Inference of population relations using ancient genomes: Challenges, biases and imputation approaches in Dimensionality Reduction methodologies

by

Angeliki Papadopoulou

Master of Science in
Molecular Biology and Biomedicine

2022

Department of Biology
University of Crete

ABSTRACT

Ancient DNA (aDNA), derived mainly from archaeological findings, is a snapshot into the past, enabling researchers to obtain insights into the human evolutionary history. Due to its age, aDNA is highly degraded and damaged, leading to limited, often low-quality, extracted information. aDNA studies unravel the complex evolutionary history of human demography, such as the past population size changes, gene flow events as well as adaptive processes that contributed to the survival of our species. The demographic inference of ancient populations sheds light on the genetic relationships between them assisting us to understand the present-day structure and the common origins of human populations. In this study, we describe and evaluate widely-used methods for the inference of population structure, focusing on the Dimensionality Reduction techniques. We highlight the biases, introduced by the inevitable missing data in aDNA. We propose a novel imputation approach that is based on the phylogeny of the samples under study. Using simulations, we tested the accuracy of our imputation approach and we showed that is superior to the widely-used mean imputation and has a similar performance to the state-of-the-art kNN imputation. In conclusion, this thesis draws attention to the challenges accompanying the usage and analysis of aDNA data to infer population relationships and, in addition, proposes a novel imputation approach to retrieve the missing information of aDNA genotypes.

TABLE OF CONTENTS

CHAPTER 1 : **INTRODUCTION**

1.1 GENETIC VARIATION

If we compare the nuclear DNA between any two humans, it will be approximately 99.9% identical. The differences of 0.1% shape the total variation, which, combined with environmental factors, is reflected in phenotypic level. The human genetic diversity is substantially lower than that of many other species, including our nearest evolutionary relatives (Jorde, 2020). The knowledge of genetic variation has broadened the research in fields such as understanding evolutionary history or finding the genetic causes of diseases. The markers of genetic variation can vary, from protein polymorphism to restriction fragment length polymorphism (RFLP) and lastly to single nucleotide polymorphism (SNP). Current sequencing technology advances, the so-called 'next generation sequencing' (NGS) technology enabled the sequencing of whole genomes fast and at a low cost. Differences in frequencies of genetic variants within the population over time can provide insights into variant selection, while differences between populations can reflect either population structure or local adaptation. In any case, the origin and evolution of the variability should be the central key in human population genetics studies.

1.1.1 Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) refer to substitution of a single nucleotide at a specific position in the genome and they are the most prevailing class of genetic variation. Approximately 90% of variation in a population is due to SNPs (Bush and Moore, 2012). Their sites are spread across the genome, indicating that they can be found both in coding and non-coding regions. SNPs constitute the mainly used markers in population and association studies because they are easy to genotype on large scale and they can give answers to research questions

such as the demographic history of populations or the detection of causative variants of diseases or traits. The source of SNP data is either SNP arrays or Next Generation Sequencing. SNP arrays are designed to capture particular positions in the genome and they are based on preexisting knowledge of mapped genetic variants in a wide range of populations. Their cost is relatively low, enabling the capture of many samples. However, they might be subject to ascertainment bias (Clark et al., 2005). In particular, informative SNPs for the SNP array are discovered on a subset of the entire population and SNPs indicative of non-sampled groups will be missed, leading to biased interpretation of genetic variation in these groups. The most widely used human SNP arrays contain polymorphisms, ascertained in samples with ancestry of Europe, Asia or West Africa. On the other hand, NGS captures the entire genetic variation in the sampled individuals and it can be used for the detection of new, non prior characterized, SNPs. The cost, though, is raised and there is still a remarkable rate of sequencing error, keeping the debate of quantity versus quality open.

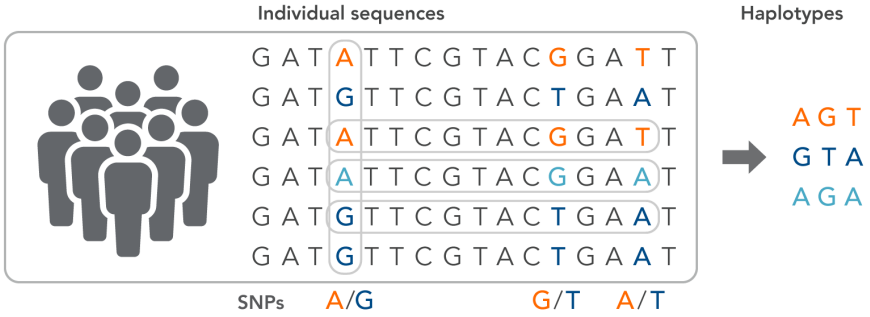

Figure 1.1: Single Nucleotide Polymorphisms (SNPs) among individuals. (Source: `https://www.idtdna.com/pages/education/decoded/article/genotyping-terms-to-know`)

1.1.2 Origin of genetic variation

The main sources of haplotypic variability in the human genome are the processes of mutation and recombination. The first generates new alleles and so is considered crucial in evolution, while the second combines pre-existing alleles at

different loci, generating novel haplotypes. A mutation is defined as a random change in the nucleotide sequence of a genome and it can be due to error during DNA replication or external, usually environmental, factors. Mutations can affect a single nucleotide, which take the form of an insertion, deletion or substitution or wider fragments of the genome, which are referred to as structural variants and include deletions, insertions, duplications and translocations of DNA segments. When mutations occur in germ cells, they are transmitted to offspring, contributing to changes in future generations. Consequently, these polymorphisms are highly significant from an evolutionary perspective. The mutation rate across the human genome is estimated at $1 - 1.2 \cdot 10^{-8}$ per base pair per generation (Consortium et al., 2010).

The second source of genetic variation, recombination, occurs in (at least) diploid organisms during meiosis when homologous chromosomes cross-over and exchange genetic material. This process leads to new combinations of variants at loci, increasing the genetic diversity. Recombination does not occur uniformly across the genome, while there are regions in which recombination is not allowed at all, such as the mitochondrion, the centromere and the Y chromosome. Other than ensuring the proper segregation of chromosomes, recombination also serves as a repairing mechanism for damaged DNA molecules. Moreover, it contributes to the adaptation of organisms to changing environments, by combining advantageous alleles at different loci.

1.1.3 Evolutionary processes

Genetic variation, which is generated as we mentioned above, passes down through generations and under the pressure of different evolutionary forces a complex history is created. The main processes that shape the genetic variation are genetic drift, gene flow and selection.

Genetic drift refers to the change in allele frequencies in a finite population

due to the random sampling of its alleles from generation to generation. Since the number of offspring of an individual is random, genetic drift is driven by stochasticity. The magnitude of frequency changes depends mainly on the population size. In a small populations allele frequencies are more easily disturbed from generation to generation since slight disturbances in sampling can cause considerable allele frequency changes. In the absence of new mutations, genetic drift leads to decrease of intra-population variation (diversity) because of the existence of two absorbing points for the allelic frequencies (at 0 and 1). On the other hand, it may lead to an increase of inter-population variation. According to the neutral theory most variation is shaped by drift, rather than natural selection, but this still remains controversial in the evolutionary community.

Natural selection refers to the evolutionary force which causes differences in the survival and reproductive rates of individuals in populations. These differences are due to mutations, whose frequency tends to be differentiated, because of their effect. There are mainly three types of selection, affecting the allele frequencies in a population. Positive selection acts on alleles with a selective advantage, increasing the fitness of the individual carrying it. As a consequence, its frequency rises and the allele spreads through the population until its fixation. Conversely, negative selection tends to decrease the frequency of alleles that impair the fitness of the individual until their removal of the population. The third type of selection is balancing selection, which favors multiple alleles at a locus and maintains their frequency at higher levels than expected from genetic drift. One of the mechanisms of balancing selection is heterozygous advantage, where heterozygous individuals are better adapted than homozygous. Signatures of natural selection can be detected and the advent of whole-genome sequencing facilitated this attempt, making available a dense map of markers in order to be analyzed in the context of genome-wide empirical distribution. Gene flow describes the migration of genetic variants between populations, following the migration movements of

individuals. By this process, new alleles can be introduced within a population, pointing out its evolutionary significance. At the intra-population level, gene flow increases the diversity due to the movement of alleles. However, genetic differentiation among populations is on decline, enabling diverged populations to increase their genetic similarity. In the current world, humans are spread around the world, shaping thousands of populations, while cases of totally isolated populations are very rare. Thus, the population set forms a network with interconnections influenced by social structure and culture. Population interconnections are governed by varying levels of gene flow.

1.1.4 Linkage Disequilibrium (LD)

The non-random association between loci is termed linkage disequilibrium (LD). LD arises from genetic drift, population admixture and selection, while it decays by recombination in each generation. It is, therefore, clear that close loci will be in higher LD and it will decrease with increasing physical distance. However, even really distant markers have been found to be under LD, either due to selection or non-adaptive stochastic processes (Reich et al., 2001). The pattern of LD is characteristic of the population, since the rate of the decay is related to the number of generations for which the population has existed. Consequently, LD can be used to study both recombination rate and demographic history. There are different ways to measure LD, but a popular one is with r^2. This statistic measures the squared correlation between the alleles of two positions. r^2 ranges between zero and one, indicating no linkage to 'complete' linkage, respectively.

1.2 POPULATION GENETICS

The field of population genetics is an expansion of evolutionary biology that studies the genetic diversity and structure within and among populations and the factors that influence the distribution of this diversity (Hartl et al., 1997). The field was born in the 1920s, when the debate about inheritance was still ongoing.

There were two schools of thought; the one viewed inheritance as the mixing of the parental traits into the offspring and it was focused on continuous traits, while the other in was influenced by Mendel's work and considered the the transmission of traits to be done via discrete characters, segregating with equal probability. The work of Fisher (1919) was clarifying, demonstrating how multiple genes of small quantitative effect could segregate according to Mendel's law of inheritance but still create seemingly continuous traits. The boost of population genetics took several decades and it was marked by the investigation of genetic variation at the molecular level. A classical work was the one of Lewontin and Hubby (1966), revealing much more genetic variation within populations than previously anticipated. This result, consequently, challenged the view that natural selection was the main driving force of evolution, because selection would lead to reduced variation. This study raised the question of selection versus neutrality and a notable paper in this discipline was that of Kimura (1968), which showed that the large amount of genetic variation within the population could only be explained by the abundance of neutral or nearly neutral mutations, giving rise later to the neutral theory of evolution (Kimura, 1983). Since then, population genetics is at the center of evolutionary biology, studying the processes that shape the allele frequencies over time with mathematical models.

1.2.1 The Wright-Fisher model

The Wright-Fisher model, developed by S. Wright and R. Fisher, describes the transmission of alleles from the parental pool to the offspring. It refers to a random-mated population with discrete non-overlapping equal size generations and absence of selection and migration. Real populations do not meet all these assumptions, but the provided model is a simplification which can be used to study how complex evolutionary forces affect a simple model. Given that the population size is N, the probability of a parental allele to be passed in the offspring

is $\frac{1}{2N}$. More generally, the probability of two particular alleles to share their first common ancestor at exactly k generations back in time is $(1-\frac{1}{N})^{k-1} \cdot \frac{1}{N}$. Since the population size is finite and offspring are chosen at random, some parental alleles do not contribute to the next generations, while others can contribute multiple times. Thanks to its simplicity, the Wright-Fisher model can be used to simulate a population over time and observe the fate of the different alleles in the population.

1.2.2 Coalescent Theory

The coalescent (Kingman, 1982) is a model describing how gene variants sampled from a population have originated from a common ancestor and it is based on the Wright-Fisher model. When the ancestral lineages of two gene copies meet in a common ancestor, it is said that they coalesce and such an event is called coalescence. Coalescent theory seeks to estimate the expectation of this time period and its variance. The probability of coalescence at each generation is geometrically distributed, while when time is considered in continuous scale the probability is well estimated by the exponential distribution. The coalescent, due to its property of following the lineages backwards in time, allows for simulations of population genetics data. Even though it was first proposed for the simplest population model, it has since been advanced to include almost any possible scenario.

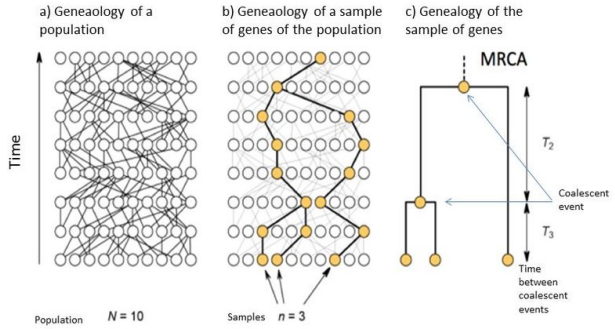

Figure 1.2: Coalescent principle. a) The complete genealogy of a population. b) The genealogy of 3 samples from the population backwards to their common ancestor. c) The genealogy of the samples, showing the coalescent events and the time to their Most Recent Common Ancestor (MRCA). (Credit: R. Leblois)

1.2.3 Population structure

The grouping of individuals in discrete subpopulations is called population structure or population stratification. Broadly, population structure refers to any deviation from random mating, leading to accumulated genetic and phenotypic differences between populations. These differences are noticed as differences in allele frequencies and the study of their source is crucial for understanding the genetic ancestry of populations. Some of the issues that population structure research addresses are: how to distinguish a structured from an homogenous population, what is the evidence for substructure in the data and how it can be quantified. Next we mention metrics and methods used to detect and describe population structure. F statistics, described by Sewall Wright (Wright et al., 1950) are well- established metrics in population genetics, partitioning genetic variability as measured by levels of heterozygosity into components of within and between population variation. The most cited statistic is F_{ST}, which describes the proportion of total heterozygosity (HT) that is explained by within population heterozygosity (HS). The formula for F_{ST} is given below (Equation 1.1). F_{ST} ranges between 0 and 1. When the subpopulations are genetically close, as in continuous admixture, high gene flow or recent split, F_{ST} should be close to 0. In the case of highly differentiated subpopulations, the measure should be closer to 1.

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} \tag{1.1}$$

The common methodologies for detection of population structure could be broadly categorized into two approaches; parametric and non parametric. Parametric methods utilize statistical models to infer population structure and assign individuals into subpopulations (Pritchard et al., 2000; Purcell and Sham, 2004). The assignment is based on the calculated likelihood that each individual belongs to a specific subpopulation. An issue with parametric approaches

is that they are based on several assumptions. Especially regarding genotype datasets, they assume Hardy-Weinberg equilibrium for populations, as well as linkage equilibrium among the genetic sites within each population. Such a parametric method is ADMIXTURE, which infers ancestral proportions for each individual and consequently multiple individuals are grouped based on their similar patterns of ancestry (Alexander et al., 2009). On the other hand, non parametric approaches include the techniques for dimensionality reduction with the well-established Principal Component Analysis (PCA), which places the data on inferred axes of maximum variation. PCA has been widely used for the detection of population structure from genetic data. Liu and Zhao (2006) have proposed a two-stage non parametric approach for analyzing population structure; dimensionality reduction followed by clustering applications, in order to reveal substructure in the reduced dataset.

1.3 ANCIENT DNA & ARCHAEOGENOMICS

1.3.1 Characteristics of aDNA

Ancient DNA (aDNA) refers to the genetic material found in remains, dated back hundreds or even thousands of years. This material can come from hominins, other animals, plants or microbes and is obtained from archaeological or palaeontological findings. Hard tissues, such as bones and teeth, are very common remains of humans and animals, due to their resistance and preservation (Hagelberg et al., 1991). In 1984 the aDNA research was marked by the DNA extraction and sequencing from a specimen (dried muscle) of *Equus quagga*, an extinct species of zebra (Higuchi et al., 1984). Since then, aDNA research is growing steadily, extending our knowledge of genetic variation beyond the present-day populations across the world. Human demographic history, animal and plant domestication and characterization and evolution of pathogens and microbes are some of the topics that aDNA has attempted to address. aDNA due to its post-mortem age and its long

exposure to the environment has specific characteristics, which contribute to its low quality and reduce the retrievable genetic information (Dabney et al., 2013). DNA fragmentation is one of them, which leads to ultra-short fragments usually shorter than 100bp in almost all ancient remains. The mechanism of fragmentation in aDNA consists of hydrolytic depurination which results in an abasic site, followed by beta elimination causing the fragmentation of the DNA strand (Lindahl, 1993). Moreover, chemical damages can occur in the nucleobases, leading to miscoding lesions in aDNA. The most common is the deamination of cytosine to uracil, by an hydrolytic degradation reaction. Such damages are accumulated in the single-stranded overhangs of aDNA fragments. The process of deamination has a significant effect because it leads to sequencing artifacts observed as C to T and also G to A misincorporations. In order to limit these effects, there is a common approach of treating extracted aDNA with uracil-DNA-glycosylase (UDG) to remove uracils prior to library construction for sequencing. Other than the chemical damages, a notable issue in the studies of aDNA, which has led to erroneous inferences (Pääbo et al., 2004) and affects the authenticity of the samples is their contamination by environmental microorganisms or human modern DNA during the handling procedures. Several improved experimental protocols have been developed to keep introduced contamination as low as possible (Poinar and Cooper, 2000). Additionally, from an in-silico point of view, developed frameworks target to isolate exclusively endogenous aDNA sequences based on their post-mortem degradation patterns (Skoglund et al., 2014).

1.3.2 Applications of aDNA studies

The field of human population history has received heightened attention with the advent of NGS technologies and the corresponding flourishing of aDNA studies. The wealth of available data, due to the relative ease in its production, has given rise to answer fundamental evolutionary questions. A key point was the finding

and sequencing of two archaic hominins, the Neanderthals (Green et al., 2010) and Denisovans (Reich et al., 2010), which revealed their genetic contribution to the ancestry of modern non-African populations and modern populations of Australasia and Oceania, respectively (Reich et al., 2011). In the last decade, the human ancient genomics field has grown rapidly and many sequences, some of them in high-coverage, have been available for genome-wide population studies. Thus, there is now well established knowledge about the demographic human history of various regions, such as America, East and Southeast Asia and Africa (Lipson et al., 2018; Raghavan et al., 2014; Wang et al., 2020; Yang et al., 2020). Other studies have given insights into natural selection, utilizing information about genomic variation from aDNA (Dehasque et al., 2020; Fehren-Schmitz and Georges, 2016). Great progress has been made in the field of animal domestication, as well. The genomic data of both ancient and modern animal samples contributed to the establishment of domestication models, by identifying signatures of introgression between wild and domesticated forms. Such models have been described for dogs (Bergström et al., 2020), wolves (Skoglund et al., 2015), cats (Ottoni et al., 2017) and pigs (Frantz et al., 2019). Paleomicrobiology and paleopathology are two related areas of research, which exploit ancient material and focus on microorganisms, in order to study their evolution or to provide insights into the life of ancient humans, by characterizing their dietary habits or the pathogens affecting them. By such studies, pathogens such as *Mycobacterium tuberculosis* and *Yersinia pestis*, responsible for tuberculosis and plague respectively, have been detected and contributed to the origin and expansion of infectious diseases (Bos et al., 2011, 2014).

1.3.3 Sequencing approaches

The first studies concerning aDNA used bacterial cloning for the amplification of small DNA sequences, retrieved from ancient specimens. Later, the develop-

ment of PCR (polymerase chain reaction) enabled the amplification even from low copies of initial material. However, with this approach, modern DNA from contamination was easily amplified, causing misleading interpretations. For this reason, many studies of that period have been disputed. The great growth of the field of aDNA, though, is associated with the advent of next-generation sequencing (NGS) technologies, which replaced the classical methodology of Sanger sequencing. Since 2005, when NGS was first introduced (Margulies et al., 2005), the methodologies have evolved, enabling fast and relatively low cost sequencing. The most common NGS method used in aDNA is the sequencing by synthesis, provided by Illumina Solexa Genome Analyzer. It is preferred because it can sufficiently handle the highly fragmented aDNA. The first step of Illumina sequencing is the library preparation, during which in the extracted DNA fragments, adapters are attached to both ends. The adapters contain an anchor sequence, necessary for the sequencing process. The resulting fragments are then amplified by PCR. The sequencing process consists of libraries transfer to a flowcell with complementary nucleotide sequences to the sequences of the adapters. Thus, the DNA fragments are immobilized on the surface of the flowcell. These fragments serve as template for the binding of fluorescently labeled nucleotides, which when binded in each cycle, are detected by a laser that induces the fluorescence and excites light, which is measured. For the next cycle, the excess nucleotides are washed off and the procedure is repeated. The DNA fragments in the genomic libraries can be single or double-stranded, leading to the corresponding nature of reads. In the studies of aDNA, different sequencing approaches can be preferred, depending on the project goal and the available material. Shotgun sequencing is a very common approach, in which the extracted material is sequenced without any *a priori* selection of the target. The short-length reads, raised from the sequencing, are identified by matching to sequence databases and they are assembled. This approach has also been used widely for metagenomic analysis of ancient samples, in order to identify

all possible microorganisms in a specimen. With shotgun sequencing we are able to reconstruct whole ancient genomes, if the depth of sequencing is sufficient. However, regarding aDNA whole-genome sequencing (WGS) is not always advisable. Mainly, in cases when the endogenous DNA is very low, WGS would not be suitable because the cost would be increased in order to sufficiently sequence the endogenous target. Thus, approaches for capture and enrichment have been developed and have gained the acceptance of the researchers. The aim of these approaches in aDNA is to enrich endogenous target DNA (Bos et al., 2015; Schuenemann et al., 2013) and specific markers of interest in a population genetics perspective (Haak et al., 2015). The most widely used are the in-solution hybridization capture methods, in which biotinylated oligonucleotide baits (probes) are used to capture specific regions of interest in the DNA libraries. After hybridization, streptavidin is used to separate the baits, resulting in target enriched libraries that can be sequenced in the same way as normal DNA libraries. The 1240K SNP capture panel has been an established approach since its development in 2015 (Haak et al., 2015) with an ever-increasing number of aDNA publications utilizing it.

