Imperial College London

IMPERIAL COLLEGE LONDON

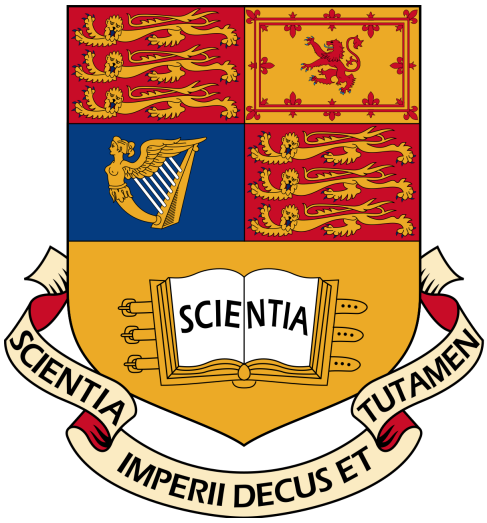
MASTER THESIS

Evaluation of statistical methods for identifying demographic processes from human genetic variation data

---

Author:
Aaron MAMANN

Supervisor:
Dr. Christopher Hallsworth



Master Thesis submitted in the context of the Master's degree - Master of Science in Statistics of Imperial College London

The work contained in this report is my own work unless otherwise stated.

Aaron MAMANN
September 2021

Abstract

Using genetic data has enabled to discover some crucial information about ancient demographic processes [(7),(11)]. Some statistical methods use variations at specific positions in individuals' DNA sequences (i.e. SNPs) in order to identify these demographic processes. However, different demographic scenarios can have similar outcomes in terms of variations obtained in individuals' DNA sequences, which turns out to be a problem for these statistical methods. Therefore, we have simulated the genetic variation corresponding to specific demographic scenarios (these simulations of variations in DNA sequences have been done by the coalescent-based simulator SCRM [(4)] via the R package Coala [(5)]) and have applied these statistical methods on the simulated genetic data in order to assess their ability to identify those demographic scenarios.

Contents

1 Introduction

Variations in human genetic data provide evidence of historical demographic processes such as migrations, population expansion and population splits. Therefore, many ancient demographic processes have been identified using DNA sequences. Some scientists have identified major migrations in Northeastern Siberia during the Late Pleistocene period (period that lasted from about 129,000 to 11,700 years ago) using 34 newly recovered ancient genomes that date to between 31,000 and 600 years ago [(7)]. Similar studies have been done on other migration events [(8),(9)]. Other studies have focused on admixture events [(10)].

The detection of patterns of variations in DNA sequences (as a result of these demographic processes) requires some statistical methods. Some scientists have used Principal Component Analysis [(8),(12)], model-based clustering [(13),(14)] or other methods. However, different demographic scenarios can have similar outcomes in terms of variations obtained in individuals' DNA sequences, which turns out to be a problem for these statistical methods.

Therefore, we have simulated the genetic variation corresponding to specific demographic scenarios and have applied these statistical methods on the simulated genetic data in order to assess their ability to identify those demographic scenarios. These simulations (of the genetic variation corresponding to specific demographic scenarios) have been done by the coalescent-based simulator SCRM [(4)] via the R package Coala [(5)].

In the section 3, we have explained a statistical method which consists in using Principal Component Analysis (PCA) on genetic data to estimate the age of a population split [(2)]. Then, we have simulated genetic data which correspond to a demographic model where there are migrations between populations, and we have applied Principal Component Analysis on these simulated genetic data in order to assess the ability of PCA to estimate the age of a population split in the presence of migrations.

In the section 4, we have precisely described the model-based clustering STRUCTURE (the version without admixture) [(1)] which has been specifically designed for identifying individuals' population of origin (in non-admixed models), using genetic variation data as the input. Then, we have simulated genetic data which correspond to a demographic model where one population has recently split into two different populations. We knew that the more recent the split is, the more similar the genetic sequences of the two populations will be. Therefore, with these simulated genetic data, we wanted to see if the model-based clustering STRUCTURE could find the true population of origin of each individual in the context of recently separated populations.

In the section 5, we have explained the version of STRUCTURE for admixed models [(1)] , which has been designed for identifying all the populations of origin of each individual and for estimating the admixture proportions (i.e. the relative contribution from each population of origin to an individual's ancestry) of each individual. For many reasons, we have suspected that there would be problems in the estimation of admixture proportions in a demographic model where there are unidirectional migrations. Therefore, we have simulated

genetic data corresponding to this demographic model and we have applied STRUCTURE on these simulated genetic data to estimate individuals' admixture proportions in the presence of unidirectional migrations.

In the section 6, we have performed population identification using real data. We have randomly selected 100 different samples of 500 SNPs from the 1105538 SNPs included in the original (real) data, creating 100 different datasets of 500 SNPs. In the same way, we have created 100 different datasets of 100 SNPs, 2000 SNPs, 5000 SNPs and 100 000 SNPs from the original data. On each dataset, we have used a 3-dimensional PCA, and we have applied the K-Means (where K=5) on the 3-dimensional data to perform population identification. Thus, we want to know to what extent the ability (of this statistical method) to find individuals' true population of origin is influenced by the amount of available genetic variation data.

2 The basics of the Coalescent theory

The theory of the coalescent can be primarily attributed to Kingman J.F.C. (1982) [(18)]. In this section, we will explain the basics of this theory.

We are going to focus on one specific position in the chromosome i. At this specific position, from one copy of the chromosome i to another copy of the chromosome i of the same individual, the base pair of nucleotides may be different. Moreover, from one copy of the chromosome i of one individual to another copy of the chromosome i of another individual, the base pair of nucleotides (at the same position) may also be different. This is the result of mutations occurring at this same position. Therefore, every existing base pair of nucleotides at this same position shares a common history which can be represented by a tree (as below).

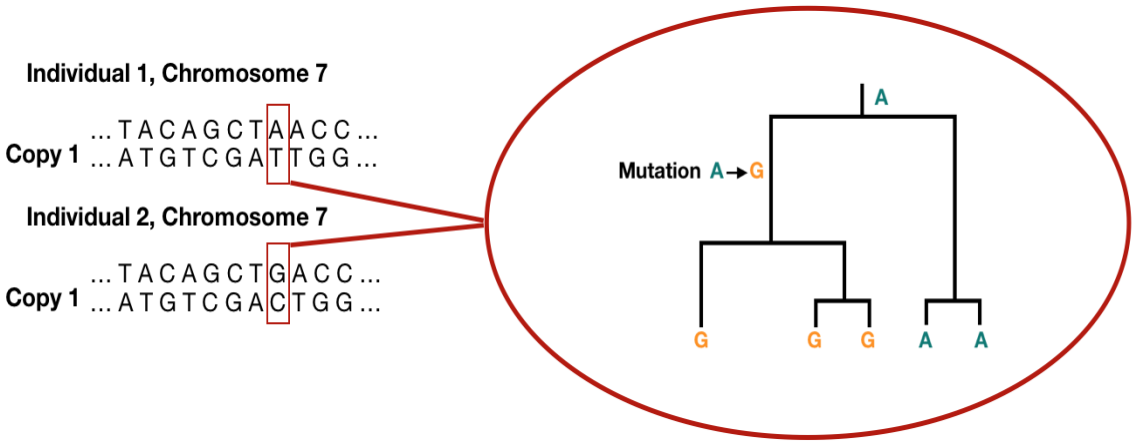


Figure 1: **Genealogical tree representing the history of a specific position in the chromosome 7.** Before the mutation, there was only one existing base pair of nucleotides (at that position of the chromosome 7) which was A-T. After the mutation, there are two existing base pairs of nucleotides which are G-C and A-T.

Looking back in time, we can see (on the Figure 1) that the lineages merge into a single lineage

which is the most recent common ancestor (with respect to this position in the chromosome 7). We say that the lineages coalesce into a single lineage.

The genealogical tree (above) representing the history of one specific position can be generalised to a genealogical tree representing the history of thousands of positions in different chromosomes.

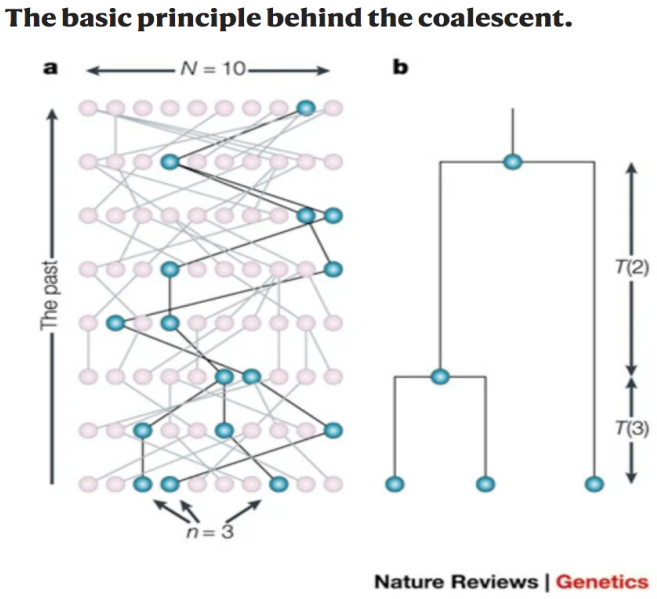


Figure 2: **Diagram provided by Rosenberg N., Nordborg M. (2002)** [(3)] (a) Diagram representing the genealogy of 10 genetic sequences (e.g. 10 entire genomes). The blue points correspond to the genealogy of a sample of 3 genetic sequences which coalesce into a single ancestral genetic sequence. (b) Genealogical tree of the sample of 3 genetic sequences. T(2) and T(3) correspond to the times between consecutive coalescence events.

The **classic model** simulates the genealogy of the **population** (the genealogy of 10 genetic sequences on the Figure 2) **going forwards in time** and after picks a sample, whereas **the coalescent** only simulates the genealogy of the **sample** (blue points on the Figure 2) **going backwards in time**.

In the coalescent model, in the absence of selection, the genetic sequences (in the sample) select randomly their ancestor in the previous generation ("randomly picking their parents" [(3)]): on the Figure 2, each layer of points corresponds to one generation, therefore it means that the black lines connecting each layer have been traced randomly.

We have explained the simplest coalescent model. There are more complicated models as the coalescent with recombination. The coalescent-based simulation software SCRM [(4)], which uses the approximated coalescent with recombination, has enabled us to simulate (in the following sections) the genetic variation corresponding to specific demographic scenarios.

