Demographic inference in complex populations: the North African case.

Jose Miguel Serradell Noguera

TESI DOCTORAL UPF / 2023

DIRECTORS DE LA TESI

Dr. David Comas
Dr. Oscar Lao

DEPARTAMENT DE MEDICINA Y CIÈNCIES DE LA VIDA

Universitat Pompeu Fabra Barcelona

“Never be certain of anything. It's a sign of weakness.”

The Doctor – The Face of Evil: Part One

AGRADECIMIENTOS

Los cuatro años que ha durado esta tesis han sido probablemente los más caóticos posibles para llevar a cabo cualquier proyecto de esta magnitud, así que quiero aprovechar esta página para agradecer a todas las personas que han estado ahí durante este tiempo y que han hecho un poco más agradable el camino. Ahora que me he quitado de encima a todo el resto podemos empezar con los agradecimientos específicos que son para lo que habéis venido, cotillas.

Primero quisiera agradecer a David y a Oscar su labor, a veces la comunicación no ha sido lo más fluida posible (entono el *mea culpa* por la parte que me toca), pero que sepáis que os considero grandes mentores y siempre he sentido que estabais allí para lo que necesitara. A continuación, quisiera agradecer a toda la gente del grupo David-Francesc-Elena que ha pasado durante estos cuatro años, aquellos con los que compartí poco por la pandemia y los que acaban de llegar. Al resto… Empecemos con Julen, gracias por las charlas de futbol y baloncesto durante el curro, no sabéis lo complicado que es encontrar a alguien con quien hablar de deportes hoy en día, jajajaja. A Jorge le tengo que agradecer, además de ser un gran colega y estar siempre dispuesto a ayudar que me haya descubierto que hay alguien que pueda ser más pupas que yo. Laura y Nerea (sé que no forma parte del grupo, pero venía más que Marcel) gracias por las charlas en los sofás, me habéis ahorrado una pasta en terapia y por darle un poco de vida al lab. Del grupo quisiera también agradecer especialmente a Gerard, Lara y Marcel sin los cuales esta tesis probablemente nunca hubiera visto la luz del día.

Fuera del laboratorio quisiera agradecer a Aitor, Xavi, Miguel y Llorenç por compartir piso conmigo y aguantarme los últimos cuatro años gritarles a 11 inútiles persiguiendo una pelota por la tele. También quiero al resto de Ludópatas, Alejandro, Marc y Sandra por ser amigos y compartir charlas de medianoche, viajes y fiestas que han hecho más entretenido estos cuatro años, y los que quedan.

Los próximos dos agradecimientos son especiales. Primero quiero agradecerte a ti, Mireia por aparecer en vida a mitad de esta tesis y darle luz a mi vida. Gracias por estar ahí, sobre todo estás últimas semanas en las que he estado más ausente de lo habitual, que ya es mucho. Te amo. Y segundo, quisiera agradecer a mi familia, a Pau y en especial a mi madre. Gracias por cuidar de mí, criarme y apoyarme incondicionalmente en todo, este logro es principalmente tuyo. Sé que no te lo digo lo suficiente, pero te quiero mamá. También quiero agradecer a mi padre, que sé que estaría muy orgulloso de esto y que allá donde esté cuida de nosotros.

Para terminar, y como no me gustaría que empezarais a leer la tesis con una lagrimilla en la mejilla, quisiera hacer una ronda de agradecimientos a gente que no sabe ni que existo, pero me mantiene cuerdo. Empezando por todo el Futbol Club Barcelona, algún día me provocareis un ataque al corazón, pero así se disfruta más la victoria. También agradecer a la BBC y Marvel Comics por darme Doctor Who y Spiderman para pasar las tardes y finalmente al Mortdog por casi hacer que esta tesis no sea posible por culpa de tu videojuego.

ABSTRACT

Since the first peopling of the region, multiple demographic events have occurred generating a complex genetic landscape in North Africa characterized by high genetic heterogeneity and constant gene flow from surrounding regions. In this thesis we have attempted to condense this in a single demographic model that could explain the diversity observed in North Africa. The analysis reveals clear different demographic histories for both the majority groups in North Africa, Imazighen & Arabs, pointing to a back-to-Africa, Upper Palaeolithic origin for the first and an Arab expansion origin for the latter. Moreover, the results points to continuous soft splits as drivers of divergence rather than hard splits followed by strong, punctual admixture events. This thesis presents advances on the exploration of the demogenomics in very complex populations, centered on the North Africa scenario.

RESUMEN

Desde los primeros pobladores de la región, múltiples eventos demográficos han generado un paisaje genético complejo en el norte de África, caracterizado por una alta heterogeneidad genética y un flujo constante de genes desde regiones circundantes. En esta tesis, hemos intentado condensar esto en un solo modelo demográfico que explique la diversidad observada en el norte de África. Los análisis revelan historias demográficas claramente diferentes para los dos grupos mayoritarios en el norte de África, los Imazighen y los árabes, señalando un origen en el Paleolítico Superior como consecuencia de un movimiento de vuelta a África para los primeros, y un origen como resultado de la expansión árabe en la región para los segundos. Además, los resultados indican migraciones suaves continuadas desde la separación de dos poblaciones como impulsores de la divergencia en lugar de divisiones fuertes seguidas de migraciones fuertes y puntuales. Esta tesis presenta avances en la exploración de la demogenómica en poblaciones muy complejas, centrada en el escenario del norte de África.

PREFACE

The study of human societies has been of focal interest since ancient Greece, to the point that we have a specific field in science exclusively dedicated to it, Anthropology. In anthropological studies we explore the origin and evolution of *Homo sapiens* from several areas of knowledge such as archaeology, linguistics, philosophy, or biology in the most clear and unbiased way possible. Biological anthropology focuses on how the biological characteristics of living people are related to their social and cultural practices and can be divided in physical and molecular anthropology, with the latter describing how these sociocultural events affect humans at the most basic level, how they are able to alter the genetic structure of a population. Population genetic studies intend to explore and analyze these genetic differences among human individuals and populations in a quantifiable and statistically significant manner. Demogenomics is the field of population genetics that utilizes genomic data to infer the demographic processes that led to the patterns of genetic diversity observed in a given population. It gathers them all together in a null demographic model that allow us to understand the history of a population (bottleneck events, population expansions, migrations…), set the neutral background to perform selection studies, and to have a model that can be used in conservation analysis to preserve the genetic diversity of a species.

In this PhD thesis we aim to apply demogenomics to the North African populations to explain the origins and effect of the different demographic events to model the genetic landscape observed in the region. Although historically underrepresented in population genetic studies, North Africa is of high interest for demogenomics. Its geographical location, isolated to

the south by the Sahara Desert, already differentiates its population history from the rest of Africa. On top of that, due to its proximity to Europe and the Middle East multiple migrations towards the region have occurred through the course of history. This extensive movement of people across the region has led to a very complex pattern of genetic diversity that presents an interesting and challenging opportunity to explore the effect of demographic events into the genetic landscape of a population.

By applying different machine learning algorithms and developing some new demographic inference methodologies we tried to disentangle how the effects bottlenecks, population expansions or migrations shaped the current populations in North Africa, represented by Amazigh and non-Amazigh groups, which was their origin and how they interacted with their surroundings.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Human Population Genetics

Population genetics is the branch of biology that deals with explaining the patterns of genetic diversity between populations, how they originated, and how they change over time. Through the integration of genetics and statistics the field unveils the intricate processes that have shaped the diversity of populations over time. To examine the genetic makeup of individuals and populations it's key to understand the effect of the four basic evolutionary forces and how they shape the population dynamics of genetic variants with a given demographic history, spatial structure, and mating system. These four drivers of evolution are:

1. **Mutation**: A mutation is a change in the DNA sequence of an organism. Mutations are a result of errors in the DNA replication during cell division, exposure to mutagens or viral infections. If this happens in germline cells, the new allele is introduced to the population and becomes an agent in evolution.
2. **Natural selection**: Differential transmission of alleles from one generation to the next due to functional differences that favor/hinder some genotypes over others resulting in an increased reproduction success (fitness) of some individuals over others.
3. **Genetic Drift**: Differential transmission of alleles from one generation to another due to random sampling. In small populations the effect of drift is stronger and can lead to fixation of non-beneficial alleles.
4. **Gene flow**: exchange of genetic material between populations because of migration of individuals. Allows the introduction of new alleles and phenotypes to a population.

Approximately 0.08% of the nucleotide base pairs in human DNA vary among individuals from which only 15% seem to be population specific (Relethford & Harding, 2001). Human population genetics, analyses this diversity and infers the evolutionary processes that led to it. The study of genetic variation in humans mainly tries to answer three questions: What does the level of variation imply about the genetic structure of a population? Can we reconstruct the history of a population? and how is this diversity affected by the local environments?

To answer these questions, population genetic studies apply statistical, biometrical, biochemical and bioinformatical procedures to the analysis of genetic markers. The first studies involved the observation of phenotypes, like eye colour or protein markers such as ABO blood groups system (Yamamoto et al., 2012) and HLA antigen analysis (Thorsby, 2009). From there, molecular markers began to be fundamental in population genetics. First with microsatellites or short tandem repeats (STRs), later with single nucleotide polymorphisms (SNPs), both allowing the exploration of variation at a fundamental state.

Uniparental markers like mitochondrial DNA (mtDNA) and Y chromosomal DNA, allowed the study of genetic history with a more direct interpretation than autosomal data due to the lack of recombination and ease of phylogeographic inference (general reference of mtDNA and Y). Nonetheless, they present limitations especially due to that uniparental markers represent a limited fraction of the genome, meaning that the genetic history analyzed is the one specific to the marker and can lead to incomplete results in the genetic history of the population, moreover due to its small size and high variability present a lower effective population sizes contrary to other genetic markers.

Autosomal markers, in contrast, provide higher resolution and no sex biased studies as a trade-off of a higher computational cost, and the need to consider recombination when analysing the data. In autosomal studies, as the segment of the genome analyzed is bigger, the number of polymorphisms increases dramatically from a few hundred ,in mtDNA, up to millions in the bigger genome-wide arrays currently available (Affymetrix & Inc, n.d.). Genome-wide array data allows the exploration of the genome by genotyping sets of SNPs specifically defined to tackle concrete problems, although not covering all the sites like in whole-genome sequencing. In human population genetics the most widely used array is the Human Origins Array that has been uniquely catered to ancestry specific analysis with over 900 hundred current worldwide individuals from multiple populations and archaic hominids as reference (Affymetrix & Inc, n.d.). Array-based analysis presents inherent biases depending on the criteria used to design the array. For example, an array design to target a specific disease would have a different set of variants that one used in population structure analysis. On top of that, arrays tend to be biased towards more represented populations (European ascertainment bias). This ascertainment bias could generate errors in the inferences when dealing with understudied populations (Eller, 2009).

Whole genome sequencing (WGS) tackles this limitation by exploring the whole extension of the human genome. This base-by-base view of the genome allows the discovery of previously unknown SNPs plus provides information on other kinds of genetic variation, such as indels, inversions, translocations or copy number variants. Another advantage that WGS presents over arrays is the exploration of rare variants, of special interest in biomedical studies or demogenomics (Pervez et al., 2022).

WGS techniques can be separated into three generations:

- Sanger sequencing: The original sequencing methodology. Based on the addition of labelled oligonucleotides to a complementary chain of DNA during the replication process. Produces medium-length reads (>500 bp) with up to 99.999% accuracy (Shendure & Ji, 2008) but is slow and expensive for large-scale projects.
- Second generation sequencing: Produce short reads (100-150 bp) with higher errors than Sanger (over 100 times higher) (Shendure & Ji, 2008), but at a higher throughput and lower cost. Currently, the go to methodology for WGS analysis.
- Third generation sequencing: Also known as long read sequencing, consists of reading the DNA molecule of a single molecule instead of breaking it into small fragments. With these methods we can generate longer sequence reads (> 10kb) at expense of a higher per-base cost rate and significantly higher error (Pervez et al., 2022).

In this thesis, 2nd generation WGS data is used in most of the analysis, specifically Illumina based methods to produce high coverage genomes. It is also important to, when available, try to homogenize the source of the genomes to minimise possible batch effects caused by the use of data from different sequencing technologies (Maceda & Lao, 2021).

1.1.1 Contextualization of population genetics

Population genetics is a crucial component of the study of human evolution; however, gaining a comprehensive understanding of this intricate process requires the integration of cultural and environmental studies into the analysis. Gene flow is a key evolutionary factor, but its interpretation goes beyond the mere exchange of genes between populations; it requires contextualization to grasp the underlying reasons for these genetic movements. Environmental factors, such as glaciation events or natural disasters, and cultural factors like wars or agricultural innovations, significantly impact the genetic composition of a population. Additionally, cultural elements like religion or language can influence the internal structure of a population. These various factors collectively challenge the detection of signals related to demographic processes within populations. Consequently, a holistic approach that considers the interplay between genetics, culture, and the environment is essential for a comprehensive understanding of human evolution (Creanza & Feldman, 2016)

.

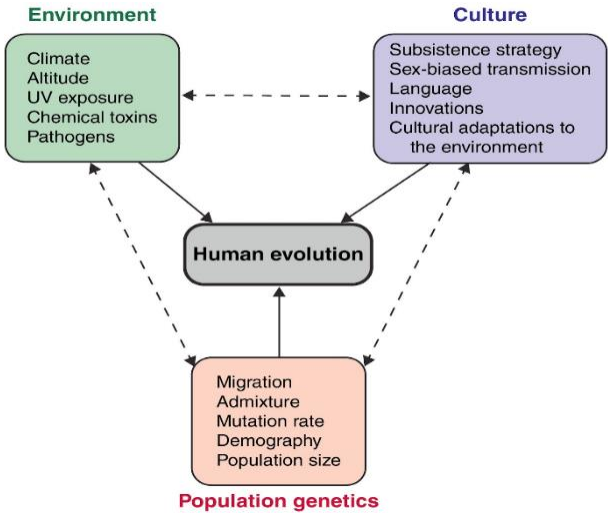


Figure 1: Genetic, environmental, and cultural factors influence one another, and all have an impact on human evolution*. Figure and caption from Creanza and Feldman 2016*

1.1.2 History of Human Population Genetics

The field of population genetics originated back in the middle of the 20th century when researchers began exploring the genetic basis of human variation. The rediscovery of Gregor Mendel's work (Mendel, 1886) on the inheritance laws, by botanists Hugo DeVries (De Vries, 1900), Carl Correns (Carl Correns, 1900) and Erich von Tschermark (Tschermak, 1900), presented the foundations of the discipline. Mendel's work was then confirmed by Thomas Hunt Morgan studies on Drosophila melanogaster (Morgan et al., 1922) who defined the inheritance units, named genes, kickstarting a century of marvels on the study of heredity and genetics.

The development of three ideas, extracted from the exploration of Mendelian heredity, before 1918 would become the basis of the fundamental works of Fisher, Wright, and Haldane, that later would be considered the steppingstones of population genetics. First, the Hardy-Weinberg equilibrium law back in 1908 (Hardy, 1908; Weinberg, 1908). The Hardy-Weinberg equilibrium states that alleles and genotype frequencies in a population will remain constant from one generation to the next if there is no influence from evolutionary forces (genetic drift, mutation, assortative mating, natural selection…), essentially defining the conditions under which there is no evolution. The second idea was the work on mathematical consequences of inbreeding mainly driven by H. S. Jennings (Jennings, 1914, 1916; Kimura & Crow, 1964), which in short, end up with a measure of linkage disequilibrium defined by Robbins (Robbins, 1918). And finally, we have the analysis of the effects of selection over many generations formulated by H.T.J Norton (Punnet, 1915).

In the following years, the contributions to the field were mainly dominated by Haldane, Fisher, and Wright. Fisher introduced the diffusion equations into population genetics (Fisher, 1922, 1923), Haldane developed an approximation of change of number of copies of very rare variants by branching processes (Haldane, 1927) and Wright presented a theory on the effects of random changes on small populations, what later was called genetic drift (Haldane, 1927). The contributions from these three scientists come together in the Wright-Fisher diffusion model (WF model). The Wright–Fisher diffusion is a central model for the temporal fluctuation of allele frequencies in a large population which assumes a population of a constant size, random mating, and non-overlapping generations in the absence of the effects of mutation, gene flow and natural selection. The WF model provides a tool for studying how the introduction of any complex evolutionary forces can affect a relatively simple model like in descriptions of coalescence theory, where a WF model is used as the standard model for the coalescent framework and is then modified to account for the effect of selection, changes in population size or migration (see in *1.2.1 - Coalescence Theory*).

Two milestones for genetic research occurred during the next few years. In 1944, Oswald Avery and colleagues discovered the DNA as a carrier of genetic information (Avery et al., 1944). Nine years later, the double helical structure of the molecule was described (Watson & Crick, 1953), allowing the development of molecular genetics. The introduction of DNA sequencing and the ability to analyse specific regions of the genome allowed scientists to explore the genetic variation within the populations.

During the 1960s population genetics continued to expand. From a theoretical perspective, there were two major achievements. On one side,

the work of Lewontin (Hubby & Lewontin, 1966) and Harris (H. Harris, 1966) on quantification of genetic variability by using electrophoresis of soluble proteins. On the other side, Motoo Kimura developed his neutral theory of evolution which shows that the probability of fixation of any variant is dependent on the effective population size (Ne) and a selection coefficient (s), proposing that genetic variation mainly arises as to the balance between mutations and genetic drift (Kimura, 1964). Kimura also proposed the "infinite sites model" (Kimura & Crow, 1964), explained later in this thesis (see in *1.2.1 - Coalescence Theory*). In a more technical aspect, in the 1960s the use of computer-based simulations became a more relevant in biological studies (Allen & Fraser, 1968; Garfinkel et al., 1964; Sheppard, 1969). In population genetics, one of the first simulation-based studies was the application of Monte-Carlo experiments to a stepwise mutation model comparing it to an infinite sites model (Kimura & Ohta, 1974). From this point forward, there has been a tight relationship between new dry lab methodologies and advances in population genetics.

In the 1970s the exploration of simulation-based studies continued (Kimura & Ohta, 1974) and it was also the decade where the first sequencing method was developed by Frederick Sanger (Sanger et al., 1977). Also, during the 70s it was defined that the difference between human populations constituted only a minor variation of the total variation between individuals (Lewontin, 1972).

The 80s marked an important advancement in population genetics, especially for this thesis, with the introduction of coalescence theory (Hudson, 1983; Kingman, 1982; Tajima, 1983). The coalescence treats a set of alleles as the product of a bifurcating genealogy, simplifying the

analysis of sequence variation under neutral models and allows estimation of parameters of interest by rapid simulation of populations that have undergone past population size changes or are subdivided (see in *1.2.1 - Coalescence Theory*).

The late 80s and early 90s reignited the population genetics field. On top of the theoretical aspects developed in the prior decades, molecular genetics advancements in the production of mtDNA sequences, array chips (Chee et al., 1996) and microsatellites (Litt & Luty, 1989) were a trigger to the exploration of diversity between human populations. In 1987, Rebecca Cann, Mark Stoneking and Allan Wilson proposed the out-of-Africa hypothesis by analysing the genetic diversity present in the mtDNA of 147 individuals, proposing the maternal origin of *Homo sapiens* in Africa (Cann et al., 1987). Another seminal work of human population genetics was the one produced by Cavalli-Sforza, Menozzi and Piazza in 1993 where they compiled all classical markers studies, defining the population structure of all major population groups (Cavalli-Sforza & Piazza, 1993). Finally, the 90s suppose the start of two major projects that would transform population genetics from then onwards; the extraction and characterization of ancient DNA molecules (Pääbo, 1989) and the Human Genome Project (HGP) that launched in the 1990 (National Human Genome Reasearch Institute, 2022).

The turn of the 21st century marked a transformative period for human population genetics, driven by the completion of the HGP (International Human Genome Sequencing Consortium et al., 2001) and the development of high-throughput genotyping technologies. These breakthroughs enabled large-scale studies, including genome-wide association studies (GWAS), which linked specific genetic variants with

complex traits and diseases, and large databases of WGS like the 1000 Genomes Projects (The 1000 Genomes Project Consortium, 2015) or the Simons Genome Diversity Project (Mallick et al., 2016) that defined fine-scale population structures and migration events. By examining the genetic architecture of diverse populations, researchers uncovered population-specific genetic risk factors and the influence of genetic background on disease susceptibility.

Demographic inference analysis also became common, especially in the 2010s. Some relevant papers on the matter explain the out-of-Africa expansion event (Gravel et al., 2011), some specific demographic histories, specially focusing on Austronesian populations (Malaspinas et al., 2016; Mondal et al., 2019) or African populations (Lorente-Galdos et al., 2019), ancient Neolithic (Marchi et al., 2022) and bronze age populations (Clemente et al., 2021) and the presence of weak population substructure at the beginning of *H. sapiens* (Ragsdale et al., 2023). Outside humans, demographic history analysis on other primates (Kuhlwilm et al., 2019; Pawar et al., 2023; Peter et al., 2010) and other mammals, like killer-whales (Foote et al., 2019) or grey wolfs (Bergström et al., 2022) among many others, have also gained relevance in the last few years.

1.2 Demographic inference in complex populations.

Demographic history inference involves reconstructing the past population dynamics based on genetic data from present-day populations. By analyzing patterns of genetic variation, we can make inferences about historical events such as population expansions, population contractions (bottlenecks as a combination of both), splits and unions of populations, admixture events and migration patterns. The next generation sequencing revolution supposes an exponential multiplication in the amount of genetic polymorphism data available and has allowed scientists the development of new inferential methods, many of them inspired by coalescence theory.

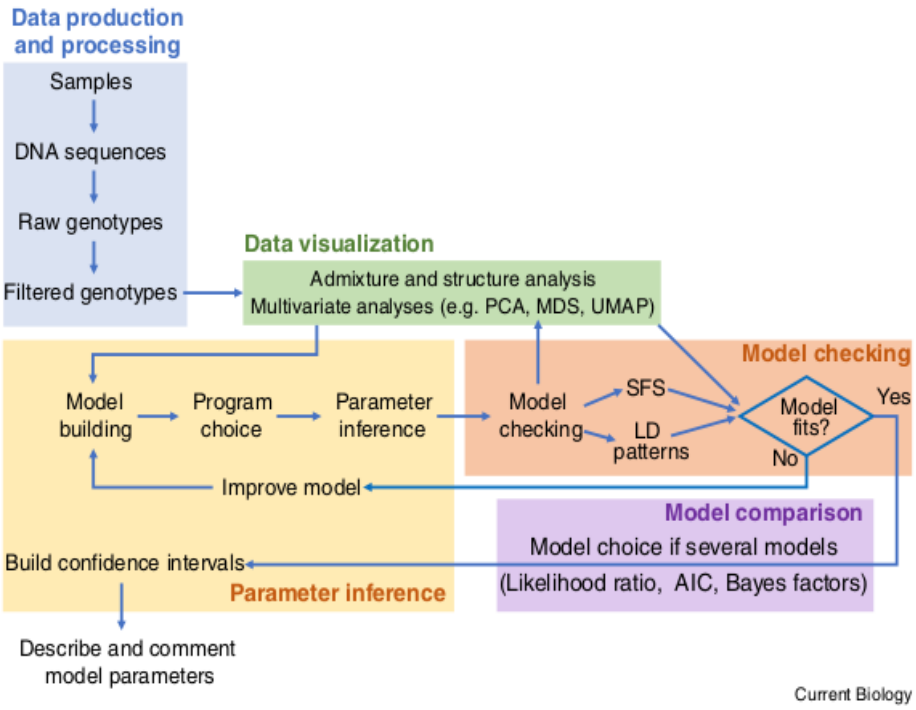


***Figure 2: Flow chart of the demographic inference process**. From Marchi et al 2021.*

1.2.1 Coalescence Theory

Coalescence theory is a theoretical framework that models the process of genetic lineages tracing back to a common ancestor. The main goal is to understand the forces which produce and maintain genetic variation. The coalescence was first described by Kingman (1982) (Kingman, 1982), but also discovered independently by Hudson (1983) (Hudson, 1983) and by Tajima (1983) (Tajima, 1983). The coalescent approach is based on two fundamental insights (M Nordborg, 2007):

The first insight is that since neutral variants do not impact reproductive success, we can separate the neutral mutation process from the genealogical process. To illustrate this, let's consider an example using a population of N clonal organisms that reproduce following the neutral Wright-Fisher (WF) model (Fisher, 1922; Haldane, 1927; Wright, 1937). In this model, each generation is discrete, and N parents are randomly selected with replacement and without mating. The number of offspring contributed by a specific individual to the new generation follows a binomial distribution with parameters N (number of trials) and 1/N (probability of selection). Now, let's examine the genealogical relationships in this reproductive context. When observing forward in time, lineages diverge when an individual produces more than one offspring and terminate when there are no offspring. Conversely, when tracing back in time, lineages merge or coalesce when multiple individuals descend -are copies- from the same parent. If we track backwards in time a group of chromosomes across generations, the number of distinct lineages will gradually decrease until reaching a single lineage, which represents the most recent common ancestor (MRCA) of the chromosomes under consideration. All of this happens independently of the neutral allelic differences between individuals, so we can model the

evolutionary dynamics of neutral allelic variants using coalescence and mutation dropping. The allelic state of any group of individuals can be generated by assigning an ancestral state to their MRCA and then 'dropping' mutations along the branches of the tree that leads to them.

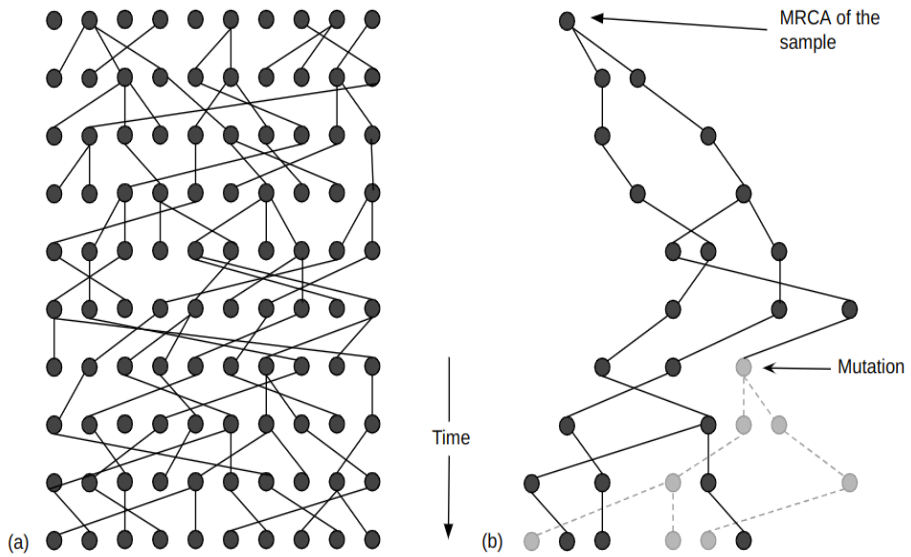


***Figure 3: Coalescence trees**. a) Neutral mutation process can be separated from the genealogical process. A realization of the genealogical relationships under a neutral Wright-Fisher model with N = 10. b) Genetic composition of a group of individuals is completely determined by the group's genealogy and the mutations that occur on it.*

The second insight is that it is possible to generate a genealogy of a group of individuals backwards in time without worrying about the rest of the population. The genealogy of a group of individuals may be generated by simply tracing back generation by generation, keeping track of coalescences between lineages until we found the MRCA.

These two insights help us realize that the pattern of neutral variation that we can observe in a population can be viewed as the result of random mutations on a coalescence tree. Therefore, we can understand how

model parameters affect polymorphism data by understanding how they affect genealogies. Another crucial concept we extract from coalescence approaches is that no matter how many individuals we sample, there is only a single underlying genealogy to estimate meaning that we could obtain an equally good inference with a sample size of one than with a lot of data.

The original Kingman paper (Kingman, 1982) described the coalescence as a continuous-time Markov process that arises naturally as a large population approximation to the Wright-Fisher model.

$$\prod_{i=0}^{k-1} \frac{N-i}{N} = \prod_{i=0}^{k-1} \left(1 - \frac{i}{N}\right) = 1 - \frac{\frac{k}{2}}{N} + O\left(\frac{1}{N^2}\right) \text{ (eq.1)}$$

The equation above describes the probability that no lineages coalesce in a previous generation for k lineages. By employing this approximation, we describe the coalescent models of a sample of n haploid chromosomes as a random bifurcating tree. In this tree, the n-1 coalescence events (representing all coalescence events until reaching the MRCA of the entire population) are treated as mutually independent and exponentially distributed random variables. The trees resulting can be extremely variable, in topology and branch lengths. Due to the independent exponential distribution of branch lengths and the random selection of lineages for coalescence regardless of branch length, the number of potential trees escalates exponentially with an increasing number of individuals. Although we can obtain multiple coalescent trees from a set of individuals, there is only one underlying genealogy. Consequently, sampled gene copies from a population must be considered dependent, rendering the impact of increasing the sample size rather ineffective.

We can extend the coalescent function to calculate what would be the probability of two lineages coalescing under a population with constant size *2N* (diploid WF model) (Hahn, 2018).

$$P(k \rightarrow k-1) = \frac{k(k-1)}{4N} \qquad \text{(eq.2)}$$

This probability assumes that at most two chromosomes can choose the same ancestor in any generation and implies that larger samples of lineages have higher chances to present a coalescent event at any given generation. The next three equations are derived from this one and give more insight on the coalescent process. The first one, indicates which should be the time for which *k* lineages remain.

$$E(T_k) = \frac{4N}{k(k-1)} \qquad \text{(eq.3)}$$

The time between coalescent events is smaller as the number of samples increases (*N*). The second equation allows us to explore at which time all the lineages present in our sample coalesce. We can determine the average tree height by summing the waiting times at each *k*.

$$E(T_{MRCA}) = 4N\left(1 - \frac{1}{n}\right) \qquad \text{(eq.4)}$$

Finally, we can calculate the total length of the tree. The total length of a coalescence tree represents the number of mutations observed in the sample.

$$E(T_{total}) = \sum_{k=2}^{n} k\frac{4N}{k(k-1)} = 4N\sum_{k=1}^{n-1}\frac{1}{k} \quad \text{(eq.5)}$$

The model of coalescence we have presented is an approximation using the WF model of evolution, that assumes drift as the only driver of fluctuations of the allele frequency. After the original representation of the coalescence, multiple generalizations have been made that take into several levels of biological complexity not present under WF. The first generalizations were applied to take into account non-haploid organisms

and sexual differentiation (Möhle, 1998). After that other important generalizations to do demographic inferences include:

- **Variation in the Effective Population size (Ne)**

Although the coalescence is not robust to variation in population size, incorporating changes in the Ne is relatively easy. Let N(t) be the population size at t generations ago. Lineages are more likely to coalesce when the population size is small than in generations with large population size (Figure 4).

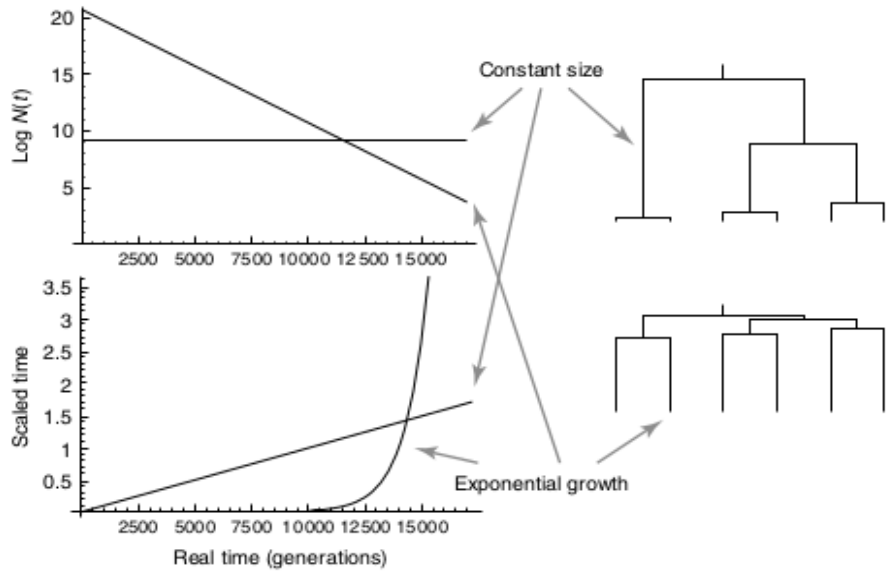


Figure 4: Variable population size can be modelled as a standard coalescence with a nonlinear time scale*. As the population that grows exponentially shrinks back in time, the scaled time runs faster making more likely the coalescence between branches when Ne is smaller. From Nordborg 2007, in Handbook of Statistical Genetics.*

- **Extensions to non-neutral models:**

Initially, the coalescent method was developed for neutral genetic variation. However, subsequent generalizations have incorporated various forms of selection, recombination, and other non-WF model processes into the coalescent framework. Both the inclusion of selection and recombination cause substantially increasing the complexity of modelling the coalescent process, especially selection. By definition, under selection some genotypes reproduce more than the others, because backwards in time there is no random picking of the parents. Even in this case, extensions to take into account selection have been proposed (Kaplan et al., 1988; Neuhauser & Kronet, 1997; SLATKIN, 2001). Recombination, on the other hand, profoundly affects the coalescence process. Each recombination event between two sites causes a rearrangement in the coalescent tree. The more recombination events there are, the more a genealogy changes. Each individual genealogy we obtain from a

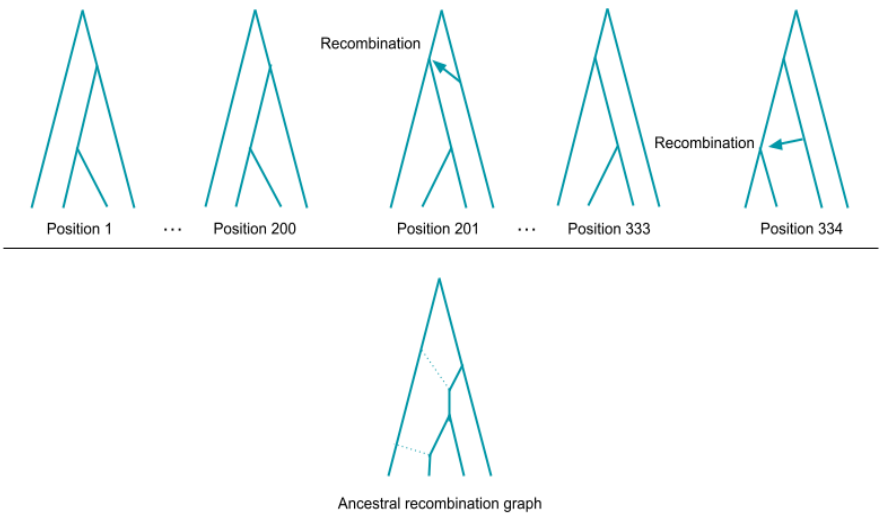


Figure 5: Effect of recombination on coalescent genealogies*. The ancestral recombination graph summarizes the three unique genealogies represented above.*

recombination event is called **marginal genealogy**. Given all the marginal genealogies we can construct an **ancestral recombination graph** (ARG) (Figure 5) that models the effect of recombination instead of a tree by representing the history of all the non-recombining segments of the genome (Griffiths & Marjoram, 1996; Hudson, 1983; Magnus Nordborg & Tavaré, 2002).

- **Multiple populations:**

The coalescent model has been extended to accommodate multiple populations or subpopulations. This allows for the analysis of population structure, migration, and admixture events.

- **Infinite-sites coalescent model:**

With the advent of high-throughput sequencing technologies, the coalescent method has been adapted to analyse genomic data, including whole-genome sequencing and genotyping data.

As the amount of data increases the new models need to consider these data arise. The best sort of model would be a four-state mutation model that could account for the differences in mutation rates among all four nucleotides and multiple mutations at the same site. However, due to the rare nature of polymorphisms the infinite-site models (Ewens, 1974; Wakeley, 2020) assume that multiple mutations at single sites do not happen. This approximation works especially well in low diversity species such as Humans. The incorporation of the infinite-sites generalization provides a simplified framework for estimating mutation rates and inferring evolutionary relationships based on the observed DNA sequence differences.

Under an infinite-sites model and the assumption that mutations occur randomly with probability μ per generation, the expected number of segregating sites in N diploid individuals in a sample is:

$$E(S) = \mu * 4N \sum_{k=1}^{n-1} \frac{1}{k} \quad \text{(eq.6)}$$

We can see that the expected number of segregating sites can be explained as the total length of a coalescent tree multiplied by the mutation rate μ. This highlights the relationship between the infinite site model and the coalescence and that, although the addition of more sequences increases the number of segregating sites, each extra sequence will contribute less and less to the total length of the coalescence tree.

The structure of coalescent genealogies can graphically represent the frequency of alleles and the Site Frequency Spectrum (SFS) *(see in 1.2.2.1 - The Site Frequency Spectrum)*. The exact topology of the genealogy will determine what allele frequencies are possible and the possible SFS. However, averaging over all genealogies we can predict the expected number of *segregating* sites at any frequency *i* with:

$$E(S_i) = \frac{4N\mu}{k} = \frac{\theta}{k} \quad \text{(eq.7)}$$

This equation implies that there are θ singletons, $\theta/2$ doubletons, $\theta/3$ tripletons, and so on, been the singletons the most common type of site and generating the expected graph of a Site Frequency Spectrum.

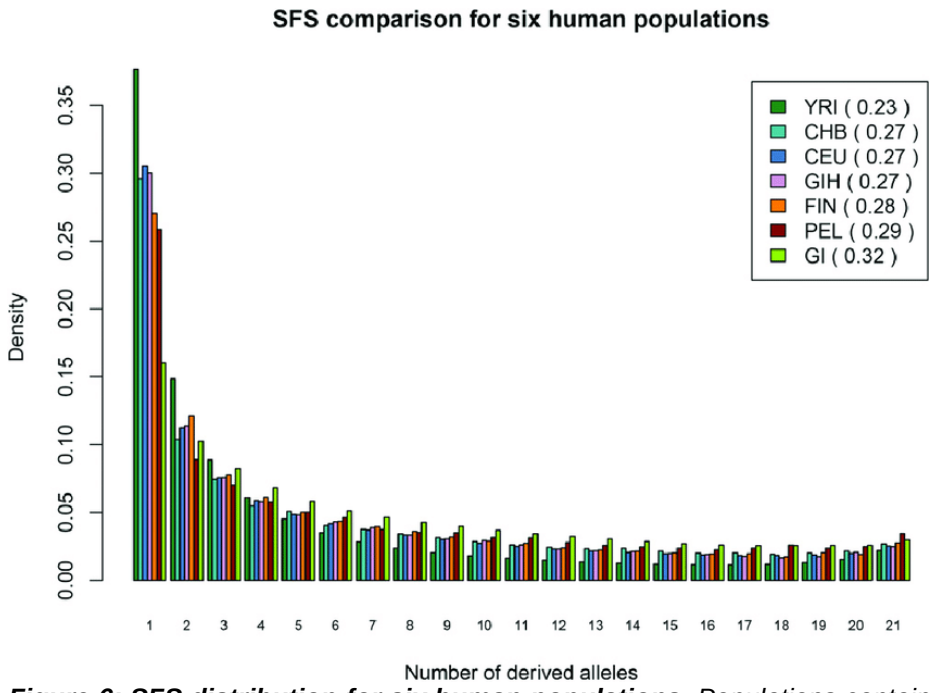


Figure 6: SFS distribution for six human populations. *Populations contain individuals with the following ancestry: Finns from Finland (FIN), Peruvians from Lima, Peru (PEL), Gujarati Indians from Houston, Texas (GIH), Utah Residents (CEPH) with Northern and Western Ancestry (CEU), Yoruba in Ibadan, Nigeria (YRI), and Han Chinese in Bejing, China (CHB) Each population is followed by a p var estimate per variable site. Figure and caption from Pedersen et al 2017.*

The coalescence theory has given us the theoretical framework to transform predictive population genetics into a rather inferential discipline. It allows the possibility to link past demographic events with genetic diversity by observing how migrations patterns, divergence times and changes in the effective population alter the calculation of MRCA that we can observe in a pair of neutrally evolving sequences. In the following section we are going to discuss different methodologies to perform demographic inferences and their pros and cons.

1.2.2 Methods of Demographic Inference

There is no straight recipe for sound demographic inference. The many different approaches and programs available, as well as the data, model selection and validation with observed data difficult our ability to do proper inferences.

Data selection is a crucial aspect of demographic inference analysis, and whole genome sequences (WGS) offer a valuable opportunity to enhance our understanding of demographic histories. Over time, WGS has become more accessible, enabling researchers to delve deeper into population genetics. Compared to other genetic data like SNPs microarrays or microsatellite data, WGS provides a higher resolution, facilitating more precise and detailed demographic inferences. Complete genomes allow the search of rare and structural variants across the genome (Hinds et al., 2005; T. L. Newman et al., 2006; Sharp et al., 2006). This information contributes to the study of populations at a finer scale and more in depth understanding of demographic history of specific groups. Moreover, WGS can identify subtle patterns of admixture enabling the study of historical interactions between populations and uncovering the demographic history of more complex populations. Finally, the popularization of WGS due to a decrease in the costs of production, has led to an increase in the amount of samples available per population. This larger sample size provides a more representative depiction of population diversity and enhances the robustness of demographic inferences.

Although the great advantages that whole genome data confers in the study of demographic history, it also presents some limitations mainly because of the high dimensionality of the data and the computational challenges that come with it. A typical solution to bypass the high

dimensionality problem is using some type of summary statistics (SS) (Tajima's D, F-statistics, nucleotide diversity - Π -, SFS, LD, …) that can capture the relevant information to carry out demographic inference analysis, with Site Frequency Spectrum (SFS) being the one more widely used in inference due to its high efficiency to capture large scale genome diversity and the fact that we understand how the underlying demography modifies the spectrum.

1.2.2.1 The Site Frequency Spectrum

The Site Frequency Spectrum (SFS), or Allele Frequency Spectrum (AFS), is essentially a histogram of the frequency of certain alleles in a dataset. In each bin in the histogram, we observe the proportion of sites in an alignment of multiple sequences with a given minor allele frequency (Figure 6). In a neutrally stationary population, the expected SFS relative frequencies are given by the expected segregating sites function (Eq. 7).

In a one-dimensional SFS (1D-SFS), the simplest form of SFS, we construct the histogram by assigning values to each Single Nucleotide Variant (SNV) depending on the reference allele. We give 0 if the SNP is homozygous for the reference allele, 1 to the heterozygous and 2 if homozygous for the alternative allele. Then we count how many SNVs show each possible minor allele frequency and build the histogram with it.

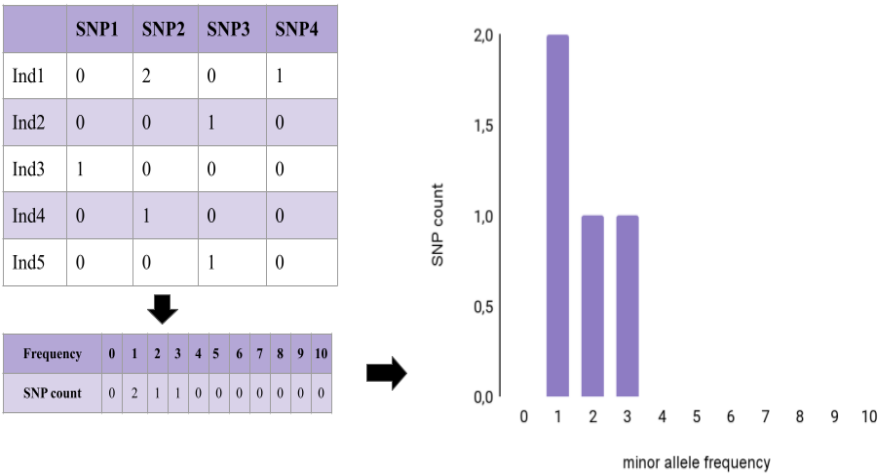


Figure 7: Diagram on how a 1D-SFS is built*. (1) Assign values to genotypes depending on minor allele presence. (2) Count the number of SNPs with different frequencies of minor alleles across the population. (3) Plot the SFS.*

For most demographic history analysis, instead of defining the bins relative to a reference sequence we do it by utilizing one or more closely related species to define the ancestral state of each site. By considering the ancestral state we could identify the source population during a population split event. This form of SFS is what is known as an *unfolded SFS*. In this thesis we use the *unfolded SFS* when we apply this summary statistic to our analysis.

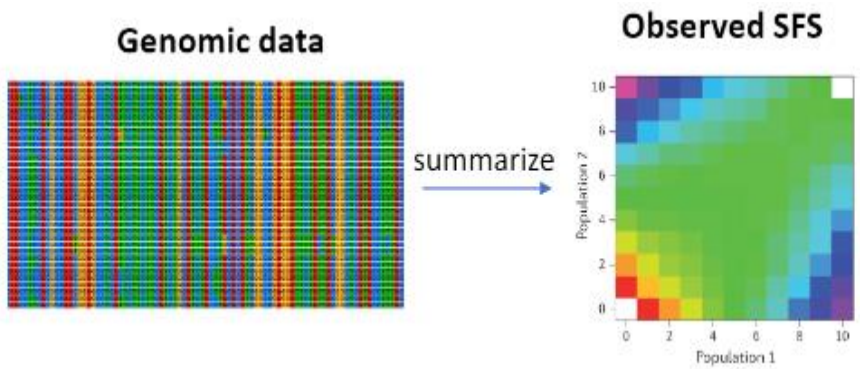


Figure 8: 2D-SFS extracted from a genomic dataset of 2 populations*. The total number of cells in the SFS matrix is 9. In a multi-SFS with 10 populations the total number of cells would be 59,049. Modified from Sousa & Hey 2013*

Sometimes we want to apply a SFS to compare between two or more populations. The SFS can be extended to two or more populations forming matrices where the entry (i,j), in the case of a 2D-SFS, corresponds to the count of sites with frequency i in population 1, and frequency j in population 2 (Sousa & Hey, 2013). Following this premise, we can theoretically build SFS that can compare as many populations as we need, but at some point we would again face a curse of dimensionality problem since the matrix increases exponentially as the number of populations increases (Blum, 2010) (Eq. 8).

$$SFS_{cells} = 3^n - 2 \qquad \text{(eq.8)}$$

By computing the observed SFS of a set of populations and comparing it to the SFS calculated from simulations we can infer demographic parameters and natural selection but we must consider possible caveats of this approximation. The first important limitation is that SFS ignores linkage disequilibrium information. The SFS captures genetic information from an independent set of SNVs. This, although it is valuable since every SNV becomes highly informative, misses the information extracted from analysing the LD patterns which can uncover fine structure of large, subdivided populations and recent migration events. In addition, the SFS minimax rate of convergence[1] is poor (Terhorst & Song, 2015), meaning that the number of samples needed to estimate a population size history function, for example, needs to be exponentially bigger to obtain a similar magnitude of error than other classical estimators such as density function or non-parametric regression (Terhorst & Song, 2015). Finally, Myers, Fefferman & Patterson (Myers et al., 2008) showed that for any population size function, one can construct many smooth functions

[1] Minimax theory provides a rigorous framework for establishing the best possible performance of a procedure under given assumptions. There are a set of techniques for finding the minimum, worst case behaviour of a procedure (Devroye & Lugosi, 2001).

generating the same population SFS. This results in the possibility of obtaining the same expected SFS for multiple trees which can be a problem when inferring the demographic history of a population if the sample size of the population is not sufficient (Baharian & Gravel, 2018; Bhaskar et al., 2015).

Once we have understood how the SFS works as a summary statistic for identifying diversity patterns across the genome, now we are going to see some methods that use this SS to do inferences about a population demographic history.

Inference from the SFS

There are three main methods that take advantage of the distribution of allelic frequencies to infer effective sizes, splitting events and migrations between populations. Those methods utilize different extensions and optimizations of the maximum likelihood function to compare the simulated data to the observed SFS.

The first method to infer demographic histories based on the use of SFS is the diffusion approximation for demographic inference (***∂a∂i***) from Gutenkunst et al. (Gutenkunst et al., 2009). ***∂a∂i*** efficiently simulates a multipopulation SFS by adopting the diffusion approach explained in (Ewens, 2000; Fisher, 1922; Kimura, 1964). Following, this approach utilizes a composite likelihood function[2] to compare the simulations with

[2] Composite likelihood is an inference function derived by multiplying a collection of component likelihoods; the collection used is often determined by the context, in this case, the linkage disequilibrium between sites. Because each individual component is a conditional or marginal density, the resulting estimating equation obtained from the derivative of the composite log-likelihood is an unbiased estimating equation (Reid et al., 2011)

the observed data to ensure that correlated (linked) allele frequencies do not bias the parameter estimation.

The next two methods, the stairway plot (Liu & Fu, 2015) and fastsimcoal2 (Excoffier et al., 2013, 2021), obtain the expected SFS probability using a coalescent based approach. The stairway plot method optimizes the observed SFS composite likelihood by means of a genetic algorithm *(see 1.2.3.1 – Genetic Programming)*. The algorithm is not restricted to specific demographic models and can explore a larger model space therefore allowing inference of more detailed demographic histories. The genetic algorithmic nature of the stairway plot makes it especially useful for exploratory and hypothesis-generating analysis.

Excoffier et al. (Excoffier et al., 2013, 2021) proposed the fastsimcoal2 framework as an algorithm that approximates the likelihood of the SFS by using coalescent simulations. Fastsimcoal2 uses coalescent simulations to estimate the expected SFS, and a conditional expectation maximization algorithm[3] (CEM) to estimate the parameters, one at a time over several optimization cycles. This approach has been shown to be very robust and can be applied to an arbitrarily large number of populations.

The use of SFS allows to explore complex scenarios and apply the inferences to large numbers of data, but lacks the information provided by linkage disequilibrium. Moreover, even though these methods can be extended to many samples, the SFS can become intractable and

[3] A CEM is an extension of the two-step Expectation Maximization (EM) algorithm. In the E-step, the algorithm computes the expected value of the complete-data log-likelihood function, given the observed data and the current estimate of the parameters. In the M-step, the algorithm maximizes the expected log-likelihood function with respect to the parameters (Dempster et al., 1977; Meng & Rubin, 1993).

simulation-based estimations become slower as the sample size increases, and the number of populations grow.

1.2.2.2 IBD & IBS based methods.

Identity by descent (IBD) and Identity by State (IBS) are two other types of summary statistics that, contrary to SFS, take into account linkage information between sites. IBS describes fragments that are identical, whereas IBD describes segments of the genome that have been inherited from a common ancestor without any recombination. The more closely individuals are related, the higher the percentage of their shared IBD, since they share a more recent common ancestor in their genealogical history compared to two randomly sampled individuals. As populations undergo both divergence and admixture over time, the lengths of IBD segments will gradually degrade due to recombination (Carmi et al., 2013; Palamara & Pe'er, 2013). Therefore, longer haplotype segments tend to indicate more recent relatedness, as there is a lower probability of recombination inducing a decay in their length over shorter periods of genealogical time (Henn et al., 2012). Demographic factors can be estimated based on the distribution of observed IBD in a contemporary population (Browning & Browning, 2015). Analysing IBD segments provides valuable insights into the demographic history and relatedness among individuals within a population. Some of the methods that make use of IBDs to infer demographic history are, GERMLINE (Gusev et al., 2009), RaPID (Naseri et al., 2019) or ILASH (Shemirani et al., 2021) among others. IBS methods also contribute to infer the parameters of a demographic model including population size changes, divergence events and admixture pulses giving accurate prediction where SFS inference methods, such as **∂a∂i,** failed to converge (K. Harris & Nielsen, 2013).

IBS has also been used to show that at least 2 migration pulses are needed to account for admixture from Europeans into Native American populations (Gravel, 2012).

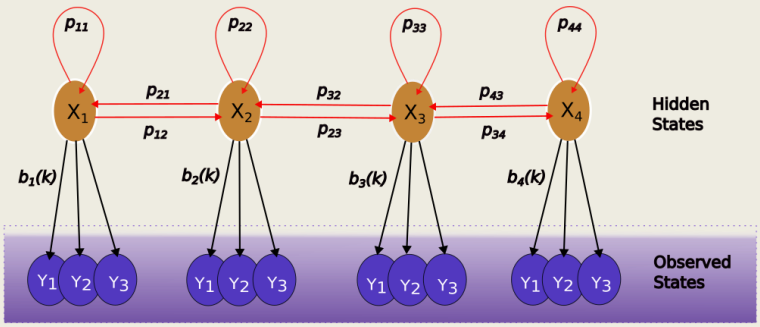
These methods rely on good ascertainment of the segments, which is specially complicated for small segments that correspond to historical events that occurred further in time. Most approaches that perform inference based on IBD or IBS blocks explicitly condition tract size to be larger than a specific cut-off when calculating the likelihood; thus, using a minimum length cut-off will not result in biased inferences (Sticca et al., 2021).

1.2.2.3 Markovian coalescent methods.

Markovian coalescent methods, pairwise (PSMC) and sequential (MSMC), are statistical approaches derived from the sequential coalescent (Wiuf & Hein, 1999), itself derived from the coalescent with recombination (Hudson, 1983), used to infer past changes in effective population size based on genetic data. Both methods are particularly useful when studying deep population timescales and when working with a very limited number of samples. Studies using this approach have been given great results in an ancient horse (Orlando et al., 2013), an ancient wolf (Skoglund et al., 2015) and two woolly mammoths (Palkopoulou et al., 2015), among many others (Mather et al., 2020).

Box 1. Hidden Markov models

A Hidden Markov model (HMM) consist of a double stochastic process, in which the hidden stochastic process (Xt) can't be directly observed, but can be inferred by analysing the sequence of observation symbols (Yt) of another set of stochastic processes that depend of Xt. The HMM can be represented as a stochastic model of discrete events and a variation of the Markov chain, a chain of linked states, in which the next state (Xt) depends only on the current state of the system (Xt-1). After Xt has moved to its new state, the value of Yt is generated by a probability distribution that depends on the value that Xt takes on that time.



Hidden Markov model with four hidden states (Xt) and three observed states (Yt). Each hidden state presents a probability *pt-1t* of transitioning from a hidden state (Xt-1) to another hidden state (Xt) and an observation probability *bt(k)* distribution for state t.

Extended from the sequential Markov coalescent (SMC) (McVean & Cardin, 2005), they make use of Hidden Markov models [Box1] under a coalescent framework that allows the tracking of the coalescent event between the two alleles at every locus in a single genome. By tracking this process in a limited time frame across the genome, PSMC and MSMC reconstruct the effective population size through time provided an assumption of the mutation rate (Mather et al., 2020) (Figure 9). The central concept behind these models is that recombination events of a population sample can be represented as transitions between the marginal genealogies along the genome. We can take a look at an ARG (Figure 4, Figure 9), each node signifies either coalescence or recombination events. The sequential coalescent associates each locus of the sample alignment with a genealogy embedded in the ARG. Consequently, a recombination

event in the ARG corresponds to a change in the genealogy along the genome.

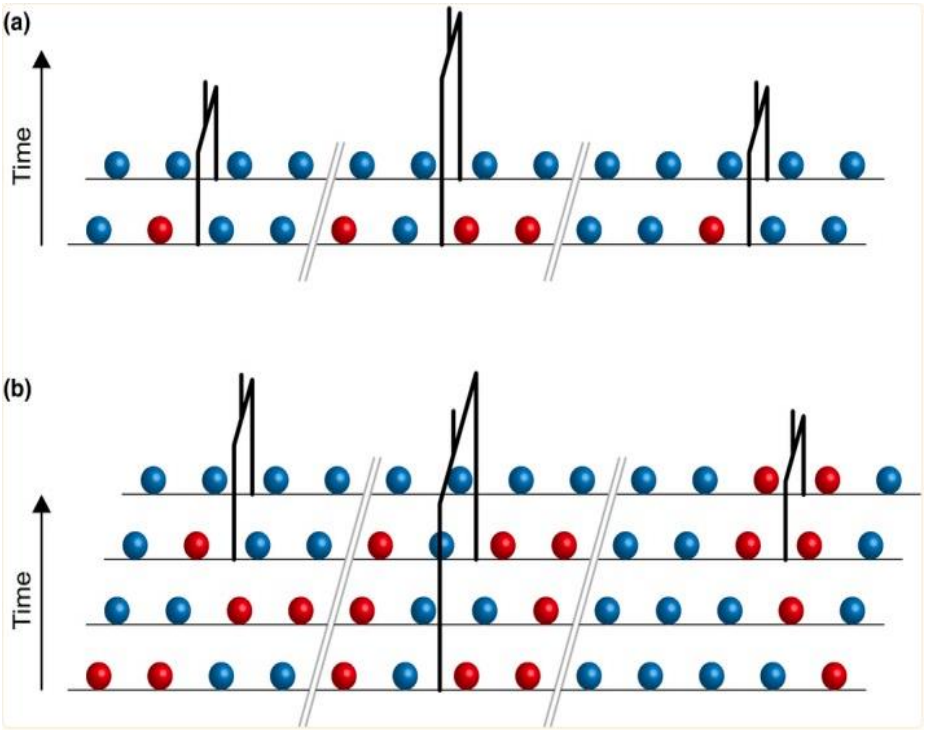


Figure 9: The sequentially Markovian coalescent. *The colored circles represent nucleotide states belonging to the alleles at each locus and double grey lines denote recombination breakpoints. The TMRCA of the two alleles at each locus is reflected in the local tree. (a) In PSMC, there are only two haplotypes. Thus, the topology of the local tree is fixed, but the time to the most recent common ancestor differs among loci. (b) In MSMC, there are multiple haplotypes. MSMC ignores most of the local tree topology, focusing only on the most recent coalescence event at each locus. From Mather et al. 2020.*

In the PSMC (Li & Durbin, 2011), one diploid sequence is needed as it uses two haplotypes to generate the inferences. From those haplotypes, PSMC infers the TMRCA based on the local density of heterozygous. From there the TMRCA at each segment/locus is used to create a

TMRCA distribution across the genome. And since the rate of coalescent events is inversely proportional to effective population size (Ne), PSMC identifies periods of Ne change. By analysing the fluctuations in coalescent rates across the genome, PSMC reconstructs changes in effective population size over time.

The MSMC (Schiffels & Durbin, 2014) extends PSMC analysing multiple genomes at the same time and calculating the MRCA of the two alleles that coalesce first a given locus. MSMC models the coalescent rates between pairs of lineages, considering the joint coalescence of multiple lineages, because this the input data must be phased. Phasing allows the method to distinguish between the two parental haplotypes and accurately infer the coalescent events occurring within each time interval. MSCM can be applied to estimate more recent evolutionary histories than PSMC, which is limited to 20-30 Kya.

Although these methods are robust and powerful for inferring changes in the population size through time, they do present some limitations we have to consider when deciding the method to use.

- Even though MSCM can deal with recent events better than PSMC, it still presents limitations in capturing very recent demographic events.
- SMC methods have limited sensitivity to migrations and very complex demographic scenarios. MSCM captures better migration events but still presents problems in very complex models, such as situations where population substructure is important.
- Both present strict cut off values in the sample quality to run reliably (Nadachowska-Brzyska et al., 2016).

Following the release of MSCM, new methods have expanded the SMC that outperform and deal with certain limitations of PSMC and MSCM. For example, the SMC++ allows a larger number of genomes and unphased data to be used, plus presents higher accuracy in recent past population size inferences as it incorporates the distributions of the allele frequency as a summary statistic of the remaining n-2 haplotypes (Terhorst et al., 2017).

The methods discussed so far rely on defining a likelihood equation to compare different models. This approach provides a strong statistical representation of the likelihood of a model in relatively simple demographic scenarios. However, it becomes challenging to apply these methods to very complex scenarios because the calculations and analyses become impractical and the equations that define the likelihood become intractable. To deal with this issue, at the end of the 20th century, methods based on Bayes statistics were developed. In the following section, we are going to explore the Approximate Bayesian Computation and some of its extensions to do demographic inferences.

1.2.2.4 Approximate Bayesian Computation

Approximate Bayesian computation (ABC) is a statistical framework extended from the Bayes Theorem (Bayes & Price, 1763) [Box2] used for inferring posterior distributions of parameters given a demographic model, and or compare models of interest if a function to estimate the likelihood of the data given the models does not exist. The first ABC-related ideas originated back in the 1980s, when Rubin (Rubin, 1984) described a hypothetical sampling mechanism that yields a sample from the posterior distribution. This description coincides exactly with that of

the ABC-rejection scheme. In 1997, two main studies settled the basis of ABC by proposing simulating artificial datasets and selecting among them by comparing summary statistics (SS) (Fu & Li, 1997; Tavare et al., 1997) and selecting those that have the same exact number of differences as the observed data. Tavare et al., 1997, explicitly can be considered the father of modern ABC as they introduced the Bayesian component in the inference as the $\theta = 4N\mu$ parameter was not fixed but sampled from a prior distribution. This first approach yielded two main problems that were addressed the next year by Weiss and Von Hassler (Weiss & Von Haeseler, 1998). The first problem is regarding the SS. Weiss & Von Hassler proposed that to capture all the information of the data a single SS was not enough. They addressed this by proposing that the distance between observations and simulations should be computed by combining multiple summary statistics instead of using just one. In the following years a lot has been written about the use of summary statistics, especially regarding the selection of optimum SS (Beaumont, 2019; Joyce & Marjoram, 2008; Marjoram et al., 2003; Nunes & Balding, 2010) and methods to overcome the SS selection bias have been developed that rely on the use of Deep Learning [Box 3] (Mondal et al., 2019). The second problem was about the retention of only those simulations that yield identical results to the observed data. If we only select those, a large proportion of the simulations we are using is discarded affecting the predictability of the methodology. Weiss and Von Hassler proposed that instead of using the distance = 0, between observed and simulated data, the selection criteria would depend on a threshold of tolerance (ε) that represents a quantile and depends on the investigator criteria that usually intends to minimize the bias-variance trade off (Beaumont, 2019; Beaumont et al., 2002).