1.4 HUMAN DEMOGRAPHIC HISTORY

1.4.1 Origin and expansion of modern humans

The study of genetic variation and how it is shaped through the years can unravel the demographic history of populations. Human demographic history has been a wide field of research, since it would provide insights on the history of present-day populations and their connections in the past. Two key models have been proposed to explain modern human evolution; the multiregional model and the model of replacement. Both models agree with the African origin of *Homo erectus* and its expansion to Eurasia about one million years ago. Although, they diverge on the origin and expansion of modern humans, i.e. *Homo sapiens sapiens*. According to the multiregional model of evolution, the regional lineages of

early human populations remained interconnected through gene flow and modern humans evolved in different regions of the Old World at different rates, depending on factors such as selection, gene flow and drift (Thorne and Wolpoff, 1992). In contrast, the replacement model, also known as the out of Africa hypothesis, supports the idea that modern humans have a relatively recent African origin and dispersed throughout the Old World, by completely replacing the existing archaic populations (Disotell, 1999). A combination of these two stands as a third model, known as the assimilation model. Assimilation model accepts the monocentric African origin, but argues, as well, for the contribution of Eurasian archaic populations to early modern humans. Both archaeological and genetic data tend to provide supportive evidence for the out of Africa model. Fossils record indicates that the first migration event from Africa occurred around 100,000 years ago (Schwarcz and Grün, 1992). It is believed that this migration into Eurasia occurred via the Levantine corridor. Additional evidence suggests a second expansion dated around 50,000 years ago (Henn et al., 2012; Stringer, 2012) with its route remaining controversial (Balter, 2011). The most prevalent scenario is the route over the Arabian peninsula at the southern end of the Red Sea (Fernandes et al., 2012; Melé et al., 2012). Two archaic genomes from the groups of Neanderthals and Denisovans questioned the monocentric African origin, providing evidence for limited gene flow of these two species to modern humans in two discrete events (Green et al., 2010; Reich et al., 2010). The first occurred at an early stage of out of Africa expansion, while the second concerned the ancestors of Oceanian populations.

1.4.2 Demographic history of Europe

Modern humans arrived in Europe by 45kya and it is estimated that they were mixed with Neanderthals, already dominant in the region. At 15kya Europe was dominated by an homogeneous group of hunter-gatherers, while in the eastern

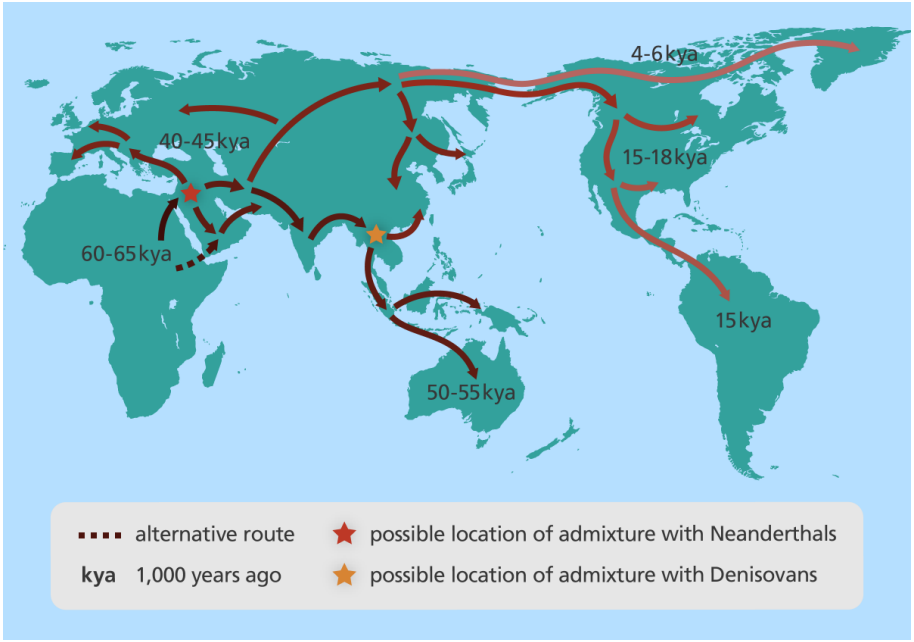

Figure 1.3: The out-of-Africa expansion of modern humans with reference to possible admixture events with Neadernthals and Denisovans.

part of the continent local hunter-gatherers were admixed with a discrete group of Siberian hunter-gatherers. Thus, European hunter-gatherers could be better described if splitted to two groups; Western hunter-gatherers (WHG), inhabited across western and southeastern Europe (González-Fortes et al., 2017; Mathieson et al., 2018) and Eastern hunter-gatherers (EHG) with the contribution of Upper Paleolithic Siberian ancestry which mainly contributed to the ancestry of northern Europeans (Günther et al., 2018). A major event which is of great interest not only in evolutionary biology but also in a historical and social perspective, is the Neolithic Revolution, i.e. the development of agriculture in Europe. This event led to the transition of hunter-gatherer to agriculturist lifestyle and based on genetic evidence is related to population migrations. In particular, migrants from the Near East, also referred to as Anatolian farmers, settled in southeastern regions of Europe at 8-9kya and expanded throughout most of mainland Europe, replacing the dominant WHG (Hofmanová et al., 2016; Mathieson et al., 2015). Other than the population movements, the Neolithic Revolution is associated with cul-

tural differentiation (Gronenborn, 1999), drastically altered diet (Richards et al., 2005), the spread of previously unseen infectious diseases and the emergence of the Indo-European language (Renfrew, 1989). Europe at 5kya (Eneolithic period) experienced a second migration wave, in which migrants from the Steppe reached eastern Europe, making a massive impact in central and northern regions until the start of the Bronze Age (Allentoft et al., 2015; Haak et al., 2015). These migrations had a crucial role in shaping the mosaic of ancestry in contemporary Europeans. According to Lazaridis et al. (2014), most present Europeans have derived from at least three ancestral populations; the WHG who contributed to all Europeans except of Near Easterners, the Early European Farmers (EEF) originated in Anatolia and the Ancient North Eurasians (ANE) who were related to Upper Paleolithic Siberians from Steppe.

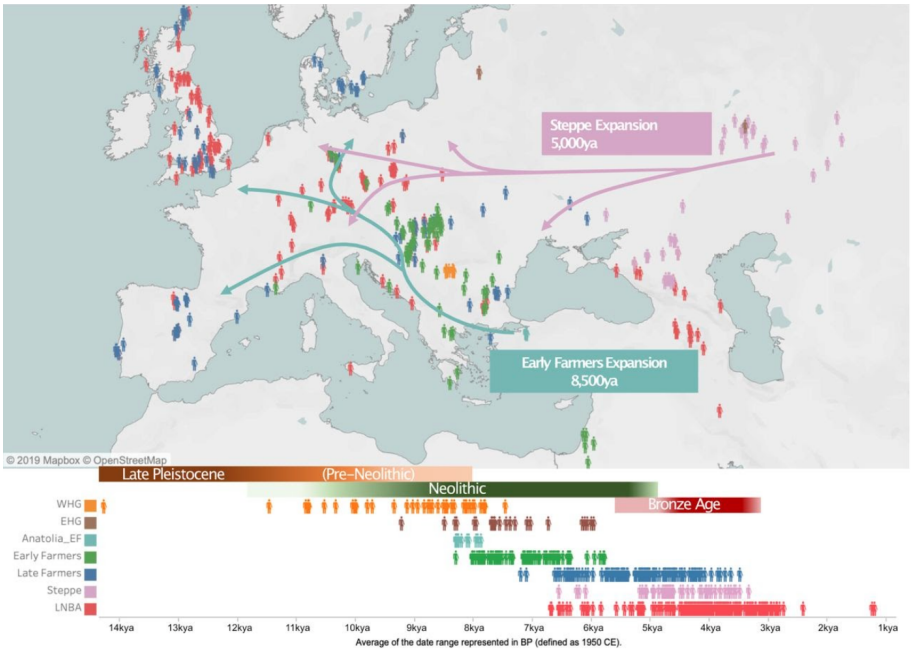

Figure 1.4: The known expansions in Europe during Neolithic and early Bronze Age.

1.5 MISSING DATA AND IMPUTATION

The issue of missing data is common in different fields of data science, since it can limit and potentially bias downstream analyses. The structure of missing

data is decisive for its effect and it is shaped by the data collection. Broadly, four mechanisms of missing data have been described; structurally missing (SMD), missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Rubin, 1976). In SMD a missing entry is not supposed to have a value. In such a case, the missing values are excluded from the analysis. MCAR in a dataset is independent from both observed and unobserved, yet it does not introduce bias but may affect only the statistical power of downstream analyses. MAR, in contrast, depends on observed and unobserved values, indicating a structure behind missing entries. Finally, MNAR is shaped by factors, not measurable by the researcher, usually in a systematic manner. This kind of missing data derives from the collection process. The growth of Next Generation Sequencing (NGS) technologies has yielded huge amounts of data. In genomic datasets, multiple missing entries might be introduced, belonging mainly to MNAR and reflecting systematic errors or artifacts from the sequencing and genotyping processes. Moreover, the capability of sequencing to high coverage, especially regarding aDNA, is often limited by sample quality or cost. A classical approach to handle missing data is the complete-case analysis, in which exclusively individuals and features with no missing data are included in the analysis. This approach, however, other than it may introduce bias, it leads to significant loss of valuable information. Thus, the need to retrieve missing data has arisen. Imputation is the process that predicts a missing value and it usually creates a predictive distribution based on the observed data, leading to inferred values that fill in the missing ones. A common imputation approach is the single imputation, which imputes missing values by a unique value. In this category is the mean imputation, which fills in the missing data of a variable with the mean of the observed values for the same variable. The issue with this approach is that it leads to several biases even in the pattern of missing data is MCAR (Jamshidian and Bentler, 1999). More advanced approaches include imputation methods, based

on machine learning (ML) techniques. These techniques have been found to perform better than the traditional statistical approaches (Rahman and Davis, 2012; Silva-Ramírez et al., 2011). A recent work by Petrazzini et al. (2021) evaluated the performance of five imputation approaches in genomic datasets, including two ML-based; a Random Forest and a Nearest Neighbors based framework. They showed that the ML-based methods had the best performance and they gave preference to the kNN approach, due to its computational simplicity. Especially regarding genotype data, the most widely used imputation methods rely on sequential probabilistic models, in which missing genotypes are inferred based on reference panels of haplotypes. Such methodology is used by the tools IMPUTE2 (Howie et al., 2009), PHASE (Stephens et al., 2001) and the recent one GLIMPSE (Rubinacci et al., 2020). Another established tool is Beagle (Browning and Browning, 2007), which is based on a model of local haplotype clusters, derived from the similarity of the reference haplotypes at nearby markers. These approaches, as mentioned, are based on a reference panel of haplotypes, which may not be an optional idea for imputation in aDNA data because they capture solely the present-day variation. An empirical evaluation of genotype imputation for aDNA data, including these methods, is provided by Ausmees et al. (2021). The issue of genotype imputation in aDNA has to be handled with caution, ensuring that it will not introduce any bias and will lead the research to robust results. In that context, the ML-based approaches as well as the development of new imputation methods, utilizing exclusively the information of the observed genotypes should be taken into serious consideration.

CHAPTER 2 : **MATERIALS & METHODS**

2.1 DATA AND PREPROCESSING

The Allen Ancient DNA Resource (AADR) is a repository for published ancient and present-day DNA data, released by David Reich lab (`https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data`). It consists of two datasets; the 1240K from shotgun or target capture sequencing at approximately 1 million sites and the 1240K + HO dataset with data from the 1240K merged with present-day individuals from the Human Origins array with 597,573 sites. The 1240K dataset includes approximately 1.24M SNP sites first described in 2015 (Mathieson et al., 2015) and used in numerous studies since. It contains sites polymorphic in diverse modern and ancient populations. The set includes over 1.15M autosomal sites, 49 K on the X chromosome and 33K on the Y. The Human Origins array has been designed specifically for studies of human population and evolutionary genetics, the array includes SNPs selected using a simple and clean ascertainment strategy that permits evolutionary hypotheses to be studied in a straightforward and quantitative way, thus enabling valuable inferences about human history. For the purposes of this thesis, we downloaded the 1240K + HO dataset on v.44.3 release, consisting of 13,197 ancient and present-day individuals and 597,573 SNPs. The data was in Eigenstrat format, organized in four files; the binary file of genotypes, two files of information about the SNPs and the individuals respectively and an annotation file with rich meta-information of each individual. Each dataset that we used for the analyses was a subset of the 1240K + HO dataset. Initially, the binary genotype file, followed by the files of SNP and individual information was transformed into binary plink format, using the convertf program of Eigenstrat software. The subsetting of individuals, as well as the following preprocessing,

were performed in PLINK 2.0 software. We pruned for linkage disequilibrium (LD pruning) in order to remove correlated sites, using a r^2 threshold of 0.5 and a sliding window of 50 variant counts shifted each time by 5. LD pruning is suggested for downstream population genetics analyses to reduce the computational burden since the SNPs that remain in the dataset are nearly uncorrelated. For PCA analysis, LD pruning is used to avoid capturing too much variance of linkage disequilibrium (LD) regions (Prive et al. 2018). Lastly, we excluded the sites of the chromosome Y and sites totally missing across all individuals. The ancient DNA genotype data was in pseudo-haplotype format, which means that a single allele appears in each site. This is a typical step in ancient DNA analysis to avoid misidentification of heterozygotes due to low sequencing coverage. In order to analyze together ancient and modern genetic data, we pseudo-haploidize the modern data as well, by randomly selecting one allele in cases of heterozygous sites.

2.2 POPULATION STRUCTURE AND ORIGIN ANALYSIS; ADMIXTURE ANALYSIS

During their evolutionary history, populations have experienced complex demographic nonadaptive (neutral) processes, consisting of population size changes and transfer of genetic material between populations. As a result, the genetic material of one individual in a given population might comprises fragments that originate in different (ancestral) populations. The detailed sequence of demographic events is usually too complex in real populations to be captured entirely. Admixture events are only one of the factors that determine population structure. In admixture analysis, partitions of ancestry are estimated from multi-locus genotype data and can be used to obtain insights into the origin of populations as well as for more accurate inference of population structure. This analysis was implemented by the ADMIXTURE software (Alexander et al., 2009). ADMIXTURE performs the estimation of ancestries in a random set of unrelated individuals using

a model-based approach and delivers directly admixture fractions. The underlying admixture coefficients and ancestral allele frequencies are estimated based on the maximum likelihood methodology.

The model does not take into consideration the linkage disequilibrium (LD), thus it is preferable to thin the dataset before running the software. The user also has to provide the value of K, which is the number of the ancestral populations from which the analyzed samples have derived. In case of unknown K, the software provides a cross validation procedure, in which multiple values of K are tested and for each of them the cross validation error is calculated. The K value that yields the minimum error should be the most appropriate for the analysis. The output estimates consist of the ancestry fractions and the allele frequencies of the inferred ancestral populations.

2.3 DIMENSIONALITY REDUCTION TECHNIQUES

Due to the rapid development of NGS technology, large biological datasets are increasingly analyzed to obtain insights into the evolutionary processes of species. The dimensionality of such datasets is represented by the number of SNPs, i.e. several millions for realistic datasets of most well-studied organisms. Dimensionality reduction techniques aim at reducing the dimensions of the data, thus making feasible to plot them in two or three dimensions, thus obtaining a visual representation of the data. Therefore, computational methods that reduce the dimensionality of such datasets lead to increased interpretability. Inevitably, however, some amount of data information will be lost as a consequence of the dimensionality reduction. Dimensionality reduction techniques focus on reducing the data dimensions in a way that minimizes loss of information. Such methods are widely used in modern genomic analyses, mainly because they allow the visualization of the data in the inferred lower dimensions.

2.3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the oldest and most widely used dimensionality reduction techniques, aiming to preserve as much variability as possible. This translates into creating new uncorrelated variables that are linear functions of those in the original dataset and successively incorporate an increasing amount of data variance. Such new variables, named Principal Components, represent the variance in descending order, i.e. the first Principal Component depicts the axis with the maximum variance of the data and so on. One of the first applications of PCA in genetics was conducted in 1978, in a study which confirmed the hypothesis of the demic spread of early farming from the Near East in Europe, by creating 'synthetic' maps using Principal Components in gene frequency data (Menozzi et al., 1978). In population genetics PCA has been widely used for studying population structure, visualizing genetic variation and probing demographic history (Fumagalli et al., 2013; Novembre and Stephens, 2008; Patterson et al., 2006).

The role of PCA is more descriptive and exploratory, rather than inferential. Regarding population genetics inference, there are different, often opposing, views. Many studies support the efficiency of the method in the recognition of ancestry and migrations events (Paschou et al., 2007; Reich et al., 2008), while others stand cautious about PCA, mentioning that it is affected by the sampling schemes as well as that it is difficult to interpret underlying demographic processes (McVean, 2009; Novembre and Stephens, 2008). Recently, Elhaik (2021)supports that PCA is inaccurate in both simple and complex scenarios and Yi and Latch (2022) raises the issue of missing data, proving that nonrandom missingness leads to biased PCA-based inference of population structure.

PCA of real data was implemented by the *smartpca* software of the EIGENSOFT package (Patterson et al., 2006; Price et al., 2006). A common approach, when handling both ancient and modern samples, is to construct the Principal

Components only from the high quality modern samples and then project the ancient, usually low-coveraged data, onto these PCs. The projection step is carried out by solving least squares equations, rather than an orthogonal projection. This approach is considered to give unbiased inference of the position of samples in the presence of missing data that are extremely common in ancient samples and it is applied with the `lsqproject:YES` option in the parameter file for executing the program. In data without missingness, such as the simulated data or real data after imputation, the PCA was implemented by the `prcomp` function in R.

2.3.2 Multidimensional Scaling (MDS)

Multidimensional scaling is an established multivariate analysis technique for obtaining quantitative estimates of similarity/dissimilarity of the data. The pairwise distances among the objects of the dataset are used for their configuration into an optimal low-dimensional space. When the data configuration is based on their geometric coordinates, e.g. the Euclidean distance, the type of MDS is metric and it is also known as Principal Coordinate Analysis (PCoA) (Cox and Cox, 2007; Gower, 1966). The non-metric MDS deals with non-numerical distances and it is preferred for ordinal data. For sequencing data, metric MDS is applied, using the pairwise distances of sequences to map the samples based on them. MDS, similarly to PCA, has been used in sequence data for the detection of population structure (Clemente et al., 2021; Maceda et al., 2021; Verdu et al., 2014). Metric MDS was implemented by the `cmdscale` function in R.

2.3.3 EMU

The population structure is used for many downstream analyses, such as for understanding population demography (Patterson et al., 2006) or for association studies (Marchini et al., 2004). Thus, an accurate inference of population structure is crucial. Large-scale sequencing studies are becoming more prevalent, since the advent of whole-genome sequencing technologies because they enable population

genetics analysis on a much broader scale. Such large-scale datasets, though, accumulate greater levels of missing information. The problem with PCA is that it cannot handle missing data in an appropriate manner, thereby leading to biased results as individuals are projected into the PC space based on their amount of missingness.

Recently, a method that deals with large-scale genetic data with high levels of missingness was proposed (Meisner et al., 2021). The algorithm, called EMU (EM-PCA for Ultra-low Coverage Sequencing Data), performs PCA with an accelerated expectation-maximization (EM) algorithm for modeling the missingness in an iterative manner.