3 Identifying the age of a population split

3.1 Context

First, let us imagine that there are two populations living at two different places without migrations between these two places. These two populations came from an ancestral population which split at a specific time. Our aim is to estimate that time. The age of the split (of the ancestral population into the derived populations) determines how different (on a genetic level) the two populations are. As the two populations are geographically separated and as there are no migrations between these two populations, each population becomes more and more unique as time goes on. In the population 1, all new offspring have both parents from population 1 (because there are no migrations and the two populations are geographically separated) and recombine their parents' genomes. Therefore, as time goes on, there will be recombinations of the genomes of different families from population 1, which is the reason why population 1 becomes more and more unique on a genetic level (it is also the case for population 2).

There are other reasons why two geographically separated populations become more and more different: mutations occur at random, and are unlikely to occur at the same place by chance in both populations, therefore one would expect genetic differences to accumulate between isolated populations.

3.2 Genealogical Interpretation of Principal Components Analysis

Some methods have been used to identify the time at which one population splits into two (or more) geographically separated populations. McVean G. (2009) [(2)] has used Principal Component Analysis (PCA) on a dataset which contains individuals' genotypes. In this dataset, the author already knew that there were only two populations and he also knew the population of origin of each individual. We said (above) that after the split each population becomes more and more different (on a genetic level) as time goes on. Therefore, the idea of the method of McVean G. [(2)] is to find a mathematical relationship between the age of the split and the genetic differences between the two populations, so that we can estimate the age of the split if we measure the genetic differences between the two populations. In order to understand the work of McVean G., we first have to understand precisely the Principal Component Analysis (PCA).

Let X be the matrix which contains the individuals' genotypes (each genotype can be coded by an integer). x_i corresponds to the ith row of the matrix X. $x_i \in R^p$ can also be seen as the p genotypes (located at p different loci) of the ith individual. Therefore, each individual can be distinguished by its p genotypes (located at p different loci), so each individual can be seen as a point in a space of p dimensions where the p coordinates of the point correspond to its p genotypes. Thus, the difference between each individual is defined by the difference of their p genotypes which corresponds to the difference of the p coordinates of the points.

Principal Component Analysis (PCA) aims to reduce the number of dimensions of this p-dimensional space while preserving as much as possible the differences between the N points. Therefore, PCA aims to maximise the variance of the projected data. We are now going to define that in mathematical terms.

Recall that X is the matrix containing the individuals' genotypes and x_i corresponds to the ith row of the matrix X. Let $\tilde{x_i} = x_i - \bar{x}$ $(i \in \{1, .., N\})$. We want to project the data onto a k-dimensional space. Therefore, this space is spanned by a family of vectors $(v_1, v_2, .., v_k)$. First, let us imagine that we project the data onto one vector that we call v_1, the variance of these projected points can be expressed by: $\frac{1}{N}\sum_{i=1}^{N}(v_1^T\tilde{x_i})^2$. We can notice that: $\frac{1}{N}\sum_{i=1}^{N}(v_1^T\tilde{x_i})^2 = v_1^T(\frac{1}{N}\sum_{i=1}^{N}\tilde{x_i}\tilde{x_i}^T)v_1$. Therefore, we have to find the vector v_1 for which the variance of the projected points is maximised:

$$argmax_{v_1, ||v_1||^2=1} \; v_1^T(\frac{1}{N}\sum_{i=1}^{N}\tilde{x_i}\tilde{x_i}^T)v_1 \tag{1}$$

We search now for a vector that we will call v_2 that has to be orthogonal to v_1 and that also has to maximise the variance of the projected points. Then, for every $j \in \{2, .., k\}$, we will search for a vector v_j which has to be orthogonal to $v_1, .., v_{j-1}$ and has to maximise the variance of the projected points. In the mathematical terms, it can be expressed as:

$$argmax_{v_j, ||v_j||^2=1, v_j \in Vect(v_1, .., v_{j-1})^\perp} \; v_j^T(\frac{1}{N}\sum_{i=1}^{N}\tilde{x_i}\tilde{x_i}^T)v_j, \;\; \forall j \in \{1, 2, .., k\} \tag{2}$$

The solution to the optimisation problem 1 is the eigenvector associated with the largest eigenvalue of the matrix $\frac{1}{N}\sum_{i=1}^{N}\tilde{x_i}\tilde{x_i}^T$. The solution to the optimisation problem 2 is the eigenvector associated with the jth largest eigenvalue of the matrix $\frac{1}{N}\sum_{i=1}^{N}\tilde{x_i}\tilde{x_i}^T$.

McVean G. applied 1-dimensional PCA on the dataset containing the individuals' genotypes [(2)]. It is equivalent to projecting each individual who corresponds to a point in R^p onto a 1-dimensional space spanned by v_1 which is the eigenvector associated with the largest eigenvalue of the matrix $\frac{1}{N}\sum_{i=1}^{N}\tilde{x_i}\tilde{x_i}^T$. We can call v_1 the first principal component.

Let D be the distance between two populations projected onto the first principal component (e.g. the distance between the set of pink points and the set of green points on the Figure 3). Therefore, this distance D can be defined as the distance between the two centroids (e.g. the distance between the two blue squares on the Figure 3).

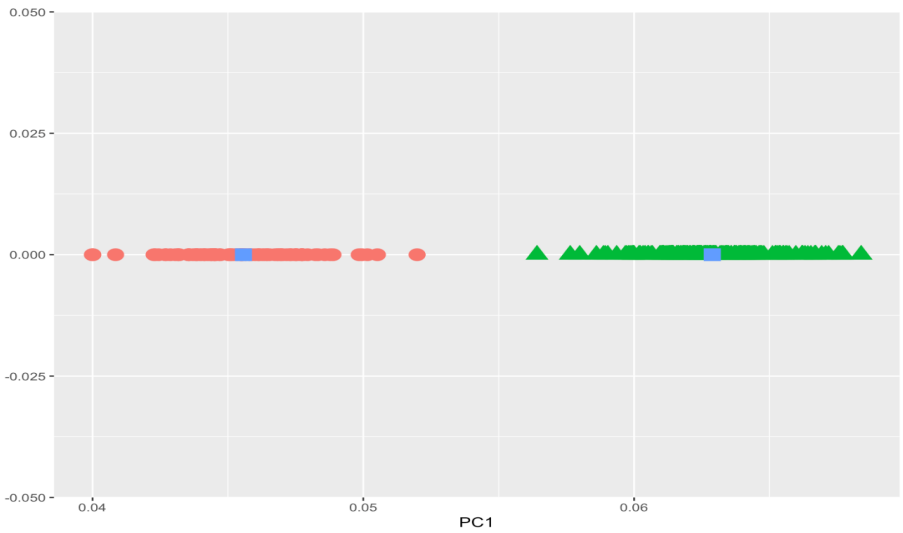


Figure 3: **Example of the projection of the genetic sequences of two populations onto the first principal component.** We have simulated these genetic sequences via the R package "Coala" [(5)]. On this Figure, each point corresponds to one individual, and each color corresponds to one population. Each blue square corresponds to the centroid of each population (here these centroids are defined as the mean of all the points in the set).

Now, we are going to describe the following demograghic model. We have two geographically separated populations: the population 1 and the population 2. These two geographically separated populations were created from the split of an ancestral population. McVean G. found that we can estimate the age of this split using the distance D (i.e. the distance between the genetic sequences of the two populations projected onto the first principal component) [(2)]. He proved that $D = \sqrt{2\Delta/T}$ where Δ is the age of the split and T is the total branch length in the coalescent tree. Therefore, he found a way to estimate the age of the split from the distance between the two projected populations.

3.3 Problems for identifying the age of a population split in the presence of migrations

3.3.1 Simulating the data

We have simulated the genetic data by the R package called "coala" [(5)]. The "coala" R package is based on the coalescent-based simulator "SCRM" [(4)] to simulate the genetic sequences of individuals in different demographic models. Here, the demographic model corresponds to two geographically separated populations. We have simulated the genetic sequences of 100 individuals from the Population 1 and 200 individuals from the Population 2. For each of these individuals, we have simulated around 2000 SNPs (we obtained them by simulating 500 sets of 1000 base pairs of nucleotides each). The Population 1 and the Population 2 were created from a split of an ancestral population. From the time of the population split to the present, we set a mutation rate per locus at 10^{-5} and the probability that a recombination event within the locus occurs in one generation is also set at 10^{-5}. The "coala" R package gives us the opportunity to modify the age of this split. Therefore, we are

going to simulate demographic models where the population split occurs at different times, so that we can study the evolution of the distance between the two populations projected onto the first principal component when the age of the split changes. Moreover, we have simulated 100 different genetic datasets so that our results do not rely on one simulated dataset.

3.3.2 Results

We have projected (onto the first principal component) the simulated genetic data of the 100 individuals from the Population 1 and the simulated genetic data of the 200 individuals from the Population 2. We have computed the distance between the two projected populations (the meaning of this distance has been illustrated in the Figure 3). We have computed this distance for demographic models in which there were migrations between the Population 1 and the Population 2 and for demographic models in which there were no migrations between the two populations, so that we can understand how migrations affect the relationship between the age of the split and the distance between the two populations projected onto the first principal component.

Figure 4: **The effect of the age of the population split on the distance between the two populations projected onto the first principal component** . In order to obtain the graph on the left, we have simulated two geographically separated populations without migrations between them. In order to obtain the graph on the right, we have simulated two geographically separated populations with symmetric migrations between them (10^{-3} corresponds to the fraction of each population that is replaced by migrants from the other population each generation since the population split). The genetic sequences (more precisely 2000 SNPs) of this demographic model have been simulated 100 times, generating 100 different genetic datasets. Therefore, each empirical distribution on these two graphs corresponds to the results from the 100 simulated genetic datasets. Moreover, we may notice that the two graphs have different y-axis scale.

On the graph on the left (Figure 4), we can see that the average distance between the two projected populations increases as the age of the population split increases. However, on the graph on the right (Figure 4), we can see that the age of the population split has no longer a clear effect on the distance between the two projected populations when there are migrations between these two geographically separated populations. We can explain this intuitively. If there were migrations between the two populations, it means that some individuals living in the population 2 have migrants (from the population 1) in their ancestry. Therefore, it means that some individuals, living in the population 2, have inherited some genetic sequences from migrants from the population 1 and some genetic sequences from individuals from the population 2. As there were migrations in both directions, there are also individuals living in the population 1, who have inherited some genetic sequences from migrants from the population 2 and some genetic sequences from individuals from population 1. This phenomenon creates similarities on a genetic level between the two populations (which is also consistent with the fact that the average distance between the two projected populations is globally smaller on the graph on the right compared to the graph on the left on the Figure 4). Thus, in the presence of migrations, there is no longer a clear relationship between the age of the population split and the distance between the two projected populations. Therefore, in the presence of migrations, we can no longer identify the age of a population split from the distance between the two populations projected onto the first principal component.

4 Population identification for non-admixed models

In this section, we are interested in finding the population of origin of individuals, that is what we call "population identification". For this section, we are going to work on non-admixed models, which means we assume that each individual has only one population of origin. We will deal with admixed models in the next section.