Beaumont et al. (Beaumont et al., 2002) formalized and generalized the ABC approach. They introduced some improvements and evaluated the performance of the ABC against full-likelihood methods. But the main improvement was the introduction of a regression step. Beaumont et al. proposed that a linear regression between a parameter and the vector of SS, estimated using the retained simulations, with more weight on the simulations closer to the observed data, could modify the retained parameters thus mimicking a situation where all simulations produce SS equal to the observed values. This approach deals with some of the problems of the trade-off between bias and variance especially when the ε is high and the number of accepted simulation increases.

Box 2. Bayes Theorem

Bayesian statistics, named after the British mathematician Thomas Bayes, are a statistical school of thought that holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Bayes theorem allows to compute the posterior distribution for an unknown from the observed data and its assumed prior distribution.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

The Bayes theorem formula expresses the posterior probability of an event, P(A|B) as the modification of the prior probability of the event, P(A) by multiplying it to the result of a division between the likelihood of the observed data given the belief of the prior is true, P(B|A) and the marginal probability of the observed data, P(B). It's important to remark that to this formula to be true, events must be independent.

We can describe the ABC (Figure 10) in an eight-step process (Bertorelle et al., 2010; Sunnåker et al., 2013):

1 **Set the scene:** Bibliographic step where we define the models and the parameters of the models we want to compare. The parameters used to specify the models for an ABC analysis are the classic demographic parameter (Ne, migration/admixture rates, split

times…) and the genetic parameters (mutation and recombination rates).

2 **Incorporating the prior information**: Prior distribution of each parameter should be large enough to include all possible values. Prior should reflect previous knowledge on some parameters and can be fixated if we know it with relatively high precision. Defining the parameters and the prior distributions can be difficult in some cases and lead to biases and errors in the model ascertainment, so metaheuristic solutions could be implemented to search for those distributions *(see 1.2.3 - Metaheuristics & Genetic Programming)*.

3 **Simulating the models**: huge number of datasets must be simulated under the models defined to have sufficient power. The higher the complexity of the model and number of populations the higher the number of simulations. Nowadays, efficient simulators that produce hundreds of thousands of simulations on relatively complex models are available both backwards coalescence (fastsimcoal2 (Excoffier et al., 2021), msprime (Baumdicker et al., 2022)) and forward in time, SLim (Haller & Messer, 2023).

4 **Filtering the simulations:** Simulations are retained when the distance between observed and simulated SS is below a certain threshold (Ɛ) (Eq 9). This threshold is determined by the investigator and usually tends to minimize the bias-variance trade-off.

$$p(\theta \mid X) \approx p(\theta_i \mid d\,(SS_{sim}, SS_{obs}) < \varepsilon) \quad \text{(eq.9)}$$

5 **Model Selection**: The probability of a model is equivalent to the proportion of simulations from that model that are retained given Ɛ. Additionally we can calculate an index to assess how much better a model predicts the data than another model, the Bayes Factor. The Bayes Factor can be computed as the ratio between the posterior

probabilities of a pair of models divided by the ratio of their prior probabilities.

6 **Quality control and model selection:** The ABC can be used to investigate the robustness by simulating datasets under scenarios with known parameter values and comparing the estimations to the real data.

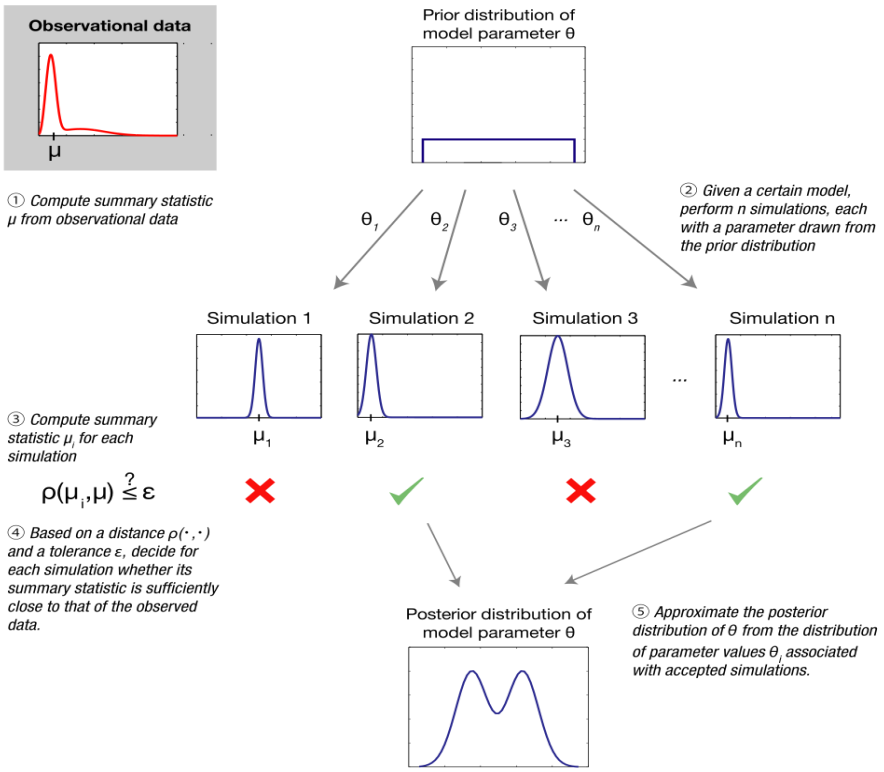


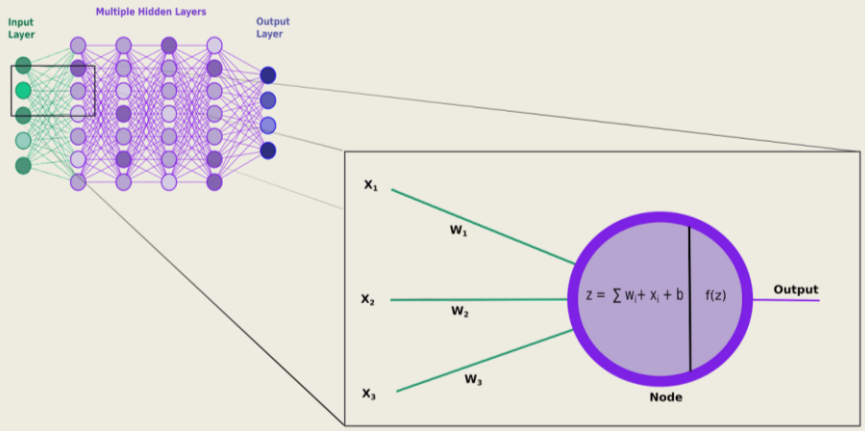
Figure 10: Conceptual overview of Approximate Bayesian Computation *from Sunnaker et al 2013. (1) Compute the SS that is going to be used to compare between the observed and simulated data; (2) Extract* n *simulations, each with a parameter drawn from the prior distribution; (3) Compute the SS for each simulation; (4) Decide for each simulation whether its SS is sufficiently close to that of the observed data based on a distance* ρ *and a threshold* ε *; (5) Approximate the posterior distribution of a parameter from the values associated to the accepted simulations.*

The 7th and 8th steps are modifications of steps 5 and 6 but specifically designed for parameter estimation (Bertorelle et al., 2010). Several programs have been developed that perform ABC analysis but the one that currently is more used and the one we use in this thesis is the "abc" package for R (Csilléry et al., 2012).

Box 3. Deep Learning

Deep Learning (DL) is a technique of Machine Learning (ML) which is a part of Artificial Intelligence (AI). DL layers algorithms and computing units - neurons - into artificial neural networks (ANN) that mimic the human brain enabling a computer to learn from observational data. DL is composed of simple but non-linear modules that each transform the representation at one level into a representation at a higher, more abstract level (Godfellow et al., 2016).

One of the most basic types of DL architecture is called multilayer perceptron (MLP). A MLP is an ANN where each neuron is connected to all the neurons of the upper and lower layer and is defined by a first layer, the input layer, where we have a vector X representing one data point in the dataset. After that, several hidden layers of neurons are added sequentially. Each layer is defined by their weight (w) and bias vector (b). The information of a layer is transmitted to the next by the means of an activation function (f(z)). It adds non-linearity to the network and transforms the value obtained, into an input to the next hidden layer or as an output (Awad & Khanna, 2015).



ANN have a wide range of applications across various fields, from facial recognition and image classification to music composition or climate modelling. In genetics, deep learning techniques has been applied to predicting the effects of noncoding variants (Zhou & Troyanskaya, 2015), detecting alternative splicing sites (Jaganathan et al., 2019) or predicting the protein 3D structures (Jumper et al. 2021). Specifically in population genomics, ANNs has been used in inference problems. Villanea and Scharaiber (Villanea & Scharaiber, 2019) successfully classify the interaction between neanderthal and two sapiens populations between five admixture models using an MLP. Mondal et al. (Mondal et al. 2019) incorporate a MLP to select summary statistics from the SFS in ABC inferences. For an exhaustive review on the use of deep learning techniques for population genetic inference studies see Huang et al. (Huang et al. 2023)

1.2.3 Metaheuristics: Genetic Programming

Model comparison by ABC requires defining which are the models to consider. These models are, by definition, simplifications of reality. However, basic assumptions about the demographic events, and particularly population substructure, can significantly bias the model ascertainment. To bypass this issue, in this thesis we have developed a new demographic inference method based on genetic programming (GP), a branch of evolutionary algorithms.

Evolutionary algorithms are a type of metaheuristic algorithms (MAs), which are optimization algorithms used to solve complex optimization problems that are not effectively solvable by traditional methods. Metaheuristics are problem independent processes that aim to guide the search process efficiently through the solution space (Sörensen, 2015). The objective is not to find the optimal solution, but rather to obtain a good solution in a reasonable amount of time for a problem that is too complex or too big. Any MA success depends on the proper handling of the exploration - exploitation trade-off. This dilemma has been a crucial issue in the field of metaheuristics since its beginnings. The exploration component of MA is responsible for the detection of the most promising regions of the search space. While the exploitation promotes the convergence of solutions (Sarhani et al., 2022). In other words, during the latter stage, the search is concentrated in a smaller space of the solution space, comparing between neighbouring solutions. Meanwhile, in the exploration stage, it is encouraged a more general search process across the whole solution space to examine unvisited regions and to generate solutions that differ in significant ways from those seen before.

There are multiple types of MA that can be classified in multiple ways. For example, we can classify the MA by the source of inspiration. Depending on the source of inspiration, MA can be classified into four categories; as swarm intelligence (SI) based algorithms, like Ant Colony optimization as physic-chemistry based algorithms, like Simulated Annealing algorithm; as Evolutionary Algorithms, like Genetic Programming and a fourth group comprised by the rest of algorithms identified as Miscellaneous (Rajwar et al., 2023) (Figure 11).

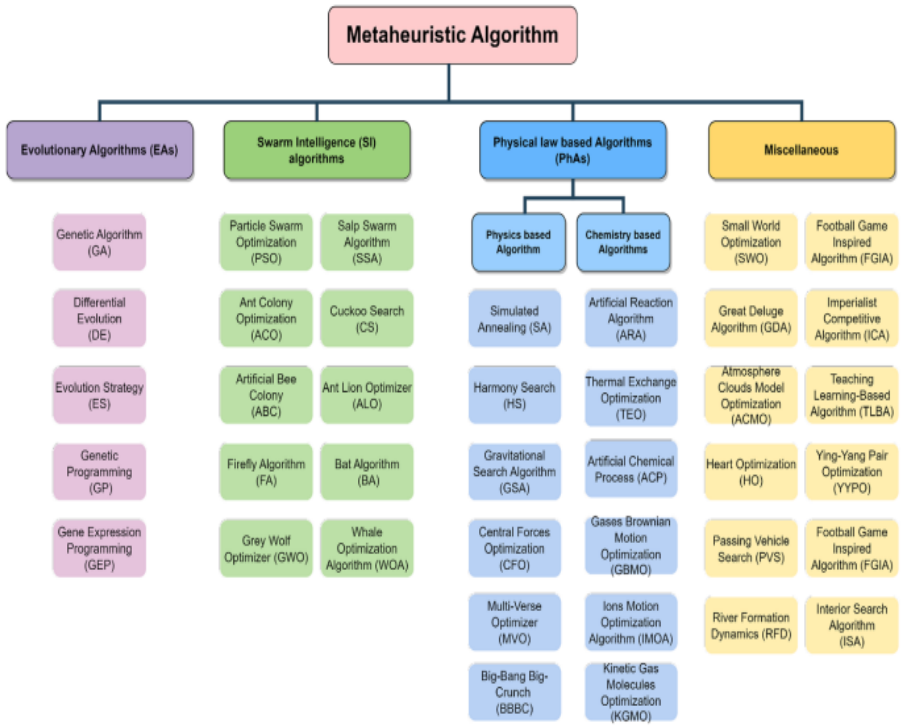


Figure 11: Classification of MAs based on the source of inspiration*. From Rajwar et al 2023*

Of the different types of MAs, we are going to focus on the ones inspired by the Darwinian ideas of natural selection, the Evolutionary Algorithms (EAs). EAs start with a population of individuals (solutions) and simulate sexual reproduction with recombination and mutation to create a

generation of offspring. This practice is repeated along a selection process that eliminates the weaker solutions while maintaining increasingly stronger offspring across the generations. Genetic algorithm (GA), differential evolution (DE), gene expression programming (GEP) or genetic programming (GP) are all subtypes of Evolutionary algorithms.

1.2.3.1 Genetic Algorithm & Genetic Programming

Genetic Algorithms were first introduced by John Holland in the early 1960s (Holland, 1962a, 1962b). In a GA each solution in the space (individual in a population) is represented as a binary coded string (chromosome) (Figure 12) with an associated fitness measure (Awad & Khanna, 2015). Successive solutions are built as combinations of the selected individual solutions. A standard GA follow a process that can be illustrated in the following pseudo-code:

```
generate initial population, G(0);
evaluate G(0)
t:=0;
repeat
     t=t+1;
     generate G(t) using G(t-1);
     evaluate G(t);
repeat until a solution is found
```

Figure 12: Basic pseudo-code that explains how a standard genetic algorithm works*. From an initial population G(0) evaluation on how close the population is to real data is done. After that, multiple iteration (t) are done where, in each iteration, the population G(t) is a modification of the one before G(t-1) and re-evaluated against the observed data until we reach a solution.*

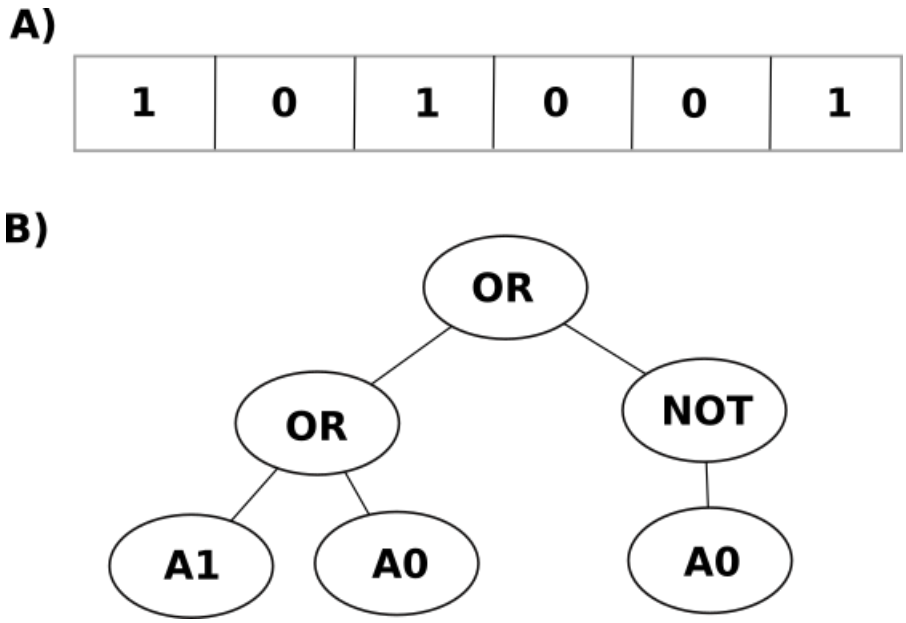


***Figure 13: Genetic algorithm vs Genetic programming structure**. A. Binary coded string that represents a chromosome in a GA. B. Tree structure of a GP program where each node is a parameter to infer.*

The first step implies a random generated initial population. Each individual of the population is represented by a string of characters. Next, we apply a fitness function to each one of the chromosomes to quantify the quality of the solution. Once we know each chromosome's fitness, a selection process takes place to select the individuals that are going to be parents of the following generation. There are multiple selection schemes, but here I am going to present some of the more used ones (Carlos A. Coello Coello, 2005; Sivanandam & Deepa, 2008):

- **Roulette Wheel Selection**: In this selection technique, each solution is presented as a slot in a wheel weighted in proportion to the individual's fitness value. Is a moderately strong technique, since fit individuals are not guaranteed to be selected for, but somewhat have a greater chance.

- **Random Selection:** Parents are selected completely random from the population.
- **Rank Selection:** The Roulette wheel selection will have a problem when the fitness values differ too much. For example, if the best solution fitness is 90%, the rest of the chromosomes will have very few chances to be selected. In rank selection, each solution receives a rank based on its fitness value. Under this scheme convergence is slower and diversity is preserved from one generation to the next, hence leading to more successful search of the space.
- **Tournament Selection:** This strategy provides selective pressure by holding a tournament among the solutions. The best individual from the tournament is the one with the highest fitness and is the one inserted in the mating pool. Multiple tournaments are held until the mating pool is filled. The mating pool at the end is composed of all tournament winners and has, on average, a higher population fitness than the original population.
- **Steady State Selection:** This technique works individual by individual, replacing the worst individual in the current generation with the best individual in the next generation. Only a few individuals are replaced in each generation. It is used in evolving rule-based systems in which incremental learning and remembering what has already been learned is important.

Following the selection of the solutions that are included in the mating pool. The offspring generation phase starts. During this stage two main mechanisms work as drivers of evolution (genetic operators) in a GA, recombination (crossover) and mutation. Mutation is an important genetic operator that randomly changes a gene of a chromosome. In a binary coded string, a mutation occurs when a 1 is changed to a 0 o vice versa. This operator introduces diversity to the population assuring the

exploration of the entire search space. The recombination operator creates new solutions by exchanging genetic material between a pair of individuals. There are several ways of performing a crossover operation with the three most common being (Sivanandam & Deepa, 2008; Vikhar, 2017):

1. Single-point crossover: Pairs of individuals are recombined by a random selected point in the chromosome.
2. Two-point crossover: Two positions in the chromosome are selected at random to exchange chromosomic material.
3. Uniform crossover: Each gene in the offspring is created by copying the corresponding gene from one or the other parent chosen according to a random generated mask, where 1 corresponds to the first parent and 0 to the second. Offspring are a mixture of genes from each parent with a number of crossing points that will average L/2 (where L is the length of the chromosome) (Figure 13).

The final step is the search termination step, or the convergence criteria. Since in most cases a true solution is not reached, there could be multiple stopping conditions to finalize a GA. Some of the stopping conditions are the reach of a maximum number of generations, the elapsed time, a no change in fitness criteria, a stop if there is no improvement in the objective functions after a number of consecutive generations (Stall generations) and no improvement during an interval of time (Stall time limit). Also, we can reach termination of search if the best individual fitness is below the convergence value or when the worst individual of the population has a fitness less than the convergence criteria. The stopping criteria depends on the problem we are tackling and the resources we have available (Carlos A. Coello Coello, 2005; Sivanandam & Deepa, 2008).

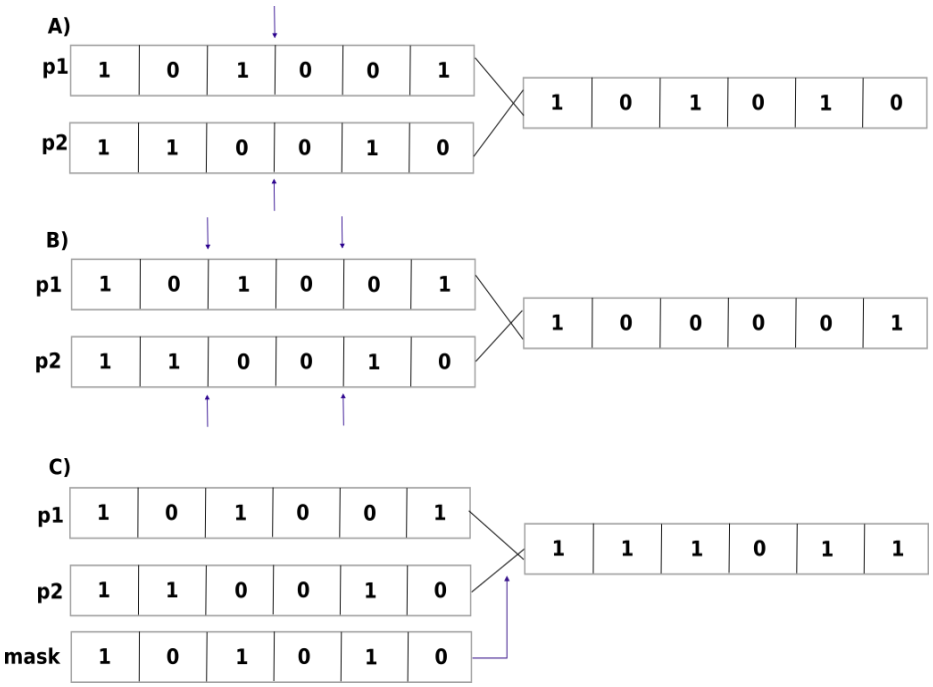


***Figure 14: Types of crossover operations**. A. One point crossover. B. Two-point crossover. C. Uniform crossover*

Genetic Programming on the other hand is better suited for generating formulae and population relationships as the algorithm codes solutions in the form of a graphical (tree) structure whose nodes or edges represent parameters (Figure 12) (John R Koza, 1994, 1990). GP is an automated invention machine, unlike GAs that are passive structures, which routinely delivers high-return human competitive machine intelligence, duplicating the functionality of previously patented inventions, infringed a previously issued patent, or created a patentable new invention (J R Koza et al., 2003). The basic workflow of a GP is similar to that of a GA but instead of modifications of genes in a chromosome, the mutation and recombination processes occur at a node level by modifying a specific node or by exchanging subtrees at random crossover points. In this thesis, we applied a GP algorithm to infer the demographic history of North

African populations because it allowed us to reduce biases at the parameter definition level on each of the tested models.

The main application areas of GP are computer science, science, or engineering. In the last few years, GP has been used to develop prediction models to forecast COVID-19 in India (Salgotra et al., 2020), to design steel-concrete composite floors (Shariati et al., 2019), to infer the energy dissipation variables in cascade spillways (Salmasi et al., 2021) or a tax revenue forecast model in a heterogeneous population (Alexi et al., 2023) just to name a few examples.

1.2.3.1.1 Genetic Programming in demographic inference studies

In canonical GP, the exploration of the space of solutions is mainly accomplished by creating new solutions from the recombination of the most successful ones. The offspring is a combination of both parental trees using a subtree-crossover operator (Sivanandam & Deepa, 2008), leaving modifications of the parental structures – i.e., mutations – as a less frequent event to generate new solutions (Sivanandam & Deepa, 2008). However, in demographic models, where time plays a pivotal role in the definition of the trees, classical recombination approaches could lead to non-compatible solutions where the root node of the replaced subtree from one tree could have older times than the preceding node. Because of this, in this thesis we have applied mutation as the solely driver of change in the GP framework, since the mutation process can be constrained to be in between the ranges of the previous and next demographic events.

1.2.3.2 Advantages and disadvantages of Evolutionary Algorithms

MAs and especially EAs have been widely used to solve many complex problems due to its advantages over classical deterministic techniques. EAs, as they are inspired by natural evolution, are conceptually simple and flexible. They use prior knowledge, which restricts the search space, speeding analysis time. Evolution is a parallel process, so each evaluation in EA performs parallel operations. EAs are robust and made to adapt solutions in changing environments. Finally, EA can solve problems without the need of human expertise (Vikhar, 2017).

Although they are quite advantageous, metaheuristics suffer from some significant disadvantages against deterministic algorithms. The main comparison is that MAs do not guarantee optimal solutions, but rather satisfactory solutions. This can be seen as a trade-off between optimum solution in simple problems and satisfactory solutions in very complex problems. Some MAs also are affected by the "curse of dimensionality problem" that affects their performance as the problem size increases. Finally, there is a lack of mathematical analysis in many MAs. We obtain results but we cannot analyse the degree of trust we have on the result. There is no strong theoretical notion that overcomes this limitation, but we need to consider that metaheuristics as a field is still in its infancy compared to physics or mathematics (Rajwar et al., 2023).

1.3 North Africa, a complex scenario

1.3.1 Human population history of North Africa

1.3.1.1 Pre-historic North Africa

1.3.1.1.1 Palaeolithic

The presence of hominids in North Africa dates back to at least 2.4 million years ago, as indicated by the discovery of stone artifacts and cutmarked bones at sites like Ain Hanech and Ain Boucherit in Algeria, estimated to be approximately 1.9 million and 2.4 million years old, respectively (Sahnouni et al., 2018).

The oldest *Homo* remains with anatomically modern human (AMH) features were found in the Jebel Irhoud cave in eastern Morocco, dating back to 314,000 ± 34,000 years ago (Jean Jacques Hublin et al., 2017; Richter et al., 2010). This discovery emphasizes the significant role of North Africa, or non-eastern Africa, in the early stages of *Homo sapiens*, potentially shedding light on the importance of early population structure in the origin of AMH (Gibbons, 2017; Hollfelder et al., 2021; Ragsdale et al., 2023). Artifact remains extracted from the site are associated with a Levallois-based Middle Stone Age stone tool assemblage (Jean Jacques Hublin et al., 2017) which can be defined as the Mousterian industry (Richter et al., 2017). The Mousterian industry eventually gave way to the Aterian culture, characterized by the presence of pedunculated tools (Klein, 2000). Aterian sites in North Africa have been dated from 150,000 years ago (Richter et al., 2010), coinciding with the beginning of the last Interglacial period, up to 20,000 years ago.

The Later Stone Age is defined by the Iberomaurisian tradition in the coastal Maghreb region, extending from 22,000 years ago to 9,500 years ago and characterized by microlithic backed, partially backed, and obtuse-ended bladelets (Irish, 2000). One of the most significant sites associated with the Iberomaurisian culture is the Taforalt site in eastern Morocco, where ancient DNA, dating back to approximately 15,000 years ago, was extracted from fossils (Van De Loosdrecht et al., 2018).

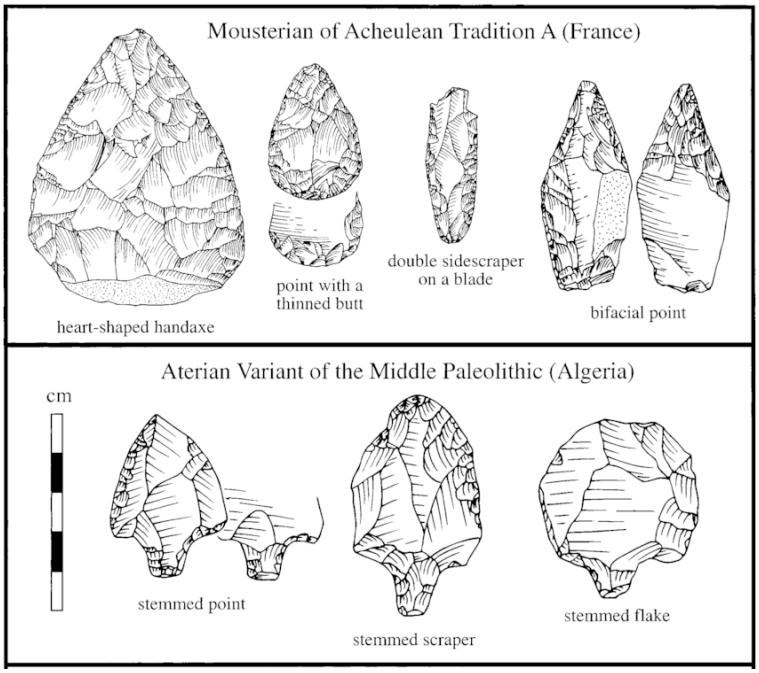


***Figure 15: Mousterian and Aterian lithic cultures**. Aterian lithic is characterized by the presence of pedunculated tools in contrast to the simpler shaped Mousterian industry. Figure modified from Klein 2000*

Around 10,000 years ago, coinciding with a wet period in the Sahara, a new culture emerged in North Africa known as the Capsian (J. L. Newman, 1997). Temporarily overlapping with the Iberomaurisian culture but preferring inland territories, Capsian communities were larger and more sedentary, occupying numerous *rammadiyat* (snail-shell mound) sites

in Algeria and Tunisia (Dachy et al., 2023). Despite their sedentary nature, Capsian communities remained hunter-gatherer societies and can be divided into two phases: the Typical Capsian, characterized by a large number of tools with backed blades, and the Upper Capsian, which featured a reduced number of tools but an emphasis on geometric microlithic and a rich bone industry (Rahmani, 2004). Dental nanometrics on Capsian samples suggest an affinity with West Asian and European traits but with deviations toward Sub-Saharan traits, indicating a complex origin of North African populations with influences from multiple sources (J. J. Hublin et al., 2012; Irish, 2000).

1.3.1.1.2 Neolithic

The origin of the Neolithic period in North Africa remains a subject of debate, with no definitive conclusion reached. While the most widely accepted hypothesis suggests a Middle Eastern origin followed by a demic diffusion into North Africa (Morales et al., 2013; Mulazzani et al., 2016), there is also evidence of Neolithic traits being acculturated by epipaleolithic communities in the region (Linstädter et al., 2012; Mulazzani et al., 2016)

In eastern North Africa, there is controversial evidence of cattle domestication dating as far back as 10,000 years ago at the Bir Kiseiba site in Lower Nubia (Brass, 2013; Wendorf et al., 1985). By the 7th millennium BCE, the Neolithic had spread across North Africa to the est in at least three distinct traditions (Figure15):

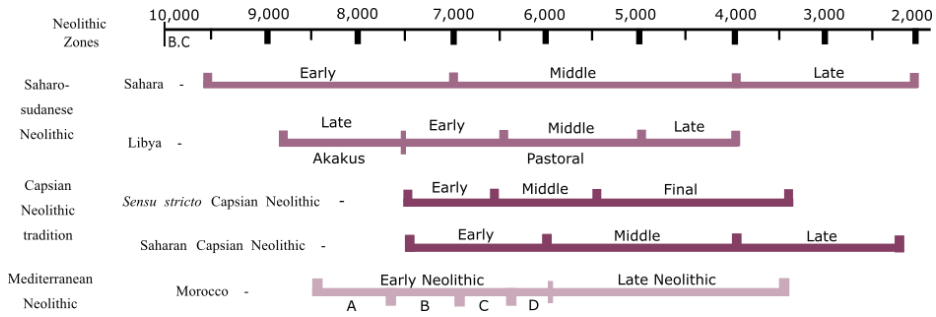


Figure 16: Chronology of the different Neolithic traditions present in North Africa. *The Neolithic in North Africa is divided into three major traditions, Saharan-Sudanese Neolithic, Capsian Neolithic and Mediterranean Neolithic, each with different geographical distributions and different eras. Translated and modified from SiAmmour et al 2022*

1. **The Saharan-Sudanese Neolithic** (8,000 - 3,000 years ago): Primarily found in the Hoggar region of southern Algeria, this tradition is characterized by a nomadic pastoralist lifestyle (Smith, 2001) and lake and river fishing (Si-Ammour, 2022)
2. **The Capsian Neolithic** (7,500 - 3,500 years ago): This tradition was widespread across the lower Sahara up to the Sahara Atlas, from Libya to Western Sahara. It initially began as Upper Capsian groups incorporated Neolithic traits. The early Neolithic Capsian tradition is marked by the rise of pastoralism, with attempts at domesticating sheep, dogs, and goats. In the late Neolithic, earlier Capsian traits were lost, and a fully developed Neolithic economy emerged, featuring agriculture and cattle domestication. This Neolithic group seems to precede the North Africa proto-historic Berber tradition (J. L. Newman, 1997; Si-Ammour, 2022).
3. **The Mediterranean Neolithic:** This tradition is associated with the periodization of the Moroccan Neolithic from around 7,000 BCE, evidenced by the documentation of domesticated plants and animals (Linstädter et al., 2012) and the spread of Cardium-

impressed pottery from the Iberian Peninsula (Si-Ammour, 2022). There is controversy whether this tradition originated in North Africa or was later introduced from the Iberian Peninsula (Perrin et al., 2022).

It is important to note that this chronology is somewhat fragmented, with many of the dates lacking a clear stratigraphic sequence. Radiocarbon dates have been obtained from less reliable samples, so the dates for the Neolithic transition should be interpreted with caution.

1.3.1.1.3 The Importance of the Sahara

The Sahara has played a crucial role as a biological corridor during several humid periods in its history. One of the most extensive of these periods occurred during the last Interglacial period, approximately between 130,000 and 117,000 years ago (Larrasoaña et al., 2013). At that time, the Sahara was characterized by a network of watercourses that flowed towards the Mediterranean. Within this landscape, three major rivers—Irharhar, Sahabi, and Kufrah—were present. Among these, the Irharhar river is considered the most likely route for the dispersion of hominins, as Middle Stone Age (MSA) artifacts have been discovered in its basin (Coulthard et al., 2013).

This network of biological corridors played a pivotal role in separating the industries of the Nile Valley from other North African regions (Osborne et al., 2008) and in the migration of flora and fauna across the Sahara. The presence of similarities in fauna between the northern and southern parts of the Sahara further supports the hypothesis of a pan-African origin of *Homo sapiens* (Drake et al., 2011; Geraads, 2010).

The last humid period in the Sahara took place during the early Holocene, around 11,000 to 8,000 years ago. During this time, various species expanded, and fishing traditions developed in the lakes scattered across the Sahara (Si-Ammour, 2022). This humid period coincided with the Neolithic expansion in North Africa. However, by the end of the Neolithic era, the aridification of the Sahara had already commenced, compelling populations to adopt a more nomadic pastoralist way of life throughout the Sahara, except for the Nile River basin. This shift in lifestyle was driven by the changing environmental conditions of the region.

1.3.1.2 Historic North Africa

By the end of the Neolithic, small proto-Amazigh-speaking (Lybico-Berbers) groups had been formed across North Africa, expanding from nomadic herders to groups with more agricultural economies, either in coastal sites or oasis at the north and center of the Sahara (Camps, 1982, 1998; J. L. Newman, 1997). The interaction of them with the different populations that would occupy the region at different points in time would result in the current demographic picture we observe today in North Africa.

1.3.1.2.1 Historical Mediterranean Contacts

The first reported non-African civilization that contacted the proto-Amazigh groups were the Phoenicians. Phoenicians established a small outpost at Utica, in the Algerian coast about 1,100 years BC (McEvedy, 1995; J. L. Newman, 1997). It seems that they never had colonization as a priority and the different posts were only to serve as trading sites with

other European civilizations. By the 6th century BC, Carthage, once a trading outpost, had grown and became a major economic and political center in the ancient world. Libyco-Berber groups were slaves working on the fertile plains close to Carthage as slave trade was part of the economic fabric of the Phoenicians. Trans-Saharan trades between Phoenicians and some proto-Amazigh groups, like the Garamantes were constant in southwestern parts of Libya. The presence of Phoenicians brought North Africa to the Iron Age and a series of states modelled along Carthaginian lines sprung throughout the Maghreb, introducing the area to the Mediterranean world (Naylor, 2009; J. L. Newman, 1997).

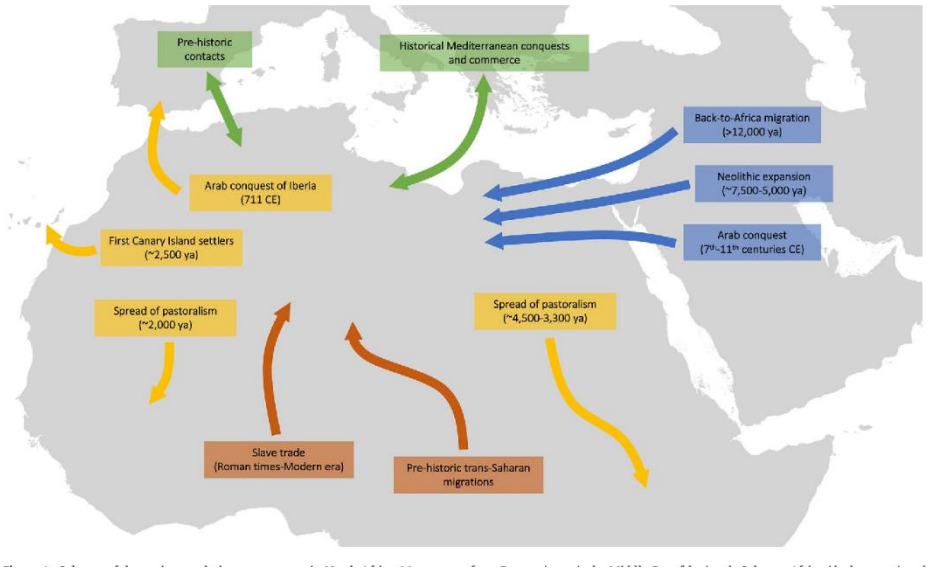


Figure 17: Scheme of the main population movements in North Africa. *Movements from Europe (green), the Middle East (blue), sub-Saharan Africa (dark orange), and North Africa (yellow) are shown. Arrows are approximations and show direction rather than specific migration routes for the major migrations, although additional migrations may have occurred. Figure and caption from Lucas-Sanchez et al 2021.*

At the east of North Africa, Mediterranean groups also started to influence. Ionian (Greeks and Macedonians) presence began in the 7th century BC, when Egypt recruited Ionian mercenaries to assist against the

Assyrian. From there, merchants established in Naukratis and started to do business along the Nile. In the 4th century BC, the Ionian took control of Egypt after banishing the Persians and began the construction of Alexandria. From there, Hellenistic language and culture began spreading, mainly in the Egyptian elite. Outside of Egypt, there were few Ionian settlements to the west, mainly related to agriculture and trade. The expansion to the west was mainly limited by the Phoenician settlements.

By the end of the 1st century BC, both the Phoenician (Carthage 146 BC) and the Ionian (Egypt 30 BC) had fallen at the hands of the Roman Empire (J. L. Newman, 1997). Although Rome had control all over North Africa their presence was not homogenous. In Egypt, Roman presence was largely administrative, mainly concerned in grain extraction and tax collection. Very few officials were Roman and Greek remained the official language. On the other hand, the presence of the Roman Empire in the Maghreb was much larger. Large number of colonists arrived to west North Africa and a great number of infrastructures were built. Legions were deployed to protect the states from Libyco-Berber attacks, and "Italians" were sent to Africa as consequence of land shortage on the peninsula (J. L. Newman, 1997).

Some Amazigh lured by Roman culture also went to live in the Roman cities at the coast while others remained in the hinterland forming different Berber Kingdoms. The Berber and Roman relationships fluctuated a lot, from trading partners, mainly exchanging wool, leather, and sub-Saharan slaves to not so cordial interactions. Some Amazigh groups revolted at different times weaking Roman authority in the region until the invasion of the Vandals (Germanic invaders) in the 420 CE. The Vandals did not take much of the Roman territory, allowing for the occupation by Imazighen Kingdoms of the North African Roman

provinces. This situation was maintained for over a hundred years until in 553 CE, the Byzantine Roman Empire took back control of North Africa from the Vandals (Arauna-Rubio, 2017; Naylor, 2009; J. L. Newman, 1997).

1.3.1.2.2 Arab Conquest

In the 610 CE, Muhammad founded Islam in the city of Mecca in current Saudi Arabia. By 630, Muslims dominated Mecca and started an expansion from the Arabian Peninsula. North African expansion started right after the conquest of nowadays Palestine and Syria, reaching Byzantine Egypt in 639 CE. From there the Arabs conquered Cyrenaica (West Libya) and Tripolitania (Northeast Libya) in the 642 & 643 CE respectively (Arauna-Rubio, 2017; J. L. Newman, 1997). They continued their advance to the west until they had the last Amazigh controlled city fall in 698 CE. In 711 CE Tariq ibn Ziyad crossed the Mediterranean and entered Iberia culminating the expansion with the conquest of Al-Andalus[4] (Naylor, 2009).

Two different but linked processes occurred in North Africa: Islamization and Arabization. The first implies the assimilation of the religious beliefs by the conquered population, which occurred rather quickly, with most of the population adopting Islam a couple centuries after the first arrival of Muslims. Arabization, in contrast, concerns the acculturation of language, culture and identity and is a separated process that not necessarily was tied together with the religious conversion, although non-Arab individuals

[4] Etymologically, al-Andalus is hypothesized to have originated from the Tamazight term “tamort u-andalus” which translates to “land of the vandals” referring to the Germanic groups that inhabited the south of the Iberian peninsula until the 5th century BC (Anders, 2023)

were discriminated against even if they converted to Islam (2). Even nowadays it is still an ongoing process in several parts of North Africa (Coffman, 1992)

Although originally the Arab expansion was mainly a cultural movement, with only small groups from the elites arriving to North African cities from the Arabian Peninsula, this would change in the 10th and 11th centuries. Driven by an increased population pressure in Arabia, two large Bedouin groups, the Banu Hilal and the Banu Sulaym, migrated first to Egypt and then expanded to the Maghreb, occupying the oases that were until now ruled by Imazighen. In the next centuries the number of Arab people highly increased both in coast and desert regions as well as the pressure to embrace Islam and Arab culture forcing the Imazighen to recede to the remote regions (Lucas-Sánchez, 2023; Naylor, 2009; J. L. Newman, 1997).

1.3.1.2.3 Ottoman Empire & Colonialism

Arab rule in North Africa was uninterrupted for almost a thousand years, until the Ottoman Empire conquered Egypt in 1517 expanding as far as Algeria with Cairo and Algiers being the most prominent cities under Ottoman rule. From that point onwards, a series of revolts followed by a new ruler arisen from time to time. The first being the Mamluk overthrowing the Ottoman in the 18th century followed the French seizure of Algeria in 1830 which later expanded to Tunisia and Morocco. Egypt was under British intervention, as Italy took over Ottoman Tripolitania and Spain claimed territories in Morocco and the Western Sahara. At the start of the 20th century these regions under European control started to declare their independence. By 1962, with the exception

of Spanish western Sahara, North Africa was decolonized but with most countries suffering from the consequences of, in most cases, incomplete (Gritzner & Gritzner, 2006; Naylor, 2009).

1.3.1.3 The Amazigh

Historical context shows North Africa as a very tumultuous region, with extensive migrations from different civilizations and times that have influenced the autochthonous populations since the Neolithic. This autochthonous group, from an anthropological point of view, seems to be formed by a heterogeneous set of groups, grouped together by a linguistic similarity. The Tamazight is a language, or group of languages of the Afroasiatic branch characterized by the use of a particular consonantal alphabet, the tifinagh. Genetic and cultural evidence consider the Amazigh as the direct descendants of the Epipaleolithic populations in North Africa (Van De Loosdrecht et al., 2018).

Figure 18: Tifinagh alphabet and its corresponding Latin-Amazigh alphabet characters*. The thirty-three characters that compose the tifinagh alphabet as stipulated by the Royal Institute of Amazigh Culture in Morocco (IRCAM 2023).*

Amazigh relationships with foreigners have been very diverse but mostly following a relationship of commerce and rebellions (J. L. Newman, 1997). The deepest cultural transformation in Imazighen in historic times happened with the Arab expansion. Huge cultural conversion processes led to large Amazigh empires like the Almohades and extensive migrations from Middle East like the Bedouin migrations of the 10th and 11th century were big turning points in the Arabization of the Amazigh population and had an impact that can be observed in the current genetic landscape of North African (Arauna-Rubio, 2017; Lucas-Sánchez et al., 2021; Serra-Vidal et al., 2019). On top of that, Imazighen were slave traders since their first contacts with populations of the Mediterranean. This control over slave migratory routes from the south of the Sahara could have also had an impact on the genetics of the Imazighen in particular, and on the North African in general (Lucas-Sánchez, 2023; Lucas-Sánchez et al., 2021).

1.3.2 Genetics of North Africa

North Africa has historically been neglected from genetic studies even though having a principal role in human history (Lucas-Sánchez et al., 2021). Sampling has been very limited, with North African groups having residual participation in public genetic databases in contrast to other populations at the south of the Sahara (Yoruba individuals in the 1000 genomes projects, n = 208; North African individuals in the 1000 genomes project, n = 30) (Bergström et al., 2020; Mallick et al., 2016; The 1000 Genomes Project Consortium, 2015) .

1.3.2.1 Current day North-African genomics

Initial studies with classical markers identified a complex pattern of genetic diversity, with extensive admixture and a clear differentiation of North Africa from the rest of the African continent (Cavalli-Sforza & Piazza, 1993). Genomic studies centered on North African populations (Arauna-Rubio et al., 2017; Fadhlaoui-Zid et al., 2013; Henn et al., 2012; Lucas-Sánchez et al., 2021, 2023; Serra-Vidal et al., 2019) point to the presence of at least four major admixture sources, resulting in a mosaic of variation at different proportions of (i) an autochthonous genetic component, (ii) a Middle Eastern-like component, (iii) a European-like component and (iv) a component that can be attributed to a mix of populations at the south of the Sahara (Henn et al., 2012; Lucas-Sánchez et al., 2023). This Middle East-like ancestry appears to be stronger in the eastern regions of North Africa, highly influenced by pre-historic and historical migrations from Arabia and the Levant, and declines moving westward in a opposite direction to the autochthonous Maghrebi component that is maximized in isolated populations of Tunisia (Arauna-Rubio et al., 2017; Serra-Vidal et al., 2019). Some authors have related this

Maghrebi component to a back-to-Africa migration more than 12.000 years ago by a pre-Holocene population splitting from the rest of out-of-Africa groups (Fregel et al., 2018; Henn et al., 2012), although is also possible that this component can be attributed to a group that remained isolated in North Africa from over ~300,000 years indicating a population continuity since the MSA due to similar lithic industry found next to the AMH of 300,000 years and an Aterian sites dated as young as 20,000 years ago.

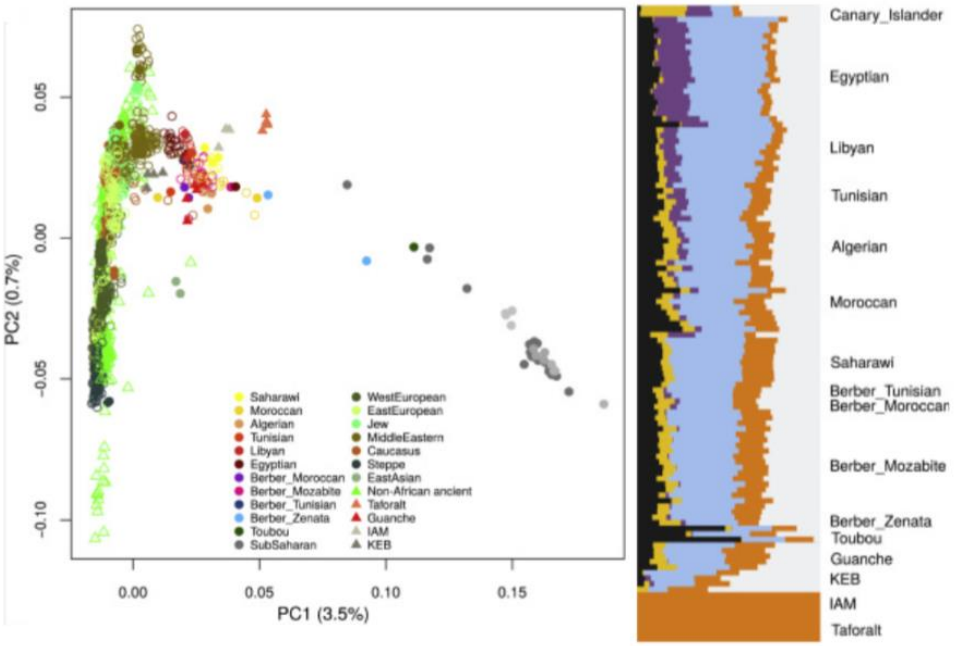


Figure 19: Genetic structure of North African populations. *Left: First and second principal components of North African samples together with a worldwide panel. Right: ADMIXTURE analysis (K=6) of North African samples. Autochthonous components in orange, Middle Eastern-like component in blue, European Neolithic-like component in grey, western European hunter-gatherer-like component in yellow, Steppe-ancestry in purple and south Saharanan-like component in black. Modified from Serra-Vidal et al 2019*

Studies on uniparental markers reveal evidence on the uniqueness of North Africa within the continent. The presence of uniparental lineages (haplogroups) that originated in sub-Saharan Africa, the Middle East, or Europe, suggests a complex pattern of gene flow towards North Africa (Brakez et al., 2001; Ennafaa et al., 2009; Fadhlaoui-Zid et al., 2004; Font-Porterias et al., 2018; González et al., 2007; Plaza et al., 2003; Turchi et al., 2009). However, autochthonous lineages have also been described. Mitochondrial DNA analyses identify two main lineages in North Africa, U6 and M1 whose origin has been traced back to Upper Palaeolithic times (Pennarun et al., 2012; Secher et al., 2014; Van De Loosdrecht et al., 2018), with presence of this lineages later in the Epipaleolithic and Early Neolithic suggesting a back-to-Africa migration from Southwestern Asia (Fregel et al., 2018; Hervella et al., 2016; Olivieri et al., 2006; Van De Loosdrecht et al., 2018). A second hypothesis supports that M1 and U6 haplogroups could be related to the spread of Iberomaurisian culture (Pennarun et al., 2012) as oldest mtDNA samples present these lineages (Kefi et al., 2018; Van De Loosdrecht et al., 2018), making a North Africa origin of the haplogroups viable. Regarding Y-lineages, two autochthonous haplogroups, E1b1b1a-M78 and E1b1b1b-M81 (Bosch et al., 2001; Solé-Morata et al., 2017), have been detected at the highest frequency in the region. Both present high levels of genetic diversity but with opposite clinal geographical distributions, with M78 being maximized in Eastern North Africa (Egypt) and M81 having the highest frequency towards the Maghreb (Fadhlaoui-Zid et al., 2013; Fregel et al., 2009; Solé-Morata et al., 2017; Van De Loosdrecht et al., 2018). An in-situ origin of both haplogroups seems to be the most plausible scenario with origin dates ranging from Epipaleolithic to historic times (Cruciani et al., 2006; Fadhlaoui-Zid et al., 2004, 2013; Solé-Morata et al., 2017).

1.3.2.2 Ancient North Africa genomics

The introduction of ancient DNA studies in North Africa revealed a genetic continuity that dates backs, at least, to the Late Stone Age, as genomic data extracted from 15,000 years old individuals from the Iberomaurisian site of Taforalt, in Morocco, share genomic segments with present-day North Africans (Serra-Vidal et al., 2019; Simões et al., 2023; Van De Loosdrecht et al., 2018). The Taforalt site in Morocco (dated between 15,100–13,900 calibrated years before present) is the oldest site to date to yield DNA data, not only in North Africa but in Africa as a whole (Lucas-Sánchez et al., 2021). Taforalt genomes present similarities with early Holocene Near Eastern Natufians (Levantine Natufians) suggesting pre-Neolithic gene flow between North Africa and the Middle East. They also present influence from populations south of the Sahara in higher proportion than current day North African groups although no good proxy of the ancestral sub-Saharan component is currently available. Following the Taforalt samples, an individual from the Epipaleolithic site of Ifri Ouberrid (Morocco), dated around 7,600 years ago showcased similar genetic structure as the Taforalt sample (Simões et al., 2023), confirming a long-lasting continuity for the last 15 thousand years. In addition to these ancient North African Epipaleolithic genomes, nine individuals from Early Neolithic, five from the Ifri n'Amr or Moussa (IAM) (Fregel et al., 2018) site and four from the coastal Moroccan site of Kaf Taht el-Ghar site show genetic similarities with the Taforalt samples with the latter presenting a maximized Early Neolithic European component suggesting an early arrival of European Neolithic groups to the coasts of the Maghreb possibly introducing Neolithic concepts to local communities (Linstädter, 2013; Simões et al., 2023). Middle and Late Neolithic samples from the Moroccan Atlantic sites of Skhirat-Rouazi (3)

(~4,000 BCE) and Kelif el Boroud (Fregel et al., 2018) (KEB) (~3,000 BCE) have been analyzed showing different proportions of Levant Neolithic and European Neolithic admixture suggesting independent Neolithic expansion processes in North Africa from Middle Eastern Levant and Europe crossing the Gibraltar Strait (Fregel et al., 2018; Simões et al., 2023). Historic ancient North African studies have been very scarce. The analysis of 3, 1st millennium BCE human remains from the archaeological site of Abusir-el Meleq in Egypt suggest a closer relationship of ancient Egyptian with Near Easterns than current day Egyptian population, with a higher influence of populations at the south of the Sahara (Schuenemann et al., 2017). Finally, a recent study on Iron Age samples from current-day Tunisia demonstrate contacts and gene flow between the different shores of the Mediterranean Sea (Moots et al., 2023).

Very little has been studied about the interaction between Neanderthals and North African groups. The difficulties in separating the autochthonous North African component to the Middle East-like and the European-like ancestral components in North African samples difficult the analysis of any specific interaction between Neanderthal and ancient North Africans. Even though, some analysis has shown signals of Neanderthal introgression towards North Africans with it not being due to recent Near Eastern or European migrations (Sánchez-Quinto et al., 2012).

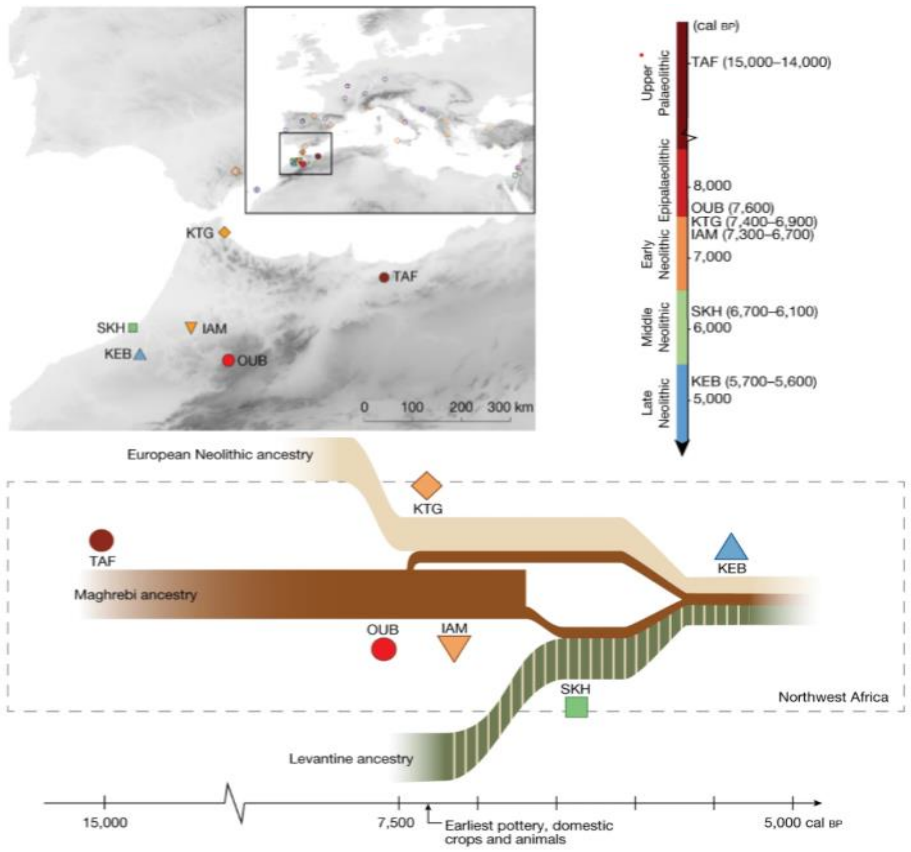


Figure 20: Geographic location and inferred population history of ancient western Maghreb samples*. Sites abbreviations stand for: Taforalt (TAF), Ifri Ouberrid (OUB), Kaf Taht el-Ghar (KTG), Ifri n'Amr or Moussa (IAM), Skhiriat-Rouazi (SKH), and Kelif el Bouroud (KEB). Modified from Simões et al 2023*

1.3.2.3 A source of gene flow

North Africa has not only been a sink but also an important source of gene flow to its surrounding regions (Lucas-Sánchez et al., 2021). Prehistoric contacts with Europe have been attested by the presence of North African uniparental markers (Brakez et al., 2001) and autosomal

segments (Arauna et al., 2019) in Iberian individuals, as well as the presence of archaeological evidence (Linstädter et al., 2012; Perrin et al., 2022). The Arab expansion greatly shaped the genetic landscape of North Africa but also heavily influenced southern Europe, especially Iberia where the Arab occupation (mainly Amazigh people) lasted more than 700 years contributing to the genetic pool of current day Iberians (4).

Ancient Guanche samples from the Canary Island suggest North African as the first settlers of the archipelago (Arauna et al., 2019; Botigué et al., 2013; Fregel et al., 2009; Maca-Meyer et al., 2003, 2004; Rodríguez-Varela et al., 2017). Genome-wide data revealed that the geographical location of the source of admixture differs between the Canary Islands and southern Europe, of the Atlantic coast for the former, and the Mediterranean front for the latter (Arauna et al., 2019). North African influence has also been detected at the south of the Sahara mainly related to pastoralism and the presence of lactase persistence alleles (Tishkoff et al., 2007; Vicente et al., 2019). This gene flow towards the south has been linked to the absence of a particular Iran Neolithic genetic component in Fulani individuals suggesting that the Eurasian-like component present in those populations had to arrive before the expansion of Neolithic into North Africa, pointing to migrations of autochthonous North African individuals south during before 8000 years BP corresponding with the last humid period of the Sahara (Castañeda et al., 2009; D'Atanasio et al., 2023).

2 OBJECTIVES

As stated in the introduction North Africa is characterized by high genetic heterogeneity and the presence of an amalgam of genetic components, because of extensive gene flow from different areas and different time frames. This complex scenario has hampered the establishment of a demographic model for North African groups. The main goal of this PhD thesis is to overcome this challenge and reconstruct a demographic model of the Amazigh and the Arab population that could be used as a neutral demographic model in future studies. To reach this objective, whole-genome sequences from different North African groups has been analyzed using already established demographic inference methods such as Approximate Bayesian Computation coupled with Deep Learning (Mondal et al., 2019) as well as newly developed methods that utilize the power of metaheuristics - Genetic Programming for Population Genetics (GP4PG) - to overcome possible biases in the identification of the demographic events and associated parameters that explain the genetic variation observed in North Africa, as well as limitations in the reproducibility of the models. The specific objectives pursued are as follow:

1) Revise the population demography data and the proposed migration hypothesis in North Africa (Results section 3.1)

2) Define if the North African groups, generally defined by the linguistically and cultural difference of Amazigh and Arabs, present a different genetic origin. *(Results section 3.2)*

3) Obtain the split time between each of the North African groups and their closest group. *(Results section 3.2)*

4) Describe specific demographic history parameters such as effective population size, migrations, and admixture events, mainly focusing on the impact of Middle Eastern migrations at Neolithic and historic (Arabization) times. *(Results section 3.2)*

5) Identify the possible effect of genetic substructure in the demographic history of a population. *(Results section 3.2)*

3 RESULTS

3.1 Population history of North African based on modern and ancient genomes

Marcel Lucas-Sánchez, Jose M. Serradell, David Comas.

Human Molecular Genetics. 2021.

doi: 10.1093/hmg/ddaa261

Lucas-Sánchez M, Serradell JM, and Comas D. Population history of North Africa based on modern and ancient genomes. Hum Mol Genet. 2021 Mar 1; 30(R1): R17-R23.

This section contains a review paper written during the PhD thesis aiming to compile and summarize the state of the art in population genetics of North Africa to the date of publishing. My contribution was focused on writing the sections "The North African Genetic Component", "Ancient Genomes in North Africa" and "Population Replacement versus Demographic Continuity Hypothesis" as well as the refining and production of the final version of the whole manuscript.

OXFORD

Human Molecular Genetics, 2021, Vol. 30, No. 2 R17–R23

doi: 10.1093/hmg/ddaa261
Advance Access Publication Date: 8 December 2020
Invited Review Article

INVITED REVIEW ARTICLE

Population history of North Africa based on modern and ancient genomes

Marcel Lucas-Sánchez[‡], Jose M. Serradell[‡] and David Comas[*,†]

Departament de Ciències Experimentals i de la Salut, Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, 08003 Barcelona, Spain

*To whom correspondence should be addressed at: Doctor Aiguader 88, 08003 Barcelona, Spain. Tel: +34 93 3160843; Fax: +34 93 3160901; Email: david.comas@upf.edu

Abstract

Compared with the rest of the African continent, North Africa has provided limited genomic data. Nonetheless, the genetic data available show a complex demographic scenario characterized by extensive admixture and drift. Despite the continuous gene flow from the Middle East, Europe and sub-Saharan Africa, an autochthonous genetic component that dates back to pre-Holocene times is still present in North African groups. The comparison of ancient and modern genomes has evidenced a genetic continuity in the region since Epipaleolithic times. Later population movements, especially the gene flow from the Middle East associated with the Neolithic, have diluted the genetic autochthonous component, creating an east to west gradient. Recent historical movements, such as the Arabization, have also contributed to the genetic landscape observed currently in North Africa and have culturally transformed the region. Genome analyses have not shown evidence of a clear correlation between cultural and genetic diversity in North Africa, as there is no genetic pattern of differentiation between Tamazight (i.e. Berber) and Arab speakers as a whole. Besides the gene flow received from neighboring areas, the analysis of North African genomes has shown that the region has also acted as a source of gene flow since ancient times. As a result of the genetic uniqueness of North African groups and the lack of available data, there is an urgent need for the study of genetic variation in the region and its implications in health and disease.

