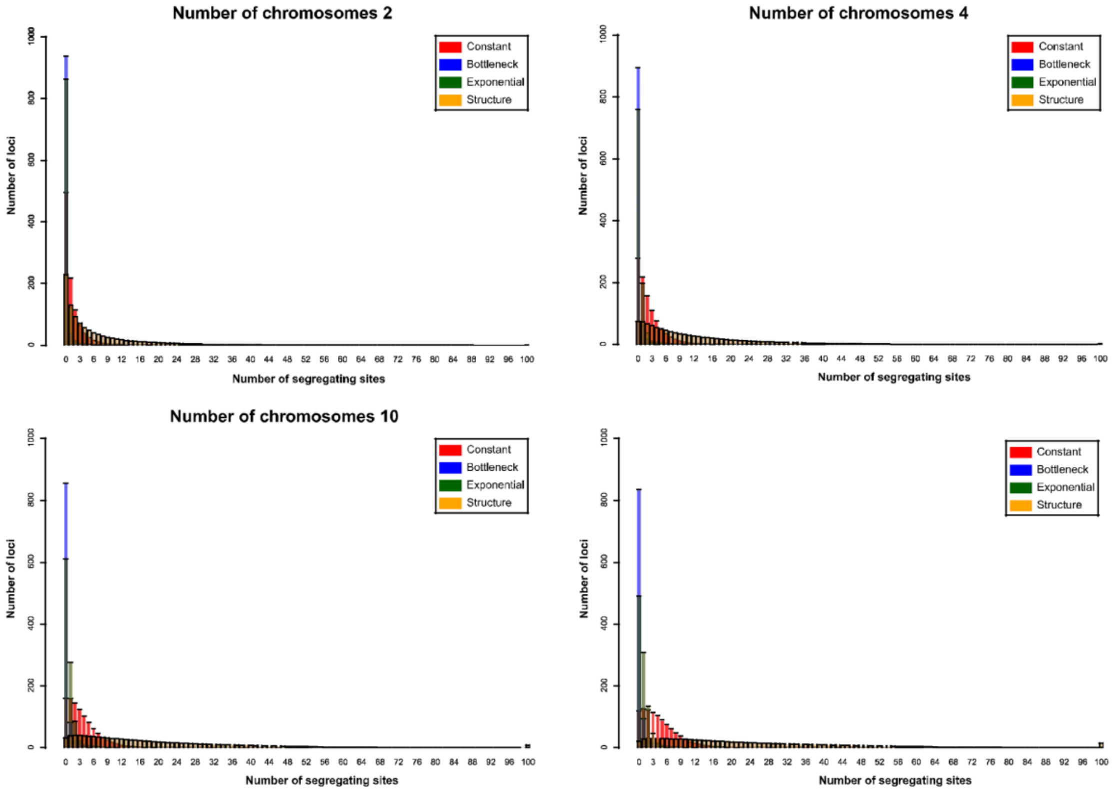


Fig S4. Proportion of True Positives for the two-populations models summarized through the *unfolded SFS*. The plots have the same features of Fig 1.

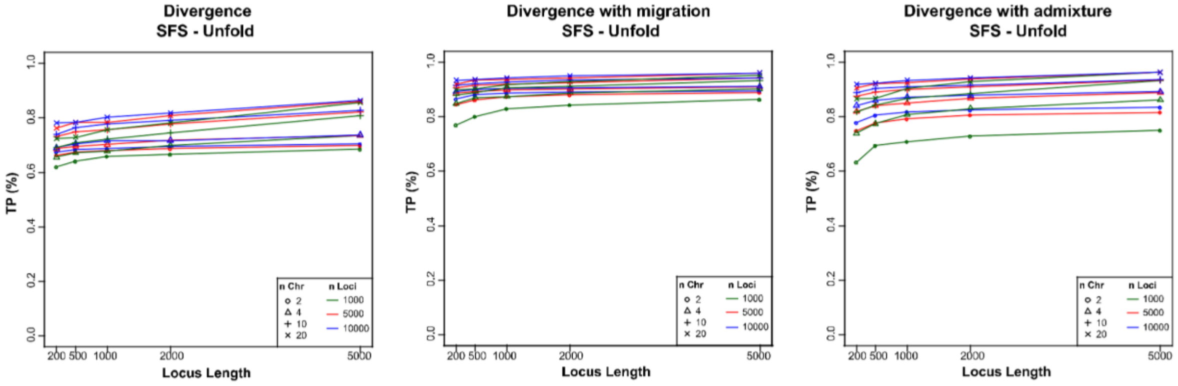


Fig S5. Proportion of True Positives for the multi-populations models summarized through the *unfolded SFS*. The plots have the same features of Fig 1.

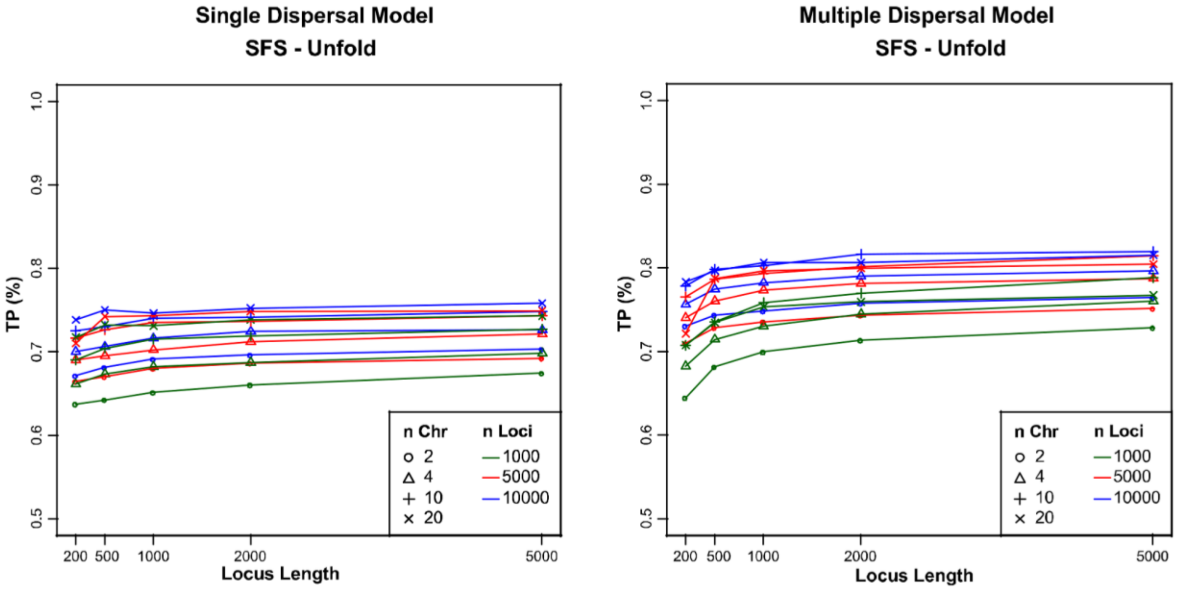


Fig S6. Proportion of True Positives for the MDM a function of the time span between the divergence time of the African ghost populations and the second exit (Delta tdYGOOA2). The time difference between the divergence time of the African ghost populations and the second exit from Africa is on the x-axes and it is expressed in years (considering a generation time of 29 years). Each plot reports the results for a different locus length.

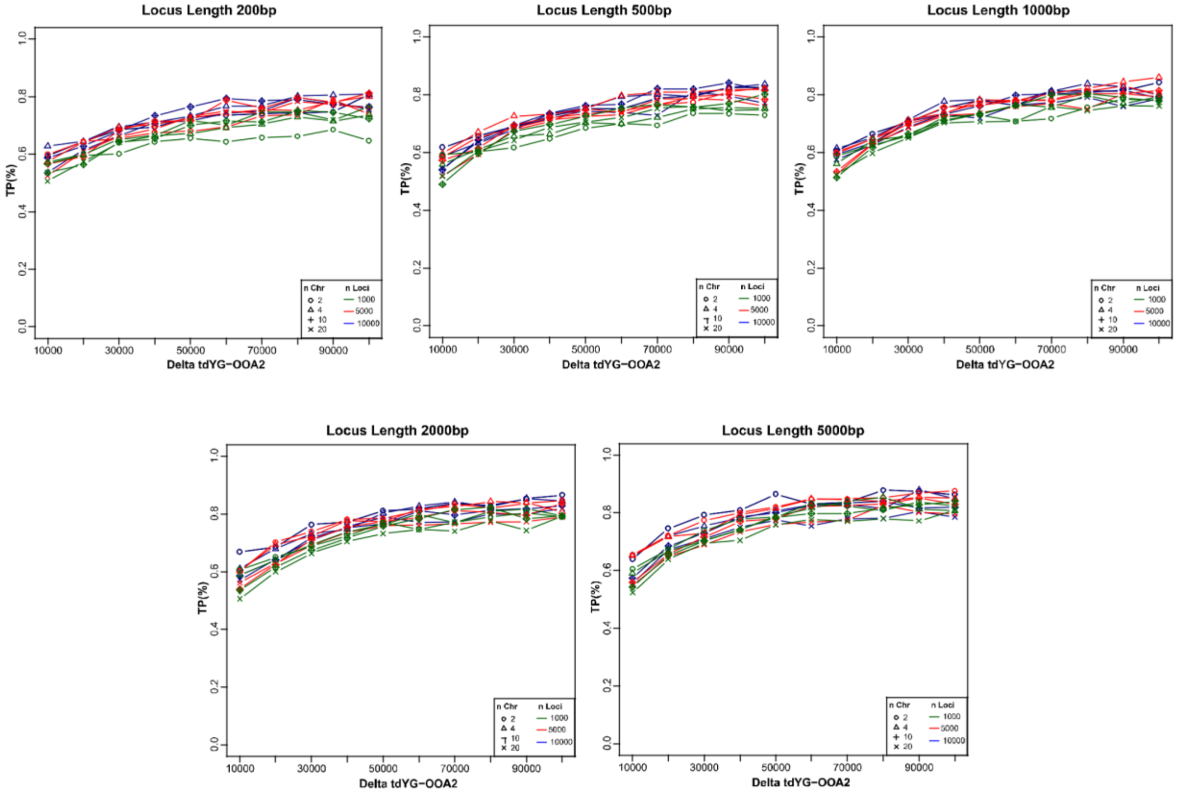


Fig S7. Principal Component Analysis (PCA) of the simulated and observed data. PCA of the simulated data generated under the MDM (orange points) and A) the observed data from Pagani et al (2016) (black points, 6 observed datasets)**,** B) the observed data from Malaspinas et al (2016) (black points, 25 observed datasets).

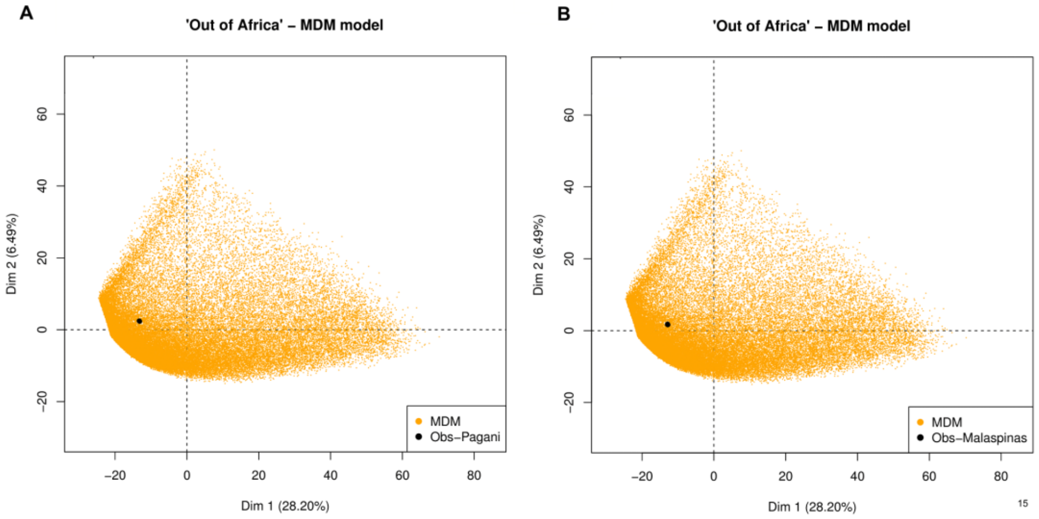


Fig S8. Principal Component Analysis (PCA) of the simulated and observed data. PCA of the simulated data generated under the model 1a (orange points) and the observed data from Nater et al. (2017) (black point).

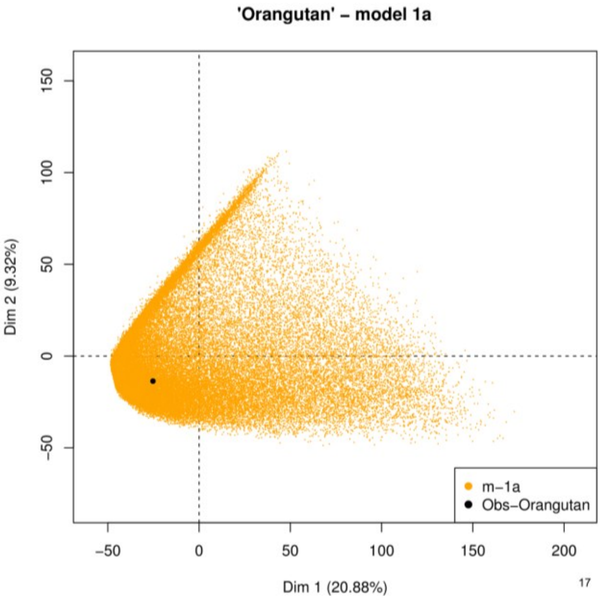


Fig S9. Proportion of True Positives for the one-population models summarized through the full *FDSS* distributions (A) and the first two moments of the *FDSS* distributions (B). The plots have the same features of Fig 1.

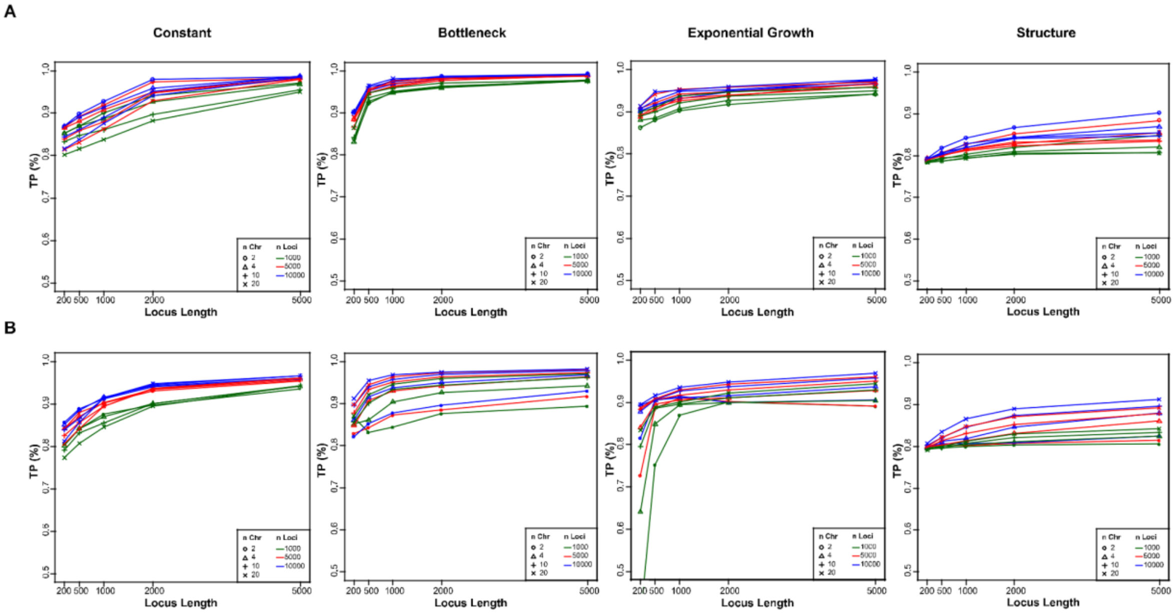


Fig S10. Proportion of True Positives for the two-populations models summarized through full *FDSS* distributions (A) and the first two moments of the *FDSS* distributions (B). The plots have the same features of Fig 1.

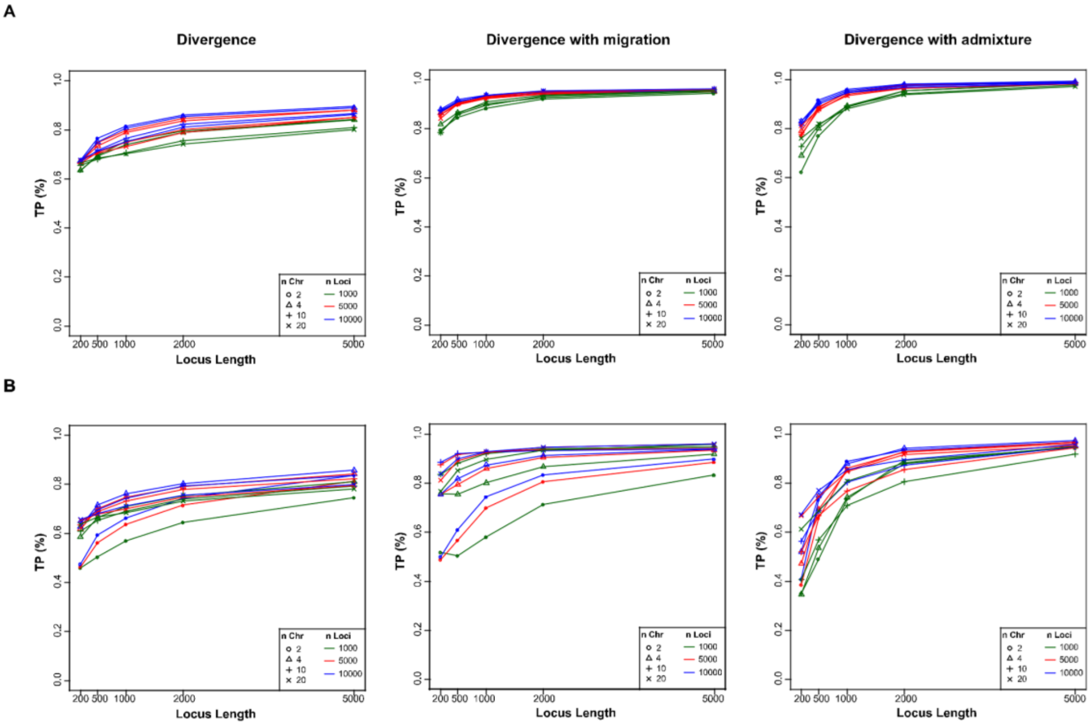


Fig S11. Proportion of True Positives for the multi-populations models summarized through full *FDSS* distributions (A) and the first two moments of the *FDSS* distributions (B). The plots have the same features of Fig 1.