4.1 Model-based clustering STRUCTURE for population identification for non-admixed models

In order to search for the population of origin of individuals from their genetic data, Pritchard, Stephens and Donnelly (2000) [(1)] developed a model-based clustering called "STRUCTURE". This clustering was specifically made for using genotypes of individuals at different loci in order to find the population from which they originated. In our case, the genotypes were generated from the SCRM algorithm [(4)] (via the R package "Coala" [(5)]). Thus, the input of the clustering algorithm STRUCTURE is a dataset where each row corresponds to an individual's genotypes at different loci (each genotype can be coded by an integer).

Here we use the version of STRUCTURE without admixture [(1)], which means that we suppose that each individual has only one population of origin (we will use the version of STRUCTURE with admixture in one of the following chapter).

Let Z correspond to the populations of origin of the individuals in the sample, X correspond to the genotypes of the individuals in the sample, and P be the allele frequencies in every population.

For all individuals in the sample, we are looking for their populations of origin (Z) and their allele frequencies (P) given their (known) genotypes. Therefore, we are interested in the posterior distribution $Pr(Z, P|X)$. We know that:

$$Pr(Z, P \mid X) \propto Pr(Z)Pr(P)Pr(X \mid Z, P) \tag{3}$$

Before dealing with $Pr(Z)$, $Pr(P)$ and $Pr(X|Z, P)$, we have to define some notations. Let $(x_l^{(i,1)}, x_l^{(i,2)})$ correspond to the genotype of the ith individual at the lth locus ($i \in \{1, 2, .., N\}$ where N is the number of individuals in the sample, $l \in \{1, 2, .., L\}$ where L is the number of loci). Let p_{klj} frequency of allele j at locus l in population k ($k \in \{1, 2, .., K\}$ where K is the number of populations and $j \in \{1, 2, .., J_l\}$ where J_l is the number of different alleles observed at locus l). Let $z^{(i)}$ be the population of origin of the individual i.

Now, we are going to focus on $Pr(P)$. Here the allele frequencies correspond to a (multivariate) random variable. At the lth locus for a population k, the allele frequencies can be denoted by $p_{kl} = (p_{kl1}, p_{kl2}, .., p_{klJ_l})$. According to Balding and Nichols (1995) [(15)], a

good choice for the distribution of the allele frequencies might be the Dirichlet distribution. Therefore, independently for each p_{kl}, we have:

$$p_{kl} \sim \mathcal{D}(\lambda_1, \lambda_2, .., \lambda_{J_l}), \forall k \in \{1, 2, .., K\}, \forall l \in \{1, 2, .., L\} \tag{4}$$

Here, the authors have taken $\lambda_1 = \lambda_2 = .. = \lambda_{J_l} = 1$. It implies that the allele frequencies are uniformly distributed.

STRUCTURE assumes that genotypes belonging to individuals of a same population have the same distribution, which means that each data point (i.e. each individual's genotype) in a same cluster (i.e. in a same population) has the same distribution. Therefore, we define the likelihood $Pr(X|Z, P)$, independently for each $x_l^{(i,a)}$, as:

$$Pr(x_l^{(i,a)} = j|Z, P) = p_{z^{(i)}lj}, \forall l \in \{1, .., L\}, \forall j \in \{1, .., J_l\}, \forall a \in \{1, 2\} \tag{5}$$

We can notice that we said "independently for each $x_l^{(i,a)}$", which means that the clustering STRUCTURE assumes that, within populations, genotypes have been independently generated between loci (this is what we call complete linkage equilibrium between loci within populations). Moreover, we have to notice that the notation $p_{z^{(i)}lj}$ has been defined in the equation 4 as a random variable (which follows a specific prior distribution).

When it comes to $Pr(Z)$, as there is no relevant information about the population from which each individual originated, we may choose the following prior distribution (independently for each $z^{(i)}$):

$$Pr(z^{(i)} = k) = \frac{1}{K}, \forall k \in \{1, 2, .., K\}, \forall i \in \{1, 2, .., N\} \tag{6}$$

Therefore, we can compute the right-hand side of the result 3 but we are often unable to compute the normalising constant of $Pr(Z, P|X)$. Thus, we use the Markov Chain Monte Carlo (MCMC) method in order to generate samples $(Z^{(0)}, P^{(0)}), .., (Z^{(m)}, P^{(m)})$ from the posterior distribution $Pr(Z, P|X)$. The algorithm used for simulating this markov chain is a Gibbs sampler. With these samples, it is then possible to compute summary statistics like the posterior mean (by computing the mean of $Z^{(0)}, .., Z^{(m)}$ and the mean of $P^{(0)}, .., P^{(m)}$).

4.2 Problems of population identification for recently separated populations

4.2.1 Context

The demographic model we have simulated is explained on the Figure 5 .

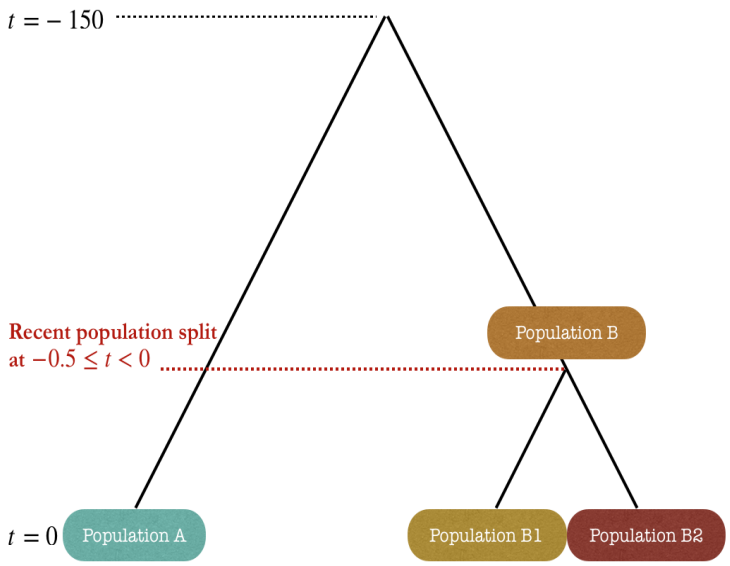


Figure 5: **Simplified diagram of a specific demographic model.** At time t=0 (now), there are 3 geographically separated populations: the population A, the population B1 and the population B2. There were no migrations between the 3 populations. The population B1 and the population B2 were created from the recent split of the population B.

On the Figure 5, we can see the recent split of the population B into two populations which are the population B1 and the population B2. We know that the more recent the split of the population B is, the more similar the genetic sequences of the population B1 and the population B2 will be.

Let t_S be the age of the split of the population B into the population B1 and the population B2. We want to know if for some t_S close to 0, the model-based clustering STRUCTURE (which we described in the previous section) constantly puts the individuals from the population B1 and the individuals from the population B2 in the same cluster. It would mean that for some t_S close to 0, STRUCTURE would not be able to distinguish between individuals from the population B1 and individuals from the population B2.

4.2.2 Simulating the data

Here are some details about how we have simulated our genetic data which we have used for assessing the ability of STRUCTURE to find individuals' true population of origin in the context of recently separated populations. We have simulated the genetic data by the coala R package [(5)]. The coala R package uses the simulator SCRM [(4)] in order to simulate the genetic variation corresponding to a specific demographic model. The demographic model that has been simulated is illustrated on the Figure 5: 3 geographically separated populations (the population A, the population B1 and the population B2) with no migrations between them, and where two populations (the population B1 and the population B2) have been created from a recent split (of the population B).

From t=-150 coalescent time units to the present, we set a mutation rate per locus at 10^{-5} and the probability that a recombination event within the locus occurs in one generation is also set at 10^{-5}. We have simulated the genetic sequences (more precisely around 1000 SNPs) of 500 individuals from the population A, 500 individuals from the population B1 and 500 individuals from the population B2.

We have simulated 100 different genetic datasets and have used the clustering STRUCTURE on each of these 100 genetic datasets so that our results do not rely on one simulated dataset.

4.2.3 Results

In order to assess the ability of the clustering STRUCTURE to find individuals' true population of origin, we have used the Rand Index. The Rand Index is defined in the following way:

$$Rand\ \ Index = \frac{x_1 + x_2}{x_1 + x_2 + x_3 + x_4} = \frac{x_1 + x_2}{\binom{n}{2}} \tag{7}$$

where x_1 is the number of pairs of individuals who have the same population of origin and are in the same cluster, where x_2 is the number of pairs of individuals who have different populations of origin and are in different clusters, where x_3 is the number of pairs of individuals who have different populations of origin but are in the same cluster, where x_4 is the number of pairs of individuals who have the same population of origin but are in different clusters, and where n is the number of individuals in the sample.

Figure 6: **Empirical distribution of the Rand Index (which assesses the ability of the clustering STRUCTURE to find individuals' true population of origin) for 3 different t_S (where t_S is the age of the split of the population B into the population B1 and B2).** The demographic model that has been simulated is illustrated on the Figure 5. We have simulated the genetic sequences (more precisely around 1000 SNPs) of 500 individuals from the population A, 500 individuals from the population B1 and 500 individuals from the population B2. Then, we have used the model-based clustering STRUCTURE on these genetic data in order to find the population of origin of the individuals in the sample. We have assumed in this clustering that there must be 3 clusters. In the computation of each Rand Index, we have removed the individuals from population A because we want this Rand Index to reflect only the ability of STRUCTURE to distinguish between the individuals from the population B1 and the individuals from the population B2. We have repeated 100 times this process using each time a different simulated genetic dataset, which has given the empirical distribution that we see on this Figure.

In the computation of each Rand Index, we have removed the individuals from population A because we want this Rand Index to reflect only the ability of the clustering STRUCTURE to distinguish between the individuals from the population B1 and the individuals from the population B2. Therefore, the worst case scenario would be that all the individuals from the population B1 and all the individuals from the population B2 are in one single cluster. Given that there are 500 individuals from the population B1 and 500 individuals from the population B2 in the sample, the Rand Index corresponding to the worst case scenario would be equal to: $\frac{x_1+x_2}{\binom{n}{2}} = \frac{\binom{500}{2}+\binom{500}{2}+0}{\binom{500+500}{2}} = 0.4994995$. It explains why we do not see any Rand Index value smaller than 0.4994995 on the Figure 6.

Moreover, a Rand Index equal to 1 means that all the individuals from the population B1 are in one cluster and all the individuals from the population B2 are in another cluster, which would mean that from the genetic data the clustering STRUCTURE has perfectly

distinguished the individuals from the population B1 from the individuals from the population B2. We know that when the Rand Index gets closer to 1, it means that STRUCTURE identifies better individuals' true population of origin.

