

Assessing ancestry through genetics: an attempt using skeletonized human remains

Ana Sofia Alves Botelho

Mestrado em Genética Forense

Departamento de Biologia

2020

Orientador

Eugénia Cunha, PhD, INMLCF/LFA/UC

Coorientador

Rui Pereira, PhD, IPATIMUP/i3S

Coorientador

Armando Serra, Especialista Superior, INMLCF

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

“If we knew what it was we were doing, it would not be called research, would it?”

Albert Einstein

ACKNOWLEDGMENTS

2020 foi definitivamente um ano atípico. Assim, não podia deixar de agradecer aos que me ajudaram e apoiaram durante este ano.

Em primeiro lugar, à Professora Eugénia Cunha por ter aceitado orientar-me sem qualquer reserva mesmo não sendo a sua especialidade. Por me ter dado a oportunidade de desenvolver este projeto, e por me ter dado a confiança para o fazer de forma autónoma.

Ao Dr. Rui Pereira por todo o conhecimento transmitido e todas as horas de discussão que melhoraram significativamente esta tese. Pode não acreditar, mas aprendi muito. A sério.

Ao Dr. Armando Serra, palavras não chegam para agradecer todo o apoio que me deu em Coimbra. Não só pela partilha de conhecimento, mas por todas as conversas que me fizeram sentir um bocadinho mais em casa. Será das melhores memórias que trago dessa cidade.

À Dra. Lisa Sampaio por me ter aberto as portas do INMLCF e por me ajudar a desenvolver este projeto. A todos do Serviço de Biologia e Genética Forense de Coimbra – Filipa, Ana Margarida, Nair, Pedro, Vanessa, Patrícia, Marta, Dra. Maria João, Dra. Virgínia, Celestinha, Paula e Tiago – obrigada por me ter recebido tão bem, e por estarem sempre disponíveis para me ajudar.

Tenho também de agradecer à Professora Teresa Ferreira do Laboratório de Antropologia Forense por toda a ajuda e acompanhamento e por me ter permitido ser a "destruidora de ossos" durante este ano. Também à Catarina Coelho por toda a partilha de conhecimentos e principalmente por me ter disponibilizado os seus resultados. À Helena Dias por todo o apoio inicial e conselhos dados.

Às minhas girls Ana Lu, Lameiras, Lu e Nico por estarem sempre lá e por serem as melhores! Por todos os jantares, férias, conversas e desabafos, e por me aceitarem mesmo com os meus horários complicados. Crescer com vocês ao meu lado tem outro significado.

À minha Feijoquinha. Porque apesar de teres escolhido o pior ano para nascer, foste o meu escape quando eu mais precisei.

A todas as meninas do MGF por estes dois anos, especialmente à brazuca Mayla por me ouvir e estar sempre lá com uma atitude positiva. Austrália nos espera!

À minha famelga “Alves e Companhia” por serem os maiores chatos e por me fazerem sentir a mais sortuda por vos ter. Pelos almoços de domingo, e por compreenderem quando eu não estou. Esta “Doutora” está quase a acabar!

Ao pessoal da BlueClinical por todos os “então como vai essa tese?” e por toda a compreensão nesta altura mais difícil.

A todos os outros que me ajudaram e apoiaram durante este ano, um grande obrigada. Agora sim, podemos ir sair!

ABSTRACT

Ancestry or biogeographical origin of an individual gives information not only about the probable place of origin, but, indirectly, about phenotypic information that the police can eventually use to direct an investigation. Forensic Genetics uses Ancestry Informative Markers (AIMs) that have a high difference in the alleles' frequencies between human major population groups to infer the probable biogeographical origin of a given sample. Forensic Anthropology uses mainly metric and non-metric evaluations of the skull to do the same inference.

Therefore, this study aims to infer the genetic ancestry of 56 skeletonized human remains using a multiplex PCR reaction of 46 AIM-Indel and then classifying each individual as African, European, East Asian, Native American, or Oceanian using a likelihood ratios approach. Ancestral membership proportions were also estimated using STRUCTURE software. Thereafter, the results will be compared with *ante-mortem* records available with information about the place of birth, and with the ancestry assessment made with anthropological methods.

DNA extraction was possible in all 56 individuals with DNA concentrations that ranged from 0.0003 ng/ µL to 0.1791 ng/ µL. The lowest number of amplified markers was 15 while the highest was 46. The success rate of this genotyping method was 81%, which demonstrated the amount of DNA degradation present in the studied samples. Some individuals presented markedly low LR values probably due to the high levels of missing data, or due to the fact that those were admixed individuals.

European probable origin was inferred for 43 individuals, while only one individual was classified as having an African probable origin. Twelve individuals revealed an admixed biogeographical origin according to the cluster membership proportions estimated by STRUCTURE. Four of these individuals were thought to have a South American ancestry – most likely from Brazil; five individuals had admixture between the European and African groups; one individual had shared membership with the European and Oceanian ancestral groups, and another had contributions from the African and Oceanian ancestral groups. These inferences are in accordance with the Portuguese history of colonization and with the recent events of migration.

Ante-mortem records were available for five individuals, corroborating the genetic findings for three individuals, while the other two shown incompatibilities between the place of birth and the genetic profile estimated.

The anthropological assessment did not always match the one made by genetics, but from an overall perspective, when non-admixed individuals are concerned, the two sciences are in agreement. When it comes to admixed individuals, anthropological methods seem to have more difficulty to detect all ancestry components involved.

According to these results, genetic assessment of ancestry with this panel of AIM-Indels can be successful even in cases when DNA is severely degraded, which can be extremely useful in forensic cases. Anthropological methods, although successful when there is only one major ancestral contribution to detect, should be applied cautiously when it comes to admixed individuals.

Nevertheless, the evaluation of the biogeographical origin of an individual should be an interdisciplinary work between the two sciences, since each can gather information that can complement the other.

KEYWORDS: Ancestry, Forensic Genetics, Forensic Anthropology, Skeletal Remains, DNA, Ancestry Informative Markers (AIMs), Indels.

RESUMO

A ancestralidade ou origem biogeográfica de um determinado indivíduo fornece, não só informação sobre a provável região de origem, mas também, indiretamente, sobre características fenotípicas que, eventualmente, a polícia pode usar para conduzir uma investigação. A Genética Forense utiliza Marcadores Informadores de Ancestralidade (AIMs) que possuem uma elevada diferença nas frequências alélicas entre os principais grupos populacionais para inferir sobre a origem biogeográfica mais provável de uma amostra. A Antropologia Forense executa avaliações métricas e não-métricas no crânio, maioritariamente, para estimar a mesma informação.

Assim, este estudo visa inferir a ancestralidade genética de 56 restos humanos esqueletizados usando uma reação de PCR em multiplex com 46 AIM-Indels, e seguidamente, classificar cada indivíduo como Africano, Europeu, Este Asiático, Nativo Americano, ou Oceânico utilizando uma abordagem de likelihood ratios. As proporções de contribuições ancestrais foram estimadas usando o software STRUCTURE. De seguida, os resultados obtidos foram comparados com os registos *ante mortem* disponíveis que continham a naturalidade dos indivíduos estudados, e com a estimativa de ancestralidade conseguida através de métodos antropológicos.

A extração do DNA foi conseguida em todos os 56 indivíduos com concentrações de DNA que variaram entre 0.0003 ng/ µL e 0.1791 ng/ µL. O menor número de marcadores amplificados foi de 15 e o maior foi 46. A taxa de sucesso deste método de genotipagem foi de 81%, o que é demonstrativo do nível de degradação de DNA presente nestas amostras. Alguns indivíduos apresentaram valores de LR bastante baixos, provavelmente devido aos elevados níveis de loci não reportados, ou, devido à sua provável natureza miscigenada.

Para 43 indivíduos foi estimada uma provável origem Europeia, e apenas um revelou uma provável origem Africana. Doze indivíduos revelaram uma origem miscigenada de acordo com as proporções estimadas pelo STRUCTURE. Estima-se que 4 destes indivíduos tenham provavelmente origem Sul Americana – possivelmente do Brasil; cinco indivíduos revelaram mistura variável entre os componentes Europeus e Africanos; um indivíduo revelou contribuição Europeia e Oceânica; enquanto outro apresentou componentes ancestrais Africanas e Oceânicas. Estas inferências ancestrais estão geralmente de acordo com a história da colonização portuguesa e com os recentes fenómenos de migração.

Os registos *ante-mortem* foram apenas disponibilizados para 5 indivíduos, sendo que estes corroboraram os resultados genéticos obtidos para 3 dos indivíduos, enquanto outros 2 apresentaram incompatibilidades entre a naturalidade e a ancestralidade estimada.

A avaliação antropológica nem sempre correspondeu à avaliação genética, mas numa perspetiva geral, quando se trata de indivíduos não miscigenados, as duas ciências estão de acordo. No caso de indivíduos miscigenados, os métodos antropológicos parecem ter mais dificuldades para detetar os componentes ancestrais envolvidos.

De acordo com os resultados obtidos, pode concluir-se a avaliação genética de ancestralidade com este painel de AIM-Indels pode ser bem-sucedida mesmo em casos em que o DNA está seriamente degradado, o que pode ser extremamente útil em casos forenses. Os métodos antropológicos, apesar de bem-sucedidos no caso de haver apenas uma componente ancestral, devem ser aplicados com prudência no que toca a indivíduos com mistura de componentes ancestrais.

Assim, a estimativa da origem ancestral de um indivíduo deve ser um trabalho interdisciplinar entre as duas ciências, visto que cada uma pode recolher informações complementares.

Palavras-chave: Ancestralidade, Genética Forense, Antropologia Forense, Restos Mortais, DNA, Marcadores Informativos de Ancestralidade (AIMs), Indels.

CONTENTS

INTRODUCTION 1
Human Genome 1
Patterns of Human Variation 1
Forensic Science 3
I. A Brief History 3
II. Forensic Genetics 5
a. DNA recovery 5
b. Genetic Markers 6
i. Short Tandem Repeats 7
ii. Single Nucleotide Polymorphisms 8
iii. Insertion/ deletion polymorphisms 9
III. Forensic Anthropology 10
Ancestry 11
I. Through Genetics 13
a. AIMs 13
II. Through Anthropology 14
a. Metric Assessment 15
b. Non-metric assessment 15
AIMS 17
MATERIALS AND METHODS 18
Sample selection and collection 18
Sample Preparation 18
DNA extraction 19
DNA quantification 20
AIM-Indel Genotyping 20
Statistical Analysis 21
Anthropological data 23
RESULTS AND DISCUSSION 24
DNA extraction and genotyping performance 24
Inference of Genetic Ancestry 27
I. Non-Admixed Individuals 32
II. Admixed individuals 35
a. Major European Ancestral Component 37

b. Major African Ancestral Component 42
Comparison with Anthropological data 45
I. Non-admixed individuals according to genetics 45
II. Admixed individuals according to genetics 46
CONCLUDING REMARKS 49
REFERENCES 52
SUPPLEMENTARY MATERIAL 60

LIST OF FIGURES

Figure 1: an example of two alleles in an STR' locus. From Goodwin (2011). 7
Figure 2: Example of a window cut in the femur of the individual CEI/XXI_111. 19
Figure 3: Representation of the variants (allele 3) of the markers MID2264 and MID360 found in samples CEI_100, CEI_262, and CEI_209 (markers are identified by MID number). 25
Figure 4: Correlation between the number of reportable AIM-Indel loci and DNA concentration in sample extracts. Orange bars represent the number of reportable markers; blue dots represent DNA quantification values; green line represents the maximum number of reportable markers. 26
Figure 5: STRUCTURE plot showing CEI samples for K varying from 2 to 5. Each vertical bar represents an individual. Differently coloured zones on each vertical bar show the proportion of the membership of each population component. 31
Figure 6: Principal Component Analysis (PCA) plot showing the two components of variation for the individuals with only one major membership according to STRUCTURE, compared to the reference data of five populations. 32
Figure 7: Principal Component Analysis (PCA) plot showing the two components of variation for the admixed individuals according to STRUCTURE, compared to the reference data of five populations. 35
Figure 8: Principal Component Analysis (PCA) plot showing the two components of variation for CEI_239, CEI_242, CEI_254, and CEI_290, using the reference samples of Africans, East Asians, and Europeans. 38
Figure 9: Principal Component Analysis (PCA) plot showing the two components of variation for the samples CEI_133, CEI_203, and CEI_296, using the reference samples of Africans, Europeans, and Native Americans. 40
Figure 10: Principal Component Analysis (PCA) plot showing the two components of variation for the sample CEI_202, using the reference samples of Europeans, East Asians, and Oceanians. 41
Figure 11: Principal Component Analysis (PCA) plot showing the two components of variation for the sample CEI_281, using the reference samples of Africans, Europeans, and East Asians. 44
Figure S1.a – k: Different expressions of the eleven non-metric traits according to Hefner (2009). Each expression is associated with a 'stereotypical association' - the ancestral group commonly associated with this trait (AFR, EUR, EAS). Adapted from Iscan and Steyn (2013). 67

LIST OF TABLES

Table 1: Information about the place of birth (nationality) that was possible to retrieve from records. 18

Table 2: Reference samples used in this study that belong to the HGDP-CEPH panel, subset H952. Total number of samples: 584. 22

Table 3: Percentage of amplified markers when the size length was lower than 140 bp or higher than 140 bp. All replicas were used to calculate percentages. 25

Table 4: Ancestry analysis from STRUCTURE (cluster membership proportions) for K = 5, and lowest likelihood ratios and classification obtained by Snipper. 28

Table 5: Relation between highest percentages of missing data obtained and the corresponding lowest likelihood ratios. 30

Table 6: STRUCTURE membership proportions for K = 2 to 5 clusters, for the individuals with more than one significant membership component. 36

Table S1: Cranial Measurements used in AncesTrees software. 61

Table S2: AIM-Indels used in the multiplex according to Pereira et al. (2012). 68

Table S3: DNA concentration (ng/µL) and respective number of markers amplified, missing data observed, and number of replicas made for each individual. 69

Table S4: STRUCTURE cluster membership proportions for K = 2, 3, and 4. 71

Table S5: Values of -log likelihood calculated by Snipper for each main cluster (lower is better). Probability of finding each profile, in each cluster, was calculated using the formulae EXP(-log(lk)). The Lowest Likelihood Ratio was calculated by dividing the higher probability by the second higher. 73

Table S6: Ancestry assessment through anthropology. AncesTrees software was used to calculate probabilities with the metric approach, while hefneR was used to calculate probabilities for the morphological approach. 75

LIST OF ABBREVIATIONS

AFR	Africa/ African
AIM	Ancestry Informative Marker
BC	Before Christ
bp	Base pairs
CE	Capillary Electrophoresis
CEI	*Coleção de Esqueletos Identificados*/ Collection of Identified Skeletons
DNA	Deoxyribonucleic Acid
EUR	Europe/ European
EAS	East Asia/ East Asian
FA	Forensic Anthropology
FG	Forensic Genetics
HGDP-CEPH	Human Genome Diversity Cell Line Panel
IBD	Identical by descent
Indel	Insertion Deletion Polymorphism
LCN	Low copy number
LR	Likelihood Ratio
NAM	Native American
OCE	Oceania/ Oceanian
PALOP	Portuguese-speaking African countries
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
SNP	Single Nucleotide Polymorphisms
STR	Short Tandem Repeats

INTRODUCTION

Human Genome

The genome of an individual has two different sources of information: one comes from the mother and another from the father. It is also known, that 99.7% of the human genome is the same for all individuals, while the differences make up for only 0.3% (Butler, 2010). A very important contribution to the differences in the human genome is recombination. Recombination or genetic reshuffling happens during meiosis when the genetic information from the father shuffles with the information from the mother, originating an individual with a combination of alleles completely different from the parents (Ayub, Haber & Tyler-Smith, 2016).

The ultimate source for genetic variation is mutation. A mutation is defined as any alteration that can occur in the DNA sequence, and can be of different types, mechanisms and even have different rates. From a change in a single nucleotide or an insertion, deletion, duplication, or inversion of segments of DNA, or expansions/ contractions in the number of tandemly repeated DNA motifs, translocation of DNA segments, to alterations in the number of chromosomes, is all considered mutation (Jobling et al., 2014). The average mutation rate in humans is considered low which means that the same mutation in two individuals is considered identical by descent (IBD) coming from the same ancestor where the mutation occurred (Ayub, Haber & Tyler-Smith, 2016).

Although DNA is inherited from the parents following a Mendelian pattern, through recombination, mutation, and some evolutionary forces, each genome is unique (with the exception of homozygotic twins).

Patterns of Human Variation

A population can be defined as a spatial-temporal group of interbreeding individuals who share a common gene pool. Unscrambling this definition, a population is a group of people that live in a specific region, inter-breeding together as their ancestors have been doing for generations. Nevertheless, it is not easy to physically define a population nowadays due to the lack of physical boundaries. Depending on the study, a boundary can be geographical, linguistic, or cultural, and it might not mirror the genetic structure (Phillips, 2015; Rosenberg et al., 2002; Serre & Pääbo, 2004; Stoneking, 2017). It is

crucial to be aware that different groups can be defined according to different levels, influencing the results of a study.

When studying populations or individuals it is important to understand where and how the diversity is shaped. As mentioned above, mutation is one of the biggest sources of variation since it adds new alleles to the populations, increasing the heterozygosity (diversity) within and between populations.

However, there are more evolutionary forces that mold population diversity such as migration that can also be called “gene flow”. When dealing with population genetics, the concern is not the individuals themselves but the gene pool (the compilation of alleles) of each population. Therefore, when an individual migrates from a population to another, he/she is in fact carrying its alleles to another population, changing the gene pool in both populations (Stoneking, 2017). In the current days with the development of technology, migration increased substantially because the so-called “boundaries” became easier to override. A specific mutation from one population will be carried to another population through migration, decreasing the differences in variation in both populations. A consequence of this phenomenon is admixture. Admixture happens when there is interbreeding between individuals from two different populations resulting in new haplotypes and variants (Ayub, Haber & Tyler-Smith, 2016). Admixture is ever-present throughout the world being the two most recognized examples of it the African American population, and the Latin population. In the first one, there was an admixture between African slaves and Europeans, and the latter between Native Americans and Europeans (with or without an African mix) (Ayub, Haber & Tyler-Smith, 2016).

Another event responsible for diversity in human populations is genetic drift which characterizes as random changes in the allele’s frequencies. This drift usually results in loss or fixation of a determined allele leading, over time, to loss of diversity (generally) (Stoneking, 2017). However, the amount of drift depends on factors like the effective population size and growth.

Natural selection leads to changes in alleles frequency depending on the fitness of an individual. The fitness of an individual is a collection of characteristics such as viability (probability of living to the reproductive age), fertility, mating success, and mortality (Stoneking, 2017). Therefore, fitness measures the capacity of an individual to survive and reproduce. Selection will try to increase the general fitness of the population, changing the alleles frequencies over time. This can be made by fixating the favourable allele, or losing the unfavourable allele, or maintaining the heterozygous genotype if this has the highest fitness.

Another phenomenon to consider is non-random mating that is typical in human populations. There are two types of non-random mating: assortative mating or inbreeding. Assortative mating can be positive or negative, and mainly means that there is a biased or preferential mating (Stoneking, 2017). Humans chose their mates according to their phenotype like skin colour, height, hair colour, etc. When mating happens more frequently between individuals that look similar to each other, then it is called positive assortative mating, which is the most common kind of mating in humans. However, when the opposite happens, in other words when the most common type of mating happens between people who have different phenotypic traits, then it is a negative assortative mating, which is rare in humans for the exception of the choice of gender (Stoneking, 2017). Inbreeding or consanguinity is the mating of individuals that are related to some extent leading to a decrease of diversity in the population.

All these evolutionary forces change the frequency of alleles all over the world, in different manners, leading to a constant state of evolution of the human genome. These changes are studied by a branch of Biology called Population Genetics.

Forensic Science

According to Houck & Siegel (2010), Forensic Science is described as “the science of associating people, places, and things involved in criminal activities; these scientific disciplines assist in investigating and adjudicating criminal and civil cases.” It is also depicted as a “mixed science” that is responsible for answering questions for the court by means of testimonies and reports (Houck & Siegel, 2010, page 4).

Different sciences contribute to the forensic field like medicine, pathology, criminology, genetics, anthropology, odontology, engineering, toxicology, chemistry, art, psychology, biology, entomology, among others.

I. A Brief History

Forensic science is a very recent concept, although some methods have been applied since very early. In 3000 BC, the Egyptians would remove and examine the internal organs of humans after their death, as a religious practice, which is thought to be the earliest example of the autopsy. However, the first recorded autopsy happened in 44 BC after the death of Julius Ceaser, when the physician Antistius revealed that, although he

was stabbed 23 times, the cause of death resulted from only one wound in the chest (Wilson & Wilson, 2003).

The first book published to help determine the cause of death was written in 1248 by Song Ci, in China, and it was entitled “Xi Yuan Lu” (“The Washing Away of Wrongs”). The book described how the author used knowledge in medicine and entomology to solve a crime and reported several methods that could be used to discern an accidental death from a murder.

Later, in the 16th century, two Italian surgeons started to study the modifications in the structure of the body after suffering a disease. This became the foundation of today’s pathology. A Swedish chemist developed a chemical capable of detecting arsenic in corpses in the 18th century, introducing toxicology to the field. Later, in the 19th century, Mathieu Orfilla determined the cause of death of an individual as poisoned by arsenic, getting a conviction for that murder (Houck & Siegel, 2010).

Anthropometry was developed by Bertillon and was used to record the physical features of individuals in a way that each record would be unique. It was highly used, especially in the United States, but got discontinued afterward (Houck & Siegel, 2010).

A major milestone in the history of forensic sciences was the fingerprint analysis (19th century). Edward Henry developed a system of classification of fingerprints, The Henry Classification System, which is still used nowadays in some countries (Houck & Siegel, 2010).