2.4 F-STATISTICS

A common approach in population genetics for phylogeny and admixture inference is the toolkit, known as f-statistics, developed by the group of Reich (Patterson et al., 2012; Reich et al., 2009). The basic concept of the statistic is to measure correlations in allele frequencies among sets of two, three or four populations, corresponding to f2, f3 and f4 statistics respectively. Their interpretation can be related to population split orders and past gene flow events. Usually, f-statistics are used for formally testing hypotheses about admixture and constructing admixture graphs, i.e. phylogenetic trees augmented with admixture events. In our study, we utilized f4-statistic values to obtain insights into the population structure. Thus, the f4-statistic values were exploited in a PCA framework. This approach was originally developed in a study about the population history of prehistoric dogs, in which PCA was performed on all possible f4-statistics among ancient and modern dogs (Bergström et al., 2020). In this way, we can study population structure among ancient samples, without having to project them into axes of differentiation defined by the modern samples.

The f4-statistic is of the form `f4(A,B;C,D)` and measures the average cor-

relation in allele frequency differences between (i) populations A and B and (ii) populations C and D. The allele frequencies are typically averaged over many biallelic SNPs. When the f4-statistic is zero, or not statistically different from zero, it corresponds to the phylogenetic relation depicted in Figure 2.1 ,indicating that no admixture has occurred. Thus, the allele frequency differences between populations A and B should be completely independent from the differences between populations C and D. Otherwise, when the statistic is different from zero, gene flow events can be deduced. Specifically, negative value indicates gene flow between either A and D or B and C, while positive implies gene flow between either A and C or B and D. In another perspective, f4 can be considered as the branch length, derived from the intersection between the path from A to B with the path from C to D.

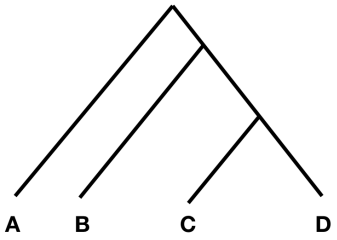

Figure 2.1: The phylogenetic relation among 4 populations (A,B,C,D), when the corresponding f4-statistic is zero

Respectively, f3 statistic is of the form `f3(A,B;C)` and measures allele frequency correlations among three populations. More precisely, the average over all genotyped sites of the product of allele frequency differences between the target population C to A and B, respectively is calculated. The f3-statistic can be used as a test of whether the population C is admixed between A and B, referred to as admixture f3, or as a measure of shared drift between populations A and B from an outgroup population C, referred to as outgroup f3. In the first case, negative value of f3 indicates admixture between the two source populations, A and B, while in the case of outgroup f3 the higher the value, the more genetic similarity

between A and B exists.

2.5 IMPUTATION METHODOLOGIES

2.5.1 Mean imputation

The imputation by mean is a common practice for handling missing data in many data science fields Lee (2011). It is achieved by replacing the missing observations of a variable with the mean of the observed values for the same variable. Mean imputation method, however, can lead to biased estimates, mainly if the number of missing data is large, because the variance following the imputation will be strongly underestimated and data with many missing sites will tend to exhibit small distances between them.

2.5.2 kNN imputation

The kNN imputation is based on the weighted k nearest neighbors (kNN) classification algorithm. Implementation of this algorithm for imputation was firstly proposed for microarrays data Troyanskaya et al. (2001), while in 2012 it was implemented for sequence genetic data (Schwender, 2012). Later, a kNN-based imputation algorithm was developed, which took into account the linkage disequilibrium (LD) between SNPs when choosing the nearest neighbors (Money et al., 2015). The imputation algorithm broadly consists of three steps: (i) the construction of a distance matrix of the data, using a distance metric, e.g. the Euclidean distance; (ii) the definition of the number k and finally (iii) the estimation of a missing observation, using a distance-weighted voting scheme by the k nearest neighbors.

In the example below (Figure 2.2), the genotype of sample X for a specific SNP is missing and has to be imputed. The number k of neighbors, i.e. other samples from the dataset, is arbitrarily defined as 4. The distances of the neighbors from the sample X are calculated and they are used as weighting factors of their genotypes, as shown in equation 2.1. Finally, the genotype state (0, 1, or 2) with

the maximum score will be selected for the imputation of X.

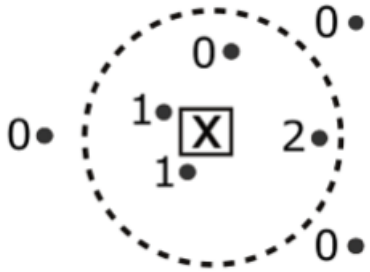

Figure 2.2: An example of kNN imputation. As X is depicted the missing genotype at a specific site. In the dotted circle is the neighborhood of genotypes, which contributes to the estimation of the genotype X in a distance-weighted manner.

$$g_i(s_i,p_j) = \arg\max \sum_{a\in(0,1,2)} \frac{I(g(s,p_j) = a)}{d_n(s_i,s)} \tag{2.1}$$

2.5.3 Phylogeny-based imputation

We propose a tree-based imputation method for imputing missing genotypes that utilizes the existing information of the data and its structure in a phylogenetic tree. First, the phylogenetic tree of the sequences is constructed using the maximum likelihood method. Then, in each site with a missing genotype, every possible genotype state is tested and in each trial the likelihood across the tree is calculated. The likelihood calculation is based on Falsenstein's pruning algorithm (Felsenstein, 1973). Thus, for every missing genotype in a given site, the whole tree as well as the site itself is taken into account. We use the generalized time-reversible (GTR) model of sequence evolution, even though other models can be readily employed.

In each node of the tree, the conditional likelihood for each genotype state is calculated (equation 2.2) and this process is iterated via a post-order traversal of the tree. The conditional likelihood of a genotype at an ancestral node is

the probability of obtaining the descendant states given the state of that genotype at the ancestral node. The likelihood value depends on the transition rates between the four nucleotides. Therefore, different evolutionary models describing such transition ratios have been developed. The simplest evolutionary model (the Jukes-Cantor model) Jukes et al. (1969) states that the rates for all possible transitions are equal. In other words, the Jukes-Cantor model assumes that the substitution of a base with any other base occurs with equal probability. In contrast, the most complex model (GTR) states that each transition might have a different value Tavaré et al. (1986). Therefore, a transition probability matrix is used for the calculation of the probability terms in 2.2. The final calculation includes the conditional likelihoods at the root, resulting in the likelihood across the whole tree using the equation 2.3.

The predicted state for the missing genotype is the state (A,C,G or T) that yields the maximum likelihood of the given tree. This procedure is repeated for every missing genotype and it is unaffected by the imputation in previous sites, since the already imputed genotypes are not used in the calculations as imputed genotypes but in their original missing state. The likelihood calculations were conducted using the pml function of the phangorn package in R.

$$L_p(i) = (\sum_{x \in k} P(x|i,t_L) L_L(x))(\sum_{x \in k} P(x|i,t_R) L_R(x)) \tag{2.2}$$

In equation 2.2, the conditional likelihood calculation in a node for state i. The two pieces of the equation refer to the left and right descendant, while x is the index of the k states. Each piece represents the probability of observing the state x in the descendant, given the state i in the node and the branch length t.

$$L = \sum_{x \in k} \pi_x L_{root}(x) \tag{2.3}$$

The likelihood calculation across the tree. $L_{root}(x)$ is the conditional likelihood

for the state x at the root of the tree, while π_x is the equilibrium probability of the state x.

In Fig 2.3, we demonstrate an example of eight samples (sequences), from which a phylogenetic tree has been already constructed using the whole sequence information. In a particular site (SNP), there is a missing genotype for one (or more) sample(s). Given the tree, every possible genotype state (0,1 and 2) will be tested and the likelihood for the tree will be re-calculated using each imputed genotype. Eventually, the genotype that maximizes this likelihood score will be reported as the imputed genotype.

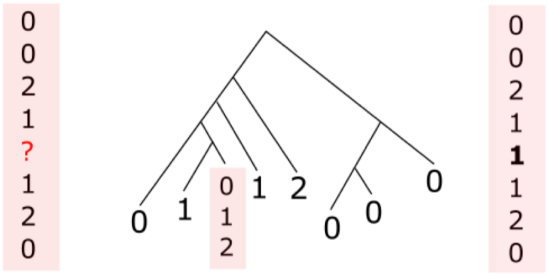

Figure 2.3: An example of phylogeny-based imputation. This is the phylogenetic tree of 8 sequences, with the genotypes at a specific site, in which the third one is missing. The red boxes on the left and the right represent the genotypes at the specific site before and after imputation, respectively.

2.6 GENERATING ARTIFICIAL DATA TO TEST IMPUTATION METHODS; THE *MS* SIMULATOR

The *ms* software (Hudson, 2002) is used to generate samples from evolving populations according to the Wright-Fisher neutral model. The model suggests finite population size, discrete generations and multinomial sampling to produce successive generations (Ewens, 2004). The simulator is based on the coalescent, a stochastic process to generate genealogies from a population by tracing randomly sampled alleles backwards in time (Kingman, 1982). The program assumes an infinite-site model of mutation, and thus no recurrent mutation can occur.

The following argument flags of the *ms* software were employed to generate the

simulated datasets:

MUTATION (-T THETA) The number of mutations across the simulated locus is determined by the parameter θ. The parameter θ is defined as the product of the mutation rate, μ, and 4 times the diploid population size N_0, $\theta = 4N_0\mu$.

RECOMBINATION In order to include recombination in the model, the flag `-r` was used. Similarly to the mutation parameter, the population recombination rate ρ is defined as $\rho = 4N_0r$, where r is the probability of recombination per generation across the simulated locus. Additionally to the parameter ρ, the number of possible recombination breakpoints should be defined.

MIGRATION Besides the simulation of populations and their structure, one could include migration in the model so as each population could receive migrants at the same rate from each of the other populations. The migration parameter is defined as $M = 4N_0m$, where m is the fraction of each population made up of new migrants in each generation. The flag `-eM` of the *ms* was employed.

POPULATION SPLITS In order to simulate more realistic scenarios, a range of past demographic events can be included in the simulator. One of these is to define a population split. Using the flag `-ej t i j`, all lineages in the population i are moved to the population j at time t (backward in time). The time is measured from the present in units of $4N_0$ generations.

The simulations used for the purpose of this study were generated with the following commands and more information about the parameters can be found in 2.6:

1. ms 100 1 -I 5 20 20 20 20 20 1 -ej 0.2 2 1 -ej 0.2 3 4 -ej 0.5 1 4 -ej 0.8 5 4 -t 500 -r 0 100

2. ms 60 1 -I 3 20 20 20 0 -ej 2 2 1 -ej 4 1 3 -t 500 -r 0 100

3. ms 60 1 -I 3 20 20 20 0 -ej 0.2 2 1 -ej 0.4 1 3 -t 500 -r 1500 100

METHODS APPENDIX

Table 2.1: Parameter values used in the simulations: Sample size, number of populations, mutation rate, migration rate and recombination rate

Parameter	**Simulation 1**	**Simulation 2**	**Simulation 3**
Sample size (individuals)	100	60	60
Number of populations	5	3	3
Mutation rate (per i,n,g*)	$1.25 \cdot 10^{-8}$	$1.25 \cdot 10^{-8}$	$1.25 \cdot 10^{-8}$
Migration rate (per i,n,g*)	$2.5 \cdot 10^{-5}$	-	-
Recombination rate (per i,n,g*)	-	-	$3.75 \cdot 10^{-8}$

* per i.n.g: per individual, nucleotide and generation

Table 2.2: Software used and their objective.

Software	**Objective**
convertf	convert Eigenstrat to Plink file format
plink2.0	Missingness filtering and LD pruning
smartpca	PCA and F_{ST} calculations
emu	EMU (EM-PCA)
admixtools	f3 and f4-statistic calculations
ms	generation of simulated genetic data

CHAPTER 3 : **RESULTS**

The results are organized as follows: (i) we present the results from the application of the dimensionality reduction techniques, demonstrating potential sources of misinterpretation of results applicable to aDNA sequence analysis. Then, (ii) we show the results of admixture analysis and (iii) the application of the f3-statistic and F_{ST} statistical approaches. In the dimensionality reduction section we have included two recently developed state-of-the-art methods in order to compare their results with more established approaches. Lastly, we demonstrate the results regarding the imputation of missing data. We present the outcome of imputation approaches, followed by the evaluation of the proposed phylogeny-based imputation and the application in real data.

3.1 DIMENSIONALITY REDUCTION TECHNIQUES

3.1.1 Principal Component Analysis (PCA)

Here, we used a dataset of 225 ancient human samples from regions of Southeastern Europe, described in Mathieson et al. (2015) and performed PCA using the Eigensoft software. We can notice some discrete population clusters, such as the clusters of hunter-gatherers from Latvia and the Iron Gates hunter-gatherers, which both have a relatively big sample size. However, the group of Neolithic in Ukraine, over-represented as well, is more disperse across the PC2 (Figure 3.1). Overall, the fact that the samples are originated from the same geographic region (Southeastern Europe) and are exclusively ancient from a range of time periods might not be sufficient to infer more reliable population relationships.

In the context of a comprehensive analysis of ancient genetic data, it is preferred to use a merged dataset of ancient and modern samples. In this way, modern variation can be taken into account. Furthermore, the relationships between an-

cient and modern populations can be assessed. For this purpose, we used a merged dataset of 15 populations, both ancient and modern. We performed PCA by projecting the ancient samples on the principal components calculated by the modern samples.

Surprisingly, we noticed that the ancient samples are mainly placed in a small region in the center of the PCA space (in the proximity of the origin $(0, 0)$). In contrast, modern genetic variation is well depicted in discrete population clusters (Figure 3.2). Even though it is plausible that ancient variation is considerably decreased in relation to the present-day genetic variation, a more plausible explanation suggests an artifact due to the mean imputation of the EIGENSOFT in aDNA data that are characterized by a significantly higher amount of missing data.

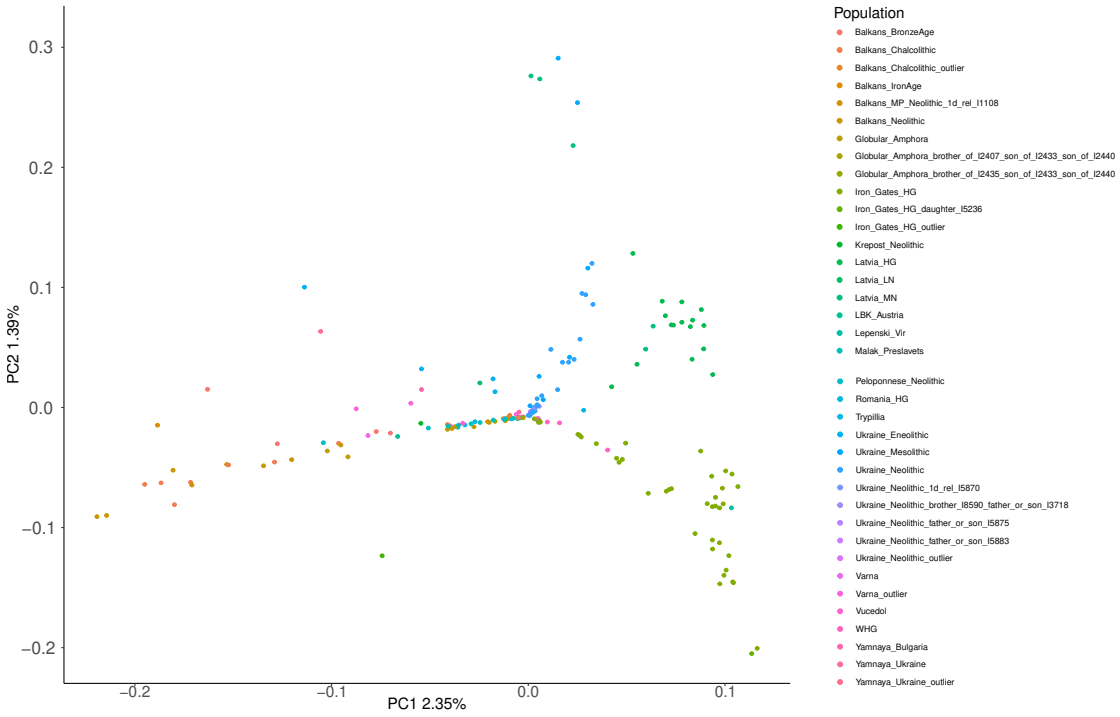

Figure 3.1: PCA of a genotype dataset of ancient individuals, originated in Southeastern Europe, implemented in smartpca software.

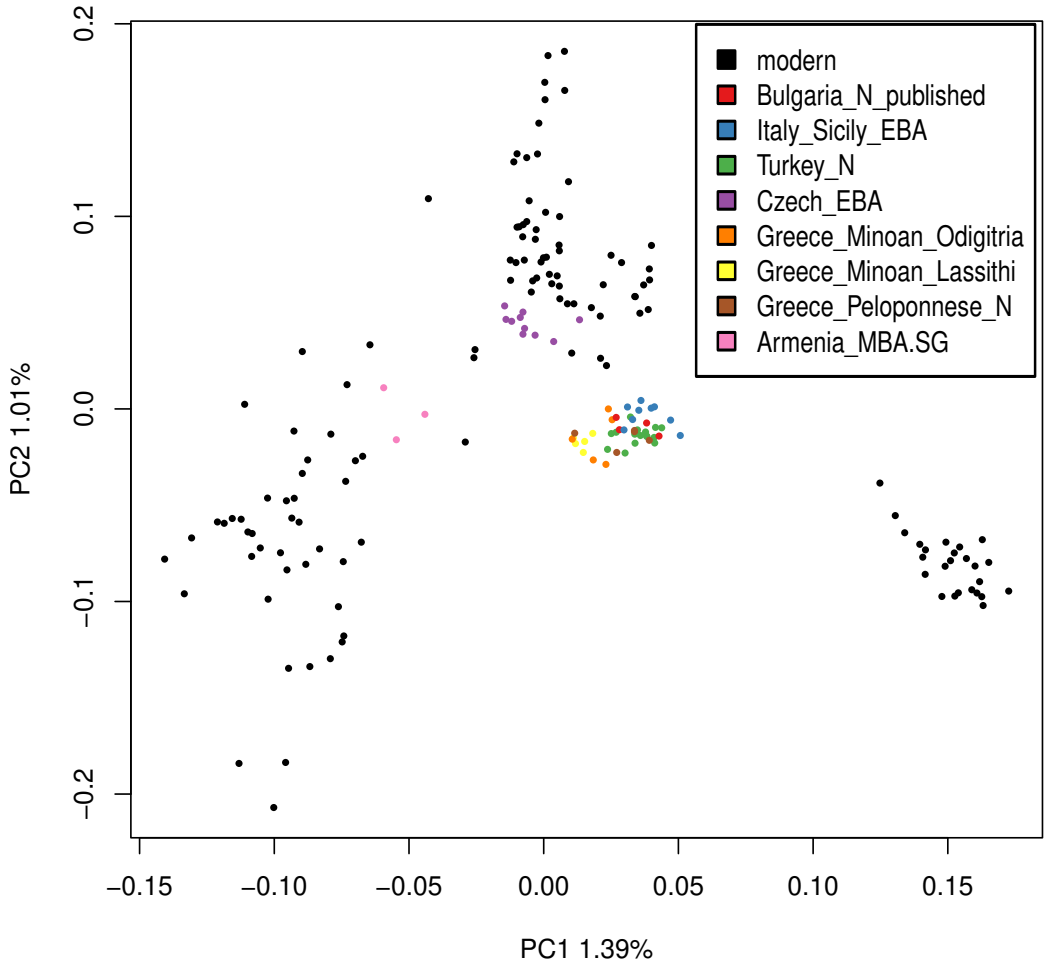

Figure 3.2: PCA of a dataset of both ancient and present-day individuals. The ancient samples are projected onto the Principal Components inferred by the present-day samples, implemented in smartpca software.

3.1.2 Multidimensional Scaling (MDS)

MDS is an alternative technique for dimensionality reduction, based on the pairwise similarities (or distances) among the samples. Since the distances between the samples are calculated in a pairwise manner, MDS might be a more appropriate method than PCA when the amount of missingness increases.

We performed MDS on both the ancient and the merged datasets. Interestingly, a completely different structure emerges compared to the PCA. Regarding the ancient dataset, we can observe the discrete clusters of the PCA, but also other smaller clusters are created (Figure 3.3), which in PCA were linearly depicted. The greatest difference, though, between MDS and PCA, was observed in the dataset of both ancient and modern samples, in which across the Coordinate 1 a clustering between modern and ancient samples is formed. However, other than this clustering there is discrete subgrouping, allowing for population structure inference.

To understand whether the amount of missingness has contributed to the observed structure patterns, we labeled the samples with a color gradient scheme based on their percentage of missingness. We noticed that the samples with high missingness -actually the ancient samples- are slightly spread, compared with their clustering in PCA. However, a cline of missingness along the first Coordinate is still apparent (Figure 3.4).