We can see on the Figure 6 that when the age of the split of the population B (into the population B1 and the population B2) is set at 0.3 coalescent time units, the median Rand Index is approximately equal to 0.94 and the first quartile is approximately equal to 0.83, which means that 75% of the 100 Rand Indexes (computed on 100 simulated genetic datasets) are higher than 0.83. Therefore, when the age of the split of the population B (into the population B1 and the population B2) is set at 0.3 coalescent time units, STRUCTURE has a good ability to distinguish the individuals from the population B1 from the individuals from the population B2.

On the Figure 6, we can notice that when the age of the split of the population B (into the population B1 and the population B2) is set at 0.2 coalescent time units, the median Rand Index is approximately equal to 0.69 and the third quartile is approximately equal to 0.75, which means that 75% of the 100 Rand Indexes (computed on 100 simulated genetic datasets) are smaller than 0.75. We know that a lower value of the Rand Index indicates that STRUCTURE makes more mistakes on the population of origin of individuals. Therefore, the fact that the split of the population B occurred at 0.2 coalescent time units in the past instead of occurring at 0.3, has negatively affected the ability of STRUCTURE to distinguish the individuals from the population B1 from the individuals from the population B2.

On the Figure 6, we have observed that when the age of the split of the population B (into the population B1 and the population B2) is set at 0.1 coalescent time units, there is a large number of simulations for which the Rand Index is approximately equal to 0.5 and the median Rand Index is approximately equal to 0.51 . Moreover, we have looked at each output of STRUCTURE for each of the 100 simulations, and in 43 simulations (when the age of the split was set at 0.1 coalescent time units) all the individuals from the population B1 and all the individuals from the population B2 were put in one single cluster. These 43 simulations gave the same Rand Index: 0.4994995 . Therefore, when the age of the split goes from 0.3 to 0.1 coalescent time units, the ability of STRUCTURE to distinguish between the two populations derived from the split is drastically deteriorated.

5 Population identification for admixed models

In a realistic representation of our world, individuals have several populations of origin. Individuals who have at least two populations of origin are called admixed individuals. Pritchard, Stephens and Donnelly (2000) [(1)] also developed methods for finding for every individual all their populations of origin. These authors have created another version of the model-based clustering STRUCTURE for admixed models.

5.1 Version of STRUCTURE with admixture: model-based clustering for population identification in admixed models

In admixed models, individuals may have two (or more) populations of origin. A version of STRUCTURE was developed for admixed models [(1)]. This admixed version of STRUCTURE enables to find all the populations of origin of individuals and it also estimates the admixture proportions (i.e. the relative contribution from each population of origin to an individual's ancestry) of individuals.

Each admixed individual has some alleles at some positions of his genome that originated from a population and has other alleles at other positions of his genome that originated from other populations. Therefore, Z has no longer the same meaning as in the version of STRUCTURE for non-admixed models (see the section on STRUCTURE for non-admixed models). In this admixed version of STRUCTURE, Z corresponds to: for each individual i, $z^{(i)} = (z_l^{(i,1)}, z_l^{(i,2)})_{l\in\{1,\ldots,L\}}$ where $z_l^{(i,1)}$ corresponds to the population of origin of the copy 1 of the allele at the lth locus (and $z_l^{(i,2)}$ corresponds to the population of origin of the copy 2 of the allele at the lth locus). The reason why we say "copy 1" of the allele and "copy 2" of the allele (at the lth locus) is because every human has two copies of each of their alleles.

In the version of STRUCTURE for admixed models, there is an additional multivariate random variable to infer (compared to the non-admixed version of STRUCTURE). This additional multivariate random variable is Q which is defined in the following way: for each individual i, $q^{(i)} = (q_k^{(i)})_{k\in\{1,\ldots,K\}}$ where $q_k^{(i)}$ is the proportion of the genome (of the individual i) which originated from the population k.

Recall that X corresponds to the genotypes of the individuals in the sample, and P corresponds to the allele frequencies in every population.

The posterior distribution of (Q, Z, P) can be written in the following way:

$$Pr(Q, Z, P|X) \propto Pr(X|Z, Q, P)Pr(Z|Q, P)Pr(Q)Pr(P) \tag{8}$$

We are going to explain how the authors of STRUCTURE have defined $Pr(X|Z, Q, P)$, $Pr(Z|Q, P)$, $Pr(Q)$ and $Pr(P)$. The prior distribution $Pr(P)$ has been defined in the same way as in the non-admixed version of STRUCTURE, and therefore has been already defined in the mathematical expression 4.

When it comes to the prior distribution $Pr(Q)$, the authors of STRUCTURE have chosen a Dirichlet distribution: independently for each individual i,

$$q^{(i)} = (q_k^{(i)})_{k\in\{1,...,K\}} \sim D(\alpha, \alpha, ..., \alpha) \tag{9}$$

We will explain after how α is chosen.

The distribution $Pr(Z|P,Q)$ has been defined in the following way: independently for each individual i and each locus l,

$$Pr(z_l^{(i,a)} = k|Q,P) = q_k^{(i)},\ \forall k \in \{1,2,..,K\}, \forall a \in \{1,2\} \tag{10}$$

The equation 10 seems logical as it says that (for an individual i) the probability that the allele copy at the lth locus originated from the population k given P and Q is equal to the proportion of the genome (of this individual i) which originated from the population k.

Therefore, the likelihood is given by: for each individual i and each locus l,

$$Pr(x_l^{(i,a)} = j|Z,Q,P) = p_{z_l^{(i,a)}lj},\ \forall j \in \{1,..,J_l\}, \forall a \in \{1,2\} \tag{11}$$

We can compute the right-hand side of the result 8 but we are often unable to compute the normalising constant of $Pr(Q,Z,P|X)$. Thus, we use the Markov Chain Monte Carlo (MCMC) method in order to generate a sample $(Z^{(0)},P^{(0)},Q^{(0)}),..,(Z^{(m)},P^{(m)},Q^{(m)})$ from the posterior distribution $Pr(Q,Z,P|X)$.

Here is the pseudo-code for the MCMC algorithm.

Algorithm 1 MCMC algorithm for the version of STRUCTURE with admixture

input : starting value $Z^{(0)}$
output: samples $(Z^{(0)},P^{(0)},Q^{(0)}),..,(Z^{(m)},P^{(m)},Q^{(m)})$ from the posterior distribution $Pr(Q,Z,P|X)$

for *t = 1,2,...* **do**
  Draw: $(Q^{(t)},P^{(t)}) \sim Pr(Q,P|X,Z^{(t-1)})$
  Draw: $Z^{(t)} \sim Pr(Z|X,P^{(t)},Q^{(t)})$
  Update α (Metropolis-Hasting step)
end

After looking at the pseudo-code 1, we can now understand how α is chosen in the mathematical expression 9. In this pseudo-code, we can see that the information from $Q^{(t)},P^{(t)},Z^{(t)}$ and the data X is used to update α at the iteration t via a Metropolis-Hastings step.

The MCMC gives us a sample $(Z^{(0)},P^{(0)},Q^{(0)}),..,(Z^{(m)},P^{(m)},Q^{(m)})$ from the posterior distribution $Pr(Q,Z,P|X)$. Q, Z and P are inferred using their posterior mean: the posterior mean of Z can be calculated by computing the mean of the sample $(Z^{(0)},...,Z^{(m)})$, and we can compute the posterior mean of P and Q in the same way.

Thus, for each individual i, STRUCTURE computes the posterior mean of the proportion of the genome (of the individual i) which originated from the population k, for every $k \in \{1, .., K\}$. Therefore, STRUCTURE can estimate the admixture proportions (i.e. the relative contribution from each population of origin to an individual's ancestry) of each individual.

5.2 Problems of admixture detection in the presence of unidirectional migrations

5.2.1 Context

With these two simplified diagrams (Figure 7), we can explain the differences (in terms of consequences on populations' genetics) between unidirectional migrations and bi-directional migrations.

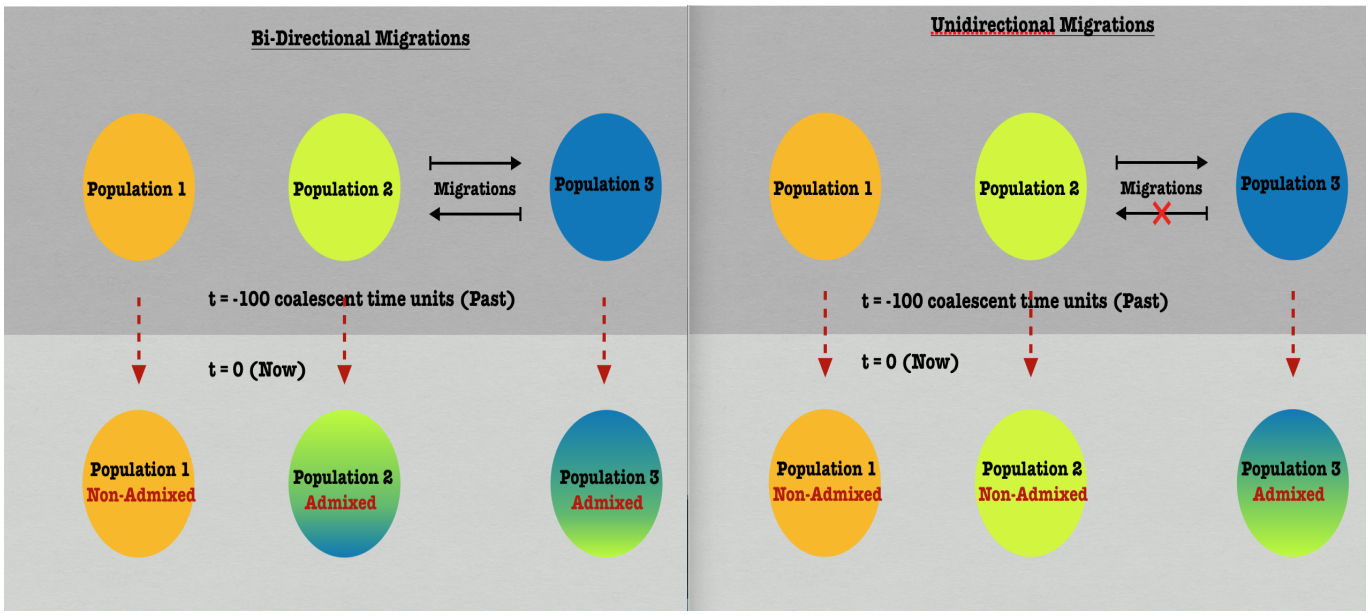


Figure 7: **Simplified diagrams of the effect of bi-directional and unidirectional migrations in terms of admixture.** The 3 populations are geographically separated. The upper part of each diagram corresponds to the past and the bottom part of each diagram corresponds to the present.