Introduction

The genetic study of North African human groups has been generally neglected. Instead the focus of population genetic analyses has been placed on neighboring areas, thereby overshadowing the relevance of North Africa. On one hand, the African continent has captured most of the attention for being the cradle of humankind, however the focus of genetic studies have been mainly set on Eastern and South Africa as the suggested geographical origins of our species (1–4). The expansion of Bantu-speaking groups from western Africa, associated with one of the major population movements also received much attention (5–7). Therefore, North Africa has been neglected in genetic studies compared with the rest of the continent. Moreover, North Africa has been considered as an extension of the Middle East into the African continent, and therefore received little recognition as a unique entity until recently (8,9). Thus, a lower amount of genetic data in North Africa has been collected as compared with other regions. Even recent global genome databases, such as the Human Genome Diversity Project (10) and the Simons Genome Diversity Project (11), only considered a single population (the Mozabite) and four individual genomes (two from the Mozabite and two from the Saharawi), respectively. Fortunately,

[†]David Comas, http://orcid.org/0000-0002-5075-0956
[‡]Both authors have contributed equally to this work.
Received: June 17, 2020. **Revised:** November 30, 2020. **Accepted:** December 2, 2020

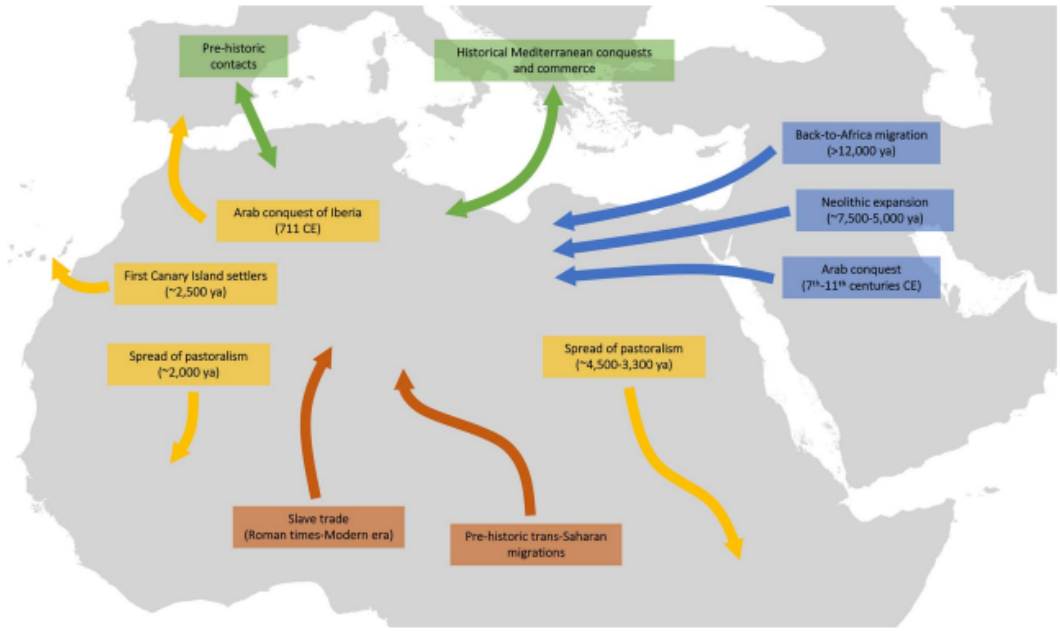


Figure 1. Scheme of the main population movements in North Africa. Movements from Europe (green), the Middle East (blue), sub-Saharan Africa (dark orange) and North Africa (yellow) are shown. Arrows are approximations and show direction rather than specific migration routes for the major migrations, although additional migrations may have occurred.

in the last few years, genetic data, including ancient and current-day whole genomes, have been analyzed in order to refine the population history of North Africans.

North African Data: From Classical Markers to Whole Genomes

Despite the limited population data in North Africa, most analyses have shown a complex pattern of genetic diversity, characterized by extensive admixture, and differentiation of the North African area from the rest of the African continent. The study of classical genetic markers, compiled in the seminal work by Cavalli-Sforza *et al.* (12), evidenced the differentiation of North Africa from the rest of the continent. This is shown in the first component of the African principal component analysis, which suggested a North African demographic history more related to the out-of-Africa (OOA) populations. In a specific North African compilation of classical markers, Bosch *et al.* (13) also showed the distinction of North Africa in comparison to other African groups, and pointed to a gradient of genetic diversity in an east–west axis, as a result of human movements limited by the Mediterranean Sea and the Sahara Desert. The uniparental marker analyses (mitochondrial DNA and Y chromosome) have also evidenced the uniqueness of North Africa within the continent and the admixture of lineages from neighboring areas. The presence of uniparental lineages that originated in sub-Saharan Africa, the Middle East or Europe, suggests a complex pattern of gene flow toward North Africa; however, autochthonous lineages have also been described in the region, pointing to extensive admixture of local and external groups with different gradients of lineages in the area (14–18). During the last decade, the analyses of genome-wide SNPs refined our knowledge about the North African genetic landscape (8,19,20) and reinforced the idea of complex demographic patterns of admixture and isolation in the region that differentiated it from the rest of the African continent. This idea has been also corroborated by the analysis of a still limited data on complete genomes from ancient and modern samples (21–23).

The North African Genetic Component

Genetic data from current-day populations suggest a complex pattern of admixture, with a minimum of four main sources of genetic ancestry for North African people. Henn *et al.* (8) first showed the presence of an autochthonous ancestral component (also known as the Maghrebi component), as well as European, Middle Eastern and sub-Saharan components, in the current North African populations (8). This result showed that North African populations exhibit their own ancestral component and cannot be considered as the result of a mere admixture of exogenous ancestries from neighboring regions. This component is related to an early North Africa population that diverged from the rest of the OOA groups predating the Holocene, more than 12 000 years ago. The component was possibly introduced in a back-to-Africa movement, already suggested by the mitochondrial DNA (mtDNA) haplogroups U6 and M1 (24–26), and more recently nuclear data (27); it is distributed in a west-to-east declining gradient across the region (8). Later studies confirmed the presence of this autochthonous component by comparing current-day genomes with data from ancient anatomically modern human samples recovered from different locations in North Africa (23). This analysis refined the sources of

ancestry in current North Africa populations, adding a Caucasian hunter–gatherer/Neolithic Irani-related component and locating the possible origin of the autochthonous North African component in Epipaleolithic or Early Neolithic times, given that it is prevalent in Moroccan Epipaleolithic and Early Neolithic samples.

Ancient Genomes in North Africa

The retrieval of stone artifacts and cutmarked bones from an archeological site in Algeria places the first peopling of North Africa around 2.4 million years ago (28), whereas direct bone dating of the oldest human remains from the Moroccan site of Jebel Irhoud points to 300 thousand years ago (ka) (9). Many more fossils have been recovered in North Africa (29) but only for a few of them it has been possible to extract and analyze their genome. The Taforalt site in Morocco (dated between 15 100 and 13 900 calibrated years before present) is the oldest site to date to yield DNA data, not only in North Africa but in Africa as a whole. The analyzed Taforalt individuals show high affinity toward Near Eastern populations, especially Epipaleolithic Natufians, with whom they share 63.5% of their ancestry on average. These individuals present mtDNA haplogroups U6 and M1, concordant with the pre-Holocene back-to-Africa event (22,26,30). An ancient sub-Saharan ancestral component is also present, showing a higher affinity with Taforalt than with any combination of Yoruba–Natufian ancestry. Also, no gene flow from Paleolithic Europeans is observed (22).

In addition to these ancient North African Epipaleolithic genomes, five individuals from the Early Neolithic Ifri n'Amr or Moussa (IAM) site were analyzed together with four Late Neolithic samples from Kelif el Boroud (KEB) (21). IAM individuals (7000 years old) showed close genome-wide affinities with the Tarofalt individuals. This was also supported by the presence of similar mtDNA haplogroups (U6, M1) associated with the back-to-Africa migration (30), suggesting a continuity between Later Stone Age and Early Neolithic populations in the Maghreb (21). On the other hand, the genome analysis of the KEB population suggests that it can be modeled as a mixture of IAM and Anatolian/European Neolithic, and it also presents a lower sub-Saharan component than IAM or Tarofalt (21). Mitochondrial and Y-chromosome haplogroups in these samples are prominently found in Anatolian and European Neolithic samples (31,32). Recently, two 7000 year-old mtDNA samples have been extracted from the Takarkori Rockshelter site (Libya) and attributed to a newly identified haplotype in Africa in the basal branch of haplogroup N (33). This haplotype could have arrived with a back-to-Africa event in the spread of pastoralism from the Levant, or it could have differentiated from the L3 haplogroup inside Africa and later spread out of the continent. The Sahara aridification could have caused the isolation and survival of the haplotype in Takarkori while being replaced in other parts of Africa (33).

Regarding Egypt, Schuenemann *et al.* (34) analyzed 151 individuals from the Abusir el-Meleq settlement, carbon-dated from 1388 BCE to 395 CE. Ninety mtDNA and three genome-wide SNP data samples show highly similar haplogroup profiles with low genetic distances between all samples, supporting the idea of genetic continuity in the region. The absence of sub-Saharan African mtDNA haplogroups in ancient samples as compared with modern Egyptians (with 20% sub-Saharan mtDNA) may be explained by recent sub-Saharan gene flow. Nuclear DNA data from these ancient Egyptians further supports these results and reveals a larger Neolithic Near Eastern component than in modern Egyptians, in agreement with the rest of North Africa ancient samples (34).

Outside mainland Africa, other ancient DNA samples have been useful to assess the genetic ancestry of North African populations. The ancient Guanche samples from 7th–11th centuries CE analyzed by parental markers and whole-genome data suggest a North African origin of the Canary Islands settlers, based on the presence of mtDNA U6 and Y-chromosome E1b1b1b1 haplogroups, which are autochthonous to North Africa (32,35), and a significant genetic component shared with Epipaleolithic North Africans in the whole-genome data (23). Autosomal data show a similar admixture profile as Moroccan KEB (21) and are consistent with a single ancestral North Africa origin, but with possible small introgression events after the first settlement of the islands (32). Additionally, European ancient DNA samples from Iberia and Mediterranean Islands confirm a widespread sporadic gene flow from North Africa to the north during the early Bronze Age (36,37).

Population Replacement versus Demographic Continuity Hypotheses

North Africa has been populated since the early stages of humankind (28), but the (still limited) ancient genomic data are only available since the Epipaleolithic population of Taforalt, who might be considered direct descendants of the autochthonous population of North Africa. The continuity of this autochthonous component until the present time has been challenged by constant gene flow into the region from neighboring populations (Fig. 1), which has partially replaced the original population in North Africa at different times (during the Paleolithic and Neolithic ages, and in historical times).

Gene flow in prehistoric times from Middle Eastern Natufians has been observed (21,22). This gene flow coincided with the last humid period of the region (38), which could have eased the connection between both populations (17). The Sahara experienced strong climatic oscillations during the Late Pleistocene and Early Holocene. In what is known as the 'Holocene climatic optimum', warmer and wetter environmental conditions appeared after the Last Glacial Maximum (from 12 to 5 ka), leading to an increase of waterways, flora and fauna that facilitated the spread of human groups across the Sahara (39–42). Later arid periods could have isolated some of these populations in refugia, causing the disappearance of genetic lineages due to genetic drift (33).

Concerning the Neolithic transition, controversy exists with hypotheses defending either cultural diffusion of agriculture (43) or replacement of the indigenous hunter–gatherer populations with Neolithic groups (44,45). The demic diffusion hypothesis has traditionally been accepted as the Neolithization mechanism in North Africa, with Middle Eastern populations suggested as the source of the transition (46,47), although contact with Iberian populations during Late Neolithic has also been observed (21,47,48) (Fig. 1). Nonetheless, recent analysis of contemporary genomes and their comparison with the Taforalt remains, as well as the endemic element shared between Taforalt and early Neolithic Ifri n'Amr or Moussa genomes (21), has shown a continuity of the Paleolithic component in North Africa (23), although this autochthonous Paleolithic component is much lower than the Paleolithic component observed in current Europeans. Therefore, although the impact of the Neolithic was dramatic in North Africa, it did not completely erase the autochthonous component, thereby also suggesting that cultural diffusion had taken place before the demic diffusion.

Eurasian gene flow after Neolithization seems to have had a lower genetic impact, as shown by Serra-Vidal *et al.* (23). The post-Neolithic movements with high genetic impacts on the region are: (i) a sub-Saharan gene flow, which was mainly due to trans-Saharan slave trade routes from the Roman period (1st century BC) through the Arab conquest and lasting until the 19th century (8,20) and (ii) the Arabization, which started in the 7th century and introduced gene flow from the Middle East across all of North Africa, thereby contributing to shape the east to west cline of the Middle Eastern component found in current North Africans (23,49) (Fig. 1). Other historical movements had only minor impacts on the genetic history; these movements include the arrivals of Phoenicians, Romans, Vandals, Byzantines, Ottoman Turks and other Mediterranean European populations (49).

Cultural and Genetic differentiation: Arabs and Imazighen (Berbers)

From a cultural point of view, populations in North Africa have been traditionally divided into Arabs and Berbers. Notably, 'Berbers' is a misnomer that traces back to Greco-Roman times (from the Latin word *barbarus*) for the original inhabitants of the region (50,51), who call themselves Amazigh (sing.)/Imazighen (pl.) (free people) (52). This differentiation has its origin in the Arab conquest of North Africa, when Arab groups occupied the region and imposed a new language, religion and customs (between the 7th and 11th centuries) (53–55). Most North Africans incorporated the new culture, admixed with the newcomers, and began to identify themselves as Arabs (56). But others escaped this influence and receded to the mountains and to remote villages, where they maintained their previous way of life, along with an Amazigh identity and language (Tamazight) (15,16,52). Imazighen are considered the autochthonous inhabitants of North Africa, as historical records account for their existence before the arrival of Phoenicians (814 BCE) (54,57), and an archeological link with the pre-Holocene North African Capsian culture has been suggested (58). As mentioned previously, the comparison between Epipaleolithic and modern North African genomes has reinforced the idea of genetic continuity in the region (23).

Different studies have assessed the cultural differentiation in North Africa from a genetic point of view, targeting several markers like classical markers (13), mtDNA lineages (15), Y-chromosome haplogroups (16,18,59) and more recently, genome-wide data (8,20) and whole genomes (23). These studies revealed a remarkable heterogeneity within North African populations as well as a lack of clear differentiation between the Arab and the Amazigh populations as a whole. Although some Amazigh groups show differences to Arab populations, others share more genetic similarities with certain Arabs than with other people with whom they share a cultural identity. Nonetheless, some Tamazight-speaking populations are outliers and show sharp genetic differences with their neighboring Arab-speaking or even Tamazight-speaking populations. This can be attributed to processes of isolation and genetic drift (15,16,60), as well as to asymmetrical sub-Saharan genetic influence (20). Defining cultural and genetic populations in North Africa is thus challenging. Differential contact with arriving populations, and acceptance of the newcomers' cultures, have led to heterogeneous admixture and local isolation processes, creating a complex mosaic of genetically diverse populations with dissimilar roles of culture in different parts of the region.

North Africa as a Source of Gene Flow

North Africa, the destination of diverse demographic movements throughout history, has not only been a sink but also an important source of gene flow to its surrounding regions (Mediterranean Europe, the Canary Islands and some sub-Saharan populations) (Fig. 1). Historical and archeological evidence exist for North African influence over its neighboring regions, and recent genetic studies using both present-day and ancient samples corroborate this. Due to its proximity and relatively recent historical events, the Iberian Peninsula is one of the main recipients of North African gene flow (14,61) (Fig. 1). The Arab expansion brought mainly Amazigh people to the Peninsula, where they stayed for more than 700 years (62,63). Nonetheless, archeological and anthropological findings account for much older contacts between both shores of the western Mediterranean, pointing to the Neolithic or even the late Paleolithic (64,65) (Fig. 1). The presence of mtDNA (14,66,67) and Y-chromosome sequences (59,68) of North African origin in Iberia, as well as evidence of admixture revealed with genome-wide data (19,61), support this trans-Mediterranean gene flow. Dates inferred with present-day samples place the Iberian admixture pulse in the Arab conquest (19,61), which probably masks older events; however, ancient DNA studies provide genetic evidence for the previously reported prehistoric contacts between both coasts (36,69). Other regions of southern Europe, such as Italy and the south of France, were also destinations of North African gene flow (19,61), although to a lesser degree than the Iberian Peninsula. Dating of the North African components in such regions places the migration events at least 5–7 generations ago (61), but a recent study estimated that the admixture pulse in Italy was much older, suggesting movements from North Africa coinciding with the fall of the Roman Empire around the fourth century (61).

The Canary Islands, consistent with their closeness to the western African coast, also show traces of migrations from the continent (70,71) (Fig. 1), and strong evidence in current and ancient genomes corroborates the North African origin of their first settlers (19,32,35,61,72,73). Genome-wide data revealed that the geographical location of the source of admixture differs between the Canary Islands and southern Europe, of the Atlantic coast for the former, and the Mediterranean front for the latter (61).

Southward gene flow from North Africa into sub-Saharan populations has been related to the spread of pastoralism (Fig. 1). Cattle domestication appeared in North Africa during the Neolithic (74–76), and contacts with southern populations introduced lactase persistence alleles of North African origin, which can be detected in some current-day sub-Saharan populations (4,77–79). Pastoralist migrations from North to East Africa date to around ~4.5–3.3 ka (74,79), whereas contacts between North and West Africa seem younger (around ~2 ka) and are also contemporary with the first traces of pastoralism in western Africa (77).

Concluding Remarks

One main challenge for studying the genetic landscape of North Africa is the insufficiency of available data. No data existed for individual complete genomes until recently, and there is still a lack of whole genomes at the population level. Ancient genome data are also limited despite recent efforts. There is an urgent need for genomic data in the region, not only to unravel the questions related to its population history, but also to understand

the genetic variants and genomic regions involved in health and disease conditions. Given the extensive and bi-directional connections between North Africa and its surrounding regions, studying the genetic and genomic variation as well as disease risk patterns of its populations could also have an impact outside North African borders; in European, Middle Eastern, and sub-Saharan populations.

Acknowledgements

Authors thank Gerard Serra-Vidal and Lara R. Arauna for helpful comments on the manuscript.

Conflict of Interest statement. None declared.

Funding

Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación (PID2019-106485GB-I00/AEI/10.13039/501100011033) and "Unidad de Excelencia María de Maeztu" (AEI – CEX2018-000792-M); and Agència de Gestió d'Ajuts Universitaris i de la Recerca (2017SGR00702).

References

1. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O. *et al.* (2009) The genetic structure and history of Africans and African Americans. *Science (80-)*, **324**, 1035–1044.
2. Scheinfeldt, L.B., Soi, S. and Tishkoff, S.A. (2010) Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 8931–8938.
3. Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., MacPherson, J.M., Kidd, J.M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbini, A. *et al.* (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 5154–5162.
4. 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.* (2019) African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.*, **20**, 82.
5. Li, S., Schlebusch, C. and Jakobsson, M. (2014) Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proc. R. Soc. B Biol. Sci.*, **281**, 20141448.
6. Quintana-Murci, L., Quach, H., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L., Mouguiama-Daouda, P., Comas, D., Tzur, S. *et al.* (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 1596–1601.
7. Patin, E., Laval, G., Barreiro, L.B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K.K., Kidd, J.R., Der, V., Van, L., Hombert, J.M. *et al.* (2009) Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet.*, **5**, e1000448.
8. Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetit, J. *et al.* (2012) Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.*, **8**, e1002397.
9. Hublin, J.J., Ben-Ncer, A., Bailey, S.E., Freidline, S.E., Neubauer, S., Skinner, M.M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K. *et al.* (2017) New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, **546**, 289–292.
10. Cann, H.M., De Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A. *et al.* (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.
11. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A. *et al.* (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, **538**, 201–206.
12. Cavalli-Sforza, L.L. and Piazza, A. (1993) Human genomic diversity in Europe: a summary of recent research and prospects for the future. *Eur. J. Hum. Genet.*, **1**, 3–18.
13. Bosch, E., Calafell, F., Pérez-Lezaun, A., Comas, D., Mateu, E. and Bertranpetit, J. (1997) Population history of North Africa: evidence from classical genetic markers. *Hum. Biol.*, **69**, 295–311.
14. Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J. and Comas, D. (2003) Joining the pillars of hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann. Hum. Genet.*, **67**, 312–328.
15. Fadhlaoui-Zid, K., Plaza, S., Calafell, F., Amor, M.B., Comas, D. and El Gaaied, A.B. (2004) Mitochondrial DNA heterogeneity in Tunisian Berbers. *Ann. Hum. Genet.*, **68**, 222–233.
16. Fadhlaoui-Zid, K., Khodjet-el-khil, H., Mendizabal, I., Benammar-elgaaied, A. and Comas, D. (2011) Genetic structure of Tunisian ethnic groups revealed by paternal lineages. *Am. J. Phys. Anthropol.* **280**, 271–280.
17. Fadhlaoui-Zid, K., Haber, M., Martínez-Cruz, B., Zalloua, P., Elgaaied, A.B. and Comas, D. (2013) Genome-wide and paternal diversity reveal a recent origin of human populations in North Africa. *PLoS One*, **8**, e80293.
18. Cherni, L., Pereira, L., Goios, A., Loueslati, B.Y., El Khil, H.K., Gomes, I., Gusmão, L., Alves, C., Slama, A., Amorim, A. *et al.* (2005) Y-chromosomal STR haplotypes in three ethnic groups and one cosmopolitan population from Tunisia. *Forensic Sci. Int.*, **152**, 95–99.
19. Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C. *et al.* (2013) Gene flow from North Africa contributes to differential human genetic diversity in southern europe. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 11791–11796.
20. Arauna, L.R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhlaoui-Zid, K., Zalloua, P., Hellenthal, G. and Comas, D. (2017) Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. *Mol. Biol. Evol.*, **34**, 318–329.
21. Fregel, R., Méndez, F.L., Bokbot, Y., Martín-Socas, D., Camalich-Massieu, M.D., Santana, J., Morales, J., Avila-Arcos, M.C., Underhill, P.A., Shapiro, B. *et al.* (2018) Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, 6774–6779.
22. Van De Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E.H., El Hajraoui, M.A. *et al.* (2018) Pleistocene North African genomes link near eastern and sub-saharan African human populations. *Science (80-.)*., **360**, 548–552.

23. Serra-Vidal, G., Lucas-Sanchez, M., Fadhlaoui-Zid, K., Bekada, A., Zalloua, P. and Comas, D. (2019) Heterogeneity in Palaeolithic population continuity and Neolithic expansion in North Africa. *Curr. Biol.*, **29**, 3953–3959.e4.
24. González, A.M., Larruga, J.M., Abu-Amero, K.K., Shi, Y., Pestano, J. and Cabrera, V.M. (2007) Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics*, **8**, 223.
25. Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D.M., Dugoujon, J.M. *et al.* (2006) The mtDNA legacy of the levantine early Upper Palaeolithic in Africa. *Science (80-.).*, **314**, 1767–1770.
26. Hervella, M., Svensson, E.M., Alberdi, A., Günther, T., Izagirre, N., Munters, A.R., Alonso, S., Ioana, M., Ridiche, F., Soficaru, A. *et al.* (2016) The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa. *Sci. Rep.*, **6**.
27. Haber, M., Gauguier, D., Youhanna, S., Patterson, N., Moorjani, P., Botigué, L.R., Platt, D.E., Matisoo-Smith, E., Soria-Hernanz, D.F., Wells, R.S. *et al.* (2013) Genome-wide diversity in the Levant reveals recent structuring by culture. *PLoS Genet.*, **9**, e1003316.
28. Sahnouni, M., Parés, J.M., Duval, M., Cáceres, I., Harichane, Z., Van Der Made, J., Pérez-González, A., Abdessadok, S., Kandi, N., Derradji, A. *et al.* (2018) 1.9-million- and 2.4-million-year-old artifacts and stone tool–cutmarked bones from ain boucherit, Algeria. *Science (80-.).*, **362**, 1297–1301.
29. Scerri, E.M.L. (2017) The north African middle stone age and its place in recent human evolution. *Evol. Anthropol.*, **26**, 119–135.
30. Pennarun, E., Kivisild, T., Metspalu, E., Metspalu, M., Reisberg, T., Moisan, J.P., Behar, D.M., Jones, S.C. and Villems, R. (2012) Divorcing the late upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. *BMC Evol. Biol.*, **12**, 234.
31. Haak, W., Balanovsky, O., Sanchez, J.J., Koshel, S., Zaporozhchenko, V., Adler, C.J., der Sarkissian, C.S.I., Brandt, G., Schwarz, C., Nicklisch, N. *et al.* (2010) Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. *PLoS Biol.*, **8**, e1000536.
32. Rodríguez-Varela, R., Günther, T., Krzewińska, M., Storå, J., Gillingwater, T.H., MacCallum, M., Arsuaga, J.L., Dobney, K., Valdiosera, C., Jakobsson, M. *et al.* (2017) Genomic analyses of pre-European conquest human remains from the Canary Islands reveal close affinity to modern North Africans. *Curr. Biol.*, **27**, 3396–3402.e5.
33. Vai, S., Sarno, S., Lari, M., Luiselli, D., Manzi, G., Gallinaro, M., Mataich, S., Hübner, A., Modi, A., Pilli, E. *et al.* (2019) Ancestral mitochondrial N lineage from the Neolithic 'green' Sahara. *Sci. Rep.*, **9**, 3530.
34. Schuenemann, V.J., Peltzer, A., Welte, B., Van Pelt, W.P., Molak, M., Wang, C.C., Furtwängler, A., Urban, C., Reiter, E., Nieselt, K. *et al.* (2017) Ancient Egyptian mummy genomes suggest an increase of sub-Saharan African ancestry in post-Roman periods. *Nat. Commun.*, **8**, 15694.
35. Maca-Meyer, N., González, A.M., Pestano, J., Flores, C., Larruga, J.M. and Cabrera, V.M. (2003) Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. *BMC Genet.*, **4**, 15.
36. Olalde, I., Mallick, S., Patterson, N., Rohland, N., Villalbamouco, V., Silva, M., Dulias, K., Edwards, C.J., Gandini, F., Pala, M. *et al.* (2019) The genomic history of the Iberian Peninsula over the past 8000 years. *Science (80-.).*, **363**, 1230–1234.
37. Fernandes, D.M., Mittnik, A., Olalde, I., Lazaridis, I., Cheronet, O., Rohland, N., Mallick, S., Bernardos, R., Broomandkhoshbacht, N., Carlsson, J. *et al.* (2020) The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean. *Nat. Ecol. Evol.*, **4**, 334–345.
38. Claussen, M., Dallmeyer, A. and Bader, J. (2017) Theory and Modeling of the African humid period and the Green Sahara. Theory and Modeling of the African humid period and the Green Sahara. *OXFORD Res. Encycl. Clim. Sci.*, **1**, 1–38.
39. Drake, N.A., Blench, R.M., Armitage, S.J., Bristow, C.S. and White, K.H. (2011) Ancient watercourses and biogeography of the Sahara explain the peopling of the desert. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 458–462.
40. Xue, Y. (2018) A history of male migration in and out of the Green Sahara. *Genome Biol.*, **19**, 30.
41. Podgorná, E., Soares, P., Pereira, L. and Černý, V. (2013) The genetic impact of the lake chad basin population in North Africa as documented by mitochondrial diversity and internal variation of the L3e5 haplogroup. *Ann. Hum. Genet.*, **77**, 513–523.
42. D'Atanasio, E., Trombetta, B., Bonito, M., Finocchio, A., Di Vito, G., Seghizzi, M., Romano, R., Russo, G., Paganotti, G.M., Watson, E. *et al.* (2018) The peopling of the last Green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages. *Genome Biol.*, 19, 20.
43. Fort, J. (2012) Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 18669–18673.
44. Ammerman, A.J. and Cavalli-Sforza, L.L. (1984) *The Neolithic Transition and the Genetics of Populations in Europe*. Princeton University Press, Princeton, New Jersey.
45. Sokal, R.R., Oden, N.L. and Wilson, C. (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature*, **351**, 143–145.
46. Mulazzani, S., Belhouchet, L., Salanova, L., Aouadi, N., Dridi, Y., Eddargach, W., Morales, J., Tombret, O., Zazzo, A. and Zoughlami, J. (2016) The emergence of the Neolithic in North Africa: a new model for the Eastern Maghreb. *Quat. Int.*, **410**, 123–143.
47. Pimenta, J., Lopes, A.M., Comas, D., Amorim, A. and Arenas, M. (2017) Evaluating the neolithic expansion at both shores of the mediterranean sea. *Mol. Biol. Evol.*, **34**, 3232–3242.
48. Hernández, C.L., Pita, G., Cavadas, B., López, S., Sánchez-Martínez, L.J., Dugoujon, J.M., Novelletto, A., Cuesta, P., Pereira, L. and Calderón, R. (2020) Human genomic diversity where the Mediterranean joins the Atlantic. *Mol. Biol. Evol.*, **37**, 1041–1055.
49. Elkamel, S., Cherni, L., Alvarez, L., Marques, S.L., Prata, M.J., Boussetta, S., Benammar-Elgaaied, A. and Khodjet-El-Khil, H. (2017) The Orientalisation of North Africa: new hints from the study of autosomal STRs in an Arab population. *Ann. Hum. Biol.*, **44**, 180–190.
50. Pellat, C., Yver, G., Basset, R. and Galand, L. (2020) Berbers. In Bearman, P., Bianquis, Th., Bosworth, C.E., van Donzel, E., Heinrichs. W.P. (eds), *Encyclopaedia of Islam*, Second Edition. http://dx.doi.org/10.1163/1573-3912_islam_COM_0114.
51. Ghaki, M. (2003) Els Berbers. In IEMed-MuPCVa (ed), *Tunísia, terra de cultures*. Tunisia, Land of Cultures, Barcelona, pp. 39–42.
52. Camps, G. (1994) Els Berbers, mite o realitat? In Roque, M.A. (ed), *Les Cultures del Magreb*. Enciclopèdia Catalana, Barcelona, pp. 41–74.
53. McEvedy, C. (1995) *The Penguin Atlas of African History*. Penguin Books, London.

54. Newman, J.L. (1995) *The Peopling of Africa: A Geographic Interpretation*. Yale University Press, New Haven.
55. Hiernaux, J. (1975) *The People of Africa*. Encore Editions, New York.
56. Ibn-Khaldoun, A. (1968) *Histoire des Berberes et des dynasties muslmanes de l'Afrique Septentrionale: Traduction de Le Baronde Slane*. Paul Geuthner, Paris.
57. Camps, G. (1998) *Los bereberes: de la orilla del mediterráneo al límite meridional del Sáhara*, Icaria, Barcelona.
58. Camps, G. (1995) *Les Berbères: mémoire et identité*. Errance, Paris.
59. Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P.A. and Bertranpetit, J. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am. J. Hum. Genet.*, **68**, 1019–1029.
60. Djaït, H. (1994) Les cultures Magrebines a través de la història. In Roque, M.A. (ed), *Les Cultures del Magreb*. Enciclopedia Catalana, Barcelona, pp. 75–95.
61. Arauna, L.R., Hellenthal, G. and Comas, D. (2019) Dissecting human North African gene-flow into its western coastal surroundings. *Proc. R. Soc. B Biol. Sci.*, **286**, 20190471.
62. Bosch-Vilà, J. (1988) *Andalus (Les Berbères en Andalus)*. Encyclopédie berbère, 5, pp. 641–647.
63. Hitti, P.K. (1996) *The Arabs: A Short History*. Gateaway, Washington, DC.
64. Linstädter, J., Medved, I., Solich, M. and Weniger, G.C. (2012) Neolithisation process within the Alboran territory: models and possible African impact. *Quat. Int.*, **274**, 219–232.
65. Manen, C., Marchand, G. and Carvalho, A.F. (2004) Le Néolithique ancien de la péninsule Ibérique: vers une nouvelle évaluation du mirage africain? *XXVIe Congrès Préhistoriuque de France: "Un siècle de construction du discours scientifique en préhistoire"*. Société préhistorique française Avignon, pp. 133–151.
66. Secher, B., Fregel, R., Larruga, J.M., Cabrera, V.M., Endicott, P., Pestano, J.J. and González, A.M. (2014) The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. *BMC Evol. Biol.*, **14**, 109.
67. Casas, M.J., Hagelberg, E., Fregel, R., Larruga, J.M. and González, A.M. (2006) Human mitochondrial DNA diversity in an archaeological site in al-Andalus: genetic impact of migrations from North Africa in Medieval Spain. *Am. J. Phys. Anthropol.*, **131**, 539–551.
68. Adams, S.M., Bosch, E., Balaresque, P.L., Ballereau, S.J., Lee, A.C., Arroyo, E., López-Parra, A.M., Aler, M., Grifo, M.S.G., Brion, M. *et al.* (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *Am. J. Hum. Genet.*, **83**, 725–736.
69. González-Fortes, G., Tassi, F., Trucchi, E., Henneberger, K., Paijmans, J.L.A., Díez-Del-Molino, D., Schroeder, H., Susca, R.R., Barroso-Ruíz, C., Bermudez, F.J. *et al.* (2019) A western route of prehistoric human migration from Africa into the Iberian Peninsula. *Proc. R. Soc. B Biol. Sci.*, **286**, 20182288.
70. Navarro, J. (1997) Arqueología de las Islas Canarias. *Espac. tiempo y forma Ser. I, Prehist. y Arqueol.*, **10**, 447–478.
71. Atoche, P. (2013) Consideraciones en relación con la colonización protohistórica de las Islas Canarias. *Anu. Estud. Atlánticos*, **59**, 521–564.
72. Maca-Meyer, N., Arnay, M., Rando, J.C., Flores, C., González, A.M., Cabrera, V.M. and Larruga, J.M. (2004) Ancient mtDNA analysis and the origin of the Guanches. *Eur. J. Hum. Genet.*, **12**, 155–162.
73. Fregel, R., Gomes, V., Gusmão, L., González, A.M., Cabrera, V.M., Amorim, A. and Larruga, J.M. (2009) Demographic history of Canary Islands male gene-pool: replacement of native lineages by European. *BMC Evol. Biol.*, **9**, 181.
74. Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M. *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, **39**, 31–40.
75. Bower, J. (1991) The pastoral Neolithic of East Africa. *J. World Prehistory*, **5**, 49–82.
76. Camps, G. (1982) Beginnings of pastoralism and cultivation in north-west Africa and the Sahara: origins of the Berbers. In Clark, J.D. (ed), *The Cambridge History of Africa*. Cambridge University Press, Cambridge, pp. 548–623.
77. Vicente, M., Priehodová, E., Diallo, I., Podgorná, E., Poloni, E.S., Černý, V. and Schlebusch, C.M. (2019) Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. *BMC Genom.*, **20**, 915.
78. Hassan, H.Y., Underhill, P.A., Cavalli-Sforza, L.L. and Ibrahim, M.E. (2008) Y-chromosome variation among sudanese: restricted gene flow, concordance with language, geography, and history. *Am. J. Phys. Anthropol.*, **137**, 316–323.
79. Ranciaro, A., Campbell, M.C., Hirbo, J.B., Ko, W.Y., Froment, A., Anagnostou, P., Kotze, M.J., Ibrahim, M., Nyambo, T., Omar, S.A. *et al.* (2014) Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am. J. Hum. Genet.*, **94**, 496–510.

3.2 Modelling the demographic history of human North African genomes points to soft split divergence between populations.

Jose M Serradell, Jose M Lorenzo-Salazar, Carlos Flores, Oscar Lao and David Comas

Under revision.

Serradell JM, Lorenzo-Salazar JM, Flores C, Lao O and Comas D. Modelling the demographic history of human North African genomes points to soft split divergence between populations. Under revision

Modelling the demographic history of human North African genomes points to soft split divergence between populations.

Authors

Jose M Serradell[1], Jose M Lorenzo-Salazar[2], Carlos Flores[2,3,4,5,6], Oscar Lao[1] and David Comas[1*]

1 Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Departament de Medicina i Ciències de la Vida, Carrer del Doctor Aiguader 88, Barcelona 08003, Spain

2 Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Granadilla de Abona s/n, 38600 Santa Cruz de Tenerife, Spain

3 Plataforma Genómica de Alto Rendimiento para el Estudio de la Biodiversidad, Instituto de Productos Naturales y Agrobiología (IPNA), Consejo Superior de Investigaciones Científicas, San Cristóbal de La Laguna, 38206 Santa Cruz de Tenerife, Spain

4 Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Carretera del Rosario 145, 38010 Santa Cruz de Tenerife, Spain

5 CIBER de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Av. de Monforte de Lemos, 3-5, 28029 Madrid, Spain

6 Facultad de Ciencias de la Salud, Universidad Fernando de Pessoa Canarias, Calle de la Juventud s/n, Santa María de Guía, 35450 Las Palmas de Gran Canaria, Spain

*Corresponding author: david.comas@upf.edu

ABSTRACT

Background: North African human populations present a complex demographic scenario. The presence of an autochthonous genetic component and population substructure, plus extensive gene flow from the Middle East, Europe, and sub-Saharan Africa, have shaped the genetic composition of its people through time.

Results: We conducted a comprehensive analysis of 364 genomes to construct detailed demographic models for the North African region, encompassing its two primary ethnic groups, the Arab and Amazigh populations. This was achieved through the utilization of the Approximate Bayesian Computation with Deep Learning (ABC-DL) framework and a novel algorithm called Genetic Programming for Population Genetics (GP4PG). This innovative approach enabled us to effectively model intricate demographic scenarios, utilizing a subset of 16 whole-genomes at >30X coverage. The demographic model suggested by GP4PG exhibited a closer alignment with the observed data compared to the ABC-DL model. Both methods point to a back-to-Africa origin of North African individuals and a close relationship of North African with Eurasian populations. Results support different origins for Amazigh and Arab populations, with Amazigh populations originating back in Epipaleolithic times, as early as 22.3 Kya. GP4PG model supports Arabization as the main source of Middle Eastern ancestry in North Africa. The GP4PG model better explaining the observed data includes population substructure in surrounding populations (sub-Saharan Africa and Middle East) with continuous gene flow after the split between populations (migration decay). In contrast to what we observed in the ABC-DL, the best GP4PG model does not require pulses of admixture from surrounding populations into North Africa pointing to soft splits as drivers of divergence in North Africa.

Conclusions: We have built a demographic model on North Africa that points to a back-to-Africa expansion and a differential origin between Arab and Amazigh populations, emphasizing the complex demographic history at a population level**.**

Keywords: Human population genetics, Whole-genome sequences, North Africa, Demographic history, Genetic programming, Deep-learning

BACKGROUND

The North African region has a complex human demographic history with multiple migration events that have shaped the genetic makeup of its populations. Stone artifacts found in Algeria suggest that the first peopling of North Africa occurred around 2.4 million years ago [1]. However, the oldest human remains found in the region, at the Moroccan site of Jebel Irhoud, date back to 300,000 years ago [2]. Nonetheless, up to now there is no evidence that points to a continuity from these ancient humans to current North African people.

The oldest ancient DNA samples retrieved from North African individuals in the Taforalt site in Morocco date to Epipaleolithic times (15.1-13.9 Kya) [3]. The analysis of these Taforalt individuals shows a high affinity with Near Eastern Natufian populations. The presence of mitochondrial DNA haplogroups U6 and M1 in North Africans are consistent with a back-to-Africa event [3–5]. When compared with current inhabitants, the Taforalt ancestry component is present in all current North African populations following a West to East cline with highest frequencies observed on Amazigh individuals [6], pointing to a genetic continuity in the region at least since Epipaleolithic times.

Multiple migrations from surrounding regions have occurred in North Africa, leaving their genetic imprints on the local populations, which are characterized by an amalgam of ancestry components [7]. Most of these migrations originated in the Middle East, such as the Neolithic expansion associated with the spread of agriculture, which had a dramatic impact on the genetic makeup of North Africans, diluting the autochthonous Palaeolithic component in a similar demic process as in Europe [8]. In addition to the Neolithic migration event from the Middle East, recent data on ancient remains have shown that North Africa has also experienced gene flow from Europe in Neolithic times, as evidenced by the presence of Iberian genetic ancestry in Western North Africans in early Neolithic sites, suggesting a complex and heterogeneous gene flow in the region [9]. Furthermore, post-Neolithization gene flow has been observed in North Africa, resulting from events such as Arabization during the 7th to 11th centuries [10], the trans-Saharan slave trade from Roman times to the 19th century, and contacts with Mediterranean populations [7,11,12]. This complex pattern of migrations, with different temporal and geographical origins, has challenged the demographic reconstruction of North African population history.

Linguistics broadly classify the present populations of North Africa into two major groups: the Imazighen (Amazigh in singular), also known as the misnomer of Berbers, and the North African Arabs. After the Arab arrival to North Africa during the Arabization, most North African autochthonous groups adopted the Arab culture, and mixed with the immigrants [13]. However, a few groups retreated to remote places and maintained their customs, along with the Amazigh identity and language (i.e., Tamazight) [14,15]. Previous studies showcase large genetic heterogeneity within North African populations [6,7,15–18] with no clear differentiation between Arab and Amazigh populations as a

whole. However, some Tamazight-speaking populations are genetic outliers with remarkable differences with their neighbouring Arab-speaking populations. This was attributed to isolation and drift, as well to differences in their demographic histories [6,18].

The genetic heterogeneity of North African populations plus the presence of an amalgam of genetic components, as a result of extensive gene flow coming from different areas and different time frames, have hampered the establishment of a demographic model for North African groups. To overcome this challenge, in this study we aim to reconstruct a demographic model of the Amazigh and Arab populations in North Africa that might be used as a neutral demographic model for future studies. We aim to tackle this issue by addressing three main questions. First, estimate the origin of North African groups and how they evolved. Secondly, address the number and amount of admixture events that have happened in the region. And finally, assess whether Imazighen and North African Arabs share the same demographic history. Overall, all these questions relate to the development of a demographic model that reproduces the rich and complex history of the region.

To answer these questions, we applied two different computational approaches. We first capitalized on the whole genome data of North Africa and used the Approximate Bayesian Computation framework coupled with Deep Learning (ABC-DL) [19]. Upon recognizing that the model identified by ABC-DL exhibited limited reproducibility with the observed data, we embarked on the development of a novel approach, Genetic Programming for Population Genetics (GP4PG), rooted in metaheuristics. GP4PG uses natural computing algorithms to infer the most optimum set of demographic events and associated parameters to explain the genetic variation observed in a given dataset.

RESULTS

Genetic Structure Analysis of North African samples.

The Principal Component Analysis (PCA) performed on the whole genome dataset (see Methods) of North African individuals (N=30) and reference samples from Africa and Eurasia (N = 364), comprising 1.58 M SNPs, explains 8.64% of total variance on the first two principal components and recapitulates a similar population structure as previously described [6,7,12], with North African individuals clustered together in-between sub-Saharan African and Eurasian populations (Fig.1_a). A second PCA, focused on North African samples, shows in more detail the genetic structure of North Africa. East North Africa individuals cluster closer to Middle Easterns than West North Africans. Nonetheless, we observe a very heterogeneous pattern with the exception of the two North African Imazighen from Chenini (Tunisia) that cluster together isolated from the rest of North Africans (Fig. 1_b) (for full analysis see Additional file 1: Fig S1). The ADMIXTURE analysis (Fig. 1_c) identifies similar patterns to the PCA analysis. The lowest cross validation errors were found in the range between K=3 and K=9 (Additional file 1: Fig S2), with K=3, K=6, K=9 showing the least number of common modes among the different runs in pong [20]. At K=3 we observe the differentiation between sub-Saharan African (red), European (dark blue) and East Asian (light grey) components. All North African samples show similar ancestry patterns exposing an exclusive North Africa component at K = 9 (light blue) except for the Egyptian samples that show higher proportions of a purple component that is maximized in the Middle Eastern samples in K = 9 (Fig. 1_c). The rest of results at different K are in the Supplementary file (Additional file 1: Fig S3).

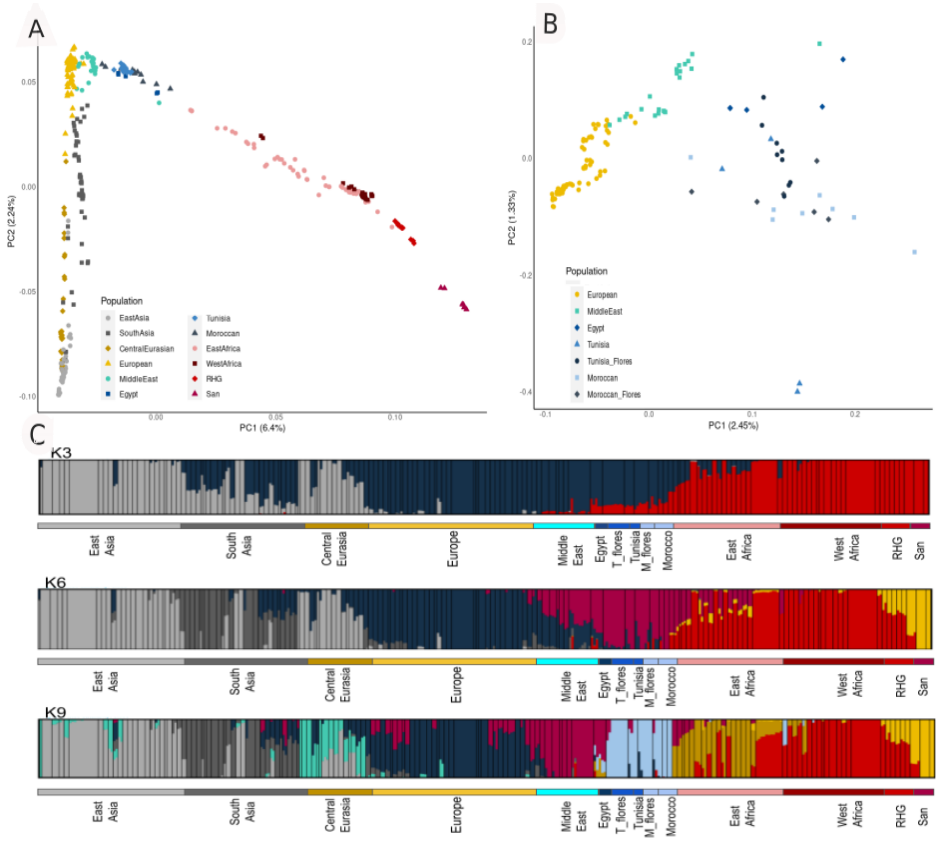


Fig. 1: Genetic structure of North Africans and reference groups. A. PCA plot of all North African samples and the reference dataset. **B**. PCA plot focused on North African population substructure. **C**. Estimated ancestry proportions for the whole dataset in the three Ks most supported by pong [20].

North African demographic model using an ABC-DL approach.

Given these previous results showing the presence of population substructure in North Africa and a complex amalgam of genetic components, we conducted an Approximate Bayesian Computation (ABC) analysis coupled to a Deep Learning (DL) framework [19].

Our aim was to delve into the origins of contemporary North African populations and discern variations in the scenario based on the cultural backgrounds of these populations.

On one side, our dataset included North Africa Amazigh populations, considered the descendants of the autochthonous inhabitants of the region. On the other hand, we have also analysed North African Arab populations resulting from multiple cultural and migration events from the Middle East that vastly changed the genetic background of the populations from the region. We were also interested in measuring the genetic impact of migrations from neighbouring populations (Middle East, Europe, and sub-Saharan Africa) in both Amazigh and Arab groups.

We implemented seven demographic models where the North African groups diverge at different moments in time from the surrounding populations (Additional file 1: Fig S4): Model A considers North African groups as sister branches of West Africa. Model B shows North African populations splitting from the Middle East in a back-to-Africa movement. This is the model mostly supported in previous analyses by population structure and admixture-f3 approaches [6,7]. Model C implements a third variation of the possible origin of North Africa, placing it also in a back-to-Africa event prior to the split between Europeans and Middle Easterns. In Model D North African populations are split between Arab and Amazigh. Imazighen show a deeper origin splitting from the Eurasian branch, whereas Arab North Africans split more recently from the Middle Eastern branch. In Model E North African groups are separated, where Imazighen are a sister branch of West African populations while Arabs split from Middle Easterns. Model F represents the origin of Imazighen after the split of European populations but before the divergence between Middle Easterns and North African Arabs. Lastly, in Model G Imazighen appear as a sister branch of East African populations while Arabs split from Middle

Easterns. Since all models comprised migration between North African populations and their surroundings, we did not include these parameters in the model ascertainment in order to improve the power of the DL prediction to discriminate between competing topologies.

After training the DL neural network for identification of the seven demographic models using as input the joint site frequency spectrum (jSFS) of each of the simulations to the observed data jSFS (see materials and methods), cross validation analyses using simulated data as observed showed that the ABC-DL can efficiently distinguish among the competing models. The minimum success rate for the discrimination is 67% and 76% for models F and D respectively, and the larger confusion is shown between these two models, which is expected as these are the two more similar topologies evaluated (Additional file 1: Table S1). The fact that F and D show such a level of misclassification is not surprising as both models show similar demographic topologies, only distinguishing on the branch where North African Imazighen diverge from the surrounding populations. In Model D, Imazighen have an origin that predates the European split while in Model F the Imazighen origin is slightly more recent, after Europeans have diverged from Middle Easterns. This similarity in the Models in addition with the fact that there is limited genetic differentiation between Middle Easterns, Europeans, and North Africans (Fig. 1), causes this misclassification rate in the ABC analysis.

When applied to the observed data, the posterior probability estimated by ABC-DL strongly supports model D (posterior probability of model given data =0.922) (Additional file 1: Table S2). Furthermore, only the models with Imazighen as a separate clade from Arabs and with a back-to-Africa movement show posterior probabilities greater than 0 (Additional file 1: Table S2). While both models are selected by the ABC, Model D represents the data 11.8 times better than Model F

(Additional file 1: Table S3), presenting a robust result on the topology discrimination.

We further explored the D topology by adding migration pulses. We differentiate between migration and admixture pulses, with migrations referring to recent and continuous gene flow between two populations and admixture pulses defined as discrete and past gene flow from a target to a source population. We evaluated five different migration scenarios for Model D (Fig. 2): Model D_1 represents Model D without migration or admixture events. Model D_2 includes gene flow of both North African populations from/to surrounding populations. The model also includes several admixture pulses from Middle East to both Arab and Amazigh groups (which could mimic historic and pre-historic admixture events [13]) and admixture pulses to both North African populations from both East & West Africa representing that could represent the effect of trans-Saharan slave trade in the region [13]. Model D_3 increases the complexity of Model D_2 by including a "ghost" population from Eurasia. This population is described elsewhere [21] as a way of explaining basal substructure Out-of-Africa. We include two admixture pulses from this Basal Eurasian population, one to its sister branch and a second one to the European-Middle Eastern branch, representing a possible weaker influence of this Basal Eurasian on Amazigh respect to European, Middle East, and North African Arab populations. Model D_4 is the most complex model since it includes a second "ghost" population in sub-Saharan Africa that represents the population substructure in Africa. This "ghost" population admixes with the "real" populations in Africa as described previously [22]. On top of that, this model also includes an admixture pulse from European populations to both North Africa [13] and one Middle Eastern admixture pulse to Imazighen. Finally, Model D_5 reduces the complexity by excluding all admixture pulses that do not

come from “ghost” populations. This model is a simplification of Model D_4 to test if admixture pulses towards our target population are needed to explain the North Africa scenario.

Cross validation analysis between these five models shows that ABC correctly identifies the model in 30-60% of the simulations (Additional file 1: Table S4). This limited sensitivity is compatible with the hypothesis that all models are remarkably similar between themselves with a lot of admixture events that may result in very similar statistics comparing one model to another.

When applied to the observed data, we obtained that Model D_4 is the “best” model with 76.2% of accepted simulations (Additional file 1 Table S5) and a Bayes factor [23] of 8.07 to the second “best” model (Additional file 1: Table S6).:

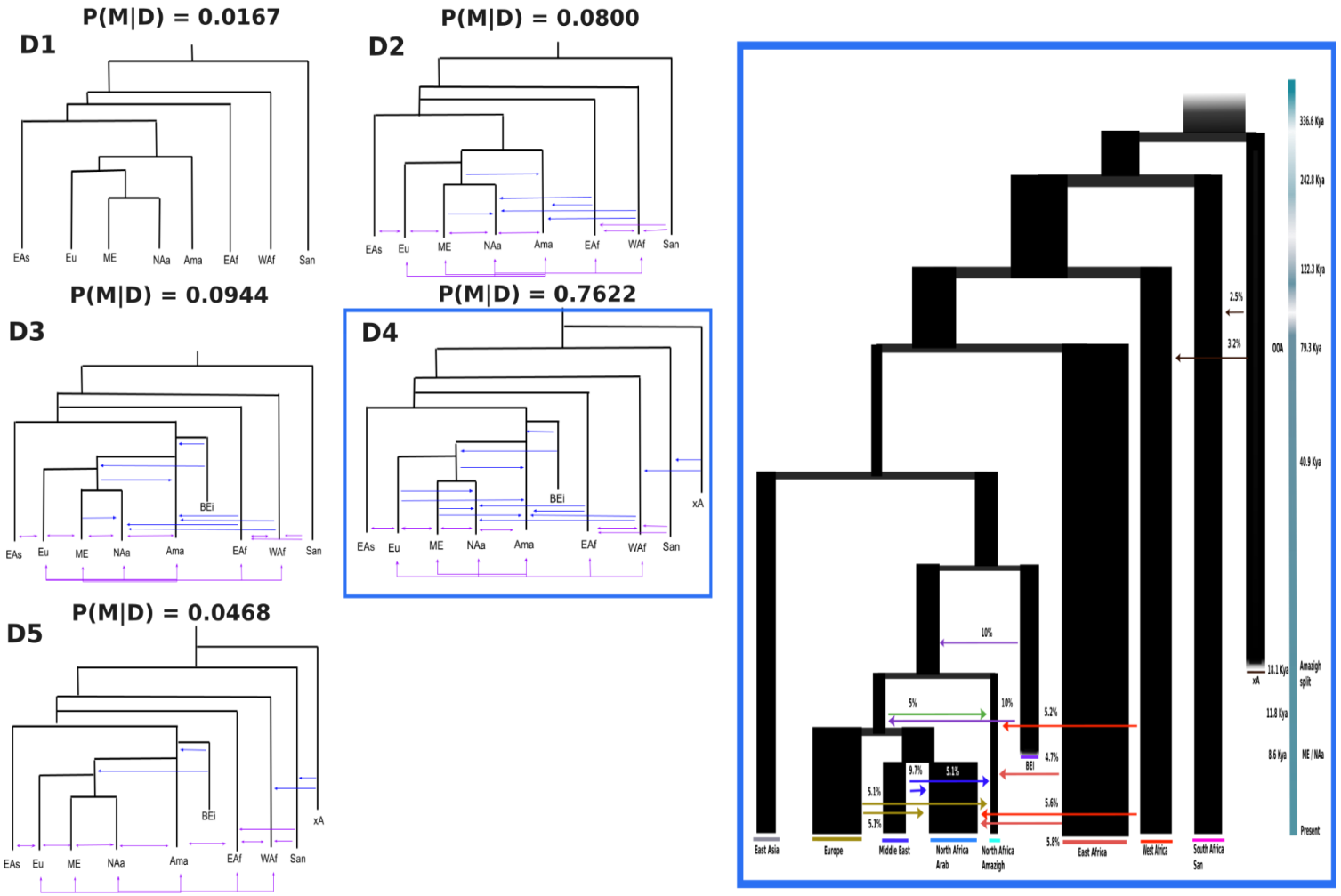

P(M|D) = 0.0167
D1
EAs
Eu
ME
NAa
Ama
EAf
WAf
San
P(M|D) = 0.0800
D2
P(M|D) = 0.0944
D3
BEi
P(M|D) = 0.7622
D4
xA
P(M|D) = 0.0468
D5
336.6 Kya
242.8 Kya
122.3 Kya
79.3 Kya
40.9 Kya
18.1 Kya
11.8 Kya
8.6 Kya
Present
Amazigh split
ME / NAa
OOA
2.5%
3.2%
10%
5%
5.2%
4.7%
9.7%
5.1%
5.6%
5.8%
East Asia
Europe
Middle East
North Africa Arab
North Africa Amazigh
East Africa
West Africa
South Africa San

Fig. 2: Tested demographic models including migration pulses. Left figures: variations of the selected topology on the first run of ABC-DL (Model D), included on the ABC-DL analyses considering North African Arab (NAa), North Africa Amazigh (Ama), Middle Eastern (ME), European (Eu), East Asian (EAs), East African (EAf), West African (WAf), and Ju/'hoansi (San) populations. Distinct levels of complexity are shown in each tested model, with purple lines representing recent migration and blue lines indicating admixture pulses from surrounding regions to both North African populations. The posterior probability obtained with our ABC-DL approach is shown on the top of each model. Right figure: fitted D_4 model with estimated parameters. Coloured lines represent admixture pulses from the specific population defined by the same color.

Next, we estimate the posterior probability for each of the 82 parameters form Model D_4 by applying the ABC-DL approach [19,22]. For each parameter, we analysed to which extent the posterior distribution captures its real value. First, we computed the factor 2 statistic (Additional file 2: page 4: Table S1), defined as the number of times that the estimated mean is within the 50% to 200% of the true value of the parameter [23]. In most of the cases, the factor 2 analysis indicates high confidence in the estimation of the true value of each parameter. Time splits show significant better results than effective population size estimates, ranging from 98% (split of Europeans from the Middle East-North African Arab clade) to 100% in the West Africa split, compared to the 75% (effective population size of North Africa_ Middle East (NA_ME)) to 99% of West Africa effective population size (Additional file 2: page 4: Table S1). Regarding introgression parameters, introgression times show better factor 2 results (98% Basal Eurasian to European, Middle East, and North African populations) than the introgression amounts (77% NA_ME to Imazighen as the best value). In fact, the less accurate performance using the mean as proxy is for migration parameters. Factor 2 statistics

show ~ 70% chance of true value being in the 50% to 200% range of the estimated mean, which is similar to the value obtained when calculating the mean from random sampling of the prior distribution (Additional file 2: page 4: Table S1). Despite the weak performance in the estimation of migration parameters, the models with migration perform better in the ABC-DL than those without migration. These results indicate that the ABC-DL framework allows us to obtain a confident set of posteriors in most of the parameters, particularly for some of the more relevant ones since they show high factor 2 values and very significant differences between prior and posterior distributions (Additional file 2: page 4: Table S2). Of particular interest is the time of the split of Imazighen to the rest of European, Middle East, and North African populations. In 99.5% of cases, the mean posterior distribution of the time split of Imazighen to the rest of European, Middle East, and North African populations is within the factor 2. This suggests that the mean of the posterior distribution of this parameter can be considered a good proxy of the real value.

When the ABC-DL is applied to each parameter of the model (Fig. 2), we observed that the posterior distributions for a large number of parameters were significantly different from the prior distributions (Additional file 3: histograms; Additional file 2: page 4: Table S2) and that in most cases there is a correlation between predicted and simulated values in the ABC analysis (Additional file 4: Spearman correlation; Additional file 2: page 4: Table S3). According to the ABC-DL analysis, the North African Arab population diverged from Middle East common ancestor 8.6 Kya (97.5% credible interval (CI) ranging from 4.65 Kya to 15.40 Kya, assuming a generation time of 29 years per generation [24]). Imazighen split from the rest of Eurasian populations at 18.12 Kya (97.5% CI of 9.68 Kya to 27.33 Kya), whereas the Out-of-Africa event is estimated at 79.28 Kya (97.5% CI

51.11 Kya to 108.39 Kya). The effective population size indicates a reduction between the split with East Africa (Ne = 23,401 (97.5% CI: 4,465 to 39,557)) and the split with East Asia (Ne = 5,619 (97.5% CI: 1,292 to 9,713)), coinciding with the Out-of-Africa bottleneck. In some cases, the effective population size is slightly larger than the expected given our prior distributions (Additional file 3: histograms; Additional file 2: page 3). All parameters are estimated considering a mutation rate of 1.61e-8 ± 0.13e-8 mutations per bp [25].

The admixture estimates obtained from Model D_4 show multiple migration pulses from Middle East, Europe, and sub-Saharan African populations towards both Amazigh and Arab populations in the last 200 generations. We observe that the amount of admixture from Middle East to North African Arab is larger than to Amazigh (9.7% [97.5% CI: 19.4% to 0.5%] vs 5.1% [97.5% CI: 9.8% to 0.3%]). Another interesting result is the 20% introgression from the “ghost” basal Eurasian population towards the MRCA of Middle East, European, and North African in at least two pulses of admixture. Other pulses of less intensity from East and West Africa to North Africa (~ 5%) and from a “ghost” African population to San and West Africa are also observed (2.5% [97.5% CI: 4.9% to 0.2%] and 3.2% [97.5% CI: 4.9% to 0.4%], respectively).

Finally, we tested the robustness of the proposed model to generate datasets compatible with the observed genetic diversity in the data. We simulated 1000 datasets using the mean estimated at each parameter from Model D_4 with fastsimcoal2. In order to quantify how similar each simulated dataset resembles the observed data, we compared the jSFS obtained for each simulation with the observed jSFS using a replication -unseen during the training of the DL- dataset by means of a PCA [26,27]. The results of the PCA indicate that the

model does not correctly replicate the data since the observed jSFS falls as an outlier in the PCA of the simulations (Additional file 1: Fig. S5; Additional file 1: Fig. S6). This suggests that the current static models used might not properly capture the whole demographic complexity present in North Africa.