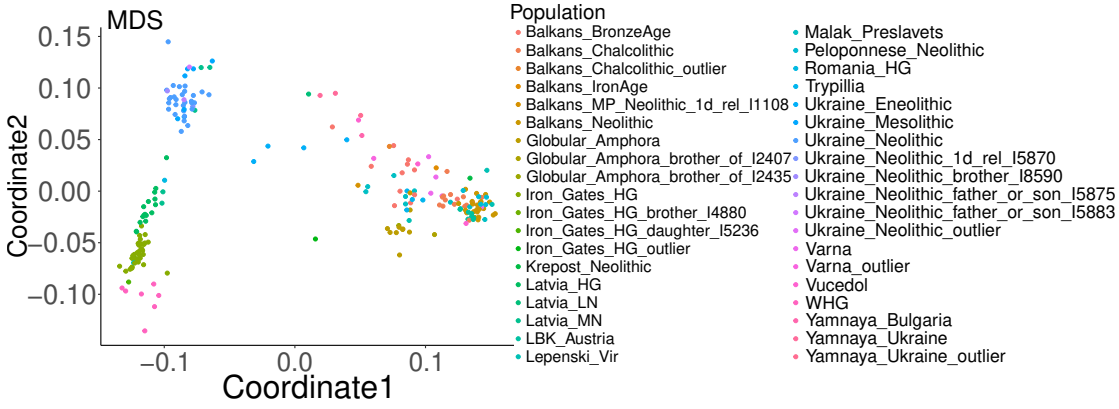

Figure 3.3: MDS in two dimensions for the dataset of ancient individuals.

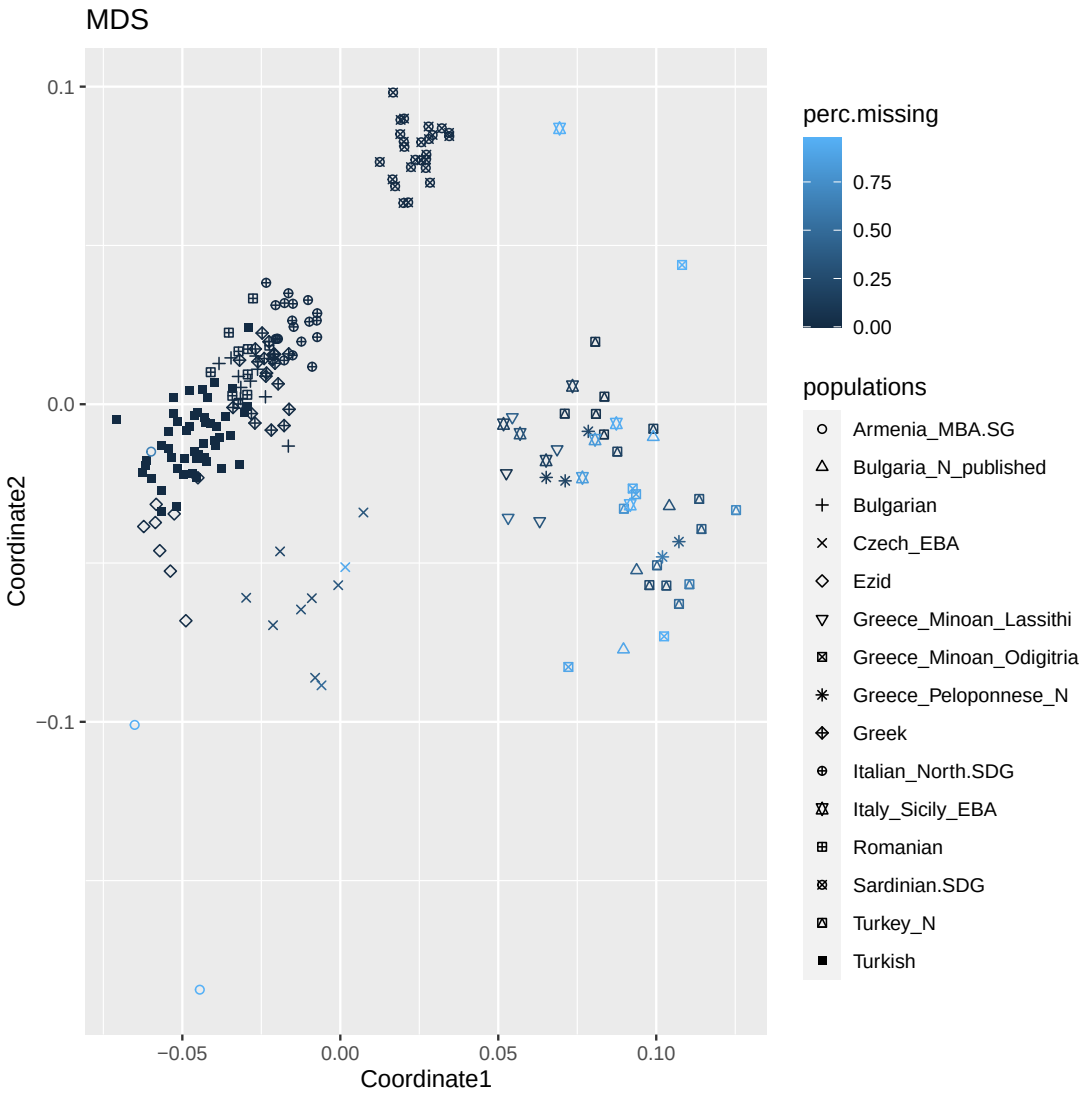

Figure 3.4: MDS in two dimensions for the dataset of both ancient and present-day individuals. The color gradient indicates the percentage of missing data.

3.1.3 State-of-the-art approaches; EMU and f4-PCA

Since no consistency between the results of PCA and MDS was observed, we further analyzed the data with two novel approaches. EMU is a PCA-based method for inference of population structure in the presence of high missingness. When running EMU for the ancient dataset we, surprisingly, observed a very similar structure with that of MDS. Actually, EMU achieved to better distinguish the clusters of Iron hunter-gatherers and hunter-gatherers from Latvia (Figure 3.5).

Other than that, the observed population structure was similar to that of MDS.

The other approach is also based on PCA, but instead of the genotype information the dimensionality reduction occurred in the space of the f4-statistics values. Since the f4 value is referred to a population, this approach allows for population depiction in the PCA space, instead of individuals as usual. We performed f4-PCA in the dataset of 15 ancient and modern populations. We, first, calculated the f4-statistics for all possible population combinations and then we organized these values in a matrix of dimensions $n * x$, where n is the number of populations and x is the number of all possible triplets of populations used in the calculation of f4. On this matrix, PCA was performed and as depicted in Figure 3.6 ancient and modern populations shape discrete clusters across the PC1, except of two ancient populations; an Armenian from the Middle Bronze Age and a Czech from the Early Bronze Age. We have not any evidence on why this occurs, but we suspect that the clustering is based on the amount of missing data and that these two ancient populations have less missing data, placing them closer to the modern populations. It might, also, be possible that we introduced a bias by substituting with zero the f4 values, for which the calculation was not supported. These were the cases with population duplication in the quadruple of populations. More cautious implementation is needed, in order to evaluate the accuracy of f4-PCA for the inference of population structure.

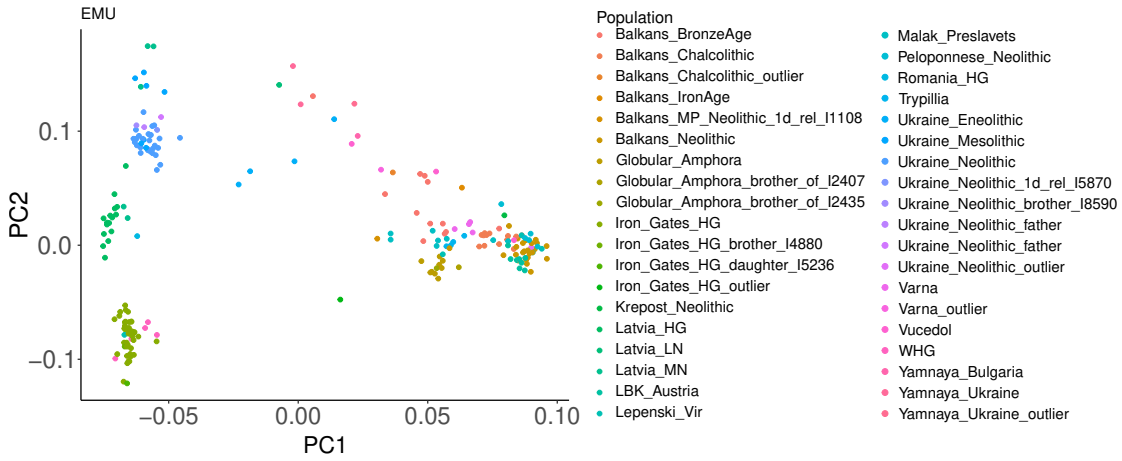

Figure 3.5: EMU in two dimensions for the dataset of ancient individuals.

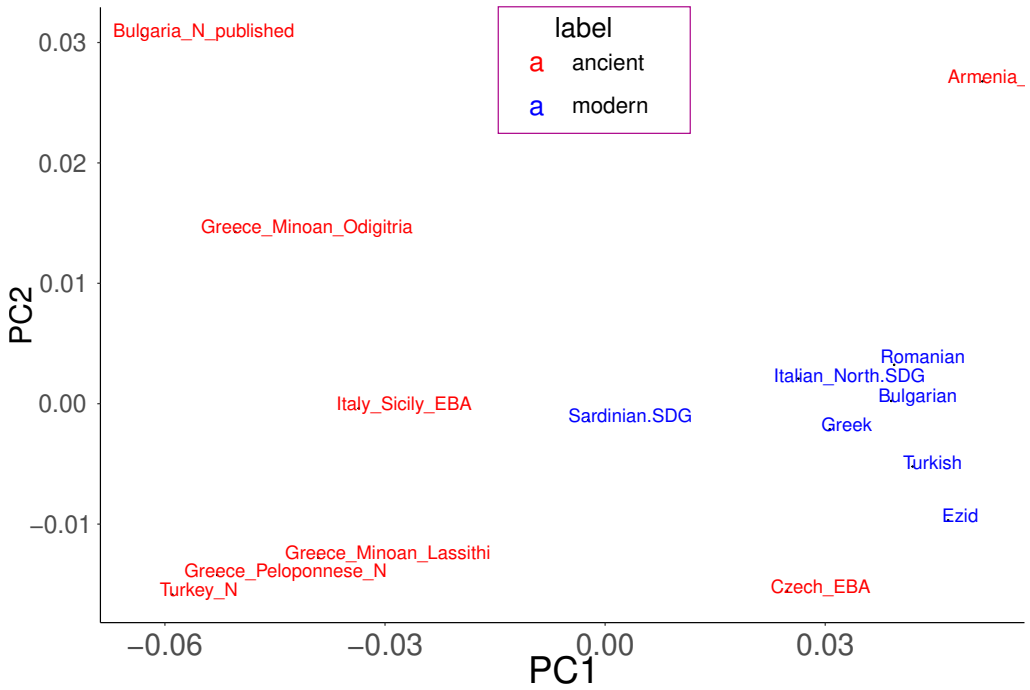

Figure 3.6: PCA based on the f4-statistic values (f4-PCA) in ancient and present-day populations

3.2 ADMIXTURE ANALYSIS

We performed admixture analysis, using the ADMIXTURE software, in the ancient dataset, assuming 4 ancestral populations. This assumption is based on the hypothesis of 3 ancestries in present-day Europeans; from early farmers, indigenous hunter-gatherers and north Eurasians. Splitting of hunter-gatherers in western and eastern forms the 4 ancestries. We can notice an ancestry-based cluster of WHG and HG from Latvia, while some the Iron Gates HG have a small proportion of this ancestry and the rest are modeled by a distinct one. Moreover, Ukrainian Neolithic and Balkans Neolithic are modeled by two discrete sources of ancestry (Figure 3.7).

However, there is a degree of uncertainty in the reliability of these results, if we take into account the accuracy of the model. The cross validation (cv) error measures how accurately the data fit into the model. When the number of ancestral populations is unknown, it is preferred to test a range of K and select the one with the minimum cv error. Surprisingly, when we evaluated the cv error for a range of K from 1 to 5, we noticed that the minimum cv error is for the K=1, which is not informative for population structure. We, then, tested the merged dataset of ancient and modern populations, resulting in the same suggestion of a single source population. Although this is not informative and probably does not reflect real ancestry, it points out a known issue of admixture modeling; the difficulty of automating the selection of K in a robust way.

Nevertheless, we estimated the proportions of ancestry assuming 2 ancestral populations. We observed a cluster of individuals having almost the entire ancestry of one population, while the ancestry of the rest was mainly from the other population with a low percentage of admixture (Figure 3.8). Surprisingly, the same pattern was observed in the percentage of missing data across the individuals (Figure 3.9). Consequently, the clustering did not represent the actual proportions of ancestries, but the amount of missing data in each individual, and

it could lead to misinterpretation. Even if the number of K was bigger, we suspect that some missingness-based structure would be hidden in the proportions of ancestry. Thus, the presence of missing data should not be ignored and the results of ADMIXTURE should be treated with caution and critical thinking.

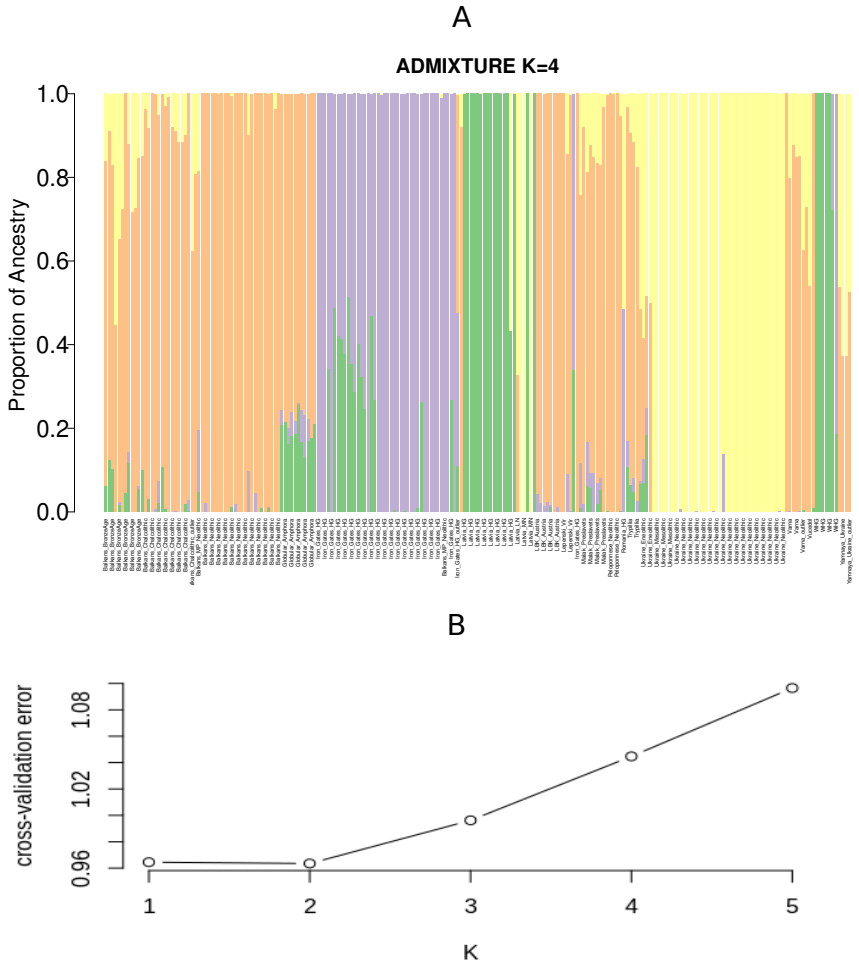

Figure 3.7: Admixture analysis of ancient individuals, supposing 4 ancestral populations (K=4) (A) and the cross-validation error for a range of K values (B).

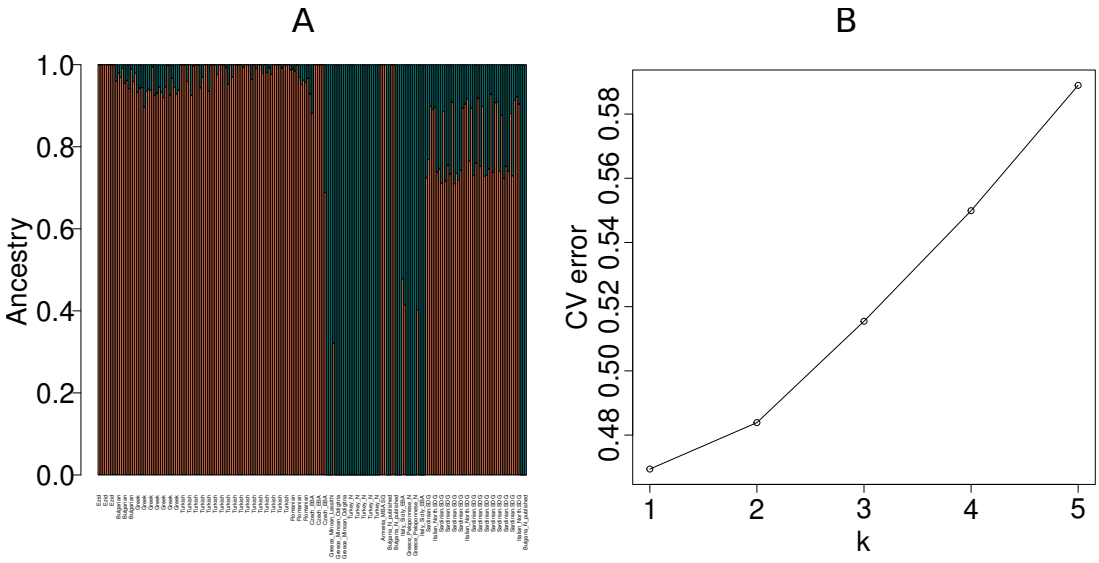

Figure 3.8: Admixture analysis of ancient and present-day individuals, supposing 2 ancestral populations (K=2) (A) and the cross-validation error for a range of K values (B).

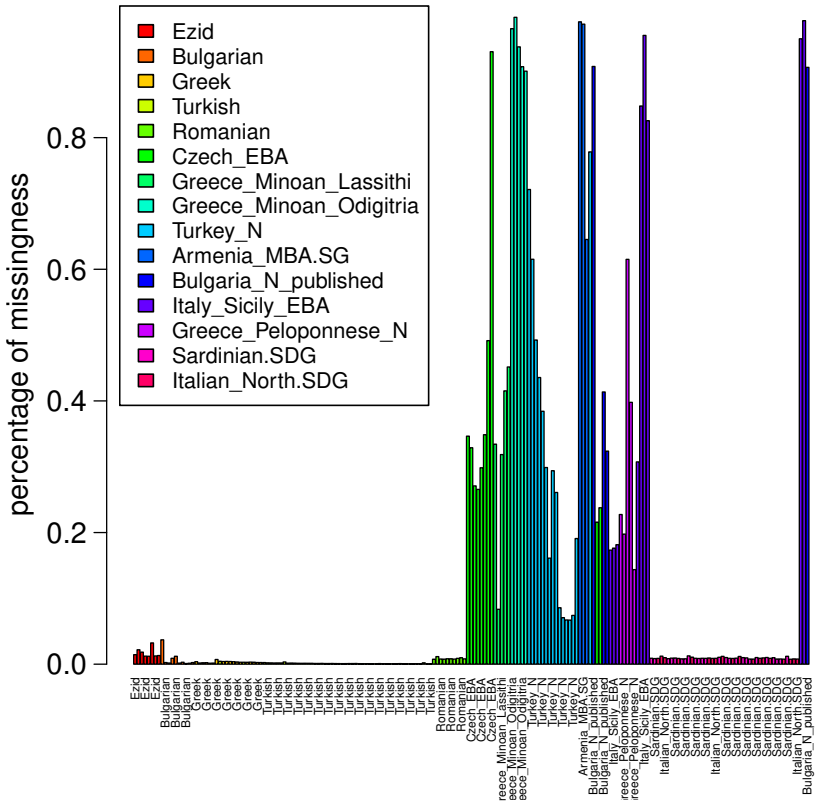

Figure 3.9: Barplot of the percentage of missing data across individuals.