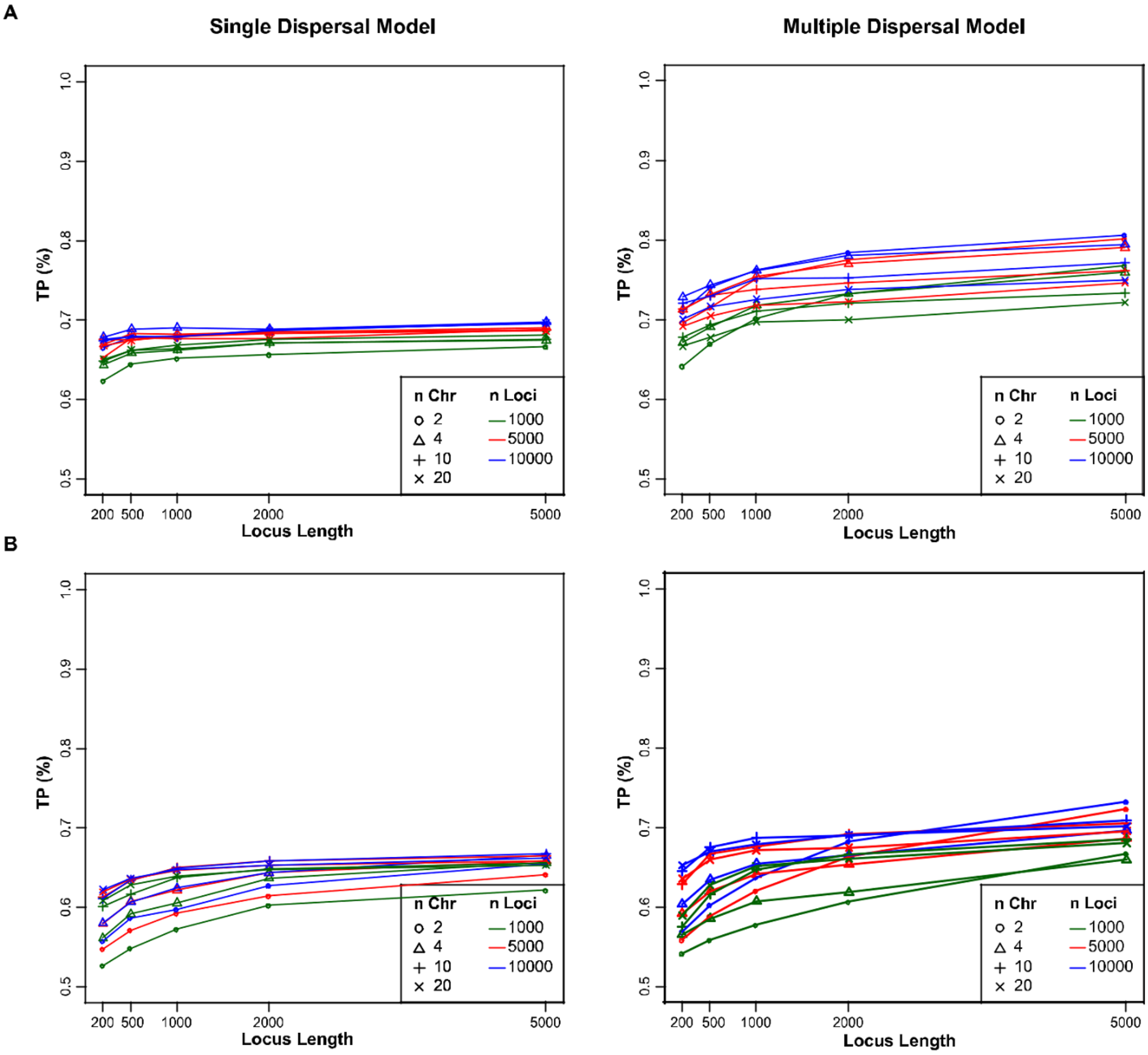


Table S1. Demographic parameters and prior distributions of One-Population models. Mutation and Recombination rates are expressed per nucleotide per generation.

Demographic Parameters	**Prior Distributions**
Effective population size (*N_1*)	Uniform {500:50,000}
Intensity bottleneck (*i*)	Uniform {10:100}
Intensity exponential growth (*i*)	Uniform {10:100}
Time bottleneck (*T*)	Uniform {100:20,000}
Time exponential growth (*T*)	Uniform {100:20,000}
Number of demes (*d*)	Uniform {2:10}
Migration rate (*m*)	Exponential {0.1}
Mutation rate	$1x10^{-8}$ {Fixed}
Recombination rate	$1x10^{-8}$ {Fixed}

Table S2. Demographic parameters and prior distributions of Two-Populations models. Mutation and Recombination rates are expressed per nucleotide per generation. Time is in generations. In the simulation step we considered a *Tadm* value only if (*Tsep*-*Tadm*)/*Tsep* was between 0.2 and 0.8.

Demographic Parameters	**Prior Distributions**
Effective population size (N_{anc}, N_1, N_2)	Uniform {500:50,000}
Time split (T_{sep})	Uniform {300:20,000}
Migration rate (m_{12}, m_{21})	Exponential {0.1}
Time admixture (T_{adm})	Uniform {50:2,500}
Admixture rate (adm_{12}, adm_{21})	Uniform {0.05:0.20}
Mutation rate	$1x10^{-8}$ {Fixed}
Recombination rate	$1x10^{-8}$ {Fixed}

Table S3. Demographic parameters and prior distributions of multi-populations models: Single Dispersal model. Migration and admixture rates are expressed per generation, times in years. We cosidered a generation time of 29 years as in Malaspinas et al. (2016). Per nucleotide per generation mutation and recombination rates are fixed as in Malaspinas et al. (2016).

Demographic Parameters	**Prior Distributions**
Effective population size (Ne)	Uniform {500:50,000}
Migration rate (ModernPop)	Uniform {10^{-6}: 10^{-3}}
Time split Africa-Ghost	Uniform {40,000:145,000}yrs
Duration time bottleneck	2,900yrs
Intensity bottleneck	Uniform {2:100}
Time split Eurasia/Papua-Ghost(OOA)	Uniform {35,000:EndBottlGhost}yrs
Time split Europe-Asia	Uniform {20,000:30,000}yrs
Time admixture Nea-Asia	Uniform {20,000:Time split Europe-Asia}yrs
Time admixture Nea-Eurasia	Uniform {Time split Europe-Asia:EndbottlOOA}yrs
Time admixture Den-Papua	Uniform {30,000:EndBottlOOA}yrs
Time admixture Arc-Papua	Uniform {Time admix. Den-Papua: EndBottl.OOA}yrs
Time admixture Nea-Ghost	Uniform {Time split. Eurs/Pap-Ghost:EndBottl.Ghost}yrs
Admixture rate	Uniform {10^{-3}:10^{-1}}
Time split Nea-NeaR	110,000yrs {Fixed}
Time split Den-DenR	393,000yrs {Fixed}
Time split Den-Nea	495,000yrs {Fixed}
Time split Arc-Nea/Den	580,000yrs {Fixed}
Time split Ancient-Modern	638,000yrs {Fixed}
Sample Time Neanderthal	85,735yrs {Fixed}
Sample Time Denisova	67,570yrs {Fixed}
Mutation rate	$1.25x10^{-8}$ {Fixed}
Recombination rate	$1.12x10^{-8}$ {Fixed}

Table S4. Demographic parameters and prior distributions of multi-populations models: Multiple Dispersals model. Migration and admixture rates are expressed per generation, times in years. We cosidered a generation time of 29 years as in Malaspinas et al. (2016). Per nucleotide per generation mutation and recombination rates are fixed as in Malaspinas et al. (2016).

Demographic Parameters	**Prior Distributions**
Effective population size (Ne)	Uniform {500:50,000}
Migration rate (ModerPop)	Uniform {10^{-6}: 10^{-3}}
Time split Africa-Ghosts(1 and 2)	Uniform {40,000:145,000}yrs
Duration time bottleneck	2,900yrs
Intensity bottleneck	Uniform {2:100}
Time split Papua-Ghost1	Uniform {40,000:Time split. Africa-Ghost1}yrs
Time split Eurasia-Ghost2	Uniform {35,000:EndBott.Papua}yrs
Time split Europe-Asia	Uniform {20,000:EndBott.Eurasia}yrs
Time admixture Nea-Asia	Uniform {20,000:Time split Europe-Asia}yrs
Time admixture Nea-Eurasia	Uniform {Time split Europe-Asia:EndBott.Eurasia}yrs
Time admixture Den-Papua	Uniform {30,000: EndBott.Papua}yrs
Time admixture Arc-Papua	Uniform {Time admix. Den-Papua:EndBott.Papua}yrs
Time admixture Nea-Ghost2	Uniform {Time split Euras-Ghost2:Time split Africa-Ghost2}yrs
Admixture rate	Uniform {10^{-3}:10^{-1}}
Time split Nea-NeaR	110,000yrs {Fixed}
Time split Den-DenR	393,000yrs {Fixed}
Time split Den-Nea	495,000yrs {Fixed}
Time split Arc-Nea/Den	580,000yrs {Fixed}
Time split Ancient-Modern	638,000yrs {Fixed}
Sample Time Neanderthal	85,735yrs {Fixed}
Sample Time Denisova	67,570yrs {Fixed}
Mutation rate	$1.25x10^{-8}$ {Fixed}
Recombination rate	$1.12x10^{-8}$ {Fixed}

Table S5. Genomes used for the comparison of SDM and MDM using real data.

Neanderthal	AltaiNea	Prufer *et al.* (2014)
Denisova	DenisovaPinky	Mayer *et al.* (2012)
African	CongPy1	Pagani *et al.* (2016)
European	Est1	Pagani *et al.* (2016)
Asian	VietN1	Pagani *et al.* (2016)
Papuan	Koinb1	Pagani *et al.* (2016)
Papuan	Koinb2	Pagani *et al.* (2016)
Papuan	Koinb3	Pagani *et al.* (2016)
Papuan	Kosip1	Pagani *et al.* (2016)
Papuan	Kosip2	Pagani *et al.* (2016)
Papuan	Kosip3	Pagani *et al.* (2016)
Papuan	EGAN00001279031	Malaspinas *et al.* (2016)
Papuan	EGAN00001279039	Malaspinas *et al.* (2016)
Papuan	EGAN00001279047	Malaspinas *et al.* (2016)
Papuan	EGAN00001279054	Malaspinas *et al.* (2016)
Papuan	EGAN00001279032	Malaspinas *et al.* (2016)
Papuan	EGAN00001279040	Malaspinas *et al.* (2016)
Papuan	EGAN00001279048	Malaspinas *et al.* (2016)
Papuan	EGAN00001279033	Malaspinas *et al.* (2016)
Papuan	EGAN00001279041	Malaspinas *et al.* (2016)
Papuan	EGAN00001279049	Malaspinas *et al.* (2016)
Papuan	EGAN00001279034	Malaspinas *et al.* (2016)
Papuan	EGAN00001279042	Malaspinas *et al.* (2016)
Papuan	EGAN00001279050	Malaspinas *et al.* (2016)
Papuan	EGAN00001279035	Malaspinas *et al.* (2016)
Papuan	EGAN00001279043	Malaspinas *et al.* (2016)
Papuan	EGAN00001279051	Malaspinas *et al.* (2016)
Papuan	EGAN00001279036	Malaspinas *et al.* (2016)
Papuan	EGAN00001279044	Malaspinas *et al.* (2016)
Papuan	EGAN00001279052	Malaspinas *et al.* (2016)
Papuan	EGAN00001279037	Malaspinas *et al.* (2016)
Papuan	EGAN00001279045	Malaspinas *et al.* (2016)
Papuan	EGAN00001279053	Malaspinas *et al.* (2016)
Papuan	EGAN00001279038	Malaspinas *et al.* (2016)
Papuan	EGAN00001279046	Malaspinas *et al.* (2016)
Papuan	EGAN00001279055	Malaspinas *et al.* (2016)

Table S6. Genomes used for the comparison of the four Orangutan evolutionary scenarios.

Pongo abelii	Elsi	Santpere *et al.* (2013)	27.39x
Pongo abelii	Suma	Nater *et al.* (2017)	25.27x
Pongo tapanuliensis	Afa	Nater *et al.* (2017)	16.92x
Pongo pygmaeus	Claus	Nater *et al.* (2017)	29.71x
Pongo pygmaeus	Panjul	Nater *et al.* (2017)	30.13x
Pongo pygmaeus	Kala	Nater *et al.* (2017)	31.06x
Pongo pygmaeus	Kajan	Nater *et al.* (2017)	22.39x

Table S7. Demographic parameters and prior distributions for Model 1a. Migration rates are expressed per generation, times in years. We used a generation time of 25 years as in Nater et al. (2017). The per nucleotide per generation mutation rate is fixed as in Nater et al. (2017).

Demographic Parameters	**Prior Distributions**
Effective population size (Ne-ModernPop)	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate (Intra BO)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (Intra NT)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (ST-strucNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancBO)	Loguniform {10^{-6}: 10^{-2}}
Time sep. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. BO-ST	Uniform {400,000:1,500,000}yrs
Time Stop migration (ST-ancBO)	Uniform {TimeBottlBO:TimeSep. BO-ST}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep. ancNT-ST	Uniform {1,500,000:4,000,000}yrs
Mutation rate	1.5x10^{-8} {Fixed}

Table S8. Demographic parameters and prior distributions for Model 2a. Migration rates are expressed per generation, times in years. We used a generation time of 25 years as in Nater et al. (2017). The per nucleotide per generation mutation rate is fixed as in Nater et al. (2017).

Demographic Parameters	**Prior Distributions**
Effective population size (Ne-ModernPop)	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate (Intra BO)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (Intra NT)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (ST-strucNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancBO)	Loguniform {10^{-6}: 10^{-2}}
Time sep. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. BO-ST	Uniform {1,500,000:4,000,000}yrs
Time Stop migration (ST-ancBO)	Uniform {TimeBottlBO:TimeSep. BO-ST}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep. ancNT-ST	Uniform {TimeStrucNT:TimeSep. BO-ST}yrs
Mutation rate	1.5x10^{-8} {Fixed}

Table S9. Demographic parameters and prior distributions for Model 1b. Migration rates are expressed per generation, times in years. We used a generation time of 25 years as in Nater et al. (2017). The per nucleotide per generation mutation rate is fixed as in Nater et al. (2017).

Demographic Parameters	**Prior Distributions**
Effective population size (Ne-ModernPop)	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate (Intra BO)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (Intra NT)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (ST-strucNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancBO)	Loguniform {10^{-6}: 10^{-2}}
Time sep. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. BO-ST	Uniform {400,000:1,500,000}yrs
Time Stop migration (ST-ancBO)	Uniform {TimeBottlBO:TimeSep. BO-ST}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep ST-ancNT	Uniform {1,500,000:4,000,000}yrs
Mutation rate	$1.5\text{x}10^{-8}$ {Fixed}

Table S10. Demographic parameters and prior distributions for Model 2b. Migration rates are expressed per generation, times in years. We used a generation time of 25 years as in Nater et al. (2017). The per nucleotide per generation mutation rate is fixed as in Nater et al. (2017).

Demographic Parameters	**Prior Distributions**
Effective population size (Ne-ModernPop)	Uniform {300:32,000}
NeStruc NT	Uniform {NeModNT:320,000}
NeAnc NT	Uniform {1,000:100,000}
NeAnc ST	Uniform {NeModST:100,000}
NeAnc BO	Uniform {NeModBO:320,000}
Migration rate (Intra BO)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (Intra NT)	Loguniform {10^{-4}: 10^{-1}}
Migration rate (ST-strucNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancNT)	Loguniform {10^{-5}: 10^{-1}}
Migration rate (ST-ancBO)	Loguniform {10^{-6}: 10^{-2}}
Time sep. modern BO	Uniform {8,750:400,000}yrs
Duration time bottleneck BO	Uniform {250:100,000}yrs
Time sep. ST-BO	Uniform {1,500,000:4,000,000}yrs
Time Stop migration (ST-ancBO)	Uniform {TimeBottlBO:TimeSep. ST-BO}yrs
Time bottleneck ST and strucNT	Uniform {250:100,000}yrs
Time structure NT	Uniform {100,000:1,500,000}yrs
Time sep. ST-ancNT	Uniform {TimeStrucNT:TimeSep. ST-BO}yrs
Mutation rate	$1.5\text{x}10^{-8}$ {Fixed}

Table S11. Model selection results using Papuan individuals from Pagani et al. (2016). The first column represents the id of the Papuan sample used in that comparison as reported in the dataset. The second column shows the model selected by the ABC procedure. The third and the fourth columns represent the proportion of votes assigned to SDM and MDM by the RF algorithm. The last column is the posterior probability of the most supported model.

ID_Papuan	Selected model	Votes SDM	Votes MDM	Post.proba
Koinb1	MDM	0.48	0.52	0.761
Koinb2	MDM	0.43	0.57	0.741
Koinb3	MDM	0.448	0.552	0.765
Kosip1	MDM	0.416	0.584	0.740
Kosip2	MDM	0.43	0.57	0.747
Kosip3	MDM	0.436	0.564	0.735

Table S12. Model selection results using Papuan individuals from Malaspinas et al. (2016). The first column represents the id of the Papuan sample used in that comparison as reported in the dataset. The second column shows the model selected by the ABC procedure. The third and the fourth columns represent the proportion of votes assigned to SDM and MDM by the RF algorithm. The last column is the posterior probability of the most supported model.

ID_Papuan	Selected model	Votes SDM	Votes MDM	Post.proba
1279031	MDM	0.264	0.736	0.734
1279039	MDM	0.24	0.76	0.737
1279047	MDM	0.244	0.756	0.726
1279054	MDM	0.246	0.754	0.704
1279032	MDM	0.234	0.766	0.721
1279040	MDM	0.248	0.752	0.699
1279048	MDM	0.238	0.762	0.728
1279033	MDM	0.248	0.752	0.730
1279041	MDM	0.234	0.766	0.705
1279049	MDM	0.24	0.76	0.703
1279034	MDM	0.252	0.748	0.700
1279042	MDM	0.248	0.752	0.711
1279050	MDM	0.254	0.746	0.720
1279035	MDM	0.25	0.75	0.697
1279043	MDM	0.234	0.766	0.739
1279051	MDM	0.256	0.744	0.689
1279036	MDM	0.246	0.754	0.703
1279044	MDM	0.254	0.746	0.719
1279052	MDM	0.252	0.748	0.723
1279037	MDM	0.25	0.75	0.732
1279045	MDM	0.266	0.734	0.719
1279053	MDM	0.246	0.754	0.724
1279038	MDM	0.244	0.756	0.707
1279046	MDM	0.25	0.75	0.710
1279055	MDM	0.248	0.752	0.700

Paper II

genes

MDPI

Article

A Revised Model of Anatomically Modern Human Expansions Out of Africa through a Machine Learning Approximate Bayesian Computation Approach

Maria Teresa Vizzari, Andrea Benazzo, Guido Barbujani and Silvia Ghirotto *

Department of Life Sciences and Biotechnology, University of Ferrara, 44121 Ferrara, Italy;
mariateresa.vizzari@unife.it (M.T.V.); andrea.benazzo@unife.it (A.B.); guido.barbujani@unife.it (G.B.)
* Correspondence: silvia.ghirotto@unife.it

Received: 5 November 2020; Accepted: 14 December 2020; Published: 16 December 2020

Abstract: There is a wide consensus in considering Africa as the birthplace of anatomically modern humans (AMH), but the dispersal pattern and the main routes followed by our ancestors to colonize the world are still matters of debate. It is still an open question whether AMH left Africa through a single process, dispersing almost simultaneously over Asia and Europe, or in two main waves, first through the Arab Peninsula into southern Asia and Australo-Melanesia, and later through a northern route crossing the Levant. The development of new methodologies for inferring population history and the availability of worldwide high-coverage whole-genome sequences did not resolve this debate. In this work, we test the two main out-of-Africa hypotheses through an Approximate Bayesian Computation approach, based on the Random-Forest algorithm. We evaluated the ability of the method to discriminate between the alternative models of AMH out-of-Africa, using simulated data. Once assessed that the models are distinguishable, we compared simulated data with real genomic variation, from modern and archaic populations. This analysis showed that a model of multiple dispersals is four-fold as likely as the alternative single-dispersal model. According to our estimates, the two dispersal processes may be placed, respectively, around 74,000 and around 46,000 years ago.