The 20th century was full of findings and developments in the field of forensics. At the beginning of the century, Karl Landsteiner discovered the ABO blood grouping system which was another milestone that allowed the use of blood to distinguish individuals (Goodwin et al., 2011). After the discovery of DNA by Watson and Crick, and the detection of highly polymorphic locus, Alec Jeffries developed the first DNA typing method designated DNA fingerprinting (Goodwin et al., 2011; Houck & Siegel, 2010). The development of PCR (Polymerase Chain Reaction) in 1983, allowed the method of DNA profiling to become easier and quicker, being used still nowadays (Goodwin et al., 2011).

Still in this century, the comparison microscope was developed by an American physician named Goddard, allowing the comparison of bullets and shell casings, hairs, fibers, and tool marks (Houck & Siegel, 2010). Methods to identify saliva, sweat, semen, and other body fluids from crime scenes and victims were also developed.

II. Forensic Genetics

This branch of Forensic Science is mainly responsible for the analysis of DNA. Forensic Genetics (FG) has revolutionized the forensic field with the so-called “DNA fingerprint”, as seen above. Therefore, a forensic geneticist is responsible for analysing the DNA whether it comes from a crime scene, human remains, or a paternity test (Goodwin et al., 2011). The forensic geneticist not only deals with human DNA, but he/she can also analyse DNA from plants, animals, or microorganisms. Anonymous (2007) defined FG as *“the application of genetics to human and non-human material (in the sense of a science with a purpose of studying inherited characteristics for the analysis of inter- and intra-specific variability in populations) to the resolution of legal conflicts”.*

The main purpose of FG is to analyse samples collected from a crime scene, generate genetic profiles, compare them with referenced suspects, and, therefrom, write a report that can be used in a court of justice (Goodwin et al., 2011). It can also be said that one of the main goals for FG is the identification either of the victim or the perpetrator. To achieve this investigation, it is mandatory to have a reference sample to compare the sample found in the crime scene. This reference sample can be either from a suspect or from a database (Amorim & Budowle, 2016). In most countries, those databases are only filled with genetic profiles of convicted people, which hardens the possibility of finding a match when the perpetrator has never been convicted of a crime. In these cases, FG can be useful to try to infer about phenotypic characteristics of the sample in question. Although this is a major help for a criminal investigation, it is not allowed in some countries due to ethic concerns. The phenotypic characteristics are biological descriptors that compose the biological profile of an individual. Through this profile, the police will try to find new investigative leads that will, hopefully, direct to the identification of the perpetrator.

a. DNA Recovery

One of the main obstacles when it comes to forensic genetics is the quality and quantity of the DNA found. Although a DNA profile can be obtained from a single cell depending on the methods used, forensic samples usually have higher levels of degradation, and thus more material is necessary to generate a good profile (Butler, 2015; Goodwin et al., 2011). DNA can be recovered from several different types of biological materials, such as blood, semen, hair, bone, epithelial cells, among others (Goodwin et al., 2011).

DNA degradation starts as soon as cells die. When oxygen stops circulating in the body, autolysis starts, and DNA begins to break down. Enzymes are released and when they come in contact with DNA, its strands start to modify and break. After that, exogenous factors start to act, like microbial activity, soil composition and pH, moisture levels, temperature, oxygen levels, and post-mortem time, among others, will contribute to DNA degradation (Gaudio et al., 2019; Latham & Miller, 2019; Misner et al., 2009).

In addition to all these exogenous factors, in the case of bone tissue, the bone type can have a very crucial role in this process. It is recommended to use dense cortical bones instead of dense spongy bone, due to microstructural differences between them (Misner et al., 2009; Prinz et al., 2007). Cortical bone has osteocytes protected by layers of hydroxyapatite and collagen, while in spongy bone, osteocytes are unprotected (White & Folkens, 2005). By being unprotected, spongy bone enables easier access of exogenous factors to DNA. Therefore, cortical bones like long bones, petrous bone, 1st distal phalange, and teeth have been used in the last years as the preferable bones for DNA extraction (Gaudio et al., 2019; Hasap et al., 2020; Prinz et al., 2007). Another contributor to the low yield of DNA concentration during extraction is PCR inhibitors that block the enzymatic reactions that occur during PCR (Latham & Miller, 2019).

When extracting DNA from a bone, it is not only the quantity of DNA present that is going to dictate the completeness of a profile, but also its purity and quality. It is counterproductive if the sample has a high quantity of DNA, but it is very degraded (low quality) and has PCR inhibitors (low purity). Although most kits available nowadays for DNA extraction are optimized and able to deal with these factors, it is still very difficult to predict if the DNA extraction will be able to provide a complete profile just by looking at the DNA quantity.

Accordingly, the choice of genetic markers used to generate a profile should depend on the conditions of the DNA recovered.

b. Genetic Markers

A polymorphism is defined as a change (or mutation) in the DNA sequence where the most common allele has a frequency of 99% or less in a population. (National Human Genome Research Institute). Genetic markers allow the detection of variability in different individuals by detecting different polymorphisms.

The DNA markers used in forensic genetics can be of two types: recombining (autosomal or X-Chromosome) or non-recombining markers (Y-chromosome or mitochondrial DNA). The markers used for identification should be recombining markers since the recombining information is unique for each individual (except in the case of an identical twin) (Butler, 2010).

The markers studied in the forensic field are usually found in the non-coding region of the human genome, thus there is no phenotypic or disease-related information revealed. The markers should also be highly polymorphic so that it is possible to discriminate between two individuals.

There are several different types of genetic markers used in forensics, being the most common the microsatellites or Short Tandem Repeats (STRs). Nevertheless, there are some other types of polymorphisms also used like single nucleotides (SNPs), insertion-deletion (Indels), polymorphisms in the mitochondrial genome, and Alu repetitive elements (Liu & Harbison, 2018). Each type of polymorphism has different characteristics and can give different information about the same individual.

It should be taken into consideration that to provide a forensic report with the probability of a sample belonging (or not) to a specific individual (likelihood ratio), the allele's frequency in that population should be known.

i. Short Tandem Repeats

Short Tandem Repeats (STR) or microsatellites are polymorphism of length that consists of variable repetitions of a core unit. The core unit is a sequence of 2 to 7 base pairs (bp), and the allele size can go from 100bp to 400bp. The number of repetitions becomes the name of the allele as demonstrated in Figure 1 (Goodwin et al., 2011). The core unit is composed of the sequence "TCTA". Allele 8 corresponds to eight repetitions of the core unit, and allele 10 corresponds to 10 repetitions.

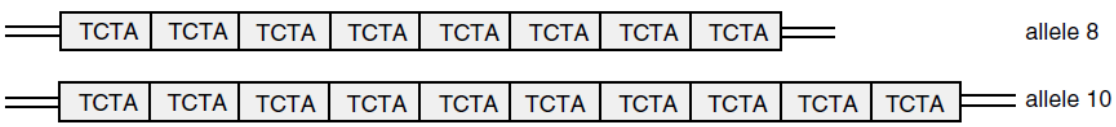


Figure 1: an example of two alleles in an STR' locus. From Goodwin (2011).

The STRs have a mutation rate of 10^{-3} per generation (Brinkmann *et al.,* 1998) which is considered high, making these markers highly polymorphic and ideal for identification (Jobling et al., 2014). This high mutation rate is due to the slippage that occurs during DNA replication (Stoneking, 2017). Generally, a new mutation happens when the DNA polymerase replicates/adds one more (or less) core unit during the replication process. This slippage is also responsible for the appearance of 'stutter bands' during PCR that can complicate the STRs' analysis. When the DNA has low quality or is degraded, these 'stutter bands' are harder to identify (V. Pereira & Gusmão, 2016). The mutation rate is, usually, dependent on the number of repeats; the more repeats, the higher the probability of occurring a mutation (Stoneking, 2017). Notwithstanding, the complexity of the core unit can also influence the mutation rate.

STRs are spread through the human genome and account for 3% of the genome (Butler, 2010), so it is possible to find thousands of STRs that are not under any selective pressure and can be used in forensic genetics (Goodwin et al., 2011).

STRs are accepted and validated as forensic markers because they satisfy all the requirements needed (Amorim & Pereira, 2005; Goodwin et al., 2011). They are highly discriminatory, reproducible, relatively cheap, easy to amplify, robust, codominant, liable to multiplex, and sensitive (Goodwin et al., 2011). STRs can be easily genotyped through a PCR reaction followed by a fragment analysis using capillary electrophoresis.

Despite all this, when it comes to degraded samples, STRs are not the foremost markers to use because of their large size length.

ii. Single Nucleotide Polymorphisms

Single nucleotide substitutions in the DNA sequence are called Single Nucleotide Polymorphisms or SNPs. These are the simplest polymorphisms that exist in the human genome (Goodwin et al., 2011). They are also the most abundant type of polymorphism in the genome accounting for 85% of the human variation, but they are not evenly distributed through it, meaning that some chromosomes can have a higher average of SNPs than others (Goodwin et al., 2011; V. Pereira & Gusmão, 2016).

Just like in STRs, mutations can occur during DNA replication, but in this case, it happens in only one single base (Goodwin et al., 2011; Jobling et al., 2014). However, because of the double-check mechanisms during replication, this misincorporation of a base is

considerably rare, being the mutation rate of an SNP of 10^{-8} (Jobling et al., 2014; Nachman & Crowell, 2000; V. Pereira & Gusmão, 2016). This low rate is the reason why SNPs are considered stable through time, and so, to have two mutations in the same position is very unlikely. Therefore, if two sequences have the same mutation in the same position, it is probable that those mutations are identical by descendent and not by state (Jobling et al., 2014).

SNPs are called biallelic polymorphisms because, generally, they can only have two possible alleles in one position (Butler, 2010), which also means that they are less informative than STRs that can have several alleles. Thus, it is necessary a much larger number of SNPs to have the same amount of information given by STRs, to get an acceptable power of discrimination to determine a unique profile (Butler, 2012).

When dealing with degraded samples, SNPs are very useful because the amplicons analysed can be smaller than 100 bp which means it is possible to recover more information than using STRs (Butler, 2012).

Another advantage of using SNPs is that there are no stutter artifacts associated with the alleles leading to an easier and simpler interpretation of the results (Butler, 2012). The detection of SNPs requires a more complex methodology than the one used for STRs. One of the most common methods is the mini-sequencing reaction which comprises a PCR reaction plus a sequencing reaction (Goodwin et al., 2011). It can also be multiplexed using different size primers and different fluorescent tags for each base.

There are four different applications for SNPs, according to Butler (2012): individual identification, lineage-informative through mitochondrial DNA and Y-chromosome, ancestry prediction, and phenotype prediction through specific physical traits. This last application has been growing in the last years because it enables the prediction of characteristics like eye colour, skin pigmentation, or even hair colour through SNPs' analysis (Maroñas et al., 2014; C. Phillips et al., 2007; Walsh et al., 2013; Walsh et al., 2017).

iii. Insertion/ deletion polymorphisms

Another length polymorphism can be an insertion or deletion of one or more nucleotides and is commonly called Indel (Insertion/Deletion) (Weber et al., 2002). It is another bi-allelic polymorphism and is the second most abundant type of polymorphism found in the human genome (explaining 16% to 20% of the genetic variation) (Mills et al., 2006).

Indels are spread throughout the human genome, although some hotspots of variation have been identified (Mills et al., 2006). Most Indels have a length of 2 to 16 bp (Mullaney et al., 2010) and the two alleles are generally classified as 'short' or 'long' (Butler, 2012). The proportion of insertions and deletions are found to be almost equal, making it 47% of insertions and 53% of deletions when compared to the chimpanzee genome (Mills et al., 2006). Most Indels are found within genes, influencing inherited traits, and even diseases, being its' study very important when dealing with personalized medicine (Mills et al., 2011; Mullaney et al., 2010).

Mills et al. (2006) divided Indels into five major classes being one of them 'Indels containing random DNA sequences. This class represents 40.8% of all Indels and has a wide range of sequence lengths going from 1bp to 10kb. More than 99% of this class of Indels have a length of less than 100bp. Curiously, these Indels are called 'small indels' (Mullaney et al., 2010).

Small indels have a low mutation rate (approximately 10^{-8}) therefore the polymorphism originated from a single mutation event hence being considered stable, and its usage as forensic markers has been increasing throughout the last decade (Pereira et al., 2009; V. Pereira & Gusmão, 2016). One of the great advantages of Indels is the fact that the amplicons are much smaller than the ones produced by STRs which is more successful in cases where the samples are degraded (Pereira et al., 2009). On the other hand, by being a bi-allelic marker, the power of discrimination of Indels is smaller than the one for STRs, requiring the use of more markers. Comparing to SNPs, the genotyping of Indels is easier and less expensive, using a single PCR reaction followed by capillary electrophoresis fragment-analysis (PCR-to-CE assays), and Indels are liable to be used in high throughput genotyping (Pereira et al., 2009b).

The interest of forensics in this polymorphism has been growing due to the advantages they bring, as mentioned above, and especially the fact that Indels can be used in cases of degraded DNA customarily found in forensic cases.

III. Forensic Anthropology

One of the earliest definitions of Forensic Anthropology (FA) was given by Stewart in 1979 who stated that it was a 'branch of physical anthropology, which for forensic purposes, deals with the identification of more-or-less skeletonised remains known to be or suspected of being human' (Iscan & Steyn, 2013). Although this was once a good definition, it does not cover all the work done by a forensic anthropologist nowadays.

Cunha and Cattaneo (2006) assembled several roles of a forensic anthropologist that include:

- the recovery of the remains from the crime scene;
- determine if the remains are of human origin or not – if they are not there is no case;
- reconstruct the biological profile of unidentified remains;
- perform facial identification through craniofacial reconstruction;
- reach personal identification using distinctive features or pathologies present in the bones;
- assist in determining the cause and manner of death when there are traumatic bone injuries;
- identifying victims of mass disaster (Disaster Victim Identification – DVI);
- assist in human rights investigations dealing with mass graves and war crimes;
- identifying living individuals through morphological and metric characteristics either in cases of undocumented people (as it happens so often with the refugees), or forensic cases where there are only images of the suspect.

The biological profile determined by a forensic anthropologist tries to answer questions about age at death, sex, ancestry, and stature of the remains found (Cattaneo, 2007). To do this, the forensic anthropologist relies mostly on methods of observation and measurements to perform the 'osteobiography' of an individual. Unfortunately, these methods depend deeply on the level of preservation of the remains, and sometimes on the level of expertise of the forensic anthropologist.

FA has an interdisciplinary nature being internally connected with taphonomy, archaeology, osteology, physical anthropology, pathology, and odontology, among others.

Ancestry

During a criminal investigation, sometimes there is no suspicion about the perpetrator of the crime. Often, eyewitnesses try to describe the suspect guessing it's sex, age, height, and ancestry – the biological profile of an individual. However, eyewitnesses are proven unreliable and can be influenced by the circumstances of the situation (Spinney, 2008). When there is some source of biological sample left at the scene it can be useful to try

to infer about the suspects' biological profile to try to provide a lead to help direct the investigation (Phillips, 2015; Sun et al., 2017).

Nonetheless, the victims' biological profile can also be required. When there is no suspicion of who the victim is, or its remains are too disfigured to provide positive identification, the inference of the biological descriptors described above, can be helpful to achieve an identification.

One of the parameters that compose the biological profile is ancestry. Ancestry can be defined as the biogeographical origin of an individual. In other words, ancestry is the genetic heritage transmitted by the recent ancestors of an individual that has occupied a specific region of origin (Phillips, 2015). Recently, population affinity or biogeographic origin are being sued instead of ancestry.

According to Phillips (2015), the inference of ancestry can be useful not only to gain information about the unknown suspect but also to corroborate eyewitnesses testimonies, to confirm self-declared ancestry when dealing with databases, to assist when trying to infer about physical characteristics of admixed individuals, to improve studies in other fields like medicine and archaeology, among other applications.

In practical terms, the assessment of biogeographical origin will provide indirect information about the phenotypic characteristics of an individual. In other words, if the individual X has a high probability of being European, then it is assumed that he will probably have light skin which would exclude suspects with darker skin. Nevertheless, in cases of admixture, this information can be completely fallacious (Jobling et al., 2014). Another hurdle when assessing ancestry is that the appearance of someone is not only influenced by genetics but also by the environment and development that the person had.

There is also the problem of 'self-identity'. For a lot of centuries, the term 'race' was used to assort people. Nowadays, this terminology is still employed (mainly) outside the scientific field as a social, biological, and cultural descriptor. However, when dealing with forensics, ancestry assessment only reveals a biological descriptor. Therefore, there is a need to be cautious when dealing with eyewitnesses and biological evidence.

For all these reasons, the assessment of ancestry is one of the hardest estimations to make in the field of forensics. The prediction of ancestry can be made through genetics and/or anthropology.

I. Through Genetics

Although the first studies about genetic variation in humans showed that most differences were found between individuals and not between populations (Lewontin, 1972), the concept of "race" was promulgated through later studies that stated that variation happened in clades represented as continents (Bowcock & Cavalli-Sforza, 1991; Cavalli-Sforza et al., 1988; Takezaki & Nei, 1996). Serre and Pääbo (2004) determined that when the sampling strategy is made homogeneously throughout the world, the results of the genetic variation confirmed that the alleles' frequency was distributed in a continuous pattern instead of in discrete clusters, and that, therefore, there was no genetic evidence that supported the idea of "races". Although alleles' frequency is distributed smoothly throughout the world, there are some differences in frequencies which is thought to be due to adaptation to a specific region throughout the years, and evolutionary forces as mentioned above. These differences are, therefore, responsible for the possible assignment of individuals to geographic regions (Rosenberg et al., 2005). To infer the biogeographical origin and estimate admixture proportions, it is most advantageous to employ markers with the largest allele-frequency differences among populations (Shriver et al., 1997). Nowadays, these markers are called Ancestry Informative Markers – AIMs.

a. AIMs

AIMs can be defined as polymorphisms that exhibit large differences in the alleles' frequency between different populations, and therefore, detect population stratification (Enoch et al., 2006). In the last years, the usage of AIMs has increased, and they have been applied in studies of genetic association, population genetics, and to aid criminal investigations (Halder et a., 2008; Jia et al., 2014; Kidd et al., 2011; Nassir et al., 2009; Pereira et al., 2012; Phillips et al., 2012; Phillips et al., 2014; Romanini et al., 2015; C. Santos et al., 2015; N. P. C. Santos et al., 2010; Tian et al., 2008; Yang et al., 2005).

Lineage markers as mitochondrial DNA and Y chromosome can be used to give information about ancestry, but only about the two lineages (maternal and paternal) separately since they are haploid and non-recombing markers. Thus, ancestry can be more thoroughly estimated when a selection of recombining autosomal markers is used.

Although most forensic genetic studies focus on STRs, it is not the preferable choice when it comes to ancestry estimation since they have high mutation rates, and not

enough differences in alleles' frequency (Kidd, 2016). Therefore, AIM-SNPs has been in the spotlight when it comes to ancestry prediction due to their efficiency (V. Pereira et al., 2017; Phillips et al., 2007; Phillips et al., 2012; Phillips et al., 2014; Sun et al., 2017; Valle-Silva et al., 2019).

More recently, Indels have been under the scope of geneticists because of the combined characteristics of both STRs and SNPs namely, the easy genotyping strategy and the option of genotyping small fragments (Pereira et al., 2012).

II. Through Anthropology

Prediction of population affinity in anthropology not only helps an on-going investigation but it also reveals the correct methods to use to infer about the other biological parameters - sex, age at death, and stature (Cunha & Ubelaker, 2019).

The most informative bone to infer about ancestry is, without a doubt, the skull (Albanese & Saunders, 2006; Cunha & Ubelaker, 2019; Hefner, 2009; L'Abbé, Van Rooyen, Nawrocki, & Becker, 2011), and the prediction can be made using two different approaches: metric and non-metric. The methods employed in this prediction are based on one single assumption – that cranial diversity exists, and it is significant (Cunha & Ubelaker, 2019; Iscan & Steyn, 2013). However, an obstacle is found when dealing with admixture which increases the difficulty of the prediction. Just like geneticists, anthropologists have been dealing with the concept of race and its misuse in forensics (Cunha & Ubelaker, 2019).

Cranial traits are phenotypic traits which means that they are influenced by genetics but also by the environment. It is important to be careful when assessing ancestry using skeletal remains because there is not a trait that is specific from one single population, making the evaluation of ancestry a very controversial issue in forensic anthropology (Iscan & Steyn, 2013). Usually, anthropology works hand-in-hand with other areas such as genetics, and when the anthropological method raises queries about the biogeographical origin, genetics is called to step-in and a bone sample is given to be analysed. In terms of anthropology, the best approach is to combine both metric and non-metric assessments.

a. Metric Assessment

The metric approach is based on the analysis of measurements of craniometric points. This approach is the most traditional one when it comes to predicting ancestry since craniometry has been employed in forensic and biological anthropology since the last century (Iscan & Steyn, 2013). Yet, it depends on how the cranial measurements are treated in terms of statistics and mathematics.

Different tools have been developed in the last years to help estimate ancestry with statistical methods. One of the most notorious tools is FORDISC, an American software that uses linear discriminant analysis to estimate ancestry, with a maximum of 34 cranial and 39 postcranial measurements (Ousley & Jantz, 2012). The database used in FORDISC contains mostly North American data which is why it is mostly used there.

AncesTrees, developed by the Laboratory of Forensic Anthropology in Portugal, applies a machine learning ensemble algorithm, random forest, with 23 cranial measurements to estimate ancestry (Navega et al., 2014). Its database has more than 3 000 individuals and is representative of six major ancestral groups (Navega et al., 2014; Cunha & Ubelaker, 2019). AncesTrees can be assessed through the Osteomics platform.