3.3 F_{ST} AND F3-STATISTIC

With the advent of aDNA sequencing technology, population structure inference can be accomplished using genome-wide information from sampled individuals. Appropriate statistical measures are calculated from the genetic polymorphism information and they can reflect population relationships. F_{ST} measures population differentiation in a pair of populations and f3-statistic measures allele frequency correlations among 3 populations. We used ancient Greek individuals, who lived in mainland Greece and the island of Crete and dated 4000 to 3250 years before present (BP), merged with a set of present-day individuals from a wide range of European countries. The ancient genomes from Crete were sampled from three distinct regions of the island. The samples of mainland Greece obtained from the Peloponnese region, in which the Mycenaean civilization was developed. F_{ST} was calculated for each pair of ancient and modern populations, using the *smartpca* software of the Eigensoft package with the parameter `fstonly:YES` in the parameter file. F_{ST} ranges between 0 and 1 and the higher the value, the greater the differentiation between the tested populations. For the calculation of the f3(A,B;C) value, population A was ancient, population B was modern and C was represented in all cases by the Mbuti population, an indigenous group in the Congo region of Africa. This approach of f3, also referred to as 'outgroup f3', is used to measure shared drift between populations A and B, compared to an outgroup population C. Since the analyzed populations are non-Africans, Mbuti can be used as the outgroup. High values of f3 indicate close relatedness between populations A and B. The f3 calculations were implemented by the `qp3Pop` program of the AdmixTools software.

Surprisingly, the maps of F_{ST} show higher differentiation between the ancient Greek samples and the Mediterranean countries, which is at first dubious. The values ranged between 0 and 0.6 and in most of the cases the F_{ST} between the present-day Greek population and the ancient one was higher than 0.4. Only the

ancient sample of Crete, Armenoi yielded an intermediate value with the present-day Greek population. However, low to indermediate F_{ST} values between ancient Greek populations and Balkan countries might reflect the close population relationships and extensive gene flow over time between the area of Greece and the Balkan peninsula. Also, we noticed an northeast to west cline, which could potentially reflect known migration waves during the Bronze Age period (Figure 3.10).

The f3-statistic, as depicted in Figure 3.11, indicates higher genetic relations of the ancient Greek samples with present-day Bulgarian and Hungarian populations. Italy, France and Great Britain yielded a relative high value, as well. Croatia, Russia and Israel had the lowest values of f3 in all of the cases, which is not in accordance with the results of F_{ST}. We suspect that the different set of variants (SNPs) used for the calculations in each pair of population tested, due to the high missingness, may introduce some bias, obscuring the real population relations.

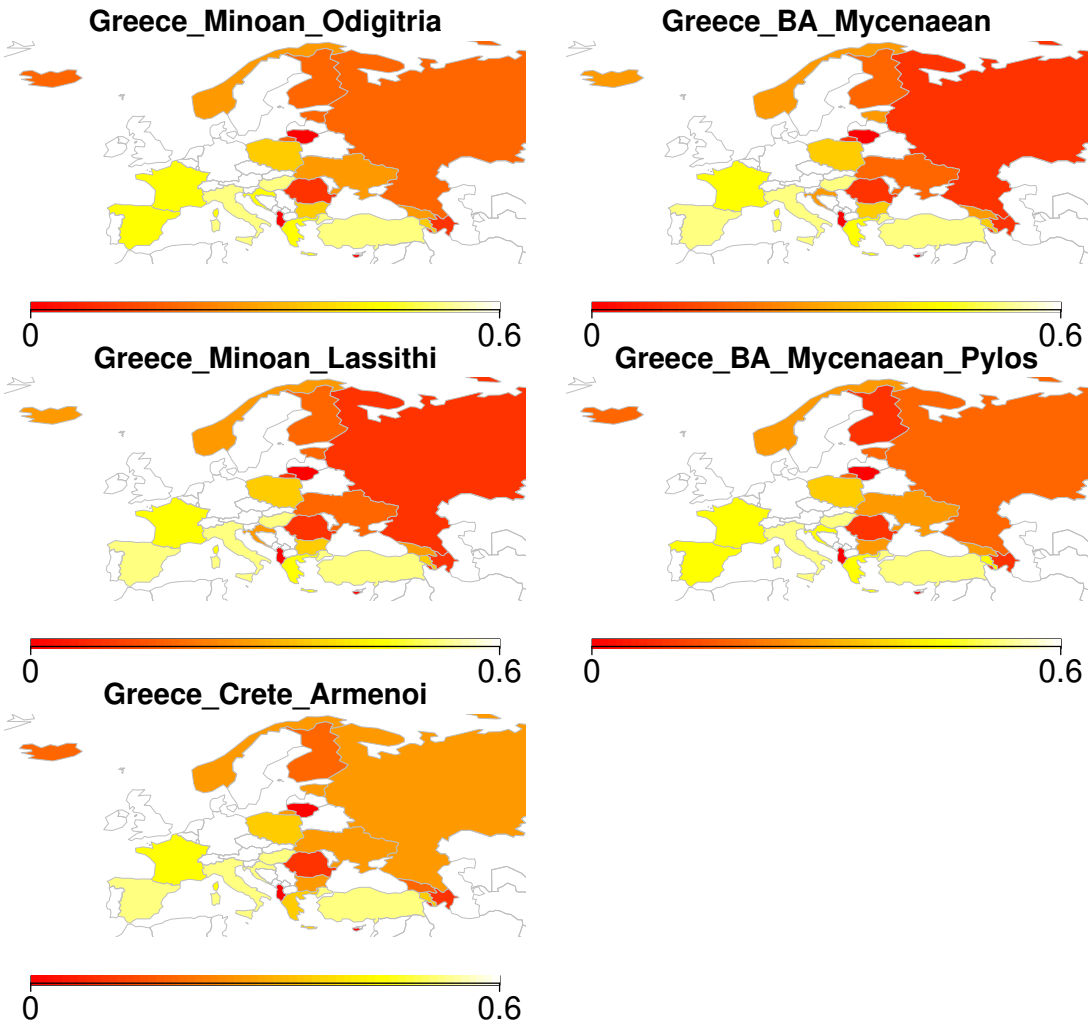

Figure 3.10: European maps of F_{ST} values between ancient Greek populations and present-day European populations.

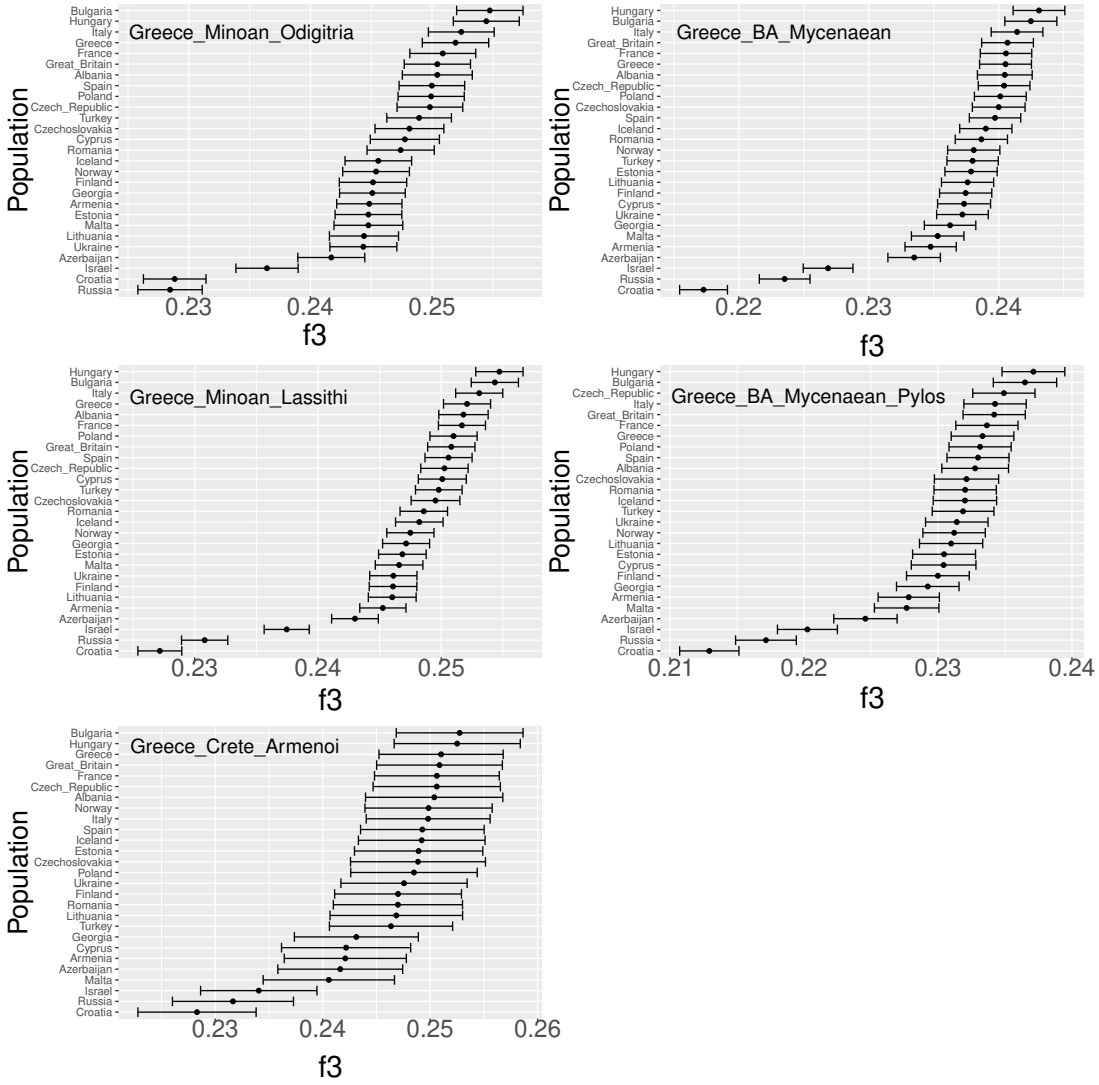

Figure 3.11: Genetic similarity between ancient Greek populations and present-day populations, measured using outgroup f3-statistic. Error bars show the standard error of the f3-statistic.

3.4 IMPUTATION APPROACHES

The aforementioned analyses are frequently used for the inference of population relations in aDNA studies. Nevertheless, as we showed, they might be influenced by the missing data, which is usually remarkably high in aDNA. Furthermore, datasets obtained from different research groups and possibly sequenced with different technologies, might be characterized by a non-uniform distribution of missing data that potentially introduces biases in the analysis (see, for example (Yi and Latch, 2022) and Figures 3.133.163.22). Consequently, the interpretation of the results could be misleading. Thus, imputation of missing data, i.e. the substitution of missing values with reliable ones, is a crucial step which is necessary and it has to be taken with caution in order to perform reliable downstream analyses.

3.4.1 Mean, kNN and phylogeny-based imputation

In this part, we are focused on imputation of missing data and we test three different approaches on simulated data and real data. The first is mean imputation, which is used by the Eigensoft software for the smartPCA function. Then, we test the kNN algorithm for imputation and lastly, an approach based on the phylogeny, developed during the current thesis. The data was produced by the ms coalescent simulator under different evolutionary scenarios (see METHODS APPENDIX 2.6). The workflow to test the efficiency of imputation approaches was as follows: i) from the simulated data we removed observations, following specific patterns of missingness ii) we performed imputation of missing observations by mean, kNN and phylogenetic tree iii) we evaluated the effect of imputation on downstream analyses, both PCA and MDS. Thus, we could compare the effect of different patterns of missingness, the performance of each imputation approach as well as the two commonly used methods for dimensionality reduction, PCA and MDS. In simulation 1, we created 5 equal-sized populations, allowing migration among them at a rate of $2.5 \cdot 10^{-5}$, while recombination was not taken into ac-

count. When removing observations, we followed a pattern of extreme missingness of 90% in the individuals of the last population, while all the rest had a percentage of missing data around 20%. The missingness across variants was random (Figure 3.12). Simulation 2, the simplest one, has three well-separated populations of equal size and neither migration nor recombination was allowed. We created two different patterns of missingness for this simulation; first, we tested a gradient increasing missingness in the individuals of each population and then we had uniform distribution across all individuals in a percentage of around 25%. Regarding the missingness across sites, we tested three different patterns; an accumulation at the start or the end of the SNPs set and a uniform distribution, similar to the previous simulation. In the first case, these three patterns were following the three populations respectively, but in the second case they were organized within each population (Figure 3.15, Figure 3.18). Lastly, in simulation 3 we introduced recombination in a model of 3 populations, while migration was not allowed. Here, two individuals of each population had an extreme missingness of 90%, while the missing data of the rest was around 20%. The missingness across SNPs was introduced randomly (Figure 3.21). We impute the missing data based on each imputation approach and perform PCA and MDS for the initial full dataset as well as for the imputed datasets. Especially regarding MDS, we can also visualize the dataset with the missing data, which is very informative about the disruption of the population structure. In the cases of extreme missingness in some individuals or even in a whole population, we observe that when the dataset is imputed by the mean, both in PCA and MDS those samples are dragged away from their real population cluster and tend to be grouped together near to the origin (the point (0,0)). Interestingly, in the case of gradient increasing missingness we can see the respective gradient pulling of the samples. Nevertheless, the imputation by kNN and tree seems to be accurate, since it conserves the real population structure even in the presence of extreme missingness. Low-rate and uniformly distributed

missing data among individuals does not introduce bias in downstream analyses, which is also the case for non random missingness among the variants. In the presence of recombination, all imputation approaches are accurate in the conservation of within population variation, but the mean imputation is limited to low percentage missingness, while the kNN and the tree-based are precise even in the samples with a high amount of missing data. In all of the cases, MDS had similar performance to PCA, but MDS still remains an advantageous method because it can be applied directly to the dataset with missing data. In that case, we observed a partial disruption of the real structure, driven by the amount of missing data.

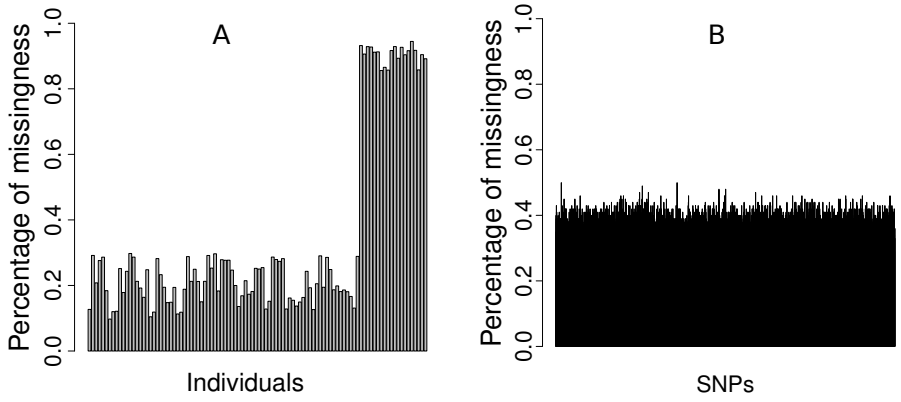

Figure 3.12: Patterns of missing data across individuals (A) and SNPs (B) in simulation 1.

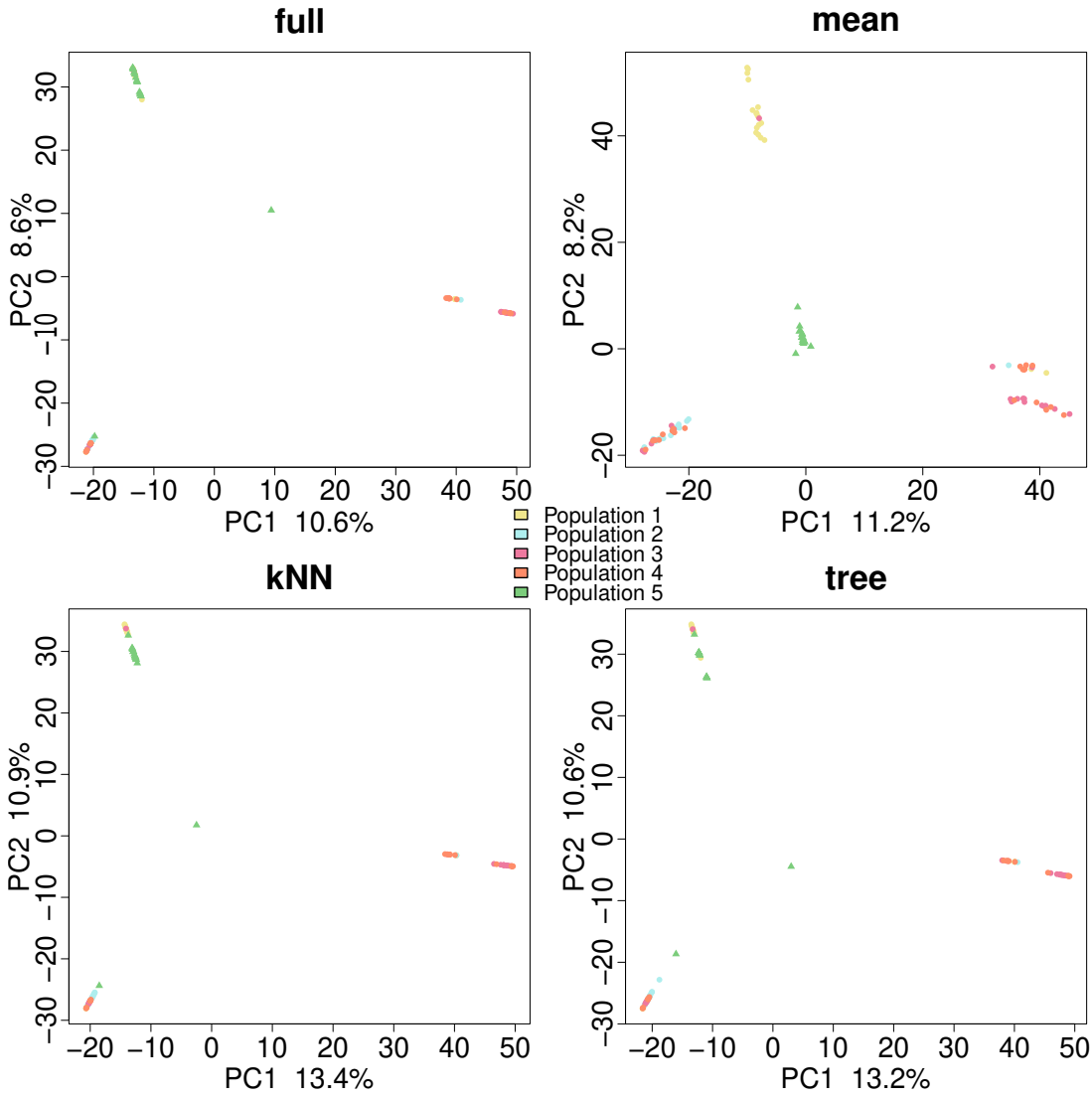

Figure 3.13: PCA of the initial and the imputed datasets from simulation 1.

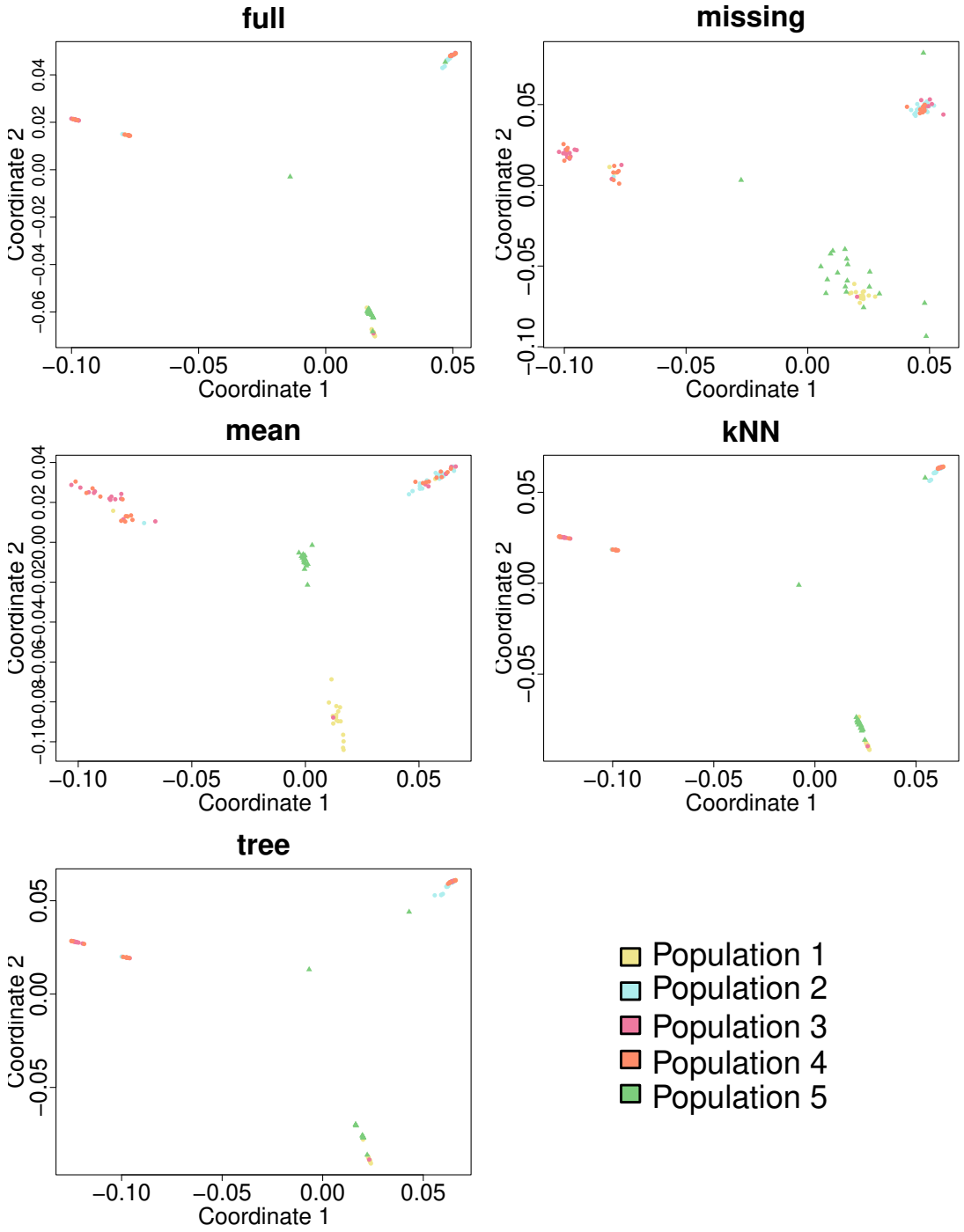

Figure 3.14: MDS before and after imputation for the simulation 1.