On the diagram on the left (Figure 7), we can see that the population 2 and 3 have become admixed because of the migrations in both directions. We can explain now why migrations are responsible for creating admixture. If there were migrations from the population 3 to the population 2, it means that some individuals living in the population 2 have migrants (from the population 3) in their ancestry. Therefore, it means that some individuals, living in the population 2, have inherited genetic sequences from migrants from the population 3 and genetic sequences from individuals from the population 2. Therefore, migrations are responsible for creating admixture.

On the diagram on the right (Figure 7) corresponding to the unidirectional migrations, we can notice that the population 3 has become admixed but it is not the case for the population 2. The reason why the population 2 is non-admixed is because migrations are going only in one direction from the population 2 to the population 3. Therefore, if we get a sample of individuals living in the population 3, their populations of origin may be the population 2 and the population 3 (because some of them have migrants in their ancestry). However, if we get a sample of individuals living in the population 2, all of the individuals in that sample can only have a unique population of origin which is the population 2 (because there were no migrations from the population 3 to the population 2). Nevertheless, having simulated the genetic sequences of this demographic model, STRUCTURE (used for identifying the populations of origin of individual from their genetic sequences) has given us consistently results saying that some individuals living in the population 2 originated from the population 2 and the population 3, which is impossible because there were no migrations from the population 3 to the population 2. We are going to show these results in one of the following sections.

5.2.2 Simulating the data

Here are some details about how we have simulated our genetic data which we have used for assessing the ability of STRUCTURE (the version of STRUCTURE with admixture) to find individuals' true populations of origin in the presence of unidirectional migrations. We have simulated the genetic data by the coala R package [(5)]. The coala R package uses the simulator SCRM [(4)] in order to simulate the genetic variation corresponding to a specific demographic model. We have simulated the genetic sequences of 3 geographically separated populations where there were unidirectional migrations from the population 2 to the population 3 (so that 10^{-3} is the fraction of the population 3 that is replaced by migrants from the population 2 each generation since t=-100 coalescent time units). From t=-100 coalescent time units to the present, we set the mutation rate per locus at 10^{-5} and the probability that a recombination event within the locus occurs in one generation is also set at 10^{-5}. We have generated the genetic data of a sample of 1000 individuals for each of the 3 populations. For each individual, 1000 SNPs are simulated.

We have simulated 100 different genetic datasets and have used the clustering STRUCTURE on each of these 100 genetic datasets so that our results do not rely on one simulated dataset.

5.2.3 Results

The model-based clustering STRUCTURE (the version of STRUCTURE with admixture) gives us the percentage of membership in each cluster for each individual. Each cluster corresponds to one population. Moreover, if for instance the population 1 corresponds to the cluster 2, the population 2 corresponds to the cluster 3 and the population 3 corresponds to the cluster 1, we have to understand the output of the clustering STRUCTURE in the following way: if for one individual the output of STRUCTURE is (0.6,0.3,0.1), it means that 60% of this individual's genome originated from the population 3, 30% of this individual's genome originated from the population 1, and 10% of this individual's genome originated

from the population 2.

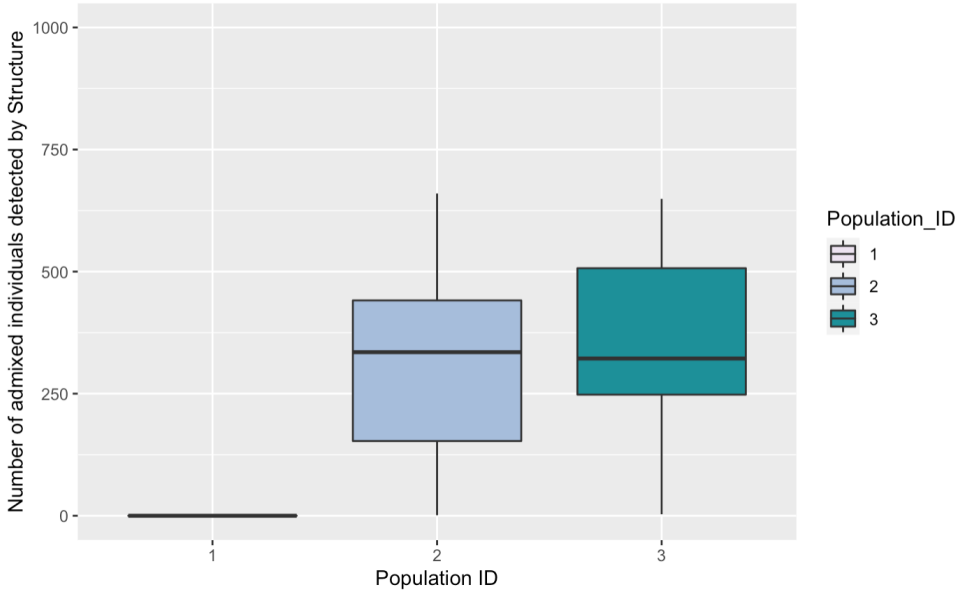


Figure 8: **Empirical distribution of the number of admixed individuals detected by the model-based clustering STRUCTURE.** We have simulated the genetic sequences of 3 geographically separated populations where there were unidirectional migrations from the population 2 to the population 3 (so that 10^{-3} is the fraction of the population 3 that is replaced by migrants from the population 2 each generation since t=-100 coalescent time units). We have sampled 1000 individuals in each of the 3 populations. For all these individuals, we have simulated 1000 SNPs. The clustering STRUCTURE uses these genetic sequences in order to find the populations of origin of the individuals. Here, this empirical distribution (on these boxplots) has been built from 100 different results given by STRUCTURE which has been used on 100 simulated genetic datasets.

We are going to clarify how we identified admixed individuals in the ouput of STRUCTURE. If less than 5% of one individual's genome originated from the population i, we should not say the population i is one of the populations of origin of this individual because it is more cautious to think that there might be a margin of error of 5% in the results from STRUCTURE. Therefore, for instance, if for one individual the output of STRUCTURE is (0.96,0.00,0.04), we will not say that this individual is admixed because only 4% of this individual's genome originated from the population corresponding to the cluster 3 so this population is not counted as one of the populations of origin of this individual.

On the Figure 8, we can notice that in the population 1 there are no admixed individuals detected by STRUCTURE, which was expected as there were no immigration in the population 1. On the Figure 8, we can also see that among the 1000 individuals in the population 3, the median number of admixed individuals detected by STRUCTURE is around 340. In the population 3, the first quartile of the empirical distribution of the number of admixed individuals detected by STRUCTURE is around 250 and its third quartile is around 510. As there were migrations from the population 2 to the population 3, it was expected that

STRUCTURE detects admixed individuals in the population 3. Moreover, having looked at each ouput of STRUCTURE, we have noticed that the populations of origin of all these admixed individuals living in the population 3 are the population 2 and the population 3, which is consistent with the fact that there were migrations from the population 2 to the population 3.

However, on the Figure 8, we can see the empirical distribution of the number of admixed individuals detected by STRUCTURE in the population 2, the median number of detected admixed individuals being around 340 among the 1000 individuals in the population 2. It means that in each of the 100 simulated genetic dataset, STRUCTURE has detected each time a significant number of admixed individuals in the population 2. Nevertheless, we know that there were no admixed individuals in the population 2 because there were no immigration in the population 2. Having looked at the results given by STRUCTURE, these results say that 100% of these detected admixed individuals in the population 2 originated from the population 2 and the population 3. Therefore, STRUCTURE behaved as if there were migrations from the population 3 to the population 2, which is not the case. We are going to explain this phenomenon intuitively. As there were migrations from the population 2 to the population 3, it means that some individuals living in the population 3 have migrants (from the population 2) in their ancestry. Therefore, it means that some individuals, living in the population 3, have inherited genetic sequences from migrants from the population 2 and genetic sequences from individuals from the population 3. Therefore, these genetic sequences from migrants from the population 2 and these genetic sequences from individuals from the population 3 have been recombined (the concept of recombination has been explained in the section "Basic knowledge in Genetics"). Thus, the genetic sequences of individuals living in the population 3 have become more and more similar to the genetic sequences of individuals living in the population 2, which may be the reason why STRUCTURE behaved as if there were migrations on both directions between the population 2 and the population 3.

6 Population identification using real data

The data were provided by Dr Christopher Hallsworth and were collected for the 1000 Genomes Project [(16)]. The dataset contains 2504 rows and 1105538 columns. Each column corresponds to a specific SNP of the chromosome 21 and each row corresponds to an individual. Therefore, in this dataset, each individual is characterized by his alleles at 1105538 SNPs of the chromosome 21. Each allele is either coded by 0 or 1.

Moreover, for all the individuals in the dataset, we know the region they come from: East Asia, South Asia, Africa, America or Europe. Therefore, as we already know the population of origin of each individual, we are able to assess the ability of a statistical method to predict correctly the population of origin of each individual using these real genetic data.

As in the previous sections we have worked with a limited amount of data (around 1000 SNPs), we asked ourselves a simple question: for a given statistical method, to what extent its ability to find individuals' true population of origin is impacted by the amount of available genetic variation data.

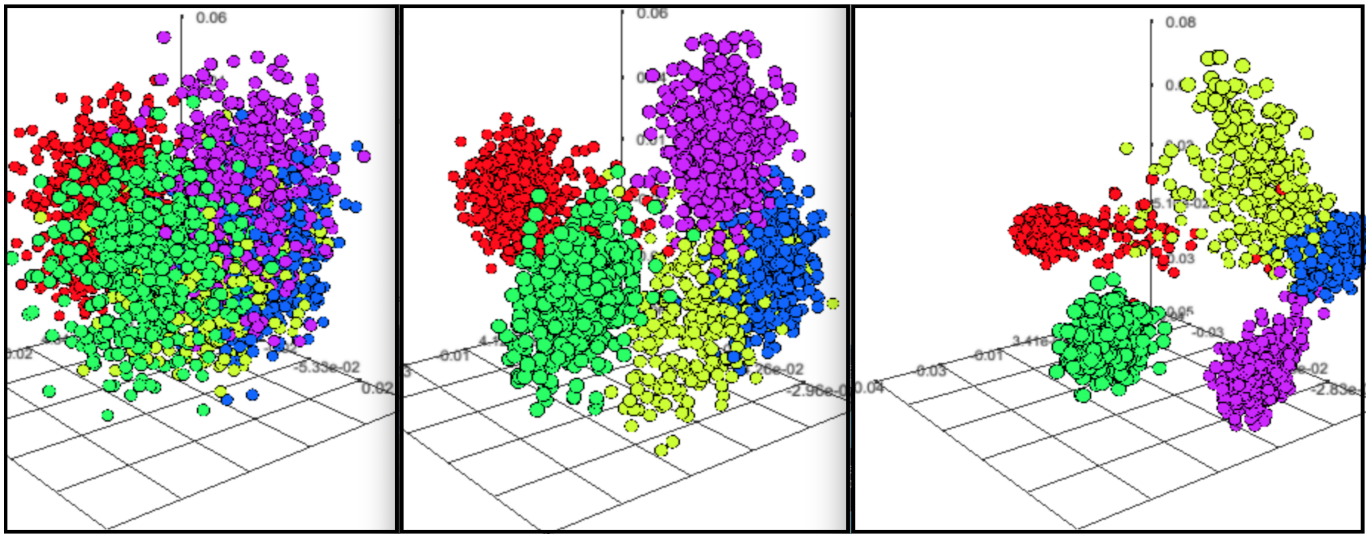


Figure 9: **The effect of limited genetic data represented by a 3-dimensional PCA.** On the three graphs, each point corresponds to one individual and each color corresponds to their population of origin. The graph on the left has been obtained by sampling randomly (without replacement) 500 SNPs from the 1105538 SNPs (from our real data) and applying PCA on this smaller dataset of 500 SNPs. The graph on the middle has been obtained by sampling randomly (without replacement) 5000 SNPs from the 1105538 SNPs (from our real data) and applying PCA on this smaller dataset of 5000 SNPs. The graph on the right has been obtained by sampling randomly (without replacement) 1000000 SNPs from the 1105538 SNPs (from our real data) and applying PCA on this smaller dataset of 1000000 SNPs.