Keywords: approximate Bayesian computation; demographic history; human evolution; migration; machine learning; random forest; whole-genome data

1. Introduction

Levels and patterns of genome diversity reflect past demographic processes, and a crucial turning point in our demographic history is the expansion of anatomically modern humans (AMH) from Africa. Some aspects of this process seem rather well established. First, what is often called the ancestral African population should not be regarded as a single, biologically homogeneous unit, but as a structured population hosting regional diversity [1]. Second, the AMH expansion was accompanied by the disappearance of preexisting archaic human forms [2,3] Third, a variable component of the genomes of most present populations—always small, seldom zero—comes from anatomically archaic ancestors [4].

Conversely, there is disagreement over other aspects of the AMH expansion out of Africa, such as the number of major dispersal events, their timing, and the geographical routes followed by migrating people. Groups of AMH may have left Africa more than 100,000 years ago [5], but genetic evidence suggests that such early phenomena were not successful and did not lead to the establishment of permanent non-African populations. One expansion left traces in modern genomes; it took place between 60,000 and 50,000 years ago, along a Northern route in the Nile valley and across the

Near East (see e.g., [6–8]). However, based on cranial morphology, Lahr and Foley [9] proposed an additional, earlier migration through a Southern route, from the Horn of Africa into the Arab peninsula, Southern Asia, and Australo-Melanesia. We shall refer to these alternative models as Single Dispersal (SD) and Multiple Dispersal (MD) hypotheses. The MD hypothesis found support in several studies, and notably in a comparison of cranial and DNA diversity data [10] but broader genomic analyses gave contradictory results. Tassi and colleagues [11] and, to a lesser extent, Pagani et al. [12] described patterns consistent with two dispersal processes, the first one overlapping in time with the proposed early Southern exit from Africa [11]. On the other hand, two studies of different genomic datasets concluded that there is little [4] or no evidence [13] for such an early dispersal process, and hence that AMH either left Africa in a single major migrational wave, or perhaps in several waves, but then only one of them contributed to the ancestry of modern populations.

Malaspinas et al. [13] conclusion in favor of SD was not really based on an explicit comparison between models. In their paper, indeed, they considered an MD model in which East Asians and Europeans have a more recent common ancestor than Aboriginal Australians and East Asians. and they estimated the models' parameters. The evidence supporting the SD model came from the overlapping estimation for the divergence times of the ancestors of Aboriginal Australians and Eurasians.

This non-straightforward procedure was due to an implicit limitation of the composite likelihood method they applied, in which model selection may be performed through likelihood ratio tests (LRT) or by the Akaike Information Criterion (AIC; [14,15]). LRT and AIC 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.

To understand which model, SD or MD, better accounts for the current levels of genome diversity, in this study we formally compare them by a recently developed Approximate Bayesian Computation framework, based on the study of the observed Frequency Distributions of four categories of Segregating Sites for pair of populations (FDSS) [16]. ABC is a powerful and flexible framework, based on computer simulations, to perform model selection and estimate models' parameters. In its original formulation [17,18] the ABC algorithm suffered from two main issues, related to the simulation effort and to the number of summary statistics used to summarize the data. These issues limited the possibility to use ABC for the analysis of complex demographic histories and/or large datasets. In 2015, the introduction of a paradigm shift in the ABC model selection procedure based on a Machine Learning approach called Random Forest (ABC-RF, [19]), allowed to overcome the above-cited limitations and paved the ground for the application of ABC to the study of complex models through the analysis of complete genomes. Under ABC-RF, the model selection procedure 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 number of simulations necessary to obtain reliable estimates passed from a few million to a few thousand; the informative statistics are systematically extracted from the pool used to summarize the data. In 2018, a similar approach, based on a machine-learning tool of regression RF, has been developed for parameter estimation [20]. In [16] we showed that the ABC-RF algorithm, combined with the inferential power provided by the FDSS, can be satisfactorily exploited to estimated past population dynamics even in case of complex demographic histories, thus making the approach particularly suitable to the analysis of SD and MD models.

Under both SD and MD models, the structure of the past populations is the same, but the tree topologies differ in that they assume, respectively, one ancestral population for the SD model, and two ancestral populations leaving Africa at different times for the MD model. As the Australo-Melanesian represent the population that might carry the signal of the first wave of migrations out of the African continent and also, to make sure that the different results obtained by [12,13] were not due to differences in the Australo-Melanesian samples available, we repeated our analyses considering genomes coming from both studies, obtaining results that seem consistent and informative.

2. Materials and Methods

2.1. The FDSS

We summarized the data through the FDSS, i.e., the frequency distributions of the four mutually exclusive categories of segregating sites for pair of populations (i.e., private polymorphisms in either population, shared polymorphisms, and fixed differences [21]). This statistic proved to be powerful for reconstructing even a complex series of demographic processes [16]. The FDSS is calculated considering each genome analyzed as subdivided into a certain number of independent fragments of a certain length, and for each fragment, the number of sites belonging to each of the four above-mentioned categories is counted. The final vector of summary statistics is thus composed by the truncated frequency distribution of fragments having from 0 to n segregating sites in each category, for each pair of populations considered. We fixed the maximum number of segregating sites in a locus of a certain length to 100, and hence the last category contains all the observations higher than 100.

We calculated the FDSS using a python script (available on Github https://github.com/anbena/ABC-FDSS) [16]. The ABC-RF model selection estimates have been obtained using the function *abcrf* from the package *abcrf* and employing a forest of 500 classification trees, a number suggested providing the best trade-off between computational efficiency and statistical precision [19]. Before proceeding with the model selection procedure, we computed the confusion matrices and evaluated the out-of-bag classification error (CE) and the proportion of True Positives (1-CE), which are representative of the power of the whole inferential procedure. The ABC-RF parameters estimation on the most supported models have been performed through the function *regAbcrf* from the package *abcrf* and employing a forest of 500 regression trees. An outline of our entire workflow is reported in Figure S1.

2.2. Simulated Models of Anatomically Modern Humans Expansion Out of Africa

We tested two alternative models of expansion of anatomically modern humans out of the African continent (Figure 1), both sharing the same structure for the archaic groups, but differing for the relationships among modern populations. To design the models, we followed the parametrization proposed by [13], with some modifications detailed below. The first model (SD) indeed accounts for a single dispersal from Africa giving rise to both modern Eurasians and Australo-Melanesians, the second model (MD) accounts for two different waves of migrations, from two different African source populations, giving rise, first, to the modern Australo-Melanesians and, later to the modern Eurasians. The archaic groups consist of three Denisovan populations, two Neanderthal populations, and an unknown archaic population ancestral to both Neandertals and Denisovans. We explicitly considered admixture pulses from archaic to modern populations: a pulse from the archaic unknown population to Australo-Melanesians (as reported in [22]), two pulses from two different Denisovan populations to Asians and Australo-Melanesians [23,24], two pulses from the same Neandertal population to modern humans just after the separation between African and non-African populations, and to the ancestor of all Eurasians [25–27]. Both models account for the presence of a Basal European population, as described in [28–30]. This (so far, unknown) population contributed genes to modern Europeans, possibly diluting the contribution of archaic Neandertal variants in European genomes. The SD and MD models have 45 and 50 free parameters (i.e., parameters whose values are defined by prior distributions), respectively. The prior distributions associated with these parameters were set following what was proposed in the recent literature by [13,23,30], and are reported in Tables S1 and S2. We considered a generation time of 29 years, and we fixed the mutation rate at 1.25×10^{-8} bp/generation [31] and the intra-locus recombination rate at 1.12×10^{-8}, all values as in [13].

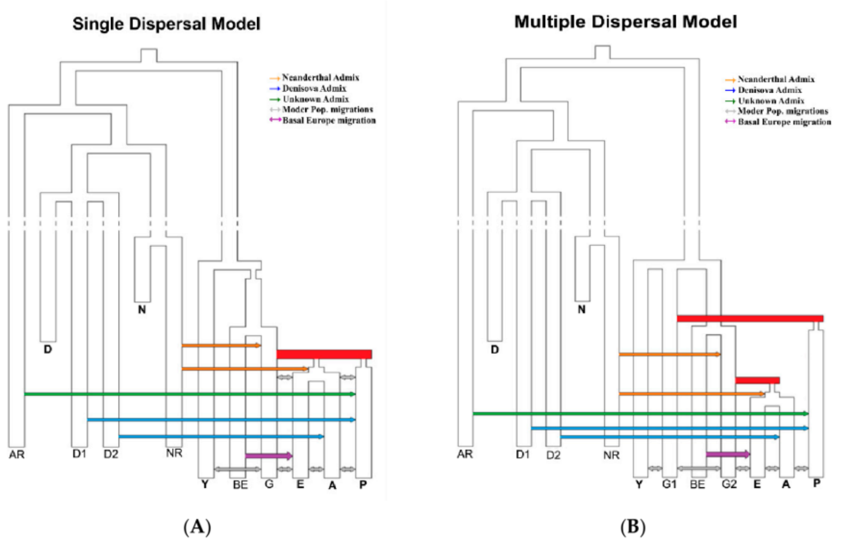


Figure 1. Demographic models compared: Single Dispersal (**A**) and Multiple Dispersals (**B**). AR: unknown archaic population; D-D1-D2: Denisovan groups; N-NR: Neandertal and Neandertal related groups; Y: African population; G1-G2: ghost populations; BE: Basal Europe population; E: European population; A: Asian population; P: Australo-Melanesian population.

We performed 20,000, 50,000, and 100,000 simulations for each model with *ms* [32], to evaluate the Prior Error Rate and identify the optimum number of simulations to use. At each iteration, we sampled six diploid genomes, one Neandertal, one Denisova, one African, one European, one Asian, and one Papuan. The FDSS was calculated from 10,000 independent genomic fragments of 500 bp length.

2.3. Observed Genomic Data

We analyzed the high-coverage genomes of Denisova [33] and Neandertal [26], together with worldwide modern human samples from [12]. 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, [34]). 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%*; [26,33]). We also included in the analysis of the 25 Papuan individuals published by [13]. For these individuals, 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* [35], 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 [13] using *vcflib* and *bcftools* [35]. The complete set of samples used for the comparison between SD and MD are reported in Table S3.

In each models' comparison, we evaluated the genomic variation of one Denisova, one Neandertal, one African (Congo-pygmies), one European (Estonians), one Asian (Vietnamese), and one Australo-Melanesian (Papuans). We decided to restrict the analysis to one high coverage diploid genome per population since previous extensive analyses showed that a single individual sampled per population has a comparable discrimination power as twenty chromosomes [16]. However, to ensure the consistency of the results, we performed several model selection procedures (a) taking into account at each run one out of six Papuans from [12] or one of 25 Papuans from [13]; (b) considering alternative individuals as representative of African, European, and Asian populations (Table S4).

2.4. Assessment of the Quality of the Parameters Estimated

One of the most interesting features of ABC is its high flexibility for model checking, i.e., for assessing the quality of the estimates inferred from real data. This is mainly achieved through the analysis of pseudo-observed data (pods), i.e., simulated datasets generated under known conditions. To determine whether the observed data would contain enough information to estimate parameters of the multi-dimensional model tested, we exploited 1000 pods, each generated from the most supported model (i.e., the MD model) and through a known combination of demographic parameters. Using these pods, for each parameter we calculated the following indices:

- The coefficient of determination (R^2). R^2 is the fraction of variance of the parameters explained by the summary statistics used to build the regression model. In the absence of an established threshold value, there is a general agreement that when $R^2 < 0.10$, the summary statistics do not convey enough information about the parameter estimates [36].
- The relative bias. To calculate the relative bias, we estimated the parameters for each pod with the same approach used for the observed data. The bias depends on the sum of differences between the 1000 estimates of each parameter thus obtained and the known (true) value, and it is calculated as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\theta_i - \theta}{\theta}$$

  where θ_i is the estimator of the parameter θ (true value), and n is the number of pods used (1000 in our case). Because bias is relative, a value of 1 corresponds to a bias equal to 100% of the true value.
- The root mean square error (RMSE). To calculate the RMSE we re-estimated parameters using pods. The RMSE depends the sum of squared differences between the 1000 estimates of each parameter thus obtained and the true value and it is calculated as:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \theta)^2}$$

- The factor 2, representing the proportion of the 1000 estimated median values lying between 50% and 200% of the true value.
- The 50% and 90% coverage, defined as the proportion of times that the known value lies within the 50% and the 90% credible interval of the 1000 estimates.

3. Results

3.1. Model Selection

Table 1 and Table S5 show the results of the power check of the comparison between SD and MD. Predictably, the Prior Error rate, which indicates the global quality of the ML classifier, decreases for increasing numbers of simulations in the reference table (from 20,000 to 100,000); for this reason, we decided to use 100,000 simulations for the subsequent analyses. The proportion of True Positives, that is the proportion of times the SD or the MD model is correctly recognized by the model selection procedure, is above 70% for both SD and MD, with a mean posterior probability associated with the true demography of about 75%.

Table 1. Power test for model comparison using a reference table with 100,000 simulations per model.

Prior Err. Rate	True Positive SD	True Positive MD	Post. Prob. SD	Post. Prob. MD
0.26	0.73	0.75	0.75	0.73

Table 2 and Table S4 show the results of the model selection. Regardless of the Papuan individual considered, and the combination of non-Australo-Melanesian tested, the model selection analyses supported the MD model as the scenario best explaining the recent evolution of anatomically modern humans out of Africa, with probabilities ranging from 78 to 84%.

Table 2. Model Selection results using Papuan individuals from [12,13]. In the first column are reported the ID of the Papuan samples used for the model choice. The second column shows the model selected by the ABC procedure. In the third and the fourth columns are reported the votes assigned to the SD and MD models by the Random-Forest algorithm. The last column shows the posterior probabilities associated with the most supported model. The samples with the highest posterior probabilities (in bold) were selected to perform the parameter estimation of the MD model.

ID_Individual	Selected Model	Votes SD	Votes MD	Post. Prob.
EGAN00001279031	MD	94	406	0.822
EGAN00001279039	MD	86	414	0.806
EGAN00001279047	MD	111	389	0.798
EGAN00001279054	MD	128	372	0.809
EGAN00001279032	**MD**	**90**	**410**	**0.825**
EGAN00001279040	MD	113	387	0.784
EGAN00001279048	MD	99	401	0.805
EGAN00001279033	MD	108	392	0.791
EGAN00001279041	MD	111	389	0.797
EGAN00001279049	MD	126	374	0.789
EGAN00001279034	MD	150	350	0.797
EGAN00001279042	MD	109	391	0.791
EGAN00001279050	MD	111	389	0.797
EGAN00001279035	MD	108	392	0.799
EGAN00001279043	MD	97	403	0.802
EGAN00001279051	MD	117	383	0.786
EGAN00001279036	MD	136	364	0.778
EGAN00001279044	MD	109	391	0.784
EGAN00001279052	MD	100	400	0.815
EGAN00001279037	MD	96	404	0.800
EGAN00001279045	MD	148	352	0.787
EGAN00001279053	MD	100	400	0.796
EGAN00001279038	MD	91	409	0.811
EGAN00001279046	MD	104	396	0.781
EGAN00001279055	MD	138	362	0.787
Koinb1	MD	165	335	0.810
Koinb2	MD	129	371	0.811
Koinb3	MD	175	325	0.820
Kosip1	MD	152	348	0.818
Kosip2	MD	136	364	0.788
Kosip3	**MD**	**123**	**377**	**0.830**

3.2. Parameters Estimation

Once identified the MD as the most probable model, we moved to estimate its parameter values maximizing the fit between observed and simulated genomic data. To do this, we exploited the recently developed ML method, based on a regression RF approach [20]. As detailed in [20], a faithful estimation of parameters' posterior distribution may be now achieved with a reduced number of simulations (i.e., a few thousand; we used 100,000 simulations), making it feasible to also perform an accurate assessment of the quality of the parameters estimated using pods.

Parameters were estimated from two observed datasets (one with a Papuan individual from [13] and one with a Papuan individual from [12]), those which produced the highest value of posterior probability for the MD model in the model selection (Tables 3 and 4). The posterior plots and the definition of the parameter's acronyms are reported in Supplementary Materials (Figures S2–S10, Table S6). The R^2, the bias, the RMSE, the Factor 2, and the 50–90% Coverage associated with each of these parameters are shown in Table 5. As expected for complex demography, many parameters are not well estimated, as indicated by low R^2, high bias, and high RMSE. The parameters showing better estimation quality are the effective population sizes, in particular those associated with the ancestral population of African and non-African modern humans (nYG, R^2 = 91%), and the ancestral population of modern and archaic groups (nAM, R^2 = 99%). The divergence times appear to have been estimated reasonably well, with most of R^2s above 10%. This is true in particular for the times of the two Out of Africa events, which also show a low bias and a high Factor2 and Coverage. On the other hand, it is evident that the data tell us very little about admixture events (their timing and admixture proportions) and migration rates. Although disappointing, this is not unexpected, and high levels of uncertainty associated with these parameters were already reported [13].

Table 3. Estimated parameters for the MD model using the Papuan samples from [13]. The mean and the median estimated values are listed, as well as the 90% and the 50% credible intervals. The parameters cited in the text are reported in bold.