These two tools give a prediction of ancestry providing probabilities of an individual belonging to a main ancestral group. Nevertheless, it is noteworthy that the results are as reliable as the database used – if a determined region is not represented in the database, the programs will assign the individual to the closest group possible which might not be the correct one.

Although the skull is the preferable bone to estimate ancestry, some postcranial bones might also be informative. Liebenberg et al., (2018), and Spradley (2016) had good results when combining the methods of different postcranial bones to estimate ancestry, which brings an advantage when the skull is not available for analysis.

b. Non-metric assessment

The non-metric approach is based on two types of variants: morphoscopic and discrete traits. The first ones are related to differences in shape (like sutures or palate shape) while the discrete traits are variants that can be present or not (like the japonicum or the Inca bone) (Iscan & Steyn, 2013; Cunha & Ubelaker, 2019). Hefner (2009) described 11 common traits useful to infer ancestry that are located mostly in the mid-facial region and

presented it using drawings and illustrations. One of the biggest issues when estimating ancestry through non-metric traits was the lack of standardized guides and methodological approaches. Hefner (2009) changed that by giving definitions, and line drawings to each trait. The software 'hefneR' was developed to calculate the probability of ancestry using the 11 morphoscopic characteristics proposed by Hefner (2009) (D'Oliveira Coelho et al., 2020).

Ancestry can also be assessed through dental morphology. Some characteristics can be population-specific like the Carabelli's cusps, incisors' shoveling, or the presence of the 5^{th} cusp in the first upper molar (Scott et al., 2018). 'rAsudas: a new web-based application for estimating ancestry from morphology' is a software developed by Scott and collaborators in 2018, to infer ancestry through teeth morphology. However, the reference sample is mainly archaeological which hinders the application of the results in a forensic context. Nevertheless, dental morphology can be used to confirm the results obtained with other methods.

Both hefneR and rAsudas use the naïve Bayes classifier for ancestry estimation and are available in the platform Osteomics.

AIMS

The inference of ancestry is one of the hardest tasks to achieve in the field of forensics. It can help direct investigations when there is no other lead about the victim and/ or perpetrator. Hence, it is important to understand what information is given when ancestry is assessed. This assessment can be performed through different sciences and methods.

In this study, a collection of 56 skeletonized human remains was used to evaluate the inference of biogeographical origin through genetic methods in such challenging samples. In order to pursue that main goal, intermediate objectives were defined:

- To perform the DNA extraction from bone successfully.
- To genotype 46 AIM-Indels through a multiplex PCR reaction followed by capillary electrophoresis.
- To infer about the probable biogeographical origin of each individual, comparing each one with reference samples from the five major continental groups – Africa, Europe, East Asia, Native America, and Oceania.
- To better understand the existence of admixed individuals and how to identify them.
- To compare the inferred genetic ancestry with the scarce *ante-mortem* records available.
- To compare the inferred genetic ancestry with ancestry data obtained through anthropological methods.

MATERIALS AND METHODS

Sample selection and collection

The 21st Century Identified Skeletal Collection is housed in the Laboratory of Forensic Anthropology, Department of Life Sciences at the University of Coimbra. It consists of 302 adult skeletons, with ages at death ranging from 25 to 101 years, which comprises Portuguese nationals who died between 1982 and 2012 (Ferreira et al., 2020). These remains are collected from the Cemetery of Capuchos in Santarém, Portugal, due to a collaboration between the Department of Life Sciences of the University of Coimbra and the City Council of Santarém (Ferreira et al., 2014). This collection comprises skeletons that were not claimed by their relatives after the required legal period.

Fifty-six skeletons were selected for analysis. This selection was made according to the possibility of having the anthropological assessment of ancestry so that a comparison would be possible. Unfortunately, *ante-mortem* records containing information on the death certificates were very scarce, and it was only possible to retrieve information about the place of birth of 5 of the fifty-six remains studied (Table 1).

Table 1: Information about the place of birth (nationality) that was possible to retrieve from records.

CEI_ID	Nationality
CEI_3	Mozambique
CEI_12	Mozambique
CEI_14	Angola
CEI_42	Brazil
CEI_62	Portugal

Sample Preparation

The bone used for the DNA extraction was the femur according to the DNA Commission of the International Society for Forensic Genetics (ISFG) (Prinz et al., 2007).

The preparation of the bone sample was always performed in laminar flow cabinets previously decontaminated using UV light to prevent contaminations. A square of 2 cm^2 of bone was removed from the diaphysis of the femur using a window cut without separating the shaft (Figure 2), cleaned by being immersed in bleach (5%) for 10 minutes, then rinsed in distilled water, and dried with paper and DNA Away. After that,

the sample was drilled to remove the outer layers of bone. The sample was cut into small pieces and put in vials, to be grounded using the SPEX CertiPrep - 6750 Freezer/Mill, with liquid nitrogen. After the grinding, the bone powder was collected in a sterile tube, and stored in a freezer at -20ºC until further processing.



Figure 2: Example of a window cut in the femur of the individual CEI/XXI_111[1].

DNA extraction

DNA extraction was performed using the PrepFiler Express BTA™ Forensic DNA Extraction Kit (Applied Biosystems). Since these samples were very degraded and difficult to extract, the protocol was modified according to Applied Biosystems' recommendations, and the triple of the quantities was used.

Therefore, approximately 150mg to 200mg of bone powder was transferred to a sterile PrepFiler™ Bone Lysate Tube and was incubated with 660µL of PrepFiler BTA™ Lysis Buffer, along with 9µL of 1 M DTT (Dithiothreitol) and 21µL of Proteinase K, in a thermal shaker at 56ºC and 1100 rpm overnight.

After the incubation, the tube was centrifuged for 120 seconds at 9900 rpm, and the clear lysate with a volume of 500 µL was transferred to a new PrepFiler™ Sample Tube. Automated DNA extraction was run using the AutoMate Express™ Instrument. After the extraction, the final volume of elution was 20 µL.

[1] For practical reasons, CEI/XXI_ID will from now on be mentioned as CEI_ID.

DNA quantification

The quantification was performed using the Quantifiler™ Human DNA Quantification Kit (Applied Biosystems). Eight quantification standards with different concentrations varying from 50 ng/µL to 0.023 ng/µL were used with one duplicate each, according to the recommendations of the producer. Quantifiler™ PCR Reaction Mix and Quantifiler™ Human Primer Mix were added making up a total of 23 µL per sample (12.5 µL and 10.5 µL respectively). The PCR mix was distributed through the wells of a 96-well reaction plate. Thereafter, 2 µL of extracted DNA samples, standard, or positive and negative control were added. The wells of the plate were set up according to the recommendations of the producer. The plate was sealed with an Optical Adhesive Cover and centrifuged at 3000 rpm for 20 seconds. The reactions were run in a 7500 Real-Time PCR System from Applied Biosystems.

Sampling, extraction, and quantification of the DNA were performed while using face masks and disposable gowns. All reagents, consumables, and workstations were sterilized before and after the procedures, either using UV light or an autoclave. These procedures were performed at Instituto Nacional de Medicina Legal e Ciências Forenses, in Coimbra, Portugal.

AIM-Indel Genotyping

Indel genotyping of the 46 loci was made in a single PCR multiplex reaction following the protocol of Pereira et al. (2012). A minimum of two amplifications was performed for each DNA extract.

For the PCR reaction, 12.5 µL of Qiagen Multiplex PCR kit at 1 x Qiagen multiplex PCR master mix was added to 3.75 µL of primer Mix (at 0.15 µM) and 5 µL of DNA extract. A reinforcement of six primers was added at the concentration of 0.15 µM. Primer details can be reviewed in Table S2.

Amplifications were performed in a thermocycler with the following conditions: an initial denaturation at 95ºC for 15 minutes, followed by 32 cycles comprising three steps (94ºC for 30 seconds, 60ºC for 90 seconds, 72ºC for 45 seconds), and a final extension at 72ºC for 60 minutes.

A purification step was added, using Sephadex™ G-50 DNA Grade Agarose Gel (Thermo Fisher Scientific). A volume of 750 µL of Sephadex was added to columns with 2 mL collection tubes. This complex was centrifuged for 4 minutes at 4400 rpm. After that, the columns were transferred to new collection tubes of 1.5 mL, and the liquid left

in the 2 mL tubes was discarded. The total volume of the PCR reaction was then added (25 µL) in the centre of the Sephadex column. Another centrifugation was made for 4 minutes at 4400 rpm, after which the columns were removed, and the purified product remained in the collection tube.

Preparation for capillary electrophoresis (CE) consisted of adding 10 µL Hi-Di™ Formamide (Applied Biosystems) with 0.3 µL of GeneScan™ 500 LIZ size standard (Applied Biosystems), to 1.7 µL of purified product. Capillary electrophoresis was performed in a 3500 Genetic Analyzer prepared with DS-33 matrix standard, POP-7™ polymer, and applying virtual filter G5 (Applied Biosystems). Electropherograms were analysed using GeneMapper 5.0 Software (Applied Biosystems).

Indel short alleles were coded as 1 while long alleles were coded as 2, for convenience. All reactions included a negative control, and a positive control (reference sample 9947). There was no evidence of contamination in the negative controls.

Two criteria were followed when analysing the electropherograms. First, an allele was only reported if it was called in more than one amplification of the same extract. Second, only compatible genotypes were reported. This means, that if in three amplifications of the same extract different genotypes were called, then the reported genotype was null.

Statistical Analysis

The reference samples used in this study were from the HGDP-CEPH diversity panel standardized subset H952 (Cann et al., 2002) belonging to the 5 major population groups (Table 2), plus extra samples from Angola (48), Portugal (48), Taiwan (48), and Brazil (48) used in Pereira et al. (2012).

Ancestry inferences were performed using STRUCTURE V2.3.4 (Pritchard et al., 2000) that uses a Bayesian clustering approach to assign individuals to populations according to their genotypes while estimating population allele frequencies. The runs were performed using the option “Admixture model” and “Correlated allele frequency” with a burnin length of 100 000 followed by 100 000 MCMC repetitions. The option “Use Population Information” was selected to assist clustering, and 3 independent runs were performed for K = 2 to K = 5, following the order of the 5 major populations – Africa, Europe, East Asia, Native America, and Oceania. The allele frequencies were updated using only the reference individuals with POPFLAG = 1 data. Through software CLUMPAK (Kopelman et al., 2015) the graphical bar plots were obtained using the total runs for each K.

Through Snipper App Suite version 2.5 (http://mathgene.usc.es/snipper/), likelihood ratio values were obtained using a naïve Bayesian classification, and Principal Component Analysis (PCA) was achieved, using the same reference genotypes.

Indel genotyping and statistical analysis were performed at i3S – Instituto de Investigação e Inovação em Saúde da Universidade do Porto.

Table 2: Reference samples used in this study that belong to the HGDP-CEPH panel, subset H952. Total number of samples: 584.

Major Group	**Population**	***n***	**Reference**
Africa	South Africa	2	HGDP-CEPH subset H952
	Lesotho	1	
	Botswana	4	
	Angola	1	
	Kenya	11	
	Senegal	22	
	Nigeria	22	
	Central Africa Republic	23	
	Congo	13	
	Namibia	6	
Europe	Orkney Islands	15	
	Russia-Caucasus	17	
	Russia	25	
	France	52	
	Italy	49	
Asia	China	165	
	Cambodia	10	
	Japan	54	
America	Colombia	7	
	Brazil	22	
	Mexico	35	
Oceania	Bougainville	11	
	New Guinea	17	

Anthropological data

The anthropological data was obtained by Ph.D. student Catarina Coelho, at the Laboratory of Forensic Anthropology, University of Coimbra. Metric and non-metric analyses of the skull were performed, and statistical probabilities of ancestry were inferred through hefneR and AncesTrees software (d'Oliveira Coelho et al., 2020; Navega et al., 2014). The cranial measurements used in AncesTrees are described in Table S1, while the traits evaluated by hefneR, and their illustrations are described in Figure S1.a-k).

RESULTS AND DISCUSSION

DNA extraction and genotyping performance

In this study, DNA extraction was possible in all 56 bone samples. The lowest DNA concentration was 0.0003 ng/µL (CEI_290), and the highest was 0.1791 ng/ µL (CEI_247) (Table S3). From 46 markers in total, the lowest number of amplified markers was 15, while in seven samples it was possible to amplify all markers. CEI_290 had the lowest DNA concentration (0.0003ng/µL) however had a complete profile (0% of missing data). Quantification was made twice for this sample, and one of the results was 'undetermined'. This might have happened due to the levels of degradation present in this sample that somehow hindered the PCR during quantification, but not the one made for AIM-Indel genotyping.

The kit of AIM-Indels used in this study gathered two great advantages of this type of genetic markers: the easy and less expensive technique combined with the amplification of smaller amplicons. Total missing data was 18.75%, meaning that the success rate of these AIM-Indels in these samples was approximately 81% which was a very good result considering the DNA concentrations obtained. The arithmetic mode demonstrated that the most frequent number of markers amplified was 44, while the arithmetic mean was 37.4 markers amplified. It is also important to highlight that, except for seven samples, all the others had more than 50% of the genetic profile obtained.

The correlation between DNA concentration and the number of markers amplified is represented in Figure 4. It shows that there is not a straightforward connection between the DNA concentration and the number of reportable markers, other than the fact that there is a slight increase in genotyping success as the DNA concentration increases. This leads to the possibility that some of these samples had very high levels of degradation, and/ or PCR inhibitors, randomly.

It is noteworthy that amplicons with a size range lower than 140 bp were more successfully amplified (93.6%), while only 68.9% of amplicons with size ranges bigger than 140 bp were amplified (Table 3) which seems to reflect the phenomenon of DNA degradation in bigger size lengths. The number of replicas obtained from each bone sample is also described in Table S3, and it did not correlate with the number of amplified markers, which was noticeable when some samples had 0% of missing data with only 2 replicas (for instance, CEI_275), while others had 67.4% of missing data with 5 replicas (CEI_3).

Table 3: Percentage of amplified markers when the size length was lower than 140 bp or higher than 140 bp. All replicas were used to calculate percentages.

% of Amplified Markers	
< 140 bp	> 140 bp
93.6%	68.9%

Two variants beyond the expected short and long alleles were described by Pereira et al., (2012) within the amplicons of markers MID360 and MID2264 and were coded as allele 3 for convenience. MID360 variant allele was reported to be found only in African samples, while MID2264 variants were found in European samples (Pereira et al., 2012). In this study, allele 3 for MID360 was reported in CEI_209, while allele 3 for MID2264 was found in CEI_100 and CEI_262 (Figure 3).

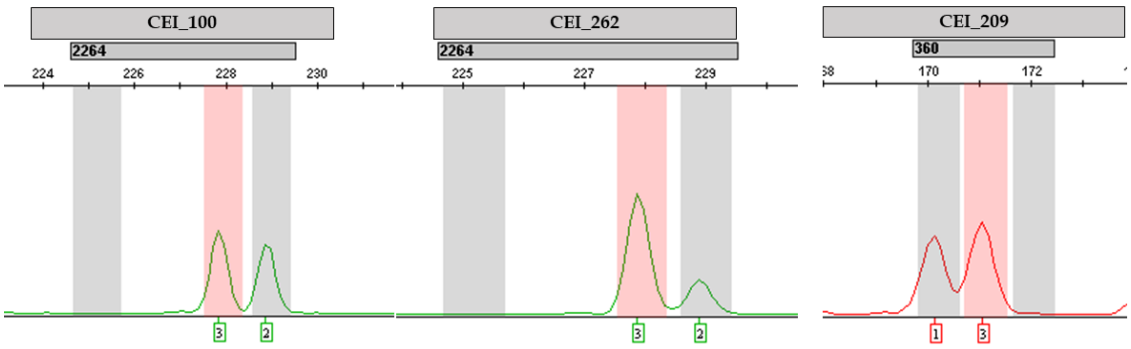


Figure 3: Representation of the variants (allele 3) of the markers MID2264 and MID360 found in samples CEI_100, CEI_262, and CEI_209 (markers are identified by MID number).

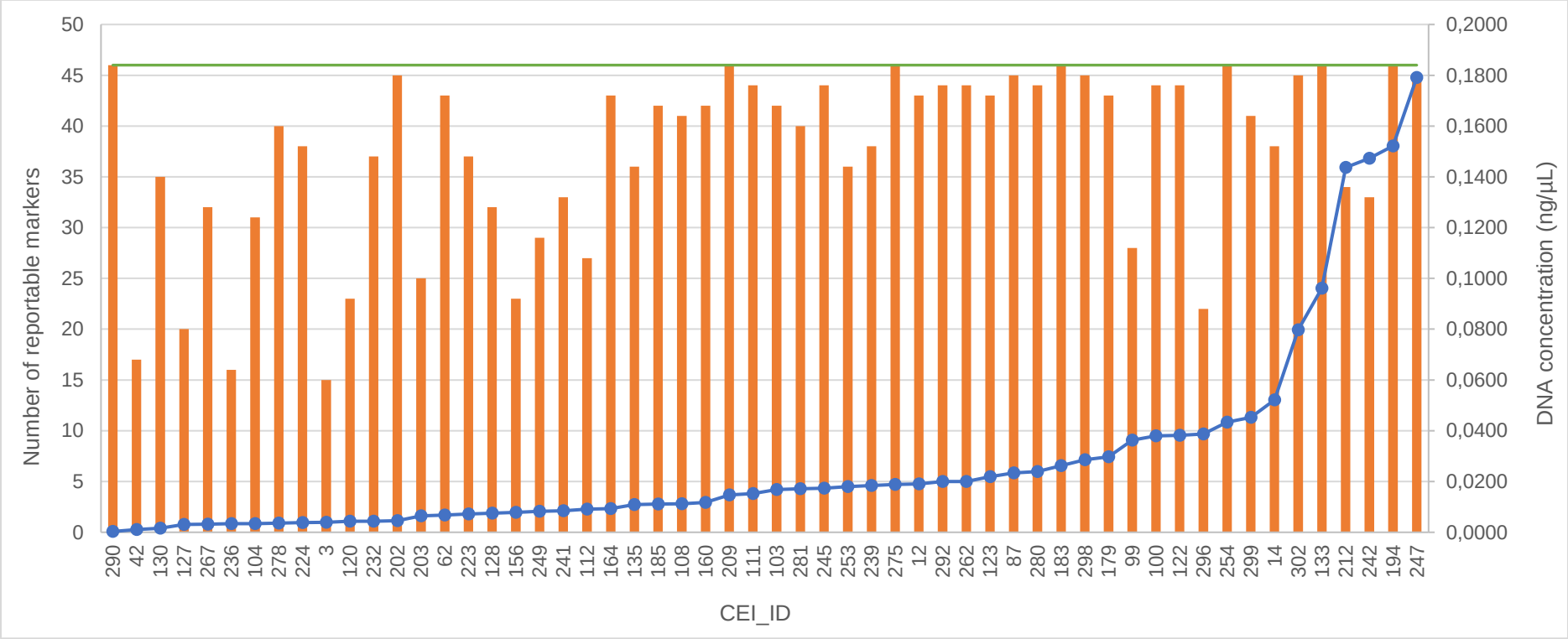


Figure 4: Correlation between the number of reportable AIM-Indel loci and DNA concentration in sample extracts. Orange bars represent the number of reportable markers; blue dots represent DNA quantification values; green line represents the maximum number of reportable markers.

Inference of Genetic Ancestry

In accordance with prior information about the provenience of the skeletons of the 21st Century Identified Skeletal Collection, it was expected that European membership proportion would be the highest and the most likely one. Indeed, the European component is present with more than 10% of membership in all but two individuals (CEI_179 and CEI_281). Below this value, membership would be considered non-significant according to Phillips (2015). Only three individuals were classified as of African origin (CEI_14, CEI_179, and CEI_281), while all others (53) were classified as European when weighing genetic evidence through Likelihood Ratios (LR).

Individual ancestry estimates computed by STRUCTURE are present in Table 4, alongside the lowest LR values and population affiliation obtained by Snipper for each profile. STRUCTURE calculated an optimal K at 5, as expected, with an estimated probability of the data [(LnP(D)] of -33753.2 (the lowest of all runs), and a mean value of *alpha* (α) of 0.053, consistent with the fact that most samples in the CEI collection had their origin mainly in one population.

The population assignments made with Snipper generally match the ancestral membership proportions estimated by STRUCTURE (except for CEI_296). Nevertheless, in cases of evident admixture in STRUCTURE ancestry estimates, Snipper usually assigned ancestry to the main contributor with markedly lower LR values than those obtained when studying individuals with essentially one single ancestral origin. It is also notable that when missing data values are considerably high (for instance more than 40%), probabilities given by Snipper are also markedly lower.

Table 5 summarizes the individual profiles with the highest levels of missing data (>39%) and the corresponding lowest LR values obtained. All these results should be carefully analysed and interpreted due to the decreased level of information that was achieved for these individuals.

Table 4: Ancestry analysis from STRUCTURE (cluster membership proportions) for K = 5, and lowest likelihood ratios and classification obtained by Snipper.