The Figure 9 highlights how easier population identification can be when we have access to a large number of SNPs.

Our dataset contains 1105538 SNPs, so we are going to randomly sample (without replacement) 100 SNPs from those 1105538 SNPs creating a smaller dataset and we are going to perform population identification on this smaller dataset. Then, we are going to randomly sample (without replacement) 500 SNPs from those 1105538 SNPs and perform population identification on this smaller data. Then, we have repeated these same steps, creating datasets of 2000 SNPs, 5000 SNPs and 100 000 SNPs.

However, performing population identification on two randomly selected samples of 500 SNPs (from the 1105538 SNPs) may give two different results. Therefore, we have randomly selected 100 different samples of 500 SNPs (from the 1105538 SNPs), creating 100 different datasets of 500 SNPs. In the same way, we have created 100 different datasets of 100 SNPs, 2000 SNPs, 5000 SNPs and 100 000 SNPs.

In order to perform population identification on each dataset, we have used PCA which has enabled us to transform our data into 3-dimensional data. Then, we have applied the K-Means (where K=5 because there are 5 populations in the data: European, East Asian, South Asian, African and American) on the 3-dimensional data to perform population identification.

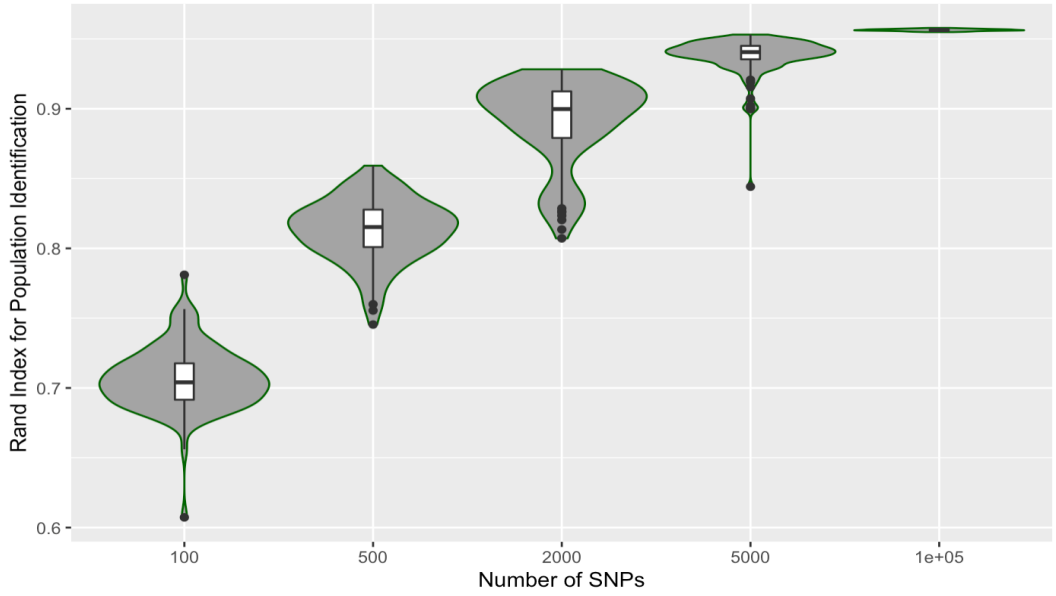


Figure 10: **The effect of the amount of genetic variation data on the ability of our statistical method to find individuals' true population of origin.** We have randomly selected 100 different samples of 500 SNPs from the 1105538 SNPs (from our real data), creating 100 different datasets of 500 SNPs. In the same way, we have created 100 different datasets of 100 SNPs, 2000 SNPs, 5000 SNPs and 100 000 SNPs. On each dataset, we have used a 3-dimensional PCA, and we have applied the K-Means (where K=5) on the 3-dimensional data to perform population identification. In order to assess the ability of this statistical method to find individuals' true population of origin, we have used the Rand Index.

In order to assess the ability of this statistical method to find individuals' true population of origin, we have used the Rand Index (which we have explained in one of the previous sections). We know that when the Rand Index gets closer to 1, it means that the corresponding statistical method identifies better individuals' true population of origin. If the Rand Index is equal to 1, it means that the corresponding statistical method has perfectly identified (i.e. has made no mistakes) individuals' true population of origin.

On the Figure 10 , we can see that the median Rand Index of our statistical method using only 100 SNPs (from the 1105538 SNPs) is approximately equal to 0.70. The median Rand Index of our statistical method using 500 SNPs (from the 1105538 SNPs) is approximately equal to 0.82. Moreover, the median Rand Index of our statistical method using 2000 SNPs (from the 1105538 SNPs) is approximately equal to 0.90. Therefore, it shows that the ability of our statistical method (to identify individuals' true population of origin) has greatly improved when it uses 500 SNPs rather than 100 SNPs, or when it uses 2000 SNPs rather than 500 SNPs.

However, from 2000 to 5000 SNPs, the median Rand Index did not increase significantly (from 0.90 to 0.94). It is even more the case going from 5000 to 100 000 SNPs (the median Rand Index went from 0.94 to 0.96). Nevertheless, from 5000 to 100 000 SNPs, the standard deviation of the empirical distribution of the Rand Indexes has drastically lowered. All the Rand Index values for datasets of 100 000 SNPs are approximately equal to 0.96. Therefore, every sample of 100 000 SNPs enables the statistical method to perform well, which is not the case for every sample of 2000 SNPs.

7 Conclusion

In the section called "Identifying the age of a population split", we have simulated genetic variation data corresponding to a demographic model with migrations between populations, and have applied Principal Component Analysis on these simulated genetic data. We have repeated this step 100 times, creating 100 different simulated genetic datasets so that our analysis does not rely on one simulated dataset. The results from this experiment have shown us that, in the presence of migrations, there is no longer a clear relationship between the age of the population split and the distance between the genetic sequences of the two populations projected onto the first principal component. Therefore, in the presence of migrations, we can no longer identify the age of a population split using only the distance between the genetic sequences of the two populations projected on the first principal component.

In the section called "Population identification for non-admixed models", we have simulated genetic variation data corresponding to a demographic model where one population has recently split into two different populations. We have applied the model-based clustering STRUCTURE on these simulated genetic data. This experiment has shown that when the age of the split goes from 0.3 to 0.1 coalescent time units, the ability of STRUCTURE to distinguish between the two populations derived from the split is drastically deteriorated.

In the section called "Population identification for admixed models", we have simulated genetic variation data corresponding to a demographic model where there are only unidirectional migrations from one population to the other one. We have applied STRUCTURE (more specifically the version of STRUCTURE for admixed models) on these simulated genetic data. This experiment has shown that STRUCTURE behaved as if there were migrations on both directions between the two populations.

We are aware of the fact that we have used a limited amount of genetic variation data for these previous experiments (between 1000 and 2000 SNPs per simulation). It could be therefore interesting to do the same experiments using a larger amount of genetic variation data. Moreover, it could be also interesting to do the same experiments with real data.

In the section called "Population identification using real data", we have randomly selected 100 different samples of 500 SNPs from the 1105538 SNPs (from the original real data), creating 100 different datasets of 500 SNPs. In the same way, we have created 100 different datasets of 100 SNPs, 2000 SNPs, 5000 SNPs and 100 000 SNPs. On each dataset, we have used a 3-dimensional PCA, and we have applied the K-Means (where K=5) on the 3-dimensional data to perform population identification. This experiment has shown that the ability of this statistical method (to identify individuals' true population of origin) has greatly improved when it uses 500 SNPs rather than 100 SNPs, or when it uses 2000 SNPs rather than 500 SNPs. However, when we used 5000 SNPs instead of 2000 SNPs, or when we used 100 000 SNPs instead of 5000 SNPs, the improvement was in the consistency of the performance: every sample of 100 000 SNPs enabled the statistical method to perform well, which was not the case for every sample of 2000 SNPs for instance.