Parameter	Mean	Median	Variance	Q (0.05)	Q (0.95)	Q (0.25)	Q (0.75)
nAR	**2822**	**2793**	$\mathbf{5.77 \times 10^{4}}$	**2540**	**3410**	**2666**	**2914**
nY	**19,077**	**14,347**	$\mathbf{1.72 \times 10^{8}}$	**4204**	**44,993**	**7976**	**29117**
nG1	26,191	26,995	$\mathbf{2.08 \times 10^{8}}$	3253	47,385	13,670	39,819
nG2	23,473	22,275	$\mathbf{1.96 \times 10^{8}}$	1903	46,649	11,151	34,663
nBE	25,612	26,269	$\mathbf{2.08 \times 10^{8}}$	2731	47,604	13,394	38,160
nE	**13,498**	**6616**	$\mathbf{2.07 \times 10^{8}}$	**627**	**42,565**	**1616**	**23,761**
nA	**16,360**	**11,553**	$\mathbf{2.25 \times 10^{8}}$	**773**	**44,620**	**2599**	**28,065**
nP	**24,268**	**24,839**	$\mathbf{2.34 \times 10^{8}}$	**1535**	**47,534**	**10,756**	**37,349**
nYG	**23,317**	**22,292**	$\mathbf{3.19 \times 10^{7}}$	**17,112**	**35,456**	**19,789**	**25,425**
nNNR	**2424**	**2343**	$\mathbf{1.22 \times 10^{5}}$	**2057**	**3001**	**2219**	**2504**
nDDR	21,360	19,680	2.00×10^{8}	1570	46,512	9482	32,332
nDN	17,025	12,576	1.77×10^{8}	2789	43,117	5312	27,001
nADN	19,733	16,531	2.28×10^{8}	2108	47,465	5770	31,455
nAM	**18,846**	**18,745**	$\mathbf{1.73 \times 10^{6}}$	**16,780**	**21,023**	**17,911**	**19,745**
rP	0.0214	0.0146	8.36×10^{-4}	0.0105	0.0532	0.0119	0.0192
rEA	0.0313	0.0179	1.91×10^{-3}	0.0109	0.0869	0.0142	0.0303
tdYG1	**101,162**	**103,842**	$\mathbf{7.61 \times 10^{8}}$	**54,830**	**140,536**	**78,262**	**125,226**
tdYG2	**99,000**	**98,925**	$\mathbf{7.13 \times 10^{8}}$	**55,038**	**137,970**	**76,482**	**124,250**
tdOA1	**77,106**	**73,566**	$\mathbf{5.86 \times 10^{8}}$	**47,019**	**120,206**	**55,392**	**96,881**
tOAbot1	73,389	66,248	6.14×10^{8}	44,341	118,942	52,082	93,165
tdOA2	**47,524**	**45,937**	$\mathbf{3.99 \times 10^{7}}$	**40,394**	**59,245**	**42,597**	**51,019**
tOAbot2	45,223	43,282	5.30×10^{7}	37,718	58,387	40,110	48,153
tdG2BE	68,415	61,497	3.78×10^{8}	50,281	113,560	53,713	75,889
tdEA	**38,187**	**37,017**	$\mathbf{4.33 \times 10^{7}}$	**30,483**	**50,076**	**33,374**	**41,444**
taNG2	52,032	49,731	8.13×10^{7}	42,680	69,758	45,402	55,444
taNEA	41,663	40,005	4.51×10^{7}	33,965	55,743	36,653	45,055
taARP	61,567	55,048	4.53×10^{8}	37,831	106,642	43,945	75,654
taD1P	51,047	44,460	3.89×10^{8}	31,094	95,155	36,207	58,088
taD2A	28,645	27,059	4.24×10^{7}	20,958	39,746	23,730	32,456
taBEE	25,269	24,844	1.00×10^{8}	11,194	45,254	16,827	31,380
paNG2	5.19×10^{-2}	4.99×10^{-2}	7.71×10^{-4}	9.44×10^{-3}	9.52×10^{-3}	2.91×10^{-2}	7.73×10^{-2}
paNEA	4.73×10^{-2}	4.73×10^{-2}	7.95×10^{-4}	5.36×10^{-3}	9.57×10^{-2}	2.30×10^{-2}	7.01×10^{-2}

Table 3. *Cont.*

Parameter	Mean	Median	Variance	Q (0.05)	Q (0.95)	Q (0.25)	Q (0.75)
paARP	4.82×10^{-2}	4.83×10^{-2}	9.00×10^{-4}	4.97×10^{-3}	9.45×10^{-2}	2.09×10^{-2}	7.71×10^{-2}
paD1P	5.21×10^{-2}	5.27×10^{-2}	8.43×10^{-4}	4.58×10^{-3}	9.53×10^{-2}	2.84×10^{-2}	7.85×10^{-2}
paD2A	4.74×10^{-2}	4.72×10^{-2}	8.46×10^{-4}	3.95×10^{-3}	9.32×10^{-2}	2.17×10^{-2}	7.24×10^{-2}
paBEE	2.78×10^{-1}	2.85×10^{-1}	1.61×10^{-2}	6.83×10^{-2}	4.79×10^{-1}	1.71×10^{-1}	3.83×10^{-1}
mYG1	4.75×10^{-4}	4.62×10^{-4}	9.64×10^{-8}	2.61×10^{-5}	9.48×10^{-4}	1.92×10^{-4}	7.54×10^{-4}
mG1Y	4.74×10^{-4}	4.64×10^{-4}	7.95×10^{-8}	4.65×10^{-5}	9.30×10^{-4}	2.25×10^{-4}	6.98×10^{-4}
mG1G2	4.93×10^{-4}	4.80×10^{-4}	8.50×10^{-8}	4.54×10^{-5}	9.41×10^{-4}	2.49×10^{-4}	7.63×10^{-4}
mG2G1	5.34×10^{-4}	5.61×10^{-4}	8.83×10^{-8}	4.77×10^{-5}	9.68×10^{-4}	2.69×10^{-4}	7.94×10^{-4}
mG2E	5.23×10^{-4}	5.29×10^{-4}	8.13×10^{-8}	5.19×10^{-5}	9.57×10^{-4}	2.84×10^{-4}	7.81×10^{-4}
mEG2	4.21×10^{-4}	3.69×10^{-4}	7.78×10^{-8}	3.73×10^{-5}	9.07×10^{-4}	1.85×10^{-4}	6.48×10^{-4}
mEA	4.19×10^{-4}	3.60×10^{-4}	8.63×10^{-8}	3.73×10^{-5}	9.66×10^{-4}	1.81×10^{-4}	6.45×10^{-4}
mAE	5.33×10^{-4}	5.69×10^{-4}	7.63×10^{-8}	5.82×10^{-5}	9.33×10^{-4}	2.90×10^{-4}	7.57×10^{-4}
mAP	1.70×10^{-4}	1.27×10^{-4}	2.26×10^{-8}	1.42×10^{-5}	5.16×10^{-4}	7.40×10^{-5}	2.10×10^{-4}
mPA	1.28×10^{-4}	1.02×10^{-4}	1.18×10^{-8}	8.01×10^{-6}	3.37×10^{-4}	4.52×10^{-5}	1.72×10^{-4}
m1G2EA	4.96×10^{-4}	5.01×10^{-4}	8.24×10^{-8}	5.60×10^{-6}	9.47×10^{-4}	2.45×10^{-4}	7.53×10^{-4}
m1EAG2	4.46×10^{-4}	4.00×10^{-4}	8.23×10^{-8}	5.18×10^{-5}	9.49×10^{-4}	1.99×10^{-4}	6.95×10^{-4}
m1EAP	4.25×10^{-4}	3.97×10^{-4}	7.57×10^{-8}	2.77×10^{-5}	9.07×10^{-4}	1.95×10^{-4}	6.39×10^{-4}
m1PEA	4.40×10^{-4}	4.02×10^{-4}	8.39×10^{-8}	4.04×10^{-5}	9.31×10^{-4}	1.77×10^{-4}	6.93×10^{-4}

Table 4. Estimated parameters for the MD model using the Papuan samples from [12]. The mean and the median estimated values are listed, as well as the 90% and the 50% credible intervals. The parameters cited in the text are reported in bold.

Parameter	Mean	Median	Variance	Q (0.05)	Q (0.95)	Q (0.25)	Q (0.75)
nAR	**2803**	**2783**	$\mathbf{4.57 \times 10^{4}}$	**2532**	**3302**	**2668**	**2900**
nY	**19,182**	**14,771**	$\mathbf{1.62 \times 10^{8}}$	**4379**	**44,930**	**8223**	**29,102**
nG1	26,722	28,003	2.18×10^{8}	2702	47,514	14,075	40,579
nG2	25,325	27,394	1.97×10^{8}	2218	47,188	13,362	36,308
nBE	25,684	26,296	2.17×10^{8}	2194	47,896	13,706	38,919
nE	**12,485**	**5373**	$\mathbf{1.94 \times 10^{8}}$	**699**	**42,194**	**1616**	**21,836**
nA	**14,543**	**8978**	$\mathbf{2.10 \times 10^{8}}$	**916**	**43,930**	**2214**	**26,207**
nP	**19,089**	**16,639**	$\mathbf{2.16 \times 10^{8}}$	**1048**	**46,319**	**4980**	**30,429**
nYG	**22,857**	**21,922**	$\mathbf{2.62 \times 10^{7}}$	**17,112**	**31,789**	**19,579**	**25,130**
nNNR	**2422**	**2336**	$\mathbf{1.24 \times 10^{5}}$	**2057**	**3023**	**2219**	**2531**
nDDR	21,778	20,572	1.94×10^{8}	1640	46291	9606	32,332
nDN	16,239	11,846	1.59×10^{8}	2879	41321	5311	25,523
nADN	19,279	16,531	2.21×10^{8}	2108	47070	4884	31,082
nAM	**18,629**	**18,574**	$\mathbf{1.57 \times 10^{6}}$	**16,671**	**20,691**	**17,779**	**19,476**
rP	0.0215	0.0143	6.10×10^{-4}	0.0104	0.0576	0.0118	0.0204
rEA	0.0314	0.0179	1.94×10^{-3}	0.0109	0.0869	0.0144	0.0310
tdYG1	**98,829**	**99,987**	$\mathbf{7.31 \times 10^{8}}$	**54,220**	**140,009**	**76,337**	**122,428**
tdYG2	**97,430**	**96,686**	$\mathbf{6.87 \times 10^{8}}$	**54,693**	**138,490**	**76,482**	**120,370**
tdOA1	**74,244**	**68,987**	$\mathbf{5.32 \times 10^{8}}$	**46,663**	**119,539**	**54,334**	**89,685**
tOAbot1	70,341	64,285	5.47×10^{8}	43,471	116,608	50,992	85,938
tdOA2	**48,554**	**46,257**	$\mathbf{7.36 \times 10^{7}}$	**40,559**	**64,865**	**42,739**	**51,453**
tOAbot2	46,366	43,475	8.49×10^{7}	37,922	63,074	40,247	50,084
tdG2BE	68,122	62,035	3.36×10^{8}	50,281	105,774	53,533	76,526
tdEA	**37,747**	**35,936**	$\mathbf{5.05 \times 10^{7}}$	**30,381**	**50,399**	**32,690**	**40,845**
taNG2	53,606	50,116	1.08×10^{8}	43,274	73,012	46,917	57,484
taNEA	42,255	40,175	7.98×10^{7}	33,449	56,376	37,030	45,231
taARP	61,203	54,697	4.60×10^{8}	37,428	106,643	43,994	73,444
taD1P	48,493	43,651	2.90×10^{8}	31,343	86,579	36,450	55,023
taD2A	29,298	27,601	5.05×10^{7}	21,090	41,451	24,133	32,700
taBEE	23,871	23,356	9.64×10^{7}	10,508	40,711	15,268	30,666

Table 4. *Cont.*

Parameter	Mean	Median	Variance	Q (0.05)	Q (0.95)	Q (0.25)	Q (0.75)
paNG2	5.29×10^{-2}	5.35×10^{-2}	7.32×10^{-4}	8.94×10^{-3}	9.52×10^{-2}	3.18×10^{-2}	7.51×10^{-2}
paNEA	5.12×10^{-2}	5.22×10^{-2}	7.83×10^{-4}	5.58×10^{-3}	9.60×10^{-2}	2.69×10^{-2}	7.44×10^{-2}
paARP	5.02×10^{-2}	5.06×10^{-2}	8.74×10^{-4}	5.45×10^{-3}	9.49×10^{-2}	2.36×10^{-2}	7.81×10^{-2}
paD1P	5.23×10^{-2}	5.50×10^{-2}	8.00×10^{-4}	6.13×10^{-3}	9.41×10^{-2}	2.78×10^{-2}	7.66×10^{-2}
paD2A	4.82×10^{-2}	4.52×10^{-2}	8.87×10^{-4}	4.93×10^{-3}	9.58×10^{-2}	2.27×10^{-2}	7.39×10^{-2}
paBEE	2.79×10^{-1}	2.91×10^{-1}	1.65×10^{-2}	6.58×10^{-2}	4.78×10^{-1}	1.68×10^{-1}	3.88×10^{-1}
mYG1	4.47×10^{-4}	4.08×10^{-4}	8.52×10^{-8}	3.74×10^{-5}	9.32×10^{-4}	1.89×10^{-4}	6.97×10^{-4}
mG1Y	4.92×10^{-4}	4.91×10^{-4}	7.55×10^{-8}	5.11×10^{-5}	9.27×10^{-4}	2.79×10^{-4}	7.28×10^{-4}
mG1G2	4.74×10^{-4}	4.59×10^{-4}	8.40×10^{-8}	4.41×10^{-5}	9.35×10^{-4}	2.31×10^{-4}	7.32×10^{-4}
mG2G1	5.20×10^{-4}	5.23×10^{-4}	9.07×10^{-8}	4.77×10^{-5}	9.67×10^{-4}	2.34×10^{-4}	7.93×10^{-4}
mG2E	5.16×10^{-4}	5.29×10^{-4}	7.87×10^{-8}	5.67×10^{-5}	9.55×10^{-4}	2.85×10^{-4}	7.60×10^{-4}
mEG2	3.77×10^{-4}	3.04×10^{-4}	8.13×10^{-8}	2.70×10^{-5}	9.11×10^{-4}	1.30×10^{-4}	5.80×10^{-4}
mEA	5.07×10^{-4}	5.15×10^{-4}	8.78×10^{-8}	4.74×10^{-5}	9.57×10^{-4}	2.52×10^{-4}	7.68×10^{-4}
mAE	4.67×10^{-4}	4.68×10^{-4}	7.94×10^{-8}	4.78×10^{-5}	9.17×10^{-4}	2.29×10^{-4}	7.07×10^{-4}
mAP	5.17×10^{-4}	5.12×10^{-4}	7.28×10^{-8}	1.04×10^{-4}	9.35×10^{-4}	2.78×10^{-4}	7.50×10^{-4}
mPA	4.05×10^{-4}	3.79×10^{-4}	5.71×10^{-8}	5.15×10^{-5}	8.70×10^{-4}	2.27×10^{-4}	5.41×10^{-4}
m1G2EA	5.20×10^{-4}	5.21×10^{-4}	8.85×10^{-8}	4.88×10^{-5}	9.74×10^{-4}	2.74×10^{-4}	7.90×10^{-4}
m1EAG2	4.56×10^{-4}	4.30×10^{-4}	7.91×10^{-8}	5.77×10^{-5}	9.24×10^{-4}	2.09×10^{-4}	7.16×10^{-4}
m1EAP	4.92×10^{-4}	5.12×10^{-4}	7.88×10^{-8}	6.32×10^{-5}	9.42×10^{-4}	2.47×10^{-4}	7.11×10^{-4}
m1PEA	4.78×10^{-4}	4.59×10^{-4}	7.42×10^{-8}	6.17×10^{-5}	9.24×10^{-4}	2.44×10^{-4}	7.02×10^{-4}

Table 5. Accuracy of the estimated parameters of the MD model assessed by 1000 pods. The parameters cited in the text are reported in bold.

Parameters	R^2	Bias	RMSE	Factor 2	Coverage 90%	Coverage 50%
nAR	**0.84**	**−0.0020**	**5.90×10^3**	**0.990**	**0.935**	**0.553**
nY	**0.54**	**0.1900**	**1.04×10^4**	**0.867**	**0.919**	**0.522**
nG1	0.08	2.0020	1.46×10^4	0.702	0.880	0.466
nG2	0.17	0.9175	1.36×10^4	0.698	0.915	0.497
nBE	0.02	2.2194	1.47×10^4	0.722	0.895	0.479
nE	**0.33**	**0.4278**	**1.25×10^4**	**0.767**	**0.908**	**0.523**
nA	**0.28**	**0.4159**	**1.20×10^4**	**0.795**	**0.922**	**0.532**
nP	**0.39**	**0.3425**	**1.21×10^4**	**0.791**	**0.908**	**0.501**
nYG	**0.91**	**0.0020**	**3.54×10^3**	**0.998**	**0.957**	**0.650**
nNNR	**0.92**	**0.0086**	**3.64×10^3**	**0.998**	**0.966**	**0.622**
nDDR	0.36	0.3529	1.18×10^4	0.800	0.923	0.522
nDN	0.54	0.1979	1.09×10^4	0.842	0.941	0.534
nADN	0.33	0.7749	1.29×10^4	0.705	0.930	0.476
nAM	**0.99**	**0.0067**	**5.40×10^2**	**0.997**	**0.995**	**0.870**
rP	0.10	0.1110	6.79×10^{-2}	0.721	0.879	0.521
rEA	0.10	0.0983	5.65×10^{-2}	0.748	0.915	0.547
tdYG1	**0.25**	**0.0629**	**2.23×10^4**	**0.998**	**0.928**	**0.576**
tdYG2	**0.25**	**0.0630**	**2.25×10^4**	**0.996**	**0.934**	**0.573**
tdOA1	**0.19**	**0.0025**	**1.99×10^4**	**0.998**	**0.911**	**0.540**
tOAbot1	0.19	0.0052	1.99×10^4	0.996	0.918	0.544
tdOA2	**0.13**	**−0.0257**	**1.24×10^4**	**0.998**	**0.883**	**0.511**
tOAbot2	0.13	−0.0261	1.24×10^4	0.995	0.881	0.512
tdG2BE	0.16	−0.0016	1.98×10^4	0.999	0.913	0.523
tdEA	**0.08**	**−0.0167**	**9.09×10^3**	**0.989**	**0.898**	**0.495**
taD2A	0.04	0.0116	7.35×10^3	0.993	0.905	0.526
paD2A	0.02	0.0010	2.88×10^{-2}	1.000	0.900	0.500
taBEE	0.03	0.1286	1.04×10^4	0.914	0.904	0.486
paBEE	0.02	0.0439	1.31×10^{-1}	1.000	0.893	0.497

Table 5. *Cont.*

Parameters	R^2	Bias	RMSE	Factor 2	Coverage 90%	Coverage 50%
taD1P	0.11	−0.0070	1.72×10^4	0.973	0.897	0.499
paD1P	0.02	−0.0002	2.85×10^{-2}	1.000	0.897	0.508
taARP	0.15	−0.0002	1.85×10^4	0.988	0.916	0.517
paARP	0.03	−0.0014	2.85×10^{-2}	1.000	0.906	0.509
taNEA	0.10	−0.0204	1.06×10^4	0.992	0.893	0.516
paNEA	0.02	0.0000	2.81×10^{-2}	1.000	0.924	0.516
taNG2	0.15	−0.0223	1.36×10^4	0.998	0.909	0.528
paNG2	0.02	−0.0003	2.89×10^{-2}	1.000	0.909	0.477
mYG1	0.15	1.2696	2.69×10^{-4}	0.709	0.927	0.521
mG1Y	0.03	1.8171	2.86×10^{-4}	0.742	0.907	0.516
mG1G2	0.05	2.0667	2.85×10^{-4}	0.737	0.895	0.519
mG2G1	0.05	2.9954	2.89×10^{-4}	0.745	0.885	0.509
mG2E	0.03	3.0547	3.01×10^{-4}	0.692	0.886	0.460
mEG2	0.19	1.5013	2.67×10^{-4}	0.722	0.908	0.503
mEA	0.12	1.4834	2.68×10^{-4}	0.744	0.902	0.543
mAE	0.11	1.9813	2.74×10^{-4}	0.731	0.908	0.523
mAP	0.27	1.4789	2.40×10^{-4}	0.766	0.910	0.548
mPA	0.37	2.2687	2.35×10^{-4}	0.773	0.908	0.546
m1G2EA	0.02	2.1201	2.90×10^{-4}	0.701	0.911	0.489
m1EAG2	0.04	2.7879	2.92×10^{-4}	0.708	0.888	0.496
m1EAP	0.06	2.5111	2.82×10^{-4}	0.728	0.901	0.528
m1PEA	0.05	3.2113	2.91×10^{-4}	0.694	0.911	0.477

The estimates for the current African effective population size (nY) is about 15,000 (median value), in agreement with previous studies [37,38]. A lower value is estimated for the Eurasians, with an effective population size of about 7000 individuals for the Europeans (nE) and of about 11,000 individuals for the Asians (nA). A bit higher is the estimate for Australo-Melanesian population: the median value of the effective population size is indeed about 25,000 individuals (nP).