CEI_ID	STRUCTURE cluster proportions					Lowest Likelihood Ratio	Classified as
	AFR	EUR	EAS	NAM	OCE		
CEI_3	0.014	0.688	0.121	0.151	0.026	23.19 times more likely European than East Asian	European
CEI_12	0.02	0.894	0.005	0.077	0.004	3,874,353,786,295 times more likely European than American	European
CEI_14	0.639	0.306	0.016	0.026	0.013	3,215 times more likely African than European	African
CEI_42	0.183	0.639	0.03	0.071	0.076	275 times more likely European than Oceanian	European
CEI_62	0.014	0.939	0.032	0.01	0.005	32,269,875,986 times more likely European than East Asian	European
CEI_87	0.027	0.94	0.011	0.007	0.014	13,224,602,017,164 times more likely European than Oceanian	European
CEI_99	0.007	0.911	0.025	0.047	0.01	2,735 times more likely European than American	European
CEI_100	0.095	0.881	0.01	0.006	0.008	1,090,982,737,484 times more likely European than African	European
CEI_103	0.005	0.948	0.01	0.026	0.012	26,691,206,406 times more likely European than American	European
CEI_104	0.006	0.971	0.006	0.008	0.01	574,666,253 times more likely European than Oceanian	European
CEI_108	0.003	0.986	0.004	0.003	0.004	66,805,063,867,654 times more likely European than Oceanian	European
CEI_111	0.007	0.969	0.008	0.004	0.011	13,292,954,566 times more likely European than Oceanian	European
CEI_112	0.057	0.908	0.014	0.011	0.01	504,488 times more likely European than African	European
CEI_120	0.024	0.924	0.013	0.009	0.031	8,920 times more likely European than Oceanian	European
CEI_122	0.009	0.974	0.003	0.008	0.006	1,541,519,049,609,530,000 times more likely European than Oceanian	European
CEI_123	0.014	0.957	0.006	0.014	0.009	153,419,942,550,404 times more likely European than Oceanian	European
CEI_127	0.007	0.94	0.011	0.038	0.005	20,649 times more likely European than American	European
CEI_128	0.005	0.963	0.011	0.01	0.012	179,574,370 times more likely European than Oceanian	European
CEI_130	0.005	0.977	0.005	0.008	0.005	2,794,571,775,820 times more likely European than East Asian	European
CEI_133	0.007	0.593	0.012	0.384	0.004	321 times more likely European than American	European
CEI_135	0.008	0.976	0.003	0.004	0.008	14,246,826,256 times more likely European than Oceanian	European
CEI_156	0.007	0.959	0.013	0.012	0.008	594,455 times more likely European than East Asian	European
CEI_160	0.008	0.951	0.007	0.003	0.031	46,837,747 times more likely European than Oceanian	European
CEI_164	0.017	0.961	0.006	0.012	0.005	280,613,670,155,687 times more likely European than East Asian	European
CEI_179	0.986	0.004	0.003	0.003	0.004	1,756,227,997,850,570,000 times more likely African than European	African
CEI_183	0.028	0.947	0.005	0.008	0.012	1,782,967,095,378 times more likely European than Oceanian	European
CEI_185	0.032	0.937	0.011	0.015	0.006	267,328,672,383,361 times more likely European than African	European

CEI_194	0.005	0.94	0.013	0.025	0.018	26,956,761,678 times more likely European than American	European
CEI_202	0.009	0.564	0.015	0.013	0.399	12.75 times more likely European than Oceanian	European
CEI_203	0.209	0.559	0.084	0.132	0.016	1,982 times more likely European than African	European
CEI_209	0.079	0.819	0.006	0.004	0.092	29,862,080 times more likely European than Oceanian	European
CEI_212	0.003	0.981	0.006	0.006	0.004	499,413,376,711 times more likely European than East Asian	European
CEI_223	0.007	0.979	0.005	0.005	0.005	88,746,046,782,835 times more likely European than East Asian	European
CEI_224	0.004	0.968	0.009	0.008	0.011	337,069,602 times more likely European than Oceanian	European
CEI_232	0.006	0.979	0.004	0.004	0.007	138,467,295,345 times more likely European than Oceanian	European
CEI_236	0.006	0.89	0.082	0.008	0.014	23.32 times more likely European than East Asian	European
CEI_239	0.182	0.71	0.035	0.005	0.068	1,316,122 times more likely European than Oceanian	European
CEI_241	0.021	0.837	0.071	0.033	0.037	22,283 times more likely European than East Asian	European
CEI_242	0.147	0.832	0.012	0.005	0.004	171,425 times more likely European than African	European
CEI_245	0.003	0.977	0.005	0.004	0.012	2,663,334,433,187 times more likely European than Oceanian	European
CEI_247	0.007	0.954	0.005	0.026	0.008	7,736,382,205 times more likely European than American	European
CEI_249	0.012	0.957	0.014	0.01	0.007	2,227,843,691 times more likely European than East Asian	European
CEI_253	0.016	0.871	0.043	0.058	0.011	6,769,244 times more likely European than East Asian	European
CEI_254	0.257	0.698	0.024	0.015	0.006	1,127,309,886 times more likely European than East Asian	European
CEI_262	0.005	0.908	0.009	0.006	0.072	32,500 times more likely European than Oceanian	European
CEI_267	0.004	0.937	0.006	0.005	0.048	2,015,545 times more likely European than Oceanian	European
CEI_275	0.01	0.972	0.004	0.003	0.01	1,285,983,255,455,230 times more likely European than Oceanian	European
CEI_278	0.014	0.906	0.014	0.018	0.049	71,618 times more likely European than Oceanian	European
CEI_280	0.09	0.858	0.017	0.012	0.023	3,796,707,193 times more likely European than Oceanian	European
CEI_281	0.524	0.02	0.066	0.007	0.383	67.64 times more likely African than Oceanian	African
CEI_290	0.113	0.847	0.009	0.021	0.01	5,386,408,757,337 times more likely European than Oceanian	European
CEI_292	0.004	0.964	0.006	0.016	0.009	1,392,608,885,575 times more likely European than American	European
CEI_296	0.462	0.194	0.034	0.241	0.069	29.36 times more likely European than Oceanian	European
CEI_298	0.004	0.959	0.008	0.018	0.012	46,216,374,562 times more likely European than Oceanian	European
CEI_299	0.007	0.958	0.009	0.02	0.006	74,689,099,892 times more likely European than East Asian	European
CEI_302	0.004	0.984	0.004	0.002	0.005	7,195,268,124,871,960 times more likely European than Oceanian	European

Table 5: Relation between the highest percentages of missing data obtained and the corresponding lowest likelihood ratios.

CEI_ID	Missing data (%)	Lowest LR
CEI_3	67	23.19
CEI_236	65	23.32
CEI_42	63	275.42
CEI_127	57	20648.54
CEI_296	52	29.36
CEI_120	50	8920.44
CEI_156	50	594454.82
CEI_203	46	1982.27
CEI_112	41	504488.34
CEI_99	39	2735.04

Through the bar plots generated by STRUCTURE (Figure 5), it is possible to observe that when setting K = 2, the two main clusters discriminated are European (blue) and African (orange). When the value of K increases to 3, a new East Asian component is considered in the analysis, and its proportions are shown (dark blue). If K is forced to 4, some of the contributions that appeared to be from the Asian group, are now revealed as Native American (green). Finally, when K = 5 (the optimal number of clusters), the Oceanian component is unveiled (dark blue), while East Asian proportions (purple) decrease slightly.

It is also possible to notice that 3 of the vertical bars have a major proportion belonging to the African cluster. These bars correspond to CEI_14, CEI_179, and CEI_281, classified as Africans as mentioned above.

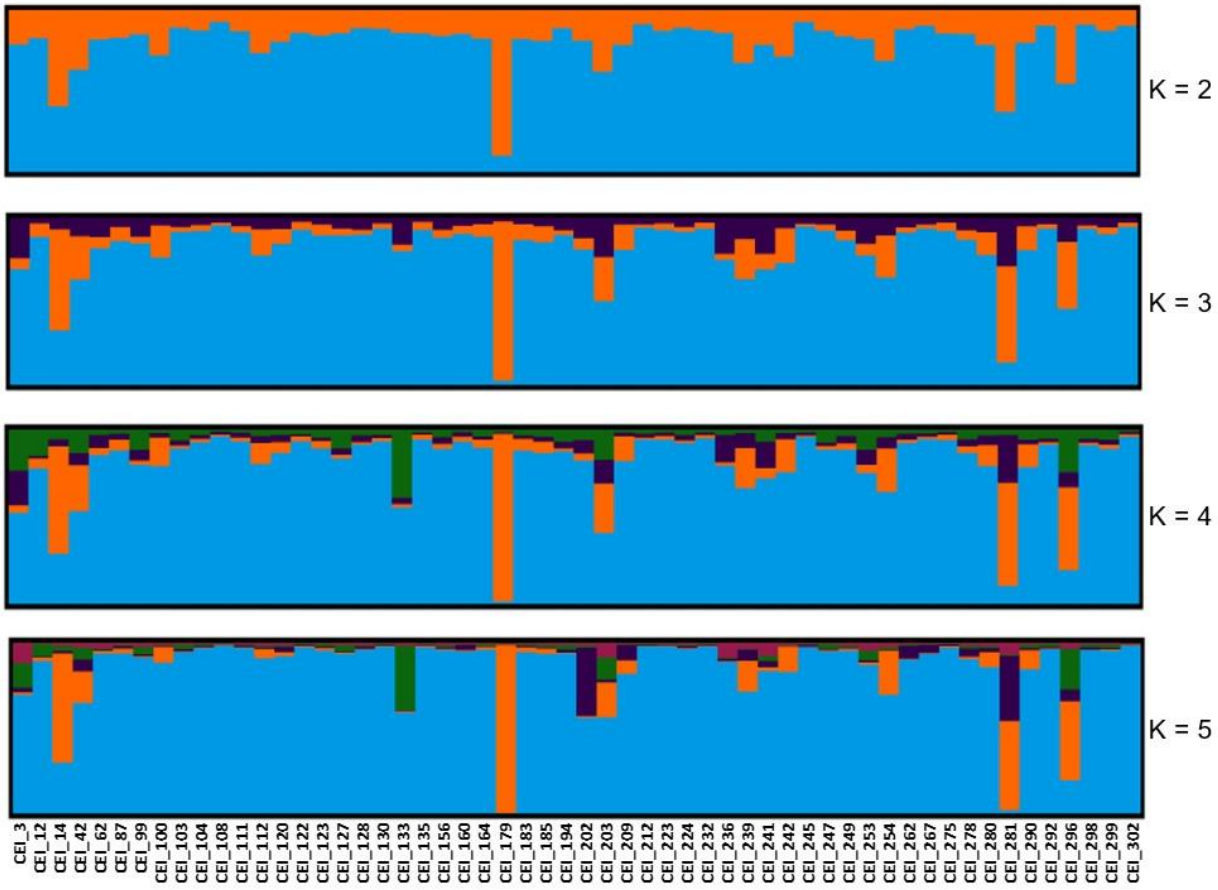


Figure 5: STRUCTURE plot showing CEI samples for K varying from 2 to 5. Each vertical bar represents an individual. Differently coloured zones on each vertical bar show the proportion of the membership of each population component.

I. Non-Admixed Individuals

Figure 6 shows a PCA using the ancestral reference samples, and the study samples that had mainly one ancestral contributor according to STRUCTURE. PC1 and PC2 capture about 38% of the variance in the data, and it is possible to identify three main clusters belonging to Africans, Europeans, and East Asians plus Native Americans and Oceanians.

As expected, most CEI samples have a major European component, which leads them to be plotted alongside this group. Without considering the samples with missing data percentages higher than 39% (which will be discussed later), the lowest likelihood ratios calculated for these individuals ranged from 2.23×10^{4} to 1.7×10^{18}, which support the classification assessed by Snipper, the high membership proportions estimated in STRUCTURE, and the PCA plot – all individuals have a likely European ancestry, except CEI_179 that has a probable African origin.

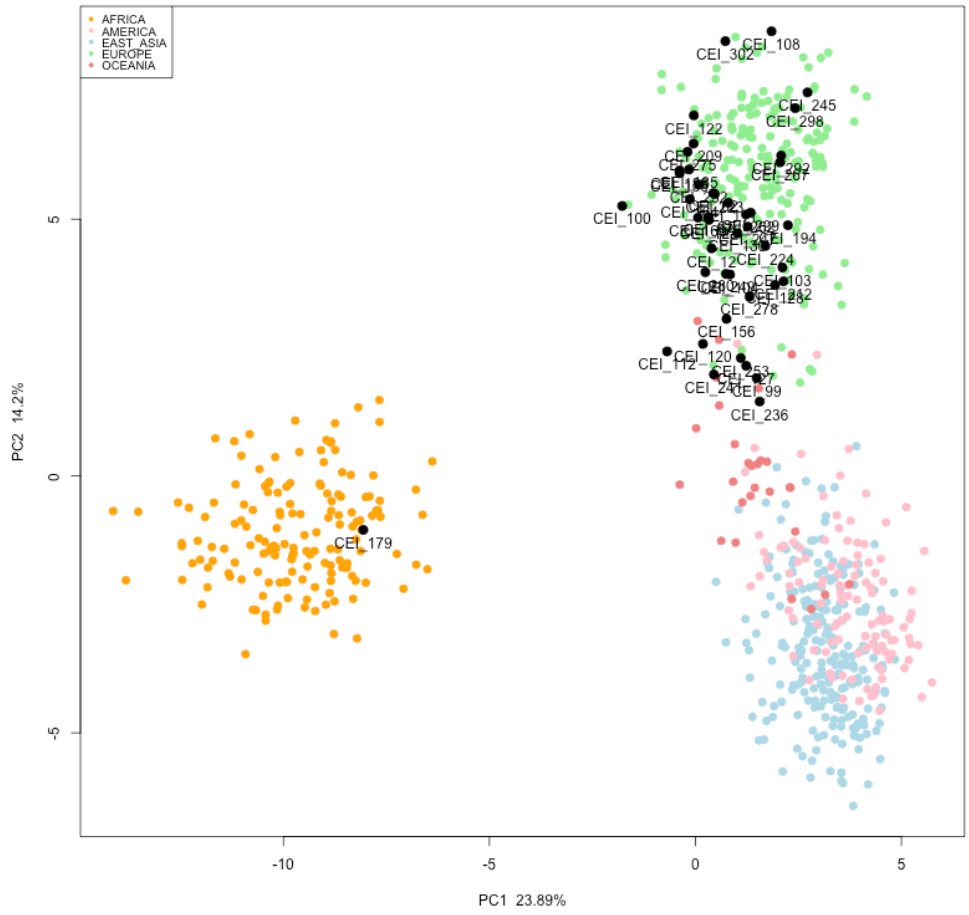


Figure 6: Principal Component Analysis (PCA) plot showing the two components of variation for the individuals with only one major membership according to STRUCTURE, compared to the reference data of five populations.

Regarding CEI_179, this sample has one major membership in the African cluster (99%), which is visible in the STRUCTURE plot, in the PCA, and in the Snipper analysis that concluded that this sample is $1.76x10^{18}$ times more likely African than European, classifying it as of African origin (Figure 5 and 6, Table 4).

For all the other individuals with a main European contributor, STRUCTURE estimated cluster membership proportions that varied from 82% to 99%. Six of these individuals (CEI_99, CEI_112, CEI_120, CEI_127, CEI_156, CEI_236) had missing data percentages higher than 39%, which led to values of LR lower than 8920 (as seen in Table 5), which can be considered low when compared to the mean value of likelihood ratios obtained ($5.91x10^{16}$). Therefore, although these individuals have essentially one major European membership proportion and were all classified as European, the results should be interpreted prudently and confronted with other methods if possible.

CEI_12 and CEI_62 had available *ante-mortem* records with the information of the place of birth, being that the first was born in Mozambique, and the latter in Portugal. Notwithstanding, they were both classified as Europeans with high values of LR ($3.87x10^{12}$ and $3.23x10^{10}$). Cluster membership proportions show that when K = 5 there is only one significant component – the European one. CEI_62s' classification is corroborated by the *ante-mortem* records.

However, there is no noticeable African component in CEI_12 in either of the analyses. Therefrom, two hypotheses arise: either mistakes were made during genotyping analysis, or this individual was born in Africa, but its parents had a European origin.

Some African countries were Portuguese colonies for about four centuries, until 1975 when they got its independence. Nowadays, these countries belong to the PALOP group (Portuguese-speaking African countries) and include Angola, Mozambique, Guiné-Bissau, São Tomé e Princípe, among others. Since the Colonial Wars (in the 60s) and the Independence of these countries, there was an increased flow of Portuguese emigrants trying to come back to Portugal (Simão et al., 2015). Hence, it is possible that this individual was born in Mozambique, but has Portuguese ancestry, and immigrated to Portugal where he/she eventually died.

Regarding the variant alleles found in CEI_100 and CEI_262, these are in accordance with the European classification for these two individuals. Conversely, CEI_209 was classified as European, although the variant allele found in MID360 had been previously reported in African samples only (Pereira et al, 2012). When assessing K = 3 and 4, it is observable that a significant African component is present which might be explained by the occurrence of this allele in this individual, as it happened with African samples used

as reference populations. Nevertheless, these three individuals are all classified as European with values of LR higher than $3.25x10^4$.

II. Admixed individuals

There are several cases of admixed individuals according to the inferences made by STRUCTURE (Table 4). For convenience, in Table 6 are presented the membership proportions for K = 2 to 5 clusters as estimated by STRUCTURE for those individuals with multiple ancestry contributors (more than 10% of cluster proportion).To this table, conditional formatting with colouring was applied to help visualize the relevant contributions of each ancestral group. The order was sorted following a decrease of the European membership proportion.

These possible admixed individuals were also plotted in a PCA (Figure 7) for better visualization.

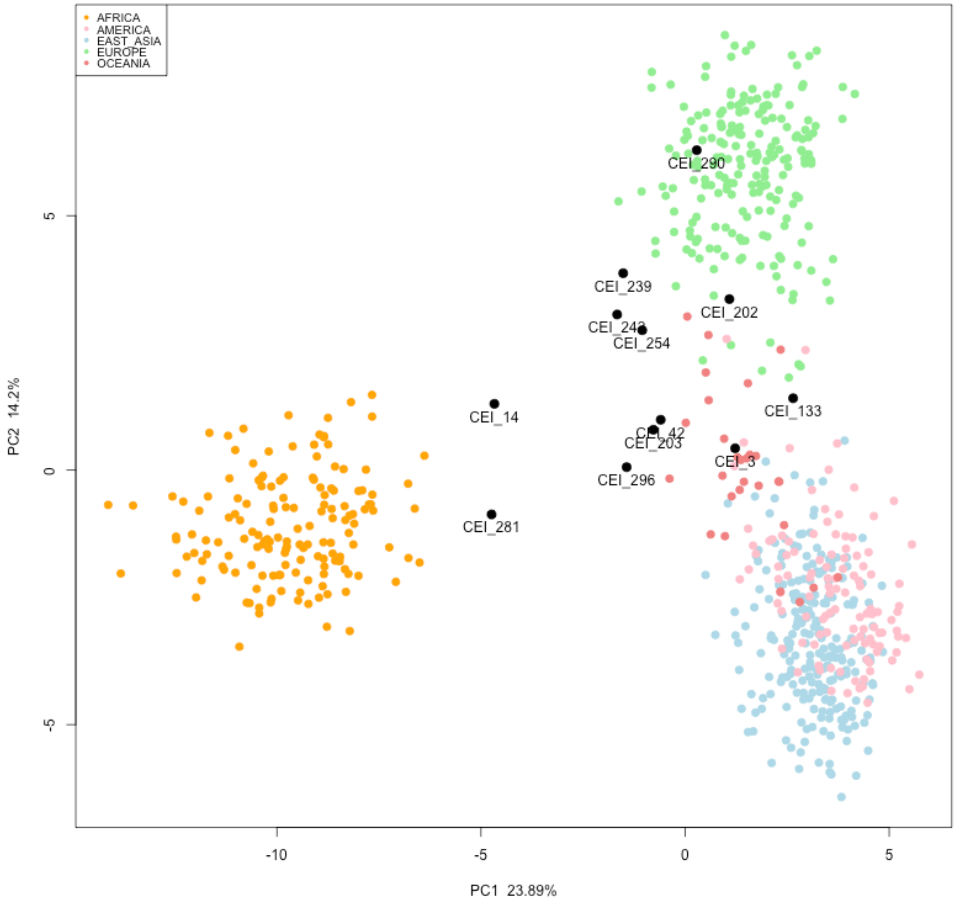


Figure 7: Principal Component Analysis (PCA) plot showing the two components of variation for the admixed individuals according to STRUCTURE, compared to the reference data of five populations.

Genetic ancestry of these individuals will be discussed in two groups: individuals with a major European ancestral contributor, and individuals with a major African ancestral contributor.

Table 4: STRUCTURE membership proportions for K = 2 to 5 clusters, for the individuals with more than one significant membership component, with respective missing data percentages (MD).

		K=2		K=3			K=4				K=5				
	MD (%)	AFR	EUR	AFR	EUR	EAS	AFR	EUR	EAS	NAM	AFR	EUR	EAS	NAM	OCE
CEI_290	0	0.202	0.798	0.144	0.808	0.049	0.132	0.787	0.025	0.056	0.113	0.847	0.009	0.021	0.01
CEI_242	28	0.285	0.715	0.208	0.732	0.06	0.189	0.757	0.036	0.018	0.147	0.832	0.012	0.005	0.004
CEI_239	17	0.327	0.673	0.24	0.633	0.127	0.228	0.665	0.089	0.017	0.182	0.71	0.035	0.005	0.068
CEI_254	0	0.311	0.689	0.253	0.647	0.101	0.25	0.645	0.064	0.041	0.257	0.698	0.024	0.015	0.006
CEI_3	67	0.213	0.787	0.064	0.694	0.242	0.043	0.527	0.196	0.234	0.014	0.688	0.121	0.151	0.026
CEI_42	63	0.369	0.631	0.259	0.632	0.11	0.26	0.541	0.07	0.129	0.183	0.639	0.03	0.071	0.076
CEI_133	0	0.141	0.859	0.039	0.805	0.156	0.021	0.555	0.036	0.389	0.007	0.593	0.012	0.384	0.004
CEI_202	2	0.186	0.814	0.064	0.813	0.122	0.041	0.824	0.076	0.059	0.009	0.564	0.015	0.013	0.399
CEI_203	46	0.379	0.621	0.264	0.505	0.231	0.285	0.407	0.139	0.169	0.209	0.559	0.084	0.132	0.016
CEI_14	17	0.594	0.406	0.607	0.326	0.067	0.617	0.29	0.041	0.053	0.639	0.306	0.016	0.026	0.013
CEI_296	52	0.454	0.546	0.4	0.454	0.146	0.472	0.203	0.087	0.238	0.462	0.194	0.034	0.241	0.069
CEI_281	13	0.625	0.375	0.578	0.134	0.288	0.591	0.104	0.276	0.029	0.524	0.02	0.066	0.007	0.383

a. Major European Ancestral Component

Most of the individuals have a major component that belongs to the European group, as discussed above. However, there are different levels of admixture that can be observed in this group.