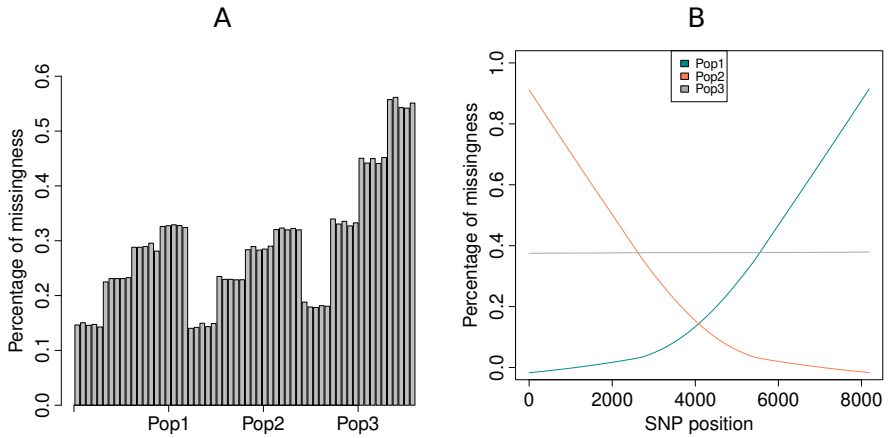

Figure 3.15: Pattern of missing data across individuals (A) and SNPs (B) in simulation 2 (scenario 1).

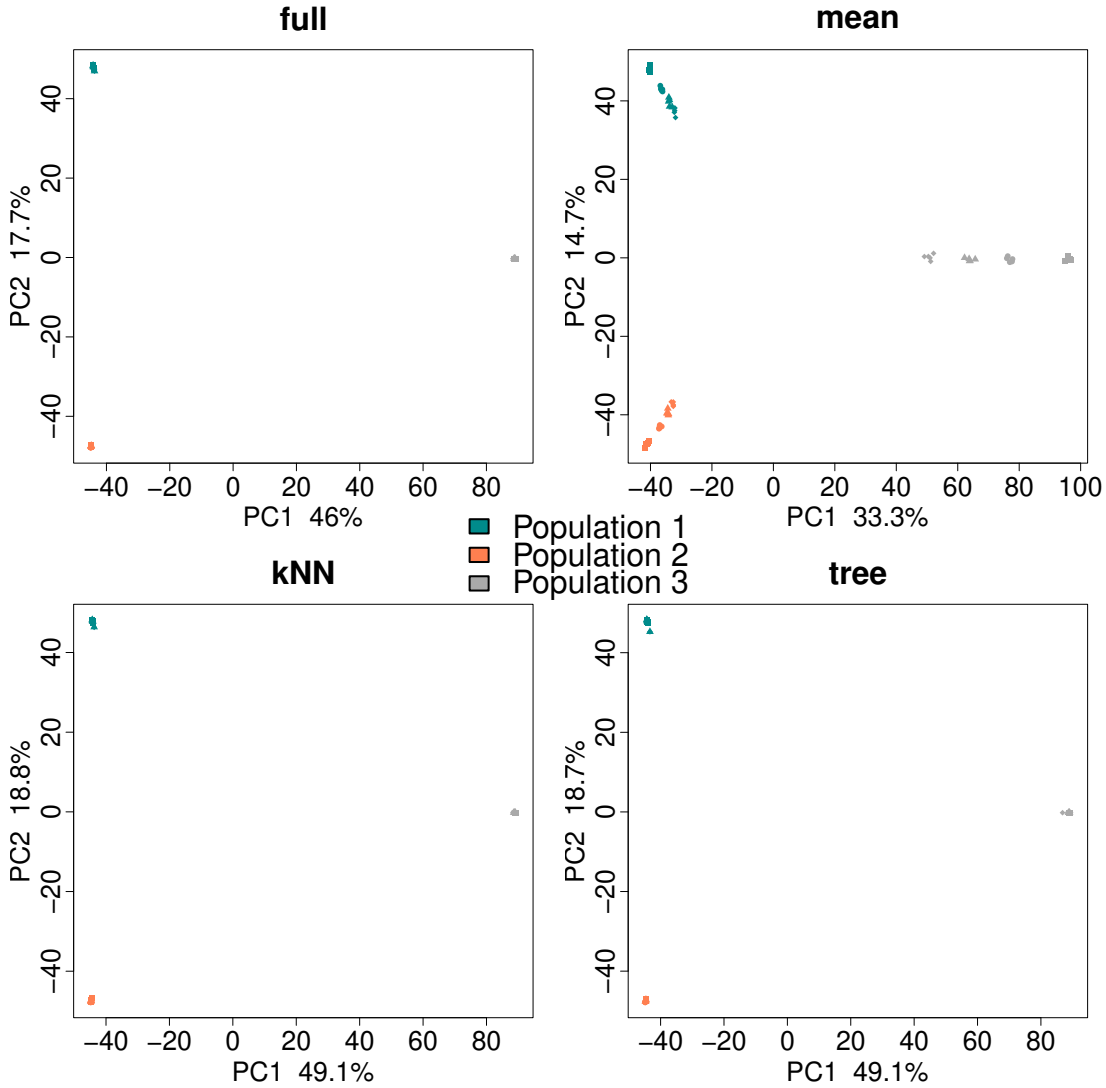

Figure 3.16: PCA of the initial and the imputed datasets of simulation 2 (scenario 1).

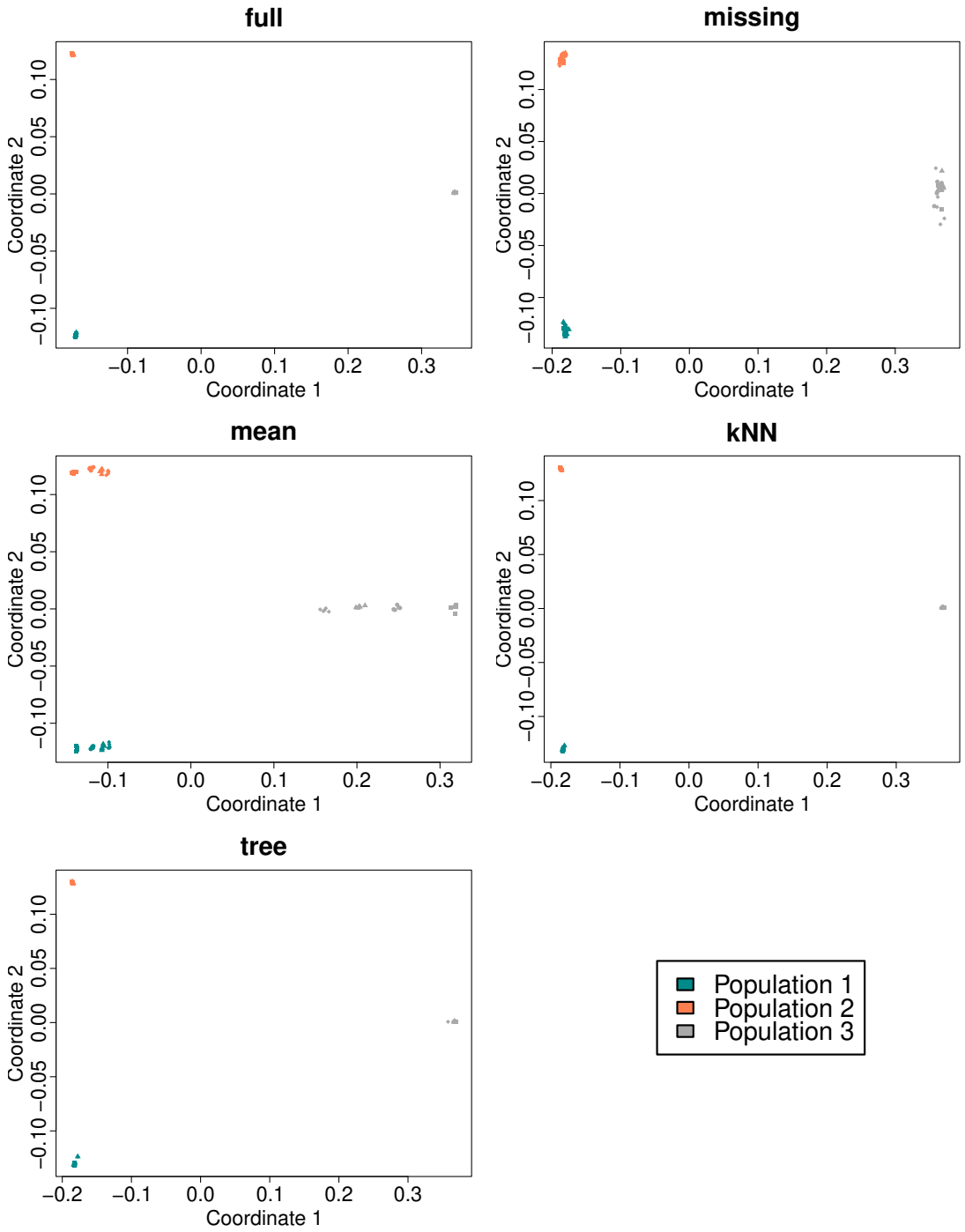

Figure 3.17: MDS before and after imputation for the simulation 2 (scenario 1).

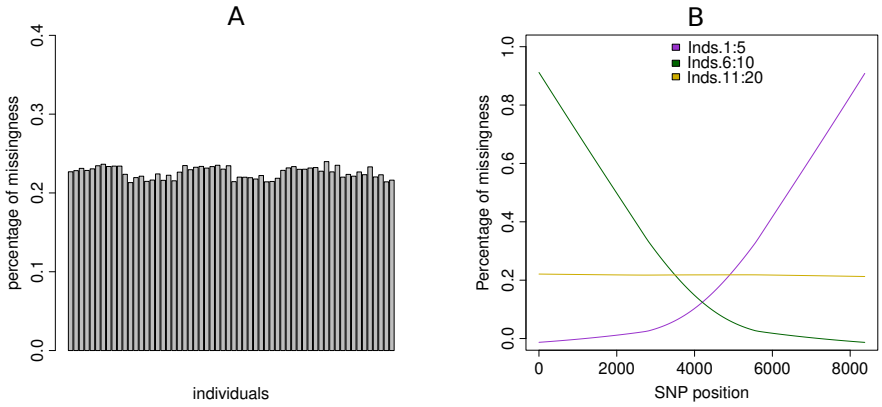

Figure 3.18: Pattern of missing data across individuals (A) and SNPs (B) in simulation 2 (scenario 2).

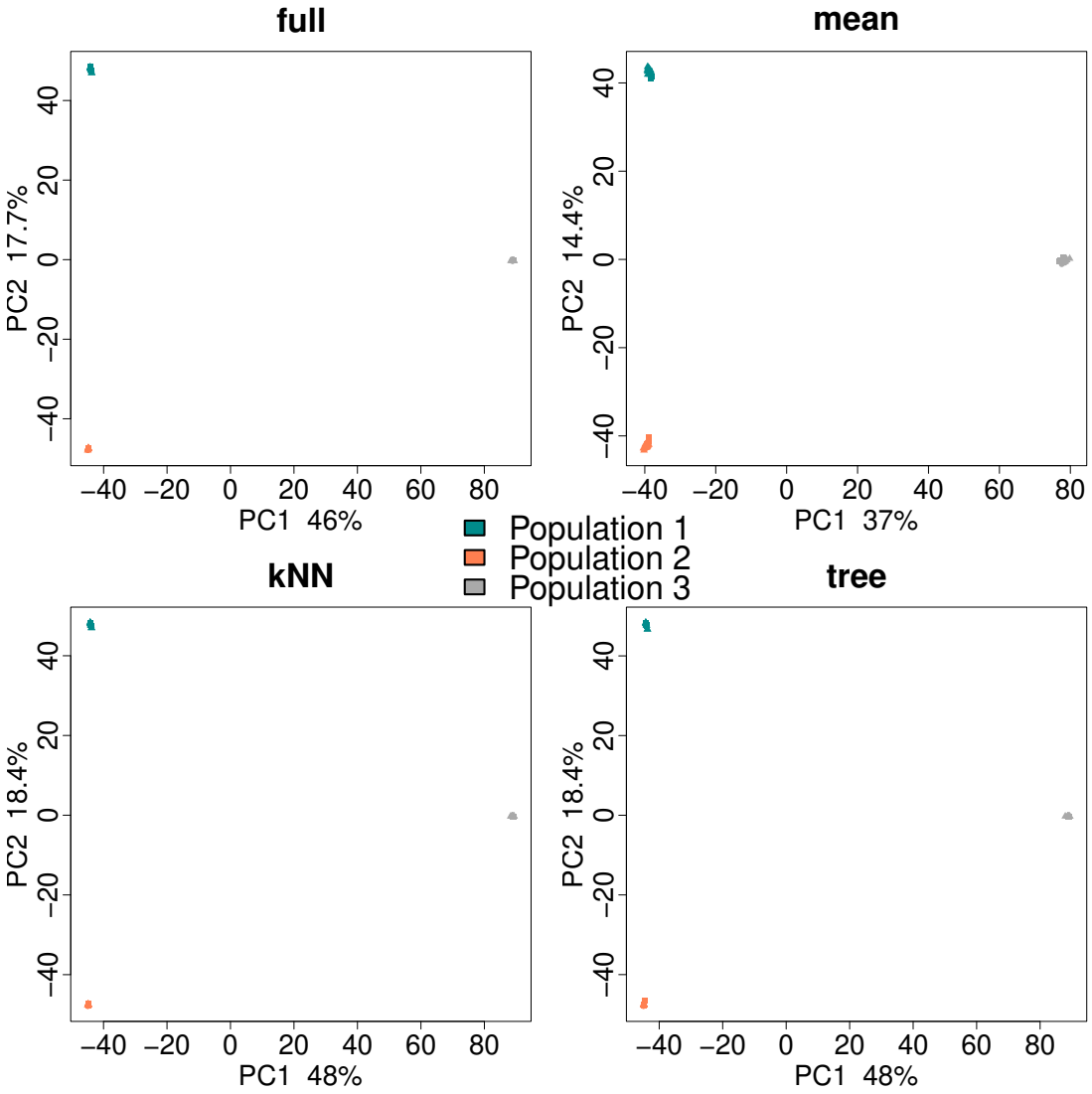

Figure 3.19: PCA of the initial and imputed datasets of simulation 2 (scenario 2).

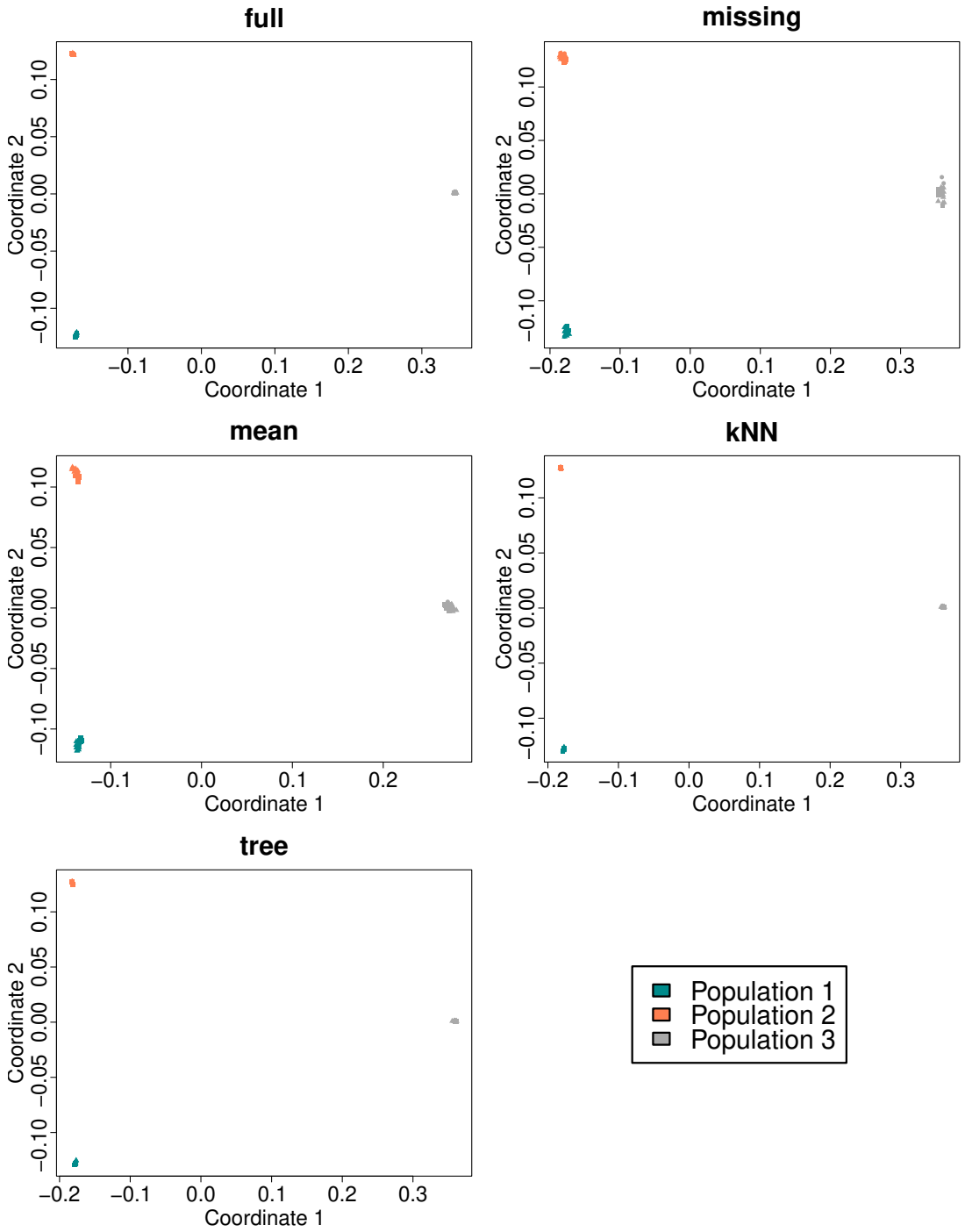

Figure 3.20: MDS before and after imputation for the simulation 2 (scenario 2).

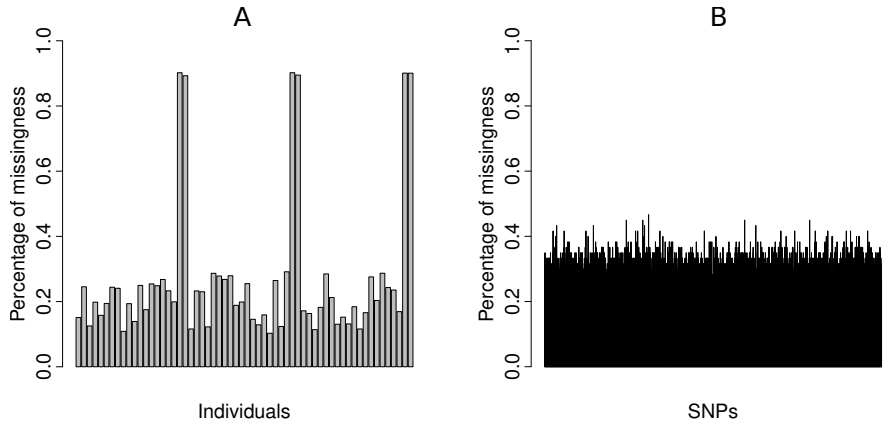

Figure 3.21: Pattern of missing data across individuals (A) and SNPs (B) in simulation 3.