The first divergence within Africa (tdYG1), that generated the source population giving rise to the first wave of migrants has been estimated about 104,000 years ago, with a 95% confidence interval between 55,000 and 141,000 years ago (and a 50% CI between 78,000 and 125,000 years ago). The first waves of migrants left Africa (tdOA1) about 74,000 years ago (95% CI: 47,000–120,000 years ago; 50% CI: 55,000–96,000 years ago), whereas the second wave of migration (tdOA2), originated from a structure generated (tdYG2) about 100,000 years ago, left Africa about 46,000 years ago (95% CI: 40,000–59,000 years ago, 50% CI: 42,000–51,000 years ago). Europeans and Asians diverged (tdEA) about 37,000 years ago. These estimates are in agreement with a previous work that considered a less realistic model and a smaller amount of genetic data [11].

4. Discussion

In this paper, we explicitly compared two models of AMH evolution through an ABC–RF approach based on the analysis of modern and ancient complete genomes. The two tested demographic models consider details of our evolutionary history that have been proposed in the recent literature, such as the presence of a (so far, unsampled) Basal European population contributing to the genome of recent Europeans [30], or the two distinct pulses of admixture from two different Denisovan populations to Asians and Papuans [23]. The main difference between the two scenarios regards the dynamics of expansion from Africa of AMH. According to the SD model, all non-African populations derive from a single major migration wave; on the contrary, the MD model assumes two migration waves, distinct in time and place, the first one giving rise to modern Australo-Melanesians and the other giving rise to Eurasians. Needless to say, successive processes of gene flow and admixture have certainly complicated the apparently simple patterns generated by the initial African dispersal(s). Yet, even these admittedly simplified models are complex (defined by up to 50 parameters), and the differences between them are relatively small; therefore, one could expect that it might be difficult to tell them apart. On the contrary, the ABC-RF procedure we chose provided a good discriminatory power, with a proportion of True Positives of about 70% for both AD and MD models. This TP proportion is comparable to, or higher than, that reported in previous works where simpler (and hence less realistic) models were analyzed (see e.g., [39,40]). When the two alternative models were compared, the MD model resulted consistently four-fold more probable than the SD model, no matter which Papuan (Table 2), African, European or Asian individuals were considered (Table S4), with a posterior probability estimated around 80%. The support for the MD model is marginally higher than in [16], where a comparison between two alternative, less up-to-date, evolutionary histories of AMH favored the MD model with a probability of about 75%. These results are robust to slight changes in the MD parametrization. We indeed tested also a version of MD in which Papuans derived part of their genomes from Eurasians, modeled as a single pulse of admixture occurring after the second exit (rather than through a process of continuous gene flow), the results are reported in Table S7. Even in this version, the MD appeared more supported by data than the SD model, although it appeared slightly less likely than the previous MD model when included in the general comparison.

In this work, for the first time, we also attempted to estimate the parameters of the supported model by ABC-RF. The MD model was defined by 50 free parameters, estimated through the regression random forest algorithm [20]. We also assessed the quality of these estimates through the calculation of statistics that gave us information about the inferential power of the parameter's estimation procedure. An assessment of the quality of the estimated parameters was prohibitive so far, due to computational limits of other inferential methods, e.g., those based on composite-likelihood [41]. With ABC-RF, instead, the same reference table (made up of just a few thousand simulations) allows one to both estimate parameters and assess their quality using a subset of the simulation as "pods". To perform the same analysis by composite-likelihood methods, one would require about 100 thousand new simulations for each pod analyzed, which means, even considering only 100 pods, billions of simulations. This large amount of simulated data often exceeds computational constraints, in particular when complex demographies are analyzed. As a consequence, in studies of complex models, no information was provided about the reliability of parameter estimates [13,42]. The procedure we applied made it possible to compensate for this drawback, as shown in Table 5.

It would have been unrealistic to expect that all 50 parameters could be reliably estimated. The migration rates among modern populations, or the proportion and timing of admixture events, for instance, proved elusive, showing a low R^2 and high bias and RMSE values. We knew that there is an almost infinite set of parameter combinations leading to the same patterns of genome diversity, with, for instance, old small-scale admixture events, and recent larger-scale admixture events, producing, in principle, the same consequences at the genomic level. Other parameters show better estimates. This is the case of the effective population sizes, or, to a lesser extent, of the divergence times. The African, European and Asian estimates of the effective population sizes are consistent with what reported in the literature [38,43]; the higher value estimated for the Australo-Melanesian group, here represented by the Papuans, may be surprising, but it is in agreement with the harmonic mean of the effective population sizes estimated over time by [12].

The most interesting parameters are those associated with the divergence/departure from Africa. These parameters show R^2 above 10%, good coverage, and a factor 2 of about 100%; however, their confidence intervals are huge and their posterior distributions often seem to reflect the prior range. This means that we should still take with caution these estimates and that the ABC inferential procedure, albeit powerful, shows room for improvement. The key advantage of the ABC estimation is that the "quality assessment" procedure allows the acquisition of consciousness about the quality of the estimates; nevertheless, having this in mind, we can still discuss the estimates obtained. We dated the structure of African groups that gave rise to the source populations of the migration waves from Africa about 100,000 years ago. The bottleneck of the first exit from Africa, associated with the origin of Australo-Melanesian groups, has been estimated at about 74,000 years ago, in line with the timing inferred from paleoanthropological data (70,000 years ago, [44]). The second exit, giving rise to Eurasian populations, was placed at about 46,000 years ago. This is in agreement with previous estimates from genomic data [4,38,45] and receives further support from the relatively recent arrival of modern humans in Europe suggested by much of the archaeological evidence (40–45 thousand years ago, [46,47]). Some authors proposed an even earlier presence of AMH in Europe [48]. Be that as it may, it is also plausible that large-scale gene flow processes, documented at least twice in Europe (in the Neolithic period and Bronze Age; see [49]) may have slightly reduced diversity and hence the apparent depth of the DNA genealogies, thus producing a bias towards more recent values in the estimation of divergence times. The two migration waves from Africa considered in the MD model appear to be separated in time, with no temporal overlap considering their 50% confidence interval (55,000–96,000 for the first exit and 42,000–51,000 for the second exit), and a limited overlap considering their 95% confidence interval (47,000–120,000 for the first exit and 40,000–59,000 for the second exit).

5. Conclusions

In this paper we extensively tested two up-to-date models of modern human expansion Out of Africa through a machine learning ABC approach. The simulated variation has been compared with those observed in ancient and modern genomes, and our results consistently supported a Multiple Dispersal Model, in which modern Australo-Melanesians derive from an earlier migration from Africa than that giving rise to Eurasians. We also estimated the parameters of the most supported model, and we concentrated our effort in assessing the quality of the estimates produced. This procedure, albeit fundamental to ensure the reliability of the estimates, it is rarely performed, due to the limitations of available inferential methods. These limitations are currently overcame by the ABC-RF procedure coupled with the FDSS statistic, which allowed us to highlight weakness and strengths of the parameters estimated. Our results indeed support that the hypothesis of two main dispersal event from Africa, separated in time and place [10–12], cannot be dismissed [4,13], but the quality assessment of the parameters we estimated certainly show that needs to be further explored.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/12/1510/s1, Table S1: Demographic parameters and prior distributions of Single Dispersal model. Table S2: Demographic parameters and prior distributions of Multiple Dispersal model. Table S3: Complete list of genomes used for the comparison of Single Dispersal model and Multiple Dispersal model using real data; Table S4: Results of model selection performed using alternative individuals from African, European and Asian populations; Table S5: Power test of model comparison for increasing number of simulations considered in the reference table.; Table S6. Complete list of acronyms of the MD model's demographic parameters.; Table S7. Model Selection results including the MD-Pulse admixture model. Figure S1: Outline of the entire workflow; Figure S2: Posterior density of the effective population sizes estimated using the Papuan sample from Malaspinas et al. (2016). Figure S3: Posterior density of the divergence times and the admixture times estimated using the Papuan sample from Malaspinas et al. (2016). Figure S4: Posterior density of the admixture rates estimated using the Papuan sample from Malaspinas et al. (2016). Figure S5: Posterior density of the migration rates estimated using the Papuan sample from Malaspinas et al. (2016). Figure S6: Posterior density of the effective population sizes estimated using the Papuan sample from Pagani et al. (2016). Figure S7: Posterior density of the divergence times and the admixture times estimated using the Papuan sample from Pagani et al. (2016). Figure S8: Posterior density of the admixture rates estimated using the Papuan sample from Pagani et al. (2016). Figure S9: Posterior density of the migration rates estimated using the Papuan sample from Pagani et al. (2016). Figure S10: The model below represents a simplified version of the most supported model (MD) showing the main demographic parameters.

Author Contributions: Conceptualization, G.B. and S.G.; formal analysis, M.T.V.; methodology, A.B. and S.G.; software, M.T.V. and A.B.; supervision, G.B. and S.G.; writing—original draft, G.B. and S.G.; writing—review and editing, M.T.V., A.B., G.B., and S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We are indebted to Francesca Tassi and Alberto Seno for technical help.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Scerri, E.M.L.; Thomas, M.G.; Manica, A.; Gunz, P.; Stock, J.T.; Stringer, C.; Grove, M.; Groucutt, H.S.; Timmermann, A.; Rightmire, G.P.; et al. Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends Ecol. Evol.* **2018**, *33*, 582–594. [CrossRef]
2. Mellars, P. Neanderthals and the Modern Human Colonization of Europe. *Nature* **2004**, *432*, 461–465. [CrossRef]
3. Higham, T.; Douka, K.; Wood, R.; Ramsey, C.B.; Brock, F.; Basell, L.; Camps, M.; Arrizabalaga, A.; Baena, J.; Barroso-Ruíz, C.; et al. The Timing and Spatiotemporal Patterning of Neanderthal Disappearance. *Nature* **2014**, *512*, 306–309. [CrossRef] [PubMed]
4. Mallick, S.; Li, H.; Lipson, M.; Mathieson, I.; Gymrek, M.; Racimo, F.; Zhao, M.; Chennagiri, N.; Nordenfelt, S.; Tandon, A.; et al. The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations. *Nature* **2016**, *538*, 201–206. [CrossRef] [PubMed]
5. Hershkovitz, I.; Weber, G.W.; Quam, R.; Duval, M.; Grün, R.; Kinsley, L.; Ayalon, A.; Bar-Matthews, M.; Valladas, H.; Mercier, N.; et al. The Earliest Modern Humans Outside Africa. *Science* **2018**, *359*, 456–459. [CrossRef]
6. Liu, H.; Prugnolle, F.; Manica, A.; Balloux, F. A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *Am. J. Hum. Genet.* **2006**, *79*, 230–237. [CrossRef] [PubMed]
7. Mellars, P.; Gori, K.C.; Carr, M.; Soares, P.A.; Richards, M.B. Genetic and Archaeological Perspectives on the Initial Modern Human Colonization of Southern Asia. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10699–10704. [CrossRef] [PubMed]
8. López, S.; Van Dorp, L.; Hellenthal, G. Human Dispersal out of Africa: A Lasting Debate. *Evol. Bioinform.* **2015**. [CrossRef] [PubMed]
9. Lahr, M.M.; Foley, R. Multiple Dispersals and Modern Human Origins. *Evol. Anthropol. Issues News Rev.* **1994**, *3*, 48–60. [CrossRef]
10. Reyes-Centeno, H.; Ghirotto, S.; Detroit, F.; Grimaud-Herve, D.; Barbujani, G.; Harvati, K. Genomic and Cranial Phenotype Data Support Multiple Modern Human Dispersals from Africa and a Southern Route into Asia. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 7248–7253. [CrossRef]
11. Tassi, F.; Ghirotto, S.; Mezzavilla, M.; Vilaça, S.T.; De Santi, L.; Barbujani, G. Early Modern Human Dispersal from Africa: Genomic Evidence for Multiple Waves of Migration. *Investig. Genet.* **2015**, *6*, 6–13. [CrossRef] [PubMed]
12. Pagani, L.; Lawson, D.J.; Jagoda, E.; Mörseburg, A.; Eriksson, A.; Mitt, M.; Clemente, F.; Hudjashov, G.; DeGiorgio, M.; Saag, L.; et al. Genomic Analyses Inform on Migration Events during the Peopling of Eurasia. *Nature* **2016**, *538*, 238–242. [CrossRef] [PubMed]
13. Malaspinas, A.S.; Westaway, M.C.; Muller, C.; Sousa, V.C.; Lao, O.; Alves, I.; Bergström, A.; Georgios, A.; Cheng, J.Y.; Crawford, G.E. A Genomic History of Aboriginal Australia. *Nature* **2016**, *538*, 207–214. [CrossRef] [PubMed]
14. Varin, C. On Composite Marginal Likelihoods. *Asta Adv. Stat. Anal.* **2008**, *92*, 1–28. [CrossRef]
15. Varin, C.; Reid, N.; Firth, D. An Overview of Composite Likelihood Methods. *Stat. Sin.* **2011**, *21*, 5–42.
16. Ghirotto, S.; Vizzari, M.T.; Tassi, F.; Barbujani, G.; Benazzo, A. Distinguishing among Complex Evolutionary Models Using Unphased Whole-genome Data through Random-Forest Approximate Bayesian Computation. *Mol. Ecol. Resour.* **2020**, 1–15. [CrossRef]
17. Beaumont, M.A.; Zhang, W.; Balding, D.J. Approximate Bayesian Computation in Population Genetics. *Genetics* **2002**, *162*, 2025–2035.
18. Beaumont, M.A. Joint Determination of Topology, Divergence Time, and Immigration in Population Trees. In *Simulations, Genetics and Human Prehistory*; McDonald Institute for Archaeological Research: Cambridge, UK, 2008; pp. 135–154.
19. Pudlo, P.; Marin, J.M.; Estoup, A.; Cornuet, J.M.; Gautier, M.; Robert, C.P. Reliable ABC Model Choice via Random Forests. *Bioinformatics* **2015**, *32*, 859–866. [CrossRef] [PubMed]
20. Raynal, L.; Marin, J.M.; Pudlo, P.; Ribatet, M.; Robert, C.P.; Estoup, A. ABC Random Forests for Bayesian Parameter Inference. *Bioinformatics* **2019**, *35*, 1720–1728. [CrossRef]
21. Wakeley, J.; Hey, J. Estimating Ancestral Population Parameters. *Genetics* **1997**, *145*, 847–855.