CEI_290, CEI_242, CEI_239, CEI_254 are individuals that have a major European ancestral component and a smaller but relevant African contribution, according to STRUCTURE. Moreover, CEI_239 and CEI_254 show a significant East Asian membership when assessing K = 3. Thus, a custom PCA was made using these three main groups as reference samples (Figure 8).

The PCA shows that CEI_290 is plotted right in the centre of the European cluster, which seems to not value the small African component estimated by STRUCTURE. An explanation for this occurrence is that sometimes, due to undetected structure in DNA genotypes, an individual can be displaced from their real position (Phillips, 2015). Since these are very degraded samples, it is possible that, during genotyping analysis, heterozygous loci were called as homozygous due to allele drop out, or homozygous loci were called as heterozygous due to allele drop-in, which led to the presence of a significant but unreal African component read by STRUCTURE but not by the PCA.

Snipper App Suite did classify these individuals as Europeans with high values of LR, hence reflecting their major membership proportion in this group (68,9% to 84,7%) with STRUCTURE. But considering their individual ancestry estimates and the intermediate positioning of CEI_242, CEI_239 and CEI_254 somewhat disassociated from the European towards the African references in the custom PCA, it is clear that these individuals harbour a relevant African background.

As already mentioned above, some African countries were Portuguese colonies. Therefore, it would not be hard to accept that individuals with these admixed ancestry proportions can be found in Portugal.

Nevertheless, it is worth mentioning that this pattern of admixed ancestry can also be observed in populations from South America. Europeans colonized most South American countries, being responsible for the transatlantic slave trade, transporting enslaved African people to these countries.

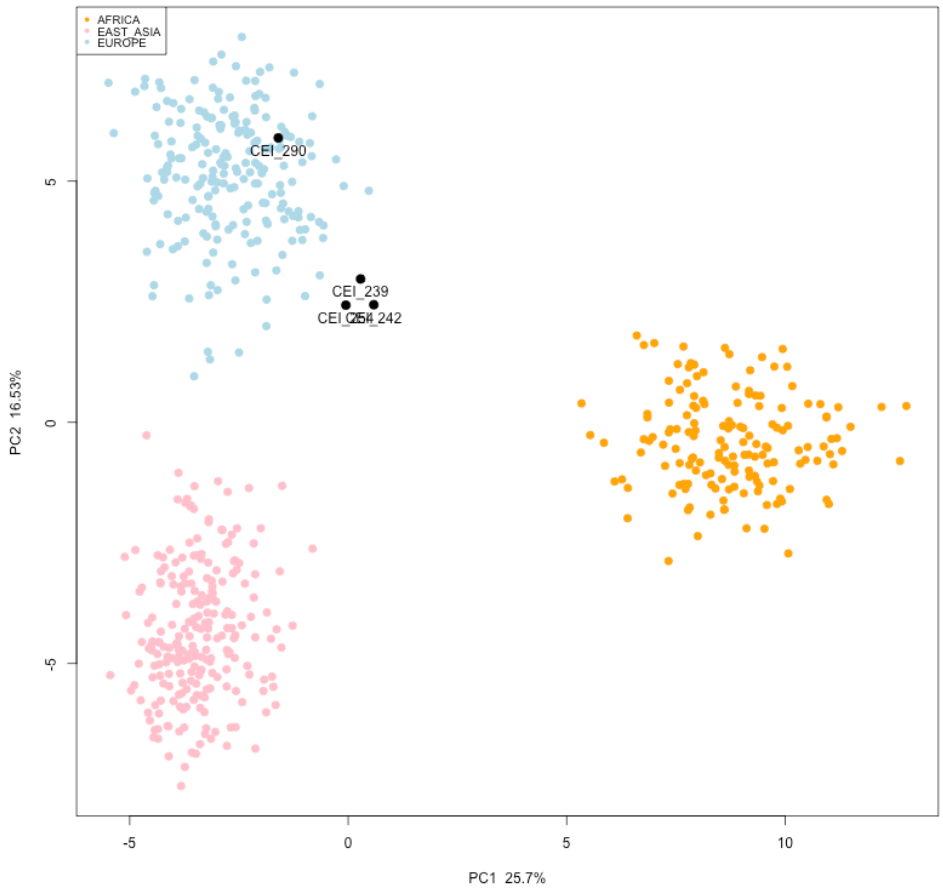


Figure 8: Principal Component Analysis (PCA) plot showing the two components of variation for CEI_239, CEI_242, CEI_254, and CEI_290, using the reference samples of Africans, East Asians, and Europeans.

Another four individuals present a European major component, with one or two smaller components, one of them belonging to the Native American group.

Regarding CEI_3, it presents admixture between the European (60%), East Asian (12%), and Native American (15%) ancestral contributors. It has a very low likelihood ratio (23.19) which makes it difficult to acknowledge the classification of European origin with confidence. According to the *ante-mortem* records, this individual was born in Mozambique, which contradicts the genetic findings since there is no significant African contribution. In this specific case and due to the high levels of missing data (67%), a repetition of the analysis (from DNA extraction to statistical analysis) would be suggested, or another method should be performed and both results confronted. There is also the hypothesis that this individual was a descendant of people that lived in Mozambique but were not African, in which case, the birthplace records become useless. This is one of the main difficulties when it comes to assessing the biogeographic origin

nowadays – migration is such a common phenomenon these days that the place of birth of an individual might have no correlation with its genetic composition.

CEI_42 has ancestral contributions from the European and the African population groups. But when analysing the cluster membership proportions for K = 4, a relevant Native American component (13%) is revealed. According to Snipper, CEI_42 is classified as European, although with a low LR (275). This implies that this individual probably has some level of admixture between the European, African, and Native American clusters. Indeed, according to the records, this individual was born in Brazil, which corroborates the genetic results since the Brazilian population has contributions of the European, African, and Native American clusters (Santos et al., 2010). Brazil was a Portuguese colony for three centuries, and its official language is Portuguese. Therefore, there is a high flow of migration between the two countries, especially in the last decades. According to *Serviço de Estrangeiros e Fronteiras* (SEF) (Machado et al., 2020), in 2019, 151 000 Brazilians were living in Portugal, with more than 50% in the Metropolitan Area of Lisbon, very close to Santarém, where these individuals were buried. Thus, this is probably a Brazilian that immigrated and died in Portugal. Even though the records corroborate the genetic findings, there is 63% of missing data in this case, and unfortunately, the skull was too fragmented for an anthropological assessment. Nevertheless, even with that level of missing data, this panel of AIM-Indels was able to capture substructure in an admixed sample.

CEI_203 also has contributions from the European (56%), African (21%), and Native American (13%) ancestral groups. When displaying a PCA with these three clusters (Figure 9), this individual is plotted distanced from any reference population. This is an indication that this is probably an admixed individual, yet, due to the high percentage of missing data (46%), interpretation should be made carefully. From the 25 markers reported in this individual, a distinctive tri-hybrid profile emerges, characteristic of the ones found in South America. Snipper calculated the lowest LR as 1982 times more likely European than African, which is not very high but encounters explanation in the major European proportion estimated by STRUCTURE, as well as in the custom PCA.

CEI_133 revealed a major contribution belonging to the European group (59%), and a smaller but relevant Native American component (38%), being 321 times more likely European than Native American according to Snipper. Moreover, if CEI_133 is displayed in a PCA with only reference samples from Europe, Africa, and America (Figure 9), it remarkably positions itself in the separation of the European and Native American

clusters. The – log (likelihood) values of occurring on European or Native American gene pools are very similar which explains the low LR and otherwise corroborates the membership proportions found. This kind of admixed profile with mainly European and Native American components is also very common to find in the Brazilian population.

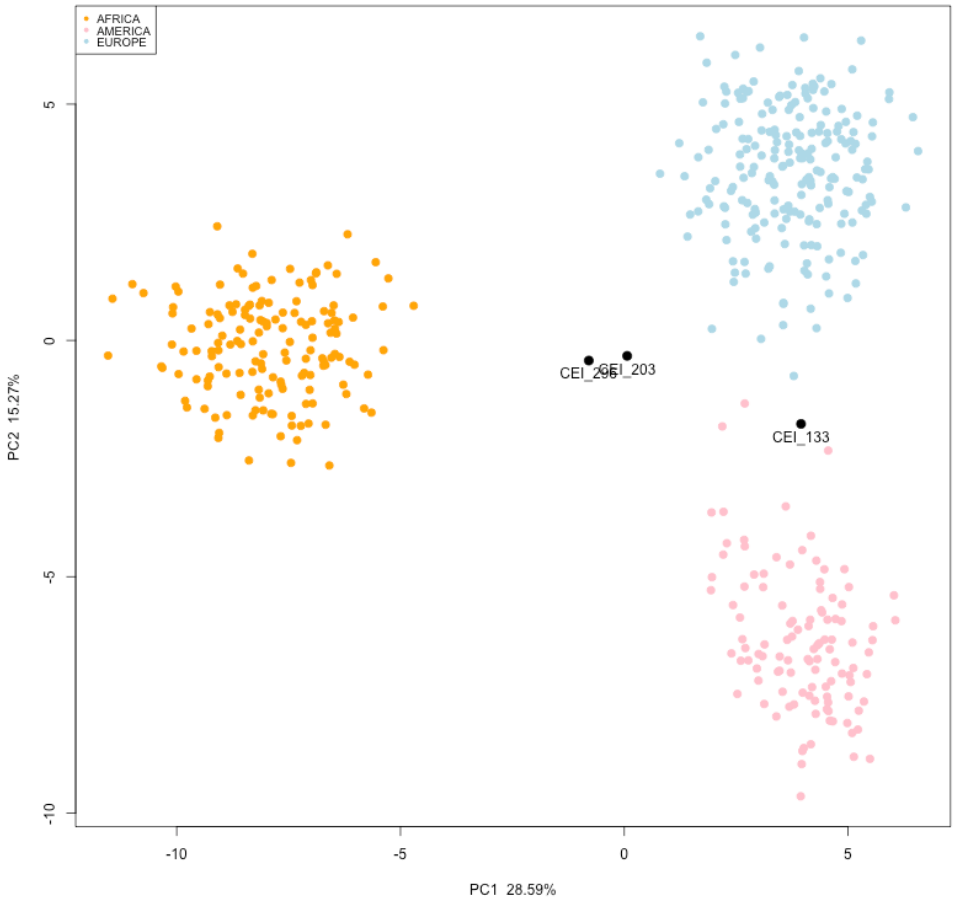


Figure 9: Principal Component Analysis (PCA) plot showing the two components of variation for the samples CEI_133, CEI_203, and CEI_296, using the reference samples of Africans, Europeans, and Native Americans.

A peculiar result was obtained regarding CEI_202. This individual reveals a major European component (60%), and a considerable proportion in the Oceanian cluster (40%), resulting in a likelihood ratio of being just 12.75 times more likely European than Oceanian, according to Snipper app suite. Given the unusual admixture patterns found in the STRUCTURE analyses (see Table 6), and in order to try to further understand the Oceanian contribution, a PCA was performed using only the European, East Asian, and Oceanian reference populations (Figure 10). The Oceanians appear more scattered than the others, and there is not a very clear distinction between them and the East Asian cluster. Noteworthy, when assessing K = 3 in STRUCTURE, there is an observable

proportion in the East Asian cluster. Then, when forcing K = 5, the fact that the Oceanian gene pool presents some affinity with the clinal genetic variation found between Europeans and East Asians (Rosenberg et al., 2002; Rosenberg et al., 2005), might constitute a confounding factor and have misled the program into assigning membership as Oceanian, when there is not a real contribution from this ancestral group in this individual.

If the second-lowest likelihood ratio is considered (using East Asian information), this individual is estimated to be 15,032,123 times more likely European than East Asian, therefore favouring a putative European ancestry, in line with the highest membership proportion estimated with STRUCTURE. Still, it can be inferred that there is ~18% of non-European genetic variability in this individual, that has more affinity with the East Asian and/or Oceanian gene pool.

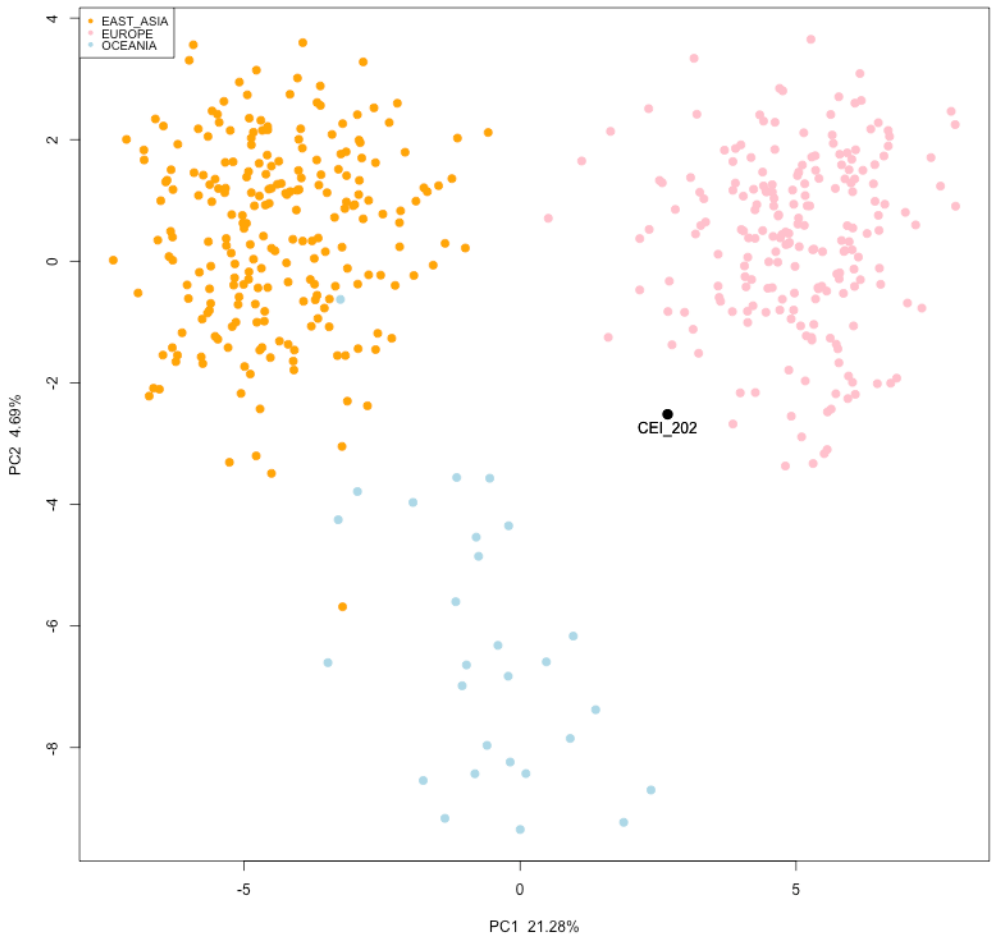


Figure 10: Principal Component Analysis (PCA) plot showing the two components of variation for the sample CEI_202, using the reference samples of Europeans, East Asians, and Oceanians.

b. Major African Ancestral Component

Three individuals revealed a predominant African contribution admixed with varied but significant non-African components.

CEI_14 has a main African contribution (64%), and a smaller proportion from European origin (31%). Snipper app suite classified this individual as of African origin while being just 3215 times more likely to be African than European, which makes it risky to classify this individual to a single ancestral contributor. *Ante-mortem* records of this individual show that this is an Angolan individual (Table 1), which is naturally compatible with the genetic ancestry results.

Angola belongs to PALOP since it was a Portuguese colony from the 15th to 20th century, and there are still Portuguese descendants living there. According to the data published by Serviço de Estrangeiros e Fronteiras (SEF) about foreigners living in Portugal (Machado et al., 2020), in 2019 there were about 90 000 immigrants from PALOP with residency in Portugal. Interestingly, most of these immigrants found residency in the Metropolitan Area of Lisbon, which is very close to Santarém, the city where these individuals were buried. Therefore, this is probably a recently admixed individual with Portuguese and Angolan ancestry, that immigrated and died in Portugal.

CEI_296 has African (46%), Native American (24%), and European (19%) contributions. Although STRUCTURE estimates show that the major contributor is African, Snipper analysis classifies this individual as European, although with a very low LR (29.36) over Oceanian. A custom PCA was computed with only those three reference clusters (Figure 9), and it shows that CEI_296 is plotted in an intermediate position among the ancestral populations. When analysing the – log (likelihood) values for this individual it is perceptible that these values are not markedly different among some ancestral groups, probably due to the high level of missing data (52%), which led to a lack of information, and also because of the admixed nature of the individual. Phillips et al., 2009 stated that likelihood ratios below 100 should be considered unreliable, which is the case for this individual. Therefore, another analysis or confrontation with other results should be proposed.

This is the only case where membership proportions given by STRUCTURE do not match the classification given by Snipper. The Snipper app suite contends only the three best likelihoods for a given profile to output the lowest LRs possible and infers the putative ancestral origin. In this case, the African hypothesis was not even considered because the -log(likelihoods) were lowest for European, Oceanian, and East Asian. When considering the African hypothesis, the LR estimated is 733.8 times more likely

European than African, which is still a low value. This anomaly happened probably due to the low number of markers reported.

Interestingly, CEI_281 shows a major African ancestral contributor (52%) with a substantial Oceanian component (38%) when considering 5 continental ancestors as possible contributors (K = 5). Just as seen above for individual CEI_202, this kind of admixture pattern is unusual considering the Portuguese context. According to Snipper, this individual is 67.64 times more likely to be African than Oceanian. When analysing the membership proportions for K = 3 and 4, it is observable that the East Asian component emerges in both scenarios (Table 6). The same can be noticed when a custom PCA is displayed with only the African, East Asian, and Oceanian reference samples (Figure 11) - CEI_281 is plotted between the reference samples, although closer to the Africans. It is, therefore, possible that this individual presents some non-African component that clusters preferentially within the Oceanians and/or East Asians, due to the fact that the Oceanian gene pool harbours subsets of the East Asian and European genetic variation (Rosenberg et al., 2002; Rosenberg et al., 2005). Thus, this individual can be considered to have an admixed ancestry with contributions from the African, European, and East Asian ancestral groups.

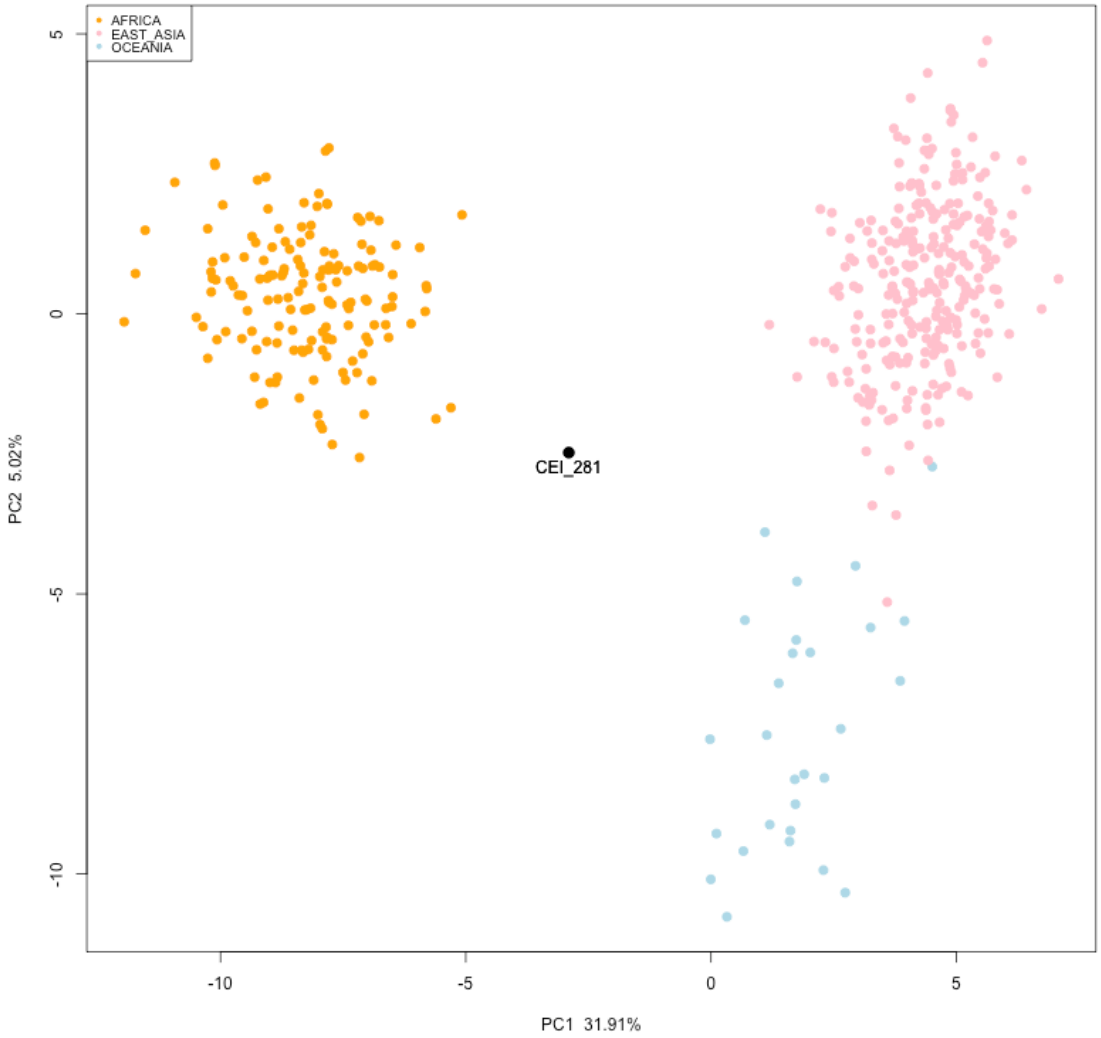


Figure 11: Principal Component Analysis (PCA) plot showing the two components of variation for the sample CEI_281, using the reference samples of Africans, East Asians, and Oceanians.