8 Appendix

8.1 Identifying the age of a population split

Here is the main part of the code corresponding to the section "Identifying the age of a population split".

```

library(coala)
library(phyclust)

N0=10000
L=1000
mu=10^(-8)*L #where mu is the mutation rate per locus.
r=10^(-8)*L #where r is the probability that a recombination event within the locus occurs in one generation


# For a given age of the population split and a given migration rate, this function computes the distance between the two projected populations onto the first principal component

PCA_distance <-function (merge_time , migration_rate,seed_sim) {

set.seed(seed_sim)

# Here we simulate the genetic variation data via the package Coala:
# We sample 500 sets of 1000 base pairs for 100 individuals from the population 1 and 200 individuals from the population 2

model <- coal_model(sample_size = c(100,200), loci_number = 500 , loci_length = 1000) +

  feat_pop_merge(time=merge_time, 1, 2) +
  feat_migration(rate=migration_rate, symmetric = TRUE) +
  feat_mutation(rate = 4*N0*mu) +
  feat_recombination(rate = 4*N0*r, locus_group = "all") +

  sumstat_seg_sites() +
  sumstat_trees()


sumstats0 <- simulate(model)

# We collect all the SNPs from the 500 different loci:

SNPmatrix0=sumstats0$seg_sites[[1]]$snps
for(i in 2:500){
  SNPmatrix0=cbind(SNPmatrix0,sumstats0$seg_sites[[i]]$snps)
}


```

```
# We project the genetic sequences of the two populations onto the first principal component
PCA_g=prcomp(t(SNPmatrix0), scale = FALSE, rank.=1)

# The first 100 rows correspond to the 100 individuals from population 1
# The next 200 rows correspond to the 200 individuals from population 2
# We compute the mean of each projected population which can be seen as centroids
# Then, we compute the distance between these two centroids:
return( abs( mean(PCA_g$rotation[1:100,]) - mean(PCA_g$rotation[101:300,]) ) )

}

# Here we compute the distance between the two projected populations when the age of the split is set at 0.2 coalescent time units.

centroid_dist_mig02=rep(0,100)
for(i in 1:100){
centroid_dist_mig02[i]=PCA_distance(merge_time=0.2 , migration_rate=0,seed_sim=1226*i)
}
mean_dist_mig0_merge02=sum(centroid_dist_mig02)/length(centroid_dist_mig02)
sd_dist_mig0_merge02=sd(centroid_dist_mig02)


# Here, we introduced migrations: we compute the distance between the two projected populations when the age of the split is set at 0.4 coalescent time units in the presence of migrations.


m=10^(-3)
centroid_dist_migrations_04=rep(0,100)
for(i in 1:100){
centroid_dist_migrations_04[i]=PCA_distance(merge_time=0.4 , migration_rate=4*N0*m ,seed_sim=1226*i)
}
mean_dist_migLow_merge04=sum(centroid_dist_migrations_04)/length(centroid_dist_migrations_04)
sd_dist_migLow_merge04=sd(centroid_dist_migrations_04)
```

8.2 Population identification for non-admixed models

8.2.1 Simulating the genetic variation data

Here is the code corresponding to one genetic variation dataset that we have simulated for the section "Population identification for non-admixed models".

```
1
2
3  library(coala)
4
5
6  N0=10000
7  L=1000
8  mu=10^(-8)*L #where mu is the mutation rate per locus.
9  r=10^(-8)*L #where r is the probability that a recombination event within
       the locus occurs in one generation
10
11
12 set.seed(9969)
13
14 # Here we simulate the genetic variation data via the package Coala:
15 # We sample 5 sets of 10000 base pairs for 500 individuals from each of
       the three populations
16
17 # Here, via the command "feat_pop_merge(time=0.3, 3, 2)" , the age of the
       split between the population 2 and the population 3 has been set at 0.3
        coalescent time units.
18
19 model1 <- coal_model(sample_size = c(500,500,500), loci_number = 5 , loci_
       length = 10000 ) +
20
21   feat_pop_merge(time=150, 1, 2) +
22   feat_pop_merge(time=0.3, 3, 2) +
23   feat_mutation(rate = 4*N0*mu) +
24   feat_recombination(rate = 4*N0*r, locus_group = "all") +
25
26   sumstat_seg_sites() +
27   sumstat_trees()
28
29 # We simulate the model
30 sumstats1 <- simulate(model1)
31
32 # We collect the SNPs
33
34 SNPmatrix = cbind(sumstats1$seg_sites[[1]]$snps,sumstats1$seg_sites[[2]]$
       snps,sumstats1$seg_sites[[3]]$snps,sumstats1$seg_sites[[4]]$snps,
       sumstats1$seg_sites[[5]]$snps)
35
36
37 # We export the dataset which contains the SNPs
38 write.table(SNPmatrix, file = "coalNoAdmix_03_7.txt", sep = " " , col.
       names=TRUE)
```

8.2.2 Using STRUCTURE (without admixture)

Here are the parameters that we have chosen in STRUCTURE for the section "Population identification for non-admixed models".

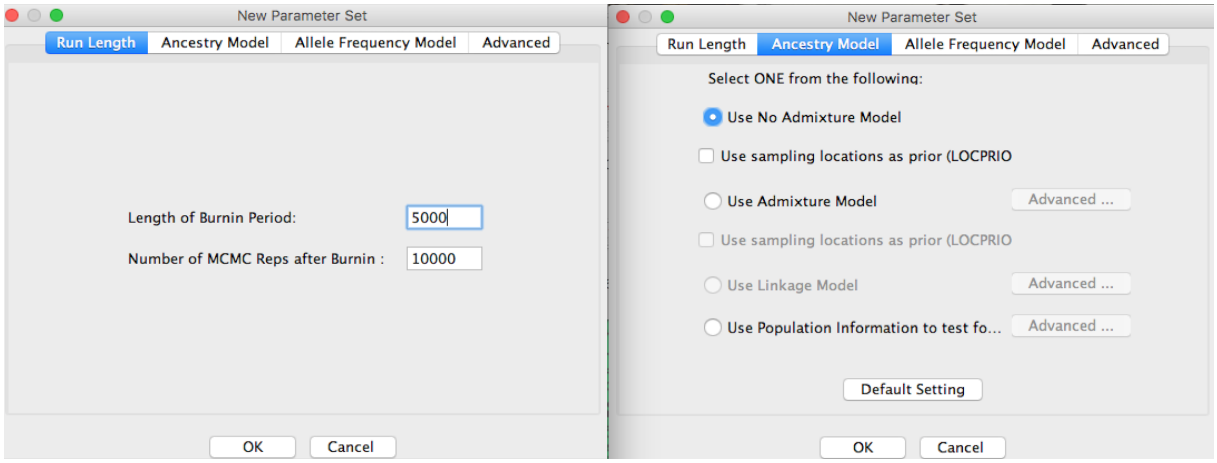


Figure 11: **Parameters chosen in STRUCTURE for the section "Population identification for non-admixed models"**

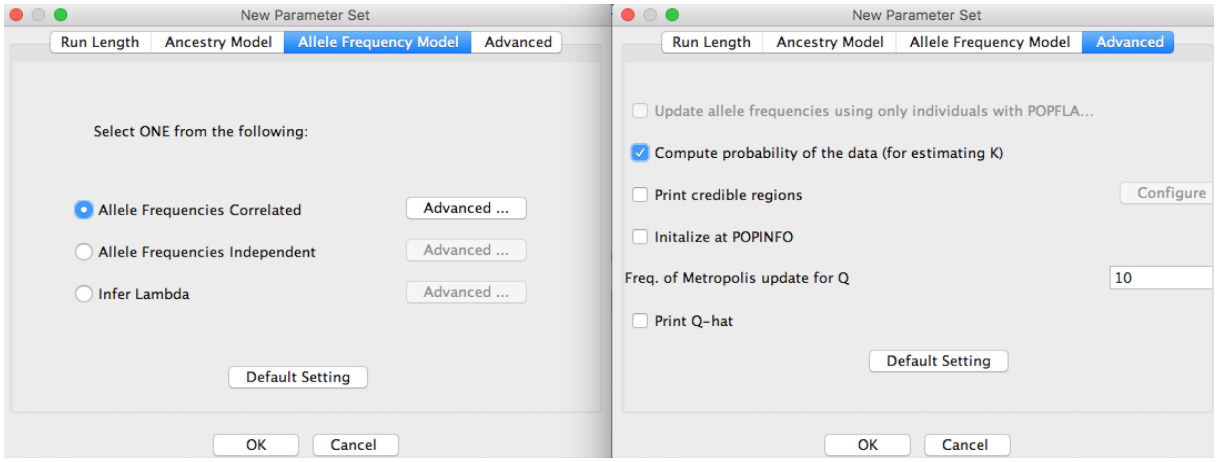


Figure 12: **Parameters chosen in STRUCTURE for the section "Population identification for non-admixed models"**

8.3 Population identification for admixed models

8.3.1 Simulating the genetic variation data

Here is the code corresponding to one genetic variation dataset that we have simulated for the section "Population identification for admixed models".

```
1
2
3  library(coala)
4
5  N0=10000
6  L=1000
7  mu=10^(-8)*L #where mu is the mutation rate per locus.
8  r=10^(-8)*L #where r is the probability that a recombination event within
       the locus occurs in one generation
9
10
11 set.seed(5398)
12
13 m=10^(-4)
14
15 # We simulate the genetic variation data via the package Coala:
16 # We sample 5 sets of 10000 base pairs for 1000 individuals from each of
       the three populations
17
18 # Here, via the command "feat_migration(rate=4*N0*m,pop_from=2,pop_to=3)",
        we have set unidirectional migrations from the population 2 to the
       population 3
19
20
21 model1 <- coal_model(sample_size = c(1000,1000,1000), loci_number = 5 ,
       loci_length = 10000 ) +
22
23   feat_pop_merge(time=100, 1, 2) +
24   feat_migration(rate=4*N0*m , pop_from = 2 , pop_to = 3 ) +
25   feat_mutation(rate = 4*N0*mu) +
26   feat_recombination(rate = 4*N0*r, locus_group = "all") +
27
28   sumstat_seg_sites() +
29   sumstat_trees()
30
31
32 # We simulate the model
33 sumstats1 <- simulate(model1)
34
35 # We collect the SNPs
36 SNPmatrix = cbind(sumstats1$seg_sites[[1]]$snps,sumstats1$seg_sites[[2]]$
       snps,sumstats1$seg_sites[[3]]$snps,sumstats1$seg_sites[[4]]$snps,
       sumstats1$seg_sites[[5]]$snps)
37
38
39
40 # We export the dataset which contains the SNPs
```

```
write.table(SNPmatrix, file = "coalAdmix3_2.txt", sep = " " , col.names=
    TRUE)
```

8.3.2 Using STRUCTURE (with admixture)

Here are the parameters that we have chosen in STRUCTURE for the section "Population identification for admixed models".