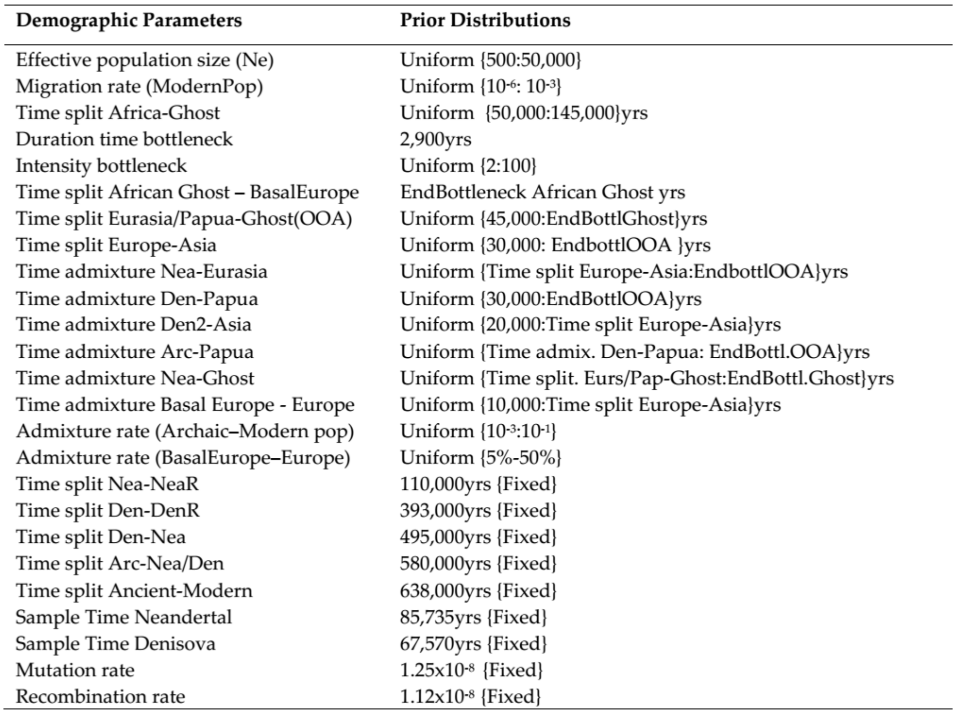
22. Mondal, M.; Casals, F.; Xu, T.; Dall'Olio, G.M.; Pybus, M.; Netea, M.G.; Comas, D.; Laayouni, H.; Li, Q.; Majumder, P.P.; et al. Genomic Analysis of Andamanese Provides Insights into Ancient Human Migration into Asia and Adaptation. *Nat. Genet.* **2016**, *48*, 1066–1070. [CrossRef] [PubMed]
23. Browning, S.R.; Browning, B.L.; Zhou, Y.; Tucci, S.; Akey, J.M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **2018**, *173*, 53–61.e9. [CrossRef]
24. Jacobs, G.S.; Hudjashov, G.; Saag, L.; Kusuma, P.; Darusallam, C.C.; Lawson, D.J.; Mondal, M.; Pagani, L.; Ricaut, F.-X.; Stoneking, M.; et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* **2019**, *177*, 1010–1021. [CrossRef] [PubMed]
25. Wall, J.D.; Yang, M.A.; Jay, F.; Kim, S.K.; Durand, E.Y.; Stevison, L.S.; Gignoux, C.; Woerner, A.; Hammer, M.F.; Slatkin, M. Higher Levels of Neanderthal Ancestry in East Asians than in Europeans. *Genetics* **2013**, *194*, 199–209. [CrossRef] [PubMed]
26. Prüfer, K.; Racimo, F.; Patterson, N.; Jay, F.; Sankararaman, S.; Sawyer, S.; Heinze, A.; Renaud, G.; Sudmant, P.H.; De Filippo, C.; et al. The Complete Genome Sequence of a Neanderthal from the Altai Mountains. *Nature* **2014**, *505*, 43–49. [CrossRef] [PubMed]
27. Vernot, B.; Akey, J.M. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science* **2014**, *343*, 1017–1021. [CrossRef] [PubMed]
28. Lazaridis, I.; Patterson, N.; Mittnik, A.; Renaud, G.; Mallick, S.; Kirsanow, K.; Sudmant, P.H.; Schraiber, J.G.; Castellano, S.; Lipson, M.; et al. Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans. *Nature* **2014**, *513*, 409–413. [CrossRef]
29. Lazaridis, I.; Nadel, D.; Rollefson, G.; Merrett, D.C.; Rohland, N.; Mallick, S.; Fernandes, D.; Novak, M.; Gamarra, B.; Sirak, K.; et al. Genomic Insights into the Origin of Farming in the Ancient Near East. *Nature* **2016**, *536*, 419–424. [CrossRef]
30. Villanea, F.A.; Schraiber, J.G. Multiple Episodes of Interbreeding between Neanderthal and Modern Humans. *Nat. Ecol. Evol.* **2019**, *3*, 39–44. [CrossRef]
31. Scally, A.; Durbin, R. Revising the Human Mutation Rate: Implications for Understanding Human Evolution. *Nat. Rev. Genet.* **2012**, *13*, 745–753. [CrossRef]
32. Hudson, R.R. Generating Samples under a Wright-Fisher Neutral Model of Genetic Variation. *Bioinformatics* **2002**, *18*, 337–338. [CrossRef]
33. Meyer, M.; Kircher, M.; Gansauge, M.T.; Li, H.; Racimo, F.; Mallick, S.; Schraiber, J.G.; Jay, F.; Prüfer, K.; De Filippo, C.; et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **2012**, *338*, 222–226. [CrossRef]
34. Hinrichs, A.S.; Raney, B.J.; Speir, M.L.; Rhead, B.; Casper, J.; Karolchik, D.; Kuhn, R.M.; Rosenbloom, K.R.; Zweig, A.S.; Haussler, D.; et al. UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics* **2016**, *32*, 1430–1432. [CrossRef]
35. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
36. Neuenschwander, S.; Largiadèr, C.R.; Ray, N.; Currat, M.; Vonlanthen, P.; Excoffier, L. Colonization History of the Swiss Rhine Basin by the Bullhead (Cottus Gobio): Inference under a Bayesian Spatially Explicit Framework. *Mol. Ecol.* **2008**, *17*, 757–772. [CrossRef] [PubMed]
37. Fan, S.; Kelly, D.E.; Beltrame, M.H.; Hansen, M.E.B.; Mallick, S.; Ranciaro, A.; Hirbo, J.; Thompson, S.; Beggs, W.; Nyambo, T.; et al. African Evolutionary History Inferred from Whole Genome Sequence Data of 44 Indigenous African Populations. *Genome Biol.* **2019**, *20*, 1–14.
38. McEvoy, B.P.; Powell, J.E.; Goddard, M.E.; Visscher, P.M. Human Population Dispersal "Out of Africa" Estimated from Linkage Disequilibrium and Allele Frequencies of SNPs. *Genome Res.* **2011**, *21*, 821–829. [CrossRef]
39. Fagundes, N.J.R.; Ray, N.; Beaumont, M.; Neuenschwander, S.; Salzano, F.M.; Bonatto, S.L.; Excoffier, L. Statistical Evaluation of Alternative Models of Human Evolution. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 17614–17619. [CrossRef] [PubMed]
40. Veeramah, K.R.; Wegmann, D.; Woerner, A.; Mendez, F.L.; Watkins, J.C.; Destro-Bisol, G.; Soodyall, H.; Louie, L.; Hammer, M.F. An Early Divergence of KhoeSan Ancestors from Those of Other Modern Humans Is Supported by an ABC-Based Analysis of Autosomal Resequencing Data. *Mol. Biol. Evol.* **2012**, *29*, 617–630. [CrossRef]

41. Excoffier, L.; Dupanloup, I.; Huerta-Sánchez, E.; Sousa, V.C.; Foll, M. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet.* **2013**, *9*, e1003905. [CrossRef]
42. Nater, A.; Mattle-Greminger, M.P.; Nurcahyo, A.; Nowak, M.G.; De Manuel, M.; Desai, T.; Groves, C.; Pybus, M.; Sonay, T.B.; Roos, C.; et al. Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Curr. Biol.* **2017**, *27*, 3576–3577. [CrossRef] [PubMed]
43. Schiffels, S.; Durbin, R. Inferring Human Population Size and Separation History from Multiple Genome Sequences. *Nat. Genet.* **2014**, *46*, 919–925. [CrossRef] [PubMed]
44. Mirazón Lahr, M.; Foley, R.A. Towards a Theory of Modern Human Origins: Geography, Demography, and Diversity in Recent Human Evolution. *Am. J. Phys. Anthropol.* **1999**, *107*, 137–176. [CrossRef]
45. Gravel, S.; Henn, B.M.; Gutenkunst, R.N.; Indap, A.R.; Marth, G.T.; Clark, A.G.; Yu, F.; Gibbs, R.A.; Bustamante, C.D.; The 1000 Genomes Project; et al. Demographic History and Rare Allele Sharing among Human Populations. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 11983–11988. [CrossRef] [PubMed]
46. Mellars, P. Why Did Modern Human Populations Disperse from Africa ca. 60,000 Years Ago? A New Model. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9381–9386. [CrossRef]
47. Reyes-Centeno, H.; Hubbe, M.; Hanihara, T.; Stringer, C.; Harvati, K. Testing Modern Human Out-of-Africa Dispersal Models and Implications for Modern Human Origins. *J. Hum. Evol.* **2015**, *87*, 95–106. [CrossRef]
48. Hublin, J.J.; Sirakov, N.; Aldeias, V.; Bailey, S.; Bard, E.; Delvigne, V.; Endarova, E.; Fagault, Y.; Fewlass, H.; Hajdinjak, M.; et al. Initial Upper Palaeolithic Homo Sapiens from Bacho Kiro Cave, Bulgaria. *Nature* **2020**, *581*, 299–302. [CrossRef]
49. Haak, W.; Lazaridis, I.; Patterson, N.; Rohland, N.; Mallick, S.; Llamas, B.; Brandt, G.; Nordenfelt, S.; Harney, E.; Stewardson, K.; et al. Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe. *Nature* **2015**, *522*, 207–211. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

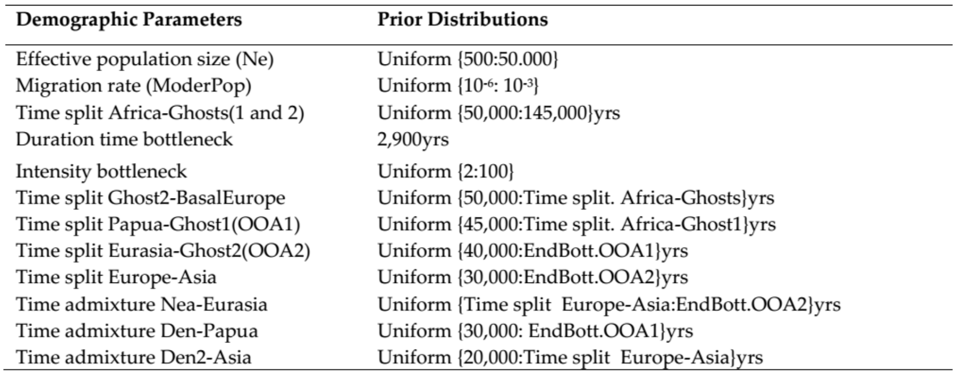
Supplementary Materials

Table S1. Demographic parameters and prior distributions of Single Dispersal model. Migration and admixture rates are expressed per generation, times in years. We considered a generation time of 29 years as in Malaspinas et al. (2016). Per nucleotide per generation mutation and recombination rates are fixed as in Malaspinas et al. (2016). Parameters defined by prior distributions having the same shape and range are indicated through the same entry.



Demographic Parameters	**Prior Distributions**
Effective population size (Ne)	Uniform {500:50,000}
Migration rate (ModernPop)	Uniform {10^{-6}: 10^{-3}}
Time split Africa-Ghost	Uniform {50,000:145,000}yrs
Duration time bottleneck	2,900yrs
Intensity bottleneck	Uniform {2:100}
Time split African Ghost – BasalEurope	EndBottleneck African Ghost yrs
Time split Eurasia/Papua-Ghost(OOA)	Uniform {45,000:EndBottlGhost}yrs
Time split Europe-Asia	Uniform {30,000: EndbottlOOA }yrs
Time admixture Nea-Eurasia	Uniform {Time split Europe-Asia:EndbottlOOA}yrs
Time admixture Den-Papua	Uniform {30,000:EndBottlOOA}yrs
Time admixture Den2-Asia	Uniform {20,000:Time split Europe-Asia}yrs
Time admixture Arc-Papua	Uniform {Time admix. Den-Papua: EndBottl.OOA}yrs
Time admixture Nea-Ghost	Uniform {Time split. Eurs/Pap-Ghost:EndBottl.Ghost}yrs
Time admixture Basal Europe - Europe	Uniform {10,000:Time split Europe-Asia}yrs
Admixture rate (Archaic–Modern pop)	Uniform {10^{-3}:10^{-1}}
Admixture rate (BasalEurope–Europe)	Uniform {5%-50%}
Time split Nea-NeaR	110,000yrs {Fixed}
Time split Den-DenR	393,000yrs {Fixed}
Time split Den-Nea	495,000yrs {Fixed}
Time split Arc-Nea/Den	580,000yrs {Fixed}
Time split Ancient-Modern	638,000yrs {Fixed}
Sample Time Neandertal	85,735yrs {Fixed}
Sample Time Denisova	67,570yrs {Fixed}
Mutation rate	1.25x10^{-8} {Fixed}
Recombination rate	1.12x10^{-8} {Fixed}

Table S2. Demographic parameters and prior distributions of Multiple Dispersal model. Migration and admixture rates are expressed per generation, times in years. We cosidered a generation time of 29 years as in Malaspinas et al. (2016). Per nucleotide per generation mutation and recombination rates are fixed as in Malaspinas et al. (2016). Parameters defined by prior distributions having the same shape and range are indicated through the same entry.



Demographic Parameters	**Prior Distributions**
Effective population size (Ne)	Uniform {500:50.000}
Migration rate (ModerPop)	Uniform {10^{-6}: 10^{-3}}
Time split Africa-Ghosts(1 and 2)	Uniform {50,000:145,000}yrs
Duration time bottleneck	2,900yrs
Intensity bottleneck	Uniform {2:100}
Time split Ghost2-BasalEurope	Uniform {50,000:Time split. Africa-Ghosts}yrs
Time split Papua-Ghost1(OOA1)	Uniform {45,000:Time split. Africa-Ghost1}yrs
Time split Eurasia-Ghost2(OOA2)	Uniform {40,000:EndBott.OOA1}yrs
Time split Europe-Asia	Uniform {30,000:EndBott.OOA2}yrs
Time admixture Nea-Eurasia	Uniform {Time split Europe-Asia:EndBott.OOA2}yrs
Time admixture Den-Papua	Uniform {30,000: EndBott.OOA1}yrs
Time admixture Den2-Asia	Uniform {20,000:Time split Europe-Asia}yrs

Time admixture BasalEurope-Europe	Uniform {10,000:Time split Europe-Asia}yrs
Time admixture Arc-Papua	Uniform {Time admix. Den-Papua:EndBott.OOA1}yrs
Time admixture Nea-Ghost2	Uniform {Time split Euras-Ghost2:Time split Africa-Ghost2}yrs
Admixture rate (Archaic–Modern pop)	Uniform {10^{-3}:10^{-1}}
Admixture rate (BasalEurope–Europe)	Uniform {5%-50%}
Time split Nea-NeaR	110,000yrs {Fixed}
Time split Den-DenR	393,000yrs {Fixed}
Time split Den-Nea	495,000yrs {Fixed}
Time split Arc-Nea/Den	580,000yrs {Fixed}
Time split Ancient-Modern	638,000yrs {Fixed}
Sample Time Neandertal	85,735yrs {Fixed}
Sample Time Denisova	67,570yrs {Fixed}
Mutation rate	1.25x10^{-8} {Fixed}
Recombination rate	1.12x10^{-8} {Fixed}

Table S3. Complete list of genomes used for the comparison of Single Dispersal model and Multiple Dispersal model using real data.

Population	ID_Individual	Reference
Neandertal	AltaiNea	Prufer *et al.* (2014)
Denisova	DenisovaPinky	Mayer *et al.* (2012)
African	CongPy1	Pagani *et al.* (2016)
African	CongPy3	Pagani *et al.* (2016)
African	CongPy6	Pagani *et al.* (2016)
European	Est1	Pagani *et al.* (2016)
European	Est2	Pagani *et al.* (2016)
European	Est3	Pagani *et al.* (2016)
European	Est4	Pagani *et al.* (2016)
European	Est5	Pagani *et al.* (2016)
European	Est6	Pagani *et al.* (2016)
Asian	VietN1	Pagani *et al.* (2016)
Asian	VietN2	Pagani *et al.* (2016)
Asian	VietC1	Pagani *et al.* (2016)
Asian	VietC2	Pagani *et al.* (2016)
Asian	VietS1	Pagani *et al.* (2016)
Asian	VietS2	Pagani *et al.* (2016)
Papuan	Koinb1	Pagani *et al.* (2016)
Papuan	Koinb2	Pagani *et al.* (2016)
Papuan	Koinb3	Pagani *et al.* (2016)
Papuan	Kosip1	Pagani *et al.* (2016)
Papuan	Kosip2	Pagani *et al.* (2016)
Papuan	Kosip3	Pagani *et al.* (2016)
Papuan	EGAN00001279031	Malaspinas *et al.* (2016)
Papuan	EGAN00001279039	Malaspinas *et al.* (2016)
Papuan	EGAN00001279047	Malaspinas *et al.* (2016)
Papuan	EGAN00001279054	Malaspinas *et al.* (2016)
Papuan	EGAN00001279032	Malaspinas *et al.* (2016)
Papuan	EGAN00001279040	Malaspinas *et al.* (2016)
Papuan	EGAN00001279048	Malaspinas *et al.* (2016)
Papuan	EGAN00001279033	Malaspinas *et al.* (2016)
Papuan	EGAN00001279041	Malaspinas *et al.* (2016)

Papuan	EGAN00001279049	Malaspinas *et al*. (2016)
Papuan	EGAN00001279034	Malaspinas *et al*. (2016)
Papuan	EGAN00001279042	Malaspinas *et al*. (2016)
Papuan	EGAN00001279050	Malaspinas *et al*. (2016)
Papuan	EGAN00001279035	Malaspinas *et al*. (2016)
Papuan	EGAN00001279043	Malaspinas *et al*. (2016)
Papuan	EGAN00001279051	Malaspinas *et al*. (2016)
Papuan	EGAN00001279036	Malaspinas *et al*. (2016)
Papuan	EGAN00001279044	Malaspinas *et al*. (2016)
Papuan	EGAN00001279052	Malaspinas *et al*. (2016)
Papuan	EGAN00001279037	Malaspinas *et al*. (2016)
Papuan	EGAN00001279045	Malaspinas *et al*. (2016)
Papuan	EGAN00001279053	Malaspinas *et al*. (2016)
Papuan	EGAN00001279038	Malaspinas *et al*. (2016)
Papuan	EGAN00001279046	Malaspinas *et al*. (2016)
Papuan	EGAN00001279055	Malaspinas *et al*. (2016)

Table S4. Results of model selection performed using alternative individual from African, European and Asian populations.

ID_Individual	Selected model	Votes SD	Votes MD	Post.Prob.
CongPy3	MD	152	348	0.83
CongPy6	MD	167	333	0.81
Est2	MD	120	380	0.82
Est3	MD	113	387	0.80
Est4	MD	132	368	0.81
Est5	MD	108	392	0.82
Est6	MD	181	319	0.80
VietN2	MD	111	389	0.83
VietC1	MD	100	400	0.84
VietC2	MD	153	347	0.84
VietS1	MD	145	355	0.83
VietS2	MD	150	350	0.82

Table S5. Power test of model comparison for increasing number of simulations considered in the reference table.

Prior Err. Rate	True Positive SD	True Positive MD	Post.Prob. SD	Post.Prob. MD	n. Sim.
0.271	0.720	0.736	0.733	0.724	20,000
0.264	0.723	0.748	0.745	0.724	50,000
0.260	0.730	0.755	0.750	0.732	100,000

Table S6. Complete list of acronyms of the MD model's demographic parameters.

Acronym	Parameters
nAR	Effective population size UnknownArchaic
nY	Effective population size Africa
nG1	Effective population size Ghost1
nG2	Effective population size Ghost2
nBE	Effective population size Basal Europe
nE	Effective population size Europe
nA	Effective population size Asia

nP	Effective population size Papua
nYG	Effective population size Ancestral Africa
nNNR	Effective population size Ancestral Neandertal
nDDR	Effective population size Ancestral Denisovan
nDN	Effective population size Ancestral Denisova-Neandertal population
nADN	Effective population size Ancestral Archaic populations
nAM	Effective population size Ancestral Archaic-Modern population
rP	Intensity Bottleneck Papua
rEA	Intensity Bottleneck Eurasia
tdYG1	Divergence time African-Ghost populations
tdYG2	Divergence time African-Ghost populations
tdOA1	Time of the first Out-of-Africa
tOAbot1	Time end bottleneck first Out-of Africa
tdOA2	Time of the second Out-of-Africa
tOAbot2	Time end bottleneck second Out-of Africa
tdG2BE	Divergence time Africa-Basal Europe
tdEA	Divergence time Europe-Asia
taD2A	Admixture time Denisova2-Asia
paD2A	Admixture rate Denisova2-Asia
taBEE	Admixture time Basal Europe-Europe
paBEE	Admixture rate Basal Europe-Europe
taD1P	Admixture time Denisova1-Papua
paD1P	Admixture rate Denisova1-Papua
taARP	Admixture time UnknownArchaic-Papua
paARP	Admixture rate UnknownArchaic-Papua
taNEA	Admixture time Neandertal- Eurasia
paNEA	Admixture rate Neandertal- Eurasia
taNG2	Admixture time Neandertal- Ghost2
paNG2	Admixture rate Neandertal- Ghost2
mYG1	Migration rate Africa-Ghost1
mG1Y	Migration rate Ghost1-Africa
mG1G2	Migration rate Ghost1- Ghost2
mG2G1	Migration rate Ghost2- Ghost1
mG2E	Migration rate Ghost2-Europe
mEG2	Migration rate Europe-Ghost2
mEA	Migration rate Europe-Asia
mAE	Migration rate Asia-Europe
mAP	Migration rate Asia-Papua
mPA	Migration rate Papua-Asia
m1G2EA	Migration rate Ghost2-Eurasia
m1EAG2	Migration rate Eurasia-Ghost2
m1EAP	Migration rate Eurasia-Papua
m1PEA	Migration rate Papua-Eurasia

Table S7. Model Selection results including the MD-Pulse admixture model. In the first column are reported the ID of the Papuan samples used for the model choice. The second column shows the model selected by the ABC procedure. In the third, fourth and fifth columns are reported the votes assigned to the SD, the MD-Continuous migration and the MD- Pulse Admixture models by the Random-Forest algorithm. The last column shows the posterior probabilities associated to the most supported model.