Comparison with Anthropological data

Ancestry assessment through anthropological methods was made for 50 individuals and is represented in Table S6. As expected, European membership is the most common one.

In some cases, the two methods used (metric and non-metric) do not corroborate each other. There are 5 individuals with different ancestry inferences in each method (CEI_14, CEI_87, CEI_156, CEI_281, and CEI_290) that will be discussed later. Another observation can be made about the probabilities given by each method: the metric approach has a mean probability value of 79.3%, while the morphological approach has a mean of 90.9%.

I. Non-admixed individuals according to genetics

CEI_179, the only individual with a putative African origin according to genetics, was also classified as having probable African origin with both metric and non-metric approach (79.60%, and 56.50%, respectively).

From the individuals with high values of missing data, only one had a different classification than the probable European origin. CEI_156 was classified as being Native American with the non-metric approach with a percentage of 62.2%. When analysing the cluster membership proportions for all the inferred K, there is no other relevant component either than the European one. The lowest likelihood ratio estimated for this individual is 594,455 times more likely European than East Asian. However, due to the high levels of missing data (50%) and applying a conservative logic, the genetic analysis must be interpreted carefully. Therefore, in this case, the anthropological data is an added value because, on one hand, it corroborates the classification given by genetics in the metric approach, and on another hand, it adds some more information that should be taken into consideration cautiously, since there is no other data to support it. Nevertheless, both methods agree to some extent with the classification as of European origin.

The metric approach calculated, for CEI_87, a probability of 51.60% of being Native American, a classification different from the one given by the non-metric methods that classified this individual as European. Nothing in the genetic analyses indicates the presence of a Native American component. In fact, the morphological approach

corroborates the genetic analysis that states that this individual is 13,224,602,017,164 times more likely European than Oceanian. Therefore, the classification given to this individual should be kept as European.

II. Admixed individuals according to genetics

Regarding the four individuals that were classified as probable Europeans with a significant African component, anthropological methods estimated a European origin for all, except for CEI_290.

CEI_290 has a probability of 57% of being Native American, and of 92% of being European, according to the metric and non-metric approaches, respectively. The Native American component is never significantly present in any of the K inferred by STRUCTURE analyses. However, there is a relevant African component detected by STRUCTURE. In this case, although both sciences agree with a most likely European ancestry, they also detected other components that did not match. Another approach should be proposed, or a repetition of the methods applied.

Concerning CEI_254, anthropological methods classified it as European although with probabilities around 60%. Genetic analysis shows that this individual has a major European component, but with a smaller African proportion. Although this individual is classified as European in all methods, it probably has some mixed traits due to the African component that explains the low probabilities given by anthropology.

CEI_3 was one of the individuals that were inferred to have mainly European ancestry (69%) and smaller East Asian (12%) and Native American (15%) components. Anthropological methods classified this individual as European with high probabilities (> 84%), which corroborates the European assignment given by Snipper app suite. According to the *ante-mortem* records, this individual was born in Mozambique, however, none of the approaches estimated an African component present here. Therefore, both genetic and anthropological methods agree that this is probably an individual born in Africa but with a probable European origin.

CEI_203 has probabilities close to 50% of being European according to anthropology. These low values are in accordance with the presence of other ancestral components found in STRUCTURE analysis, which revealed an essentially tri-hybrid profile, typical of South America. Since the major component estimated belongs to the European group,

and the anthropological methods only report the most likely putative origin, the results are corroborated by both sciences. The low probabilities given by the anthropological methods should be an indication of a probable admixed individual, which was successfully assessed through genetics.

CEII_133 has a genetic ancestry profile typically found in the Brazilian population due to its European and Native American high proportions. It is classified as European by the anthropological methods with high probabilities (95% and 99%). Although the classification is not wrong (is the same as the one given by genetics), anthropological methods failed to detect the Native American component, probably due to the fact that features of the skull are not population-specific, and there is an overlap of features through different regions (Cunha & Ubelaker, 2019).

CEI_202 revealed unusual genetic ancestral proportions of 60% European and 40% Oceanian in the 5-group analysis. It was classified as European by both genetic and anthropological methods. However, while genetic methods were able to detect a fraction of variability belonging non-European ancestors, anthropology failed to do that probably since such proportion was very small for the anthropological methods to detect.

CEI_14 has 69.80% of being Asian according to the metric approach, and 65.90% of being African according to the morphological approach. When assessing K = 3 to 5 by STRUCTURE, there were no East Asian components with significant values, which completely contradicts the result of the metric approach. Interestingly, the percentage given by the morphological approach is not very high which might be due to a great European component also present in this individual according to genetic ancestry estimates. According to genetics, this is an admixed individual with mostly African (64%) and European (31%) components, which are compatible with the *ante-mortem* records, that confirm that this individual was born in Angola. Therefore, the assessment made by the metric analysis is not in agreement with these findings, although the non-metric assessment does corroborate the genetic results.

Snipper classified CEI_296 as European, but STRUCTURE detected a predominant African component (46%). More than half of the markers in this sample were not reported (52%), which render these results unreliable. The anthropological assessment, however, classified this individual as European with 75% and 98% probability. Therefore, since there is no more information about this individual, and in a way to try to reconcile both strategies, the final classification should be a probable European origin, with a possible admixture with other ancestral groups.

CEI_281 has very interesting results. According to genetics, it is classified as African despite the 38% of non-African, probably East Asian, variability. Coincidentally, the metric approach gives a likelihood of 53% of being Asian, while the morphological approach calculated a probability of 32% of being European. Technically, it is not wrong since both ancestral contributors are present in this sample's genetic pool. However, the main contributor (African) was not detected, which demonstrates the difficulty of identifying admixed individuals using anthropological methods.

CONCLUDING REMARKS

DNA extraction from bone samples is improving, but there are still some aspects like degradation that are very hard to overcome. Even when the extraction is made successfully, there is always the matter of LCN and low-quality DNA, that increase the level of difficulty of genotyping these samples, especially for someone with little experience. The AIM-Indel panel used in this study has proven to be very effective when it comes to assessing ancestry of bone samples as previously demonstrated by Romanini et al. (2015). Even with high numbers of unreported markers, this panel was able to detect substructure in probable admixed individuals. The simple technique involved in genotyping the 46 AIM-Indel, and the ability to deal with problems usually found in forensic DNA samples (degradation, inhibitors, concentration) makes it ideal for the forensic field.

Population Genetics is crucial when it comes to ancestry analysis. It is not possible to find the putative origin of an individual if that geographic region has never been studied. The HGDP-CEPH diversity panel with samples from the African, European, East Asian, Native American, and Oceanian populations, allowed the assignment of non-admixed individuals into one ancestral contributor with high membership proportions, while still being able to assign individuals to multiple ancestral contributors in the case of probable admixture. Nevertheless, it is important to be cautious when the level of unreported markers is sufficient to raise doubt about the results.

In a general outlook, most classifications given by anthropology were the same as the ones given by genetics. When it comes to individuals with essentially one ancestral contribution, the two sciences (generally) see eye-to-eye. However, when it comes to possible admixed individuals, the anthropological assessment did not always match the one given by genetics. The non-metric assessment had more concordant results with genetics, while the metric assessment, besides classifying some individuals to different ancestry groups, gave markedly lower probabilities than the ones given by non-metric methods. One should bear in mind that the features evaluated by anthropology are phenotypic ones, which means that they are not only determined by genetics, but also by the environment. Thus, it is not a complete surprise that some results are different since anthropology “reads” more information than the one given by genetics.

Nonetheless, within the 56 skeletons evaluated it was possible to identify 43 European individuals, one African individual, four individuals with admixture pattern characteristics of South America (probably Brazil), five individuals with admixture between the European

and African groups in different proportions, one individual with probable admixture between the European and Oceanian group, and another with contributions from African and Oceanian.

Further studies, especially with admixed individuals should be carried out for anthropological and genetic ancestry assessments. The best way to improve the anthropological methods is by confronting and understanding the information given by genetics, and by the environment (using samples of known admixed individuals). Due to globalisation, the existence of mixed traits is more and more usual since migration is such an easy and common phenomenon nowadays, turning the evaluation of ancestry more complex.

The inference of biogeographical origin should never be a task performed by only one discipline since different sciences can gather different information that can be useful to aid the investigation. While genetic ancestry has a more straightforward analysis, anthropology assessment reports a more solid information about the phenotypic features of an individual. On the other hand, both approaches are depending on the biological material available. Without DNA, genetics cannot be applied; with a severely fragmented skeleton, anthropology has no way to act.

Forensic scientists should aim to achieve a multidisciplinary effort in order to retrieve all possible information. Forensic Genetics and Forensic Anthropology should complement each other, to overcome each’s weak points, and corroborate results.

REFERENCES

Albanese, J., & Saunders, S. (2006). Is It Possible to Escape Racial Typology in Forensic Identification? In *Forensic Anthropology and Medicine: Complementary Sciences From Recovery to Cause of Death* (pp. 281–316). Humana Press.

Ayub, Q., Haber, M., & Tyler-Smith, C. (2016). Evolutionary and Population Genetics in Forensic Science. In Amorim, A., & Budowle, B. (Eds), *Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation* (pp. 33 – 60). World Scientific Europe Ltd.

Amorim, A., & Budowle, B. (2016). Definition and Purpose. In Amorim, A., & Budowle, B. (Eds), *Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation* (pp. 1 – 12). World Scientific Europe Ltd.

Amorim, A., & Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: A comparative analysis with STRs. *Forensic Science International*, *150*(1), 17–21. https://doi.org/10.1016/j.forsciint.2004.06.018

Anonymous. (2007). Launching Forensic Science International daughter journal in 2007: Forensic Science International: Genetics. *Forensic Science International: Genetics*, *1*(1), 1–2. https://doi.org/10.1016/j.fsigen.2006.10.001

Bowcock, A., & Cavalli-Sforza, L. (1991). The study of variation in the human genome. *Genomics*, *11*(2), 491–498. https://doi.org/10.1016/0888-7543(91)90170-J

Brinkmann, B., Klintschar, M., Neuhuber, F., Hu, J., & Rolf, B. (1998). Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *American Journal of Human Genetics*, *62*, 1408–1415.

Butler, J. M. (2010). *Fundamentals of Forensic DNA Typing*. Elsevier.

Butler, J. M. (2012). *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier.

Butler, J. M. (2015). The future of forensic DNA testing. *Philosophical Transactions of the Royal Societ*, *370*. http://dx.doi.org/10.1098/rstb.2014.0252

Cann, H. M., Toma, C., Cazes, L., Legrand, M. F., Morel, V., & Piouffre, L. (2002). A human genome diversity cell line panel. *Science, 296*. doi:10.1126/science.296.5566.261b

Cattaneo, C. (2007). Forensic anthropology: developments of a classical discipline in the

new millennium. *Forensic Science International*, *165*(2–3), 185–193. https://doi.org/10.1016/j.forsciint.2006.05.018

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P., & Mountain, J. (1988). Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(16), 6002–6006. https://doi.org/10.1073/pnas.85.16.6002

Cunha, E., & Cattaneo, C. (2006). Forensic Anthropology and Forensic Pathology. In *Forensic Anthropology and Medicine: Complementary Sciences From Recovery to Cause of Death* (pp. 39–53).

Cunha, E., & Ubelaker, D. H. (2019). Evaluation of ancestry from human skeletal remains: a concise review. *Forensic Sciences Research*, *5*(2), 89–97. https://doi.org/10.1080/20961790.2019.1697060

d'Oliveira Coelho, J., Curate, F., & Navega, D. (2020). Osteomics: Decision support systems for forensic anthropologists. *Statistics and Probability in Forensic Anthropology*, 259–273. https://doi.org/10.1016/b978-0-12-815764-0.00005-8

Enoch, M. A., Shen, P. H., Xu, K., Hodgkinson, C., & Goldman, D. (2006). Using ancestry-informative markers to define populations and detect population stratification. *Journal of Psychopharmacology*, *20*, 19–26. https://doi.org/10.1177/1359786806066041

Ferreira, M. T., Coelho, C., Makhoul, C., Navega, D., Gonçalves, D., Cunha, E., & Curate, F. (2020). New data about the 21st Century Identified Skeletal Collection (University of Coimbra, Portugal). *International Journal of Legal Medicine*. https://doi.org/10.1007/s00414-020-02399-6

Ferreira, M. T., Vicente, R., Navega, D., Gonçalves, D., Curate, F., & Cunha, E. (2014). A new forensic collection housed at the University of Coimbra, Portugal: The 21st century identified skeletal collection. *Forensic Science International*, *245*, 202.e1-202.e5. https://doi.org/10.1016/j.forsciint.2014.09.021

Gaudio, D., Fernandes, D. M., Schmidt, R., Cheronet, O., Mazzarelli, D., Mattia, M., … Pinhasi, R. (2019). Genome-Wide DNA from Degraded Petrous Bones and the Assessment of Sex and Probable Geographic Origins of Forensic Cases. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-44638-w

Goodwin, W., Linacre, A., & Hadi, S. (2011). *An Introduction to Forensic Genetics*. Wiley.

Halder, I., Shriver, M., Thomas, M., Fernandez, J. R., & Frudakis, T. (2008). A panel of

ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: Utility and applications. *Human Mutation*, *29*(5), 648–658. https://doi.org/10.1002/humu.20695

Hasap, L., Chotigeat, W., Pradutkanchana, J., Vongvatcharanon, U., Kitpipit, T., & Thanakiatkrai, P. (2020). A novel, 4-h DNA extraction method for STR typing of casework bone samples. *International Journal of Legal Medicine*. https://doi.org/10.1007/s00414-019-02232-9

Hefner, J. T. (2009). Cranial nonmetric variation and estimating ancestry. *Journal of Forensic Sciences*, *54*(5), 985–995. https://doi.org/10.1111/j.1556-4029.2009.01118.x

Houck, M. M., & Siegel, J. A. (2010). *Fundamentals of Forensic Science* (2nd ed). Elsevier Ltd.

Iscan, M. Y., & Steyn, M. (2013). *The Human Skeleton in Forensic Medicine*. Charles C. Thomas.

Jia, J., Wei, Y. L., Qin, C. J., Hu, L., Wan, L. H., & Li, C. X. (2014). Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates. *Forensic Science International: Genetics*, *8*(1), 187–194. https://doi.org/10.1016/j.fsigen.2013.09.004

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., & Tyler-Smith, C. (2014). *Human Evolutionary Genetics*. Garland Science, Taylor & Francis Group, LLC.

Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M., & Kidd, K. K. (2011). Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics*, *2*(1), 1–13. https://doi.org/10.1186/2041-2223-2-1

Kidd, K. (2016). Thoughts on Estimating Ancestry. In Amorim, A., & Budowle, B. (Eds), *Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation* (pp. 131 – 144). World Scientific Europe Ltd.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour, 15*(5), 1179-1191. doi:10.1111/1755-0998.12387

L'Abbé, E. N., Van Rooyen, C., Nawrocki, S. P., & Becker, P. J. (2011). An evaluation of non-metric cranial traits used to estimate ancestry in a South African sample.

Forensic Science International, *209*(1–3), 195.e1-195.e7. https://doi.org/10.1016/j.forsciint.2011.04.002

Latham, K. E., & Miller, J. J. (2019). DNA recovery and analysis from skeletal material in modern forensic contexts. *Forensic Sciences Research*, *4*(1), 51–59. https://doi.org/10.1080/20961790.2018.1515594

Lewontin, R. C. (1972). The apportionment of human diversity revisited. In *Evolutionary Biology* (pp. 381–398). https://doi.org/10.1002/ajpa.22899

Liebenberg, L., Krüger, G. C., L'Abbé, E. N., & Stull, K. E. (2019). Postcraniometric sex and ancestry estimation in South Africa: a validation study. *International Journal of Legal Medicine*, *133*(1), 289–296. https://doi.org/10.1007/s00414-018-1865-x

Liu, Y. Y., & Harbison, S. A. (2018). A review of bioinformatic methods for forensic DNA analyses. *Forensic Science International: Genetics*, *33*, 117–128. https://doi.org/10.1016/j.fsigen.2017.12.005

Machado, R., Reis, S., Esteves, S., Sousa, P., Rosa, A.P. (2020) Relatório de Imigração, Fronteiras e Asilo 2019, Oeiras.

Maroñas, O., Phillips, C., Söchtig, J., Gomez-Tato, A., Cruz, R., Alvarez-Dios, J., de Cal, M. C., Ruiz, Y., Fondevila, M., Carracedo, Á., & Lareu, M. V. (2014). Development of a forensic skin colour predictive test. *Forensic science international. Genetics*, *13*, 34–44. https://doi.org/10.1016/j.fsigen.2014.06.017

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, *16*(9), 1182–1190. https://doi.org/10.1101/gr.4565806

Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., & Devine, S. E. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research*, *21*(6), 830–839. https://doi.org/10.1101/gr.115907.110

Misner, L. M., Halvorson, A. C., Dreier, J. L., Ubelaker, D. H., & Foran, D. R. (2009). The correlation between skeletal weathering and dna quality and quantity. *Journal of Forensic Sciences*, *54*(4), 822–828. https://doi.org/10.1111/j.1556-4029.2009.01043.x

Mullaney, J. M., Mills, R. E., Stephen Pittard, W., & Devine, S. E. (2010). Small insertions

and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), 131–136. https://doi.org/10.1093/hmg/ddq400

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304.

Nassir, R., Kosoy, R., Tian, C., White, P. A., Butler, L. M., Silva, G., … Seldin, M. F. (2009). An ancestry informative marker set for determining continental origin: Validation and extension using human genome diversity panels. *BMC Genetics*, *10*, 1–13. https://doi.org/10.1186/1471-2156-10-39

National Human Genome Research Institute (n.d.). Polymorphisms. Retrieved August 25 2020, from https://www.genome.gov/genetics-glossary/Polymorphism

Navega, D., Coelho, C., Vicente, R., Ferreira, M. T., Wasterlain, S., & Cunha, E. (2014). AncesTrees: ancestry estimation with randomized decision trees. *International Journal of Legal Medicine*, *129*(5), 1145–1153. https://doi.org/10.1007/s00414-014-1050-9

Ousley, S., & Jantz, R. (2012). Fordisc 3 and Statistical Methods for Estimating Sex and Ancestry. In D. C. Dirkmaat (Ed.), *A Companion to Forensic Anthropology* (pp. 311–329). Blackwell Publishing.

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, Á., & Gusmao, L. (2009). A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis*, *30*(21), 3682–3690. https://doi.org/10.1002/elps.200900274

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, Á., & Gusmão, L. (2009). Insertion/deletion polymorphisms: A multiplex assay and forensic applications. *Forensic Science International: Genetics Supplement Series*, *2*(1), 513–515. https://doi.org/10.1016/j.fsigss.2009.09.005

Pereira, R., Phillips, C., Pinto, N., Santos, C., dos Santos, S. E., Amorim, A., Carracedo, Á., & Gusmão, L. (2012). Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PloS one*, *7*(1), e29684. https://doi.org/10.1371/journal.pone.0029684

Pereira, V., & Gusmão, L. (2016). Types of Genomes, Sequences and Genetic Markers (Repeats, SNPs, Indels, Haplotypes). In Amorim, A., & Budowle, B. (Eds), *Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation* (pp. 163 – 192). World Scientific Europe Ltd.

Pereira, V., Mogensen, H. S., Børsting, C., & Morling, N. (2017). Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers. *Forensic Science International: Genetics*, *28*, 138–145. https://doi.org/10.1016/j.fsigen.2017.02.013

Phillips, C., Fondevila, M., & Lareau, M. V. (2012). A 34-plex autosomal SNP single base extension assay for ancestry investigations. *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-61779-461-2_8

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., Eduardoff, M., Børsting, C., Johansen, P., Fondevila, M., Morling, N., Schneider, P., EUROFORGEN-NoE Consortium, Carracedo, A., & Lareu, M. V. (2014). Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic science international. Genetics*, *11*, 13–25. https://doi.org/10.1016/j.fsigen.2014.02.012

Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M. C., Ballard, D., Lareu, M. V., Carracedo, A., & SNPforID Consortium (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic science international. Genetics*, *1*(3-4), 273–280. https://doi.org/10.1016/j.fsigen.2007.06.008

Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *Forensic Science International: Genetics*, *18*, 49–65. https://doi.org/10.1016/j.fsigen.2015.05.012

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics, 155*(2), 945.

Prinz, M., Carracedo, A., Mayr, W.R., Morling, N., Parsons, T.J., Sajantila, A., Scheithauer, R., Schmitter, H., Schneider, P.M. DNA Commission of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI), Forensic Science International: Genetics, Volume 1, Issue 1, 2007, Pages 3-12, ISSN 1872-4973, https://doi.org/10.1016/j.fsigen.2006.10.003.