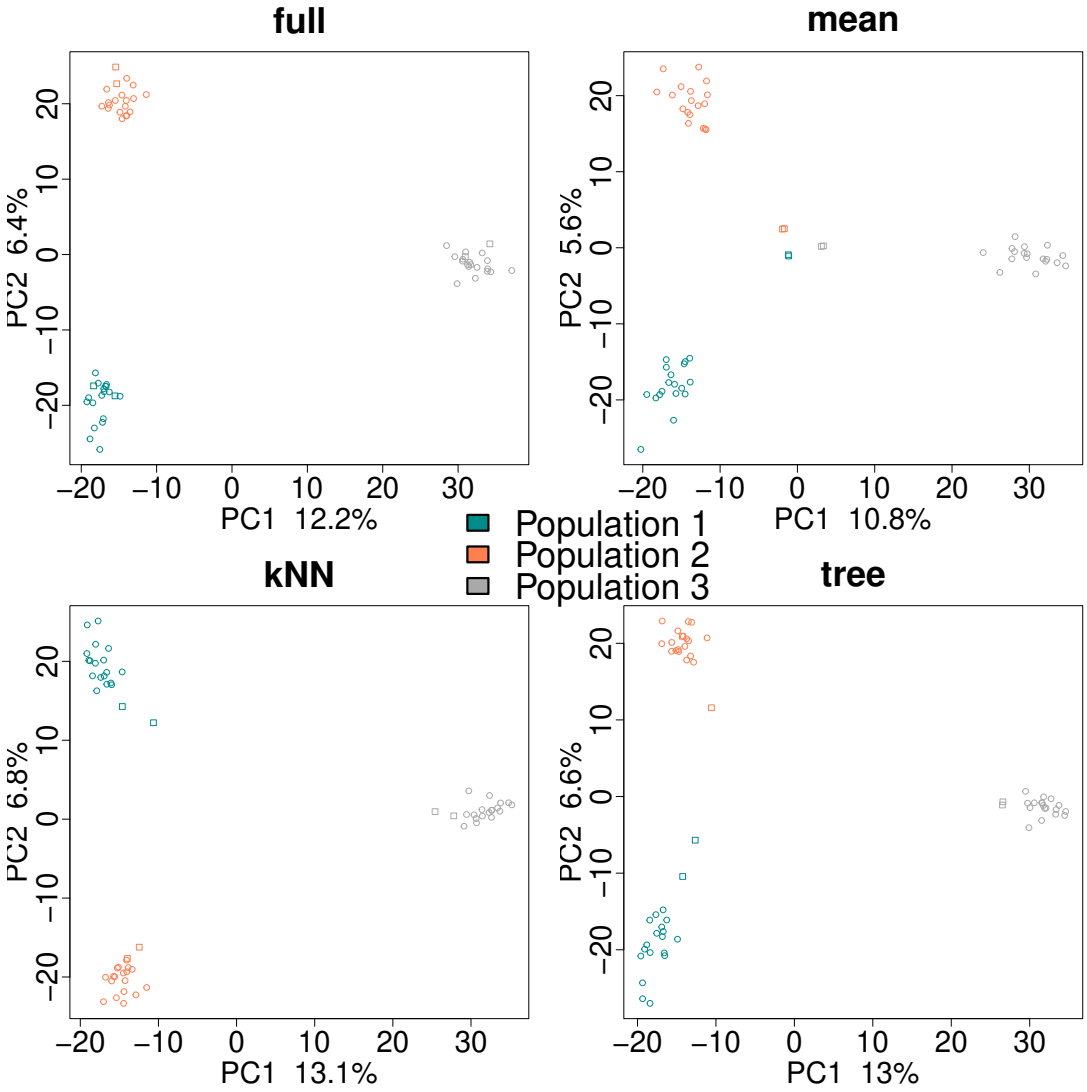

Figure 3.22: PCA of the initial and imputed datasets of simulation 3.

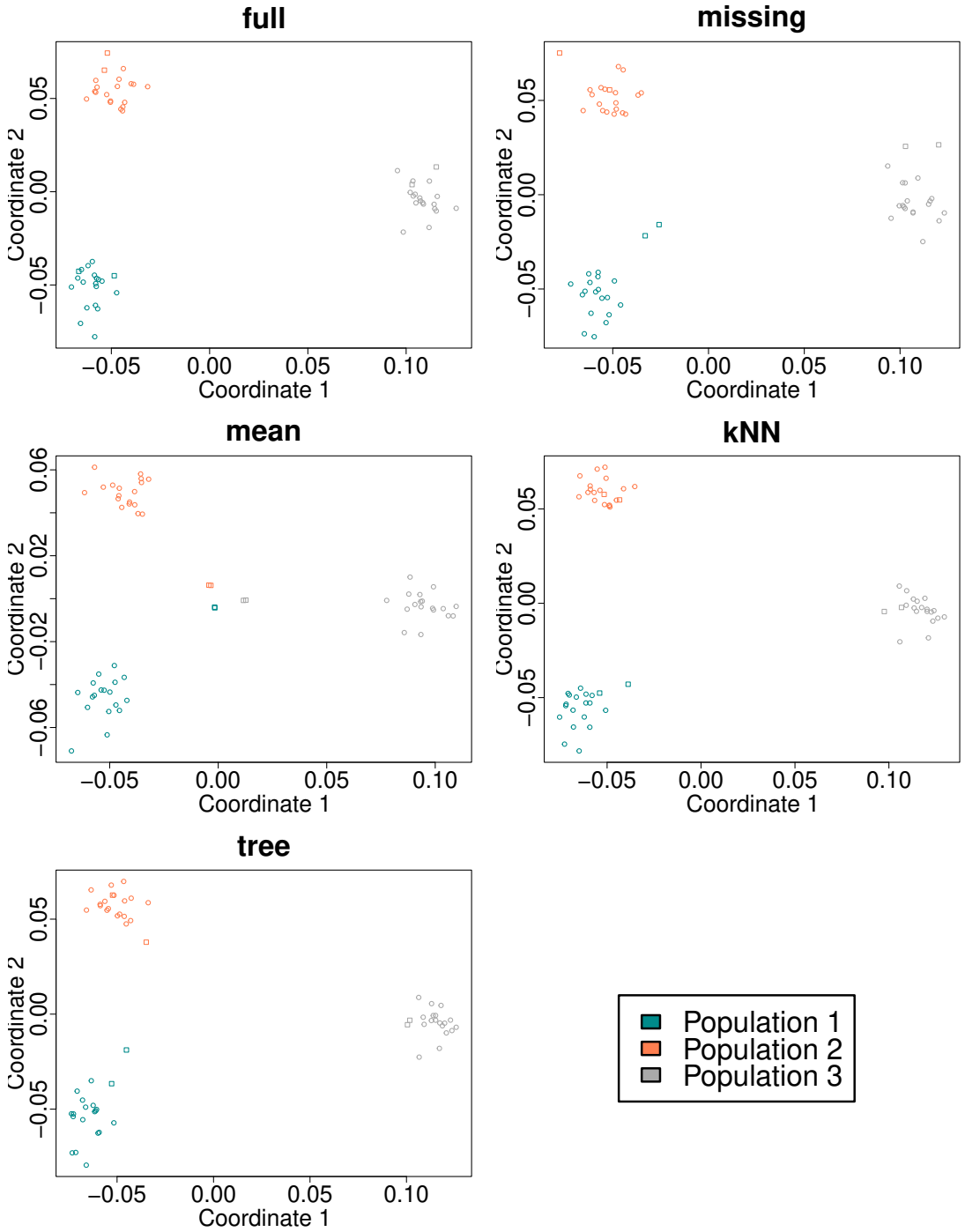

Figure 3.23: MDS before and after imputation for the simulation 3.

3.4.2 Evaluating the phylogeny-based imputation

In order to expand the evaluation of our proposed tree-based imputation approach beyond the downstream analyses of PCA and MDS, we detected and studied further the sites in which the imputation was unsuccessful. We focused on the last simulation, the one with recombination and extreme missingness in two individuals of each population. The percentage of these sites, also referred to as mis-imputed or in genotype discordance, was 7.5%. The initial dataset, simulating genotypes, had two distinct values; 0 indicating the ancestral state and 1 indicating the derived state. We, first, examined which state was most frequently imputed incorrectly. As depicted in Figure 3.24, most of the genotype discordance concerned sites of the derived state. This result is plausible, since initially the sites of the derived state were fewer and as a consequence the information about them for the imputation was hardly accessible. We, then, summarized the distribution of derived state frequencies by constructing the site frequency spectrum (SFS) (Figure 3.25). Based on that, we estimated the distribution of mis-imputed sites across the classes of the SFS (Figure 3.26). Interestingly, we noticed that intermediate frequency polymorphic sites are more frequently imputed incorrectly. This may be useful for improvement of imputation efficiency, by removing these sites in advance.

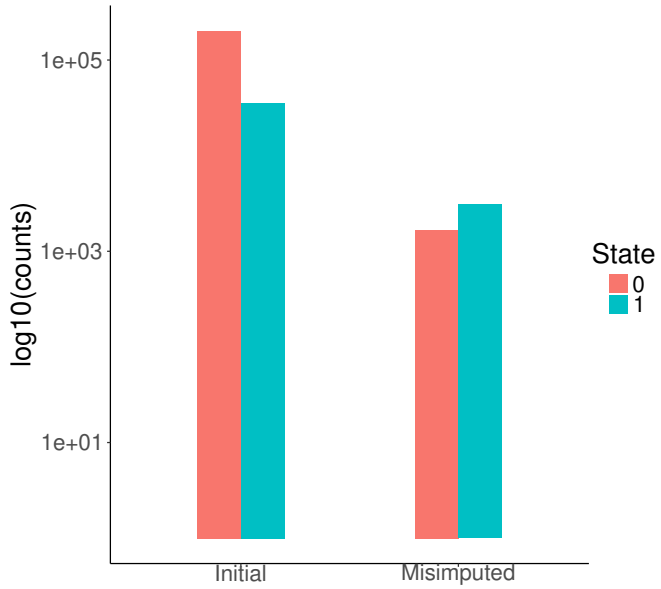

Figure 3.24: Number (in log scale) of ancestral (0) and derived (1) state in the initial dataset and and at mis-imputed sites.

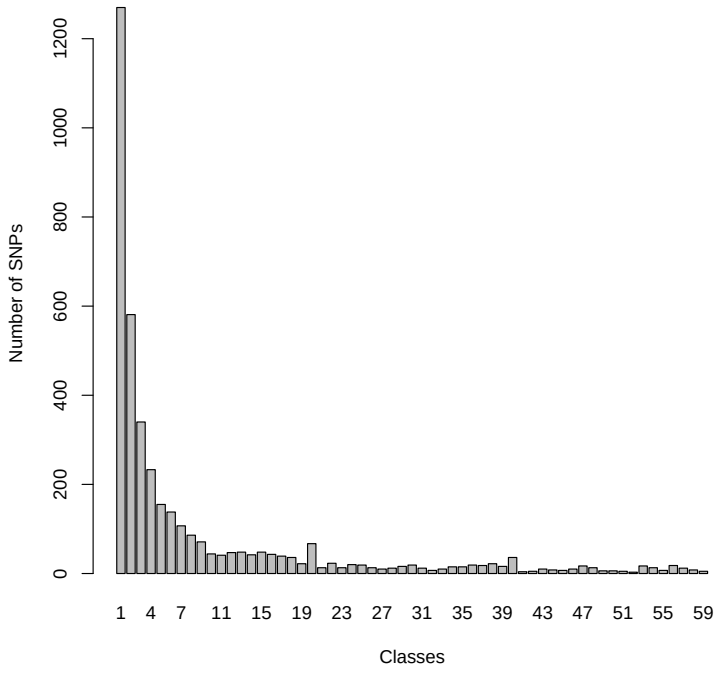

Figure 3.25: Site Frequency Spectrum (SFS) of the imputed dataset.

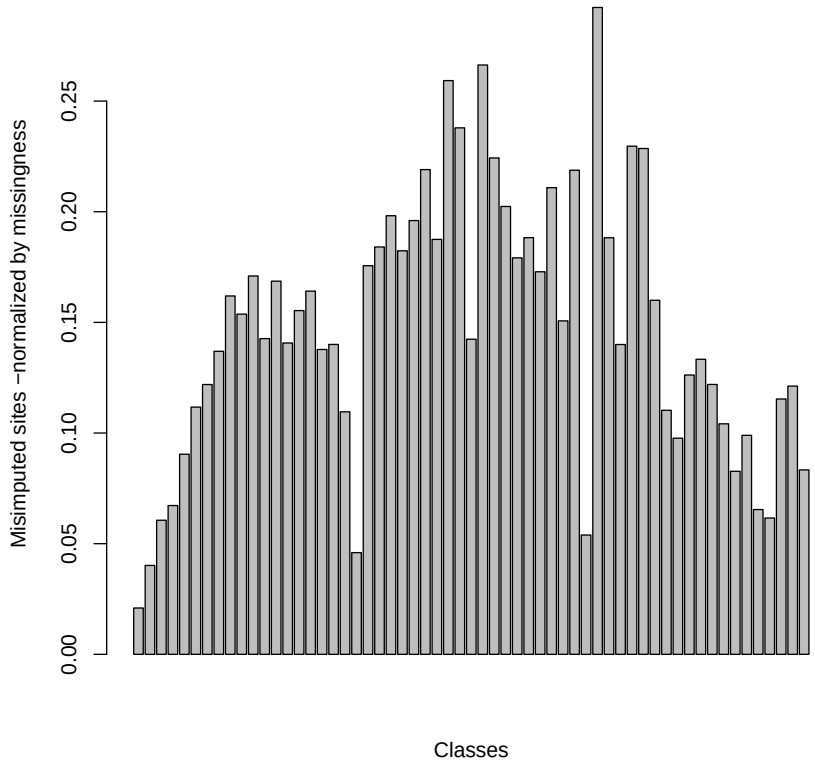

Figure 3.26: Frequency of mis-imputed sites across the classes of SFS.

3.4.3 Imputation on real data

Since kNN and tree imputation seemed to be accurate on simulated data, we implemented these approaches on real data, as well. We used a dataset of four ancient populations; Western Hunter Gatherers (WHG) and Neolithic from Anatolia, mainland Greece and Crete. The dataset consisted of 38 individuals and 10000 SNPs. The percentage of missing data was 35.5% and it was not uniformly distributed across the individuals, as depicted in Figure 3.27. We imputed the missing data by mean, kNN and tree approach and we performed PCA (Figure 3.28). The mean imputation clustered all the individuals with missingness greater than approximately 20% in the origin of PCA space, as observed on simulated data. kNN and phylogeny-based imputation had better performance, shaping the same pattern of population structure. We did not notice clear population clusters, but this does not mean that the observed representation does not reflect the real structure. As it is known, WHG used to live in Europe before the appearance of farmers from Anatolia during the early Neolithic. Thus, it is expected that these two groups were admixed. Such an admixture could easily occur also between Neolithic Anatolians and Greeks, both from mainland and Crete, as the former passed by Greece during their spread in Europe.

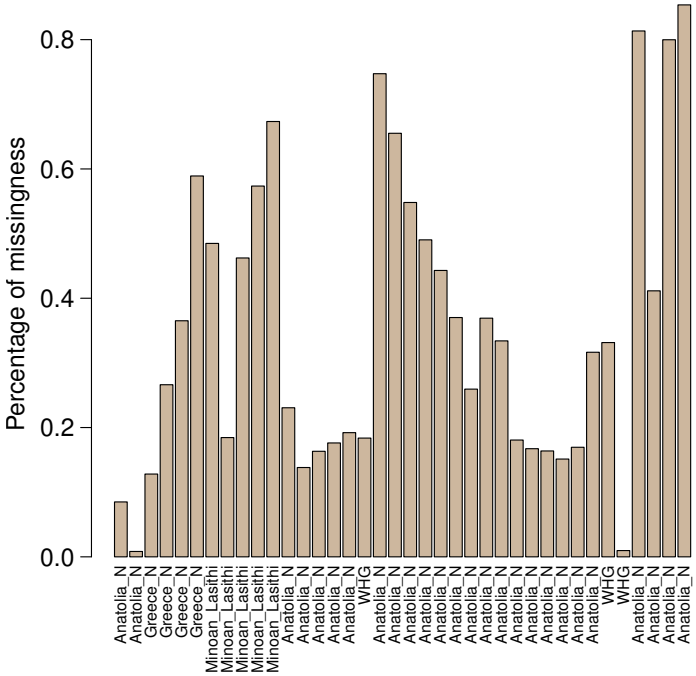

Figure 3.27: Percentage of missing data across individuals.

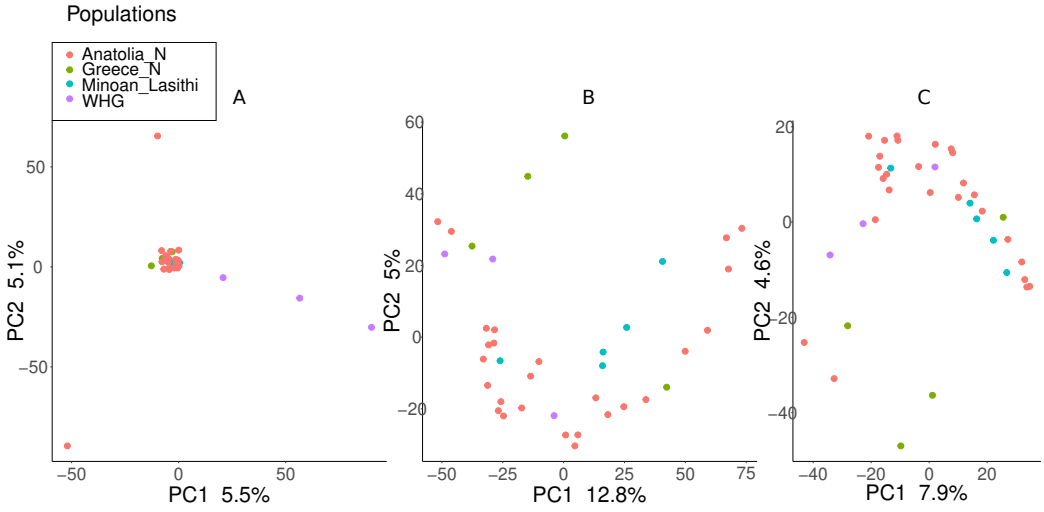

Figure 3.28: PCA of real data imputed by mean (A), kNN (B) and phylogeny (C).

CHAPTER 4 : **DISCUSSION**

4.1 DIMENSIONALITY REDUCTION METHODS

Dimensionality reduction methodologies have been widely-used for the detection of population structure in modern and ancestral populations using modern DNA and aDNA, respectively. They facilitate efforts to unravel the evolutionary history and connect the ancient genetic information with present-day variation by offering a visual representation of high-dimensional genetic data. The *smartpca* software Patterson et al. (2006) performs PCA on genotype data and is the most-widely used software to perform a PCA analysis on aDNA data. Typically, when aDNA data are analyzed, an approach for projection of ancient data on present-day data is followed. Often, such an approach results in PCA plots in which aDNA data are placed in the origin of the axes, mainly due to the vast amount of missing data that characterizes them Yi and Latch (2022). Similarly to Yi and Latch (2022), our results suggest that individuals biased with missing data would be placed progressively away from their real population clusters towards the origin of PCA plots as the amount of missingness increases, making them indistinguishable from true admixed individuals and potentially leading to misinterpreted population structure. Even though *smartpca* is extensively used in aDNA studies, in our analyses its performance was doubtful. The ancient samples were clustered together, albeit they were chronologically and geographically diverged. On top of that, the placement of their cluster in the PCA origin points out the bias introduced by the mean imputation, since the data are mean-centered prior to the eigenanalysis. This bias, which is related to the amount of missing data, is also demonstrated both in our simulated scenarios and in recent studies (Malan et al., 2020; Yi and Latch, 2022), supporting that missingness, especially non-uniformly distributed, leads to biased PCA-based inference of population structure. Fur-

thermore, Elhaik (2021) carried out an extensive empirical evaluation of PCA for a wide range of test cases and demonstrated that PCA fails to extract accurate conclusions in both simple and complex scenarios. More importantly, PCA results can be easily manipulated by intentionally introducing sampling biases. Since PCA is governed by data variance, sample size is a crucial factor of such an analysis. As Elhaik (2021) states, *'we do not believe that PCA can provide evidence of important migration events. Instead, our example shows how PCA can be used to generate multiple and alternative scenarios, all mathematically correct but, obviously, biologically incorrect'.*

Prior work of our group (Aspa Orfanou, 2021, *unpublished work*) has shown that under scenarios with high migration or bottleneck, which are more realistic, PCA becomes weak and should not be trusted. Thus, PCA, especially when it is performed with *smartpca*, on aDNA data with high amounts of missing data should be handled with caution.