ID_Individual	Selected model	Votes SD	Votes MD-Cont.Migration	Votes MD-PulseAdmx	Post.Prob.
EGAN00001279031	MD- Cont.Migration	68	235	197	0.81
EGAN00001279039	MD- PulseAdmx	50	191	259	0.82
EGAN00001279047	MD- PulseAdmx	45	212	243	0.83
EGAN00001279054	MD- PulseAdmx	34	161	305	0.82
EGAN00001279032	MD- PulseAdmx	53	195	252	0.80
EGAN00001279040	MD- PulseAdmx	39	190	271	0.83
EGAN00001279048	MD- Cont.Migration	70	247	183	0.82
EGAN00001279033	MD- Cont.Migration	73	234	193	0.83
EGAN00001279041	MD- Cont.Migration	71	247	182	0.83
EGAN00001279049	MD- Cont.Migration	65	218	217	0.83
EGAN00001279034	MD- PulseAdmx	40	177	283	0.82
EGAN00001279042	MD- PulseAdmx	43	193	264	0.84
EGAN00001279050	MD- PulseAdmx	55	203	242	0.82
EGAN00001279035	MD- PulseAdmx	29	165	306	0.82
EGAN00001279043	MD- Cont.Migration	65	238	197	0.82
EGAN00001279051	MD- PulseAdmx	36	164	300	0.81
EGAN00001279036	MD- PulseAdmx	41	171	288	0.82
EGAN00001279044	MD- Cont.Migration	66	250	184	0.83
EGAN00001279052	MD- Cont.Migration	55	249	196	0.83
EGAN00001279037	MD- Cont.Migration	72	231	197	0.81
EGAN00001279045	MD- Cont.Migration	65	233	202	0.82
EGAN00001279053	MD- PulseAdmx	54	214	232	0.81
EGAN00001279038	MD- PulseAdmx	37	205	258	0.84
EGAN00001279046	MD- Cont.Migration	70	242	188	0.82
EGAN00001279055	MD- PulseAdmx	25	149	326	0.82
Koinb1	MD- Cont.Migration	120	298	82	0.80
Koinb2	MD- Cont.Migration	123	294	83	0.80
Koinb3	MD- Cont.Migration	135	269	96	0.82
Kosip1	MD- Cont.Migration	117	289	94	0.80
Kosip2	MD- Cont.Migration	106	294	100	0.81
Kosip3	MD- Cont.Migration	112	312	76	0.80

Figure S1: Outline of the entire workflow.

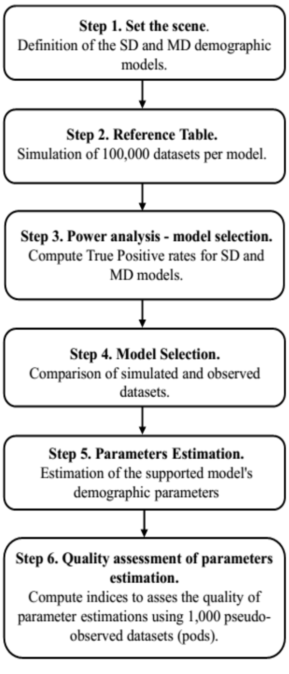

Step 1. Set the scene.
Definition of the SD and MD demographic models.
Step 2. Reference Table.
Simulation of 100,000 datasets per model.
Step 3. Power analysis - model selection.
Compute True Positive rates for SD and MD models.
Step 4. Model Selection.
Comparison of simulated and observed datasets.
Step 5. Parameters Estimation.
Estimation of the supported model's demographic parameters
Step 6. Quality assessment of parameters estimation.
Compute indices to asses the quality of parameter estimations using 1,000 pseudo-observed datasets (pods).

Figure S2: **Posterior density of the effective population sizes estimated using the Papuan sample from Malaspinas et al. (2016)**. The plots show: the posterior density (black), the mean (red) and median (blue) estimated values and the distribution of parameter's values sampled from the prior (gray).

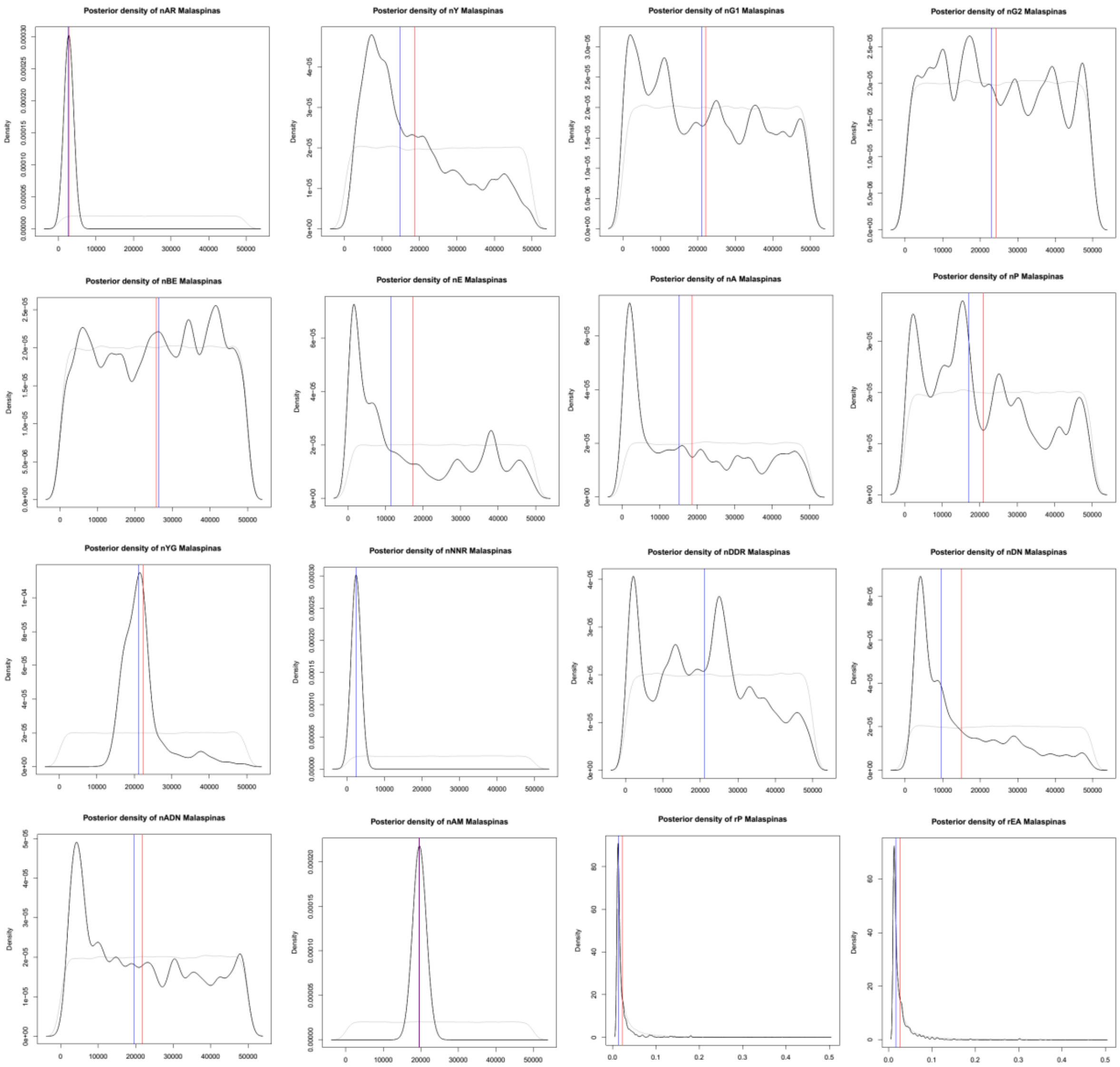


Figure S3. **Posterior density of the divergence times and the admixture times estimated using the Papuan sample from Malaspinas et al. (2016)**. The plots have the same features of Figure S2.

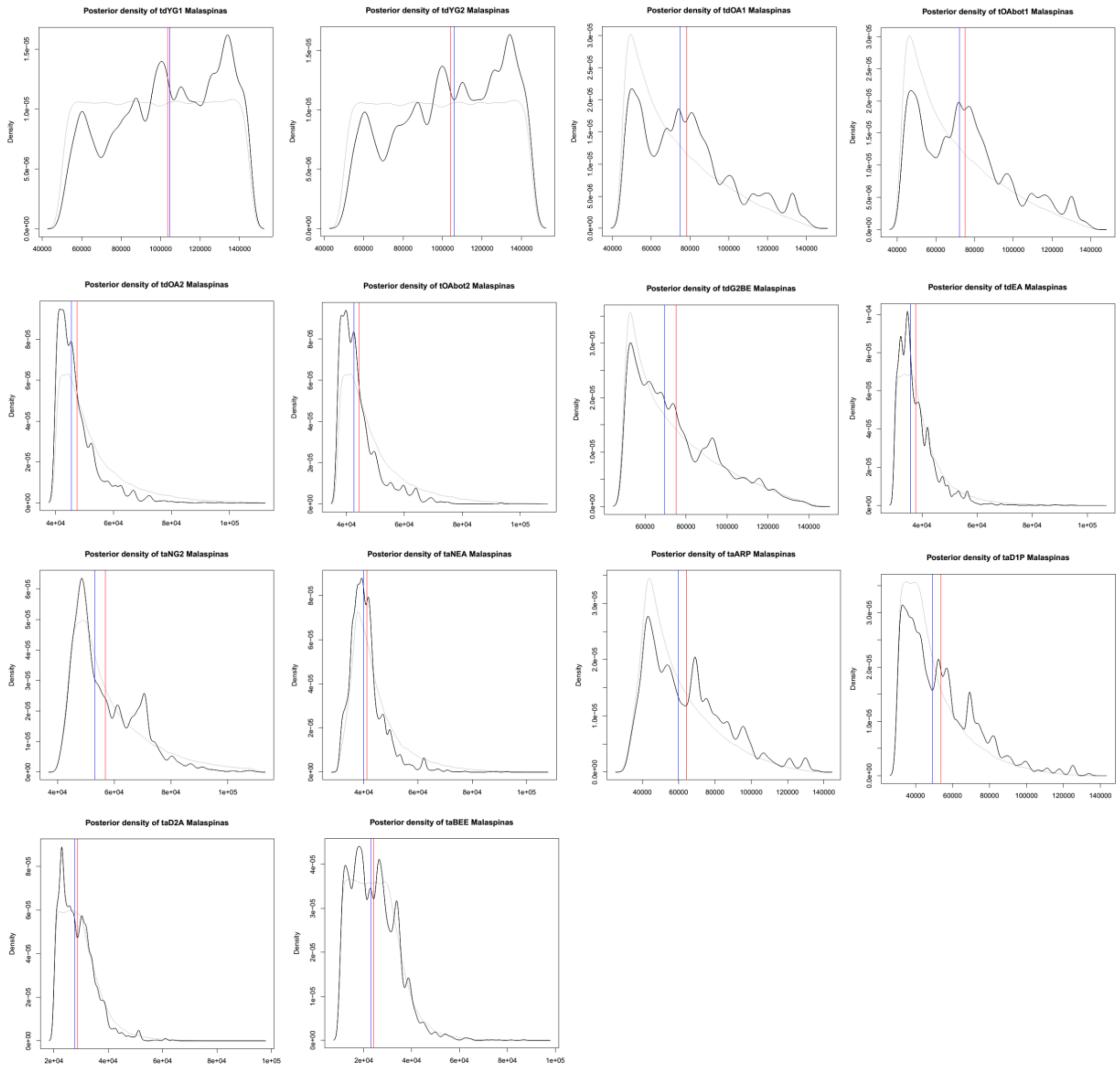


Figure S4. **Posterior density of the admixture rates estimated using the Papuan sample from Malaspinas et al. (2016)**. The plots have the same features of Figure S2.

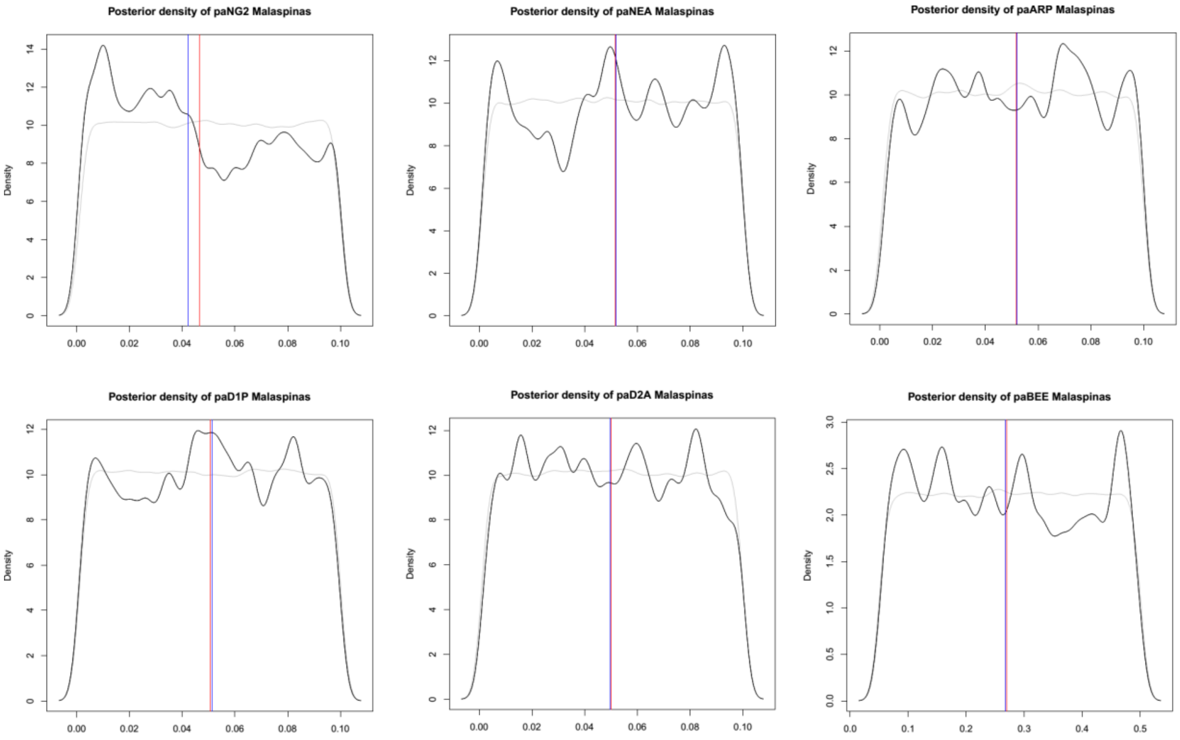


Figure S5. **Posterior density of the migration rates estimated using the Papuan sample from Malaspinas et al. (2016)**. The plots have the same features of Figure S2.

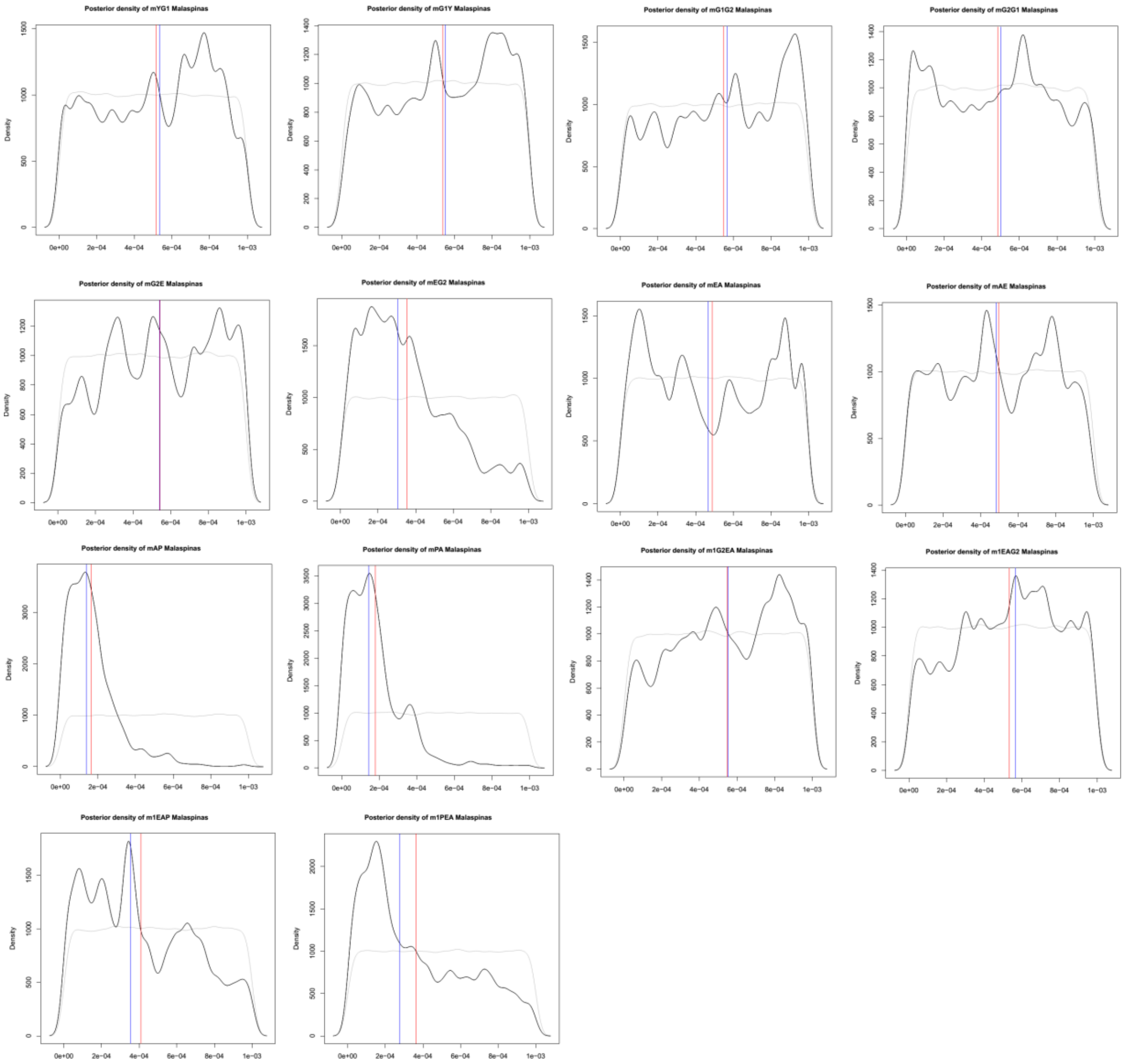


Figure S6. **Posterior density of the effective population sizes estimated using the Papuan sample from Pagani et al. (2016)**. The plots have the same features of Figure S2.