Romanini, C., Romero, M., Salado Puerto, M., Catelli, L., Phillips, C., Pereira, R., Gusmão, L., & Vullo, C. (2015). Ancestry informative markers: inference of ancestry in aged bone samples using an autosomal AIM-Indel multiplex. *Forensic science international. Genetics*, *16*, 58–63. https://doi.org/10.1016/j.fsigen.2014.11.025

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, *1*(6), 0660–0671. https://doi.org/10.1371 /journal.pgen.0010070

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*(5602), 2381–2385. https://doi.org/10.1126/science.1078311

Santos, C., Phillips, C., Oldoni, F., Amigo, J., Fondevila, M., Pereira, R., … Lareu, M. V. (2015). Completion of a worldwide reference panel of samples for an ancestry informative Indel assay. *Forensic Science International: Genetics*, *17*, 75–80. https://doi.org/10.1016/j.fsigen.2015.03.011

Santos, N. P. C., Ribeiro-Rodrigues, E. M., Ribeiro-dos-Santos, Â. K. C., Pereira, R., Gusmão, L., Amorim, A., … Dos Santos, S. E. B. (2010). Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Human Mutation*, *31*(2), 184–190. https://doi.org/10.1002/humu.21159

Scott, G. R., Pilloud, M., Navega, D., Coelho, J., Cunha, E., & Irish, J. (2018). rASUDAS: A New Web-Based Application for Estimating Ancestry from Tooth Morphology. *Forensic Anthropology*, *1*(1), 18–31. https://doi.org/10.5744/fa.2018.0003

Serre, D., & Pääbo, S. (2004). Evidence for gradients of human genetic diversity within and among continents. *Genome Research*, *14*(9), 1679–1685. https://doi.org/10.1101/gr.2529604

Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., & Ferrell, R. E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, *60*(4), 957–964.

Spinney, L. (2008). Eyewitness identification: line-ups on trial. *Nature*, *453*, 442–444.

Spradley, M. K. (2016). Metric Methods for the Biological Profile in Forensic Anthropology: Sex, Ancestry, and Stature. *Academic Forensic Pathology*, *6*(3), 391–399. https://doi.org/10.23907/2016.040

Stoneking, M. (2017). *An Introduction to Molecular Anthropology*. John Wiley & Sons, Inc.

Sun, Q., Jiang, L., Zhang, G., Liu, J., Zhao, L., Zhao, W., & Li, C. (2017). Twenty-seven continental ancestry-informative SNP analysis of bone remains to resolve a forensic

case. *Forensic Sciences Research*, *1790*(May), 1–3. https://doi.org/10.1080/20961790.2017.1306431

Takezaki, N., & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, *144*(1), 389–399.

Tian, C., Gregersen, P. K., & Seldin, M. F. (2008). Accounting for ancestry: Population substructure and genome-wide association studies. *Human Molecular Genetics*, *17*(R2), 143–150. https://doi.org/10.1093/hmg/ddn268

Valle-Silva, G. do, Souza, F. D. N. de, Marcorin, L., Pereira, A. L. E., Carratto, T. M. T., Debortoli, G., … Mendes-Junior, C. T. (2019). Applicability of the SNPforID 52-plex panel for human identification and ancestry evaluation in a Brazilian population sample by next-generation sequencing. *Forensic Science International: Genetics*, *40*, 201–209. https://doi.org/10.1016/j.fsigen.2019.03.003

Walsh, S., Chaitanya, L., Breslin, K., Muralidharan, C., Bronikowska, A., Pospiech, E., … Kayser, M. (2017). Global skin colour prediction from DNA. *Human Genetics*, *136*(7), 847–863. https://doi.org/10.1007/s00439-017-1808-5

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., … Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics*, *7*(1), 98–115. https://doi.org/10.1016/j.fsigen.2012.07.005

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., & Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. *American Journal of Human Genetics*, *71*(4), 854–862. https://doi.org/10.1086/342727

White, T., & Folkens, P. (2005). *The Human Bone Manual*. Academic Press.

Wilson, C., & Wilson, D. (2003). *Written in Blood: A History of Forensic Detection*. Retrieved from https://books.google.pt/books?id=GI-GPwAACAAJ

Yang, N., Li, H., Criswell, L. A., Gregersen, P. K., Alarcon-Riquelme, M. E., Kittles, R., … Seldin, M. F. (2005). Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: Application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Human Genetics*, *118*(3–4), 382–392. https://doi.org/10.1007/s00439-005-0012-1

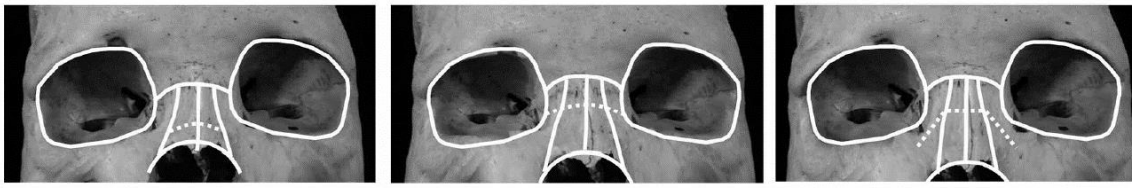
SUPPLEMENTARY MATERIAL

Table S1: cranial measurements used in AncesTrees software.

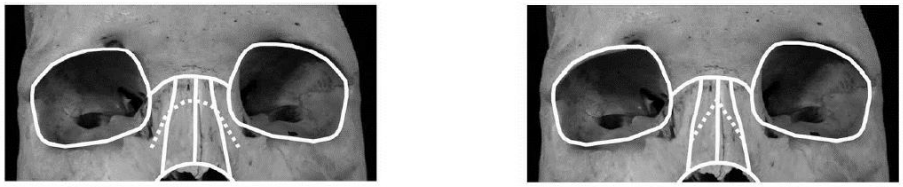
Cranial measurement	**Description**
Glabello-occipital length (GOL)	Greatest length, from the glabellar region, in the median sagittal plane.
Nasio-occipital length (NOL)	Greatest cranial length in the median sagittal plane, measured from nasion.
Basion-bregma height (BBH)	Distance from basion to bregma, as defined.
Maximum cranial breadth (XCB)	The maximum cranial breadth perpendicular to the median sagittal plane, above the supramastoid crests.
Bimaxillary breadth (ZMB)	The breadth across the maxillae, from one zygomazillare (anterior) to the other.
Biasterionic breadth (ASB)	Direct measurement from one asterion to the other.
Bizygomatic breadth (ZYB)	The direct distance between both zigya located at their most lateral points of the zygomatic arches
Biauricular breadth (AUB)	The least exterior breadth across the roots of the zygomatic processes, wherever found.
Palate breadth, external (MAB)	The greatest breadth across the alveolar borders, wherever found, perpendicular to the median plane.
Nasion-prosthion height (NPH)	Upper facial height from nasion to prosthion, as defined.
Basion-prosthion length (BPL)	The facial length from basion to prosthion, as defined.
Basion-nasion length (BNL)	Direct length between basion and nasion.
Nasal height (NLH)	The average height from nasion to the lowest point on the border of the nasal aperture on either side.
Nasal breadth (NLB)	The distance between the anterior edges of the nasal aperture at its widest extent.
Biorbital breadth (EKB)	The breadth across the orbits from ectoconchion to ectoconchion.
Interorbital breadth (DKB)	The breadth across the nasal space from dacryon to dacryon.

Orbit height (OBH)	The height between the upper and lower borders of the left orbit, perpendicular to the long axis of the orbit and bisecting it.
Nasion-bregma chord (FRC)	The frontal chord, or direct distance from nasion to bregma, taken in the midplane and at the external surface.
Bregma-lambda chord (PAC)	The external parietal chord, or direct distance from bregma to lambda, taken in the midplane and at the external surface.
Lambda-opisthion chord (OCC)	The external occipital chord, or direct distance from lambda to opisthion, taken in the midplane and at the external surface.
Orbit breadth (OBB)	Breadth from ectoconchion to dacryon, as defined, approximating the longitudinal axis which bisects the orbit into equal upper and lower parts.
Mastoid Height (MDH)	The direct distance between porion and mastoidale

a. Expressions of nasal bone contour

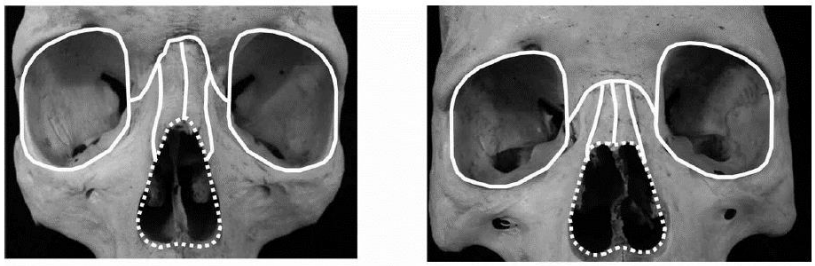


1 – Low (African) 2 – Oval (Asian) 3 – Steep (European)

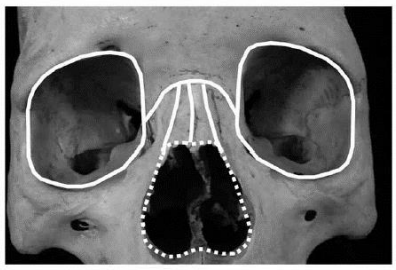


4 – Steep-sided (European) 5 – Triangular, Steepled (European)

b. Expressions of nasal aperture width

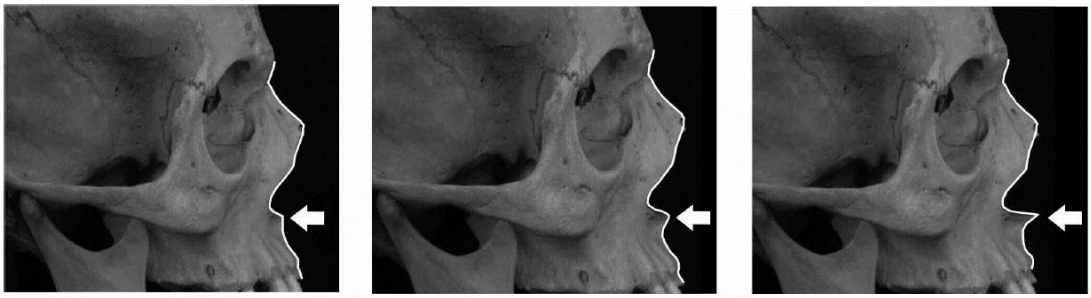


1 – Narrow (European) 2 – Medium (Asian)



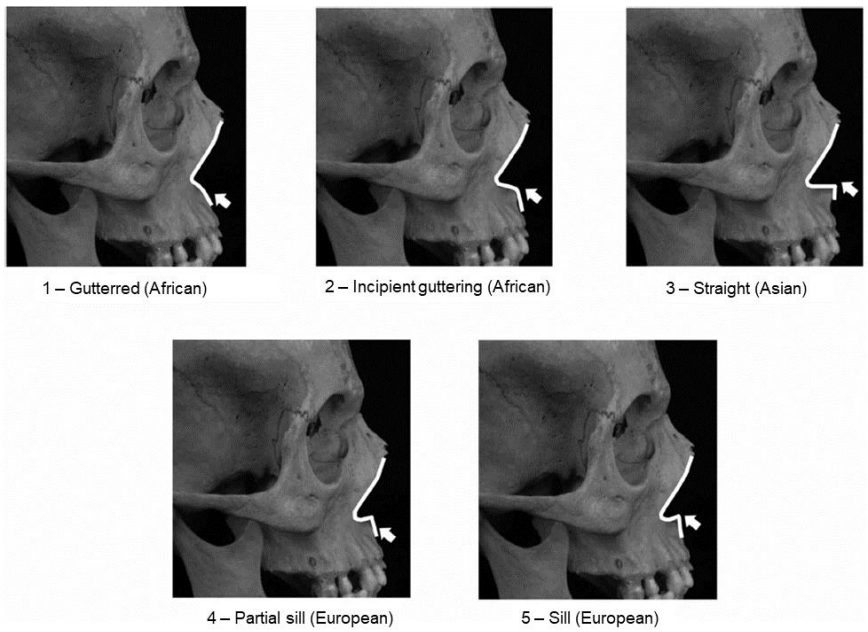
1 – Wide (African)

c. Expressions of anterior nasal spine projection



1 – Rounded, no projection (African) 2 – Medium (Asian) 3 – Long (European)

d. Expressions of the inferior nasal margin



1 – Gutterred (African) 2 – Incipient guttering (African) 3 – Straight (Asian)

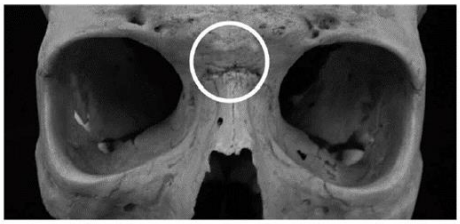
4 – Partial sill (European) 5 – Sill (European)

e. Expressions for nasal overgrowth

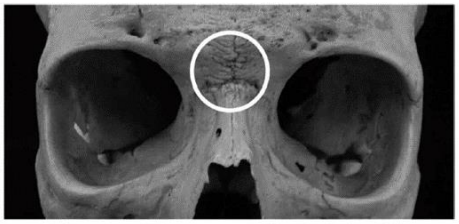


0 – Absent (African/ European) 1 – Present (Asian)

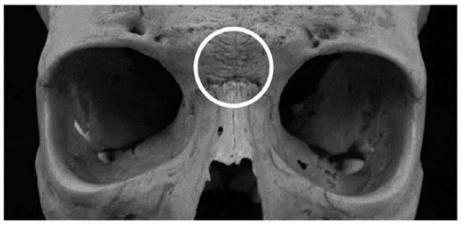
f. Expressions of supranasal suture



0 – Absent (?)

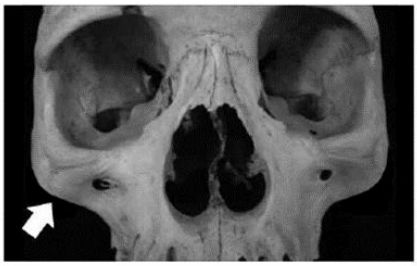


1 – Open (?)

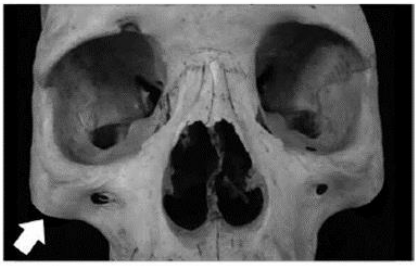


2 – Closed but visible (?)

g. Expressions of malar tubercle



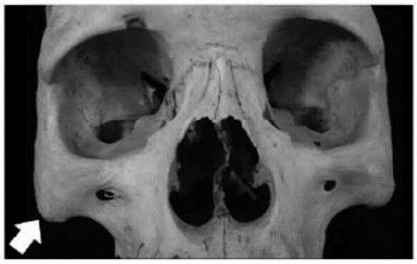
0 – Absent (African)



1 – Incipient (Asian)

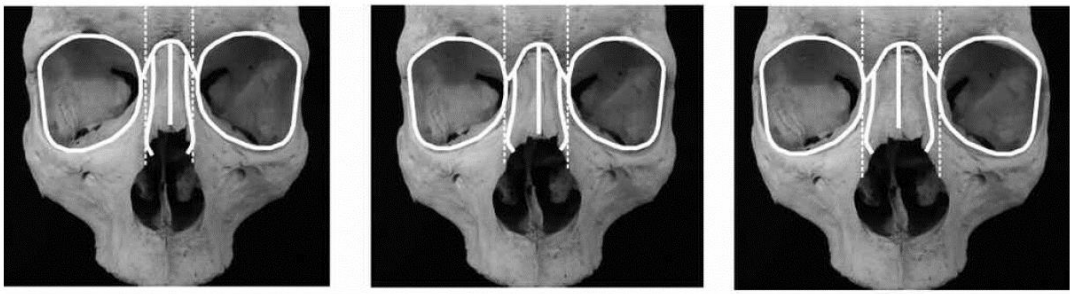


2 – Trace, medium protusion (European)



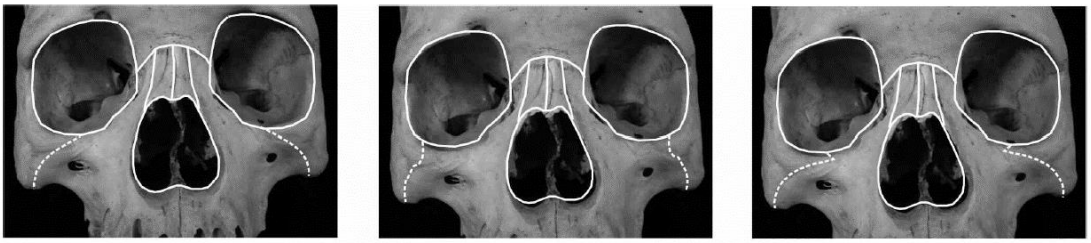
3 – Pronounced (?)

h. Expressions of interorbital breadth



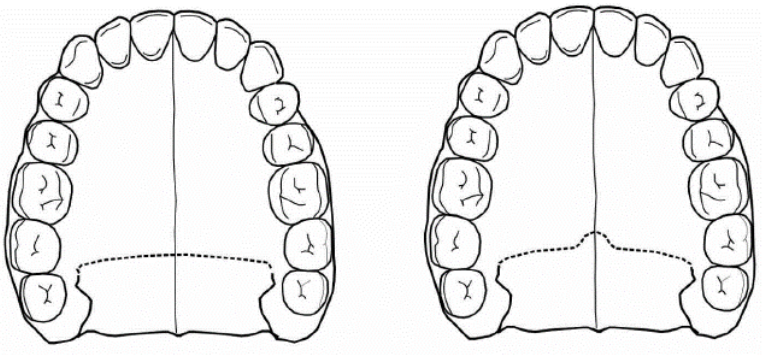
1 – Narrow (European) 2 – Intermediate (Asian) 3 – Wide (African)

i. Expressions of zygomaxillary suture

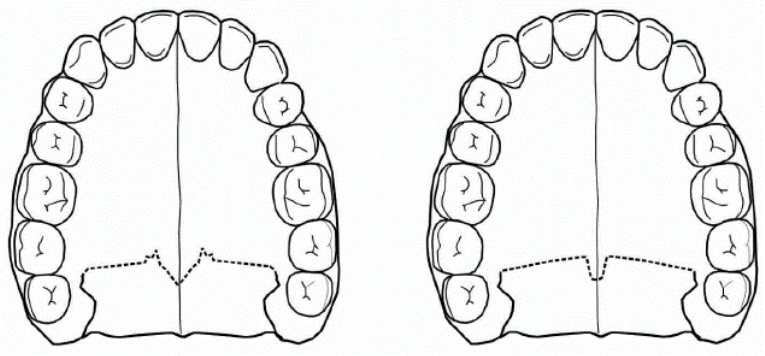


1 – Smooth (African) 2 – Angled (Asian) 1 – S-shaped (European)

j. Expressions of transverse palatine suture shape



1 – Straight (Asian) 2 – Anterior bulging (African)



3 – Anterior and posterior deviation (European) 4 – Posterior symmetrical deviation (European)

k. Expressions of postbregmatic depression

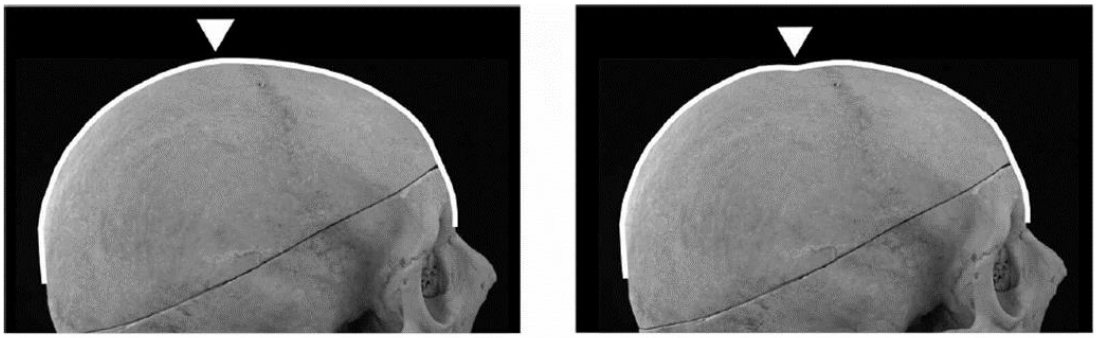


Figure S1.a – k: Different expressions of the eleven non-metric traits according to Hefner (2009). Each expression is associated with a 'stereotypical association' - the ancestral group commonly associated with this trait (AFR, EUR, EAS). Adapted from Iscan and Steyn (2013).

Table S2: AIM-Indels used in the multiplex according to Pereira et al. (2012).