The alternative approaches tested for dimensionality reduction and detection of population structure seemed more accurate. MDS is advantageous since it ignores the presence of missing data in the calculation of the pairwise distances, so it does not require imputation. In our simulations, we noticed that even in datasets with sequences harboring 60% of missing data, MDS analysis is able to reconstruct the true population structure. Therefore, our results suggest that MDS could be used as an exploratory tool prior to other more statistically elaborated analyses (for example, selective sweep detection and/or approximate Bayesian computation).

Similarly, a state-of-the-art methodology that employs f4-statistics values in a PCA framework (f4-PCA) is a valuable approach that can visualize relations directly among populations, rather than individuals as in typical PCA (Bergström et al., 2020). Other than that, it allows us to analyze both ancient and modern populations, releasing the need to project ancient samples on axes of modern variation and making it ambitious for accurate detection of population structure.

However, one should be mindful of potential bias that could be introduced both in f4-PCA and in MDS, because each pair of populations (or individuals) is characterized by a *different* set of common SNPs that are used for the pairwise analyses. Ideally, the set of SNPs with existing genotype information should be common across the dataset, but this is unrealistic due to the high amount of missing data.

The alternative, recently developed, approach for PCA for low coverage sequencing data, EMU (Meisner et al., 2021), was tested and obtained similar results with MDS. This method performs PCA but allows missing data, by modeling the missingness iteratively utilizing an expectation-maximization (EM) algorithm. We propose that both EMU and f4-PCA could be used as exploratory tools for the investigation of the relations among individuals or populations and the detection of population structure as a pre-analysis step, when necessary.

4.2 POPULATION STRUCTURE

The admixture analysis, a model-based approach, has been used for the classification of individuals based on their proportions of ancestry from K defined populations. Here, we pointed out that the estimation of the number K based on the cross-validation error of the model is inaccurate, since a single population source was suggested even if the samples were diverged. This issue of admixture modeling has been a long-standing discussion topic, as raised in Chapter 4 of Dutheil (2020). It is preferred to estimate the model for each K in a given range and observe how the patterns of ancestry proportions are shaping. In this approach, meaningful population substructure could be defined.

Conclusively, we propose *not to base the interpretation in just one type of analysis*. It is preferred to combine dimensionality reduction methods with admixture analysis, in order to reach more reliable results about population structure.

4.3 MISSING GENOTYPE DATA

The results of the above analyses have raised the issue of missing data and how it affects the inference of population structure. aDNA data is usually accompanied by extreme proportions of missing data, due to its low depth sequencing. Since the DNA from ancient specimens is valuable and can provide significant information about human history, such samples, even in low quality, cannot be ignored. Although, the requirement of full dataset in downstream analyses, such as PCA, limits the utilization of this data. Thus, the inference of unobserved genotypes, a process known as imputation is crucial for aDNA studies. We developed a novel method for imputation, which is based on the phylogenetic tree of the sequences and we compared it with the common approach of mean imputation and the clustering-based kNN imputation approach. Both tree-based and kNN approaches were efficient in our simulated scenarios. In contrast, we demonstrated the strong bias introduced by the mean imputation. In fact, samples with relatively high proportion of mean imputed data tend to be clustered in the proximity of the origin of a PCA plot, making the interpretation of the population relations misleading. It is important to note that the effect of mean imputation were solely observed when the missingness was non-randomly distributed across individuals.

During the evaluation of the tree-based imputation, studied which types of dataset sites (columns in a multiple sequence alignment) tend to give most of the erroenous imputation results. We found that classes of polymorphic sites with intermediate frequency tend to have high levels of genotype discordance. Thus, we propose *a priori* removal of such sites, in order to improve the imputation accuracy. In our third simulated scenario, recombination was taken into consideration. The phylogenetic tree-based and kNN-based imputation were efficient and the inter-population variation was conserved. This was a rather unexpected result since we strongly acknowledge the detrimental effect of the recombination in the accurate tree estimation. Even though population structure was successfully recovered,

within population diversity was probably reduced as a result of the 'guidance' provided by the phylogenetic tree. This intra-population effect will be the subject of a future work plan.

We extended our primary single-tree-based imputation by employing 'local' trees from defined regions of the genome, using information about recombination across the genome. In this way, each region would be represented by a local tree that would possibly be more accurate for the specific region from the tree inferred using the whole genome, leading to more precise imputation. Also, this approach accounts for linkage disequilibrium, since all SNPs located proximal to each other have been evolved under the same genealogy (and therefore they may have high levels of LD). However, due to the limited amount of information that is present locally, the process of tree reconstruction might be inaccurate. Thus, the idea of imputation based on local trees is still ongoing and we aspire for more robust results in the upcoming time.

CHAPTER 5 : **CONCLUSIONS**

This study set out to describe and evaluate methodologies for the inference of population structure, using data from ancient DNA (aDNA). We focused on Dimensionality Reduction techniques and admixture analysis and we pointed out the effect of missing data on these methods. A major finding concerns the widely used approach of projecting ancient individuals onto Principal Components inferred from the present-day variation in PCA, implemented by the smartpca software, which was proven doubtful and could produce misleading results. The alternative approaches of MDS and the state-of-the-art EMU and f4-PCA seemed more accurate than PCA and further studies concerning them would be worthwhile. The raised issue of the introduced bias by the missing data was followed by imputation approaches in order to overcome it. This study revealed the weakness of mean imputation and demonstrated the accuracy of kNN imputation in all of the simulated scenarios. Importantly, it contributed to the development of new imputation approaches as well, with our proposed phylogeny-based imputation which yielded as accurate results as kNN. However, it is preferable to kNN because it takes into account the evolutionary distances among the individuals. Summarizing, this work highlights the caution with which the interpretation of population relationships should be treated, especially in the presence of high missingness which is very common in aDNA and addresses the arised challenges of information loss by imputation, paving the way for more robust inferences in aDNA studies.

BIBLIOGRAPHY

D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, et al. Population genomics of bronze age eurasia. *Nature*, 522(7555):167–172, 2015.

K. Ausmees, F. Sanchez-Quinto, M. Jakobsson, and C. Nettelblad. An empirical evaluation of genotype imputation of ancient dna. *bioRxiv*, 2021.

M. Balter. Was north africa the launch pad for modern human migrations?, 2011.

A. Bergström, L. Frantz, R. Schmidt, E. Ersmark, O. Lebrasseur, L. Girdland-Flink, A. T. Lin, J. Storå, K.-G. Sjögren, D. Anthony, et al. Origins and genetic legacy of prehistoric dogs. *Science*, 370(6516):557–564, 2020.

K. I. Bos, V. J. Schuenemann, G. B. Golding, H. A. Burbano, N. Waglechner, B. K. Coombes, J. B. McPhee, S. N. DeWitte, M. Meyer, S. Schmedes, et al. A draft genome of yersinia pestis from victims of the black death. *Nature*, 478(7370):506–510, 2011.

K. I. Bos, K. M. Harkins, A. Herbig, M. Coscolla, N. Weber, I. Comas, S. A. Forrest, J. M. Bryant, S. R. Harris, V. J. Schuenemann, et al. Pre-columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis. *Nature*, 514(7523): 494–497, 2014.

K. I. Bos, G. Jäger, V. J. Schuenemann, Å. J. Vågene, M. A. Spyrou, A. Herbig, K. Nieselt, and J. Krause. Parallel detection of ancient pathogens via array-based dna capture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660):20130375, 2015.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, 15 (11):1496–1502, 2005.

F. Clemente, M. Unterländer, O. Dolgova, C. E. G. Amorim, F. Coroado-Santos, S. Neuenschwander, E. Ganiatsou, D. I. C. Dávalos, L. Anchieri, F. Michaud, et al. The genomic history of the aegean palatial civilizations. *Cell*, 184(10):2565–2586, 2021.

. G. P. Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.

Cox and Cox. *Multidimensional scaling in Handbook of data visualization.* Springer Science & Business Media, 2007.

J. Dabney, M. Meyer, and S. Pääbo. Ancient dna damage. *Cold Spring Harbor perspectives in biology*, 5(7):a012567, 2013.

M. Dehasque, M. C. Ávila-Arcos, D. Díez-del Molino, M. Fumagalli, K. Guschanski, E. D. Lorenzen, A.-S. Malaspinas, T. Marques-Bonet, M. D. Martin, G. G. Murray, et al. Inference of natural selection from ancient dna. *Evolution Letters*, 4(2):94–108, 2020.

T. R. Disotell. Human evolution: origins of modern humans still look recent. *Current Biology*, 9(17):R647–R650, 1999.

J. Y. Dutheil. *Statistical population genomics.* Springer Nature, 2020.

E. Elhaik. Why most principal component analyses (pca) in population genetic studies are wrong. *BioRxiv*, 2021.

W. J. Ewens. *Mathematical population genetics: theoretical introduction*, volume 1. Springer, 2004.

L. Fehren-Schmitz and L. Georges. Ancient dna reveals selection acting on genes associated with hypoxia response in pre-columbian peruvian highlanders in the last 8500 years. *Scientific reports*, 6(1):1–11, 2016.

J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics*, 25(5):471, 1973.

V. Fernandes, F. Alshamali, M. Alves, M. D. Costa, J. B. Pereira, N. M. Silva, L. Cherni, N. Harich, V. Cerny, P. Soares, et al. The arabian cradle: mitochondrial relicts of the first steps along the southern route out of africa. *The American Journal of Human Genetics*, 90(2):347–355, 2012.

R. A. Fisher. The causes of human variability. *The Eugenics Review*, 10(4):213, 1919.

L. A. Frantz, J. Haile, A. T. Lin, A. Scheu, C. Geörg, N. Benecke, M. Alexander, A. Linderholm, V. E. Mullin, K. G. Daly, et al. Ancient pigs reveal a near-complete genomic turnover following their introduction to europe. *Proceedings of the National Academy of Sciences*, 116(35):17231–17238, 2019.

M. Fumagalli, F. G. Vieira, T. S. Korneliussen, T. Linderoth, E. Huerta-Sánchez, A. Albrechtsen, and R. Nielsen. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, 195(3):979–992, 2013.

G. González-Fortes, E. R. Jones, E. Lightfoot, C. Bonsall, C. Lazar, A. Grandal-d'Anglade, M. D. Garralda, L. Drak, V. Siska, A. Simalcsik, et al. Paleogenomic evidence for multi-generational mixing between neolithic farmers and mesolithic hunter-gatherers in the lower danube basin. *Current Biology*, 27(12):1801–1810, 2017.

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, et al. A draft sequence of the neandertal genome. *science*, 328(5979):710–722, 2010.

D. Gronenborn. A variation on a basic theme: the transition to farming in southern central europe. *Journal of world prehistory*, 13(2):123–210, 1999.

T. Günther, H. Malmström, E. M. Svensson, A. Omrak, F. Sánchez-Quinto, G. M. Kılınç, M. Krzewińska, G. Eriksson, M. Fraser, H. Edlund, A. R. Munters, A. Coutinho, L. G. Simões, M. Vicente, A. Sjölander, B. Jansen Sellevold, R. Jørgensen, P. Claes, M. D. Shriver, C. Valdiosera, M. G. Netea, J. Apel, K. Lidén, B. Skar, J. Storå, A. Götherström, and M. Jakobsson. Population genomics of mesolithic scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLOS Biology*, 16(1):1–22, 01 2018. URL `https://doi.org/10.1371/journal.pbio.2003703`.

W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, et al. Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211, 2015.

E. Hagelberg, L. S. Bell, T. Allen, A. Boyde, S. J. Jones, and J. B. Clegg. Analysis of ancient bone dna: techniques and applications. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 333(1268):399–407, 1991.

D. L. Hartl, A. G. Clark, and A. G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman. The great human expansion. *Proceedings of the National Academy of Sciences*, 109(44):17758–17764, 2012.

R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, and A. C. Wilson. Dna sequences from the quagga, an extinct member of the horse family. *nature*, 312(5991):282–284, 1984.

Z. Hofmanová, S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, D. Díez del Molino, L. van Dorp, S. López, A. Kousathanas, V. Link, K. Kirsanow, L. Cassidy, R. Martiniano, M. Strobel, A. Scheu, K. Kotsakis, P. Halstead, S. Triantaphyllou, N. Kyparissi, and J. Burger. Early farmers from across europe directly descended from neolithic aegeans. *Proceedings of the National Academy of Sciences*, 113:201523951, 06 2016. doi: 10.1073/pnas.1523951113.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5 (6):e1000529, 2009.

R. R. Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

M. Jamshidian and P. M. Bentler. Ml estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and behavioral Statistics*, 24(1):21–24, 1999.

L. Jorde. Genetic variation and human evolution. https://education.nsw.gov.au/content/dam/main-education/teaching-and-learning/curriculum/key-learning-areas/science/s-6/biology/Genetic_variation_and_human_evolution_resource.pdf, 2020. Accessed: 2022-04-27.

T. H. Jukes, C. R. Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

M. Kimura. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetics research*, 11(3):247–270, 1968.

M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

J. F. C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3): 235–248, 1982.

I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, et al. Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 513(7518): 409–413, 2014.

S.-Y. Lee. *Handbook of latent variable and related models*. Elsevier, 2011.

R. C. Lewontin and J. L. Hubby. A molecular approach to the study of genic heterozygosity in natural populations. ii. amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura. *Genetics*, 54(2):595, 1966.

T. Lindahl. Instability and decay of the primary structure of dna. *nature*, 362(6422): 709–715, 1993.

M. Lipson, O. Cheronet, S. Mallick, N. Rohland, M. Oxenham, M. Pietrusewsky, T. O. Pryce, A. Willis, H. Matsumura, H. Buckley, et al. Ancient genomes document multiple waves of migration in southeast asian prehistory. *Science*, 361(6397):92–95, 2018.

N. Liu and H. Zhao. A non-parametric approach to population structure inference using multilocus genotypes. *Human genomics*, 2(6):1–12, 2006.

I. Maceda, M. M. Álvarez, G. Athanasiadis, R. Tonda, J. Camps, S. Beltran, A. Camps, J. Fàbrega, J. Felisart, J. Grané, et al. Fine-scale population structure in five rural populations from the spanish eastern pyrenees using high-coverage whole-genome sequence data. *European Journal of Human Genetics*, 29(10):1557–1565, 2021.

L. Malan, C. M. Smuts, J. Baumgartner, and C. Ricci. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. *Nutrition Research*, 75:67–76, 2020.

J. Marchini, L. R. Cardon, M. S. Phillips, and P. Donnelly. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512–517, 2004.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, et al. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, 528(7583):499–503, 2015.

I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, et al. The genomic history of southeastern europe. *Nature*, 555(7695):197–203, 2018.

G. McVean. A genealogical interpretation of principal components analysis. *PLoS genetics*, 5(10):e1000686, 2009.

J. Meisner, S. Liu, M. Huang, and A. Albrechtsen. Large-scale inference of population structure in presence of missingness using pca. *Bioinformatics*, 37(13):1868–1875, 2021.

M. Melé, A. Javed, M. Pybus, P. Zalloua, M. Haber, D. Comas, M. G. Netea, O. Balanovsky, E. Balanovska, L. Jin, et al. Recombination gives a new insight in the effective population size and the history of the old world human populations. *Molecular biology and evolution*, 29(1):25–30, 2012.

P. Menozzi, A. Piazza, and L. Cavalli-Sforza. Synthetic maps of human gene frequencies in europeans: These maps indicate that early farmers of the near east spread to all of europe in the neolithic. *Science*, 201(4358):786–792, 1978.

D. Money, K. Gardner, Z. Migicovsky, H. Schwaninger, G.-Y. Zhong, and S. Myles. Linkimpute: fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genetics*, 5(11):2383–2390, 2015.

J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.

C. Ottoni, W. Van Neer, B. De Cupere, J. Daligault, S. Guimaraes, J. Peters, N. Spassov, M. E. Prendergast, N. Boivin, A. Morales-Muñiz, et al. The palaeogenetics of cat dispersal in the ancient world. *Nature Ecology & Evolution*, 1(7):1–7, 2017.

S. Pääbo, H. Poinar, D. Serre, V. Jaenicke-Després, J. Hebler, N. Rohland, M. Kuch, J. Krause, L. Vigilant, and M. Hofreiter. Genetic analyses from ancient dna. *Annu. Rev. Genet.*, 38:645–679, 2004.

P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. Pca-correlated snps for structure identification in worldwide human populations. *PLoS genetics*, 3(9):e160, 2007.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3): 1065–1093, 2012.

B. O. Petrazzini, H. Naya, F. Lopez-Bello, G. Vazquez, and L. Spangenberg. Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData mining*, 14(1):1–13, 2021.

H. N. Poinar and A. Cooper. Ancient dna: do it right or not at all. *Science*, 5482(1139): 416, 2000.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

S. Purcell and P. Sham. Properties of structured association approaches to detecting population stratification. *Human heredity*, 58(2):93–107, 2004.

M. Raghavan, P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T. W. Stafford Jr, L. Orlando, E. Metspalu, et al. Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, 505(7481):87–91, 2014.

M. M. Rahman and D. N. Davis. Fuzzy unordered rules induction algorithm used as missing value imputation methods for k-mean clustering on real cardiovascular data. *Lect Notes Eng Comput Sci*, 2197(1):391–4, 2012.

D. Reich, A. L. Price, and N. Patterson. Principal component analysis of genetic data. *Nature genetics*, 40(5):491–492, 2008.

D. Reich, K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. Reconstructing indian population history. *Nature*, 461(7263):489–494, 2009.

D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, et al. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053–1060, 2010.

D. Reich, N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni, I. Pugach, A. M.-S. Ko, Y.-C. Ko, T. A. Jinam, M. E. Phipps, et al. Denisova admixture and the first modern human dispersals into southeast asia and oceania. *The American Journal of Human Genetics*, 89(4):516–528, 2011.

D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001.

C. Renfrew. Models of change in language and archaeology. *Transactions of the Philological Society*, 87(2):103–155, 1989.

M. P. Richards, R. Jacobi, J. Cook, P. B. Pettitt, and C. B. Stringer. Isotope evidence for the intensive use of marine foods by late upper palaeolithic humans. *Journal of Human Evolution*, 49(3):390–394, 2005.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

S. Rubinacci, O. Delaneau, and J. Marchini. Genotype imputation using the positional burrows wheeler transform. *PLoS genetics*, 16(11):e1009049, 2020.

V. J. Schuenemann, P. Singh, T. A. Mendum, B. Krause-Kyora, G. Jäger, K. I. Bos, A. Herbig, C. Economou, A. Benjak, P. Busso, et al. Genome-wide comparison of medieval and modern mycobacterium leprae. *Science*, 341(6142):179–183, 2013.

H. P. Schwarcz and R. Grün. Electron spin resonance (esr) dating of the origin of modern man. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 337(1280):145–148, 1992.

H. Schwender. Imputing missing genotypes with weighted k nearest neighbors. *Journal of Toxicology and Environmental Health, Part A*, 75(8-10):438–446, 2012.

E.-L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M.-D. Cubiles-de-la Vega. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, 24(1):121–129, 2011.

P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, and M. Jakobsson. Separating endogenous ancient dna from modern day contamination in a siberian neandertal. *Proceedings of the National Academy of Sciences*, 111(6): 2229–2234, 2014.

P. Skoglund, E. Ersmark, E. Palkopoulou, and L. Dalén. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, 25(11):1515–1519, 2015.

M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68 (4):978–989, 2001.

C. Stringer. The status of homo heidelbergensis (schoetensack 1908). *Evolutionary Anthropology: Issues, News, and Reviews*, 21(3):101–107, 2012.

S. Tavaré et al. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

A. G. Thorne and M. H. Wolpoff. The multiregional evolution of humans. *Scientific American*, 266(4):76–83, 1992.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

P. Verdu, T. J. Pemberton, R. Laurent, B. M. Kemp, A. Gonzalez-Oliver, C. Gorodezky, C. E. Hughes, M. R. Shattuck, B. Petzelt, J. Mitchell, et al. Patterns of admixture and population structure in native populations of northwest north america. *PLoS genetics*, 10(8):e1004530, 2014.

K. Wang, S. Goldstein, M. Bleasdale, B. Clist, K. Bostoen, P. Bakwa-Lufu, L. T. Buck, A. Crowther, A. Dème, R. J. McIntosh, et al. Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-saharan africa. *Science Advances*, 6(24):eaaz0183, 2020.

S. Wright et al. Genetical structure of populations. *Nature*, 166:247–49, 1950.

M. A. Yang, X. Fan, B. Sun, C. Chen, J. Lang, Y.-C. Ko, C.-h. Tsang, H. Chiu, T. Wang, Q. Bao, et al. Ancient dna indicates human population shifts and admixture in northern and southern china. *Science*, 369(6501):282–288, 2020.

X. Yi and E. K. Latch. Nonrandom missing data can bias principal component analysis inference of population genetic structure. *Molecular Ecology Resources*, 22(2):602–611, 2022.