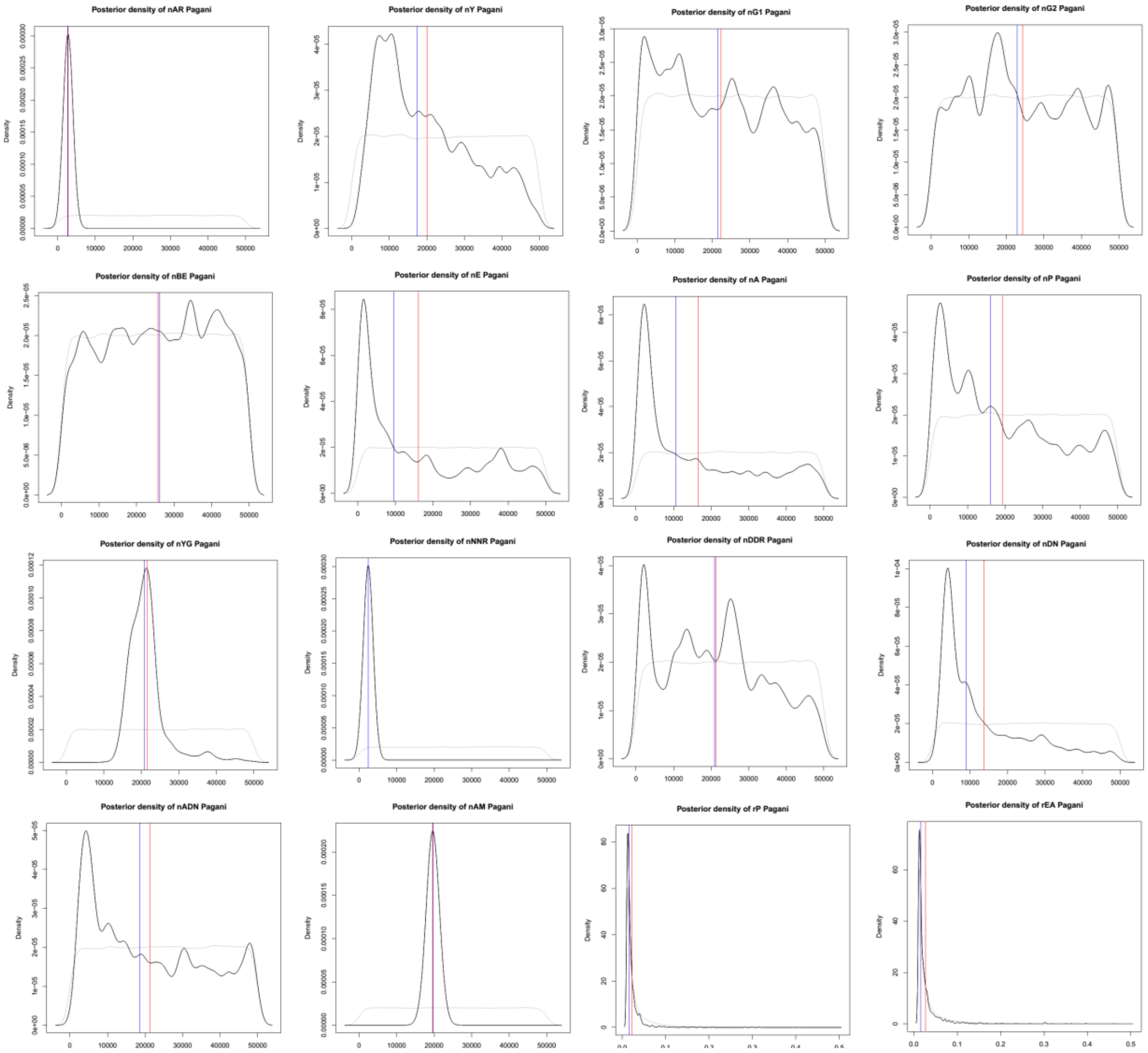


Figure S7. **Posterior density of the divergence times and the admixture times estimated using the Papuan sample from Pagani et al. (2016)**. The plots have the same features of Figure S2.

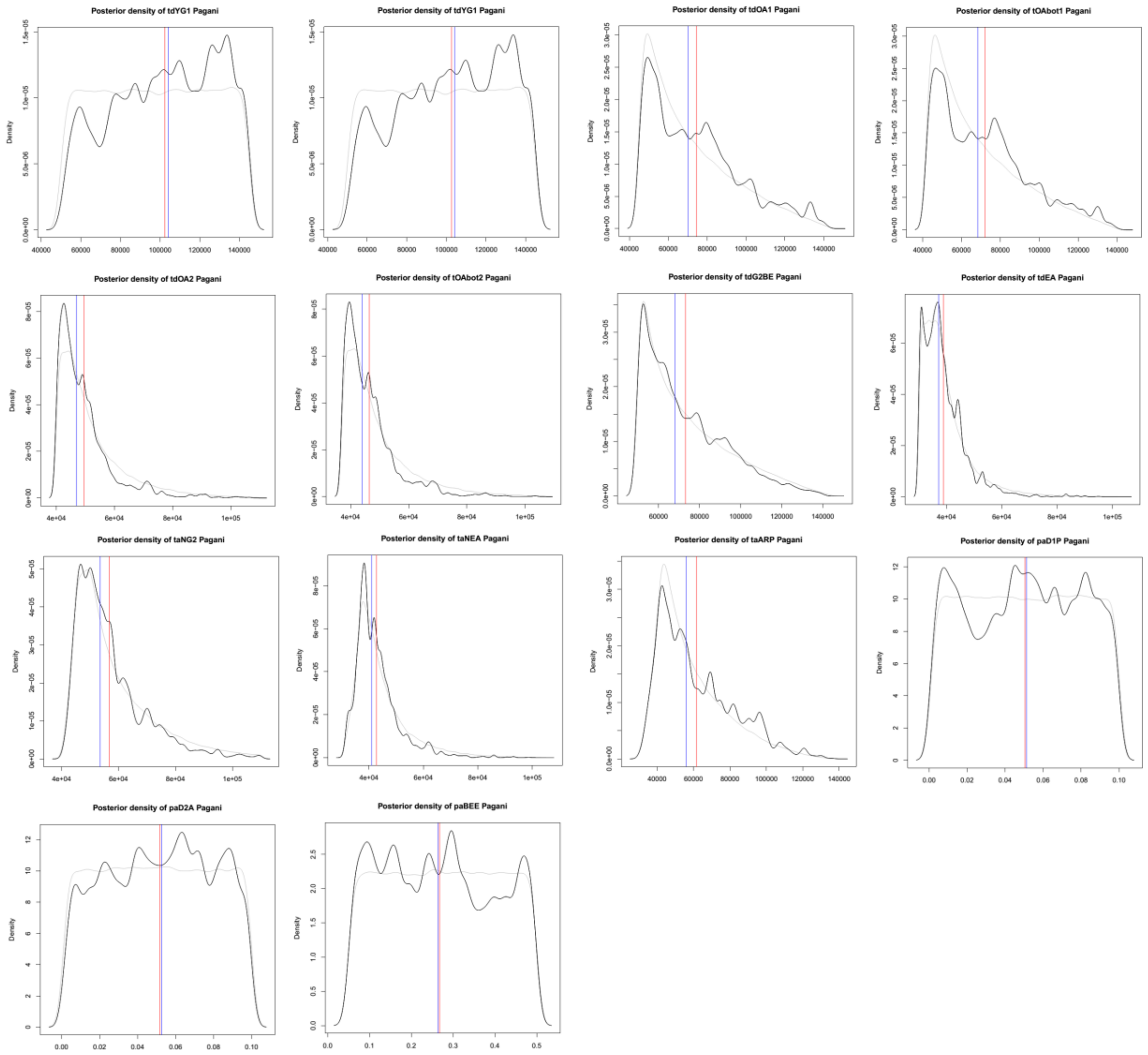


Figure S8. **Posterior density of the admixture rates estimated using the Papuan sample from Pagani et al. (2016)**. The plots have the same features of Figure S2.

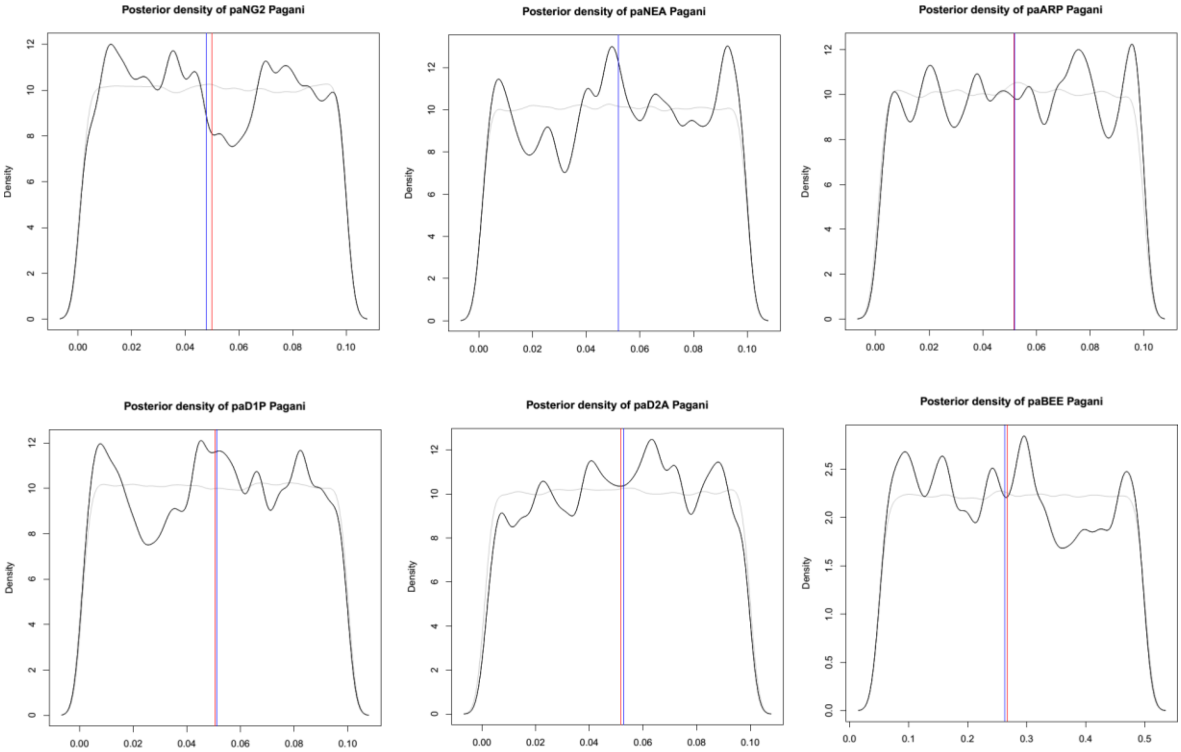


Figure S9. **Posterior density of the migration rates estimated using the Papuan sample from Pagani et al. (2016)**. The plots have the same features of Figure S2.

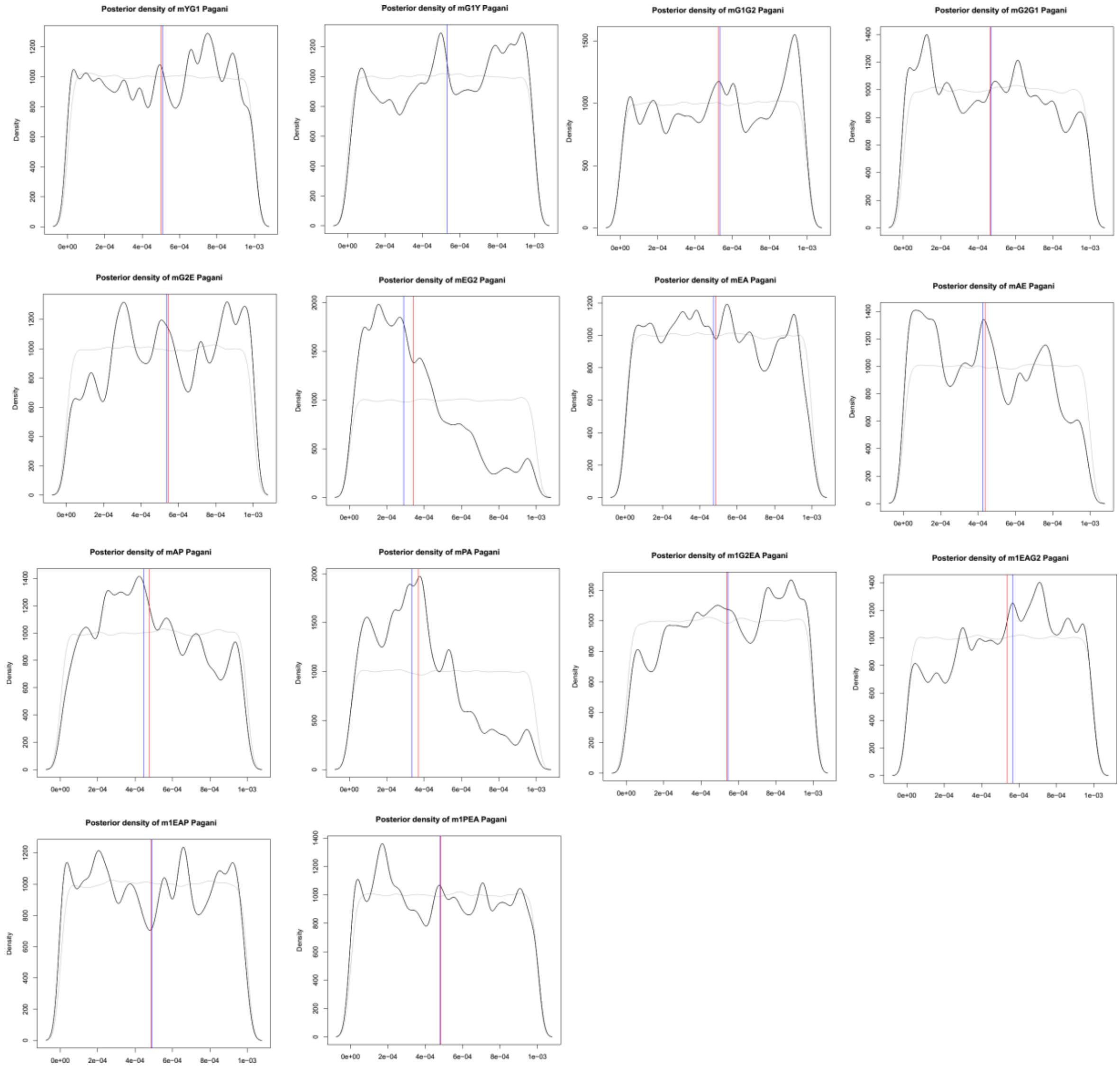
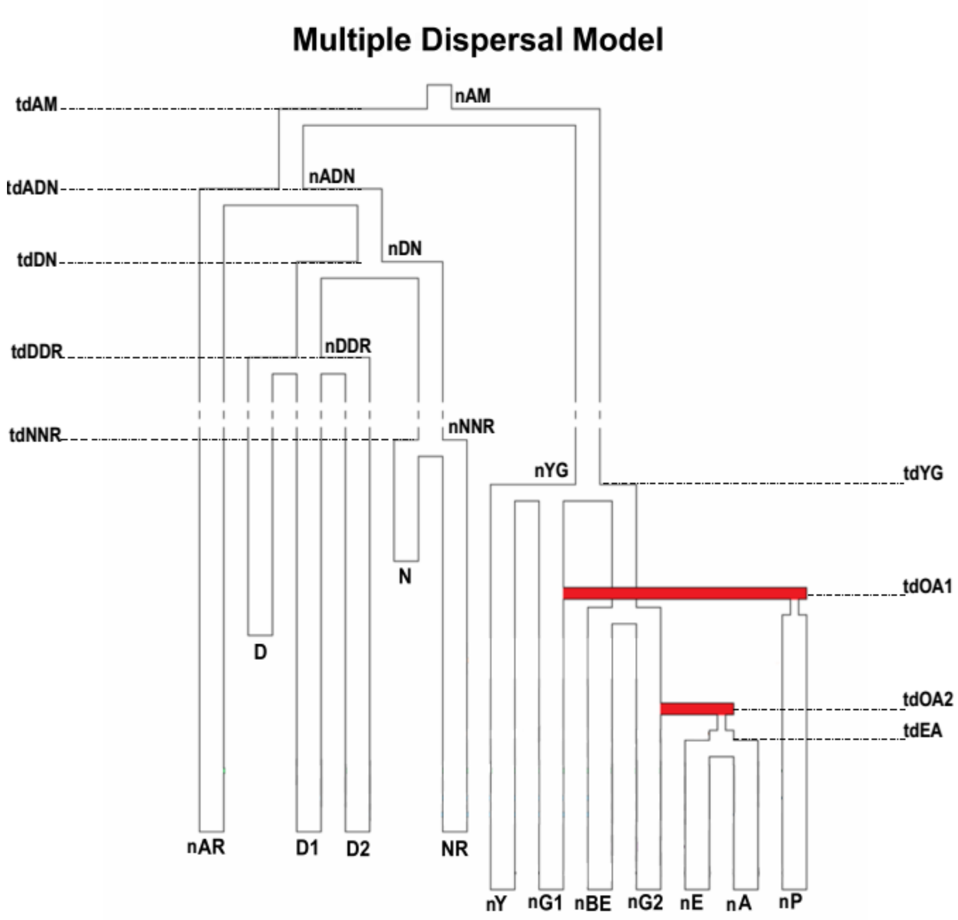


Figure S10. The model below represents a simplified version of the most supported model (MD) showing the main demographic parameters. To ensure readability migrations and admixture events are not shown.


Multiple Dispersal Model
tdAM
nAM
tdADN
nADN
tdDN
nDN
tdDDR
nDDR
tdNNR
nNNR
nYG
tdYG
N
tdOA1
D
tdOA2
tdEA
nAR
D1
D2
NR
nY
nG1
nBE
nG2
nE
nA
nP