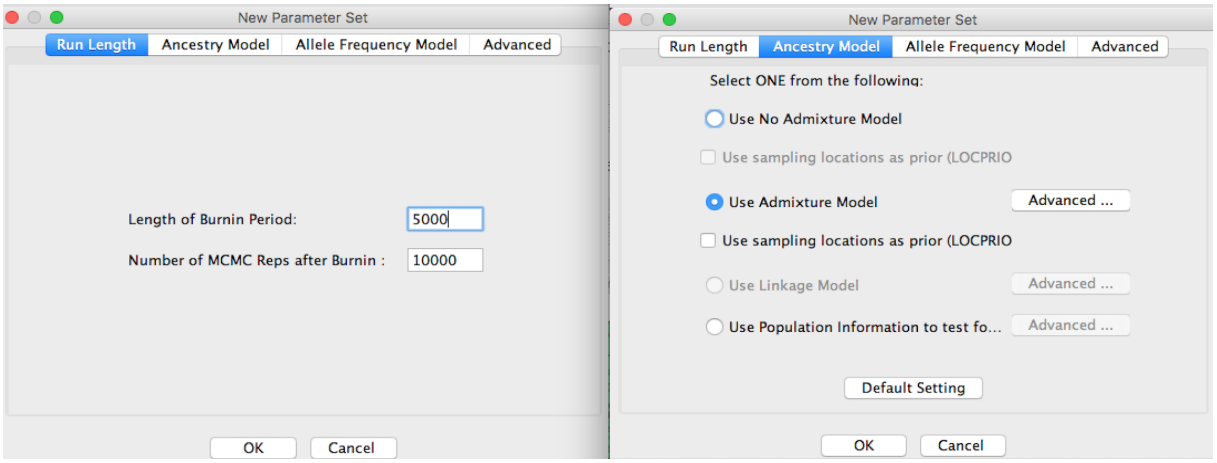


Figure 13: **Parameters chosen in STRUCTURE for the section "Population identification for admixed models"**

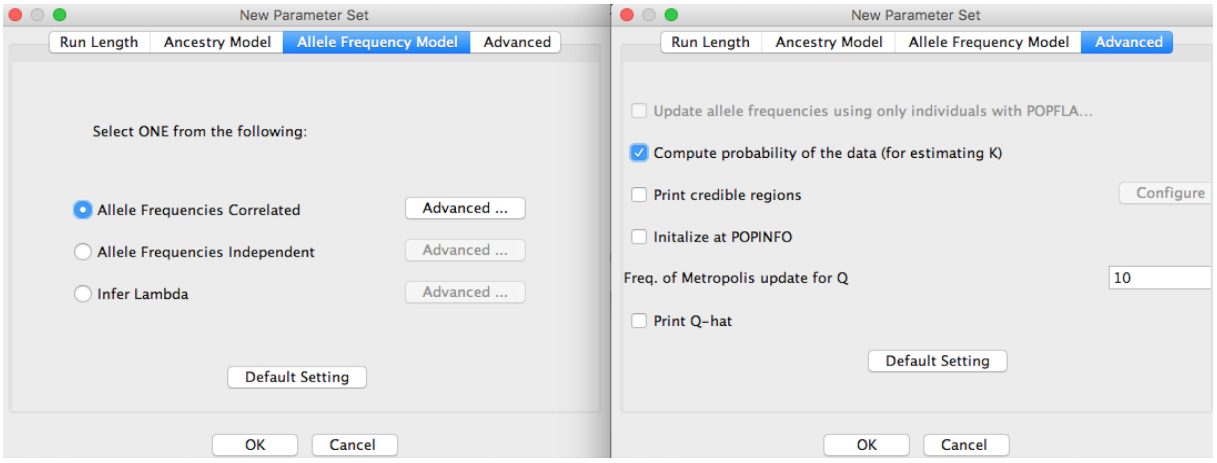


Figure 14: **Parameters chosen in STRUCTURE for the section "Population identification for admixed models"**

8.4 Population identification using real data

Here is the main part of the code corresponding to the section "Population identification using real data".

```
1
2
3
4  # This function requires to have already loaded the real data: the
       variable "chr21" contains this real data.
5
6
7  library(irlba)
8  library(stats)
9  library(fossil)
10 library(threejs)
11
12 SVD_KMeans <- function(Size_Sequence ,seed_simulation) {
13
14
15 set.seed(seed_simulation)
16
17 # We sample randomnly X SNPs from the 1105538 SNPs from the real data
18 index_sample=sample(x=1:dim(chr21)[2], size=Size_Sequence, replace = FALSE
       )
19 chr21_sample=chr21[,index_sample]
20
21 cm_sample = colMeans(chr21_sample)
22
23 # Apply 3-dimensional PCA
24 p_sample = irlba(chr21_sample, nv=3, nu=3, tol=0.1, center=cm_sample)
25
26
27 # Download the superpopulation data for each sample, order by ids
28 ped = read.table(url("ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/
       working/20130606_sample_info/20130606_g1k.ped"),sep="\t",header=TRUE,
       row.names=2)[ids,6,drop=FALSE]
29
30 # Download the subpopulation and superpopulation codes
31 # Map the sub-populations to super-populations
32 pop = read.table("ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/
       20131219.populations.tsv",sep="\t",header=TRUE)
33 pop = pop[1:26,]
34 super = pop[,3]
35 names(super) = pop[,2]
36 super = factor(super)
37 ped$Superpopulation = super[as.character(ped$Population)]
38
39 # Plot with colors corresponding to super populations
40 M = length(levels(super))
41 print(scatterplot3js(p_sample$u, col=rainbow(M)[ped$Superpopulation], size
       =0.3))
42
43 #Apply the K-Means on the 3-dimensional data
```

```
km <- kmeans(p_sample$u, centers=5, iter.max = 20, nstart = 10)

SuperPopulations=rep(0, length(ped$Superpopulation) )
SuperPopulations[which(ped$Superpopulation=="EUR")]=1
SuperPopulations[which(ped$Superpopulation=="EAS")]=2
SuperPopulations[which(ped$Superpopulation=="AMR")]=3
SuperPopulations[which(ped$Superpopulation=="AFR")]=4
SuperPopulations[which(ped$Superpopulation=="SAS")]=5

#We compute the Rand Index to assess the ability of the statistical method
    to find individuals' true population of origin

print(rand.index(SuperPopulations,km$cluster))

return(rand.index(SuperPopulations,km$cluster))

}

#We have randomly selected 100 different samples of 100 SNPs from the
   1105538 SNPs (from the original real data). On this smaller dataset, we
    have used a 3-dimensional PCA, and we have applied the K-Means (where
   K=5) on the 3-dimensional data to perform population identification.

RI_100=rep(0,100)
for (i in 1:100){
RI_100[i]=SVD_KMeans(Size_Sequence=100,seed_simulation=29*i)
}

#We have randomly selected 100 different samples of 500 SNPs from the
   1105538 SNPs (from the original real data). On this smaller dataset, we
    have used a 3-dimensional PCA, and we have applied the K-Means (where
   K=5) on the 3-dimensional data to perform population identification.


RI_500=rep(0,100)
for (i in 1:100){
RI_500[i]=SVD_KMeans(Size_Sequence=500,seed_simulation=29*i)
}

#We have randomly selected 100 different samples of 2000 SNPs from the
   1105538 SNPs (from the original real data). On this smaller dataset, we
    have used a 3-dimensional PCA, and we have applied the K-Means (where
   K=5) on the 3-dimensional data to perform population identification.


RI_2000=rep(0,100)
for (i in 1:100){
RI_2000[i]=SVD_KMeans(Size_Sequence=2000,seed_simulation=29*i)
}

#We have randomly selected 100 different samples of 5000 SNPs from the
   1105538 SNPs (from the original real data). On this smaller dataset, we
    have used a 3-dimensional PCA, and we have applied the K-Means (where
   K=5) on the 3-dimensional data to perform population identification.
```

```
85
86
87 RI_5000=rep(0,100)
88 for (i in 1:100){
89 RI_5000[i]=SVD_KMeans(Size_Sequence=5000,seed_simulation=29*i)
90 }
91
92 #We have randomly selected 100 different samples of 100 000 SNPs from the
       1105538 SNPs (from the original real data). On this smaller dataset, we
        have used a 3-dimensional PCA, and we have applied the K-Means (where
       K=5) on the 3-dimensional data to perform population identification.
93
94
95 RI_100000=rep(0,100)
96 for (i in 1:100){
97 RI_100000[i]=SVD_KMeans(Size_Sequence=100000,seed_simulation=29*i)
98 }
```

References

[1] Pritchard J.K., Stephens M., Donnelly P.J. Inference of population structure using multilocus genotype data, (2000). Genetics 155: 945-959.

[2] McVean G, A Genealogical Interpretation of Principal Components Analysis. (2009) PLoS Genet 5(10): e1000686. doi:10.1371/journal.pgen.1000686

[3] Rosenberg N., Nordborg M. ,Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. (2002) Nat Rev Genet 3, 380–390 . https://doi.org/10.1038/nrg795

[4] Staab PR, Zhu S, Metzler D, Lunter G. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. (2015) Bioinformatics. 31(10):1680-1682. doi:10.1093/bioinformatics/btu861

[5] Staab PR, Metzler D. Coala: An R framework for coalescent simulation. (2016) Bioinformatics. 32(12):btw098. DOI: 10.1093/bioinformatics/btw098

[6] Novembre J., Stephens M., Interpreting principal component analyses of spatial population genetic variation. (2008) Nature genetics 40.5, 646.

[7] Sikora M., Pitulko V.V., Sousa V.C. et al. The population history of northeastern Siberia since the Pleistocene. (2019) Nature 570, 182–188. https://doi.org/10.1038/s41586-019-1279-z

[8] Mondal M., Casals F., Xu T., et al. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. (2016) Nature Genetics, vol. 48, no 9, p. 1066-1070.

[9] Pagani L et al. Genomic analyses inform on migration events during the peopling of Eurasia. (2016) Nature 538, 238–242. (doi:10.1038/nature19792)

[10] Pugach I., Matveev R., Spitsyn V., Makarov S., Novgorodov I., Osakovsky V.,Stoneking M., Pakendorf B., The complex admixture history and recent southern origins of Siberian populations.(2016) Mol. Biol. Evol. 33, 1777–1795 . https://doi.org/10.1093/molbev/msw055

[11] Wong E.H., Khrunin A., Nichols L., Pushkarev D., Khokhrin D., Verbenko D., Evgrafov O., Knowles J., Novembre J., Limborska S., Valouev A. Reconstructing genetic history of Siberian and Northeastern European populations. (2017) Genome Res. 27(1):1-14. doi: 10.1101/gr.202945.115. Epub 2016 Dec 13. PMID: 27965293; PMCID: PMC5204334.

[12] Kılınç G.M., Kashuba N., Koptekin D., Bergfeldt N., Dönertaş H.M., Rodríguez-Varela R., Shergin D., Ivanov G., Kichigin D., Pestereva K., Volkov D., Mandryka P., Kharinskii A., Tishkin A., Ineshin E., Kovychev E., Stepanov A., Dalén L., Günther T., Kırdök E., Jakobsson M., Somel M., Krzewińska M., Storå J., Götherström A. Human population

dynamics and Yersinia pestis in ancient northeast Asia. (2021) Sci Adv. 6;7(2):eabc4587. doi: 10.1126/sciadv.abc4587. PMID: 33523963; PMCID: PMC7787494.

[13] Lawson D.J., Falush D., Population Identification Using Genetic Data, (2012) Annual Review of Genomics and Human Genetics 13:1, 337-361

[14] Patterson N., Moorjani P., Luo Y., et al. Ancient admixture in human history. (2012) Genetics, vol. 192, no 3, p. 1065-1093.

[15] Balding D. J., and Nichols R. A., A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. (1995) Genetica 96:3–12.

[16] The 1000 Genomes Project Consortium. A global reference for human genetic variation. (2015) Nature 526, 68–74 . https://doi.org/10.1038/nature15393

[17] Wakeley J., Coalescent theory. (2009) Roberts and Company.

[18] Kingman J.F.C. (1982). The coalescent. Stochastic processes and their applications, 13(3), 235-248.