Unravelling demographic history using Genetic Programming for Population Genetics (GP4PG).

To overcome the limitation of the ABC-DL approach in the reconstruction of North African population history, we have developed a novel approach to explore the demographic parameter and model topology space of a demographic model based on the paradigm of Genetic Programming. Genetic Programming is a meta-heuristic method inspired on evolution to generate formulae or programs coded as trees. Each node in the tree represents an operation or function, and the branches represent arguments or operands. An evolutionary approach Genetic Programming for Population Genetics (GP4PG) algorithm considers evolutionary events, such as changes in the effective population size or the increase or decrease of population substructure, as operations. GP4PG search the space of possible configurations of evolutionary events to define the demographic events that would generate genomic datasets similar to the observed ones (see material and methods). These demographic events include the ones already considered in ABC-DL, such as the presence of admixture, effective population sizes, or time of the demographic event. However, GP4PG allows modelling population substructure within each of the considered populations, or ecodemes [28,29], from the demographic model. In this framework, each ecodeme is formed by multiple topodemes that relate to each other following an isolation by distance pattern [30], where topodemes that are situated closer will

migrate at higher rates than topodemes that are further apart (Additional file 1: Fig. S7). This allows generating reticulated and partially reticulated demographic models [31].

In addition, this new algorithm also takes into account population substructure when exploring different demographic scenarios. In our case, we considered eight ecodemes: San, West Africa, East Africa, Middle East, East Asia, Europe, North Africa Amazigh, and North Africa Arab, all with the same size and without distance between neighbouring ecodemes to simplify the models.

We apply the GP4PG algorithm with the six of the considered topologies used with the ABC-DL. We discarded Model A since it consistently showed the worst performance in multiple iterations of the ABC-DL reducing computing times and resource allocation. For the remaining six models, we used two different versions: a variant that includes migration between topodemes (populations) and a variant without including migrations resulting in 12 different topologies (Additional file 1: Fig. S8). Since GP4PG is a metaheuristic approach based on exploring graph topologies (see material and methods), it can be trapped into local optima. Therefore, we independently run 40 times the GP4PG algorithm, each for 200 iterations, retrieving the best demographic model from the end of the 200 iterations. After the 200 iterations the fitness error between the simulated dataset and the observed dataset at each run of GP4PG reaches a plateau (Additional file 1: Fig. S9). Interestingly, we observe differences in the final error of each run, ranging between 10 to 2, thus supporting that GP4PG tends to get trapped in local optima.

Model D is the one most supported by the GP4PG with 10 out of the 40 replications performed. The model that presents the minimum error

is Model D (Model D_15 - Error = 2.56). Next, we assessed the performance of each model to predict the observed data. We selected the 10 models from the GP4PG runs with the least amount of fitness error. For each of them, we computed a thousand simulations. For each simulation we computed the 4-fold SFS of all the simulations. Finally, we compared them to a replication set of the observed data using PCA. We compared the results of the ABC-DL simulation, transforming the jSFS into a 4-fold SFS to evaluate if the performance of the GP4PG is better than the ABC-DL (Additional file 1: Fig. S9). Simulations from the ABC-DL ModelD_4 fall further to the observed data than any of the GP4PG models in the PCA indicating that the models produced by the new methodology are consistently better at describing the observed data than the models produced by the ABC-DL (Additional file 1: Fig. S9).

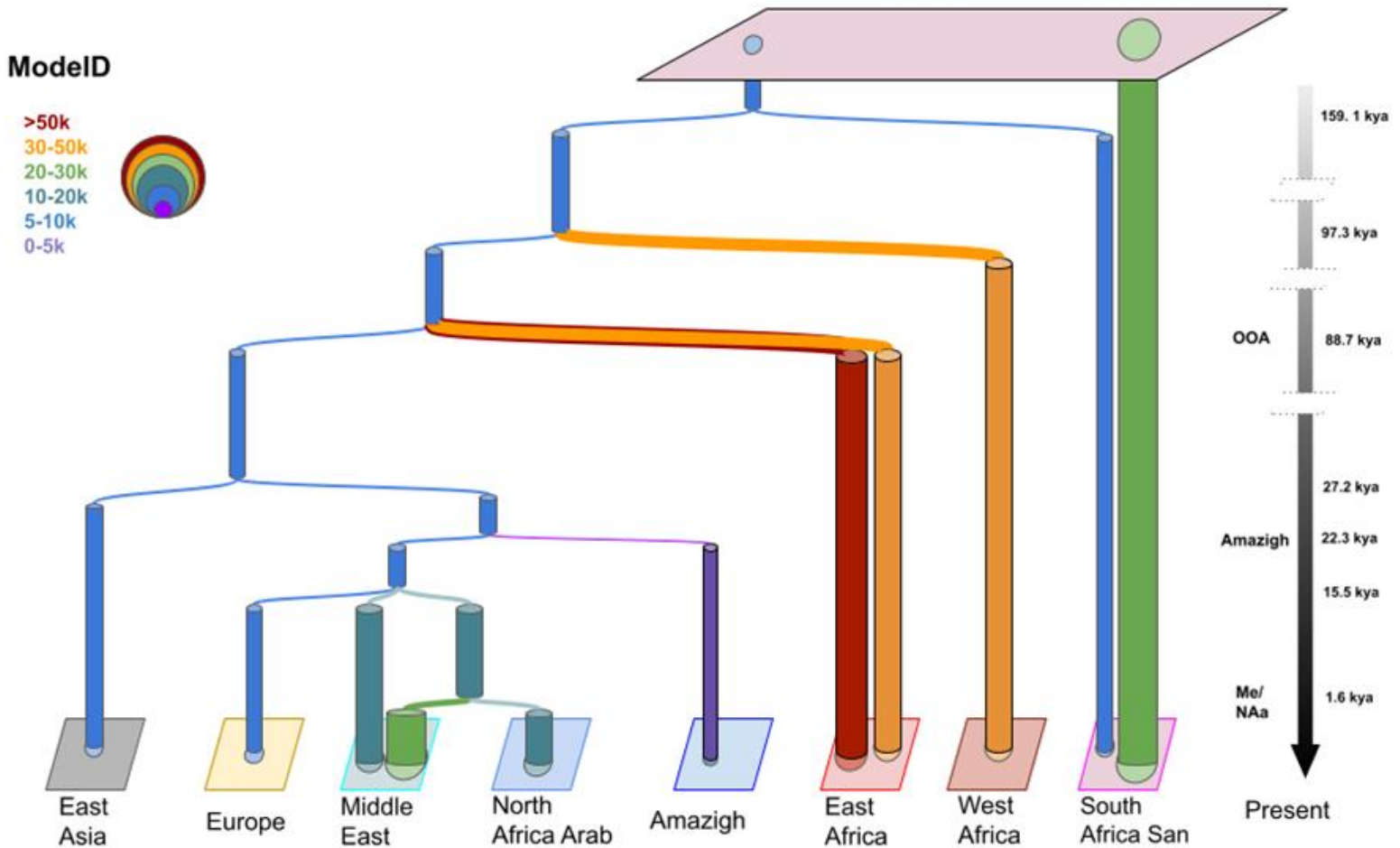

ModelID
>50k
30-50k
20-30k
10-20k
5-10k
0-5k
159. 1 kya
97.3 kya
OOA
88.7 kya
27.2 kya
Amazigh
22.3 kya
15.5 kya
Me/
NAa
1.6 kya
Present
East
Asia
Europe
Middle
East
North
Africa Arab
Amazigh
East
Africa
West
Africa
South
Africa San

Fig. 3 : Fitted model obtained with GP4PG algorithm. Demographic model of North African populations with the least amount of error in the GP4PG algorithm (E = 2.56), iteration 15 out of 40. This model presents a different origin of North African Amazigh and Arab. Each color represents a different range of Ne also represented by the width of the columns. The model shows some population substructure, especially in sub-Saharan African populations and presents the ability to explain complex demographic models with population substructure and decaying migration after split pattern.

After verifying that simulated data from models inferred by GP4PG fit the genetic diversity observed in the real data, we ascertained the model interpreted as the “best” after both ABC-DL and GP4PG (Fig. 3). This demographic model presents a topology that proposes different demographic histories between Amazigh and Arab populations, with Amazigh splitting from the MRCA around 22.3 thousand years ago. On the other hand, North African Arab individuals split from Middle Easterns around 1.6 Kya, which might be related to the Arabization process in North Africa. This result contrasts with the one obtained with the ABC-DL approach that estimated an older split of North African Arabs closer to Neolithic times. Another detail that we deduced from this demographic model of North Africa is that admixture pulses do not appear as drivers of genetic diversity in any of the best 10 models. Instead, most of the current genetic diversity can be explained by a combination of population substructure and migration decay between demes. The rest of the parameters of each model can be found in the supplementary material (Additional file 5: parameters for the best 10 models of the GP4PG).

DISCUSSION

North African populations are a demographic melting pot [7,13]. The presence of complex and recent demographic histories with multiple gene flow events from different sources difficult our ability to construct a feasible demographic model with current methodologies. The absence of model-specific statistics recapitulating particular events even for moderately complex evolutionary problems has fuelled the development of methods based on machine learning [32].

The application of the ABC-DL approach considering relatively complex demographic models supports that the Imazighen present a different demographic history compared to surrounding populations. Prior studies with autosomal markers on North Africa populations pointed to Amazigh populations presenting a unique genetic component that has been described as the autochthonous component for the region and that can be traced back to, at least, Epipaleolithic times [3,7]. This is supported by our ABC-DL model as we estimate a continuity of the Amazigh population up to 18.14 Kya (97.5% CI of 9.68 Kya to 27.33 Kya) falling in the range of the Epipaleolithic samples of Taforalt (15-13 Kya) [3]. The Arab population in North Africa, on the other hand, shows higher genetic affinity with Middle Eastern groups than to the ancient Taforalt samples. This affinity is described by an east to west cline on the amount of Middle East-like genetic ancestry across North Africa and has been hypothesized as a consequence of several migration movements from the Middle East to North Africa with the Neolithization and the Arab expansion being the more relevant ones [13]. The ABC-DL model supports the Neolithic expansion to North Africa as the main source of the Middle Eastern component. This is supported by divergence times between North African Arabs and Middle Easterns at 8.6 Kya (97.5% CI of 4.65 Kya to 15.40 Kya) which

overlaps with previous studies on Neolithic expansion in North Africa [8]. We observe at least one pulse of admixture in late Neolithic from the Middle East to both North African populations around 5.6 - 5.9 thousand years ago (97.5% CI of 0.54 Kya to 13.94 Kya to Arab; 97.5% CI of 0.63 Kya to 14.03 Kya to Amazigh), highlighting the importance of Neolithic demic diffusion in North Africa [8,33]. Other admixture events from surrounding regions are also supported by our model. European admixture in Neolithic times is present, although always placed later than the Middle Eastern admixture (Additional file 2: page 4). This is observed both in both Amazigh and non-Amazigh populations in small proportions (~5%) which is consistent with the hypothesis of European contact in western North Africa at least 5,000 years ago [8–10]. In addition to these gene flow events, it has been proposed the presence of a basal Eurasian population that could explain early diversity in West Eurasia, North Africa, and the Near East [21]. We observed that, when including this “ghost” basal Eurasian population in our models, the performance of the models increases, reinforcing the hypothesis of the basal Eurasian population proposed previously [21]. Focusing on the performance of the ABC-DL approach in North Africa, some of the main flaws are related to the estimation of effective population sizes. In some populations the posterior distribution of effective population size is larger than the initial prior distributions we provided. This might be an issue when estimating several parameters since effective population size and split times are highly correlated. Effective population size is dependent on heterozygosity [26]. As the heterozygosity increases, the effective population size or the split time of a population needs to increase in order to reach coalescence [26]. We observe that the heterozygosity in the North African Arab samples is higher compared to European and Middle Eastern populations (Additional file 1: Fig. S10) which can be a result of recent gene flow from sub-Saharan Africa. This increased

heterozygosity hinders the estimation of split times and effect sizes as we may need bigger sample sizes or more complex scenarios to take this diversity into account.

Despite the reasonable performance of the ABC-DL approach, it has replicability issues. The accepted models cannot grasp all the diversity of the data and are biased by our preconceptions of the history of the populations. Moreover, a further drawback of the methodology is the black-box nature of DL approaches [34], which limits interpreting which parameters from the model should be modified in order to improve the performance of the inferred demography. To overcome these limitations, we have developed a new algorithm, GP4PG, inspired by the field of genetic programming and evolutionary algorithms [35–38], for automatizing the exploration of complex parameter-free demographic models. GP4PG performs, in terms of reproducing the observed genetic variation, better in the North Africa scenario than the best possible model obtained with ABC-DL algorithms by minimizing the difference between our dataset and the simulations obtained from the best model (Additional file 1: Fig. 9). It also reduces the amount of bias that can be imputed due to the building of each of the competing models, given that the only inputs to the GP4PG algorithm are the topologies for each of the models.

The GP4PG analysis, like the ABC-DL, supports that Amazigh and Arab populations in North Africa have different demographic histories despite both originated as back-to-Africa movements from Eurasian populations. Amazigh groups appeared before the European, Middle Eastern, and North Africa Arab branch, splitting from a common ancestor around 22.3 Kya (Fig. 3). This date precedes the oldest genomic data available in the region [3] for at least seven thousand years, pointing to a genetic continuity even before than expected, and

which is closer to the earliest known appearance of Iberomaurisian culture in Northwest Africa (25.85 to 25.27 cal. Kyra) at Tamar Hat [39]. On the other hand, North African Arab populations appear to have split from Middle Easterns 1.6 Kya, placing this population closer to an Arabization replacement on North Africa than to a Neolithization demic diffusion. This result contrasts with the ABC-DL results that show a deeper impact of the Neolithization process in the North African Arab - Middle Eastern split. The difference in the divergence time can be attributed to the effect of migration. From a coalescent perspective, migration between two populations generates a distinction between the coalescence time between the lineages and the split time between the populations. In models where migration is included as a parameter, the split time between North African Arab and Middle East populations is closer to the one observed in the ABC-DL analysis (ModelD_7 at 6.41 Kya; ModelD_26 at 13.39 Kya; Additional file 5) implying that the inclusion of migration increases the time of coalescence between the populations.

The absence of admixture events in the best model of the GP4PG algorithm can be attributed to the fact that our algorithm is weighted towards modelling population splits by a soft process in which migration decays forward in time between related populations. Therefore, whenever a split between two populations occurs, migration between the newly formed topodemes continues for generations, decaying as the demes grow further apart until they achieve a constant migration rate. Our selected model presents population substructure in current sub-Saharan African populations that extends to ancient times (Fig. 3). These observations support the results obtained by Ragsdale et al [31], where reticulated population substructure tens of thousands years ago could explain some of the genetic diversity previously attributed to archaic introgression [19,22]. Although the most robust

model supports this idea of soft splits between populations, we do not rule out the possibility of admixture pulses as we observe in the ABC-DL analysis because we hypothesize that the effect of continuous migration after split could be mixed up with the effect of admixture pulses.

Since our analysis includes a limited number of samples (n = 2 per population), the study lacks some power to confidently corroborate some of these results, especially for the sub-Saharan population. Despite this, the results indicate that the models always perform better when including some amount of population substructure. On top of that, we must be aware that the selected Imazighen individuals are part of a very isolated population, the Chenini Amazigh. This population is an outlier in North Africa (Fig.1_b) due to isolation. This characteristic makes it useful as a proxy of the North Africa autochthonous component given that the amount of sub-Saharan and Middle Eastern genetic components is lower than in the rest of North African Imazighen groups [40]. Nonetheless, taking into account the heterogeneity within Imazighen groups due to different amounts of genetic components coming from neighbouring populations, using other Imazighen groups could lead to slight differences in some of the studied parameters.

Most of the selected topologies in the 40 runs of the GP4PG are either Model D (25%), Model C (17.5%), or Model F (20%). These three models present very similar topologies, with slight variations, mainly in setting the Amazigh origin. The current implementation of the GP4PG algorithm has enough statistical power to discriminate between competing models but falls short to detect fine scale migration and admixture events. This is due to the multimodal nature of the SFS that can lead to similar genomic patterns [26,41] with different demographic

models. Implementation of haplotype-based summary statistics to the GP4PG algorithm in the future could solve some of these issues in very complex demographic scenarios such as that of North Africa.

CONCLUSION

In sum, we have built a robust model of the demographic scenario of the North African populations. By implementing an ABC-DL algorithm and a novel GP4PG algorithm based on metaheuristics, we have defined a clear topology that proposes a back-to-Africa origin and differentiates the Amazigh-speaking population from the Arab-speaking groups in the origin and settlement in northern Africa. Our data point to a complex scenario where population substructure and admixture events had a significant impact on the genetic structure of current North African populations.

MATERIAL & METHODS

Databases

To analyse the population structure and demographic history of North Africa a dataset was compiled consisting of 32 whole genomes sequenced on deep coverage from North Africa. This dataset includes fifteen newly published samples from Morocco (n = 6) and Tunisia (n = 9) (Ref : EGAXXXX), Imazighen (n =4) and non-Imazighen (n = 6) individuals from Serra-Vidal 2019 [6], Egyptian samples (n = 3) from Pagani et al. 2015 [42] and Saharawi (n = 2) and Mozabite (n = 2) North Africa individuals extracted from the Simons Genome Diversity Project (SGDP) [43]. These North African samples were merged with a panel of world-wide populations from the SGDP (n = 295) [43], the 1000 Genomes Project (n = 38) [44] and high coverage Qatari individuals (n = 9) from Fakhro et al [45]. The final dataset used for the

population structure analysis consists of 374 whole genome high coverage individuals. Furthermore, to perform the demographic inference analysis, a subset of individuals from each of the proxy groups for every population was analysed (Additional file 1: Table S7). Two North African Arab speaking population (Tunisian Arab), two from a Tamazight speaking group in North Africa (Tunisian Chenini), two Middle Eastern representatives from Qatar, two Northern European from Utah (CEU), two East Asian from Han (CHB), two West African from Yoruba population in Ibadan (YRI), two East African from Luhya in Kenya (LWK) and two South African San from Ju/'hoansi North in Namibia (JHN). The final dataset for the demographic inference analysis comprised 16 individuals with an individual whole-genome coverage of >30X.

Read Mapping and Variant Calling

Single-nucleotide polymorphism (SNP) genotype variation of each sample was obtained by the following procedure. Read quality assessment of the fastq files was performed with fastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA-MEM v0.7.13) [46]. Reads were then sorted using SAMtools v1.2 [47] and duplicates were removed using MarkDuplicates from Picard (https://broadinstitute.github.io/picard/). Indels were realigned and quality scores were recalibrated using the Genome Analysis Toolkit (GATK 4.1.8.1) [48]. Variant calling was done using the HaplotypeCaller and merging of each sample into a multisample VCF was done using the GATK GenomicDB-GenotypeGVCFs functions [48].

We obtained the structure dataset by merging all samples, keeping the common SNPs across all samples, and applying SNP (geno = 0.1) and individual missingness (mind = 0.1) in PLINK v1.9 [49–51]. We then excluded the SNPs with MAF < 0.01 and removed the individuals with cryptic relatedness using KING [52] based on a cutoff of 0.325. The final dataset comprised 9.68 M SNPs on 365 individuals from 150 different populations.

Data filtering

For demographic modelling, the dataset was further filtered to obtain a confident set of variants by implementing the following criteria: (i) a minimum of 5 reads mapped for each locus, (ii) a quality score threshold for the alternative allele of the variant, with a minimum score of 20 in the QUAL field of the VCF file, (iii) a PASS in the genotyping quality, (iv) exclusion of regions covered by structural variants [22] using TandemRepeatMarker repeats of length greater than 80 bp (UCSC browser) and 1000 Genomes Project copy number variants (https://www.ncbi.nlm.nih.gov/dbvar/studies/studyvariants_for_estd199 .csv), (v) exclusion of regions adjacent to indels with a 6bp flanking region, and (vi) exclusion of multiallelic variants. Based on these filtering steps, we obtained a 1.962.660.202 bp-long callable genome containing a high-confidence set of 16.4 M SNPs for downstream analyses. The ancestral state of each variant in these genomes was set to the chimpanzee reference genome (panTro4 genome assembly from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro4/reciprocalBest) to avoid any discrepancy between African and non-African populations as detailed elsewhere [19,22].

Structure analysis

Principal component analysis

Principal component analysis was performed using flashPCA [53], pruning the data for linkage disequilibrium between the markers using PLINK v1.90 [49–51] based on an r2 threshold of 0.4 in every continuous window of 200 SNPs with a step of 25 SNPs (i.e., –indep-pairwise 200 25 0.4).

ADMIXTURE analysis

ADMIXTUREv1.3 [54] was applied on the whole structure dataset, which was previously pruned for linkage disequilibrium between markers (–indep-pairwise 200 25 0.4). ADMIXTURE-ready dataset had 1.58 M SNPs on 365 samples. ADMIXTURE in unsupervised mode was run assuming several ancestral clusters ranging from K = 2 to K = 12 with 10 independent runs for each K using different randomly generated seeds for each run. The cross-validation error was assessed for each run, with K = 3 to K = 9 giving the minimum error. We then ran pong [20] in the greedy mode in order to identify common modes among different runs for each K to align clusters across different values of K.

Demographic Model

To decipher the complex demographic scenario of North Africa, we used an Approximate Bayesian Computation with Deep Learning as explained more in detail in Mondal et al. [19]. The ABC-DL implementation is a three-step analysis. First, we generated thousands of simulations with fastSimcoal2 [55,56] for each of the competing models using the joint multidimensional site frequency spectrum among populations (jSFS) as a raw summary statistic. This statistic

contains the information to run most of the frequency-based statistical analyses used in population genetics which are informative for detecting most of the demographic parameters considered in our demographic models (more in [19]. Second, we trained a DL to predict, from the jSFS, the most informative summary statistic (SS-DL) of the considered parameter or set of models. A potential limitation of this approach is the fact that the DL is trained with simple data and compared to the real model generated by the observed data, possibly overfitting our models. To avoid biases in the DL prediction, we injected jSFS noise in each simulation from the real data (see [19]). Finally, we performed a classical ABC approach using the SS-DL in a new set of simulated datasets.

The callable fraction of the genome we used for the demographic modelling is a modification of that defined by Pouyet [57]. We further cleaned the original callable genome, defined by Pouyet, to identify neutral regions by masking genomic regions containing Ensembl genes within a 20 kb range and masking CpG islands as defined elsewhere [19]. After that, we excluded all regions failing to reach a SNP density over 90% on 10 Kb windows with a sliding step of 2500 bp. to obtain 53.7 Mb of callable genome that was used in the next steps.

Ten Deep Learning (DL) networks were created, with four hidden layers each, and trained with 20,000 simulations each. An additional set of 180,000 simulations per model was generated, injecting noise from the observed jSFS of an individual of each population (BTUN01, TUN01, NA18559, NA12878, SR098230, NA19037, NA19207, HGDP01032), and the probability of each model was predicted using the 10 DL networks. The results were averaged to get the summary statistic (SS-DL) for the Approximate Bayesian Computation (ABC)

analysis, carried out using the “abc” package in R [58,59]. We have assumed a mutation rate of 1.61e-8 ± 0.13e-8 [25] and a generation time of 29 years [24].

The ABC process can be divided into two steps. First, we checked how well the “abc” was able to distinguish between competing models. To do so, we applied the cv4postpr function on the “abc” package that performs a cross validation analysis on the ability of

25

distinguishing between models [58]. This was done using 50 simulations per model and the same ABC parameters we used when analysing the observed data. The cv4postpr runs the ABC using simulated data as observed data and counting the number of times that the model with the highest posterior probability was, in fact, the model that generated the simulated data. Once the “abc” was able to distinguish between the different competing models, we began the discrimination of the “best” model. To do so we applied the “postpr” function of the “abc” package [58], keeping the 1000 best simulations (out of a total of 180,000 simulations per model) under a “mnlogistic” option for model comparison (tol = 0.001). We applied this procedure twice, one for the discrimination between the 7 competing models and another one to select the best variant for the topology selected before.

The DL process and ABC calculation were repeated for the parameter estimation of each parameter in the “best” model to get the posterior range for all demographic parameters. Mean, median, mode, 95% Credible Interval (CI) and 95% Highest Density Interval (HDI) [60] were calculated for each parameter. Finally, Spearman correlation, factor2 [23], and Kullback-Leiber distance [61] were used to assess the quality of the posterior predictions.

GP4PG

Model comparison by ABC requires defining which are the models to consider. These models are, by definition, simplifications of reality. However, basic assumptions about the demographic events, and particularly population substructure, can significantly bias the model ascertainment [31]. Previous studies have dealt with the presence of hidden population substructure by modelling “ghost” populations [19,21,22] or by generating ancient weak structured stems that interact forming the current populations [31]. To bypass this issue, we have developed the Genetic Programming for Population Genetics (GP4PG), which is based on using genetic programming (GP), a branch of natural computing.

Natural computing refers to meta-heuristic algorithms inspired by nature to solve –by means of optimizing an error function– complex problems that are otherwise intractable. The underlying rationale of natural computing is that strategies used by nature to solve natural problems can be applied by reverse engineering to human-based-problems. Within the context of natural computing, Darwinian evolution inspires a broad family of algorithms called Evolutionary Algorithms that mimic the process of how evolution works to adapt an organism to its environment. Within the evolutionary algorithm family, Genetic algorithms have been already used in population genomics for demographic parameter definition [62]. However, GP is better suited for generating formulae and population relationships as the algorithm codes solutions in the form of a graphical (tree) structure whose nodes or edges represent parameters [37,38]. GP is an automated invention machine, which routinely delivers high-return human competitive machine intelligence, duplicating the functionality of previously patented inventions, infringed a previously issued patent, or created a

patentable new invention [63]. The basic workflow of a GP embraces the basics of the biological evolution of a population, including selection, recombination, and mutation. Within this machine learning framework, a proposed solution, coded in the form of a graph, is considered as a biological being, which is subject to selection according to how good is for predicting the parameter of interest –how good is the summary statistic to distinguish models, or to predict a parameter of a given model. A set of solutions define a population that evolves over generations, exchanging information through recombination and exploring the surrounding space through mutation, to minimize an error function. Within the context of demographic modelling, the nodes refer to possible demographic events – and the tree depicts a demographic model (Additional file 1: Fig. S11; Additional file 1: Table S8).

The proposed GP4PG framework considers the particular features of demographic modelling and applies some modifications to the classical GP algorithm to account for them. GP4PG organizes populations in fundamental homogeneous groups called “demes” that are nested within “topodemes" and “ecodemes” [29]. Gilmour and Gregor defined these concepts from an ecological perspective. Ecodeme refers to those topodemes sharing a given habitat, while topodeme is used to group the individuals that are from the same locality. Each ecodeme can have one or more topodemes depending on the heterogeneity of the population. In a population genetics and demographics scenario, we used topodeme as a unit of population substructure to resemble the different population nuclei that we could observe in a population and that suggest the internal diversity that be observed in a given region (“ecodeme”) [28].

The GP4PG algorithm models the migration rate of each ecodeme within itself and with adjacent ecodemes following an isolation by distance approach [30]. Thus, populations situated further apart would have less chance to exchange migrants than those that are closer together. For the sake of simplicity, we considered that populations all occupy a same size space and that the distance between adjacent populations was zero, meaning that geographical barriers, such as the Sahara Desert or the Mediterranean Sea, were not considered to order the populations in the space. All models included a migration decay function. Following the split between two ecodemes, these demes continue to exchange migrants, but the quantity is gradually reduced over time until it reaches the migration rate defined between the two populations or they stop sharing migrants if the models do not have migration. This modelling approach accounts for the fact that the migration rate between two populations is not constant from the outset, but rather fluctuates and diminishes exponentially with time as they become more isolated and distinct.

First of all, the nodes used in GP4PG represent demographic events rather than operations to be added to a formula. Each demographic event requires a time when it occurs, and such time determines the relationship with its preceding nodes in backward. The GP4PG will also choose from the several possible demographic events (admixture, addition and reduction of demes or changes in the migration rates) and different combinations to produce the simulations that are going to be tested against the observed data. All the simulations were ranked by means of a standardized error comparing the jSFS (4-wise SFS) of each of the simulations to the observed data jSFS. The best simulations will, following the concepts of GP, produce more offspring than the worst simulations, that will suffer modifications using mutations and recombination with other models, similar to an

evolutionary process, optimizing the models after each generation. In this way, at each iteration, we'll obtain better models to explain the observed data.

Exploring the space of possibilities within the GP4PG framework

In canonical GP, exploration of the space of solutions is mainly accomplished by generating solutions from the most successful ones by means of recombination: exchanging sub-trees at a given node in both parents. Thus, the offspring is a combination of both trees using a subtree-crossover operator [36]. In the GP framework, modifications of the parental structure -i.e., mutation- are less frequent [36]. However, classical recombination approaches applied to demographic models could easily produce non-compatible solutions, where the root node of the replaced subtree from one parent could have older times than its preceding node. This does not occur if only the mutation process is applied, as in this case changing the time can be constrained to be between the ranges of the previous and next demographic event. There are different evolutionary strategies and evolutionary strategies-like algorithms whose main exploration force is a type of mutation. We adapted the invasive weed optimization algorithm (IWO) [64,65] to be used for GP-tasks. IWO emulates the process of colonizing new environments by invasive plants. In IWO, each solution present in the population reproduces proportional to its fitness. First, a finite number of seeds (demographic models) is produced. Reproduction of each solution depends on a seed offspring function [Eq.1] that lets those solutions that have better fitness reproduce more than those that are further from the optimum. This reproduction technique allows the chance to survive and reproduce for unfeasible solutions, hence not discarding possible useful information carried by low fitness individuals. Once the new solutions are proposed proportional to how

good each parental solution is, the new solutions are ranked from the best to the worse and a new population of the original size is generated by disregarding worse performing solutions.

In our case, we rank all the possible solutions according to how close the simulated dataset produces summary statistics (SS) close to the summary statistics observed in the real data (Eq. 1):

$$f_{sim} = \sum_{s=1}^{nsumstat} \left(\frac{SS_{sim}^{s} - SS_{obs}^{s}}{\sigma_{obs}^{sim}} \right)^2 \quad \text{(Eq.1)}$$

Where each element s comes from the jSFS of each simulated model computed among all possible combinations of four populations (4jSFS) against the 4jSFS of the observed data. The standard error of each element s is obtained by Monte Carlo resampling with replacement from the considered genomic fragments 1,000 datasets of the same genomic size as in the training dataset and computing for each dataset the 4jSFS.

By using 4-population-fold jSFS instead of the full multidimensional SFS (mSFS) among all populations, the total number of summary statistics is reduced to Eq. 2 instead of Eq. 3. This avoids the exponential explosive nature of the full multidimensional site frequency spectrum and the associated curse of dimensionality [66], reducing the number of mSFS combinations with value 0, while allowing to recapitulate the demographic relationships among populations [67]:

$$nsumstat_{4jSFS} = (3^4 - 2) * \frac{n!}{4!(n-4)!} \quad \text{(Eq. 2)}$$

$$nsumstat_{mSFS} = (3^n - 2) \qquad \text{(Eq .3)}$$

30

Evolving the population of answers

The best performing solution in our population (i.e., the one whose demographic model produces 4jSFS close to the one in the observed data) generated eight new solutions, each of them showing in probability differences on the time of the events, the events and particular parameters related to the events. The worst performing solution in our population produced two offspring. Other solutions reproduced in proportion S (Eq.4). In our analysis we have determined a $SSminmin$ of 2 and a $SSmaxmax$ of 8, this allows the preservation of low fitness solutions that could potentially give higher fitness offspring that otherwise with a more restrictive $SSminmin$ would be lost:

$$S_i = (S_{max} - S_{min}) * \frac{f - f_{min}}{f_{max} - f_{min}} + S_{min} \quad \text{(Eq.4)}$$

Model comparison with GP4PG

We tested six competing topologies in the GP4PG algorithm (Fig. 4), which are the same topologies tested for ABC-DL except for Model A. On top of that, we constructed two models for each topology, one considering migration between “ecodemes” and the other without it. For the GP4PG algorithm, in order to speed up the algorithm, we have used half the masking regions filtered from Pouyet [57], and a second set of data to validate the analysis. The GP4PG algorithm gives us the model that presents the least amount of error to the data for each iteration. In our case we have performed 40 iterations with 200 generations for each iteration. The 10 best solutions were then compared to the observed data by performing a PCA of the 4jSFS of a thousand simulations of each model against the 40 4jSFS of the replication dataset to take into account the deviation in the SFS due to the random selection of masked regions.

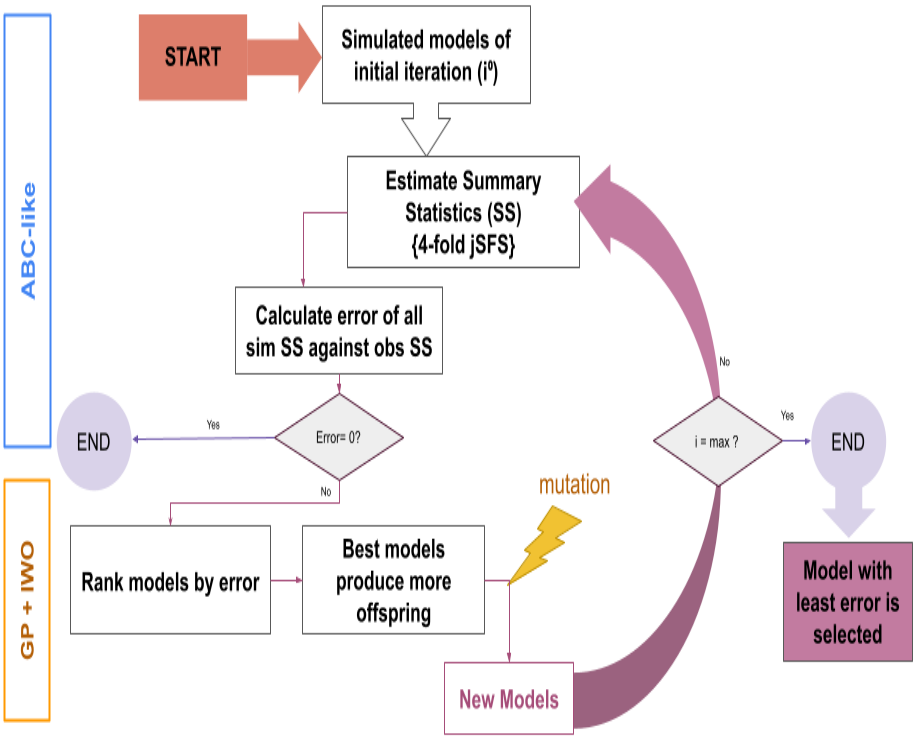


Fig. 4: Structure of the GP4PG algorithm. From an initial set of models, we compute a summary statistic (4jSFS) for each simulation and compare it to the observed data by the means of a fitness error function. Then, the errors are ranked. We produce the offspring (2nd generation) following an invasive weed optimization algorithm modifying each child model (mutation). We repeat this procedure until the error is 0 or it reaches a plateau.

DECLARATIONS

Ethics approval and consent to participate.

The present project has the corresponding IRB approval (Comitè d'Ètica d'Investigació-Parc de Salut Mar 2019/8900/I, Barcelona, Spain).

Consent for publication

The authors declare no competing interests and agree with the publication of the results.

Availability of data and materials

Funding

This work was supported by the Spanish Ministry of Science and Innovation (grant numbers PID2019-106485GB-I00 and RTC-2017–6471-1 AEI/FEDER, UE), Fundación CajaCanarias and Fundación Bancaria “La Caixa” (2018PATRI20), and “Unidad María de Maeztu” (CEX2018-000792-M) funded by the MCIN and the AEI (DOI:10.13039/501100011033). J.M.S was supported with a Formació de Personal Investigador fellowship from Generalitat de Catalunya (FI_B100135).

Authors' contributions

J.M.S., O.L., and D.C. conceived the work. J.M.S. and O.L. performed the computational analyses, generated the figures, and wrote the manuscript. D.C. contributed to analysis and/or interpretation of results. All authors approved the final manuscript.

Acknowledgements

We would like to thank the Scientific Computing Core Facility at the UPF (https://www.upf.edu/web/sct-sit) for their technical help and support.

Competing interest

The authors declare no competing interests.

REFERENCES:

1. Sahnouni M, Parés JM, Duval M, Cáceres I, Harichane Z, Van Der Made J, et al. 1.9 million and 2.4 million year old artifacts and stone tool–cutmarked bones from ain boucherit, Algeria. Science (80-). 2018;362:1297–301.
2. Hublin JJ, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature [Internet]. 2017;546:289–92. Available from: http://dx.doi.org/10.1038/nature22336
3. Van De Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, et al. Pleistocene north african genomes link near eastern and sub-saharan african human populations. Science (80-) [Internet]. 2018;360:548–52. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aar8380
4. Hervella M, Svensson EM, Alberdi A, Günther T, Izagirre N, Munters AR, et al. The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa. Sci Rep. 2016;6.
5. Pennarun E, Kivisild T, Metspalu E, Metspalu M, Reisberg T, Moisan JP, et al. Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. BMC Evol Biol. 2012;12.
6. Serra-Vidal G, Lucas-Sanchez M, Fadhlaoui-Zid K, Bekada A, Zalloua P, Comas D. Heterogeneity in Palaeolithic Population Continuity and Neolithic Expansion in North Africa. Curr Biol [Internet].

2019;29:3953-3959.e4. Available from: https://doi.org/10.1016/j.cub.2019.09.050

7. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic ancestry of North Africans supports back-to-Africa migrations. Schierup MH, editor. PLoS Genet [Internet]. 2012;8:e1002397. Available from: http://dx.plos.org/10.1371/journal.pgen.1002397

8. Pimenta J, Lopes AM, Comas D, Amorim A, Arenas M. Evaluating the neolithic expansion at both shores of the mediterranean sea. Mol Biol Evol. 2017;34:3232–42.

9. Simões LG, Günther T, Martínez-Sánchez RM, Vera-Rodríguez JC, Iriarte E, Rodríguez-Varela R, et al. Northwest African Neolithic initiated by migrants from Iberia and Levant. Nature. 2023;618:550–6.

10. Fregel R, Méndez FL, Bokbot Y, Martín-Socas D, Camalich-Massieu MD, Santana J, et al. Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. Proc Natl Acad Sci U S A [Internet]. 2018;115:6774–9. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1800851115

11. Elkamel S, Cherni L, Alvarez L, Marques SL, Prata MJ, Boussetta S, et al. The Orientalisation of North Africa: New hints from the study of autosomal STRs in an Arab population. Ann Hum Biol. 2017;44:180–90.

12. Lucas-Sánchez M, Fadhlaoui-Zid K, Comas D. The genomic analysis of current-day North African populations reveals the existence of trans-Saharan migrations with different origins and dates. Hum Genet. 2023;142:305–20.

13. Lucas-Sánchez M, Serradell JM, Comas D. Population history of North Africa based on modern and ancient genomes. Hum Mol Genet. 2021;30:R17–23.

14. Camps G. Els Berbers, mite o realitat? In: Roque MA, editor. Les Cult del Magreb. Barcelona: Enciclopedia Catalana; 1994. p. 41–74.
15. Fadhlaoui-Zid K, Plaza S, Calafell F, Amor M Ben, Comas D, El Gaaied AB. Mitochondrial DNA heterogeneity in Tunisian Berbers. Ann Hum Genet. 2004;68:222–33.
16. Bosch E, Calafell F, Pérez-Lezaun A, Comas D, Mateu E, Bertranpetit J. Population history of North Africa: Evidence from classical genetic markers. Hum Biol. 1997;69:295–311.
17. Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J. High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet [Internet]. 2001;68:1019–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11254456
18. Arauna LR, Mendoza-Revilla J, Mas-Sandoval A, Izaabel H, Bekada A, Benhamamouch S, et al. Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa. Mol Biol Evol. 2017;34:318–29.
19. Mondal M, Bertranpetit J, Lao O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. Nat Commun [Internet]. 2019;10. Available from: http://dx.doi.org/10.1038/s41467-018-08089-7
20. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics. 2016;32:2817–23.
21. Lazaridis I, Belfer-Cohen A, Mallick S, Patterson N, Cheronet O, Rohland N, et al. Paleolithic DNA from the Caucasus reveals core of West Eurasian ancestry. bioRxiv. 2018;
22. Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, Kuderna LFK, Arauna LR, et al. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal

population of modern humans into sub-Saharan populations. Genome Biol. 2019;20:1–15.
23. Excoffier L, Estoup A, Cornuet JM. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics. 2005;169:1727–38.
24. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 2005;128:415–23.
25. Lipson M, Loh P-R, Sankararaman S, Patterson N, Berger B, Reich D. Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. PLOS Genet. 2015;11:e1005550.
26. Marchi N, Schlichta F, Excoffi Er L. Demographic inference. R276 Curr. Biol. 2021.
27. Clemente F, Unterländer M, Dolgova O, Amorim CEG, Coroado-Santos F, Neuenschwander S, et al. The genomic history of the Aegean palatial civilizations. Cell. 2021;184:2565-2586.e21.
28. Winsor MP, Gilmour SL, Gregor JW. Species, Demes, and the Omega Taxonomy: Gilmour and The New Systematics. Biol. Philos. 2000.
29. Gilmour JS., Gregor JW. Demes: A Suggested New Terminology. Nature. 1939;333.
30. Wright S. ISOLATION BY DISTANCE. Genetics [Internet]. 1943;28:114–38. Available from: https://academic.oup.com/genetics/article/28/2/114/6033172
31. Ragsdale AP, Weaver TD, Atkinson EG, Hoal EG, Möller M, Henn BM, et al. A weakly structured stem for human origins in Africa. Nature [Internet]. 2023;617:755–63. Available from: https://www.nature.com/articles/s41586-023-06055-y

32. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet. Elsevier Ltd; 2018. p. 301–12.
33. Mulazzani S, Belhouchet L, Salanova L, Aouadi N, Dridi Y, Eddargach W, et al. The emergence of the Neolithic in North Africa: A new model for the Eastern Maghreb. Quat Int [Internet]. 2016;410:123–43. Available from: https://linkinghub.elsevier.com/retrieve/pii/S104061821501215X
34. Korfmann K, Gaggiotti OE, Fumagalli M. Deep Learning in Population Genetics. Genome Biol. Evol. NLM (Medline); 2023.
35. Vikhar PA. Evolutionary algorithms: A critical review and its future prospects. Proc - Int Conf Glob Trends Signal Process Inf Comput Commun ICGTSPICC 2016. Institute of Electrical and Electronics Engineers Inc.; 2017. p. 261–5.
36. Sivanandam SN, Deepa · S N. Introduction to Genetic Algorithms. Berlin Heidelberg New York: Springer; 2008.
37. Koza JR. Genetically breeding populations of computer programs to solve problems in artificial intelligence. Dyn Genet Chaotic Program [Internet]. Stanford University, Department of Computer Science Stanford, CA; 1990. p. 819–27. Available from: http://www.genetic-programming.com/jkpdf/soucek1992.pdf
38. Koza JR. Genetic programming as a means for programming computers by natural selection. Stat Comput. 1994;4:87–112.
39. Hogue JT, Barton RNE. New radiocarbon dates for the earliest Later Stone Age microlithic technology in Northwest Africa. Quat Int. 2016;413:62–75.
40. Arauna LR, Hellenthal G, Comas D. Dissecting human North African gene-flow into its western coastal surroundings. Proc R Soc B Biol Sci. 2019;286.

41. Lapierre M, Lambert A, Achaz G. Accuracy of demographic inferences from the site frequency spectrum: The case of the yoruba population. Genetics. 2017;206:139–449.
42. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. Am J Hum Genet [Internet]. 2015;96:986–91. Available from: http://dx.doi.org/10.1016/j.ajhg.2015.04.019
43. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–6.
44. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
45. Fakhro KA, Staudt MR, Ramstetter MD, Robay A, Malek JA, Badii R, et al. The Qatar genome: A population-specific tool for precision medicine in the Middle East. Hum Genome Var. 2016;3:16016.
46. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
47. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10.
48. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
49. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008;40:1253–60.
50. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
52. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73.
53. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33:2776–8.
54. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
55. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust Demographic Inference from Genomic and SNP Data. PLoS Genet. 2013;9.
56. Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. Fastsimcoal2: Demographic inference under complex evolutionary scenarios. Bioinformatics. 2021;37:4882–5.
57. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife. 2018;7:1–21.
58. Csilléry K, François O, Blum MGB. Abc: An R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012;3:475–9.
59. 3.6.3 RDCT. A Language and Environment for Statistical Computing. R Found Stat Comput [Internet]. 2020 [cited 2023 Mar 15];https://www.R-project.org. Available from: https://www.r-project.org/

60. Kruschke JK. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition. 2nd ed. Doing Bayesian Data Anal. A Tutor. with R, JAGS, Stan, Second Ed. Elsevier Science; 2014.
61. Kullback S, Leibler RA. On Information and Sufficiency. Ann Math Stat. 1951;22:79–86.
62. Noskova E, Ulyantsev V, Koepfli KP, O'brien SJ, Dobrynin P. GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. Gigascience. 2020;9.
63. Koza JR, Keane MA, Streeter MJ, Mydlowec W, Yu J, Lanza G. Genetic Programming IV: Routine Human-Competitive Machine Intelligence. 2003.
64. Mehrabian AR, Lucas C. A novel numerical optimization algorithm inspired from weed colonization. Ecol Inform. 2006;1:355–66.
65. Misaghi M, Yaghoobi M. Improved invasive weed optimization algorithm (IWO) based on chaos theory for optimal design of PID controller. J Comput Des Eng. 2019;6:284–95.
66. Blum MGB. Approximate bayesian computation: A nonparametric perspective. J Am Stat Assoc. 2010;105:1178–87.
67. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192:1065–93.

4 DISCUSSION

4.1 What are the contributions of this PhD thesis?

This PhD thesis presents two main contributions to the knowledge on the population genetic field. On one hand, the work here presented expands our understanding on the demographic history of the underrepresented populations of Northern Africa by presenting the first comprehensive demographic model that tackles the genetic diversity of the region. On the other hand, as a part of this thesis we have developed a novel approach to demographic inference, Genetic Programming for Population Genetics (GP4GP), that applies genetic programming and evolutionary algorithms to automate the exploration of complex parameter-free demographic models.

4.1.1 North African demographic history

As stated in the introduction and the results, North African populations are a demographic melting pot. Since the Upper Palaeolithic, multiple groups have inhabited the region, possibly interacting with themselves and with populations from their surroundings. This interaction has been accompanied by extensive gene flow resulting in the presence of complex demographic histories, challenging our ability to reconstruct a demographic model for North Africa. In this thesis, I present the first attempt at defining a model that explains the genetic variation of the region and how the different evolutionary forces (mutation, drift and migration) have affected the demographic history of the region.

The most supported models we have tested back the idea of separate demographic histories for North Africa Arab and North African Amazigh groups. Both the results obtained with the ABC-DL method (Mondal et al., 2019) and the GP4PG method establish the origin of Amazigh

populations as an isolated group that originated in a back-to-Africa event between 18.14 and 22.3 thousand years ago, respectively. Both dates situate the presence of a North African autochthonous group, identified in previous studies as a Maghrebi autochthonous genetic component (Arauna-Rubio et al., 2017; Henn et al., 2012; Serra-Vidal et al., 2019), before the current oldest North African available genomes, the fossils extracted from the Taforalt cave in Morocco, dated around 15,000 years ago (Van De Loosdrecht et al., 2018). The Arab population in North Africa, on the other hand, shows higher genetic affinity with Middle Eastern groups. This affinity is represented in both methods by a demographic model that shows the North African Arab individuals as a sister clade of the Middle Eastern one with a recent split time between the two groups. This Middle Eastern-like genetic component has been hypothesized to originate as a consequence of several migration movements from the Middle East to North Africa with the Neolithization and the Arab expansion being the more relevant ones. The ABC-DL methodology supports a more ancient split of North African Arab and Middle Eastern followed by admixture waves with smaller impact during the Arabization. These results are in concordance with previous studies that point to a higher impact of the ancestral differentiation (Ammerman & Cavalli-Sforza, 1984; Pimenta et al., 2017; Sokal et al., 1991) rather than the hypothesis supported by the GP4PG analysis that backs the Arabization process as the main driver of the appearance of Middle Eastern genetic affinity in Northern Africa. The hypothesis that the Arab conquest has been the main driver of the observed differences between North African groups has gained momentum in recent years (5) and the results obtained by the GP4PG analysis increase the support on this idea that the impact of Arab migrations (specially the Bedouin expansion of the 11th century) have had a lasting genetic impact on current North

African populations. The discrepancy between the ABC-DL & the GP4PG results regarding the split times of the North African Arab population can be attributed to the effect of migration to coalescence. Migration occurs at a fastest time scale than coalescence, so in models with stronger migration, like the one obtained by the ABC-DL approach, the coalescence time between the Middle Eastern and the North African lineages would be further in the past than in model with weaker migration, like the one selected as the best by the GP4PG approach (6).

4.1.1.1 On the effective population size

The concept of effective population size (Ne) is central to population genetics and evolution as it quantifies the magnitude of genetic drift and inbreeding in real populations (Charlesworth, 2009; Wang et al., 2016). *N*e determines the amount and distribution of genetic variation in a population in interaction with several evolutionary forces like mutation, recombination, selection, and migration making it dependent on the demographic history of a population. So, the value of the effective population size is a good indicator of the demographic dynamics of a certain population.

The results of our thesis (*see Results 3.2*), indicate that North African Arab and Amazigh groups present different demographic dynamics as pointed out by the different effective population sizes we observe in both populations. For the Amazigh group, the effective population is maintained small, probably because of isolation of these autochthonous groups at different times since the Epipaleolithic. North Africa is a vast region, currently characterised by deserts and mountainous areas, an orography that is very suitable for the appearance of shelters that could have isolated populations protecting them from climate change and

cultural, political and military events, all of them very present in the history of North Africa. This has led to different isolation degrees in the population, which, as observed by the results of this thesis, have had a lasting impact on the genetic diversity of Imazighen. The Arabs, on the other hand, present an effective population size like that of surrounding populations like the European and Middle Easterns. Being that North African Arab groups present higher levels of admixture due to inhabiting more cosmopolitan areas than the isolated Imazighen, it is expected that diversity of the Arab population is higher resulting in a bigger effective population size.

4.1.2 Genetic Programming for Population Genetics

One of the major challenges of this thesis has been related to the replicability issues of the models. Although the ABC-DL presented robust results regarding the demographic inference for North Africa, when trying to replicate the observed diversity pattern with the simulations obtained by the "best" model we kept failing because of the Bayesian nature of the approach. As explained in the ABC section of the introduction *(see 1.2.2.4)*, the ABC defines the best suitable model/parameter for a given dataset by comparing the simulated models/parameters extracted from a prior distribution to the observed data, and applying a threshold to keep those simulations that are considered the "best". Although a more proper term would be to keep those simulations that are closer to the data. Meaning that, even if we are taking the "best" model from the possible pool of priors, this model could be extremely non-representative of the observed dataset. As we kept finding in multiple iterations of the ABC-DL algorithm that the model that ended up being the "best" was unable to properly represent the diversity of North Africa, probably due the

complex nature of the demographic scenario at hand. We decided that a new approach was needed, one that would detect the "best" possible model without the restrictions of defining prior demographic models that are simplifications of reality and may not explore the whole of the possible demographic events that could occur in a population. After this thought process the Genetic Programming for Population Genetics (GP4GP) is born.

GP4PG intends to explore this space of demographic models, reducing the inherent bias that human defined demographic models suppose. This Genetic Programming algorithm, it is extensively explained in this thesis *(see 1.2.3 & 3.2)*, gives us results that when compared to an observed dataset show higher replicability than the ABC-DL approach. On top of that, it also allowed us to address the substructure-ghost introgression issue, which is a hot topic on the demographic inference field (Mondal et al., 2019; Ragsdale et al., 2023). When defining demographic models for a population there is always some level of variation that we are not able to explain just with the samples we are using to test the models. In most cases, population substructure is the reason behind this variation, and not taking it into account can significantly bias the model ascertainment (Ragsdale et al., 2023). Previous studies have dealt with hidden population substructure by introducing "ghost" populations (Lazaridis et al., 2018; Loog, 2021; Lorente-Galdos et al., 2019; Mondal et al., 2019) or by generating ancient weak structured stems that account for the excess of diversity we observed in current day populations (Ragsdale et al., 2023). We proposed a different approach in the GP4PG framework. Here we define a population as a "ecodeme" and then simulate the population substructure by considering smaller "topodemes" that have higher

migration rates between themselves than with the demes of other populations on an isolation by distance basis.

4.2 Caveats, limitations, and possible biases

4.2.1 Data availability

Limitations in North Africa genetic studies start at the sampling step. Although recently there has been an effort to increase the representation of North Africa in the worldwide pool of whole genomes (Bergström et al., 2020; Pagani et al., 2015; Serra-Vidal et al., 2019), the amount of individuals and populations available is still scarce. Very few genetic studies include North African samples, with even a smaller number focusing on North Africa. Many of the studies are based on re-analyzing already available data rather than generating new samples. This situation presents a problem, especially in highly heterogeneous populations like North Africa, where the lack of genomic data of populations with different lifestyles and cultural backgrounds may lead to extracting biased conclusions. An extensive genetic panel of North Africa with enough representation of its different cultural and demographic groups could be a solution to explore in depth the genetic and biomedical patterns in the region.

4.2.2 Ancient North African genomes

The inclusion of ancient genomes when inferring the demographic history of a population is highly informative, allowing to fill gaps in the genetic history that otherwise would be impossible to. During the preparation of this thesis, I have thought several times about including the ancient Taforalt samples to the demographic models of North Africa. Although

methodologically feasible the inclusion of ancient genomes would incorporate another layer of complexity to the models as we had to consider contamination and DNA damage to the simulations. On top of that, and the main reason we discarded the possibility of including aDNA to the demographic inference is that there were no available ancient whole genomes until the ones from Simões et al (2023) (Simões et al., 2023) and the trying perform demographic inferences using current whole genomes and ancient enrich array panels (Van De Loosdrecht et al., 2018) increased the error probability of the simulations making it un-assumable to include ancient samples in the models (Clemente et al., 2021).

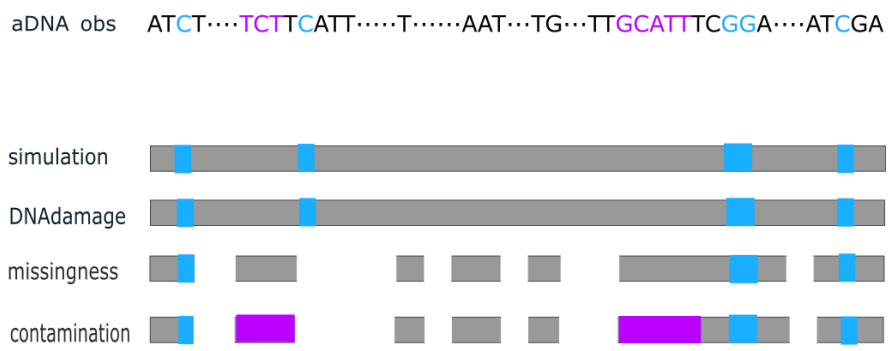


Figure 21: Diagram on the process a simulation must have to adapt to the characteristics of aDNA*. A lot of noise must be added to a simulation to resemble an aDNA molecule, including deamination processes, missingness and noise related to possible contamination from exogenous DNA.*

As explained above, there are very few ancient genome studies from North Africa (Fregel et al., 2018; Simões et al., 2023; Van De Loosdrecht et al., 2018) with nearly null exploration of historic ancient samples in North Africa (Schuenemann et al., 2017). Many conclusions of prehistoric North Africa are based on this very limited number of samples, which may not be representative of the people living in the region in their time. Also, due to climatic and political constraints most of the samples come

from the extremes of North Africa (Morocco and Egypt), with no data from the land in between. More aDNA data would provide a more comprehensive picture of prehistoric North Africa as well as more information on the current North African scenario. The historic aDNA situation is even worse, with basically no data available after Egyptian times. With data from different historical eras, we would be able to fill the gaps and learn about the genetic composition and the effect of different demographic events in the region.

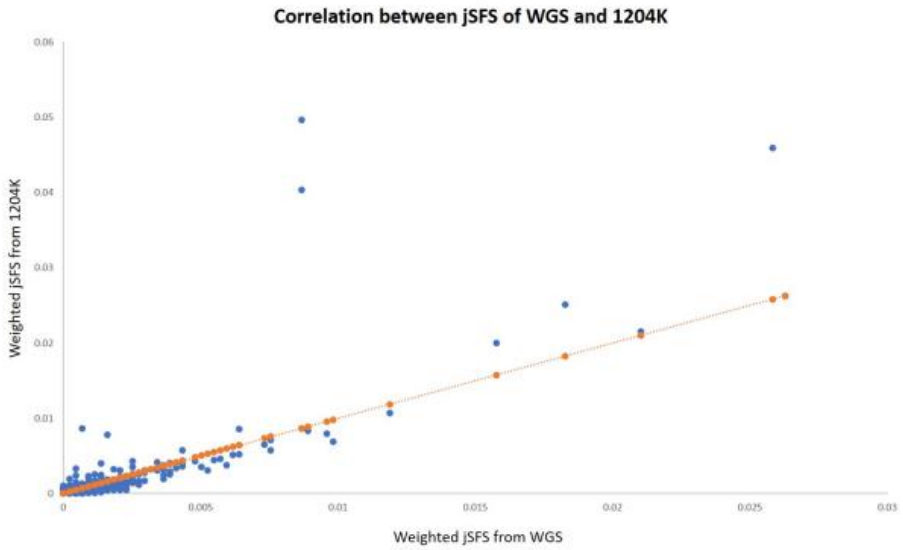


Figure 21: Correlation between observed jSFS of whole genome and the 1240K datasets. The orange line indicates complete correlation. *WGS and 1240K jSFS do not correlate so we cannot use data from both methodologies to compare SFS. Modified from Supplementary of Clemente 2022.*

4.2.3 Population definition in demographic inference

During this thesis we've widely made use of the term 'population' to describe the groups of individuals we have analysed. But defining what

constitutes a population is somewhat challenging and we must be treated carefully. Defining populations as non-interbreeding, phenotypically different individuals, like in interspecies studies is simple. However, in human studies this becomes more complex, as, although humans have the potential to breed with anyone around the world, they clearly do not form one panmictic population, but rather exhibit substructure. Defining populations is difficult and no criteria is absolute, but we tend to group individuals by geographical proximity, common language and shared ethnicity or cultural background (Jobling et al., 2004). In this thesis we analyze individuals that follow a criterion of four grandparents coming from the same geographical zone and cultural background to be identified as members of a population. Although it has been the most standard method of population identification it still can present problems, especially as the distances between birthplaces of parents increases, the number of individuals that meet these criteria will decrease in the next years increasing the difficulties on population determination in future studies.

In demographic inference methods the definition of the populations and the samples that form these populations is of extreme importance. Most population genetic analysis, where many individuals are used, assume that those populations are discrete and homogeneous without any subdivision or genetic substructure (Loog, 2021). However, this rarely happens, as individuals inside a population are often sampled from structured populations or from relatively broad geographical locations. Such heterogeneity inside a population do not impact the inference of the genetic affinity between individuals such as in a principal component analysis (PCA) or admixture approaches but can be a greater issue when

studying the demographic history leading to underestimations of TMRCA or overestimations of population size (Marchi et al., 2023).

In this thesis we apply an SFS-based demographic inference approach that deals with this problem by treating each single individual as a population and estimating the past demography of each population separately, with the help of hundreds of thousands of simulations. We have performed the demographic inference of the North African population using just two individuals for each Tunisian group (Tunisian Arabs & Tunisian Amazigh) as a proxy for the whole North Africa, one as a test and the second as a replica. Both the ABC-DL and the GP4PG frameworks only allow a limited number of sequences and populations in the tested models (Clemente et al., 2021; Lorente-Galdos et al., 2019; Mondal et al., 2019) and rely on the testing of many simulations to estimate the parameters in each model. This limitation on the population and number of individuals makes the definition of the proxies for each group extremely important.

The first population we define was the proxy population for the autochthonous North African group. The Tunisian Chenini population, due to their isolated nature (Fadhlaoui-Zid et al., 2011) shows the highest genetic affinity to the ancient Taforalt population making it an ideal proxy for the autochthonous populations of North Africa. After that, we defined the North African Arab population. To reduce the possible geographical differences between both North African groups, we took a Tunisian cosmopolitan population as the proxy for North African Arab. The results obtained in this thesis have this in mind, and we are aware that we have analysed two extremes of the whole diversity present in North Africa. As stated before, the demography of the region has been complex and the simple dichotomy of Arab and Amazigh populations does not reflect this complex pattern, since there is an extensive heterogeneity

within both groups. If other individuals had been used as proxies for both North African Arab and North African Amazigh the demographic histories that we are explaining could have been slightly different.

4.3 Concluding remarks and future research

4.3.1 Concluding remarks

This PhD thesis explores the understudied region of North Africa from an approach that has not been tested before, and for a reason. Inferring the demographic model of an extremely complex region, such as the North African, has been a struggle. Trying to identify which parameters better suit each model and the prior distributions for each of them has given me more than one headache. Other demographic inferences done with the same ABC-DL methodology infer as much as 45 parameters for the Aegean civilization demographic model (Clemente et al., 2021), 46 for the study of introgression in Asia and Oceania (Mondal et al., 2019) or 51 parameters in a single model of sub-Saharan Africa (Lorente-Galdos et al., 2019) while, the selected best model in our ABC-DL study infers up to 82 different parameters applying a level of complexity to demographic inference methods we have not seen until this thesis. This level of complexity was what drove us to develop this new methodology to infer demographic histories that do not rely on previous knowledge of the population as it explores that demographic space after each iteration. The GP4PG has proven to present more reliable and replicable results than the ABC-DL, at least for complex scenarios like the North African one. Results from this analysis, conclude that the genetic continuity in North Africa is far deeper in time than previously detected, with the autochthonous genetic component emerging during the Upper

Palaeolithic before any available ancient genome analysed in the region. Moreover, it also reflects the importance of isolation and its effect on the genetic diversity of a population. Finally, the GP4PG plants a seed for future exploration regarding the effect of punctual gene flow and how can it be masked by genetic substructure and soft splits, as the results obtained, especially for the North African Arab - Middle Eastern splits as soft split models show less error and more affinity to the observed data than models with discrete gene flow events.

4.3.2 Future research

In this last chapter I will present some points to further investigate regarding North African genetics:

a) *What is the extent of the genetic heterogeneity in North Africa as the number of sampled populations increases? How does it correlate with the cultural diversity and the geographical location of the groups?*

   As we dive deeper into the knowledge of the North African region, we realise that the complexity and genetic heterogeneity of the groups in the region is even greater than what we expected. In most studies, including the one of this theses, North African has been studied from an Arab-Amazigh dichotomic perspective. However, this is an oversimplification of the socio-cultural landscape. As we increase the number of samples and reach non-studied groups, we could obtain more robust answers that could present us with a more extensive understanding of the relationships between different Amazigh and Arab groups in North Africa.

b) *How are selection studies in North Africa affected with the presence of a null demographic model of the region?*

After this thesis, we now possess a null demographic model for North Africa. This model sets a basal genetic diversity for Arab and Amazigh populations and can be used in selection studies (Font-Porterias et al., 2021). Up until now, selection analyses in North Africa rely on a rather simplistic demographic model to assess the effect of natural selection and drift to the population. From now on, a more extensive model that considers the effect of surrounding populations is available to explore these effects as well as the effect of archaic introgression into the region.

c) *Can this model be applied in biomedical studies?*

Demographic models have already been used to detect risk variants for diseases. For example, a demographic model of Eurasian groups was used to describe the evolutionary trajectory of a risk variant (*TYK2* P1104A) for tuberculosis in Europeans (Kerner et al., 2021). So, the application of the North African demographic model on the detection of the evolutionary trajectories for endemic variants could help define population-based medicine for North Africa.

d) *Will this model need updates?*

Finally, further updates on these models should be taken into consideration when better archaic genomes are available in the region as including aDNA to the analysis may give more

robustness to some of the inferences that have been made in these models.

5 REFERENCES

Affymetrix, & Inc. (n.d.). *A. Select 1.35M of 1.81M candidate SNPs for validation*. http://www.cephb.fr/en/hgdp/

Alexi, A., Lazebnik, T., & Shami, L. (2023). Microfounded Tax Revenue Forecast Model with Heterogeneous Population and Genetic Algorithm Approach. *Computational Economics*. https://doi.org/10.1007/s10614-023-10379-2

Allen, R., & Fraser, A. (1968). Simulation of genetic systems. *Theoretical and Applied Genetics*, *38*(6), 223–225. https://doi.org/10.1007/BF01245621

Ammerman, A. J., & Cavalli-Sforza, L. L. (1984). The Neolithic Transition and the Genetics of Populations in Europe. In *Princeton: Univ. Press*. Princeton University Press. https://doi.org/10.2307/2803105

Anders, V. (2023). *Etímologia de Andalucia*. https://etimologias.dechile.net/?Andaluci.a

Arauna-Rubio, L. (2017). *Genetic structure of North African human populations*. Universitat Pompeu Fabra.

Arauna-Rubio, L., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhlaoui-Zid, K., Zalloua, P., Hellenthal, G., & Comas, D. (2017). Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa. *Molecular Biology and Evolution*, *34*(2), 318–329. https://doi.org/10.1093/molbev/msw218

Arauna, L. R., Hellenthal, G., & Comas, D. (2019). Dissecting human North African gene-flow into its western coastal surroundings. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1902). https://doi.org/10.1098/rspb.2019.0471

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *Journal of Experimental Medicine*, *79*(2), 137–158. https://doi.org/10.1084/jem.79.2.137

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers andSystem Designers*. ApressOpen.

Baharian, S., & Gravel, S. (2018). On the decidability of population size histories from finite allele frequency spectra. *Theoretical Population Biology*, *120*, 42–51. https://doi.org/10.1016/j.tpb.2017.12.008

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3).

https://doi.org/10.1093/genetics/iyab229
Bayes, T., & Price, R. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053
Beaumont, M. A. (2019). Approximate Bayesian Computation. *Annual Review of Statistics and Its Application* , *6*, 379–403. https://doi.org/10.1146/annurev-statistics-030718
Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, *162*, 2025–2035. https://academic.oup.com/genetics/article/162/4/2025/6050069
Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484). https://doi.org/10.1126/science.aay5012
Bergström, A., Stanton, D. W. G., Taron, U. H., Frantz, L., Sinding, M. H. S., Ersmark, E., Pfrengle, S., Cassatt-Johnstone, M., Lebrasseur, O., Girdland-Flink, L., Fernandes, D. M., Ollivier, M., Speidel, L., Gopalakrishnan, S., Westbury, M. V., Ramos-Madrigal, J., Feuerborn, T. R., Reiter, E., Gretzinger, J., … Skoglund, P. (2022). Grey wolf genomic history reveals a dual ancestry of dogs. *Nature*, *607*(7918), 313–320. https://doi.org/10.1038/s41586-022-04824-9
Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. In *Molecular Ecology* (Vol. 19, Issue 13, pp. 2609–2625). https://doi.org/10.1111/j.1365-294X.2010.04690.x
Bhaskar, A., Wang, Y. X. R., & Song, Y. S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, *25*(2), 268–279. https://doi.org/10.1101/gr.178756.114
Blum, M. G. B. (2010). Approximate bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, *105*(491), 1178–1187. https://doi.org/10.1198/jasa.2010.tm09448
Bosch, E., Calafell, F., Comas, D., Oefner, P. J., Underhill, P. A., & Bertranpetit, J. (2001). High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *American Journal of Human Genetics*, *68*(4), 1019–1029. https://doi.org/10.1086/319521

Botigué, L. R., Henn, B. M., Gravel, S., Maples, B. K., Gignoux, C. R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., Bertranpetit, J., Comasa, D., & Bustamante, C. D. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern europe. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11791–11796. https://doi.org/10.1073/pnas.1306223110

Brakez, Z., Bosch, E., Izaabel, H., Akhayat, O., Comas, D., Bertranpetit, J., & Calafell, F. (2001). Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Annals of Human Biology*, *28*(3), 295–307. https://doi.org/10.1080/030144601300119106

Brass, M. (2013). Revisiting a hoary chestnut : the nature of early cattle domestication in North-East Africa. *Sahara (Segrate)*, *1*(24), 65–70.

Browning, S. R., & Browning, B. L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *American Journal of Human Genetics*, *97*(3), 404–418. https://doi.org/10.1016/j.ajhg.2015.07.012

Camps, G. (1982). Beginnings of pastoralism and cultivation in north-west Africa and the Sahara: origins of the Berbers. In J. D. Clark (Ed.), *The Cambridge History of Africa* (pp. 548–623). Cambridge University Press. https://doi.org/https://doi.org/10.1017/CHOL9780521222150.009

Camps, G. (1998). *Los bereberes: de la orilla del mediterráneo al límite meridional del Sáhara*. Icaria.

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*(6099), 31–36. https://doi.org/10.1038/325031a0

Carl Correns. (1900). G. Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. *Berichte Der Deutschen Botanischen Gesellschaft*, *18*(4), 158–168.

Carlos A. Coello Coello. (2005). An Introduction to Evolutionary Algorithms and Their Applications. *Advanced Distributed Systems Lecture Notes in Computer Science Volume 3563*, 425–442.

Carmi, S., Palamara, P. F., Vacic, V., Lencz, T., Darvasi, A., & Pe'er, I. (2013). The variance of identity-by-descent sharing in the wright-fisher model. *Genetics*, *193*(3), 911–928. https://doi.org/10.1534/genetics.112.147215

Castañeda, I. S., Mulitza, S., Schefuß, E., Dos Santos, R. A. L., Damsté, J. S. S., & Schouten, S. (2009). Wet phases in the Sahara/Sahel region and human migration patterns in North Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20159–20163. https://doi.org/10.1073/pnas.0905771106

Cavalli-Sforza, L. L., & Piazza, A. (1993). Human genomic diversity in

Europe: a summary of recent research and prospects for the future. *European Journal of Human Genetics : EJHG*, *1*, 3–18. https://doi.org/10.1159/000472383
Charlesworth, B. (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. In *Nature Reviews Genetics* (Vol. 10, Issue 3, pp. 195–205). https://doi.org/10.1038/nrg2526
Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., & Fodor, S. P. A. (1996). Accessing Genetic Information with High-Density DNA Arrays. *Science*, *274*(5287), 610–614. https://doi.org/10.1126/science.274.5287.610
Clemente, F., Unterländer, M., Dolgova, O., Amorim, C. E. G., Coroado-Santos, F., Neuenschwander, S., Ganiatsou, E., Cruz Dávalos, D. I., Anchieri, L., Michaud, F., Winkelbach, L., Blöcher, J., Arizmendi Cárdenas, Y. O., Sousa da Mota, B., Kalliga, E., Souleles, A., Kontopoulos, I., Karamitrou-Mentessidi, G., Philaniotou, O., … Papageorgopoulou, C. (2021). The genomic history of the Aegean palatial civilizations. *Cell*, *184*(10), 2565-2586.e21. https://doi.org/10.1016/j.cell.2021.03.039
Coffman, J. M. (1992). Arabization and Islamization in the Algerian university. In *PhD Confluences Méditerranée* (Vol. 45).
Coulthard, T. J., Ramirez, J. A., Barton, N., Rogerson, M., & Brücher, T. (2013). Were rivers flowing across the Sahara during the last interglacial? Implications for human migration through Africa. *PloS One*, *8*(9), e74834. https://doi.org/10.1371/journal.pone.0074834
Creanza, N., & Feldman, M. W. (2016). Worldwide genetic and cultural change in human evolution. *Current Opinion in Genetics & Development*, *41*, 85–92. https://doi.org/10.1016/j.gde.2016.08.006
Cruciani, F., La Fratta, R., Torroni, A., Underhill, P. A., & Scozzari, R. (2006). Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. *Human Mutation*, *27*(8), 831–832. https://doi.org/10.1002/humu.9445
Csilléry, K., François, O., & Blum, M. G. B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x
D'Atanasio, E., Risi, F., Ravasini, F., Montinaro, F., Hajiesmaeil, M., Bonucci, B., Pistacchia, L., Amoako-Sakyi, D., Bonito, M., Onidi, S., Colombo, G., Semino, O., Destro Bisol, G., Anagnostou, P., Metspalu, M., Tambets, K., Trombetta, B., & Cruciani, F. (2023). The genomic echoes of the last Green Sahara on the Fulani and Sahelian people. *Current Biology*. https://doi.org/10.1016/j.cub.2023.10.075

Dachy, T., Guéret, C., Green, W., & Perrin, T. (2023). Rethinking the Capsian: Lithic Variability Among Holocene Maghreb Hunter-Gatherers. *African Archaeological Review*, *40*(1), 169–203. https://doi.org/10.1007/s10437-023-09514-z

De Vries, H. (1900). Sur la loi de disjonction des Hybrides. *Comptes Rendus de l'Académie Des Sciences*, *130*, 845–847.

Dempster, A. P., Laird, ; N M, & Rubin, ; D B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 39, Issue 1).

Devroye, L., & Lugosi, G. (2001). *Minimax Theory* (pp. 150–176). https://doi.org/10.1007/978-1-4613-0125-7_15

Drake, N. A., Blench, R. M., Armitage, S. J., Bristow, C. S., & White, K. H. (2011). Ancient watercourses and biogeography of the Sahara explain the peopling of the desert. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(2), 458–462. https://doi.org/10.1073/pnas.1012231108

Eller, E. (2009). Effects of Ascertainment Bias on Recovering Human Demographic History. *Human Biology*, *81*(5–6), 735–751. https://doi.org/10.3378/027.081.0618

Ennafaa, H., Cabrera, V. M., Abu-Amero, K. K., González, A. M., Amor, M. B., Bouhaha, R., Dzimiri, N., Elgaaïed, A. B., & Larruga, J. M. (2009). Mitochondrial DNA haplogroup H structure in North Africa. *BMC Genetics*, *10*(1), 8. https://doi.org/10.1186/1471-2156-10-8

Ewens, W. J. (1974). A Note on the Sampling Theory for Infinite Alleles and Infinite Sites Models. In *POPULATION BIOLOGY* (Vol. 6).

Ewens, W. J. (2000). *Mathematical Population Genetics. 1. Theoretical Introduction*. Springer Verlag.

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). https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). Fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Fadhlaoui-Zid, K., Haber, M., Martínez-Cruz, B., Zalloua, P., Elgaaied, A. B., & Comas, D. (2013). Genome-wide and paternal diversity reveal a recent origin of human populations in north africa. *PLoS ONE*, *8*(11), e80293. https://doi.org/10.1371/journal.pone.0080293

Fadhlaoui-Zid, K., Khodjet-el-khil, H., Mendizabal, I., Benammar-elgaaied, A., & Comas, D. (2011). *Genetic Structure of Tunisian Ethnic Groups Revealed by Paternal Lineages*. *280*(August), 271–280. https://doi.org/10.1002/ajpa.21581

Fadhlaoui-Zid, K., Plaza, S., Calafell, F., Amor, M. Ben, Comas, D., & El Gaaied, A. B. (2004). Mitochondrial DNA heterogeneity in Tunisian Berbers. *Annals of Human Genetics*, *68*(3), 222–233. https://doi.org/10.1046/j.1529-8817.2004.00096.x

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009

Fisher, R. A. (1923). XXI.—On the Dominance Ratio. *Proceedings of the Royal Society of Edinburgh*, *42*, 321–341. https://doi.org/10.1017/S0370164600023993

Font-Porterias, N., Caro-Consuegra, R., Lucas-Sánchez, M., Lopez, M., Giménez, A., Carballo-Mesa, A., Bosch, E., Calafell, F., Quintana-Murci, L., & Comas, D. (2021). The Counteracting Effects of Demography on Functional Genomic Variation: The Roma Paradigm. *Molecular Biology and Evolution*, *38*(7), 2804–2817. https://doi.org/10.1093/molbev/msab070

Font-Porterias, N., Solé-Morata, N., Serra-Vidal, G., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., Calafell, F., & Comas, D. (2018). The genetic landscape of Mediterranean North African populations through complete mtDNA sequences. *Annals of Human Biology*, *45*(1), 98–104. https://doi.org/10.1080/03014460.2017.1413133

Foote, A. D., Martin, M. D., Louis, M., Pacheco, G., Robertson, K. M., Sinding, M. H. S., Amaral, A. R., Baird, R. W., Baker, C. S., Ballance, L., Barlow, J., Brownlow, A., Collins, T., Constantine, R., Dabin, W., Dalla Rosa, L., Davison, N. J., Durban, J. W., Esteban, R., … Morin, P. A. (2019). Killer whale genomes reveal a complex history of recurrent admixture and vicariance. *Molecular Ecology*, *28*(14), 3427–3444. https://doi.org/10.1111/mec.15099

Fregel, R., Gomes, V., Gusmão, L., González, A. M., Cabrera, V. M., Amorim, A., & Larruga, J. M. (2009). Demographic history of Canary Islands male gene-pool: Replacement of native lineages by European. *BMC Evolutionary Biology*, *9*(1), 181. https://doi.org/10.1186/1471-2148-9-181

Fregel, R., Méndez, F. L., Bokbot, Y., Martín-Socas, D., Camalich-Massieu, M. D., Santana, J., Morales, J., Avila-Arcos, M. C., Underhill, P. A., Shapiro, B., Wojcik, G., Rasmussen, M., Soares, A. E. R., Kapp, J., Sockell, A., Rodríguez-Santos, F. J., Mikdad, A., Trujillo-Mederos, A., & Bustamante, C. D. (2018). Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(26), 6774–6779. https://doi.org/10.1073/pnas.1800851115

Fu, Y.-X., & Li, W.-H. (1997). Estimating the Age of the Common Ancestor of a Sample of DNA Sequences. *Molecular Biology and Evolution*, *114*, 195–199. https://academic.oup.com/mbe/article/14/2/195/1083938

Garfinkel, D., Arthur, R. H. Mac, & Sack, R. (1964). COMPUTER SIMULATION AND ANALYSIS OF SIMPLE ECOLOGICAL SYSTEMS*. *Annals of the New York Academy of Sciences*, *115*(2), 943–951. https://doi.org/10.1111/j.1749-6632.1964.tb00068.x

Geraads, D. (2010). Biogeographic relationships of Pliocene and Pleistocene North-western African mammals. *Quaternary International*, *212*(2), 159–168. https://doi.org/10.1016/j.quaint.2009.06.002

Gibbons, A. (2017). World's oldest Homo sapiens fossils found in Morocco. *Science*. https://doi.org/10.1126/science.aan6934

Godfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

González, A. M., Larruga, J. M., Abu-Amero, K. K., Shi, Y., Pestano, J., & Cabrera, V. M. (2007). Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics*, *8*(1), 223. https://doi.org/10.1186/1471-2164-8-223

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, *191*(2), 607–619. https://doi.org/10.1534/genetics.112.139808

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., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., … McVean, G. A. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, *108*(29), 11983–11988. https://doi.org/10.1073/pnas.1019276108

Griffiths, R., & Marjoram, P. (1996). Ancestral Inference from Samples of DNA Sequences with Recombination. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *3*, 479–502. https://doi.org/10.1089/cmb.1996.3.479

Gritzner, J. A., & Gritzner, C. F. (2006). *North Africa and the Middle East*. Chelsea House Publishers.

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., & Pe'Er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, *19*(2), 318–326. https://doi.org/10.1101/gr.081398.108

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10).

https://doi.org/10.1371/journal.pgen.1000695
Hahn, M. W. (2018). *Molecular Population Genetics* (Sinauer Associates (ed.); 1st ed.). Oxford University Press.
Haldane, J. B. S. (1927). A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *23*(7), 838–844. https://doi.org/10.1017/S0305004100015644
Haller, B. C., & Messer, P. W. (2023). SLiM 4: Multispecies Eco-Evolutionary Modeling. *The American Naturalist*, *201*(5), E127–E139. https://doi.org/10.1086/723601
Hardy, G. . (1908). Mendelian proportions in a mixed population, . *Science*, *XXVIII*, 49–50.
Harris, H. (1966). C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *164*(995), 298–310. https://doi.org/10.1098/rspb.1966.0032
Harris, K., & Nielsen, R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, *9*(6). https://doi.org/10.1371/journal.pgen.1003521
Henn, B. M., Botigué, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., Zalloua, P. A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C. D., & Comas, D. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genetics*, *8*(1), e1002397. https://doi.org/10.1371/journal.pgen.1002397
Hervella, M., Svensson, E. M., Alberdi, A., Günther, T., Izagirre, N., Munters, A. R., Alonso, S., Ioana, M., Ridiche, F., Soficaru, A., Jakobsson, M., Netea, M. G., & De-La-Rua, C. (2016). The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa. *Scientific Reports*, *6*. https://doi.org/10.1038/srep25501
Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., & Cox, D. R. (2005). Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science*, *307*(5712), 1072–1079. https://doi.org/10.1126/science.1105436
Holland, J. . (1962a). Concerning efficient adaptive systems. In M. . Yovits, G. . Jacobi, & G. . Goldstein (Eds.), *Self-Organizing Systems* (pp. 215–230). Spartan Books.
Holland, J. H. (1962b). Outline for a logical theory of adaptive systems. *Journal of the Association for Computing Machinery*, *9*, 297–314.
Hollfelder, N., Breton, G., Sjödin, P., & Jakobsson, M. (2021). The deep population history in Africa. In *Human Molecular Genetics* (Vol. 30, Issue R1, pp. R2–R10). https://doi.org/10.1093/hmg/ddab005
Huang, X., Rymbekova, A., Dolgova, O., Lao, O., & Kuhlwilm, M.

(2023). Harnessing deep learning for population genetic inference. In *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-023-00636-3

Hubby, J. L., & Lewontin, R. C. (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*, *54*(2), 577–594. https://doi.org/10.1093/genetics/54.2.577

Hublin, J. J., Verna, C., Bailey, S., Smith, T., Olejniczak, A., Sbihi-Alaoui, F. Z., & Zouak, M. (2012). Dental Evidence from the Aterian Human Populations of Morocco. In J.-J. Hublin & S. P. McPherron (Eds.), *Modern Origins: A North African Perspective* (Issue 9789400729285, pp. 189–204). Springer. https://doi.org/10.1007/978-94-007-2929-2_13

Hublin, Jean Jacques, Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., & Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, *546*(7657), 289–292. https://doi.org/10.1038/nature22336

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, *23*(2), 183–201. https://doi.org/https://doi.org/10.1016/0040-5809(83)90013-8

International Human Genome Sequencing Consortium, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

IRCAM. (2023). *Alphabet Tifinaghe*. https://www.ircam.ma/fr/alphabet-tifinaghe

Irish, J. D. (2000). The Iberomaurusian enigma: North African progenitor or dead end? *Journal of Human Evolution*, *39*(4), 393–410. https://doi.org/10.1006/jhev.2000.0430

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K. H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

Jennings, H. S. (1914). Formulae for the results of inbreeding. *The American Naturalist*, *48*, 693–696.

Jennings, H. S. (1916). The numerical results of diverse systems of

breeding. *Genetics*, *1*, 53–89.
Jobling, M. A., Hurles, M., & Tyler-Smith, C. (2004). *Human Evolutionary Genetics: Origins, Peoples & Disease*. Garland Science Publishing.
Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, *7*(1). https://doi.org/10.2202/1544-6115.1389
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
Kaplan, N. L., Darden, T., & Hudson', R. R. (1988). *The Coalescent Process in Models With Selection*.
Kefi, R., Hechmi, M., Naouali, C., Jmel, H., Hsouna, S., Bouzaid, E., Abdelhak, S., Beraud-Colomb, E., & Stevanovitch, A. (2018). On the origin of Iberomaurusians: new data based on ancient mitochondrial DNA and phylogenetic analysis of Afalou and Taforalt populations. *Mitochondrial DNA Part A: DNA Mapping, Sequencing, and Analysis*, *29*(1), 147–157. https://doi.org/10.1080/24701394.2016.1258406
Kerner, G., Laval, G., Patin, E., Boisson-Dupuis, S., Abel, L., Casanova, J. L., & Quintana-Murci, L. (2021). Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans over the last 2,000 years. *American Journal of Human Genetics*, *108*(3), 517–524. https://doi.org/10.1016/j.ajhg.2021.02.009
Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, *1*(2), 177–232. https://doi.org/10.2307/3211856
Kimura, M., & Crow, J. F. (1964). THE NUMBER OF ALLELES THAT CAN BE MAINTAINED IN A FINITE POPULATION. *Genetics*, *49*(4), 725–738. https://doi.org/10.1093/genetics/49.4.725
Kimura, M., & Ohta, T. (1974). *Probability of Gene Fixation in an Expanding Finite Population (population genetics/diffusion model/fixation probability of mutant/logistic population)* (Vol. 71, Issue 9). https://www.pnas.org
Kingman, J. F. C. (1982). On the Genealogy of Large Populations. In *Source: Journal of Applied Probability* (Vol. 19).
Klein, R. G. (2000). Archeology and the evolution of human behavior. *Evolutionary Anthropology: Issues, News, and Reviews*, *9*(1), 17–36. https://doi.org/10.1002/(SICI)1520-6505(2000)9:1<17::AID-EVAN3>3.0.CO;2-A

Koza, J R, Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., & Lanza, G. (2003). *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. https://doi.org/10.1007/b137549

Koza, John R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, *4*(2), 87–112. https://doi.org/10.1007/BF00175355

Koza, John R. (1990). Genetically breeding populations of computer programs to solve problems in artificial intelligence. *Dynamic, Genetic, and Chaotic Programming*, *34*(June), 819–827. https://doi.org/10.1109/tai.1990.130444

Kuhlwilm, M., Han, S., Sousa, V. C., Excoffier, L., & Marques-Bonet, T. (2019). Ancient admixture from an extinct ape lineage into bonobos. *Nature Ecology and Evolution*, *3*(6), 957–965. https://doi.org/10.1038/s41559-019-0881-7

Larrasoaña, J. C., Roberts, A. P., & Rohling, E. J. (2013). Dynamics of Green Sahara Periods and Their Role in Hominin Evolution. *PLoS ONE*, *8*(10), e76514. https://doi.org/10.1371/journal.pone.0076514

Lazaridis, I., Belfer-Cohen, A., Mallick, S., Patterson, N., Cheronet, O., Rohland, N., Bar-Oz, G., Bar-Yosef, O., Jakeli, N., Kvavadze, E., Lordkipanidze, D., Matzkevich, Z., Meshveliani, T., Culleton, B. J., Kennett, D. J., Pinhasi, R., & Reich, D. (2018). Paleolithic DNA from the Caucasus reveals core of West Eurasian ancestry. *BioRxiv*. https://doi.org/10.1101/423079

Lewontin, R. C. (1972). The Apportionment of Human Diversity. In *Evolutionary Biology* (pp. 381–398). Springer US. https://doi.org/10.1007/978-1-4684-9063-3_14

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

Linstädter, J. (2013). Climate Induced Mobility and the Missing Middle Neolithic of Morocco. *Palaeoenvironment and the Development of Early Settlements*, *2016*, 63–80.

Linstädter, J., Medved, I., Solich, M., & Weniger, G. C. (2012). Neolithisation process within the Alboran territory: Models and possible African impact. *Quaternary International*, *274*, 219–232. https://doi.org/10.1016/j.quaint.2012.01.013

Litt, M., & Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*, *44*(3), 397–401.

Liu, X., & Fu, Y. X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*(5), 555–559. https://doi.org/10.1038/ng.3254

Loog, L. (2021). Sometimes hidden but always there: The assumptions

underlying genetic inference of demographic histories: Demographic inference from genetic DNA. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 376, Issue 1816, p. 20190719). https://doi.org/10.1098/rstb.2019.0719

Lorente-Galdos, B., Lao, O., Serra-Vidal, G., Santpere, G., Kuderna, L. F. K., Arauna, L. R., Fadhlaoui-Zid, K., Pimenoff, V. N., Soodyall, H., Zalloua, P., Marques-Bonet, T., & Comas, D. (2019). Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. *Genome Biology*, *20*(1), 1–15. https://doi.org/10.1186/s13059-019-1684-5

Lucas-Sánchez, M. (2023). *Human Population Genetics of North Africa: insights into demography and functional variation*. Universitat Pompeu Fabra.

Lucas-Sánchez, M., Fadhlaoui-Zid, K., & Comas, D. (2023). The genomic analysis of current-day North African populations reveals the existence of trans-Saharan migrations with different origins and dates. *Human Genetics*, *142*(2), 305–320. https://doi.org/10.1007/s00439-022-02503-3

Lucas-Sánchez, M., Serradell, J. M., & Comas, D. (2021). Population history of North Africa based on modern and ancient genomes. *Human Molecular Genetics*, *30*(R1), R17–R23. https://doi.org/10.1093/hmg/ddaa261

Maca-Meyer, N., Arnay, M., Rando, J. C., Flores, C., González, A. M., Cabrera, V. M., & Larruga, J. M. (2004). Ancient mtDNA analysis and the origin of the Guanches. *European Journal of Human Genetics*, *12*(2), 155–162. https://doi.org/10.1038/sj.ejhg.5201075

Maca-Meyer, N., González, A. M., Pestano, J., Flores, C., Larruga, J. M., & Cabrera, V. M. (2003). Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. *BMC Genetics*, *4*(15). https://doi.org/10.1186/1471-2156-4-15

Maceda, I., & Lao, O. (2021). Analysis of the Batch Effect Due to Sequencing Center in Population Statistics Quantifying Rare Events in the 1000 Genomes Project. *Genes*, *13*(1), 44. https://doi.org/10.3390/genes13010044

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. In *Nature* (Vol. 538, Issue 7624, pp. 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

Marchi, N., Kapopoulou, A., & Excoffier, L. (2023). Demogenomic inference from spatially and temporally heterogeneous samples. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13877

Marchi, N., Winkelbach, L., Schulz, I., Brami, M., Hofmanová, Z., Blöcher, J., Reyna-Blanco, C. S., Diekmann, Y., Thiéry, A., Kapopoulou, A., Link, V., Piuz, V., Kreutzer, S., Figarska, S. M., Ganiatsou, E., Pukaj, A., Struck, T. J., Gutenkunst, R. N., Karul, N., … Excoffier, L. (2022). The genomic origins of the world's first farmers. *Cell*, *185*(11), 1842-1859.e18. https://doi.org/10.1016/j.cell.2022.04.008

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. In *PNAS December* (Vol. 23, Issue 26). www.pnas.orgcgidoi10.1073pnas.0306899100

Mather, N., Traves, S. M., & Ho, S. Y. W. (2020). A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, *10*(1), 579–589. https://doi.org/10.1002/ece3.5888

McEvedy, C. (1995). *The Penguin Atlas of African History*. Penguin Books.

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

Mendel, G. (1886). Versuche über Pflanzen-Hybriden. In *Verhandlungen des Naturforschenden Vereines in Brünn* (Vol. 4, pp. 3–47).

Meng, X.-L., & Rubin, D. B. (1993). *Maximum Likelihood Estimation via the ECM Algorithm: A General Framework* (Vol. 80, Issue 2). https://www.jstor.org/stable/2337198

Möhle, M. (1998). Robustness Results for the Coalescent. *Journal of Applied Probability*, *35*, 438–447.

Mondal, M., Bertranpetit, J., & Lao, O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08089-7

Moots, H. M., Antonio, M., Sawyer, S., Spence, J. P., Oberreiter, V., Weiß, C. L., Lucci, M., Cherifi, Y. M. S., La Pastina, F., Genchi, F., Praxmeier, E., Zagorc, B., Cheronet, O., Özdoğan, K. T., Demetz, L., Amrani, S., Candilio, F., De Angelis, D., Gasperetti, G., … Pinhasi, R. (2023). A genetic history of continuity and mobility in the Iron Age central Mediterranean. *Nature Ecology*

and Evolution, *7*(9), 1515–1524. https://doi.org/10.1038/s41559-023-02143-4

Morales, J., Pérez-Jordà, G., Peña-Chocarro, L., Zapata, L., Ruíz-Alonso, M., López-Sáez, J. A., & Linstädter, J. (2013). The origins of agriculture in North-West Africa: Macro-botanical remains from Epipalaeolithic and Early Neolithic levels of Ifri Oudadane (Morocco). *Journal of Archaeological Science*, *40*(6), 2659–2669. https://doi.org/10.1016/j.jas.2013.01.026

Morgan, T., Sturtevant, A., & Bridges, C. (1922). *Year Book Carnegie Inst Wash* (Issue 22, pp. 283–287).

Mulazzani, S., Belhouchet, L., Salanova, L., Aouadi, N., Dridi, Y., Eddargach, W., Morales, J., Tombret, O., Zazzo, A., & Zoughlami, J. (2016). The emergence of the Neolithic in North Africa: A new model for the Eastern Maghreb. *Quaternary International*, *410*, 123–143. https://doi.org/10.1016/j.quaint.2015.11.089

Myers, S., Fefferman, C., & Patterson, N. (2008). Can one learn history from the allelic spectrum? *Theoretical Population Biology*, *73*(3), 342–348. https://doi.org/10.1016/j.tpb.2008.01.001

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. *Molecular Ecology*, *25*(5), 1058–1072. https://doi.org/10.1111/mec.13540

Naseri, A., Liu, X., Tang, K., Zhang, S., & Zhi, D. (2019). RaPID: Ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1754-8

National Human Genome Reasearch Institute. (2022, August 24). *Human Genome Project*. https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

Naylor, P. C. (2009). *North Africa: A History from Antiquity to the Present*. Univeristy of Texas Press.

Neuhauser, C., & Kronet, S. M. (1997). *The Genealogy of Samples in Models With Selection*.

Newman, J. L. (1997). The Peopling of Africa: A Geographic Interpretation. *The International Journal of African Historical Studies*, *29*(3), 619. https://doi.org/10.2307/221381

Newman, T. L., Rieder, M. J., Morrison, V. A., Sharp, A. J., Smith, J. D., Sprague, L. J., Kaul, R., Carlson, C. S., Olson, M. V., Nickerson, D. A., & Eichler, E. E. (2006). High-throughput genotyping of intermediate-size structural variation. *Human Molecular Genetics*, *15*(7), 1159–1167. https://doi.org/10.1093/hmg/ddl031

Nordborg, M. (2007). Coalescent Theory. In Balding D. J., M. Bishop, & C. Cannings (Eds.), *Handbook of Statistical Genetics* (3rd ed.,

Vol. 2, pp. 843–877). Wiley.
Nordborg, Magnus, & Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, *18*(2), 83–90. https://doi.org/https://doi.org/10.1016/S0168-9525(02)02557-X
Nunes, M. A., & Balding, D. J. (2010). On Optimal Selection of Summary Statistics for Approximate Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology*, *9*(1). https://doi.org/https://doi.org/10.2202/1544-6115.1576
Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D. M., Dugoujon, J. M., Coudray, C., Santachiara-Benerecetti, A. S., Semino, O., Bandelt, H. J., & Torroni, A. (2006). The mtDNA legacy of the levantine early Upper Palaeolithic in Africa. *Science*, *314*(5806), 1767–1770. https://doi.org/10.1126/science.1135566
Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A. S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., … Willerslev, E. (2013). Recalibrating equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, *499*(7456), 74–78. https://doi.org/10.1038/nature12323
Osborne, A. H., Vance, D., Rohling, E. J., Barton, N., Rogerson, M., & Fello, N. (2008). A humid corridor across the Sahara for the migration of early modern humans out of Africa 120,000 years ago. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(43), 16444–16447. https://doi.org/10.1073/pnas.0804472105
Pääbo, S. (1989). Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, *86*(6), 1939–1943. https://doi.org/10.1073/pnas.86.6.1939
Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., Oljira, T., Mekonnen, E., Luiselli, D., Bradman, N., Bekele, E., Zalloua, P., Durbin, R., Kivisild, T., & Tyler-Smith, C. (2015). Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *American Journal of Human Genetics*, *96*(6), 986–991. https://doi.org/10.1016/j.ajhg.2015.04.019
Palamara, P. F., & Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, *29*(13). https://doi.org/10.1093/bioinformatics/btt239
Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., Omrak, A., Vartanyan, S., Poinar, H., Götherström, A., Reich, D., & Dalén, L. (2015). Complete genomes reveal signatures of

demographic and genetic declines in the woolly mammoth. *Current Biology*, *25*(10), 1395–1400. https://doi.org/10.1016/j.cub.2015.04.007

Pawar, H., Rymbekova, A., Cuadros-Espinoza, S., Huang, X., de Manuel, M., van der Valk, T., Lobon, I., Alvarez-Estape, M., Haber, M., Dolgova, O., Han, S., Esteller-Cucala, P., Juan, D., Ayub, Q., Bautista, R., Kelley, J. L., Cornejo, O. E., Lao, O., Andrés, A. M., … Kuhlwilm, M. (2023). Ghost admixture in eastern gorillas. *Nature Ecology and Evolution*, *7*(9), 1503–1514. https://doi.org/10.1038/s41559-023-02145-2

Pedersen, C.-E. T., Lohmueller, K. E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H. R., Moltke, I., & Albrechtsen, A. (2017). The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. *Genetics*, *205*(2), 787–801. https://doi.org/10.1534/genetics.116.193821

Pennarun, E., Kivisild, T., Metspalu, E., Metspalu, M., Reisberg, T., Moisan, J. P., Behar, D. M., Jones, S. C., & Villems, R. (2012). Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. *BMC Evolutionary Biology*, *12*(1). https://doi.org/10.1186/1471-2148-12-234

Perrin, T., Dachy, T., López-Montalvo, E., Manen, C., & Marchand, G. (2022). What relations between North Africa and Europe in the early holocene? *Tabona: Revista de Prehistoria y Arqueología*, *22*, 261–281. https://doi.org/10.25145/j.tabona.2022.22.13

Pervez, M. T., Hasnain, M. J. ul, Abbas, S. H., Moustafa, M. F., Aslam, N., & Shah, S. S. M. (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *BioMed Research International*, *2022*, 1–12. https://doi.org/10.1155/2022/3457806

Peter, B. M., Wegmann, D., & Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology*, *19*(21), 4648–4660. https://doi.org/10.1111/j.1365-294X.2010.04783.x

Pimenta, J., Lopes, A. M., Comas, D., Amorim, A., & Arenas, M. (2017). Evaluating the neolithic expansion at both shores of the mediterranean sea. *Molecular Biology and Evolution*, *34*(12), 3232–3242. https://doi.org/10.1093/molbev/msx256

Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., & Comas, D. (2003). Joining the pillars of hercules: mtDNA sequences show multidirectional gene flow in the Western Mediterranean. *Annals of Human Genetics*, *67*(4), 312–328. https://doi.org/10.1046/j.1469-1809.2003.00039.x

Punnet, R. C. (1915). *Mimicry in Butterflies.* Cambridge University Press.

Ragsdale, A. P., Weaver, T. D., Atkinson, E. G., Hoal, E. G., Möller,

M., Henn, B. M., & Gravel, S. (2023). A weakly structured stem for human origins in Africa. *Nature*, *617*(7962), 755–763. https://doi.org/10.1038/s41586-023-06055-y

Rahmani, N. (2004). Technological and cultural change among the last hunter-gatherers of the Maghreb: The Capsian (10,000-6000 B.P.). In *Journal of World Prehistory* (Vol. 18, Issue 1, pp. 57–105). https://doi.org/10.1023/B:JOWO.0000038658.50738.eb

Rajwar, K., Deep, K., & Das, S. (2023). An exhaustive review of the metaheuristic algorithms for search and optimization: taxonomy, applications, and open challenges. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-023-10470-y

Reid, N., Varin, C., & Firth, D. (2011). An Overview of Composite Likelihood Methods. In *Statistica Sinica* (Vol. 21). https://www.researchgate.net/publication/228634405

Relethford, J. H., & Harding, R. M. (2001). Population Genetics of Modern Human Evolution. In *eLS*. Wiley. https://doi.org/10.1038/npg.els.0001470

Richter, D., Grün, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rué, M., Fernandes, P., Raynal, J. P., Geraads, D., Ben-Ncer, A., Hublin, J. J., & McPherron, S. P. (2017). The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature*, *546*(7657), 293–296. https://doi.org/10.1038/nature22335

Richter, D., Moser, J., Nami, M., Eiwanger, J., & Mikdad, A. (2010). New chronometric data from Ifri n'Ammar (Morocco) and the chronostratigraphy of the Middle Palaeolithic in the Western Maghreb. *Journal of Human Evolution*, *59*(6), 672–679. https://doi.org/10.1016/j.jhevol.2010.07.024

Robbins, R. B. (1918). SOME APPLICATIONS OF MATHEMATICS TO BREEDING PROBLEMS III. *Genetics*, *3*, 375–389.

Rodríguez-Varela, R., Günther, T., Krzewińska, M., Storå, J., Gillingwater, T. H., MacCallum, M., Arsuaga, J. L., Dobney, K., Valdiosera, C., Jakobsson, M., Götherström, A., & Girdland-Flink, L. (2017). Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. *Current Biology*, *27*(21), 3396-3402.e5. https://doi.org/10.1016/j.cub.2017.09.059

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Sahnouni, M., Parés, J. M., Duval, M., Cáceres, I., Harichane, Z., Van Der Made, J., Pérez-González, A., Abdessadok, S., Kandi, N., Derradji, A., Medig, M., Boulaghraif, K., & Semaw, S. (2018). 1.9-million- and 2.4-million-year-old artifacts and stone tool–cutmarked bones from ain boucherit, Algeria. *Science*, *362*(6420), 1297–1301. https://doi.org/10.1126/science.aau0008originally

Salgotra, R., Gandomi, M., & Gandomi, A. H. (2020). Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. *Chaos, Solitons and Fractals*, *138*. https://doi.org/10.1016/j.chaos.2020.109945

Salmasi, F., Sattari, M. T., & Nurcheshmeh, M. (2021). Genetic Programming Approach for Estimating Energy Dissipation of Flow over Cascade Spillways. *Iranian Journal of Science and Technology - Transactions of Civil Engineering*, *45*(1), 443–455. https://doi.org/10.1007/s40996-020-00541-3

Sánchez-Quinto, F., Botigué, L. R., Civit, S., Arenas, C., Ávila-Arcos, M. C., Bustamante, C. D., Comas, D., & Lalueza-Fox, C. (2012). North African Populations Carry the Signature of Admixture with Neandertals. *PLoS ONE*, *7*(10), e47765. https://doi.org/10.1371/journal.pone.0047765

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174)* (Vol. 74, Issue 12).

Sarhani, M., Voß, S., & Jovanovic, R. (2022). Initialization of metaheuristics: comprehensive review, critical analysis, and research directions. In *International Transactions in Operational Research*. John Wiley and Sons Inc. https://doi.org/10.1111/itor.13237

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

Schuenemann, V. J., Peltzer, A., Welte, B., Van Pelt, W. P., Molak, M., Wang, C. C., Furtwängler, A., Urban, C., Reiter, E., Nieselt, K., Teßmann, B., Francken, M., Harvati, K., Haak, W., Schiffels, S., & Krause, J. (2017). Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nature Communications*, *8*(15694). https://doi.org/10.1038/ncomms15694

Secher, B., Fregel, R., Larruga, J. M., Cabrera, V. M., Endicott, P., Pestano, J. J., & González, A. M. (2014). The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. *BMC Evolutionary Biology*, *14*(1), 109. https://doi.org/10.1186/1471-2148-14-109

Serra-Vidal, G., Lucas-Sanchez, M., Fadhlaoui-Zid, K., Bekada, A., Zalloua, P., & Comas, D. (2019). Heterogeneity in Palaeolithic Population Continuity and Neolithic Expansion in North Africa. *Current Biology*, *29*(22), 3953-3959.e4. https://doi.org/10.1016/j.cub.2019.09.050

Shariati, M., Mafipour, M. S., Mehrabi, P., Zandi, Y., Dehghani, D., Bahadori, A., Shariati, A., Trung, N. T., Salih, M. N. A., & Poi-Ngian, S. (2019). Application of Extreme Learning Machine (ELM) and Genetic Programming (GP) to design steel-concrete

composite floor systems at elevated temperatures. *Steel and Composite Structures*, *33*(3), 319–332. https://doi.org/10.12989/scs.2019.33.3.319
Sharp, A. J., Cheng, Z., & Eichler, E. E. (2006). Structural Variation of the Human Genome. *Annual Review of Genomics and Human Genetics*, *7*(1), 407–442. https://doi.org/10.1146/annurev.genom.7.080505.115618
Shemirani, R., Belbin, G. M., Avery, C. L., Kenny, E. E., Gignoux, C. R., & Ambite, J. L. (2021). Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-22910-w
Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486
Sheppard, C. W. (1969). Computer simulation of stochastic processes through model-sampling (Monte Carlo) techniques. *FEBS Letters*, *2*(S1). https://doi.org/10.1016/0014-5793(69)80071-2
Si-Ammour, S. (2022). The Chronology of the Neolithic in Northwest Africa. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki*, *164*(3), 228–242. https://doi.org/10.26907/2541-7738.2022.3.228-242
Simões, L. G., Günther, T., Martínez-Sánchez, R. M., Vera-Rodríguez, J. C., Iriarte, E., Rodríguez-Varela, R., Bokbot, Y., Valdiosera, C., & Jakobsson, M. (2023). Northwest African Neolithic initiated by migrants from Iberia and Levant. *Nature*, *618*(7965), 550–556. https://doi.org/10.1038/s41586-023-06166-6
Sivanandam, S. N., & Deepa, · S N. (2008). *Introduction to Genetic Algorithms*. Springer.
Skoglund, P., Ersmark, E., Palkopoulou, E., & Dalén, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, *25*(11), 1515–1519. https://doi.org/10.1016/j.cub.2015.04.019
SLATKIN, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genetical Research*, *78*(1), 49–57. https://doi.org/10.1017/S0016672301005183
Smith, A. (2001). Saharo-Sudanese Neolithic. In *Encyclopedia of Prehistory Volume 1: Africa* (pp. 245–259). Springer US. https://doi.org/10.1007/978-1-4615-1193-9_19
Sokal, R. R., Oden, N. L., & Wilson, C. (1991). Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature*, *354*, 143–145.
Solé-Morata, N., García-Fernández, C., Urasin, V., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., Comas, D., & Calafell, F. (2017). Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183

(M81). *Scientific Reports*, *7*(1), 15941. https://doi.org/10.1038/s41598-017-16271-y
Sörensen, K. (2015). Metaheuristics-the metaphor exposed. *International Transactions in Operational Research*, *22*(1), 3–18. https://doi.org/10.1111/itor.12001
Sousa, V., & Hey, J. (2013). Understanding the origin of species with genome-scale data: Modelling gene flow. In *Nature Reviews Genetics* (Vol. 14, Issue 6, pp. 404–414). https://doi.org/10.1038/nrg3446
Sticca, E. L., Belbin, G. M., & Gignoux, C. R. (2021). Current Developments in Detection of Identity-by-Descent Methods and Applications. In *Frontiers in Genetics* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fgene.2021.722602
Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLoS Computational Biology*, *9*(1). https://doi.org/10.1371/journal.pcbi.1002803
Tajima, F. (1983). EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS. *Genetics*, *105*(2), 437–460. https://doi.org/10.1093/genetics/105.2.437
Tavare, S., Balding, D. J., Griffiths', +j R C, & Donneuyst2, P. (1997). Inferring Coalescence Times From DNA Sequence Data. *Genetics*, *145*, 505–518.
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 of the United States of America*, *112*(25), 7677–7682. https://doi.org/10.1073/pnas.1503717112
The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393
Thorsby, E. (2009). A short history of HLA. *Tissue Antigens*, *74*(2), 101–116. https://doi.org/10.1111/j.1399-0039.2009.01291.x
Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., & Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, *39*(1), 31–40. https://doi.org/10.1038/ng1946
Tschermak, E. von. (1900). Über künstliche Kreuzung bei Pisum sativum. *Zeitschrift Für Das Landwirtschaftliche Versuchswesen*

in Österreich, *3*, 465–555.
Turchi, C., Buscemi, L., Giacchino, E., Onofri, V., Fendt, L., Parson, W., & Tagliabracci, A. (2009). Polymorphisms of mtDNA control region in Tunisian and Moroccan populations: An enrichment of forensic mtDNA databases with Northern Africa data. *Forensic Science International: Genetics*, *3*(3), 166–172. https://doi.org/10.1016/j.fsigen.2009.01.014
Van De Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E. H., El Hajraoui, M. A., Amzazi, S., Hublin, J. J., Pääbo, S., Schiffels, S., Meyer, M., Haak, W., Jeong, C., & Krause, J. (2018). Pleistocene north african genomes link near eastern and sub-saharan african human populations. *Science*, *360*(6388), 548–552. https://doi.org/10.1126/science.aar8380
Vicente, M., Priehodová, E., Diallo, I., Podgorná, E., Poloni, E. S., Černý, V., & Schlebusch, C. M. (2019). Population history and genetic adaptation of the Fulani nomads: Inferences from genome-wide data and the lactase persistence trait. *BMC Genomics*, *20*(1), 915. https://doi.org/10.1186/s12864-019-6296-7
Vikhar, P. A. (2017). Evolutionary algorithms: A critical review and its future prospects. *Proceedings - International Conference on Global Trends in Signal Processing, Information Computing and Communication, ICGTSPICC 2016*, 261–265. https://doi.org/10.1109/ICGTSPICC.2016.7955308
Villanea, F. A., & Schraiber, J. G. (2019). Multiple episodes of interbreeding between nean- derthal and modern humans. *Nature Ecology & Evolution*, *3*(1), 39–44.
Wakeley, J. (2020). Developments in coalescent theory from single loci to chromosomes. In *Theoretical Population Biology* (Vol. 133, pp. 56–64). Academic Press Inc. https://doi.org/10.1016/j.tpb.2020.02.002
Wang, J., Santiago, E., & Caballero, A. (2016). Prediction and estimation of effective population size. In *Heredity* (Vol. 117, Issue 4, pp. 193–206). https://doi.org/10.1038/hdy.2016.43
Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0
Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte Des Vereins Für Vaterländische Naturkunde in Württemberg*, *64*, 368–382.
Weiss, G., & Von Haeseler, A. (1998). Inference of Population History Using a Likelihood Approach The coalescent (Kingman 1982a,b,c) is an efficient. *Genetics*, *149*, 1539–1546. https://academic.oup.com/genetics/article/149/3/1539/5923014
Wendorf, F., Close, A. E., & Schild, R. (1985). Prehistoric Settlements in the Nubian Desert: a region that is now virtually uninhabitable

contains a record of human adaptation to arid environments that may be 500,000 years long. *American Scientist*, *73*(2), 132–141.
Wiuf, C., & Hein, J. (1999). Recombination as a Point Process along Sequences. In *Theoretical Population Biology* (Vol. 55). www.idealibrary.com
Wright, S. (1937). The Distribution of Gene Frequencies in Populations. *Proceedings of the National Academy of Sciences*, *23*(6), 307–320. https://doi.org/10.1073/pnas.23.6.307
Yamamoto, F., Cid, E., Yamamoto, M., & Blancher, A. (2012). ABO Research in the Modern Era of Genomics. *Transfusion Medicine Reviews*, *26*(2), 103–118. https://doi.org/10.1016/j.tmrv.2011.08.002
Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*(10), 931–934.

6 ANNEXES

6.1 Supplementary information: Modelling the demographic history of human North African genomes points to soft split divergence between populations.

Modelling the demographic history of human North African genomes points to a recent soft split divergence between populations.

Authors
Jose M Serradell, Jose M Lorenzo-Salazar, Carlos Flores, Oscar Lao and David Comas

Additional file 1: Supplementary figures S1 to S11 and Supplementary Tables S1 to S10

Additional file 2: Xlsx file with the posterior values of the accepted models in the ABC-DL analysis (Model D in first analysis and Model D4 in second analysis). Factor 2, Kullback-Leiber, and Spearman correlation tables for the accepted models in both ABC-DL analysis.

Additional file 3: Histograms with the posterior versus prior distributions of all parameters for the best model in ABC-DL analysis (Model D4)

Additional file 4: Spearman correlation plots for all parameters in the best model in the ABC-DL analysis (Model D4)

Additional file 5: Parameter values for the best 10 models in the GP4PG analysis.

Additional file 1.

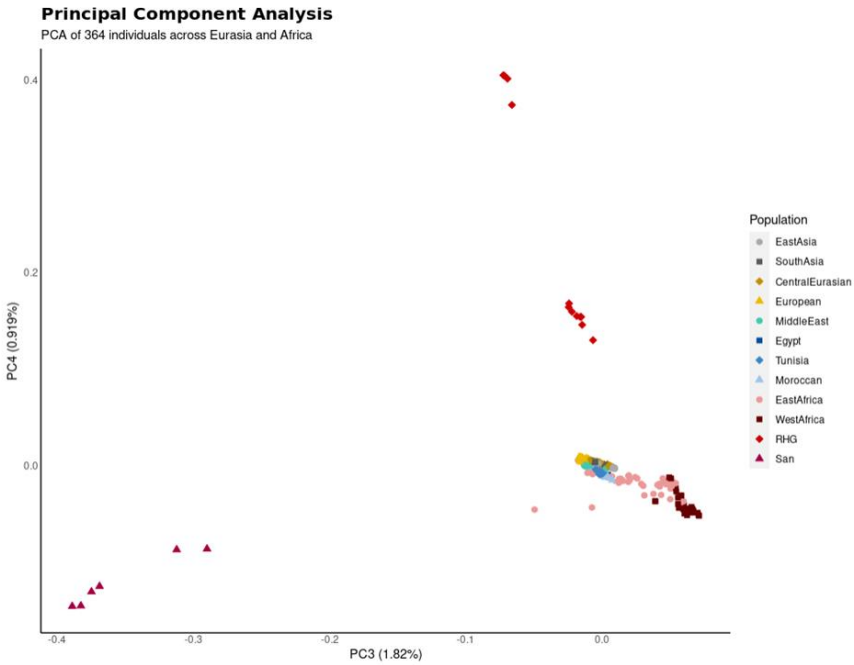


Fig. S 1: **Principal Component Analysis on genomic dataset of North Africa. Visualization of PC3 & PC4**

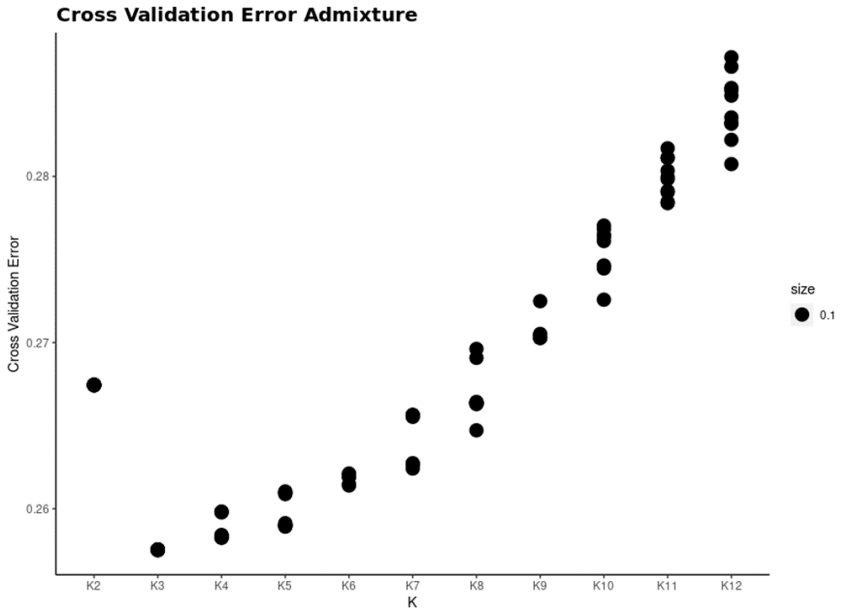


Fig. S 2: **CrossValidation error of ADMIXTURE analysis** with K=2 to K=12. Best K is K=3

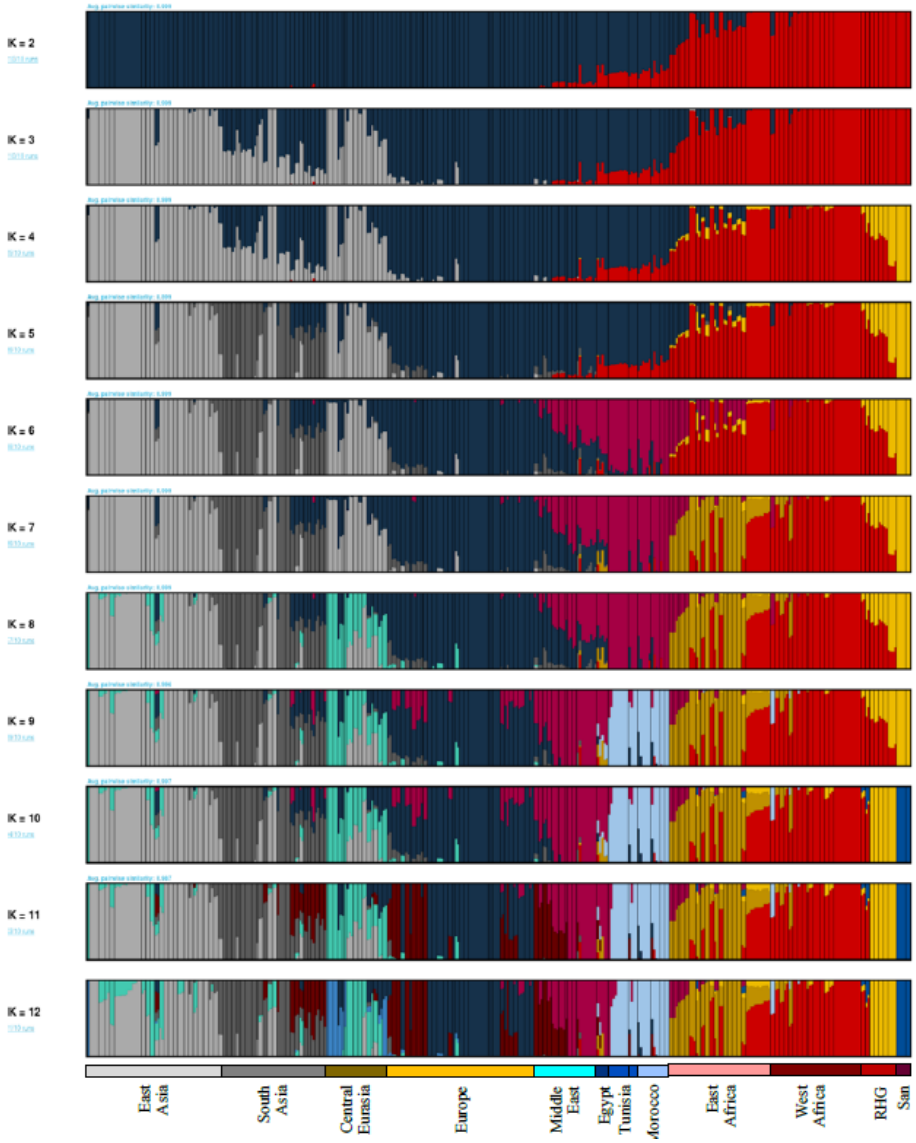


Fig. S 3: ADMIXTURE analysis on 364 individuals with K=2 to K=12

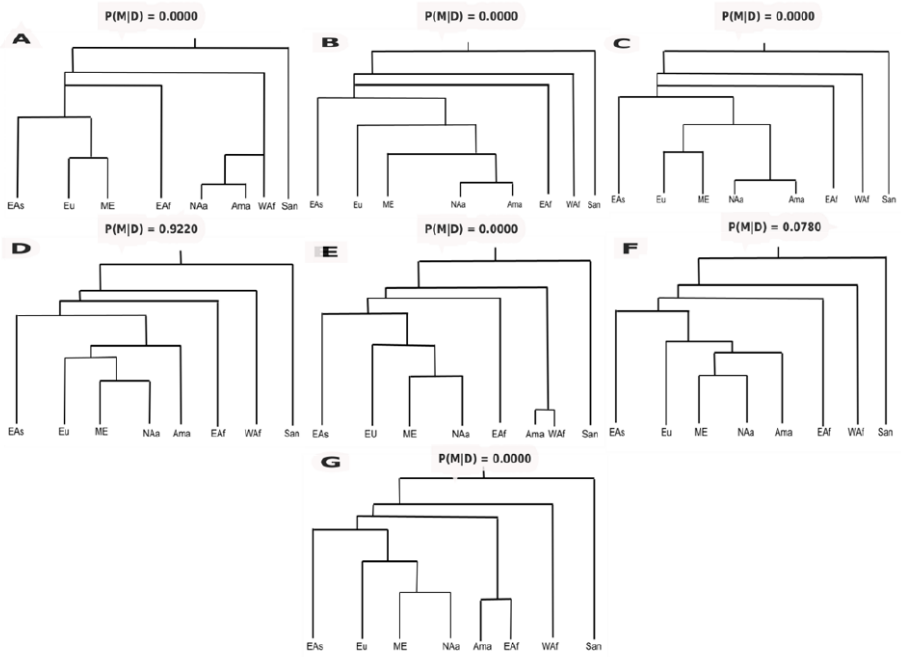


Fig. S 4: Competing topologies tested in ABC-DL analysis. Seven different topologies included on the ABC-DL analyses considering North African Arab (NAa), North Africa Amazigh (Ama), Middle Eastern (ME), European (Eu), East Asian (EAs), East African (EAf), West African (WAf), and Ju/'hoansi (San) populations.

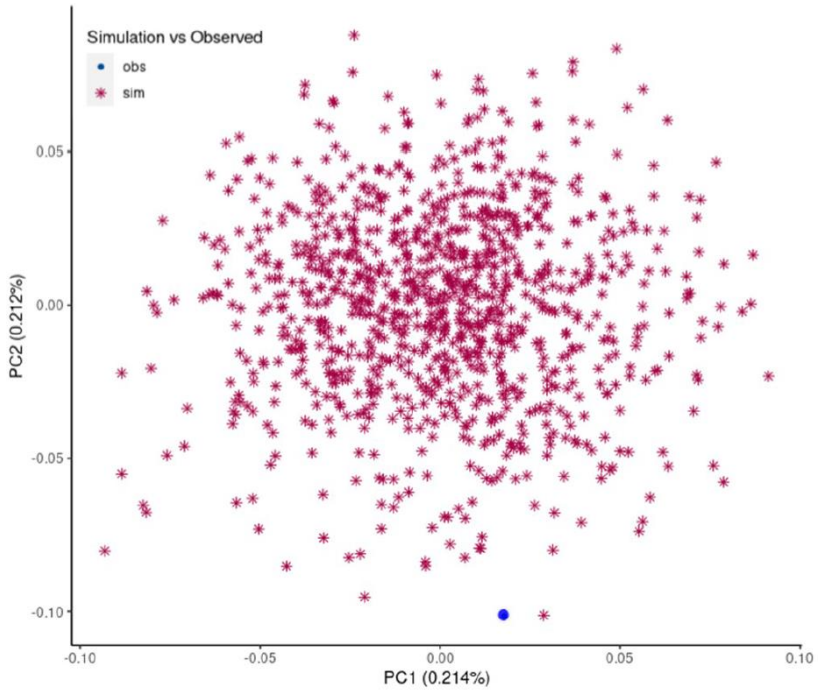


Fig. S 5: Replication PCA for Model D_4 in ABC-DL analysis. PCA for 1000 simulations of the model D_4, -the best model in the ABC-DL analysis- and the replication dataset of observed data. Observed data is an outlier in the PCA indicating that the ABC-DL model cannot properly replicate the diversity observed in the dataset.

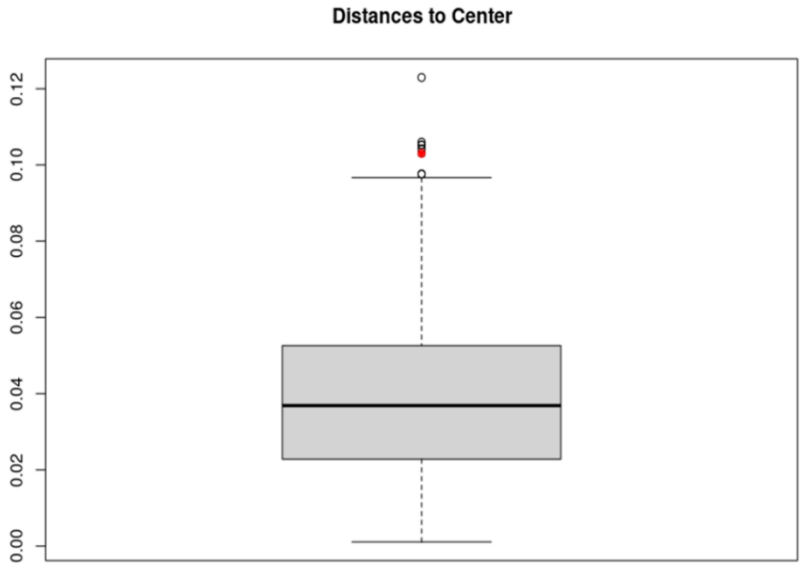


Fig. S 6: Box plot of the distances between each simulation in the PCA and the centroid of the PCA. The red dot represents the observed data as an outlier of the distances.

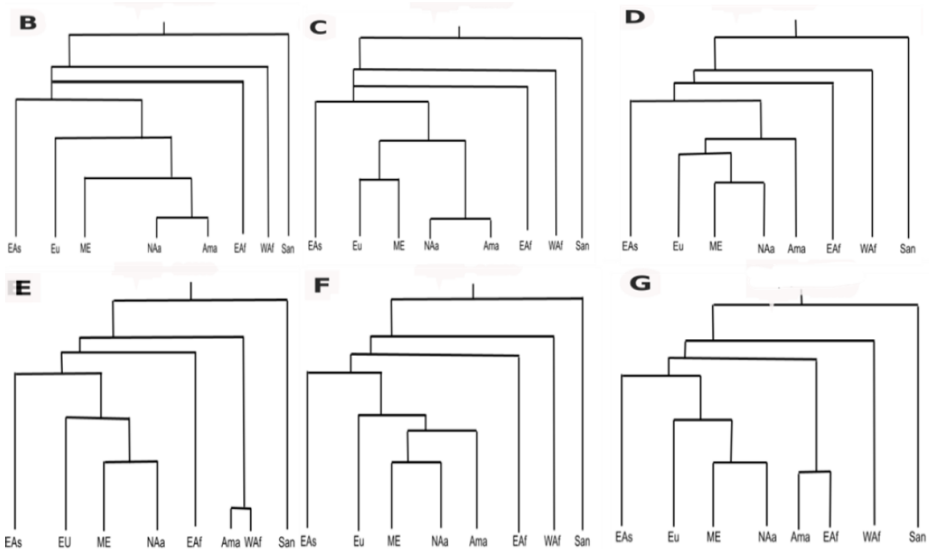


Fig. S 7: Competing topologies tested in GP4PG analysis. The competing topologies for the GP4PG analysis are the same as the ones used in the ABC-DL analysis but discarding Model A, due to being the worst performing one in previous analysis and due resource consumption.

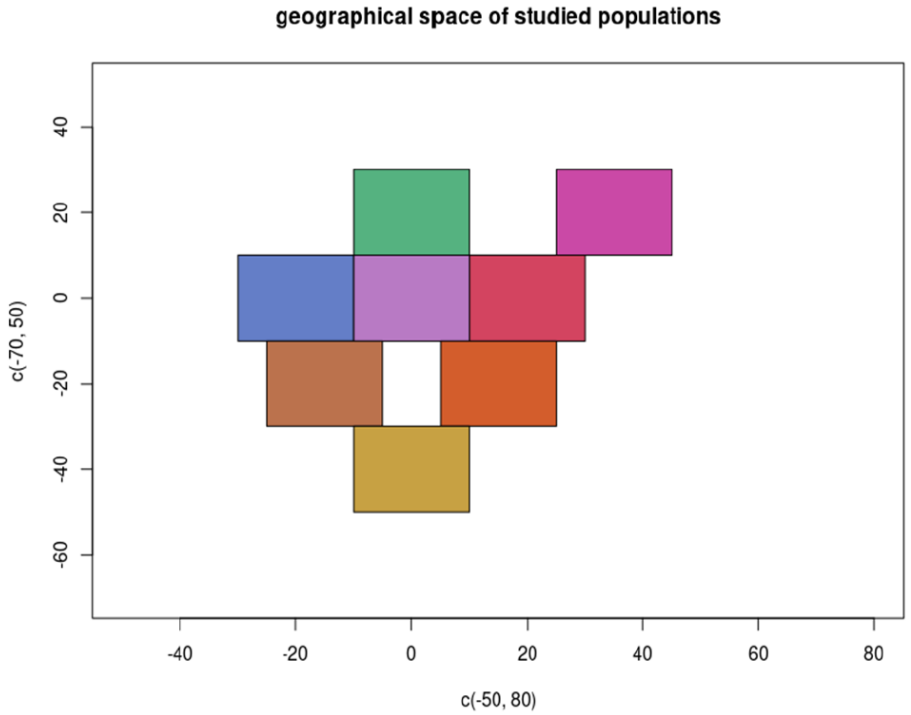


Fig. S 8: Coordinates of the different ecodemes we are testing in the GP4PG analysis. Each ecodeme has the exact same size and major geographical barriers such as seas and deserts has been removed for the sake of simplicity.

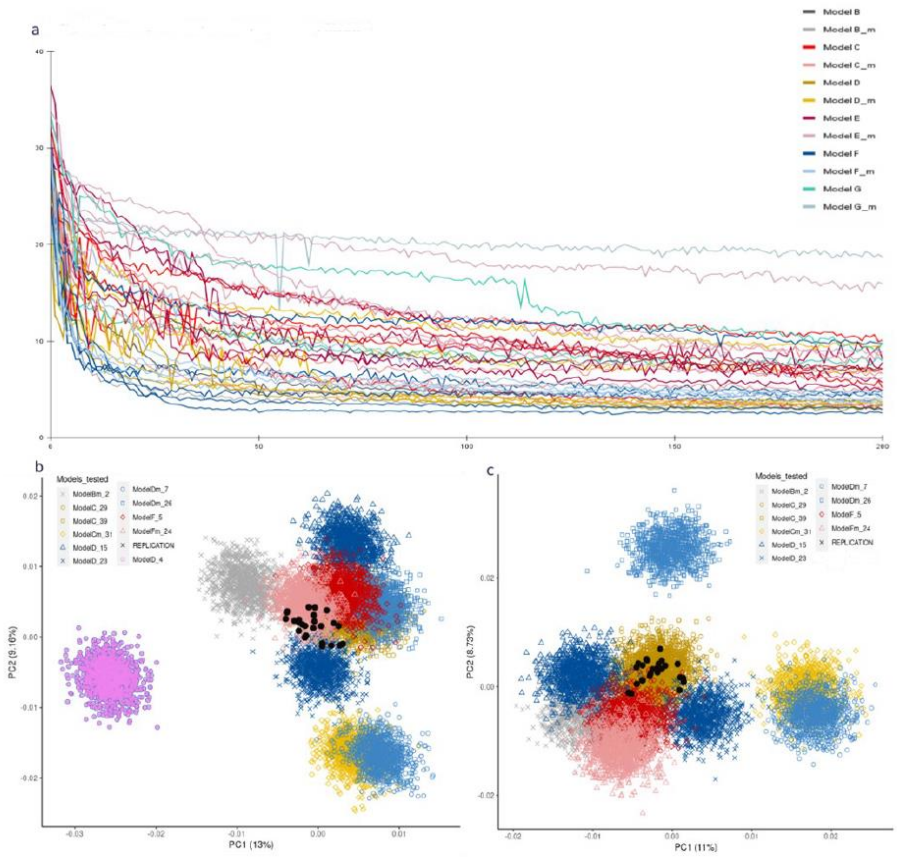


Fig. S 9: Fitness of the different runs of the genetic algorithm. **a.** Distribution of the fitness error of 40 independent iterations of the GP4PG algorithm with 6 competing topologies (B to G in ABC-DL) during 200 generations. Model D appears as the most selected model in a fourth of all the iterations, with D_15 as the model with the least error. **b.** PCA plot comparing the jSFS obtained from simulations of the best ABC-DL model with the 10 best GP4PG models. GP4PG simulations explain the observed data better than the ABC-DL. **c**. Same PCA plot as **b** but not including the simulations from ABC-DL result. Models C_29 and C_39 are the ones that show a more similar jSFS to the one produced by the observed data.

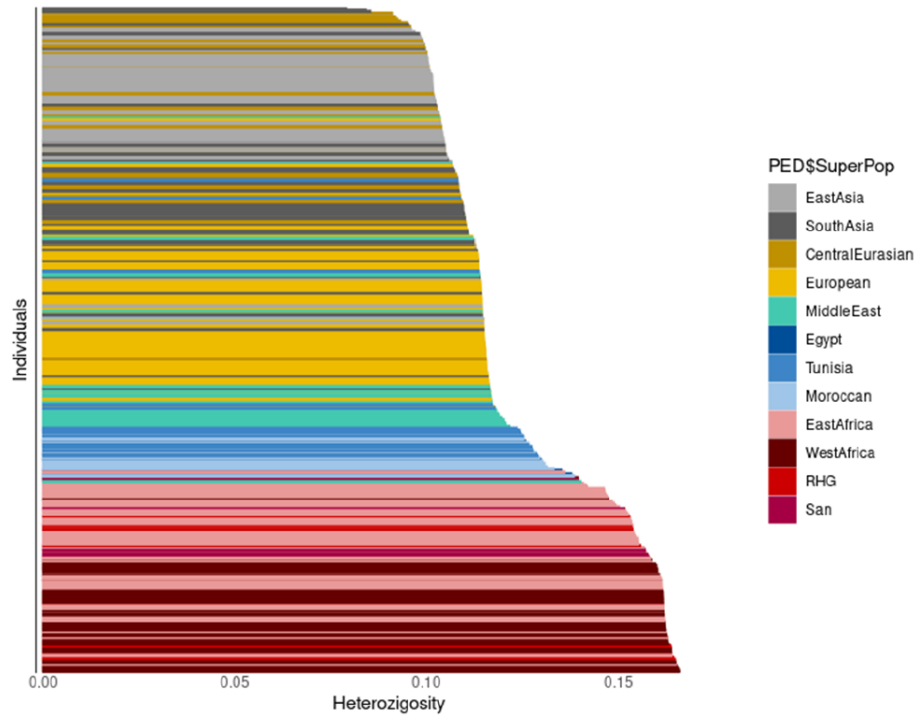


Fig. S 10: Observed heterozygosity per individual compared by superpopulation. Sub-Saharan populations present a higher heterozygosity than Eurasian populations, North African individuals have heterozygosity levels between the sub-Saharan and the Eurasians, probably due to gene flow from sub-Sharan populations to north African individuals.

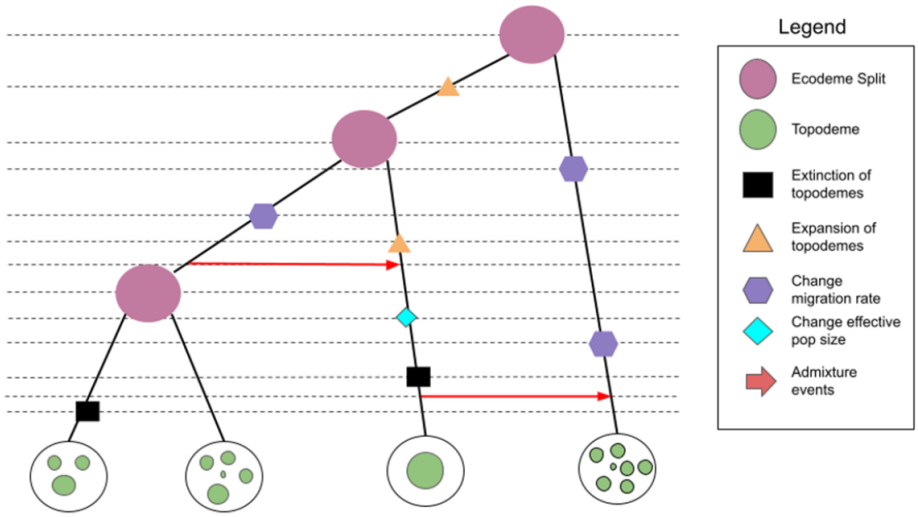


Fig. S 11: A tree-based depiction of the demographic relationships between different ecodemes and topodemes. Each node codes for a demographic event that occurs at a given time (dashed lines).

	ModelA	ModelB	ModelC	ModelD	ModelE	ModelF	ModelG
ModelA	**1**	0	0	0	0	0	0
ModelB	0	**0.7941**	0.1023	0.0122	0	0.0914	0
ModelC	0	0.0603	**0.9395**	0.0001	0	0.0001	0
ModelD	0	0.0145	0.002	**0.7579**	0	0.2256	0.0001
ModelE	0	0	0	0	**1**	0	0
ModelF	0	0.1565	0.002	0.1881	0	**0.6534**	0.0001
ModelG	0	0	0	0	0	0	**1**

Table S1: Confusion matrix computed with the 7 models under evaluation. 50 randomly sampled simulations per model were used as "observed" data for the ABC-DL algorithm. Diagonal, in bold, shows the probability of a model being correctly assigned by the A

ModelA	ModelB	ModelC	ModelD	ModelE	ModelF	ModelG
0	0	0	0.922	0	0.078	0

Table S 2: Proportion of accepted simulations using postpr function for the "abc" package with tolerance = 0.0008. Model D is present 92.2% of times in the 1000 closest simulations to the observed data.

	ModelA	ModelB	ModelC	ModelD	ModelE	ModelF	ModelG
ModelA	NA	NA	NA	Inf	NA	Inf	NA
ModelB	NA	NA	NA	Inf	NA	Inf	NA
ModelC	NA	NA	NA	Inf	NA	Inf	NA
ModelD	0	0	0	1	0	0.08459	0
ModelE	NA	NA	NA	Inf	NA	Inf	NA
ModelF	0	0	0	11.8205	0	1	0
ModelG	NA	NA	NA	Inf	NA	Inf	NA

Table S 3: Bayes factor for the ABC-DL topology discrimination analysis. Model D is 11.8 times better at explaining the observed data than the second-best model (Model F).

	Model D1	Model D2	Model D3	Model D4	Model D5
Model D1	**0.5888**	0.0058	0.0064	0.0061	0.3933
Model D2	0.0019	**0.3712**	0.3499	0.2759	0.0011
Model D3	0.0003	0.3507	**0.3557**	0.2931	0.0001
Model D4	0.0003	0.2729	0.3020	**0.4248**	0
Model D5	0.3888	0	0	0	**0.6112**

Table S 4: Confusion matrix computed with the five D models under evaluation. 50 randomly sampled simulations per model were used as "observed" data for the ABC-DL algorithm. Diagonal, in bold, shows the probability of a model being correctly assigned by the ABC.

Model D1	Model D2	Model D3	Model D4	Model D5
0.0167	0.0800	0.0944	0.7622	0.0468

Table S 5: Proportion of accepted simulations using postpr function for the "abc" package with tolerance = 0.001. Model D4 is present 76.22% of times in the 1000 closest simulations to the observed data.

	Model D1	Model D2	Model D3	Model D4	Model D5
Model D1	1	0.2087	0.1768	0.0219	0.3567
Model D2	4.7913	1	0.8471	0.1049	1.7093
Model D3	5.6560	1.1805	1	0.1238	2.0178
Model D4	45.6714	9.5321	8.0748	1	16.2932
Model D5	2.8031	0.5850	0.4956	0.0614	1

Table S 6: Bayes factor for the ABC-DL with different admixture patterns. Model D4 is 8.074 times better at explaining the observed data than the second-best model (Model D3).

SampleID	AccesionID	Population	Superpopulation	Reference
CEU01	SAME123392	Centre European Utah	European	1000 Genomes
CEU02	SAMN00801650	Centre European Utah	European	1000 Genomes
CHB01	SAME124093	Han Chinese	East Asia	1000 Genomes
CHB02	SAME123926	Han Chinese	East Asia	1000 Genomes
JHN01	SAMEA3302682	Ju|'hoan North	South Africa San	SGDP, Mallick 2016
JHN02	SAMEA3302894	Ju|'hoan North	South Africa San	SGDP, Mallick 2016
LWK04	SAMN00001054	Luhya	East Africa	1000 Genomes
LWK07	SAMN00001112	Luhya	East Africa	1000 Genomes
QTR01	SAMN03800116	Qatar	Middle East	Fakhro 2016
QTR02	SAMN03800117	Qatar	Middle East	Fakhro 2016
TUN10	SAMEA4969309	Tunisia Arab	North Africa Arab	Serra-Vidal 2019
TUN11	SAMEA4969310	Tunisia Arab	North Africa Arab	Serra-Vidal 2019
TUN12	SAMEA4969294	Tunisia Chenini	North Africa Amazigh	Serra-Vidal 2019
TUN13	SAMEA4969295	Tunisia Chenini	North Africa Amazigh	Serra-Vidal 2019
YRI01	SAME122984	Yoruba	West Africa	1000 Genomes
YRI02	SAME125386	Yoruba	West Africa	1000 Genomes

Table S 7: Samples for the demographic analysis of North Africa

Demographic Event	Prior probability of inclusion in GP4PG
Change in Ne	1
Extintion event of a topodeme	1
Expansion event of a topodeme	1
Change in migration rate	1
Admixture event from one topodeme to another	0.3

Table S 8: Possible demographic events and associated probabilities in GP4PG algorithm.

Parameter	Distribution	A	B	C	D	E	F	G
misclassification	U(0.0,5.0E-4)	X	X	X	X	X	X	X
NeSan	U(10000.0,90000.0)	X	X	X	X	X	X	X
NeYoruba (WAf)	U(10000.0,90000.0)	X	X	X	X	X	X	X
NeLuhya (EAf)	U(10000.0,60000.0)	X	X	X	X	X	X	X
NeTunisia_Chenini (NAfb)	U(1000.0,20000.0)	X	X	X	X	X	X	X
NeTunisia (NAfa)	U(1000.0,40000.0)	X	X	X	X	X	X	X
NeQatar (ME)	U(1000.0,40000.0)	X	X	X	X	X	X	X
NeCEU (EU)	U(1000.0,40000.0)	X	X	X	X	X	X	X
NeHan (EAs)	U(1000.0,40000.0)	X	X	X	X	X	X	X
tME_EU	U(200.0,800.0)	X		X				
NeME_EU	U(1000.0,20000.0)	X		X				
tNAfb_NAfa	U(10.0,150.0)	X	X	X				
NeNAfb_NAfa	U(1000.0,20000.0)	X	X	X				
tME_EU_EAs	U(tME_EU,3000.0)	X						
NeME_EU_EAs	U(1000.0,7000.0)	X						
tWAf_NAfb_NAfa	U(tNAfb_NAfa,600.0)	X						
NeWAf_NAfb_NAfa	U(1000.0,30000.0)	X						
tEAf_ME_EU_EAs	U(tME_EU_EAs,4000.0)	X						
NeEAf_ME_EU_EAs	U(1000.0,10000.0)	X						
tWAf_EAf_EAs_ME_EU_NAfb_NAfa	U(tEAf_ME_EU_EAs,6000.0)	X	X	X	X	X	X	X
NeWAf_EAf_EAs_ME_EU_NAfb_NAfa	U(1000.0,30000.0)	X	X	X	X	X	X	X
tSan_WAf_EAf_EAs_ME_EU_NAfb_NAfa	U(tWAf_EAf_EAs_ME_EU_NAfb_NAfa, 12500.0)	X	X	X	X	X	X	X

NeSan_WAf_EAf_EAs_ME_EU_NAfb_NAfa	U(1000.0,30000.0)	X	X	X	X	X	X	X
tNAfb_NAfa_ME	U(tNAfb_NAfa,600.0)		X					
NeNAfb_NAfa_ME	U(1000.0,15000.0)		X					
tNAfb_NAfa_ME_EU	U(tNAfb_NAfa_ME,800.0)		X					
NeNAfb_NAfa_ME_EU	U(1000.0,20000.0)		X	X	X		X	
tNAfb_NAfa_ME_EU_EAs	U(tNAfb_NAfa_ME_EU,3000.0)		X	X	X		X	
NeNAfb_NAfa_ME_EU_EAs	U(1000.0,7000.0)		X	X	X		X	
tNAfb_NAfa_ME_EU_EAs_EAf	U(tNAfb_NAfa_ME_EU_EAs,4000.0)		X	X	X		X	X
NeNAfb_NAfa_ME_EU_EAs_EAf	U(1000.0,30000.0)		X	X	X		X	X
tNAfb_NAfa_ME_EU	U(tME_EU,1000.0)			X				
tNAfa_ME	U(150.0,600.0)				X	X	X	X
NeNAfa_ME	U(1000.0,20000.0)				X	X	X	X
tNAfa_ME_EU	U(tNAfa_ME,800.0)				X	X		X
NeNAfa_ME_EU	U(1000.0,20000.0)				X	X		X
tNAfb_NAfa_ME_EU	U(tNAfa_ME_EU,1000.0)				X			
tNAfa_ME_EU_EAs	U(tNAfa_ME_EU,1000.0)					X		
NeNAfa_ME_EU_EAs	U(1000.0,7000.0)					X		X
tWAf_NAfb	U(150.0,600.0)					X		
NeWAf_NAfb	U(1000.0,30000.0)					X		
tNAfa_ME_EU_EAs_EAf	U(tNAfa_ME_EU_EAs,4000.0)					X		
NeNAfa_ME_EU_EAs_EAf	U(1000.0,30000.0)					X		
tNAfb_NAfa_ME	U(tNAfb_NAfa,800.0)						X	
NeNAfb_NAfa_ME	U(1000.0,20000.0)						X	
tNAfb_NAfa_ME_EU	U(tNAfb_NAfa_ME,1000.0)						X	
tNAfa_ME_EU_EAs	U(tNAfa_ME_EU,3000.0)							X
tEAf_NAfb	U(150.0,1000.0)							X
NeEAf_NAfb	U(1000.0,30000.0)							X

Table S 9: Parameters and prior distributions of the seven considered models in Fig. S4

Parameter	Distribution	D	D 2	D 3	D 4	D 5
misclassification	U(0.0, 5.0E-4)	X	X	X	X	X
NeSan	U(10000.0,90000.0)	X	X	X	X	X
NeYoruba	U(10000.0,90000.0)	X	X	X	X	X
NeLuhya	U(10000.0,60000.0)	X	X	X	X	X
NeTunisia_Chenini	U(1000.0,20000.0)	X	X	X	X	X
NeTunisia	U(1000.0,40000.0)	X	X	X	X	X
NeQatar	U(1000.0,40000.0)	X	X	X	X	X
NeCEU	U(1000.0,40000.0)	X	X	X	X	X
NeHan	U(1000.0,40000.0)	X	X	X	X	X
NeBasal_Eurasian_ghost (BEi)	U(1000.0,20000.0)			X	X	X
NeAfrican_ghost (Xa)	U(1000.0,20000.0)				X	X
migrationNAfa_NAfb	U(0.0,5.0E-4)		X	X	X	
migrationNAfb_NAfa	U(0.0,5.0E-4)		X	X	X	
migrationME_NAfb	U(0.0,5.0E-4)		X	X	X	
migrationNAfb_ME	U(0.0,5.0E-4)		X	X	X	
migrationME_NAfa	U(0.0,5.0E-4)		X	X	X	
migrationNAfa_ME	U(0.0,5.0E-4)		X	X	X	
migrationEU_ME	U(0.0,5.0E-4)		X	X	X	
migrationEU_NAfb	U(0.0,5.0E-4)		X	X	X	
migrationNAfb_EU	U(0.0,5.0E-4)		X	X	X	
migrationEU_NAfa	U(0.0,5.0E-4)		X	X	X	
migrationNAfa_EU	U(0.0,5.0E-4)		X	X	X	
migrationEAs_EU	U(0.0,5.0E-4)		X	X	X	
migrationEU_EAs	U(0.0,5.0E-4)		X	X	X	
migrationEAf_NAfb	U(0.0,5.0E-4)		X	X	X	
migrationNAfb_EAf	U(0.0,5.0E-4)		X	X	X	
migrationEAf_NAfa	U(0.0,5.0E-4)		X	X	X	
migrationNAfa_EAf	U(0.0,5.0E-4)		X	X	X	
migrationEAf_ME	U(0.0,5.0E-4)		X	X	X	

migrationME_EAf	U(0.0,5.0E-4)			X	X	X
migrationEAf_WAf	U(0.0,5.0E-4)			X	X	X
migrationWAf_EAf	U(0.0,5.0E-4)			X	X	X
migrationWAf_NAfb	U(0.0,5.0E-4)			X	X	X
migrationNAfb_WAf	U(0.0,5.0E-4)			X	X	X
migrationWAf_NAfa	U(0.0,5.0E-4)			X	X	X
migrationNAfa_WAf	U(0.0,5.0E-4)			X	X	X
migrationSan_to_WAf	U(0.0,5.0E-4)			X	X	X
migrationSan_to_EAf	U(0.0,5.0E-4)			X	X	X
tNAfa_ME	U(150.0,600.0)	X	X	X	X	X
NeNAfa_ME	U(1000.0,20000.0)	X	X	X	X	X
tNAfa_ME_EU	U(tNAfa_ME,800.0)	X	X	X	X	X
NeNAfa_ME_EU	U(1000.0,20000.0)	X	X	X	X	X
tNAfb_NAfa_ME_EU	U(tNAfa_ME_EU,1000.0)	X	X	X	X	X
tNAfb_NAfa_ME_EU_EAs	U(tNAfb_NAfa_ME_EU,20000.0)	X	X	X	X	X
NeNAfb_NAfa_ME_EU_EAs	U(1000.0,7000.0)	X	X	X	X	X
tNAfb_NAfa_ME_EU_EAs_EAf	U(tNAfb_NAfa_ME_EU_EAs,4000.0)	X	X	X	X	X
NeNAfb_NAfa_ME_EU_EAs_EAf	U(1000.0,30000.0)	X	X	X	X	X
tWAf_EAf_EAs_ME_EU_NAfb_NAfa	U(tEAf_ME_EU_EAs,6000.0)	X	X	X	X	X
NeWAf_EAf_EAs_ME_EU_NAfb_NAfa	U(1000.0,30000.0)	X	X	X	X	X
tSan_WAf_EAf_EAs_ME_EU_NAfb_NAfa	U(tWAf_EAf_EAs_ME_EU_NAfb_NAfa, 12500.0)	X	X	X	X	X
NeSan_WAf_EAf_EAs_ME_EU_NAfb_NAfa	U(1000.0,30000.0)	X	X	X	X	X
tSan_WAf_EAf_EAs_ME_EU_NAfb_NAfa_Xa	U(tNAfb_NAfa_ME_EU_BEi_EAs_WAf_San, 14000.0)				X	X
NeSan_WAf_EAf_EAs_ME_EU_NAfb_NAfa_Xa	U(1000.0,30000.0)				X	X
tNAfb_NAfa_ME_EU_BEi	U(tNAfb_NAfa_ME_EU,2000.0)			X	X	X
NeNAfb_NAfa_ME_EU_BEi	U(1000.0,20000.0)			X	X	X
tAdmxME_NA	U(60.0,tNAfa_ME)		X	X	X	X

admixtureME_NA	U(0.001,0.20)	X	X	X	X
tAdmxEU_NA	U(20.0,tNAfa_ME)			X	X
admixtureEU_NA	U(0.001,0.10)			X	X
tAdmxEU_NAb	U(20.0,tNAfa_ME)			X	X
admixtureEU_NAb	U(0.001,0.10)			X	X
tAdmxME_NAb	U(60.0,tNAfa_ME)			X	X
admixtureME_NAb	U(0.001,0.20)			X	X
tAdmxMENA_Amazigh	U(tNAfa_ME,tNAfa_ME_EU)	X	X	X	X
admixtureMENA_Amazigh	U(0.001,0.10)	X	X	X	X
tAdmxWAf_Amazigh	U(10.0,tNAfb_NAfa_ME_EU)	X	X	X	X
admixtureWAf_Amazigh	U(0.001,0.10)	X	X	X	X
tAdmxWAf_Arab	U(10.0,tNAfa_ME)	X	X	X	X
admixtureWAf_Arab	U(0.001,0.10)	X	X	X	X
tAdmxEAf_Amazigh	U(10.0,tNAfb_NAfa_ME_EU)	X	X	X	X
admixtureEAf_Amazigh	U(0.001,0.10)	X	X	X	X
tAdmxEAf_Arab	U(10.0,tNAfa_ME)	X	X	X	X
admixtureEAf_Arab	U(0.001,0.10)	X	X	X	X
tAdmxBEi_MENAU	U(tNAfa_ME_EU, tNAfb_NAfa_ME_EU)		X	X	X
admixtureBEi_MENAU	U(0.001,0.20)		X	X	X
tAdmxBEi_AMENAU	U(tNAfb_NAfa_ME_EU, tNAfb_NAfa_ME_EU_BEi)		X	X	X
admixtureBEi_AMENAU	U(0.001,0.20)		X	X	X
tAdmxXa_San	U(700.0,tNAfb_NAfa_ME_EU_BEi_EAs_WAf_San)			X	X
admixtureXa_San	U(0.001,0.05)			X	X
tAdmxXa_WAf	U(700.0,tNAfb_NAfa_ME_EU_BEi_EAs_WAf)			X	X
admixtureXa_WAf	U(0.001,0.05)			X	X

Table S 10: Parameters and prior distributions of the five considered models in Fig. 2.

Additional file 2.

Model D results

Page 1: Mean, median, mode and other centrality statistics for each parameter in best Model D in the first run of ABC-DL

Table S1: Posterior values for Model D

	Mean	Median	Mode	0,025-CI	0,975-CI	HDI_89_low	HDI_89_high
missclassification	2.45E-04	2.45E-04	1.06E-04	9.88E-06	4.83E-04	2.64E-05	4.99E-04
NeKhs	64816.15	64190.48	63169.23	41668.39	87436.56	46037.74	89413.74
NeWAf	76693.14	78428.03	87425.60	54557.21	89399.31	57429.62	89994.85
NeEAf	34567.51	34224.35	15978.81	11025.86	59038.00	12051.09	59614.27
NeNAfb	10450.62	10209.55	9834.57	5791.67	16666.30	4844.44	15092.85
NeAfa	31228.33	32616.33	37053.80	16684.26	39558.70	18157.38	39981.90
NeME	12497.94	11137.22	9673.46	3619.39	30315.09	1915.99	26001.10
NeEU	25349.57	25464.76	25235.05	10535.79	38843.84	11859.70	39587.33
NeEAs	9278.33	9277.83	9215.72	4737.34	13712.30	4637.77	13528.99
tNAfa_ME	388.32	395.51	394.17	173.05	583.31	177.00	585.83
NeNAfa_ME	13211.07	14313.61	16405.53	3278.71	19702.43	4412.65	19996.83
tNAfa_ME_EU	416.18	410.48	376.91	210.66	639.01	203.64	626.89
NeNAfa_ME_EU	6733.76	6441.63	5470.85	2637.64	12753.32	2395.70	11921.37
tNAfb_NAfa_ME_EU	712.13	718.13	715.61	440.40	940.43	478.00	961.24
NeNAfb_NAfa_ME_EU	11062.73	10822.03	8680.08	2828.23	19123.04	3364.32	19328.55
tNAfb_NAfa_ME_EU_EAs	1153.04	1123.51	1107.98	846.12	1568.56	826.71	1541.26
NeNAfb_NAfa_ME_EU_EAs	4317.97	4452.60	4800.97	1358.58	6833.54	1627.19	6997.00
tNAfb_NAfa_ME_EU_EAs_EAf	2114.31	2085.24	2074.81	1528.48	2897.56	1476.95	2775.66
NeNAfb_NAfa_ME_EU_EAs_EAf	20078.78	21451.30	25519.43	6004.06	29536.32	7601.43	29948.31
tNAfb_NAfa_ME_EU_EAs_WAf	2339.91	2331.50	2423.55	1718.92	3085.32	1697.41	3048.02
NeNAfb_NAfa_ME_EU_EAs_WAf	16592.28	16379.90	14438.44	5295.04	28190.74	5679.85	28319.34
tNAfb_NAfa_ME_EU_EAs_WAf_Khs	4546.61	4510.28	4487.75	3893.68	5298.91	3946.57	5306.81
NeNAfb_NAfa_ME_EU_EAs_WAf_Khs	18654.03	18427.99	15625.50	8796.91	28794.08	8468.90	28361.21

Page 2: Three tables with other statistics to asses quality of ABC-DL results for Model D (Factor2, KL-divergence and Spearman correlation)

Table S1: FACTOR 2

	factor2_result	fcator2_prior
missclassification	0.739	0.74
NeKhs	0.991	0.838
NeWAf	0.995	0.798
NeEAf	0.83	0.832
NeNAfb	0.978	0.784
NeAfa	0.863	0.77
NeME	0.87	0.759
NeEU	0.893	0.751
NeEAs	0.955	0.764
tNAfa_ME	0.954	0.909
NeNAfa_ME	0.777	0.778
tNAfa_ME_EU	0.984	0.962
NeNAfa_ME_EU	0.93	0.786
tNAfb_NAfa_ME_EU	0.994	0.984
NeNAfb_NAfa_ME_EU	0.884	0.777
tNAfb_NAfa_ME_EU_EAs	0.996	0.927
NeNAfb_NAfa_ME_EU_EAs	0.901	0.839
tNAfb_NAfa_ME_EU_EAs_EAf	0.998	0.968
NeNAfb_NAfa_ME_EU_EAs_EAf	0.873	0.774
tNAfb_NAfa_ME_EU_EAs_EAf_WAf	1	0.981
NeNAfb_NAfa_ME_EU_EAs_EAf_WAf	0.874	0.765
tNAfb_NAfa_ME_EU_EAs_EAf_WAf_Khs	1	0.975
NeNAfb_NAfa_ME_EU_EAs_EAf_WAf_Khs	0.919	0.764

Table S2: Kullback-Leiber Divergence

	KL-divergenc	p-value
missclassification	0.00845	0.855
NeKhs	14.50866	0
NeWAf	14.50866	0
NeEAf	14.50866	0
NeNAfb	14.50866	0
NeAfa	14.50866	0
NeME	0.67860	0
NeEU	14.50866	0
NeEAs	14.50866	0
tNAfa_ME	14.50866	0
NeNAfa_ME	14.50866	0
tNAfa_ME_EU	14.50866	0
NeNAfa_ME_EU	2.43543	0
tNAfb_NAfa_ME_EU	14.50866	0
NeNAfb_NAfa_ME_EU	14.50866	0
tNAfb_NAfa_ME_EU_EAs	14.50866	0
NeNAfb_NAfa_ME_EU_EAs	14.50866	0
tNAfb_NAfa_ME_EU_EAs_EAf	14.50866	0
NeNAfb_NAfa_ME_EU_EAs_EAf	9.31572	0
tNAfb_NAfa_ME_EU_EAs_EAf_WAf	14.50866	0
NeNAfb_NAfa_ME_EU_EAs_EAf_WAf	6.21461	0
tNAfb_NAfa_ME_EU_EAs_EAf_WAf_Khs	14.50866	0
NeNAfb_NAfa_ME_EU_EAs_EAf_WAf_K	14.50866	0

Table S3: Spearman Correlation

	Spearman correlatic	p_value
missclassification	-0.0026	0.2637
NeKhs	0.0290	0.0000
NeWAf	0.9323	0.0000
NeEAf	0.9282	0.0000
NeNAfb	0.4279	0.0000
NeAfa	0.9543	0.0000
NeME	0.7074	0.0000
NeEU	-0.1805	0.0000
NeEAs	0.8105	0.0000
tNAfa_ME	0.9113	0.0000
NeNAfa_ME	0.9320	0.0000
tNAfa_ME_EU	0.7889	0.0000
NeNAfa_ME_EU	0.3858	0.0000
tNAfb_NAfa_ME_EU	-0.2988	0.0000
NeNAfb_NAfa_ME_EU	0.8873	0.0000
tNAfb_NAfa_ME_EU_EAs	0.9494	0.0000
NeNAfb_NAfa_ME_EU_EAs	0.7370	0.0000
tNAfb_NAfa_ME_EU_EAs_EAf	0.8481	0.0000
NeNAfb_NAfa_ME_EU_EAs_EAf	0.9048	0.0000
tNAfb_NAfa_ME_EU_EAs_EAf_WAf	0.0802	0.0000
NeNAfb_NAfa_ME_EU_EAs_EAf_WAf	0.6959	0.0000
tNAfb_NAfa_ME_EU_EAs_EAf_WAf_K	0.9114	0.0000
NeNAfb_NAfa_ME_EU_EAs_EAf_WAf_	0.8420	0.0000

Model D4 results

Page 3: Mean, median, mode and other centrality statistics for each parameter in best Model D4 in the second run of ABC-DL

Table S1: Posterior values for Model D4

	Mean	Median	Mode	X0,025CI	X0,975CI	HDI_89_low	HDI_89_high
missclassification	2.54E-04	2.58E-04	3.80E-04	1.16E-05	4.85E-04	1.65E-05	4.88E-04
migrationNAfa_NAfb	2.64E-04	2.76E-04	4.69E-04	1.50E-05	4.94E-04	2.71E-05	5.00E-04
migrationNAfb_NAfa	2.55E-04	2.65E-04	2.86E-04	1.31E-05	4.84E-04	3.20E-06	4.72E-04
migrationME_NAfb	2.41E-04	2.30E-04	1.29E-04	1.17E-05	4.83E-04	2.12E-05	4.90E-04
migrationNAfb_to_ME	2.68E-04	2.79E-04	4.55E-04	2.16E-05	4.89E-04	3.56E-05	4.97E-04
migrationME_NAfa	2.53E-04	2.49E-04	2.40E-04	1.49E-05	4.87E-04	2.83E-05	4.97E-04
migrationNAfa_to_ME	2.55E-04	2.57E-04	4.10E-04	1.34E-05	4.85E-04	7.51E-06	4.78E-04
migrationEU_ME	2.44E-04	2.35E-04	1.64E-04	1.38E-05	4.89E-04	2.58E-05	4.98E-04
migrationEU_NAfb	2.84E-04	3.01E-04	3.01E-04	1.82E-05	4.89E-04	4.18E-05	4.99E-04
migrationNAfb_to_EU	2.73E-04	2.77E-04	4.02E-04	1.91E-05	4.91E-04	3.69E-05	4.99E-04
migrationEU_NAfa	2.66E-04	2.74E-04	4.95E-04	1.87E-05	4.90E-04	2.93E-05	4.95E-04
migrationNAfa_to_EU	2.51E-04	2.52E-04	2.68E-04	1.88E-05	4.92E-04	2.94E-05	5.00E-04
migrationEAs_EU	2.11E-04	1.90E-04	8.26E-05	1.06E-05	4.82E-04	1.20E-06	4.62E-04
migrationEU_to_EAs	1.73E-04	1.56E-04	9.55E-05	1.07E-05	4.43E-04	2.81E-07	3.98E-04
migrationEAf_NAfb	7.28E-05	5.93E-05	2.24E-05	2.83E-06	2.23E-04	1.90E-07	1.92E-04
migrationNAfb_to_EAf	8.92E-05	6.97E-05	3.29E-05	3.97E-06	2.66E-04	1.92E-07	2.34E-04
migrationEAf_NAfa	1.79E-04	1.49E-04	1.30E-05	6.07E-06	4.65E-04	1.29E-08	4.42E-04
migrationNAfa_to_EAf	3.24E-04	3.42E-04	4.96E-04	5.32E-05	4.95E-04	8.27E-05	4.99E-04
migrationEAf_to_ME	1.24E-04	9.62E-05	1.55E-06	2.99E-06	3.88E-04	1.50E-07	3.40E-04
migrationME_to_EAf	3.51E-04	3.64E-04	4.27E-04	1.14E-04	4.94E-04	1.55E-04	5.00E-04
migrationEAf_to_WAf	4.05E-04	4.26E-04	4.70E-04	2.18E-04	4.96E-04	2.53E-04	5.00E-04
migrationWAf_to_EAf	3.74E-04	4.01E-04	4.58E-04	9.87E-05	4.95E-04	1.72E-04	4.99E-04
migrationWAf_NAfb	1.47E-04	1.31E-04	3.39E-05	6.63E-06	3.85E-04	4.52E-07	3.34E-04
migrationNAfb_WAf	9.64E-05	8.00E-05	4.40E-05	2.89E-06	2.73E-04	2.64E-08	2.35E-04
migrationWAf_NAfa	1.56E-04	1.33E-04	7.15E-05	5.71E-06	4.47E-04	3.20E-07	3.96E-04
NeXa	10603.99	10314.06	8833.43	1358.23	19530.03	1268.75	19299.88
tNAfa_ME	296.77	283.90	228.29	160.35	531.10	150.69	487.23
NeNAfa_ME	17122.87	17804.01	20993.12	2805.71	29431.93	3867.87	29995.25
tNAfa_ME_EU	406.95	390.53	372.00	191.89	714.05	174.15	674.82
NeNAfa_ME_EU	6579.06	5922.15	4981.40	1952.13	13924.99	1348.78	12755.90
tNAfb_NAfa_ME_EU	625.00	613.45	572.80	333.76	942.39	331.81	936.89
NeNAfb_NAfa_ME_EU	12136.52	12669.97	14829.15	2905.95	19659.52	3908.28	19983.17
tNAfb_NAfa_ME_EU_BEi	1054.02	1004.89	868.46	543.36	1830.43	533.11	1815.36
NeNAfb_NAfa_ME_EU_BEi	12249.32	12656.39	14328.99	2372.20	19587.38	3413.34	19869.81
tNAfb_NAfa_ME_EU_BEi_EAs	1412.06	1395.20	1398.96	961.10	1970.08	876.64	1857.80
NeNAfb_NAfa_ME_EU_BEi_EAs	5619.97	5630.04	6233.44	1292.21	9713.15	1627.88	9882.21
tNAfb_NAfa_ME_EU_EAs_EAf	2733.82	2731.67	2599.52	1762.44	3737.65	1689.07	3657.37
NeNAfb_NAfa_ME_EU_EAs_EAf	23401.80	23944.75	31672.69	4465.12	39557.31	6243.51	39985.58
tNAfb_NAfa_ME_EU_BEi_EAs_EAf_WAf	4219.48	4124.83	4061.73	2827.96	5805.92	2983.57	5943.54
NeNAfb_NAfa_ME_EU_BEi_EAs_EAf_WAf	28846.55	30795.30	35648.67	8950.99	39618.69	12116.54	39989.92
tNAfb_NAfa_ME_EU_BEi_EAs_EAf_WAf_Khs	8375.44	8319.66	9261.56	4685.41	12220.21	4943.48	12322.71
NeNAfb_NAfa_ME_EU_BEi_EAs_EAf_WAf_Khs	20267.26	16808.13	4738.75	1641.19	48099.05	1011.17	45980.97
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_EAf_Khs_X	11608.74	12078.50	12448.77	7512.76	13894.75	8268.24	13992.55
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_EAf_Khs	36914.84	38227.40	47425.77	16847.89	49561.81	19916.92	49996.20
tAdmxME_NA	193.18649	177.69549	179.73571	18.57877	480.66699	10.83239	434.30671
admixtureME_NA	0.09715	0.09392	0.03641	0.00498	0.19430	0.00238	0.19054
tAdmxEU_NA	133.83634	122.19219	96.03760	15.61688	338.93413	10.79760	294.43174
admixtureEU_NA	0.05178	0.05061	0.04569	0.00386	0.09834	0.00643	0.09970
tAdmxME_NAb	205.71488	178.08999	150.67872	21.66275	483.75571	10.84548	447.95715
admixtureME_NAb	0.05153	0.05243	0.07131	0.00387	0.09750	0.00294	0.09631
tAdmxEU_NAb	140.99831	122.63915	69.74872	18.42806	372.19098	10.66647	329.34676
admixtureEU_NAb	0.05122	0.05113	0.08706	0.00362	0.09710	0.00127	0.09410
tAdmxMENA_Amazigh	463.66981	445.28837	381.65645	198.06215	822.36673	180.65461	768.53310
admixtureMENA_Amazigh	0.05048	0.05124	0.01873	0.00377	0.09643	0.00194	0.09342
tAdmxWAf_Amazigh	423.27768	425.76922	488.14390	30.10023	879.65827	5.97039	833.92498
admixtureWAf_Amazigh	0.05253	0.05251	0.08041	0.00417	0.09798	0.00742	0.09984
tAdmxWAf_Arab	141.95417	127.80308	53.76752	13.13984	388.75657	5.45750	335.08416
admixtureWAf_Arab	0.05622	0.05884	0.08951	0.00556	0.09799	0.00875	0.09990
tAdmxEAf_Amazigh	359.29481	334.09009	237.98858	24.11573	802.52943	7.19778	740.58313
admixtureEAf_Amazigh	0.04713	0.04547	0.01346	0.00263	0.09635	0.00101	0.09364
tAdmxEAf_Arab	137.92699	121.26281	106.03876	12.55817	396.09307	5.89307	322.52776
admixtureEAf_Arab	0.05881	0.06221	0.09419	0.00570	0.09815	0.00932	0.09978
tAdmxBEi_MENAU	457.35230	438.84897	395.88321	238.93522	771.21820	221.46163	740.90035
admixtureBEi_MENAU	0.10000	0.10079	0.11830	0.00693	0.19318	0.00560	0.19149
tAdmxBEi_AMENAU	765.87006	734.18710	668.27068	319.94226	1452.03859	252.41821	1345.21364
admixtureBEi_AMENAU	0.10115	0.10105	0.19276	0.00527	0.19590	0.01123	0.19937
tAdmxXa_Khs	5087.49937	4927.38626	4990.83887	838.11235	10934.16255	720.72508	10293.34
admixtureXa_Khs	0.02588	0.02591	0.04564	0.00214	0.04920	0.00301	0.04975
tAdmxXa_WAf	2729.35146	2663.57027	3219.85264	808.02582	5172.13206	721.94834	4955.94096
admixtureXa_WAf	0.03232	0.03486	0.04668	0.00499	0.04941	0.00686	0.04987

Page 4: Three tables with other statistics to asses quality of ABC-DL results for Model D4 (Factor2, KL-divergence and Spearman correlation)

Tabke S1: FACTOR 2

	factor2_resul	fcator2_prior
missclassification	0.739	0.735
migrationNAfa_NAfb	0.742	0.751
migrationNAfb_NAfa	0.74	0.764
migrationME_NAfb	0.741	0.749
migrationNAfb_to_ME	0.738	0.744
migrationME_NAfa	0.758	0.766
migrationNAfa_to_ME	0.743	0.738
migrationEU_ME	0.755	0.763
migrationEU_NAfb	0.709	0.726
migrationNAfb_to_EU	0.741	0.747
migrationEU_NAfa	0.76	0.768
migrationNAfa_to_EU	0.751	0.759
migrationEAs_EU	0.742	0.747
migrationEU_to_EAs	0.771	0.751
migrationEAf_NAfb	0.777	0.743
migrationNAfb_to_EAf	0.742	0.746
migrationEAf_NAfa	0.734	0.75
migrationNAfa_to_EAf	0.747	0.752
migrationEAf_to_ME	0.723	0.743
migrationME_to_EAf	0.776	0.735
migrationEAf_to_WAf	0.79	0.761
migrationWAf_to_EAf	0.737	0.745
migrationWAf_NAfb	0.797	0.748
migrationNAfb_WAf	0.756	0.732
migrationWAf_NAfa	0.731	0.738
migrationNAfa_WAf	0.733	0.727
migrationKhs_to_WAf	0.807	0.734
migrationKhs_to_EAf	0.762	0.741
NeKhs	0.908	0.813
NeWAf	0.98	0.808
NeEAf	0.849	0.849
NeNAfb	0.981	0.763
NeAfa	0.86	0.753
NeME	0.841	0.756
NeEU	0.844	0.775
NeEAs	0.931	0.769
NeBEi	0.816	0.811
NeXa	0.787	0.787
tNAfa_ME	0.972	0.913
NeNAfa_ME	0.76	0.766
tNAfa_ME_EU	0.974	0.971

Table S2: Kullback-Leiber Divergence

	KL-divergence	p-value
missclassification	0.00986	7.25E-01
migrationNAfa_NAfb	13.34251	0.00E+00
migrationNAfb_NAfa	12.91307	0.00E+00
migrationME_NAfb	13.29591	0.00E+00
migrationNAfb_to_ME	13.22645	0.00E+00
migrationME_NAfa	12.99378	0.00E+00
migrationNAfa_to_ME	0.00806	8.79E-01
migrationEU_ME	13.28899	0.00E+00
migrationEU_NAfb	12.96629	0.00E+00
migrationNAfb_to_EU	13.17653	0.00E+00
migrationEU_NAfa	12.72406	0.00E+00
migrationNAfa_to_EU	13.21237	0.00E+00
migrationEAs_EU	0.04907	0.00E+00
migrationEU_to_EAs	13.25301	0.00E+00
migrationEAf_NAfb	12.83731	0.00E+00
migrationNAfb_to_EAf	12.81136	0.00E+00
migrationEAf_NAfa	12.22954	0.00E+00
migrationNAfa_to_EAf	12.51879	0.00E+00
migrationEAf_to_ME	0.58272	0.00E+00
migrationME_to_EAf	12.42863	0.00E+00
migrationEAf_to_WAf	12.65425	0.00E+00
migrationWAf_to_EAf	12.74265	0.00E+00
migrationWAf_NAfb	12.65758	0.00E+00
migrationNAfb_WAf	12.30569	0.00E+00
migrationWAf_NAfa	0.29903	0.00E+00
migrationNAfa_WAf	12.57933	0.00E+00
migrationKhs_to_WAf	12.64947	0.00E+00
migrationKhs_to_EAf	12.17826	0.00E+00
NeKhs	14.50866	0.00E+00
NeWAf	14.50866	0.00E+00
NeEAf	14.50866	0.00E+00
NeNAfb	14.50866	0.00E+00
NeAfa	14.50866	0.00E+00
NeME	6.21461	0.00E+00
NeEU	14.50866	0.00E+00
NeEAs	14.50866	0.00E+00
NeBEi	9.55802	0.00E+00
NeXa	14.50866	0.00E+00
tNAfa_ME	14.50866	0.00E+00
NeNAfa_ME	4.60517	0.00E+00
tNAfa_ME_EU	14.50866	0.00E+00

NeNAfa_ME_EU	0.89	0.738
tNAfb_NAfa_ME_EU	0.999	0.989
NeNAfb_NAfa_ME_EU	0.742	0.747
tNAfb_NAfa_ME_EU_BEi	0.989	0.981
NeNAfb_NAfa_ME_EU_BEi	0.758	0.757
tNAfb_NAfa_ME_EU_BEi_EAs	0.997	0.978
NeNAfb_NAfa_ME_EU_BEi_EAs	0.816	0.761
tNAfb_NAfa_ME_EU_EAs_EAf	0.998	0.988
NeNAfb_NAfa_ME_EU_EAs_EAf	0.786	0.771
tNAfb_NAfa_ME_EU_BEi_EAs_WAf	0.997	0.992
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf	0.803	0.756
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs	0.995	0.978
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs	0.727	0.754
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_Xa	0.986	0.984
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_X	0.848	0.759
tAdmxME_NA	0.659	0.623
admixtureME_NA	0.752	0.748
tAdmxEU_NA	0.652	0.628
admixtureEU_NA	0.767	0.766
tAdmxME_NAb	0.632	0.62
admixtureME_NAb	0.73	0.731
tAdmxME_NAb,1	0.631	0.615
admixtureEU_NAb	0.766	0.771
tAdmxMENA_Amazigh	0.967	0.946
admixtureMENA_Amazigh	0.726	0.728
tAdmxWAf_Amazigh	0.718	0.698
admixtureWAf_Amazigh	0.744	0.748
tAdmxWAf_Arab	0.665	0.646
admixtureWAf_Arab	0.705	0.732
tAdmxEAf_Amazigh	0.706	0.701
admixtureEAf_Amazigh	0.738	0.743
tAdmxEAf_Arab	0.629	0.606
admixtureEAf_Arab	0.763	0.778
tAdmxBEi_MENAU	0.99	0.974
admixtureBEi_MENAU	0.749	0.749
tAdmxBEi_AMENAU	0.987	0.987
admixtureBEi_AMENAU	0.751	0.748
tAdmxXa_Khs	0.697	0.703
admixtureXa_Khs	0.733	0.735
tAdmxXa_WAf	0.828	0.834
admixtureXa_WAf	0.727	0.748

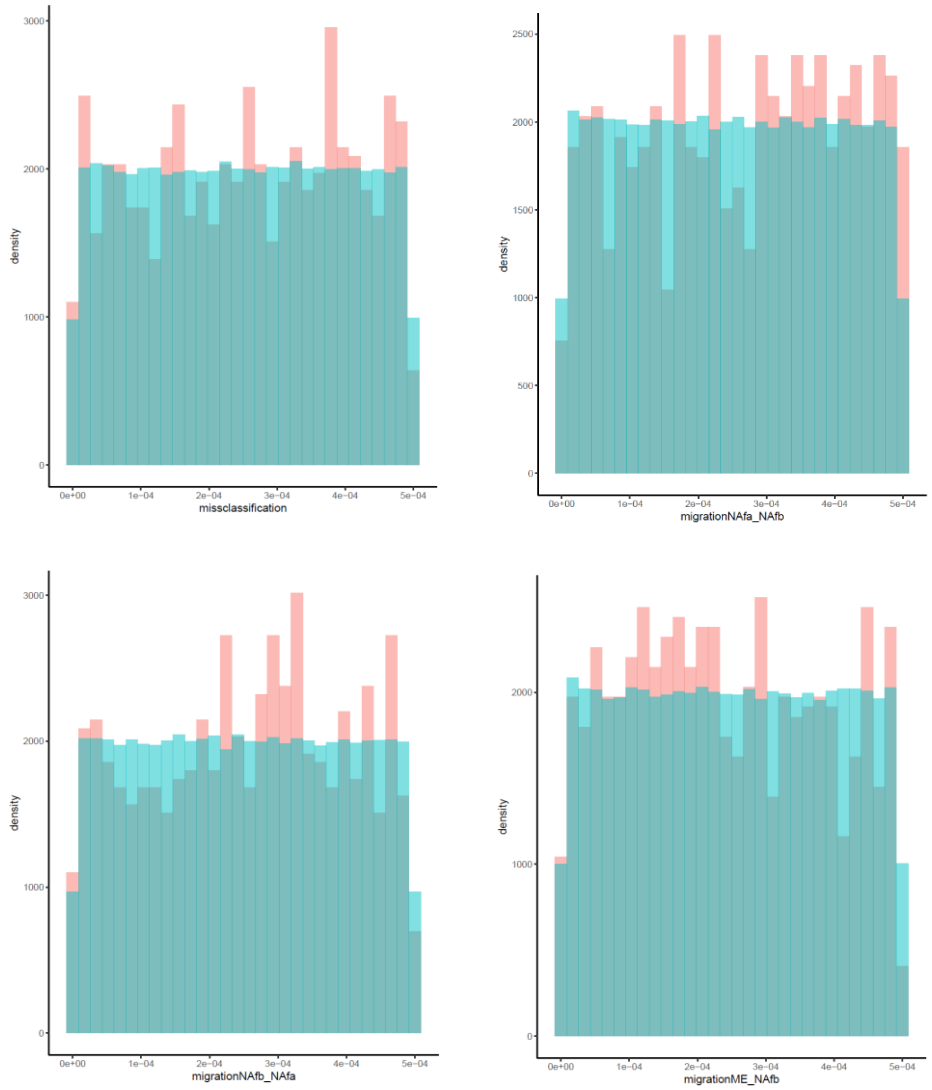
NeNAfa_ME_EU	14.50866	0.00E+00
tNAfb_NAfa_ME_EU	11.14939	0.00E+00
NeNAfb_NAfa_ME_EU	5.52146	0.00E+00
tNAfb_NAfa_ME_EU_BEi	14.50866	0.00E+00
NeNAfb_NAfa_ME_EU_BEi	5.52146	0.00E+00
tNAfb_NAfa_ME_EU_BEi_EAs	14.50866	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs	14.50866	0.00E+00
tNAfb_NAfa_ME_EU_EAs_EAf	13.52076	0.00E+00
NeNAfb_NAfa_ME_EU_EAs_EAf	5.80914	0.00E+00
tNAfb_NAfa_ME_EU_BEi_EAs_WAf	14.50866	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf	14.50866	0.00E+00
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs	14.50866	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs	3.27017	0.00E+00
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_Xa	6.14300	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_	14.50866	0.00E+00
tAdmxME_NA	2.00248	0.00E+00
admixtureME_NA	11.93900	0.00E+00
tAdmxEU_NA	1.66073	0.00E+00
admixtureEU_NA	12.46156	0.00E+00
tAdmxME_NAb	12.58715	0.00E+00
admixtureME_NAb	13.05318	0.00E+00
tAdmxME_NAb,1	1.93102	0.00E+00
admixtureEU_NAb	12.74685	0.00E+00
tAdmxMENA_Amazigh	14.50866	0.00E+00
admixtureMENA_Amazigh	12.92648	0.00E+00
tAdmxWAf_Amazigh	10.98965	0.00E+00
admixtureWAf_Amazigh	12.10416	0.00E+00
tAdmxWAf_Arab	1.76026	0.00E+00
admixtureWAf_Arab	12.87342	0.00E+00
tAdmxEAf_Amazigh	2.74887	0.00E+00
admixtureEAf_Amazigh	13.35127	0.00E+00
tAdmxEAf_Arab	0.13667	0.00E+00
admixtureEAf_Arab	12.74519	0.00E+00
tAdmxBEi_MENAU	14.50866	0.00E+00
admixtureBEi_MENAU	12.48615	0.00E+00
tAdmxBEi_AMENAU	14.50866	0.00E+00
admixtureBEi_AMENAU	13.21711	0.00E+00
tAdmxXa_Khs	14.50866	0.00E+00
admixtureXa_Khs	13.00372	0.00E+00
tAdmxXa_WAf	14.50866	0.00E+00
admixtureXa_WAf	13.01183	0.00E+00

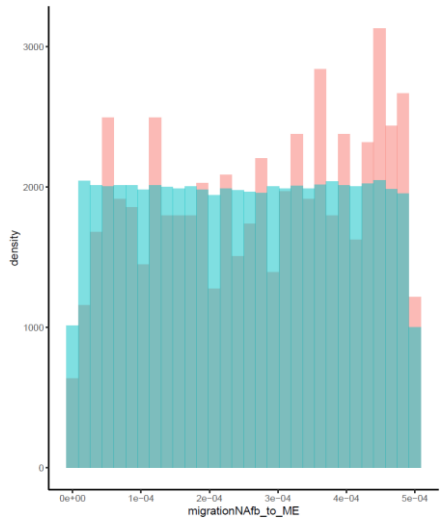
Table S3: Spearman Correlation

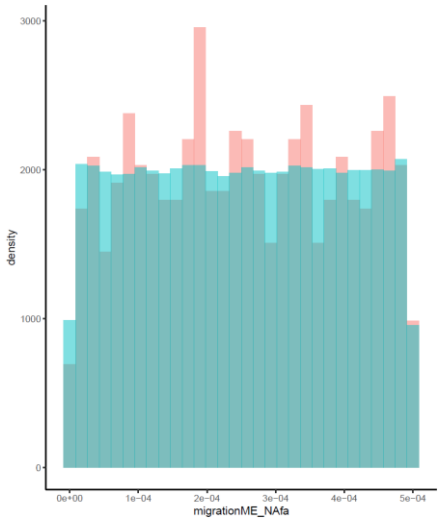
	Spearman Correlatio	p_value
missclassification	-0.00094	6.90E-01
migrationNAfa_NAfb	0.00076	7.48E-01
migrationNAfb_NAfa	0.48584	0.00E+00
migrationME_NAfb	0.55713	0.00E+00
migrationNAfb_to_ME	0.60538	0.00E+00
migrationME_NAfa	0.53837	0.00E+00
migrationNAfa_to_ME	0.01471	4.31E-10
migrationEU_ME	0.00609	9.73E-03
migrationEU_NAfb	0.47884	0.00E+00
migrationNAfb_to_EU	0.47711	0.00E+00
migrationEU_NAfa	0.73946	0.00E+00
migrationNAfa_to_EU	0.61551	0.00E+00
migrationEAs_EU	0.06104	3.66E-148
migrationEU_to_EAs	0.01348	1.08E-08
migrationEAf_NAfb	0.47263	0.00E+00
migrationNAfb_to_EAf	0.73357	0.00E+00
migrationEAf_NAfa	0.84784	0.00E+00
migrationNAfa_to_EAf	0.86652	0.00E+00
migrationEAf_to_ME	0.42856	0.00E+00
migrationME_to_EAf	0.00621	8.42E-03
migrationEAf_to_WAf	0.76715	0.00E+00
migrationWAf_to_EAf	0.79349	0.00E+00
migrationWAf_NAfb	0.77883	0.00E+00
migrationNAfb_WAf	0.88289	0.00E+00
migrationWAf_NAfa	0.56436	0.00E+00
migrationNAfa_WAf	0.13540	0.00E+00
migrationKhs_to_WAf	0.78416	0.00E+00
migrationKhs_to_EAf	0.93406	0.00E+00
NeKhs	0.88653	0.00E+00
NeWAf	0.79228	0.00E+00
NeEAf	0.44903	0.00E+00
NeNAfb	-0.00938	6.89E-05
NeAfa	0.88761	0.00E+00
NeME	0.89504	0.00E+00
NeEU	0.84941	0.00E+00
NeEAs	0.88856	0.00E+00
NeBEi	-0.00113	6.31E-01
NeXa	-0.01171	6.82E-07
tNAfa_ME	0.95043	0.00E+00
NeNAfa_ME	0.86344	0.00E+00
tNAfa_ME_EU	0.83965	0.00E+00
NeNAfa_ME_EU	0.83288	0.00E+00
tNAfb_NAfa_ME_EU	0.79134	0.00E+00
NeNAfb_NAfa_ME_EU	-0.00207	3.80E-01
tNAfb_NAfa_ME_EU_BEi	0.44146	0.00E+00
NeNAfb_NAfa_ME_EU_BEi	0.48808	0.00E+00
tNAfb_NAfa_ME_EU_BEi_EAs	0.92312	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs	0.76390	0.00E+00
tNAfb_NAfa_ME_EU_EAs_EAf	0.23626	0.00E+00
NeNAfb_NAfa_ME_EU_EAs_EAf	-0.04325	2.96E-75
tNAfb_NAfa_ME_EU_BEi_EAs_WAf	0.93613	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf	0.30877	0.00E+00
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs	0.75938	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs	0.61901	0.00E+00
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_Xa	0.42444	0.00E+00
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_	-0.32127	0.00E+00
tAdmxME_NA	0.50783	0.00E+00
admixtureME_NA	0.88161	0.00E+00
tAdmxEU_NA	0.88157	0.00E+00
admixtureEU_NA	0.75212	0.00E+00
tAdmxME_NAb	0.50235	0.00E+00
admixtureME_NAb	0.28765	0.00E+00
tAdmxME_NAb,1	0.86747	0.00E+00
admixtureEU_NAb	0.80053	0.00E+00
tAdmxMENA_Amazigh	0.65711	0.00E+00
admixtureMENA_Amazigh	0.70213	0.00E+00
tAdmxWAf_Amazigh	0.74899	0.00E+00
admixtureWAf_Amazigh	0.00688	3.49E-03
tAdmxWAf_Arab	0.81563	0.00E+00
admixtureWAf_Arab	0.43438	0.00E+00
tAdmxEAf_Amazigh	0.72543	0.00E+00
admixtureEAf_Amazigh	0.46706	0.00E+00
tAdmxEAf_Arab	0.24972	0.00E+00
admixtureEAf_Arab	-0.01171	6.76E-07
tAdmxBEi_MENAU	0.47200	0.00E+00
admixtureBEi_MENAU	0.65962	0.00E+00
tAdmxBEi_AMENAU	0.34778	0.00E+00
admixtureBEi_AMENAU	0.85832	0.00E+00
tAdmxXa_Khs	-0.00050	8.32E-01
admixtureXa_Khs	-0.01602	1.07E-11
tAdmxXa_WAf	0.76796	0.00E+00
admixtureXa_WAf	0.58245	0.00E+00

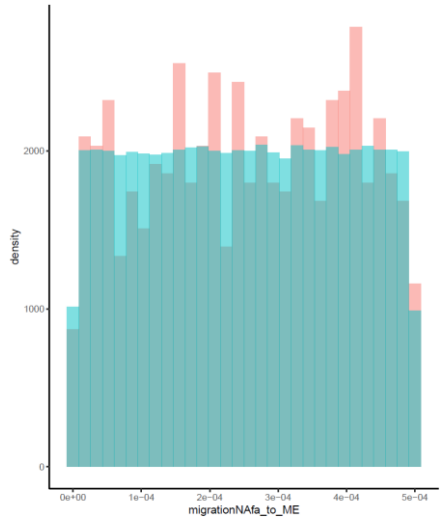
Additional file 3:

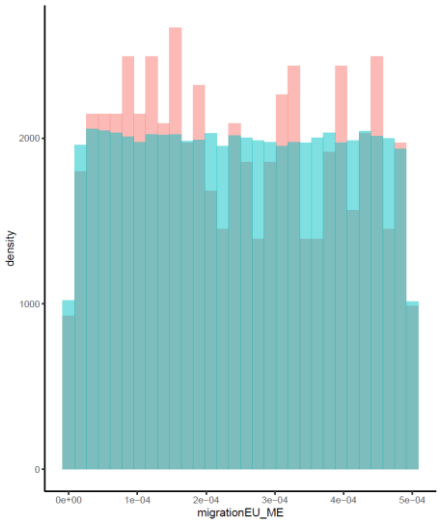
Histograms with the posterior (red) versus prior distributions (blue) of all parameters for the best model in ABC-DL analysis (Model D4).

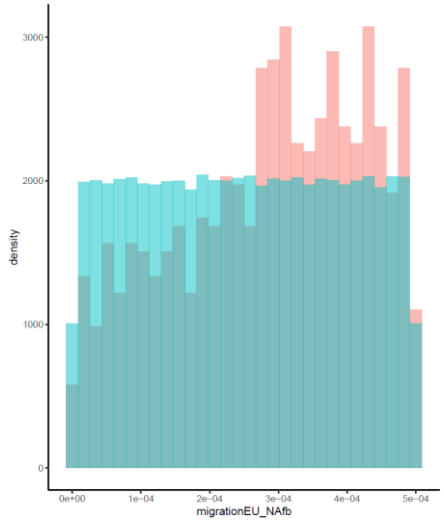


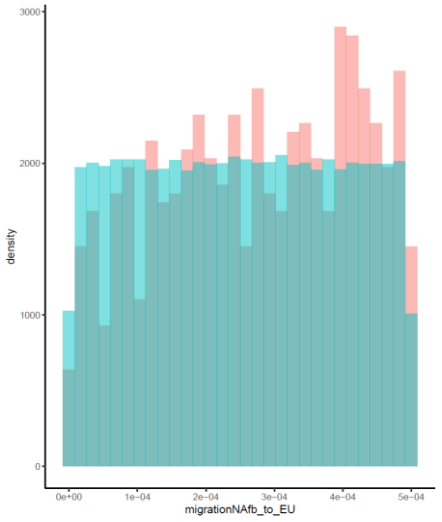

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfb_to_ME

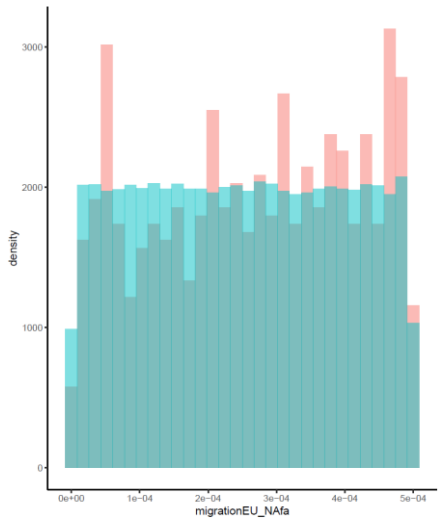

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationME_NAfa

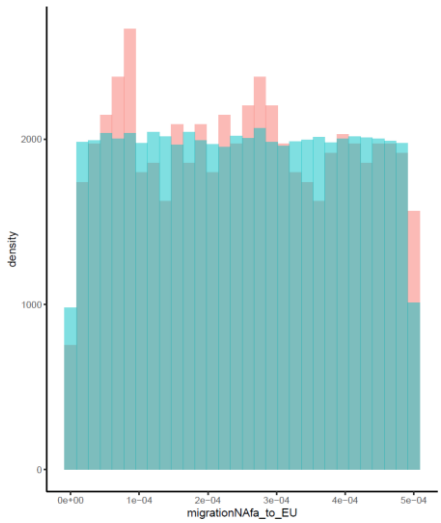

density
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfa_to_ME

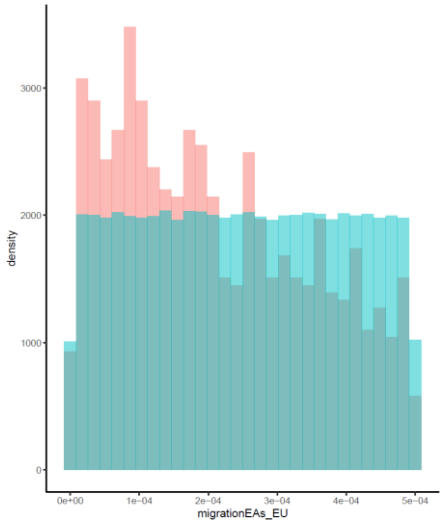

density
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEU_ME

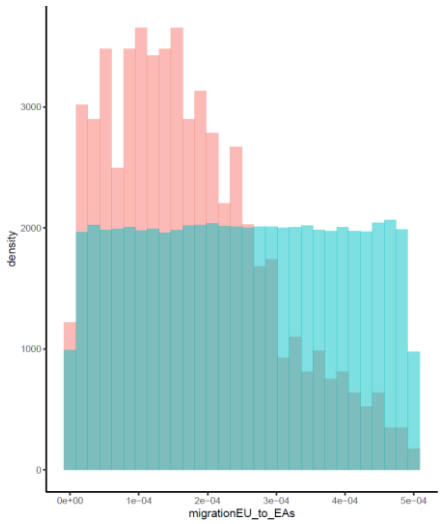

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEU_NAfb

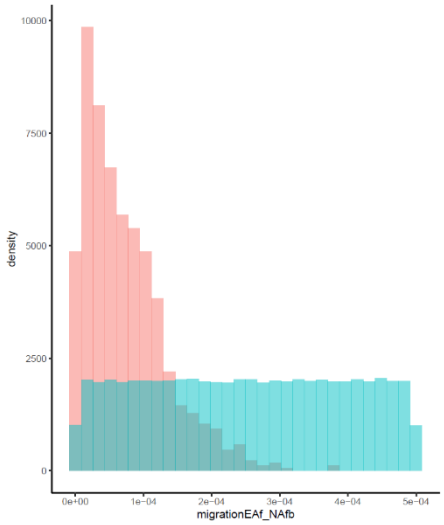

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfb_to_EU

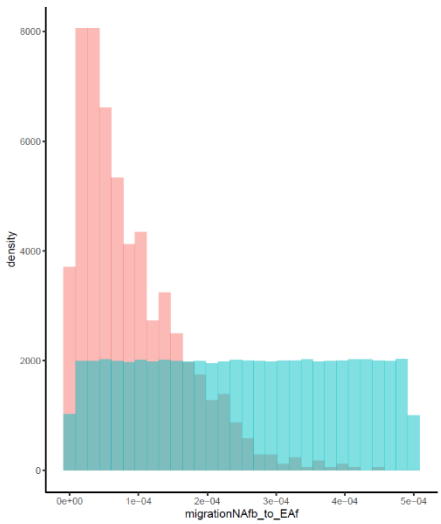

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEU_NAfa

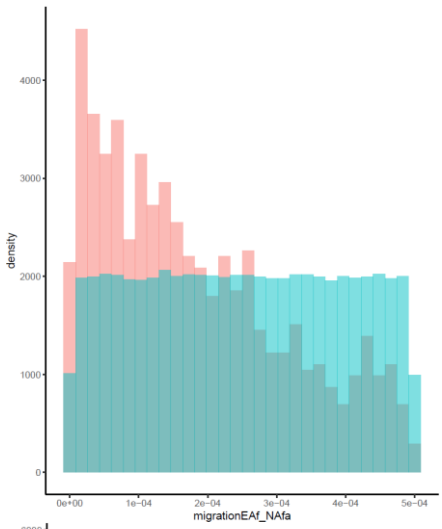

density
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfa_to_EU

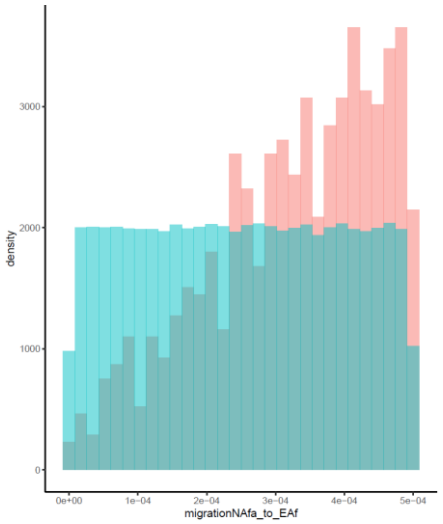

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEAs_EU

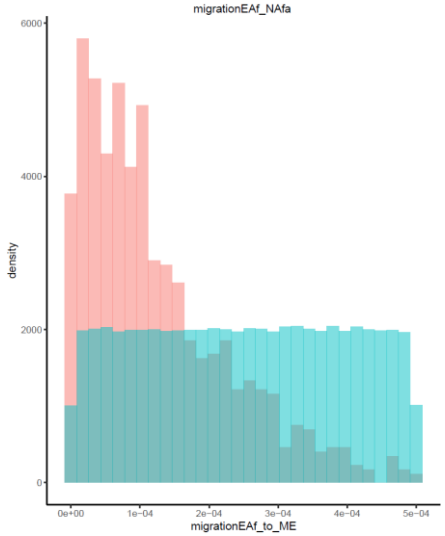

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEU_to_EAs

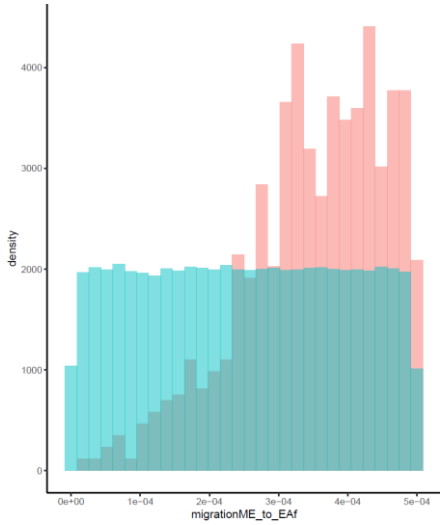

density
10000
7500
5000
2500
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEAf_NAfb

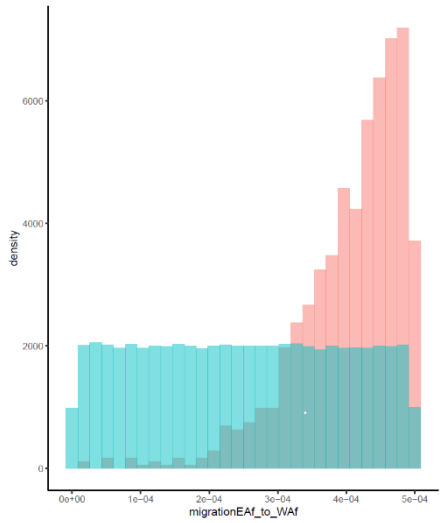

density
8000
6000
4000
2000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfb_to_EAf

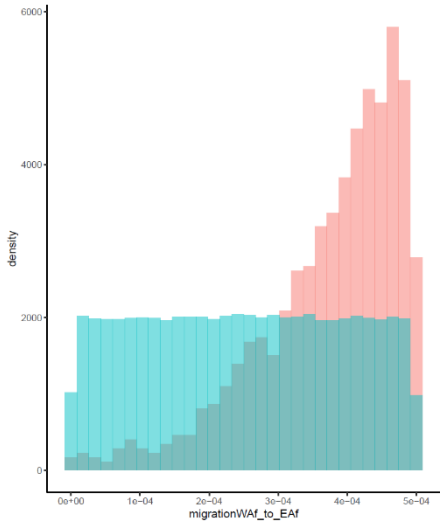

density
4000
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEAf_NAfa

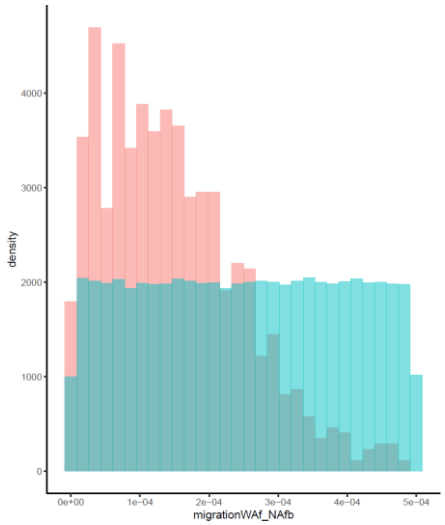

density
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfa_to_EAf

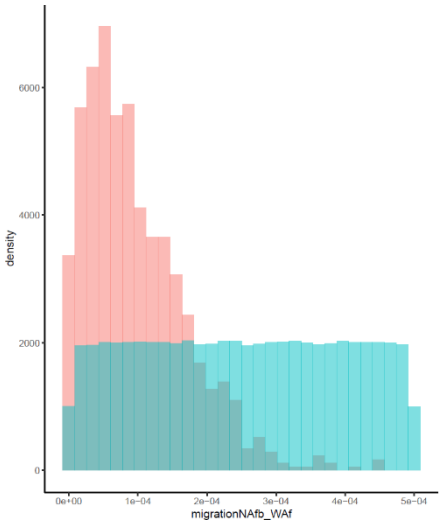

density
6000
4000
2000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEAf_to_ME

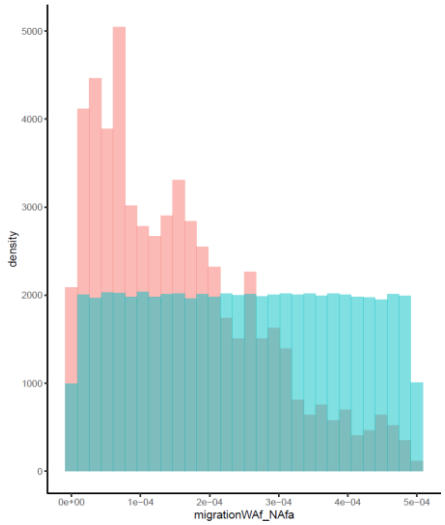

density
4000
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationME_to_EAf

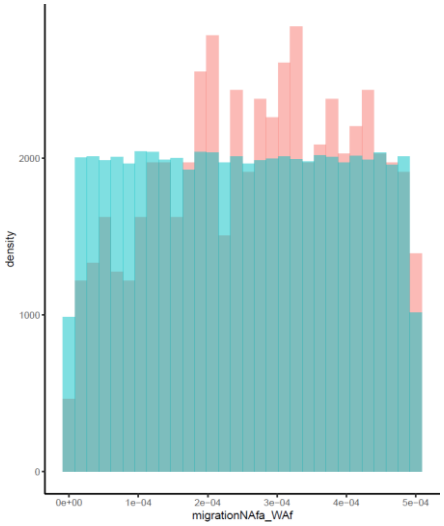

density
6000
4000
2000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationEAf_to_WAf

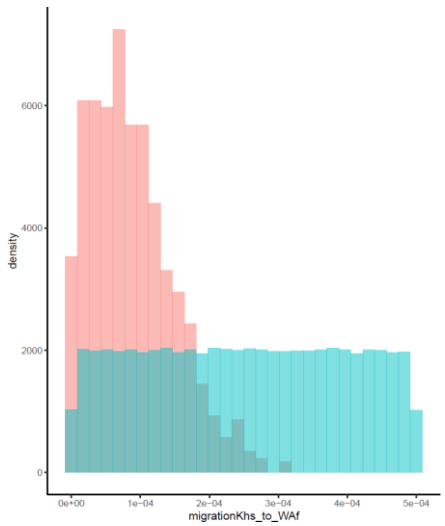

density
6000
4000
2000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationWAf_to_EAf

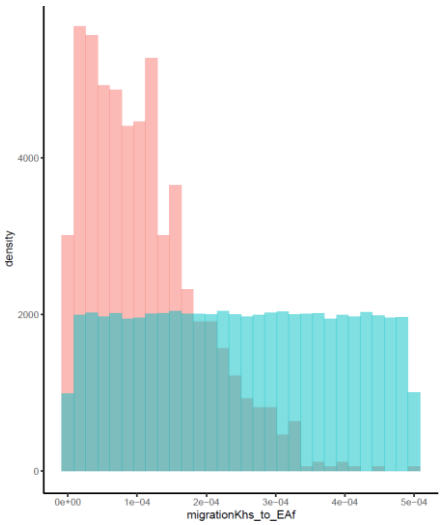

density
4000
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationWAf_NAfb

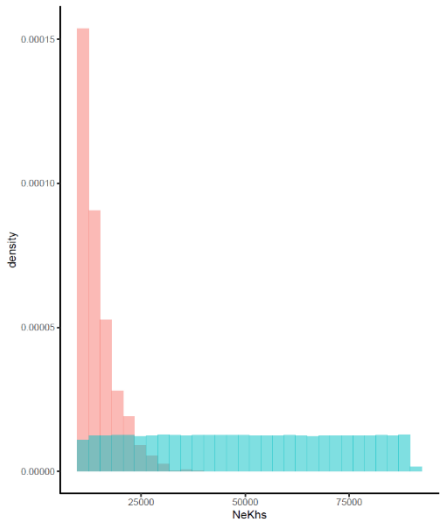

density
6000
4000
2000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfb_WAf

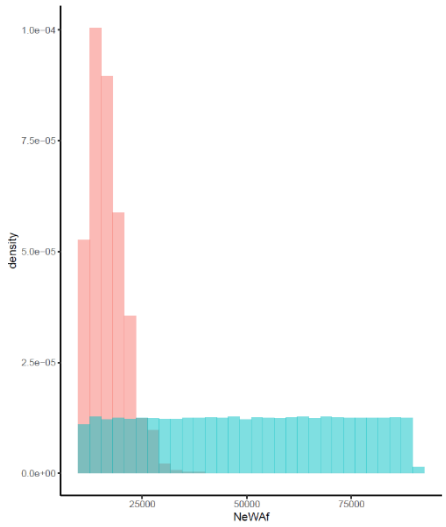

density
5000
4000
3000
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationWAf_NAfa

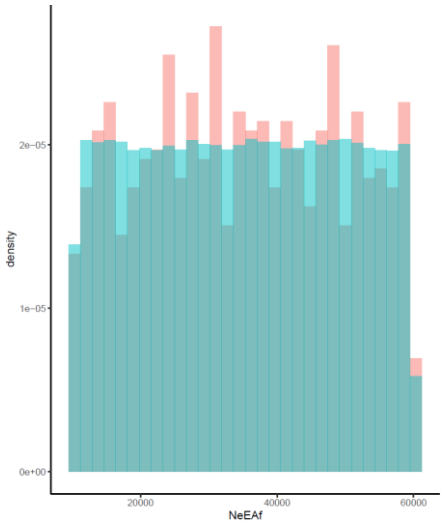

density
2000
1000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationNAfa_WAf

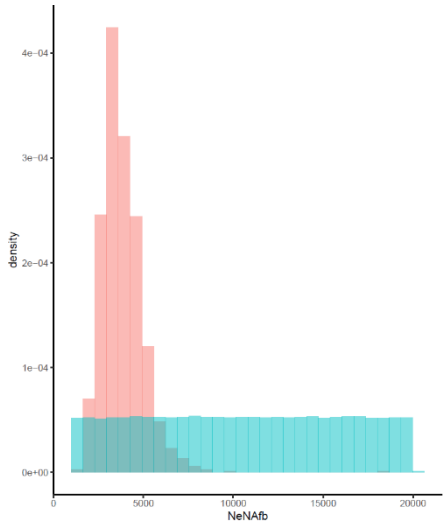

density
6000
4000
2000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationKhs_to_WAf

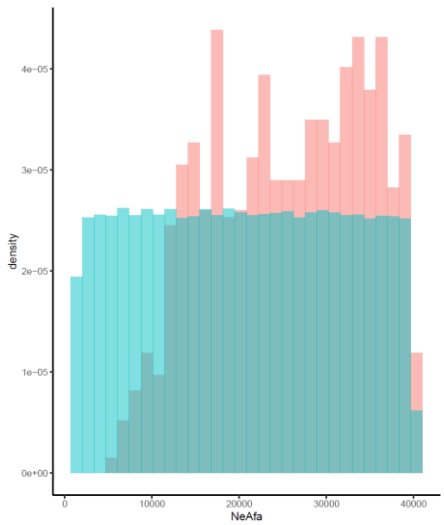

density
4000
2000
0
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
migrationKhs_to_EAf

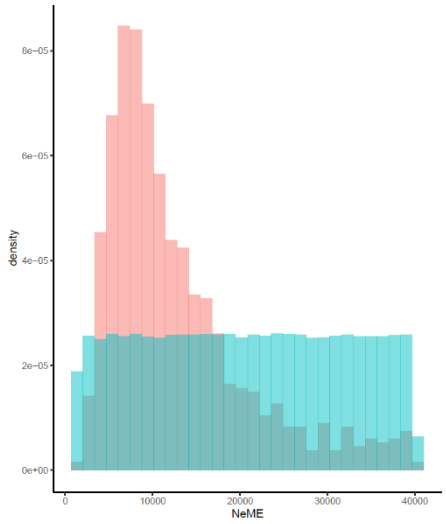

density
0.00015
0.00010
0.00005
0.00000
25000
50000
75000
NeKhs

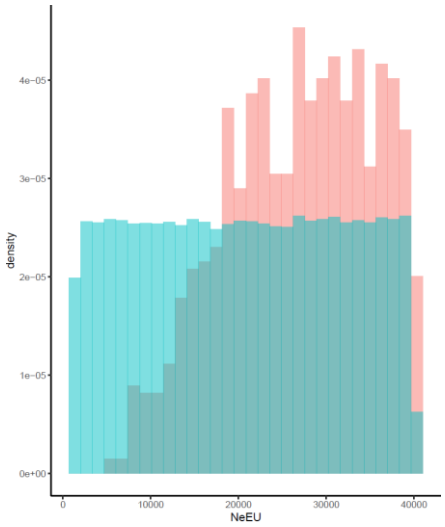

density
1.0e−04
7.5e−05
5.0e−05
2.5e−05
0.0e+00
25000
50000
75000
NeWAf

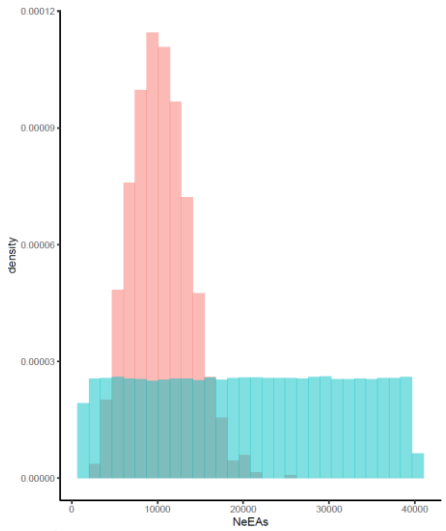

density
2e−05
1e−05
0e+00
20000
40000
60000
NeEAf

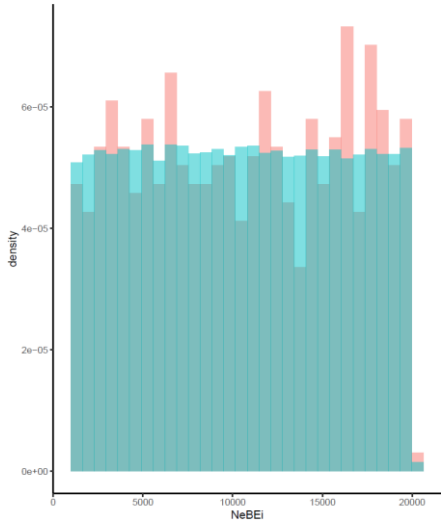

density
4e−04
3e−04
2e−04
1e−04
0e+00
0
5000
10000
15000
20000
NeNAfb

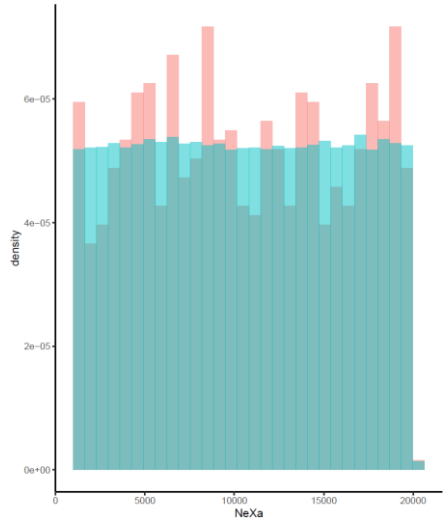

density
4e−05
3e−05
2e−05
1e−05
0e+00
0
10000
20000
30000
40000
NeAfa

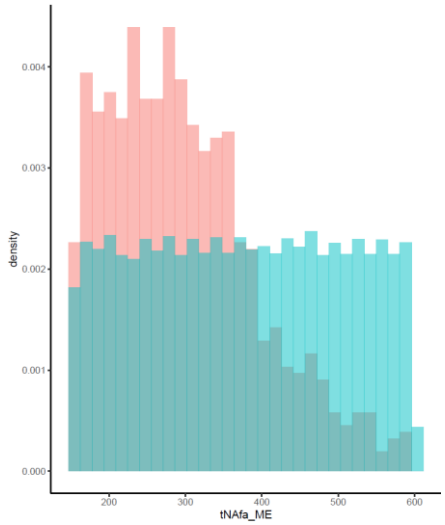

density
8e−05
6e−05
4e−05
2e−05
0e+00
0
10000
20000
30000
40000
NeME

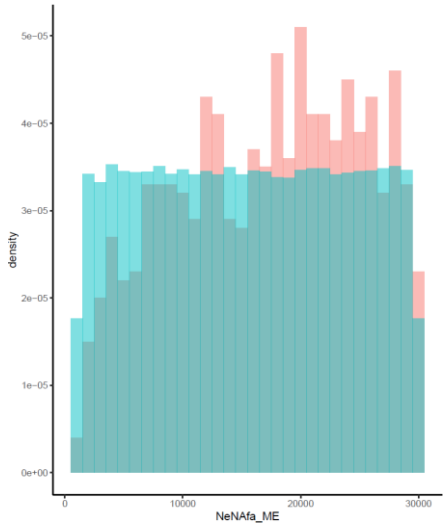

density
4e−05
3e−05
2e−05
1e−05
0e+00
0
10000
20000
30000
40000
NeEU

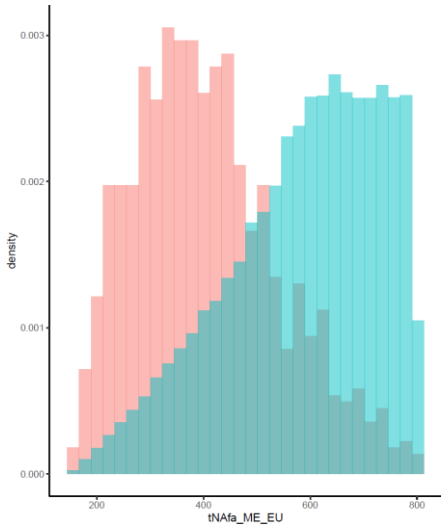

density
0.00012
0.00009
0.00006
0.00003
0.00000
0
10000
20000
30000
40000
NeEAs

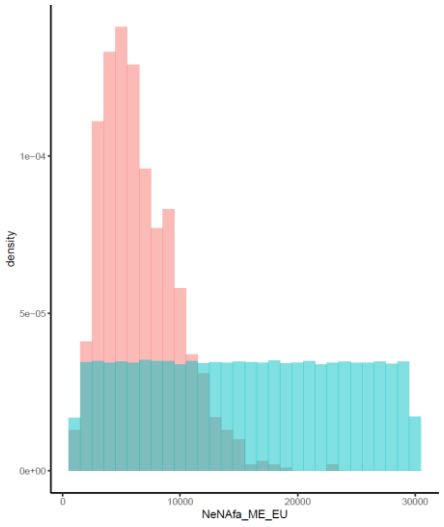

density
6e−05
4e−05
2e−05
0e+00
0
5000
10000
15000
20000
NeBEi

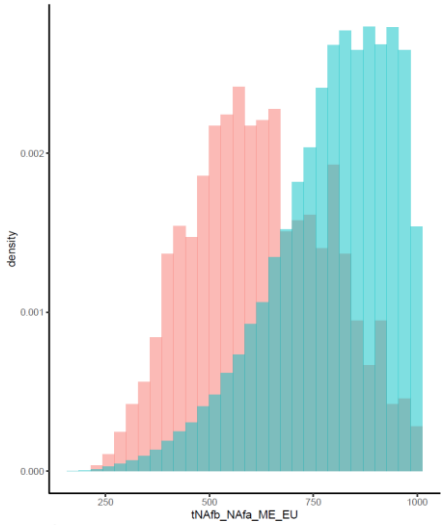

density
6e−05
4e−05
2e−05
0e+00
0
5000
10000
15000
20000
NeXa

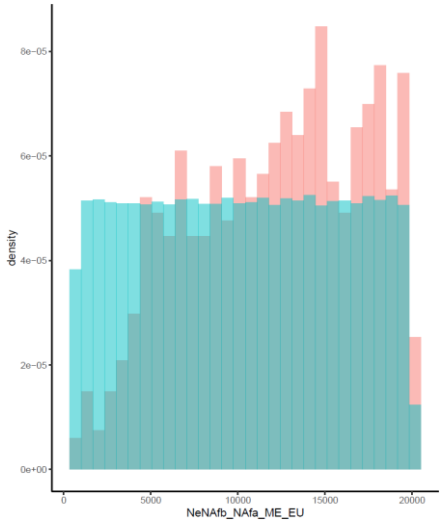

density
0.004
0.003
0.002
0.001
0.000
200
300
400
500
600
tNAfa_ME

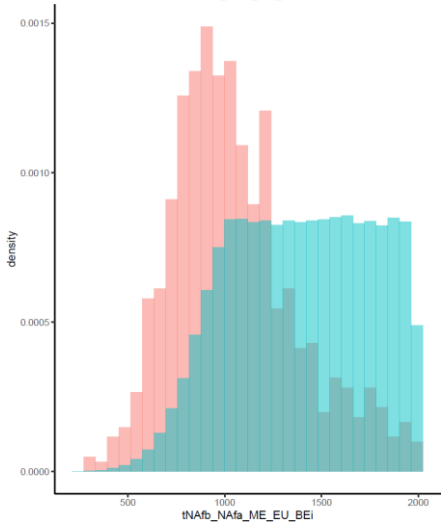

density
5e−05
4e−05
3e−05
2e−05
1e−05
0e+00
0
10000
20000
30000
NeNAfa_ME

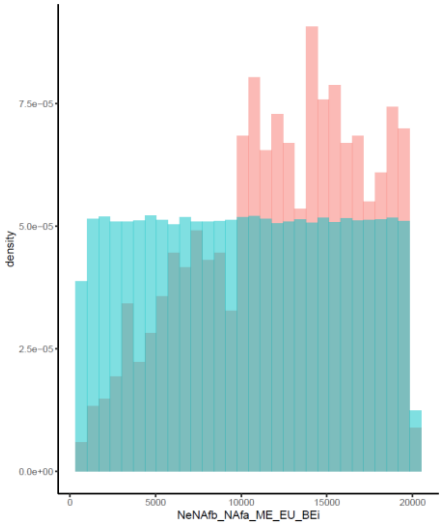

density
0.003
0.002
0.001
0.000
200
400
600
800
tNAfa_ME_EU

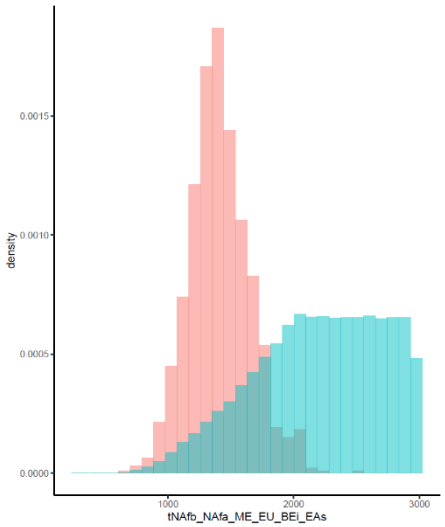

density
1e−04
5e−05
0e+00
0
10000
20000
30000
NeNAfa_ME_EU

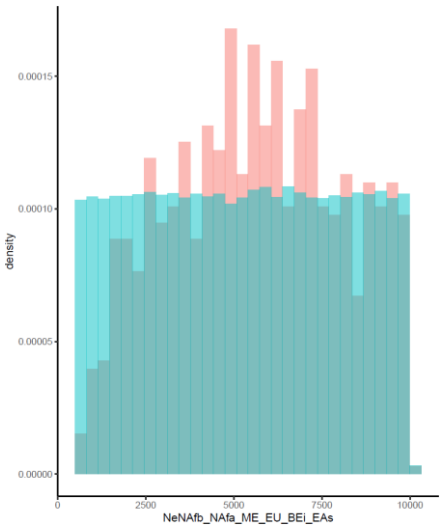

density
0.002
0.001
0.000
250
500
750
1000
tNAfb_NAfa_ME_EU

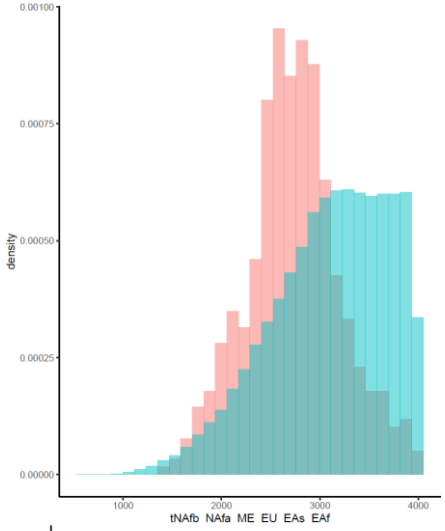

density
8e−05
6e−05
4e−05
2e−05
0e+00
0
5000
10000
15000
20000
NeNAfb_NAfa_ME_EU

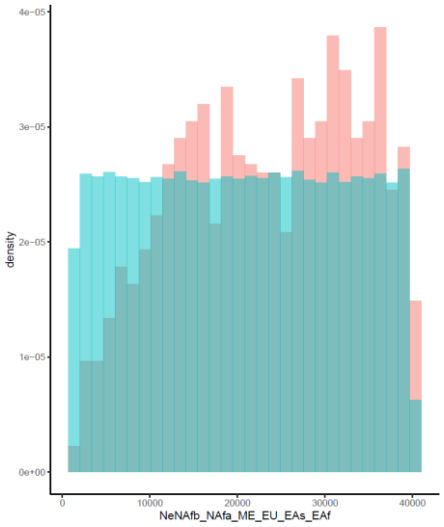

density
0.0015
0.0010
0.0005
0.0000
500
1000
1500
2000
tNAfb_NAfa_ME_EU_BEi

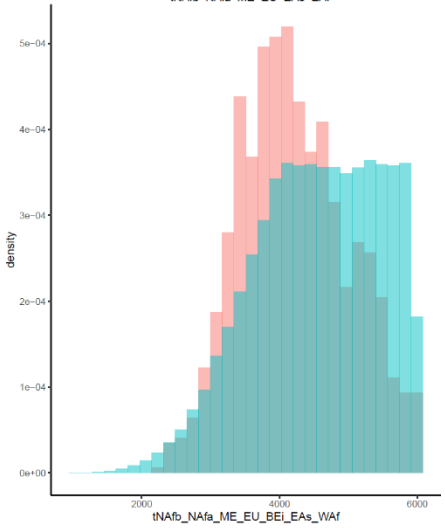

density
7.5e−05
5.0e−05
2.5e−05
0.0e+00
0
5000
10000
15000
20000
NeNAfb_NAfa_ME_EU_BEi

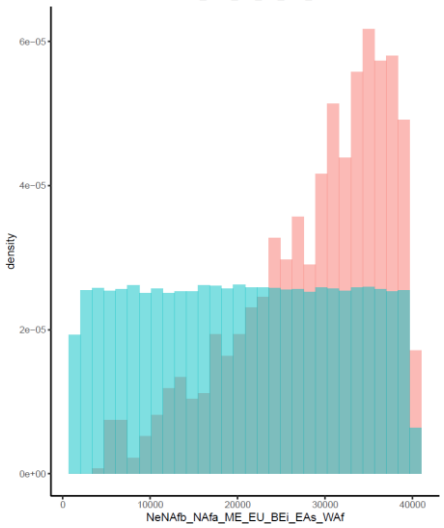

density
0.0015
0.0010
0.0005
0.0000
1000
2000
3000
tNAfb_NAfa_ME_EU_BEi_EAs

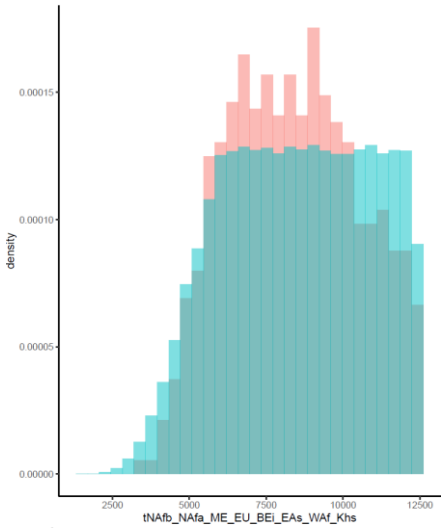

density
0.00015
0.00010
0.00005
0.00000
0
2500
5000
7500
10000
NeNAfb_NAfa_ME_EU_BEi_EAs

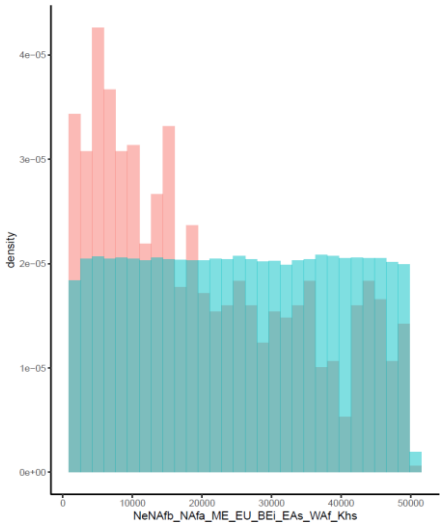

density
0.00100
0.00075
0.00050
0.00025
0.00000
1000
2000
3000
4000
tNAfb_NAfa_ME_EU_EAs_EAf

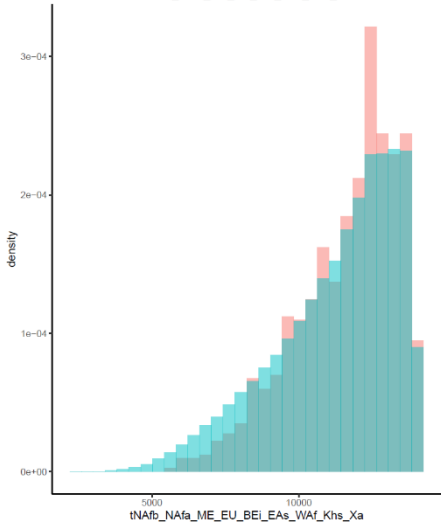

density
4e−05
3e−05
2e−05
1e−05
0e+00
0
10000
20000
30000
40000
NeNAfb_NAfa_ME_EU_EAs_EAf

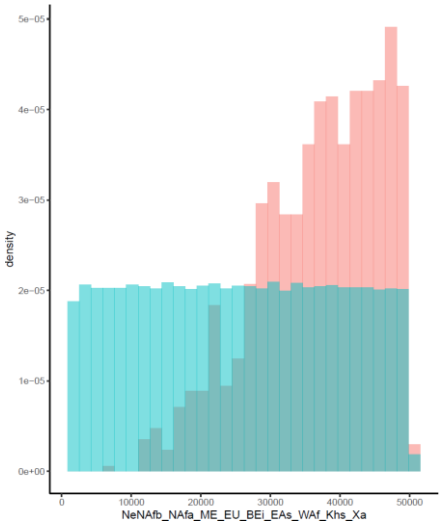

density
5e−04
4e−04
3e−04
2e−04
1e−04
0e+00
2000
4000
6000
tNAfb_NAfa_ME_EU_BEi_EAs_WAf

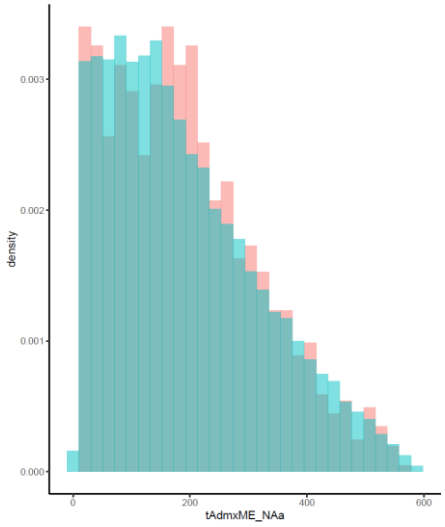

density
6e−05
4e−05
2e−05
0e+00
0
10000
20000
30000
40000
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf

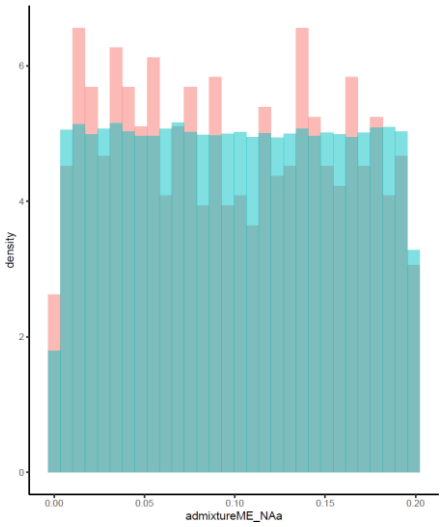

density
0.00015
0.00010
0.00005
0.00000
2500
5000
7500
10000
12500
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs

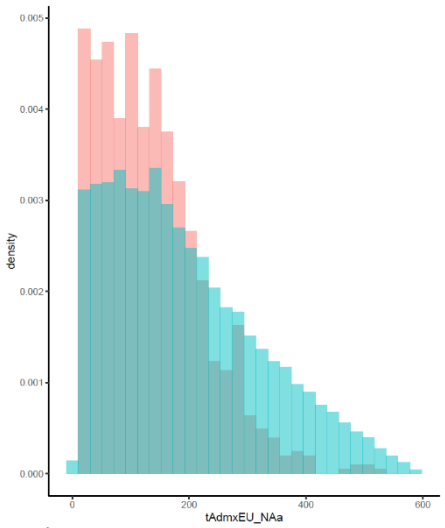

density
4e−05
3e−05
2e−05
1e−05
0e+00
0
10000
20000
30000
40000
50000
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs

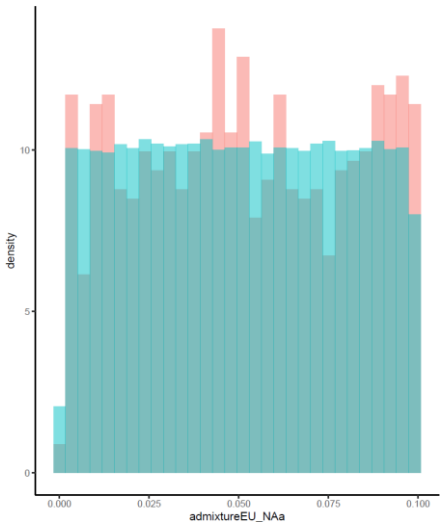

density
3e−04
2e−04
1e−04
0e+00
5000
10000
tNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_Xa

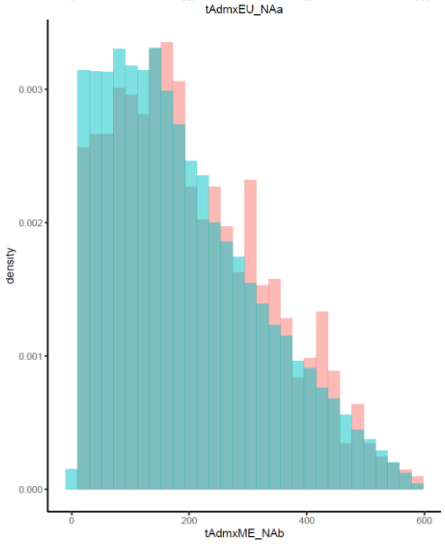

density
5e−05
4e−05
3e−05
2e−05
1e−05
0e+00
0
10000
20000
30000
40000
50000
NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_Khs_Xa

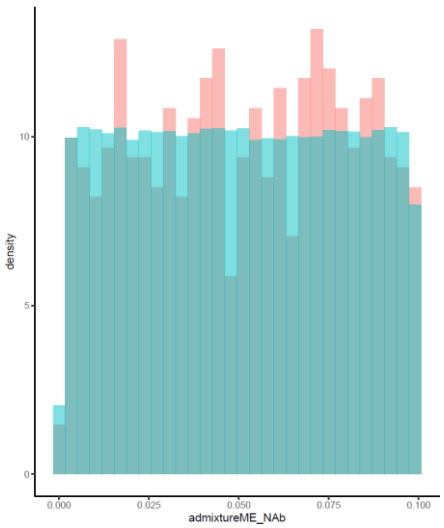

density
0.003
0.002
0.001
0.000
0
200
400
600
tAdmxME_NAa

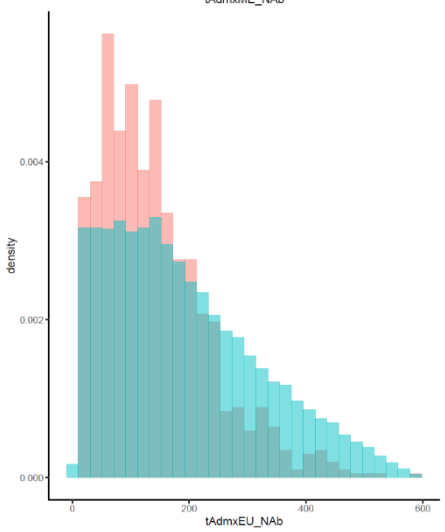

density
6
4
2
0
0.00
0.05
0.10
0.15
0.20
admixtureME_NAa

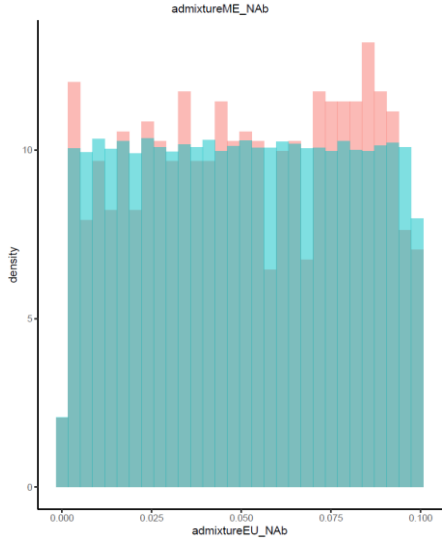

density
0.005
0.004
0.003
0.002
0.001
0.000
0
200
400
600
tAdmxEU_NAa

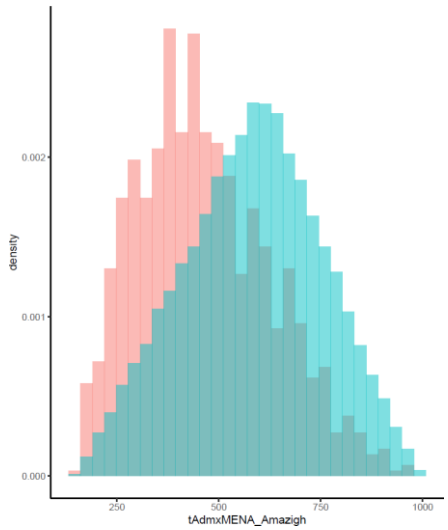

density
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureEU_NAa

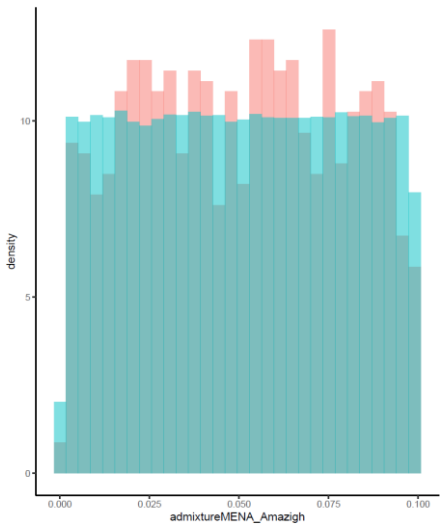

density
0.003
0.002
0.001
0.000
0
200
400
600
tAdmxME_NAb

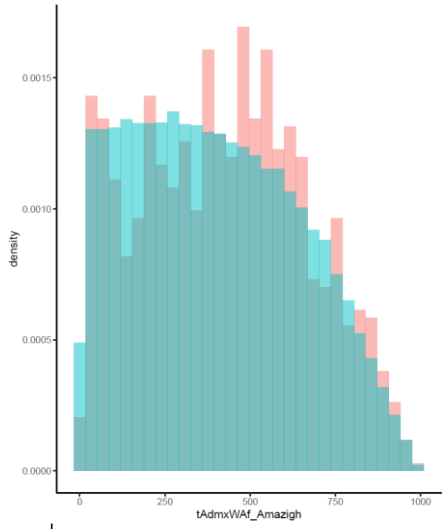

density
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureME_NAb

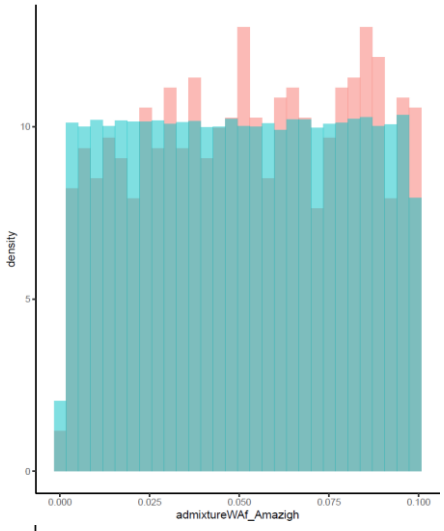

density
0.004
0.002
0.000
0
200
400
600
tAdmxEU_NAb

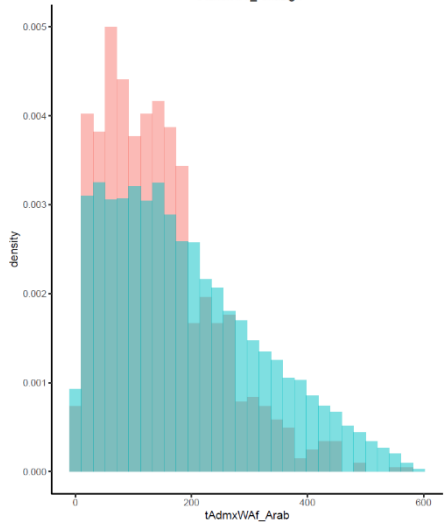

density
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureEU_NAb

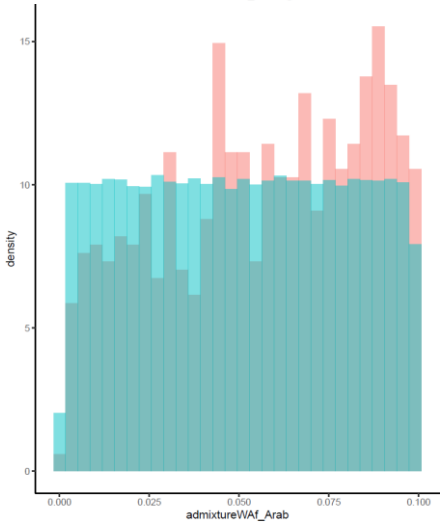

density
0.002
0.001
0.000
250
500
750
1000
tAdmxMENA_Amazigh

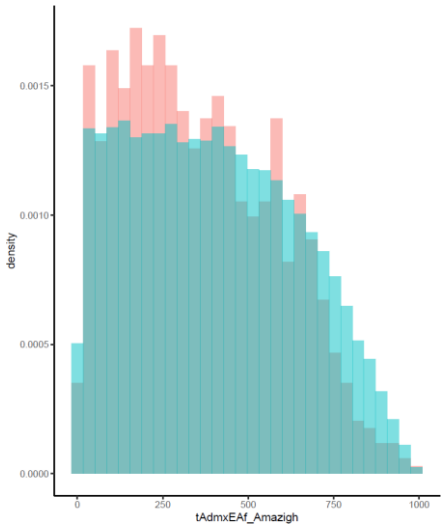

density
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureMENA_Amazigh

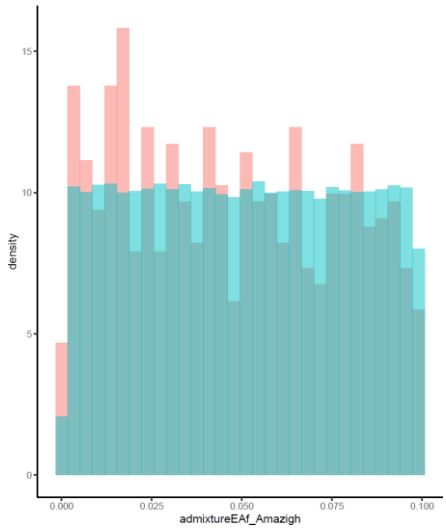

density
0.0015
0.0010
0.0005
0.0000
0
250
500
750
1000
tAdmxWAf_Amazigh

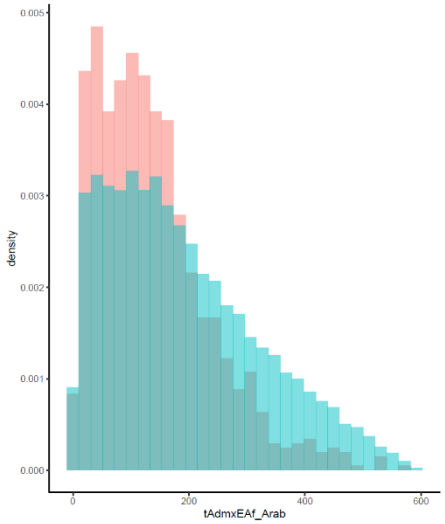

density
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureWAf_Amazigh

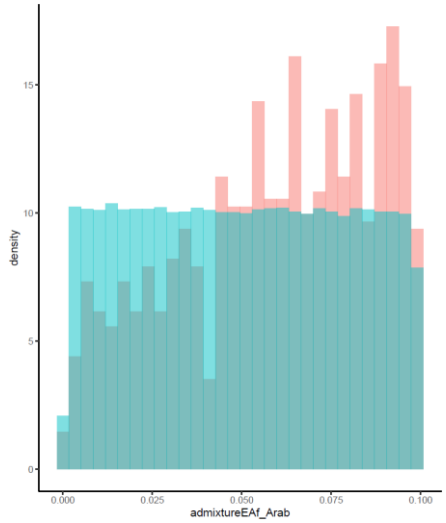

density
0.005
0.004
0.003
0.002
0.001
0.000
0
200
400
600
tAdmxWAf_Arab

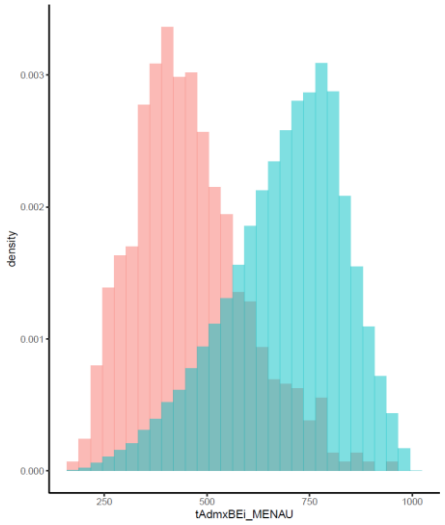

density
15
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureWAf_Arab

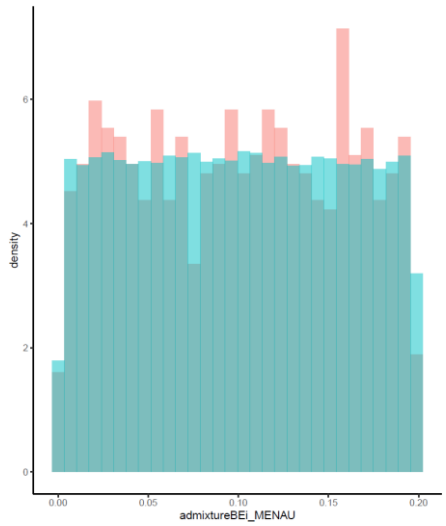

density
0.0015
0.0010
0.0005
0.0000
0
250
500
750
1000
tAdmxEAf_Amazigh

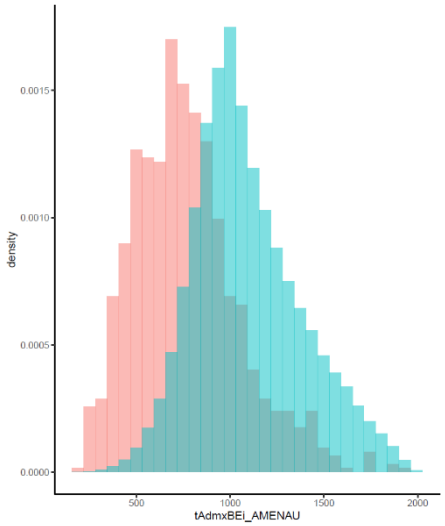

density
15
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureEAf_Amazigh

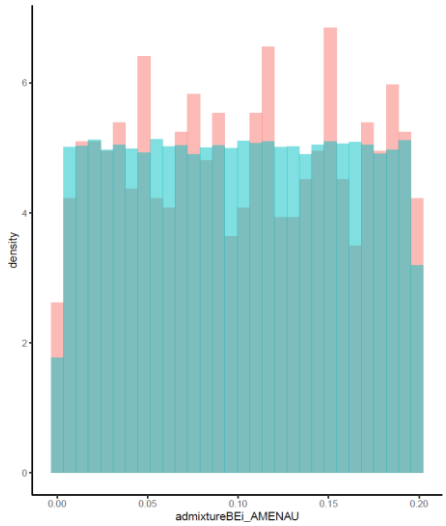

density
0.005
0.004
0.003
0.002
0.001
0.000
0
200
400
600
tAdmxEAf_Arab

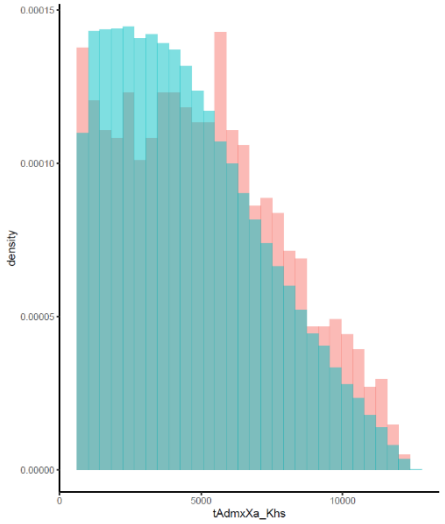

density
15
10
5
0
0.000
0.025
0.050
0.075
0.100
admixtureEAf_Arab

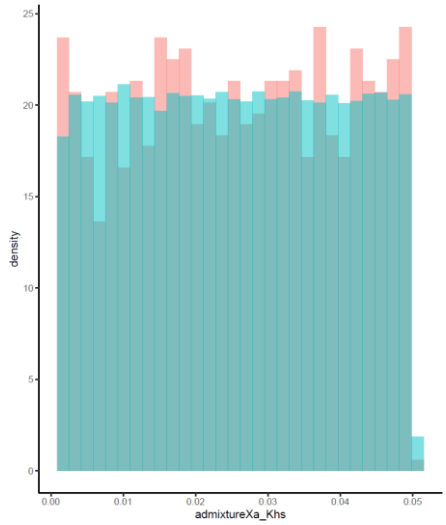

density
0.003
0.002
0.001
0.000
250
500
750
1000
tAdmxBEi_MENAU

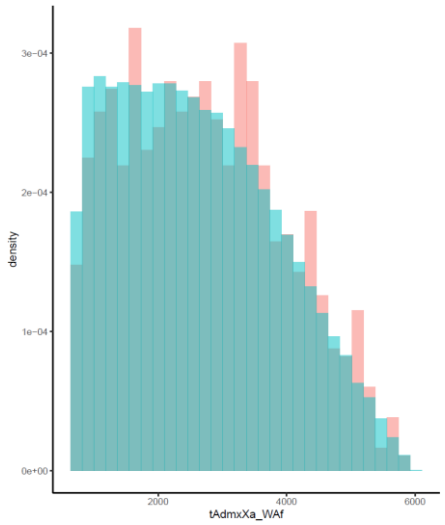

density
6
4
2
0
0.00
0.05
0.10
0.15
0.20
admixtureBEi_MENAU

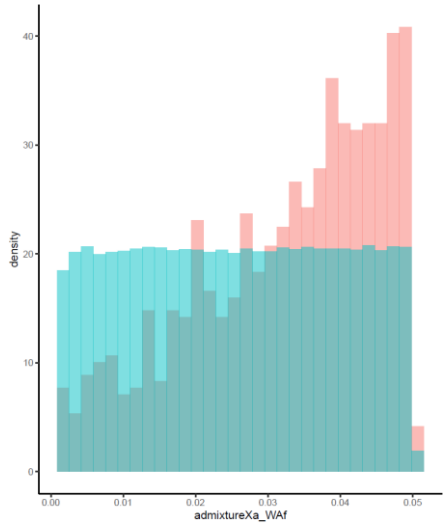

density
0.0015
0.0010
0.0005
0.0000
500
1000
1500
2000
tAdmxBEi_AMENAU


density
6
4
2
0
0.00
0.05
0.10
0.15
0.20
admixtureBEi_AMENAU

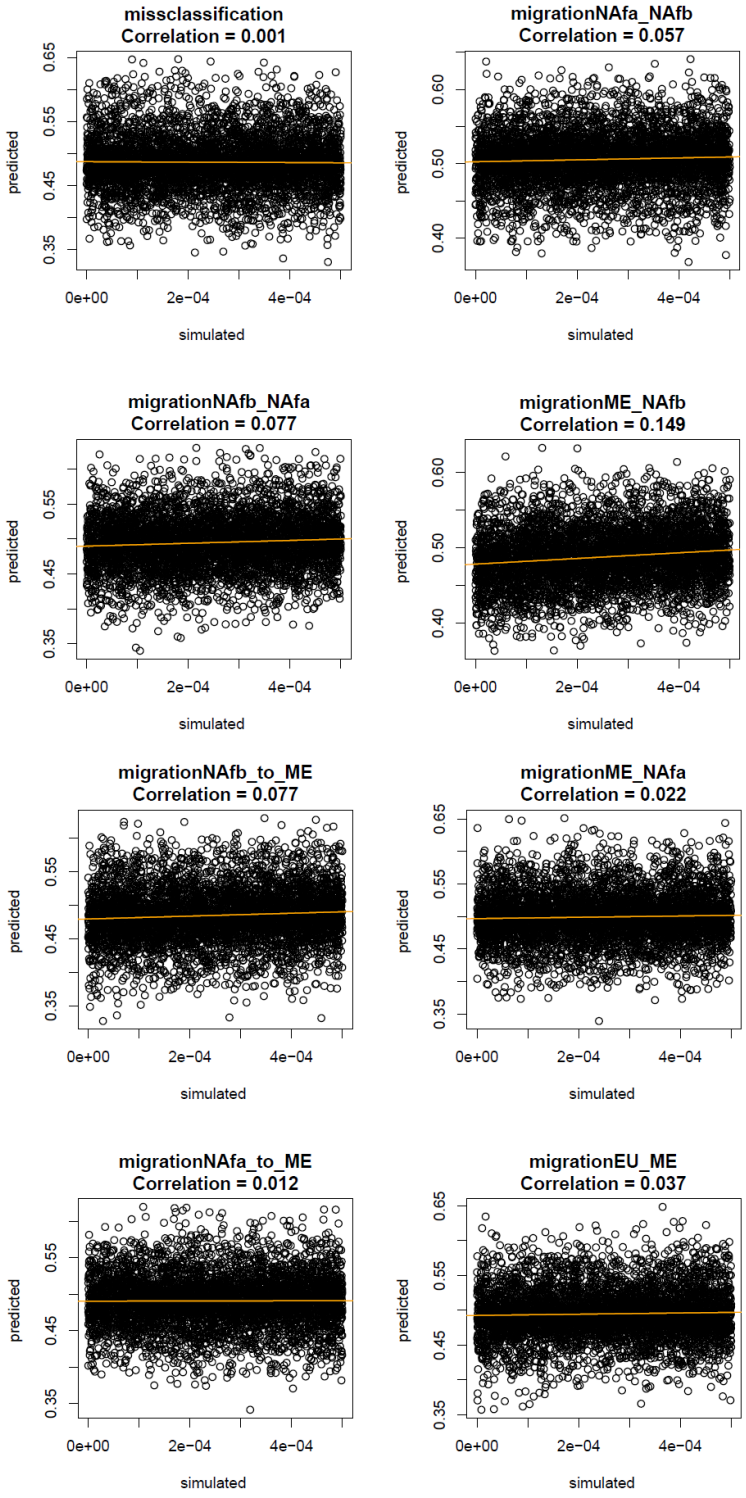

density
0.00015
0.00010
0.00005
0.00000
0
5000
10000
tAdmxXa_Khs

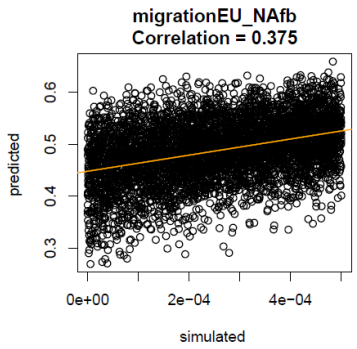

density
25
20
15
10
5
0
0.00
0.01
0.02
0.03
0.04
0.05
admixtureXa_Khs

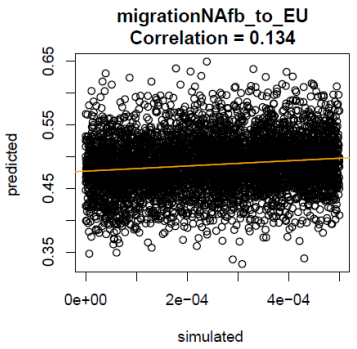

density
3e−04
2e−04
1e−04
0e+00
2000
4000
6000
tAdmxXa_WAf

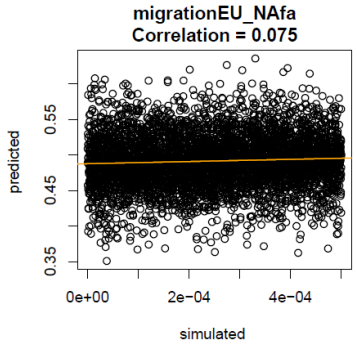

density
40
30
20
10
0
0.00
0.01
0.02
0.03
0.04
0.05
admixtureXa_WAf

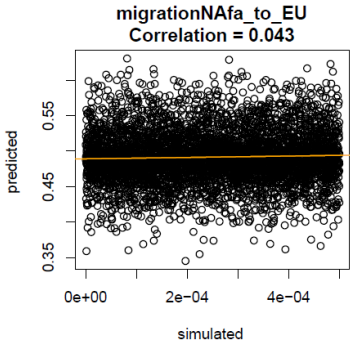
Additional file 4: Spearman correlation plots for all parameters in the best model in the ABC-DL analysis (Model D4)

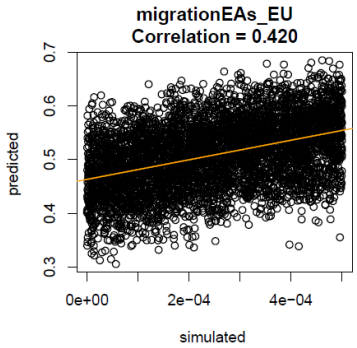


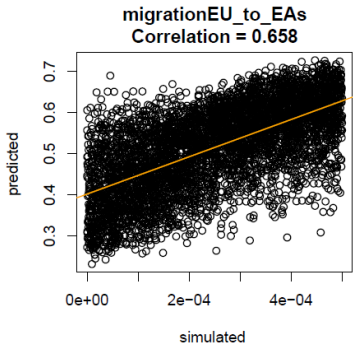

migrationEU_NAfb
Correlation = 0.375
predicted
simulated
0e+00
2e-04
4e-04
0.3
0.4
0.5
0.6

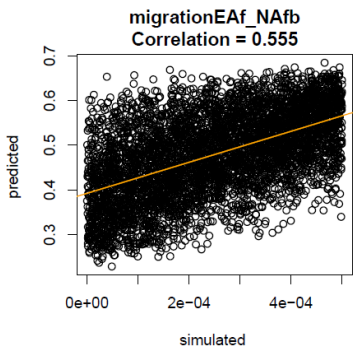

migrationNAfb_to_EU
Correlation = 0.134
predicted
simulated
0e+00
2e-04
4e-04
0.35
0.45
0.55
0.65

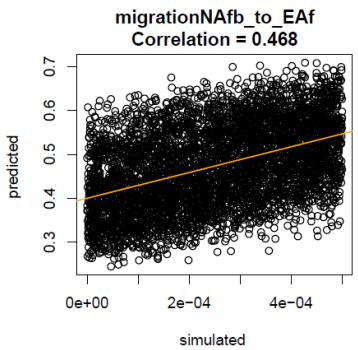

migrationEU_NAfa
Correlation = 0.075
predicted
simulated
0e+00
2e-04
4e-04
0.35
0.45
0.55

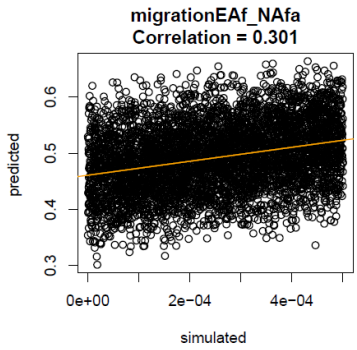

migrationNAfa_to_EU
Correlation = 0.043
predicted
simulated
0e+00
2e-04
4e-04
0.35
0.45
0.55

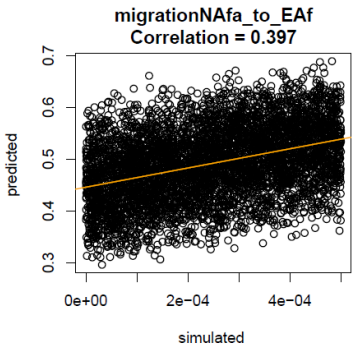

migrationEAs_EU
Correlation = 0.420
predicted
simulated
0e+00
2e-04
4e-04
0.3
0.4
0.5
0.6
0.7

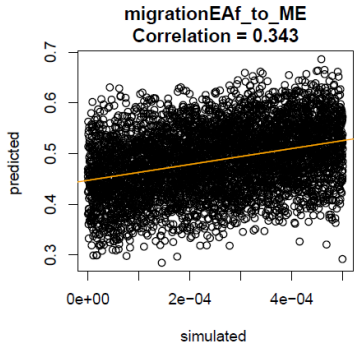

migrationEU_to_EAs
Correlation = 0.658
predicted
simulated
0e+00
2e-04
4e-04
0.3
0.4
0.5
0.6
0.7

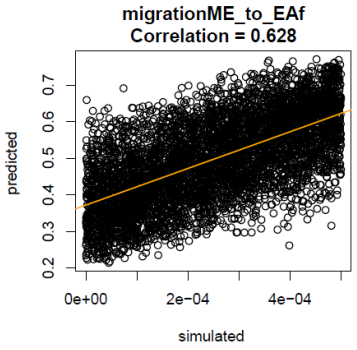

migrationEAf_NAfb
Correlation = 0.555
predicted
simulated
0e+00
2e-04
4e-04
0.3
0.4
0.5
0.6
0.7

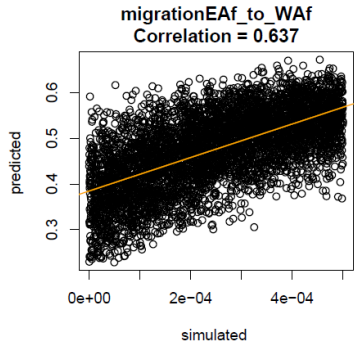

migrationNAfb_to_EAf
Correlation = 0.468
predicted
simulated
0e+00
2e-04
4e-04
0.3
0.4
0.5
0.6
0.7

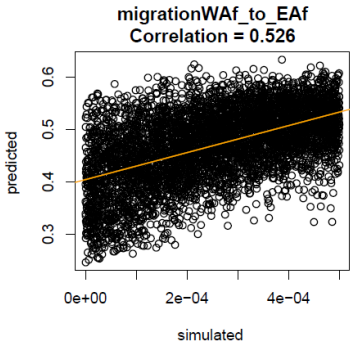

migrationEAf_NAfa
Correlation = 0.301
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6

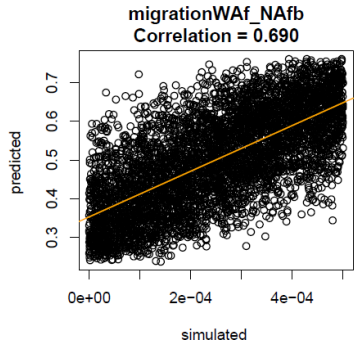

migrationNAfa_to_EAf
Correlation = 0.397
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

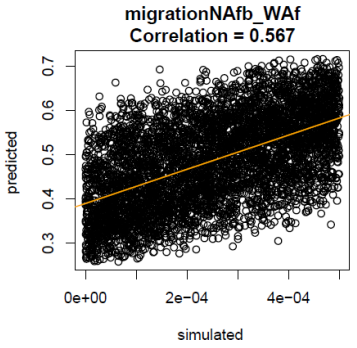

migrationEAf_to_ME
Correlation = 0.343
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

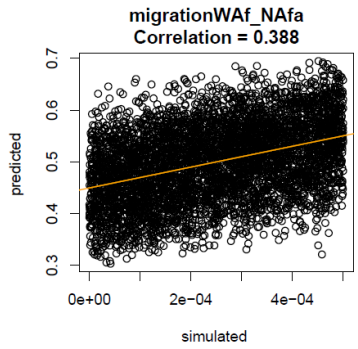

migrationME_to_EAf
Correlation = 0.628
predicted
simulated
0e+00
2e−04
4e−04
0.2
0.3
0.4
0.5
0.6
0.7

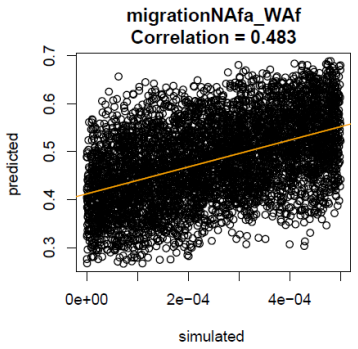

migrationEAf_to_WAf
Correlation = 0.637
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6

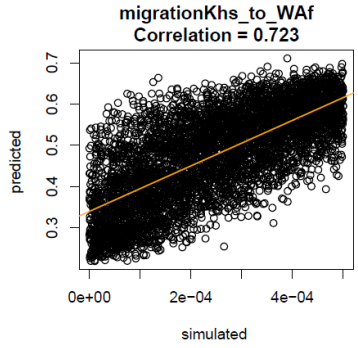

migrationWAf_to_EAf
Correlation = 0.526
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6

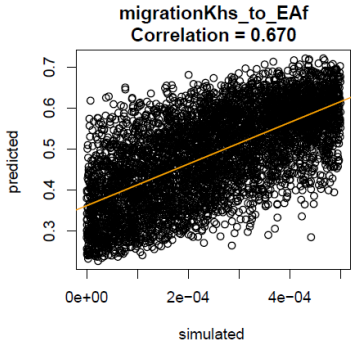

migrationWAf_NAfb
Correlation = 0.690
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

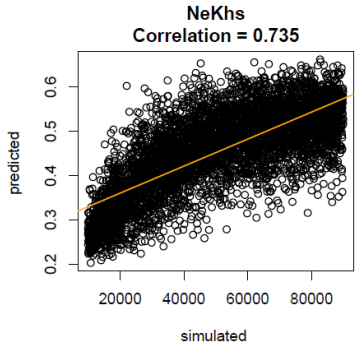

migrationNAfb_WAf
Correlation = 0.567
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

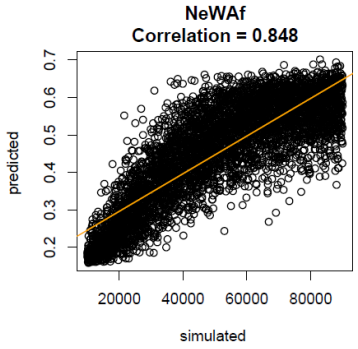

migrationWAf_NAfa
Correlation = 0.388
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

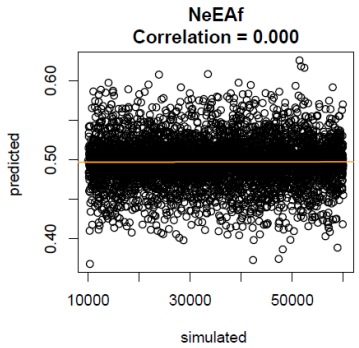

migrationNAfa_WAf
Correlation = 0.483
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

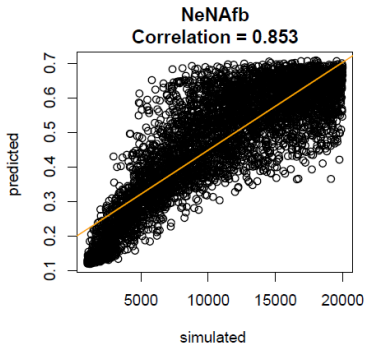

migrationKhs_to_WAf
Correlation = 0.723
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

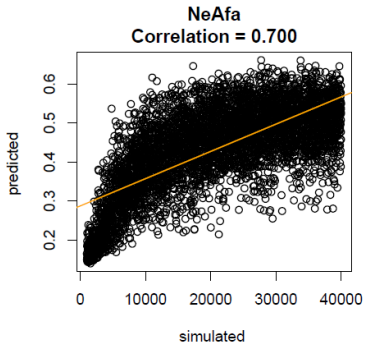

migrationKhs_to_EAf
Correlation = 0.670
predicted
simulated
0e+00
2e−04
4e−04
0.3
0.4
0.5
0.6
0.7

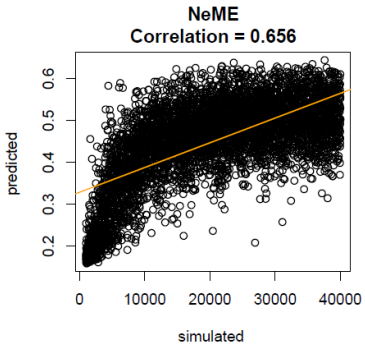

NeKhs
Correlation = 0.735
predicted
simulated
20000
40000
60000
80000
0.2
0.3
0.4
0.5
0.6

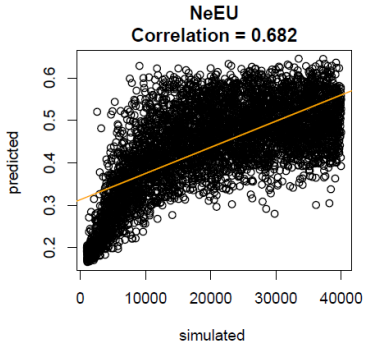

NeWAf
Correlation = 0.848
predicted
simulated
20000
40000
60000
80000
0.2
0.3
0.4
0.5
0.6
0.7

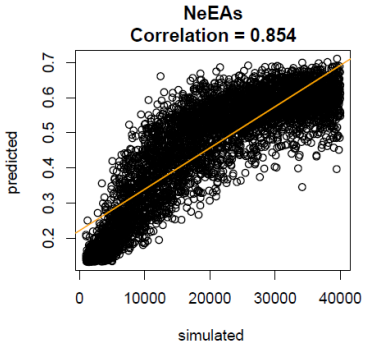

NeEAf
Correlation = 0.000
predicted
simulated
10000
30000
50000
0.40
0.50
0.60

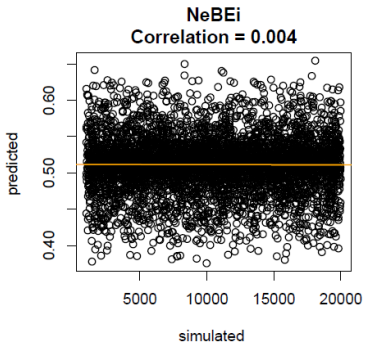

NeNAfb
Correlation = 0.853
predicted
simulated
5000
10000
15000
20000
0.1
0.2
0.3
0.4
0.5
0.6
0.7

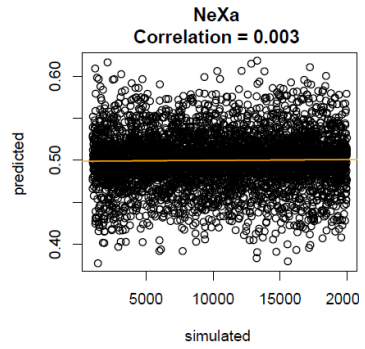

NeAfa
Correlation = 0.700
predicted
simulated

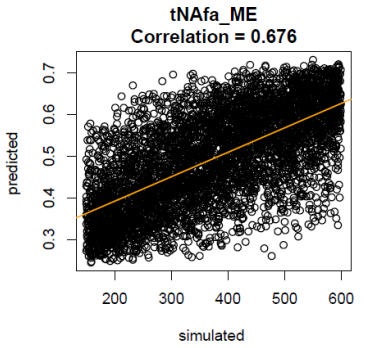

NeME
Correlation = 0.656
predicted
simulated

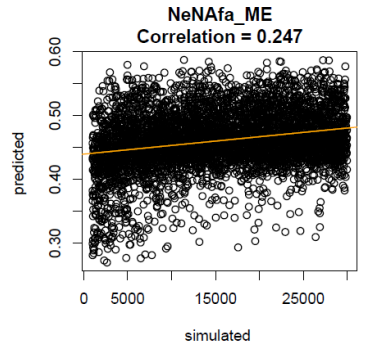

NeEU
Correlation = 0.682
predicted
simulated

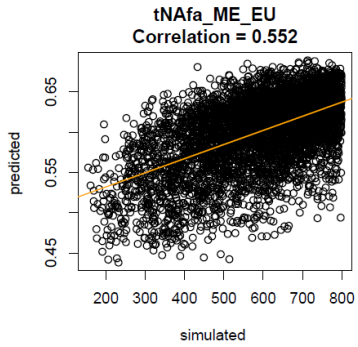

NeEAs
Correlation = 0.854
predicted
simulated

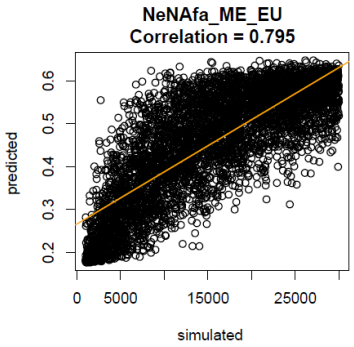

NeBEi
Correlation = 0.004
predicted
simulated

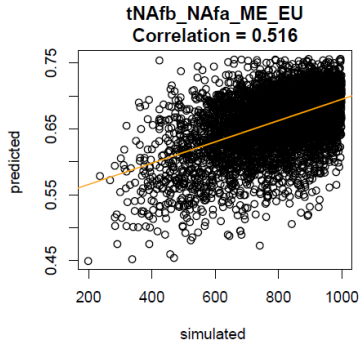

NeXa
Correlation = 0.003
predicted
simulated

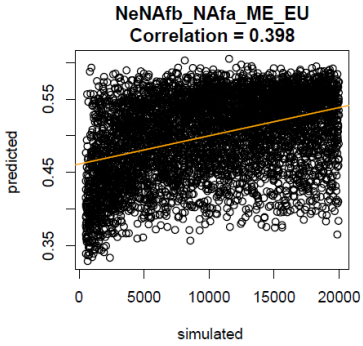

tNAfa_ME
Correlation = 0.676
predicted
simulated

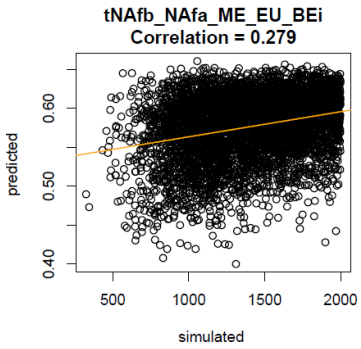

NeNAfa_ME
Correlation = 0.247
predicted
simulated

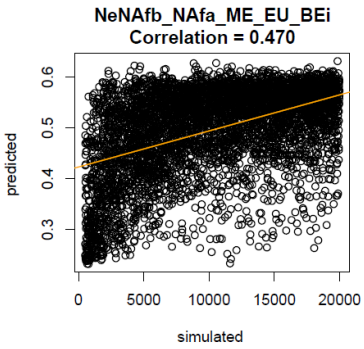

tNAfa_ME_EU
Correlation = 0.552
predicted
simulated

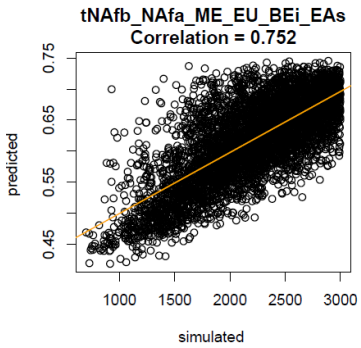

NeNAfa_ME_EU
Correlation = 0.795
predicted
simulated

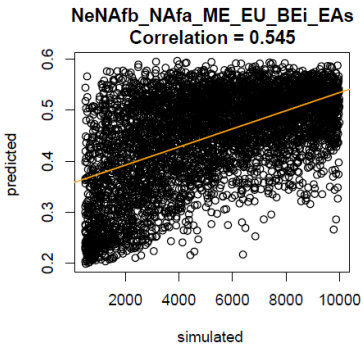

tNAfb_NAfa_ME_EU
Correlation = 0.516
predicted
simulated

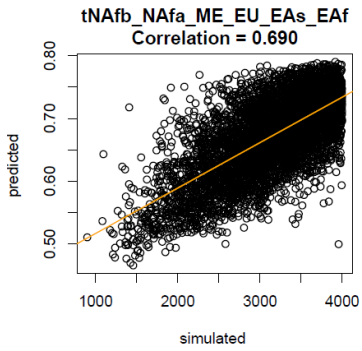

NeNAfb_NAfa_ME_EU
Correlation = 0.398
predicted
simulated

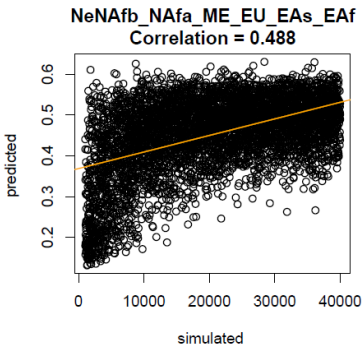

tNAfb_NAfa_ME_EU_BEi
Correlation = 0.279
predicted
simulated

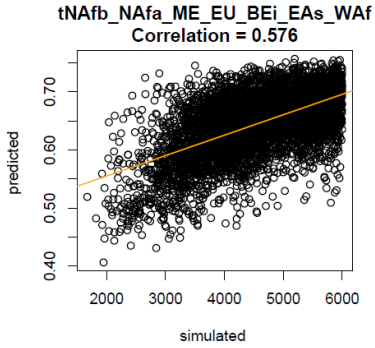

NeNAfb_NAfa_ME_EU_BEi
Correlation = 0.470
predicted
simulated

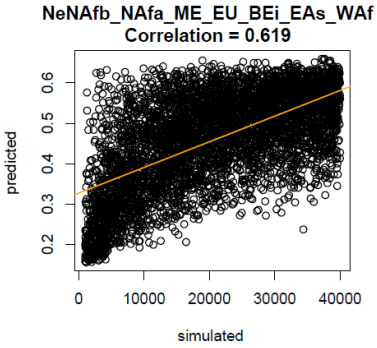

tNAfb_NAfa_ME_EU_BEi_EAs
Correlation = 0.752
predicted
simulated

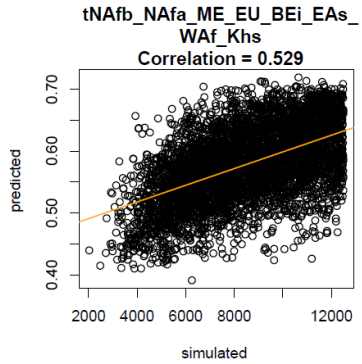

NeNAfb_NAfa_ME_EU_BEi_EAs
Correlation = 0.545
predicted
simulated

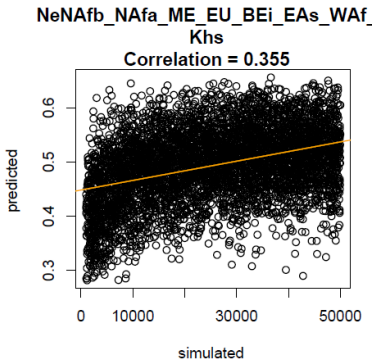

tNAfb_NAfa_ME_EU_EAs_EAf
Correlation = 0.690
predicted
simulated

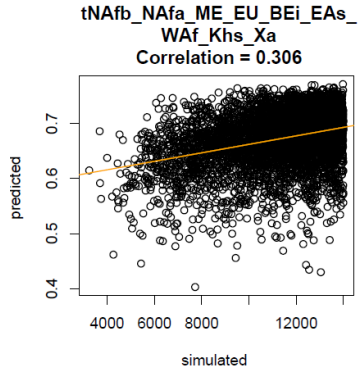

NeNAfb_NAfa_ME_EU_EAs_EAf
Correlation = 0.488
predicted
simulated

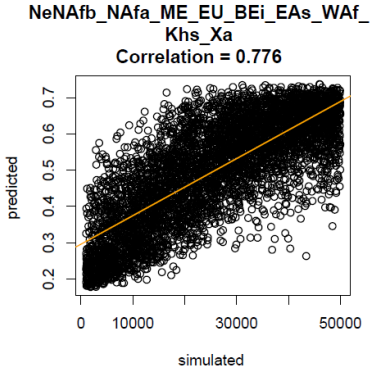

tNAfb_NAfa_ME_EU_BEi_EAs_WAf
Correlation = 0.576
predicted
simulated

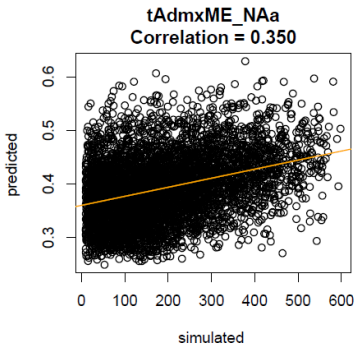

NeNAfb_NAfa_ME_EU_BEi_EAs_WAf
Correlation = 0.619
predicted
simulated

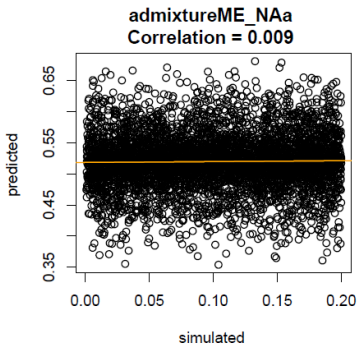

tNAfb_NAfa_ME_EU_BEi_EAs_
WAf_Khs
Correlation = 0.529
predicted
simulated

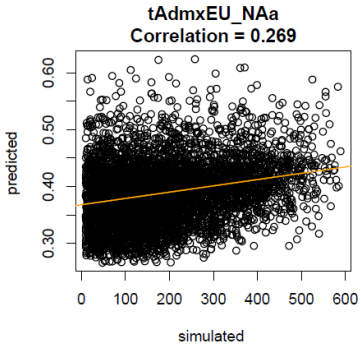

NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_
Khs
Correlation = 0.355
predicted
simulated

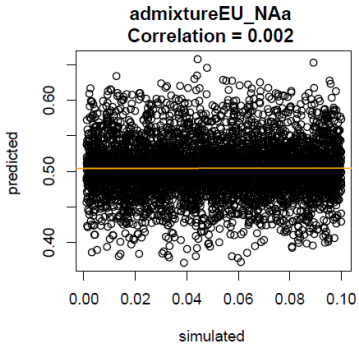

tNAfb_NAfa_ME_EU_BEi_EAs_
WAf_Khs_Xa
Correlation = 0.306
predicted
simulated

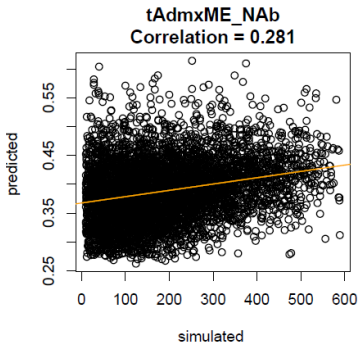

NeNAfb_NAfa_ME_EU_BEi_EAs_WAf_
Khs_Xa
Correlation = 0.776
predicted
simulated

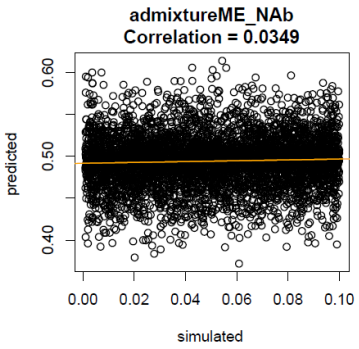

tAdmxME_NAa
Correlation = 0.350
predicted
simulated
0.3
0.4
0.5
0.6
0
100
200
300
400
500
600

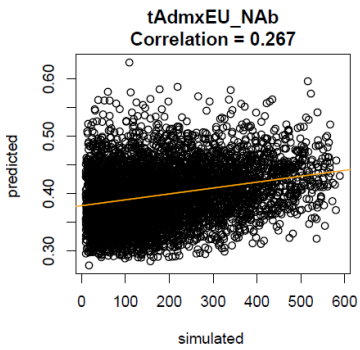

admixtureME_NAa
Correlation = 0.009
predicted
simulated
0.35
0.45
0.55
0.65
0.00
0.05
0.10
0.15
0.20

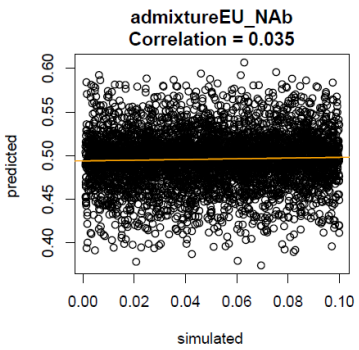

tAdmxEU_NAa
Correlation = 0.269
predicted
simulated
0.30
0.40
0.50
0.60
0
100
200
300
400
500
600

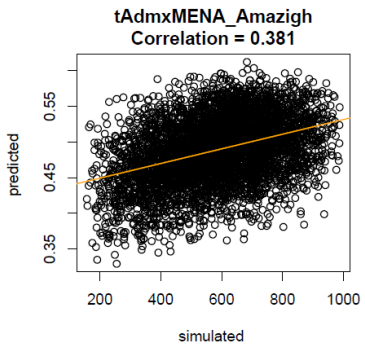

admixtureEU_NAa
Correlation = 0.002
predicted
simulated
0.40
0.50
0.60
0.00
0.02
0.04
0.06
0.08
0.10

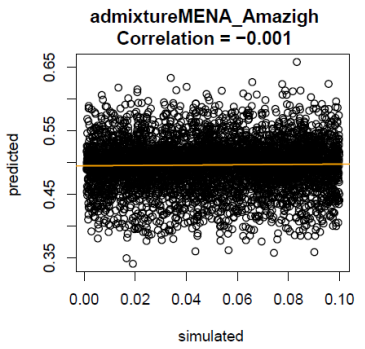

tAdmxME_NAb
Correlation = 0.281
predicted
simulated
0.25
0.35
0.45
0.55
0
100
200
300
400
500
600


admixtureME_NAb
Correlation = 0.0349
predicted
simulated
0.40
0.50
0.60
0.00
0.02
0.04
0.06
0.08
0.10

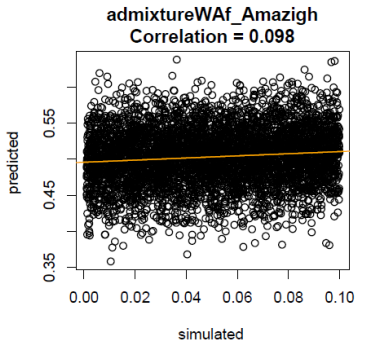

tAdmxEU_NAb
Correlation = 0.267
predicted
simulated
0.30
0.40
0.50
0.60
0
100
200
300
400
500
600

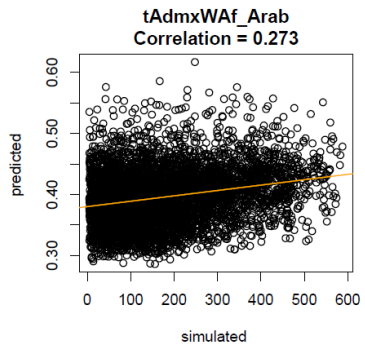

admixtureEU_NAb
Correlation = 0.035
predicted
simulated
0.40
0.45
0.50
0.55
0.60
0.00
0.02
0.04
0.06
0.08
0.10

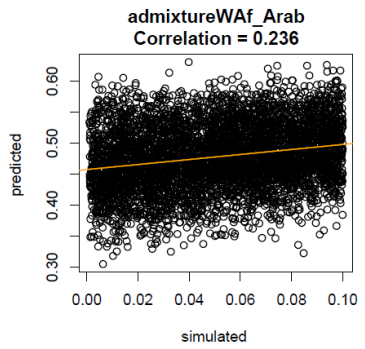

tAdmxMENA_Amazigh
Correlation = 0.381
predicted
simulated

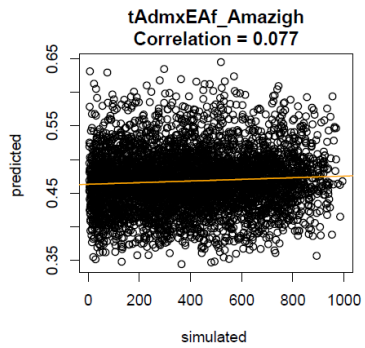

admixtureMENA_Amazigh
Correlation = −0.001
predicted
simulated

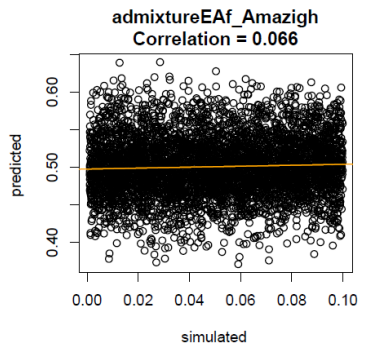

tAdmxWAf_Amazigh
Correlation = 0.354
predicted
simulated

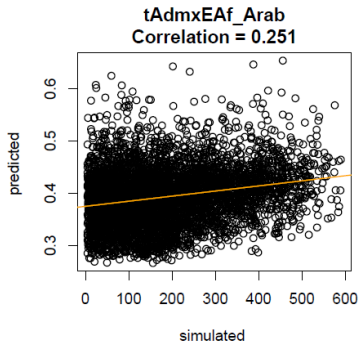

admixtureWAf_Amazigh
Correlation = 0.098
predicted
simulated

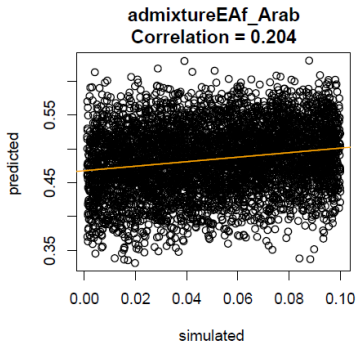

tAdmxWAf_Arab
Correlation = 0.273
predicted
simulated

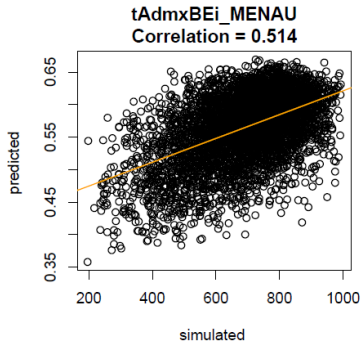

admixtureWAf_Arab
Correlation = 0.236
predicted
simulated

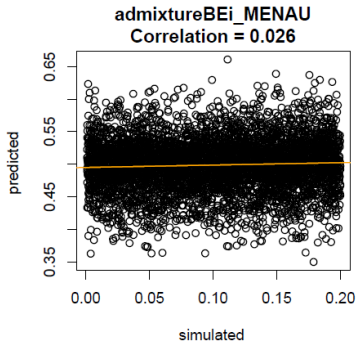

tAdmxEAf_Amazigh
Correlation = 0.077
predicted
simulated

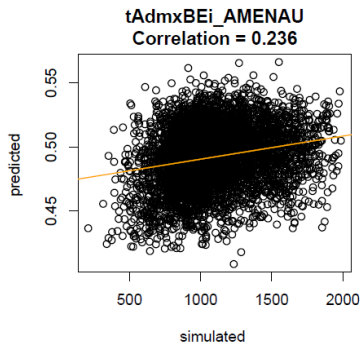

admixtureEAf_Amazigh
Correlation = 0.066
predicted
simulated

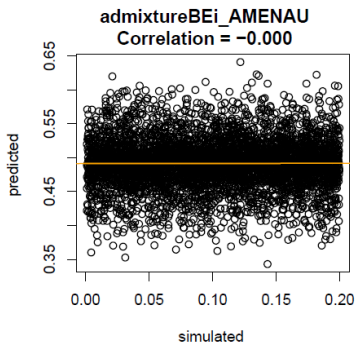

tAdmxEAf_Arab
Correlation = 0.251
predicted
simulated
0.3
0.4
0.5
0.6
0
100
200
300
400
500
600

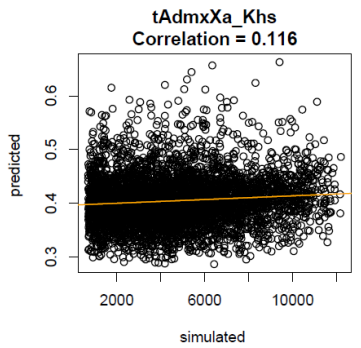

admixtureEAf_Arab
Correlation = 0.204
predicted
simulated
0.35
0.45
0.55
0.00
0.02
0.04
0.06
0.08
0.10

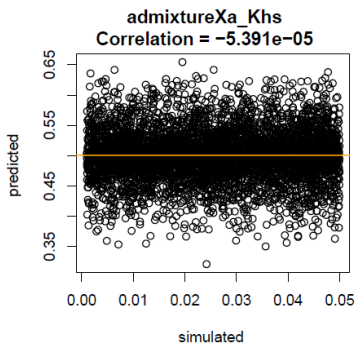

tAdmxBEi_MENAU
Correlation = 0.514
predicted
simulated
0.35
0.45
0.55
0.65
200
400
600
800
1000

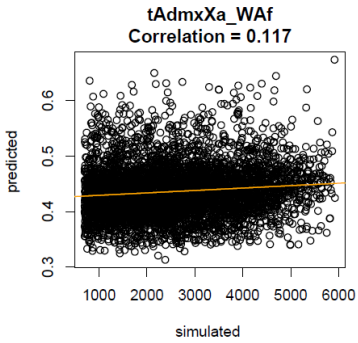

admixtureBEi_MENAU
Correlation = 0.026
predicted
simulated
0.35
0.45
0.55
0.65
0.00
0.05
0.10
0.15
0.20

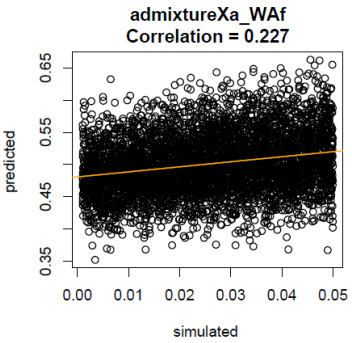

tAdmxBEi_AMENAU
Correlation = 0.236
predicted
simulated
0.45
0.50
0.55
500
1000
1500
2000


admixtureBEi_AMENAU
Correlation = −0.000
predicted
simulated
0.35
0.45
0.55
0.65
0.00
0.05
0.10
0.15
0.20


tAdmxXa_Khs
Correlation = 0.116
predicted
simulated
0.3
0.4
0.5
0.6
2000
6000
10000


admixtureXa_Khs
Correlation = −5.391e−05
predicted
simulated
0.35
0.45
0.55
0.65
0.00
0.01
0.02
0.03
0.04
0.05


tAdmxXa_WAf
Correlation = 0.117
predicted
0.3
0.4
0.5
0.6
1000
2000
3000
4000
5000
6000
simulated


admixtureXa_WAf
Correlation = 0.227
predicted
0.35
0.45
0.55
0.65
0.00
0.01
0.02
0.03
0.04
0.05
simulated

Additional file 5: Parameter values for the best 10 models in the GP4PG analysis. Time express in generations (29 years/generation)

Demes at present time

Parameters	ModelBm_2	ModelF_5	ModelDm_7	ModelD_15	ModelD_23	ModelFm_24	ModelDm_26	ModelC_29	ModelCm_31	ModelC_39
Ecodeme0(NAa)										
°TOPODEMES	16804	15962	4770	14852	33473	5970	20667	24986	16009	16565
	18665	20401			32878	36191	18849	16873		
		5522			25978	11721				
		42539			28011					
SUMATORY	2	4	1	1	4	3	2	2	1	1
migrationWithi	3.52E-05	9.52E-05	4.24E-05	1.71E-05	8.95E-05	1.73E-05	5.97E-05	4.81E-05	8.12E-05	7.13E-05
Ecodeme1(Ama)										
°TOPODEMES	7627	4320	4671	3840	5904	12670	2840	4910	3741	6212
			17971							
SUMATORY	1	1	2	1	1	1	1	1	1	1
migrationWithi	3.56E-05	4.92E-06	6.21E-05	4.49E-05	5.50E-05	6.22E-05	1.48E-05	9.61E-05	1.30E-05	4.13E-05
Ecodeme2(ME)										
°TOPODEMES	14450	20318	33539	22620	6509	14386	8986	23803	22879	6087
		10864	30998	19338		30424	4698			
			47809							
SUMATORY	1	2	3	2	1	2	2	1	1	1
migrationWithi	8.68E-05	2.93E-05	3.25E-05	5.43E-05	1.33E-05	5.23E-05	8.69E-05	9.04E-05	6.24E-05	5.01E-05
Ecodeme3(EU)										
°TOPODEME	15325	13481	26116	7779	9096	3875	25001	20203	16988	19307
	25736		15529							26640
			25401							26847
										26330
										11788
SUMATORY	2	1	3	1	1	1	1	1	1	5
migrationWithi	2.10E-05	6.21E-05	2.74E-05	1.94E-05	9.81E-05	1.49E-05	2.13E-05	2.94E-05	5.63E-05	7.75E-05
Ecodeme4(EAs)										
°TOPODEME	8088	24170	6339	8104	7004	7747	3963	10494	8408	6021
	26392					25623			20997	
SUMATORY	2	1	1	1	1	2	1	1	2	1
migrationWithi	7.78E-05	3.44E-05	3.43E-06	6.40E-05	4.60E-05	4.48E-05	1.27E-05	7.04E-05	1.87E-05	9.26E-05
Ecodeme5(EAf)										
°TOPODEME	62983	70133	79735	65695	78578	53508	19063	50517	40050	67517
	70263	53448		46038			39172		73500	68358
							12965		79903	
							70635		73739	
SUMATORY	2	2	1	2	1	1	4	1	4	2
migrationWithi	5.89E-05	7.63E-05	2.10E-05	8.24E-05	1.99E-05	9.10E-06	8.92E-05	6.56E-05	5.24E-05	2.17E-05
Ecodeme6(WAf)										
°TOPODEME	31405	42178	24780	49741	48808	19116	30541	27751	24776	51960
	7848	7253				5708	3511	2522		57634
										34831
SUMATORY	2	2	1	1	1	2	2	2	1	3
migrationWithi	7.93E-05	1.08E-04	5.25E-05	1.41E-05	2.94E-05	7.01E-05	7.44E-05	5.63E-05	1.38E-05	4.89E-05
Ecodeme7(San)										
°TOPODEME	36019	59352	37749	5842	24898	22155	41527	47458	39157	25220
		46230	39807	29326		35858		77439		7073
		71170	36141			51583				
			35460							
SUMATORY	1	3	4	2	1	3	1	2	1	2
migrationWithi	8.42E-05	5.51E-05	3.48E-05	8.98E-05	2.83E-05	5.06E-05	5.17E-05	6.14E-05	5.94E-05	9.00E-05

Each model presents a different number of topodemes in the predefined ecodemes (populations) that migrate within themselves at a defined rate. Values represent the effective population size (Ne) at each topodeme.

Time splits and migration decay

Parameters	ModelBm_2	ModelF_5	ModelDm_7	ModelD_15	ModelD_23	ModelFm_24	ModelDm_26	ModelC_29	ModelCm_31	ModelC_39
1st_Split										
t	545	322	462	56	46	385	221	260	426	79
source/sink	0(0,1)/1(2)	0(0)/2(5)	2(3,4,5)/0(0)	0(0)/2(2)	2(5)/0(0)	2(4)/0(2)	0(0,1)/2(4)	3(4)/2(3)	3(3)/2(2)	2(2)/3(4)
source/sink		0(1,2,3)/2(6)				2(5)/0(0)				
SUMATORY	1	2	1	1	1	2	1	1	1	1
min_migration_decay	7.15E-07	7.03E-05	1.80E-04	8.89E-07	5.70E-07	8.71E-07	4.33E-07	7.91E-05	6.79E-05	9.77E-07
max_migration_decay	6.94E-04	7.93E-03	4.24E-03	8.40E-03	2.18E-03	3.85E-03	4.67E-03	3.86E-04	3.58E-04	5.32E-04
2n_Split										
t	765	767	681	535	502	453	327	790	749	127
source/sink	2(3)/1(2)	2(5,6)/1(4)	3(6,7,8)/0(0)	2(2,3)/3(4)	0(0,1,2,3)/3(6)	1(3)/0(0)	2(3,4)/3(5)	0(0,1)/1(2)	0(0)/1(1)	0(0)/1(1)
SUMATORY	1	1	1	1	1	1	1	1	1	1
min_migration_decay	4.45E-05	2.39E-05	3.24E-05	6.25E-07	2.00E-08	8.79E-07	2.02E-05	3.06E-07	1.60E-05	6.08E-05
max_migration_decay	9.68E-04	9.76E-04	3.85E-04	1.60E-04	5.75E-04	5.08E-04	5.46E-05	9.16E-04	0.009877202	7.96E-03
3rd_Split										
t	904	828	699	770	815	699	675	870	926	762
source/sink	3(4)/1(2)	3(7)/1(4)	1(1,2)/0(0)	1(1)/3(4)	1(4)/3(6)	0(0,1,2)/3(6)	3(5)/1(2)	2(3)/1(2)	2(2)/1(1)	3(3,4,5,6,7)/1(
SUMATORY	1	1	1	1	1	1	1	1	1	1
min_migration_decay	8.15E-05	5.50E-05	6.27E-07	5.33E-05	4.40E-05	1.30E-04	7.79E-07	9.29E-07	8.90E-07	8.43E-05
max_migration_decay	6.12E-04	9.76E-04	3.04E-04	5.73E-04	5.69E-04	8.92E-04	4.32E-04	3.65E-04	5.73E-04	1.68E-04
4th_Split										
t	1200	1094	899	939	941	1044	876	1107	1265	1061
source/sink	4(6,7)/1(2)	4(8)/1(4)	4(9)/0(0)	3(4)/4(5)	3(6)/4(7)	4(7,8)/3(6)	4(6)/1(2)	4(5)/1(2)	4(4,5)/1(1)	4(8,9)/1(1)
SUMATORY	1	1	1	1	1	1	1	1	1	1
min_migration_decay	5.94E-07	7.60E-07	6.54E-07	7.48E-07	8.62E-07	1.12E-04	3.98E-07	6.33E-07	2.32E-05	3.91E-05
max_migration_decay	4.73E-04	4.76E-04	3.88E-05	8.35E-04	9.53E-04	2.42E-04	3.45E-05	4.12E-04	6.36E-04	9.22E-05
5th_Split										
t	3407	2527	2727	3058	3017	3001	2803	2643	2803	3393
source/sink	5(8,9)/1(2)	5(9,10)/1(4)	5(10)/0(0)	5(6,7)/4(5)	5(8)/4(7)	5(9)/3(6)	1(2)/5(9)	5(6)/1(2)	5(6,7,8,9)/1(1)	5(10,11)/1(1)
SUMATORY	1	1	1	1	1	1	1	1	1	1
min_migration_decay	6.91E-07	2.25E-07	7.73E-08	1.15E-07	9.09E-07	7.14E-07	5.10E-07	5.72E-07	2.55E-05	2.31E-05
max_migration_decay	5.12E-04	3.91E-04	6.46E-04	1.86E-04	1.65E-04	8.15E-04	9.44E-04	2.57E-05	3.02E-04	3.35E-04
6th_Split										
t	5891	2537	4225	3354	4057	5622	3080	4388	3862	3605
source/sink	1(2)/6(11)	1(4)/6(12)	0(0)/6(11)	6(8)/4(5)	6(9)/4(7)	3(6)/6(11)	5(7,8)/6(11)	1(2)/6(7)	1(1)/6(10)	6(12,13,14)/1(
source/sink							5(9,10)/6(12)			
SUMATORY	1	1	1	1	1	1	2	1	1	1
min_migration_decay	1.15E-05	8.68E-05	4.21E-05	3.28E-05	4.64E-07	1.33E-04	2.51E-04	9.32E-05	1.99E-04	3.50E-05
max_migration_decay	9.85E-04	4.84E-04	8.94E-04	7.50E-04	9.67E-04	8.88E-04	6.57E-04	6.13E-04	7.75E-04	3.16E-04
7th_Split										
t	5900	8574	4226	5487	9916	6808	3704	5119	3865	6290
source/sink	7(12)/6(10)	7(13,15)/6(12)	7(12,13,14,15)/	4(5)/7(9)	4(7)/7(10)	7(12,14)/6(10)	7(13)/6(11)	7(10)/6(8)	7(11)/6(10)	1(1)/7(16)
source/sink		7(14)/6(11)				7(13)/6(11)		7(9)/6(7)		
SUMATORY	1	2	1	1	1	2	1	2	1	1
min_migration_decay	2.87E-06	1.09E-06	2.35E-05	1.14E-07	7.70E-05	6.17E-05	9.33E-07	2.75E-08	6.97E-07	4.99E-07
max_migration_decay	4.40E-04	8.80E-04	2.93E-05	6.23E-04	5.24E-04	5.16E-04	9.30E-04	9.76E-04	1.33E-04	7.95E-04

For each model the order at which the population split is predefined but how they do it is generated by the GP4PG algorithm. The different parameters are t, time of the split in generations (29 y/gen); source/sink, specifies the original and the new demes formed after the split backwards in time; min/max_migration_decay; range at which the migration between interacting demes decays after the split of the population, it stabilizes as the 2 new populations never stop interchanging individuals.

Other demographic events (migrations, changes in Ne, modification in number of demes…)

Parameters	ModelBm_2	ModelF_5	ModelDm_7	ModelD_15	ModelD_23	ModelFm_24	ModelDm_26	ModelC_29	ModelCm_31	ModelC_39
OTHER DEMOGRAPHIC EVENTS										
1st Demographic event										
Name	Increase_InFwd									
time	68									
NeChage	3(5>3(4)									
Migrations Between Topodemes										
Ama>NAa	2.99E-04		3.00E-04			9.15E-04	2.87E-04		5.27E-05	
Ama>ME	1.91E-03		3.05E-04			1.48E-03	8.33E-04		1.87E-03	
Ama>EU	1.44E-04		3.15E-04			4.72E-05	8.31E-04		2.79E-06	
Ama>WAf	2.20E-03		2.23E-04			5.04E-05	1.66E-03		3.47E-04	
Ama>EAf	2.98E-04		6.71E-05			2.24E-04	8.23E-04		3.76E-04	
NAa>Ama	4.22E-04		8.11E-04			4.30E-04	3.56E-04		9.89E-05	
NAa>ME	2.23E-04		1.54E-04			2.84E-04	1.34E-03		8.50E-04	
NAa>EU	9.30E-05		1.10E-03			4.95E-04	2.09E-03		4.77E-04	
NAa>WAf	1.36E-03		3.54E-04			2.81E-04	9.18E-05		4.10E-04	
NAa>EAf	1.87E-04		2.43E-03			1.27E-03	2.49E-03		2.03E-03	
ME>Ama	7.54E-04		4.66E-04			1.25E-03	2.31E-05		1.42E-05	
ME>NAa	3.30E-04		7.32E-04			2.46E-05	4.41E-04		2.05E-05	
ME>EU	2.53E-04		9.00E-03			1.33E-04	2.95E-04		3.50E-04	
ME>EAf	2.73E-04		5.92E-04			6.34E-03	4.02E-03		3.94E-03	
EU>Ama	1.47E-04		1.33E-05			9.48E-04	1.06E-05		2.06E-03	
EU>NAa	1.34E-04		1.59E-04			1.26E-03	4.13E-04		1.22E-04	
EU>ME	5.99E-05		3.87E-04			3.96E-03	4.43E-04		4.98E-05	
EU>EAs	9.50E-04		2.29E-04			1.07E-02	6.47E-04		6.72E-04	
EU>WAf	5.51E-04		4.39E-04			1.84E-03	8.38E-04		1.30E-03	
EU>EAf	2.00E-04		4.01E-04			8.23E-05	1.47E-03		4.81E-04	
EAs>EU	1.31E-05		1.36E-06			1.18E-03	6.67E-05		8.86E-06	
EAf>Ama	4.54E-04		1.28E-04			8.91E-04	1.82E-03		1.28E-03	
EAf>NAa	8.96E-04		6.62E-04			5.52E-05	3.64E-04		4.36E-03	
EAf>ME	1.22E-04		9.89E-04			2.86E-03	2.10E-03		7.58E-04	
EAf>EU	6.46E-04		9.31E-04			9.63E-04	9.87E-04		3.41E-04	
EAf>WAf	3.64E-04		3.62E-05			7.65E-04	3.66E-04		7.29E-04	
EAf>San	3.68E-03		1.10E-03			1.64E-03	1.98E-03		1.44E-03	
WAf>Ama	1.40E-03		2.08E-04			1.01E-04	7.00E-04		6.79E-04	
WAf>NAa	4.34E-04		4.06E-04			1.09E-03	1.69E-05		3.76E-04	
WAf>ME	1.86E-03		2.01E-04			6.98E-04	2.32E-03		1.97E-04	
WAf>EU	3.55E-03		3.61E-04			1.21E-03	2.50E-04		2.98E-04	
WAf>EAf	1.15E-04		1.69E-04			2.24E-04	3.52E-04		9.52E-04	
WAf>San	3.74E-04		9.20E-05			2.75E-04	6.20E-04		5.68E-03	
San>EAf	2.74E-03		4.25E-04			1.18E-03	4.03E-05		1.31E-05	
San>WAf	9.07E-04		1.13E-05			1.10E-03	2.68E-04		5.94E-04	

In the 10 best models the only demographic events that we observe are migrations between different topodemes that follow an isolation by distance pattern, and an increase in forward of topodemes in a population. This means that an ecodeme gains population substructure (by gaining a topodeme), in this case deme 5 produces deme 4 at generation 68.

This thesis is a result of the I+D+I project with reference **AEI-PID2019-106485GB-I00/AEI/10.13039/501100011033**, funded by Ministerio de Ciencia, Innovación y Universidades (MCIU), and by Agencia Estatal de Investigación (AEI)