MID	rs number	Chromosome	Position (bp)	Alleles described in dbSNP
MID-1470	rs2307666	11	64729920	-/GTTAC
MID-777	rs1610863	16	6551830	-/GAA
MID-196*	rs16635	6	99789775	-/CAT
MID-881*	rs1610965	5	79746093	-/ACTT
MID-3122	rs35451359	18	45110983	-/ATCT
MID-548	rs140837	6	3708909	-/CT
MID-659	rs1160893	2	224794577	-/CT
MID-2011	rs2308203	2	109401291	-/CTAGA
MID-2929	rs33974167	8	87813725	-/TA
MID-593	rs1160852	6	137345857	-/TT
MID-798	rs1610884	5	56122323	-/GGGAAA
MID-1193	rs2067280	5	89818959	-/AT
MID-1871	rs2308067	7	127291541	-/TT
MID-17	rs4183	3	3192524	-/TAAC
MID-2538	rs3054057	15	86010538	-/AACA
MID-1644	rs2307840	1	36099090	-/GT
MID-3854*	rs60612424	6	84017514	-/TCTA
MID-2275	rs3033053	14	42554496	-/TCAGCAG
MID-94	rs16384	22	42045009	-/AAC
MID-3072	rs34611875	18	67623917	-/GCCCCCA
MID-772*	rs1610859	5	128317275	-/TAG
MID-2313	rs3045215	1	234740917	-/ATTATAACT
MID-397	rs25621	6	139858158	-/TTCT
MID-1636	rs2307832	1	55590789	-/AA
MID-51	rs16343	4	17635560	-/TTTAT
MID-2431	rs3031979	8	73501951	-/ATTG
MID-2264	rs34122827	13	63778778	-/AAGT
MID-2256	rs133052	22	41042364	-/CAT
MID-128	rs6490	12	108127168	-/ATT
MID-15	rs4181	2	42577803	-/AAATACACAC
MID-2241	rs3030826	6	67176774	-/GTCCAATA
MID-419	rs140708	6	170720016	-/AATGGCA
MID-943	rs1611026	5	82545545	-/TGAT
MID-159	rs16438	20	25278470	-/CCCCA
MID-2005	rs2308161	10	69800909	-/AACAAT
MID-250	rs16687	7	83887882	-/CA
MID-1802	rs2307998	5	7814345	-/GGA
MID-1607	rs2307803	3	108981031	-/TG
MID-1734	rs2307930	6	84476378	-/CCAT
MID-406	rs25630	6	14734341	-/AG
MID-1386	rs2307582	1	247768775	-/AAACTATTCATTTTTCACCCT
MID-1726	rs2307922	1	39896964	-/CAAGAACTATAAT/CACTATCTATTAT
MID-3626*	rs11267926	15	45526069	-/AATATAATTTCTCCA
MID-360	rs25584	12	112145217	-/AA
MID-1603	rs2307799	5	70828427	-/TTGT
MID-2719*	rs34541393	20	30701405	-/AACT

*Primers used in the 6plex-reinforcement.

Table S3: DNA concentration (ng/µL) and respective number of markers amplified, missing data observed, and number of replicas made for each individual. Sorted by increased value of DNA concentration.

CEI_ID	Concentration (ng/µL)	Amplified Markers	Missing data (%)	Number of replicas
290	0,0003	46	0,0	2
42	0,0010	17	63,0	2
130	0,0016	35	23,9	2
127	0,0031	20	56,5	3
267	0,0031	32	30,4	3
236	0,0034	16	65,2	3
104	0,0034	31	32,6	3
278	0,0036	40	13,0	3
224	0,0038	38	17,4	3
3	0,0040	15	67,4	5
120	0,0043	23	50,0	3
232	0,0044	37	19,6	3
202	0,0045	45	2,2	3
203	0,0064	25	45,7	3
62	0,0068	43	6,5	5
223	0,0072	37	19,6	3
128	0,0076	32	30,4	2
156	0,0078	23	50,0	3
249	0,0083	29	37,0	2
241	0,0085	33	28,3	3
112	0,0091	27	41,3	3
164	0,0093	43	6,5	2
135	0,0109	36	21,7	3
185	0,0111	42	8,7	3
108	0,0112	41	10,9	3
160	0,0117	42	8,7	2
209	0,0147	46	0,0	3
111	0,0153	44	4,3	2
103	0,0168	42	8,7	3
281	0,0171	40	13,0	2
245	0,0174	44	4,3	2
253	0,0180	36	21,7	2
239	0,0184	38	17,4	2
275	0,0188	46	0,0	2
12	0,0190	43	6,5	2
292	0,0200	44	4,3	2
262	0,0200	44	4,3	2
123	0,0218	43	6,5	2
87	0,0234	45	2,2	5
280	0,0240	44	4,3	2

183	0,0263	46	0,0	3
298	0,0286	45	2,2	3
179	0,0297	43	6,5	3
99	0,0363	28	39,1	5
100	0,0380	44	4,3	3
122	0,0382	44	4,3	2
296	0,0387	22	52,2	3
254	0,0434	46	0,0	2
299	0,0452	41	10,9	2
14	0,0522	38	17,4	2
302	0,0798	45	2,2	2
133	0,0961	46	0,0	3
212	0,1437	34	26,1	3
242	0,1473	33	28,3	3
194	0,1522	46	0,0	3
247	0,1791	45	2,2	3

Table S4: STRUCTURE cluster membership proportions for K = 2, 3, and 4.

	K=2		K=3			K=4			
	AFR	**EUR**	**AFR**	**EUR**	**EAS**	**AFR**	**EUR**	**EAS**	**NAM**
CEI_3	0.213	0.787	0.064	0.694	0.242	0.043	0.527	0.196	0.234
CEI_12	0.173	0.827	0.081	0.887	0.032	0.056	0.781	0.015	0.148
CEI_14	0.594	0.406	0.607	0.326	0.067	0.617	0.290	0.041	0.053
CEI_42	0.369	0.631	0.259	0.632	0.110	0.260	0.541	0.070	0.129
CEI_62	0.180	0.820	0.065	0.820	0.115	0.040	0.855	0.075	0.030
CEI_87	0.170	0.830	0.084	0.863	0.054	0.063	0.882	0.032	0.023
CEI_99	0.149	0.851	0.040	0.847	0.113	0.024	0.799	0.063	0.113
CEI_100	0.277	0.723	0.191	0.765	0.044	0.161	0.795	0.025	0.018
CEI_103	0.108	0.892	0.030	0.918	0.052	0.017	0.895	0.027	0.061
CEI_104	0.126	0.874	0.035	0.924	0.041	0.019	0.929	0.023	0.029
CEI_108	0.073	0.927	0.017	0.956	0.027	0.010	0.964	0.015	0.011
CEI_111	0.130	0.870	0.039	0.912	0.048	0.023	0.933	0.029	0.015
CEI_112	0.265	0.735	0.156	0.775	0.069	0.117	0.807	0.041	0.035
CEI_120	0.196	0.804	0.090	0.845	0.066	0.062	0.869	0.038	0.031
CEI_122	0.140	0.860	0.047	0.931	0.022	0.027	0.936	0.012	0.024
CEI_123	0.157	0.843	0.066	0.895	0.039	0.041	0.895	0.020	0.043
CEI_127	0.141	0.859	0.039	0.898	0.063	0.023	0.835	0.036	0.107
CEI_128	0.112	0.888	0.029	0.902	0.069	0.016	0.914	0.039	0.031
CEI_130	0.117	0.883	0.032	0.937	0.031	0.018	0.935	0.017	0.030
CEI_133	0.141	0.859	0.039	0.805	0.156	0.021	0.555	0.036	0.389
CEI_135	0.144	0.856	0.047	0.931	0.021	0.028	0.946	0.012	0.014
CEI_156	0.161	0.839	0.048	0.879	0.072	0.027	0.892	0.040	0.041
CEI_160	0.147	0.853	0.050	0.904	0.046	0.030	0.932	0.025	0.013
CEI_164	0.175	0.825	0.074	0.889	0.037	0.049	0.896	0.020	0.035
CEI_179	0.897	0.103	0.954	0.027	0.019	0.963	0.015	0.011	0.012
CEI_183	0.179	0.821	0.095	0.870	0.035	0.068	0.881	0.019	0.033
CEI_185	0.189	0.811	0.094	0.857	0.049	0.065	0.867	0.030	0.039
CEI_194	0.112	0.888	0.029	0.898	0.073	0.016	0.879	0.037	0.067
CEI_202	0.186	0.814	0.064	0.813	0.122	0.041	0.824	0.076	0.059
CEI_203	0.379	0.621	0.264	0.505	0.231	0.285	0.407	0.139	0.169
CEI_209	0.216	0.784	0.152	0.807	0.041	0.142	0.822	0.023	0.014
CEI_212	0.086	0.914	0.021	0.940	0.038	0.011	0.944	0.022	0.023
CEI_223	0.127	0.873	0.037	0.932	0.031	0.021	0.945	0.018	0.016
CEI_224	0.109	0.891	0.027	0.918	0.054	0.016	0.924	0.031	0.029
CEI_232	0.124	0.876	0.035	0.939	0.026	0.021	0.950	0.015	0.015
CEI_236	0.139	0.861	0.037	0.752	0.211	0.020	0.792	0.157	0.031
CEI_239	0.327	0.673	0.240	0.633	0.127	0.228	0.665	0.089	0.017
CEI_241	0.216	0.784	0.092	0.693	0.214	0.063	0.722	0.147	0.068
CEI_242	0.285	0.715	0.208	0.732	0.060	0.189	0.757	0.036	0.018
CEI_245	0.073	0.927	0.017	0.949	0.034	0.010	0.958	0.018	0.014
CEI_247	0.127	0.873	0.038	0.924	0.038	0.022	0.886	0.020	0.073

CEI_249	0.162	0.838	0.059	0.865	0.076	0.036	0.885	0.043	0.036
CEI_253	0.176	0.824	0.070	0.774	0.155	0.047	0.752	0.085	0.116
CEI_254	0.311	0.689	0.253	0.647	0.101	0.250	0.645	0.064	0.041
CEI_262	0.117	0.883	0.033	0.913	0.054	0.019	0.924	0.034	0.023
CEI_267	0.097	0.903	0.025	0.938	0.037	0.014	0.946	0.021	0.019
CEI_275	0.144	0.856	0.052	0.923	0.025	0.032	0.942	0.013	0.013
CEI_278	0.147	0.853	0.057	0.871	0.072	0.039	0.868	0.042	0.051
CEI_280	0.213	0.787	0.138	0.777	0.085	0.120	0.790	0.056	0.033
CEI_281	0.625	0.375	0.578	0.134	0.288	0.591	0.104	0.276	0.029
CEI_290	0.202	0.798	0.144	0.808	0.049	0.132	0.787	0.025	0.056
CEI_292	0.097	0.903	0.025	0.938	0.037	0.014	0.913	0.020	0.053
CEI_296	0.454	0.546	0.400	0.454	0.146	0.472	0.203	0.087	0.238
CEI_298	0.090	0.910	0.023	0.934	0.043	0.013	0.916	0.023	0.048
CEI_299	0.128	0.872	0.040	0.905	0.055	0.023	0.893	0.029	0.055
CEI_302	0.097	0.903	0.025	0.948	0.027	0.015	0.960	0.016	0.009

Table S5: Values of -log likelihood calculated by Snipper for each main cluster (lower is better). Probability of finding each profile, in each cluster, was calculated using the formulae EXP(-log(lk)). The Lowest Likelihood Ratio was calculated by dividing the higher probability by the second higher.

CEI_ID	African -log(lk)	European -log(lk)	East Asian -log(lk)	American -log(lk)	Oceanian -log(lk)		Pr(AFR)	Pr(EUR)	Pr(EAS)	Pr(NAM)	Pr(OCE)	Lowest LR
CEI_3	27.30	10.65	13.79	15.85	22.16		1.40E-12	2.37E-05	1.02E-06	1.31E-07	2.38E-10	23.19
CEI_12	78.68	40.96	77.12	69.95	86.78		6.72E-35	1.62E-18	3.21E-34	4.19E-31	2.04E-38	3874353786295.17
CEI_14	40.45	48.53	69.81	77.40	64.29		2.71E-18	8.43E-22	4.80E-31	2.43E-34	1.20E-28	3215.06
CEI_42	23.51	16.62	23.73	24.98	22.24		6.18E-11	6.04E-08	4.95E-11	1.42E-11	2.19E-10	275.42
CEI_62	76.21	36.72	60.92	73.55	71.61		8.02E-34	1.13E-16	3.49E-27	1.14E-32	7.92E-32	32269875986.30
CEI_87	71.41	34.69	67.32	73.29	64.90		9.72E-32	8.61E-16	5.80E-30	1.48E-32	6.51E-29	13224602017164.20
CEI_99	46.32	18.04	29.71	25.96	31.83		7.62E-21	1.46E-08	1.25E-13	5.34E-12	1.49E-14	2735.04
CEI_100	68.97	41.25	75.98	92.21	76.25		1.11E-30	1.21E-18	1.01E-33	8.98E-41	7.71E-34	1090982737484.16
CEI_103	75.12	28.01	56.96	52.02	53.33		2.38E-33	6.82E-13	1.82E-25	2.55E-23	6.88E-24	26691206405.58
CEI_104	48.12	18.12	38.90	45.63	38.29		1.26E-21	1.35E-08	1.28E-17	1.53E-20	2.34E-17	574666252.61
CEI_108	85.63	25.40	66.11	75.64	57.23		6.49E-38	9.31E-12	1.94E-29	1.41E-33	1.39E-25	66805063867654.60
CEI_111	71.86	29.12	58.99	67.99	52.43		6.19E-32	2.26E-13	2.41E-26	2.96E-30	1.70E-23	13292954566.84
CEI_112	31.86	18.72	32.82	38.66	36.82		1.46E-14	7.38E-09	5.60E-15	1.63E-17	1.03E-16	504488.34
CEI_120	37.72	17.43	30.51	36.64	26.53		4.15E-17	2.69E-08	5.64E-14	1.22E-16	3.02E-12	8920.44
CEI_122	84.02	36.41	85.08	78.64	78.29		3.25E-37	1.55E-16	1.12E-37	7.04E-35	1.00E-34	1541519049609530000.00
CEI_123	75.12	36.61	75.14	70.77	69.28		2.39E-33	1.26E-16	2.33E-33	1.84E-31	8.19E-31	153419942550404.00
CEI_127	37.59	11.73	23.07	21.66	35.71		4.73E-17	8.08E-06	9.61E-11	3.91E-10	3.08E-16	20648.54
CEI_128	62.78	21.74	41.99	45.41	40.75		5.46E-28	3.61E-10	5.78E-19	1.90E-20	2.01E-18	179574370.42
CEI_130	62.88	24.14	52.80	53.56	57.04		4.94E-28	3.27E-11	1.17E-23	5.51E-24	1.69E-25	2794571775820.59
CEI_133	87.79	42.44	62.14	48.21	80.39		7.48E-39	3.70E-19	1.03E-27	1.15E-21	1.22E-35	321.40
CEI_135	56.01	23.29	58.86	65.40	46.67		4.75E-25	7.67E-11	2.74E-26	3.96E-29	5.39E-21	14246826256.31
CEI_156	35.80	13.00	26.30	31.43	34.01		2.84E-16	2.25E-06	3.78E-12	2.24E-14	1.70E-15	594454.82
CEI_160	70.98	35.44	63.74	83.05	53.10		1.50E-31	4.07E-16	2.09E-28	8.57E-37	8.69E-24	46837747.21
CEI_164	72.86	34.72	67.99	73.03	74.15		2.29E-32	8.31E-16	2.96E-30	1.91E-32	6.30E-33	280613670155687.00
CEI_179	28.15	70.16	88.44	91.36	79.72		5.93E-13	3.37E-31	3.91E-39	2.11E-40	2.38E-35	1756227997850570000.00
CEI_183	83.40	43.03	84.41	88.18	71.24		6.04E-37	2.06E-19	2.19E-37	5.06E-39	1.16E-31	1782967095378.19
CEI_185	73.57	40.35	76.97	78.42	77.08		1.12E-32	2.99E-18	3.73E-34	8.75E-35	3.35E-34	267328672383361.00
CEI_194	84.34	33.30	61.62	57.32	59.30		2.35E-37	3.46E-15	1.74E-27	1.28E-25	1.77E-26	26956761678.68
CEI_202	67.28	36.21	52.74	60.94	38.76		6.05E-30	1.88E-16	1.25E-23	3.43E-27	1.47E-17	12.75
CEI_203	37.90	30.31	38.81	39.62	45.85		3.47E-17	6.88E-14	1.40E-17	6.22E-18	1.22E-20	1982.27

CEI_209	80.38	39.75	71.59	89.13	56.96		1.24E-35	5.47E-18	8.09E-32	1.95E-39	1.83E-25	29862079.82
CEI_212	66.14	20.61	47.55	50.56	56.63		1.89E-29	1.12E-09	2.24E-21	1.11E-22	2.55E-25	499413376711.98
CEI_223	71.99	31.95	64.06	77.40	65.52		5.44E-32	1.34E-14	1.51E-28	2.44E-34	3.52E-29	88746046782835.60
CEI_224	69.20	27.36	52.98	56.47	46.99		8.84E-31	1.32E-12	9.75E-24	3.00E-25	3.91E-21	337069601.62
CEI_232	56.42	22.67	54.00	65.22	48.32		3.15E-25	1.43E-10	3.54E-24	4.74E-29	1.03E-21	138467295345.49
CEI_236	33.19	7.82	10.97	22.64	18.12		3.85E-15	4.03E-04	1.73E-05	1.47E-10	1.35E-08	23.32
CEI_239	60.98	40.44	56.20	78.13	54.53		3.28E-27	2.74E-18	3.92E-25	1.17E-34	2.08E-24	1316121.88
CEI_241	55.14	30.67	40.69	52.57	46.09		1.13E-24	4.77E-14	2.14E-18	1.48E-23	9.58E-21	22283.46
CEI_242	42.35	30.30	52.80	71.63	65.80		4.05E-19	6.94E-14	1.17E-23	7.76E-32	2.66E-29	171424.81
CEI_245	91.73	27.87	65.50	76.29	56.48		1.45E-40	7.90E-13	3.56E-29	7.35E-34	2.96E-25	2663334433187.36
CEI_247	72.82	30.03	63.13	52.80	58.87		2.37E-32	9.06E-14	3.83E-28	1.17E-23	2.70E-26	7736382205.77
CEI_249	58.98	26.37	47.90	54.75	51.84		2.42E-26	3.52E-12	1.58E-21	1.67E-24	3.07E-23	2227843691.21
CEI_253	61.98	30.58	46.31	48.37	55.26		1.21E-27	5.22E-14	7.72E-21	9.86E-22	1.01E-24	6769244.21
CEI_254	67.83	46.25	67.10	73.76	77.24		3.48E-30	8.17E-21	7.25E-30	9.23E-33	2.85E-34	1127309886.45
CEI_262	76.33	31.91	61.87	69.56	42.30		7.10E-34	1.38E-14	1.36E-27	6.15E-31	4.25E-19	32500.15
CEI_267	78.27	26.73	56.15	63.06	41.25		1.02E-34	2.46E-12	4.11E-25	4.12E-28	1.22E-18	2015544.62
CEI_275	84.94	39.09	80.54	92.20	73.88		1.29E-37	1.06E-17	1.05E-35	9.10E-41	8.23E-33	1285983255455230.00
CEI_278	67.03	29.38	49.22	53.16	40.56		7.78E-30	1.74E-13	4.19E-22	8.21E-24	2.43E-18	71617.88
CEI_280	73.49	39.46	64.92	72.78	61.51		1.21E-32	7.31E-18	6.40E-29	2.47E-32	1.93E-27	3796707193.50
CEI_281	44.95	59.53	66.26	85.42	49.16		3.02E-20	1.40E-26	1.68E-29	7.97E-38	4.46E-22	67.64
CEI_290	96.48	49.85	84.27	82.44	79.16		1.25E-42	2.25E-22	2.52E-37	1.57E-36	4.18E-35	5386408757337.54
CEI_292	81.66	29.78	65.02	57.74	58.77		3.42E-36	1.17E-13	5.80E-29	8.39E-26	3.00E-26	1392608885575.52
CEI_296	32.54	25.94	31.45	33.62	29.32		7.40E-15	5.43E-12	2.20E-14	2.49E-15	1.85E-13	29.36
CEI_298	98.16	38.71	73.34	74.68	63.27		2.35E-43	1.54E-17	1.41E-32	3.67E-33	3.32E-28	46216374562.54
CEI_299	75.15	31.79	56.83	61.56	64.50		2.31E-33	1.56E-14	2.08E-25	1.85E-27	9.75E-29	74689099892.68
CEI_302	80.71	29.30	69.35	91.13	65.81		8.89E-36	1.88E-13	7.60E-31	2.64E-40	2.62E-29	7195268124871960.00

Table S6: Ancestry assessment through anthropology. AncesTrees software was used to calculate probabilities with the metric approach, while hefneR was used to calculate probabilities for the non-metric approach.

CEI_ID	Metric Approach	Probability (%)	Non-metric Approach	Probability (%)
3	European	84.2	European	99.80
14	Asian	69.8	African	65.90
87	Native American	51.6	European	99.70
99	European	68.8	European	79.00
100	European	89.0	European	99.10
104	European	53.0	European	96.10
108	European	71.8	European	95.30
111	European	92.4	European	99.60
112	European	93.4	European	93.80
120	European	96.8	European	98.40
122	European	56.4	European	99.80
123	European	87.8	European	93.40
127	European	95.0	European	99.30
128	European	88.4	European	99.20
130	European	86.6	European	96.60
133	European	95.0	European	99.40
135	European	89.2	European	92.50
156	European	73.8	Native American	62.20
160	European	98.0	European	99.90
179	African	79.6	African	56.50
183	European	73.4	European	95.10
185	European	98.6	European	99.95
194	European	84.8	European	98.60
202	European	70.8	European	99.90
203	European	62.6	European	44.60
209	European	76.4	European	86.60
212	European	82.2	European	99.80
223	European	99.8	European	99.90
224	European	62.4	European	94.90
232	European	74.2	European	99.70
236	European	89.6	European	99.80
239	European	78.8	European	99.50
241	European	75.8	European	77.10
242	European	73.4	European	99.60
245	European	93.8	European	99.60
247	European	81.0	European	99.40
249	European	85.4	European	99.40
253	European	76.2	European	99.50
254	European	69.4	European	66.10
262	European	98.4	European	99.90

267	European	68.2	European	86.50
275	European	65.0	European	98.60
278	European	98.8	European	99.80
280	European	93.4	European	77.70
281	Asian	53.0	European	31.80
290	Native American	56.8	European	92.40
292	European	68.2	European	94.10
296	European	75.4	European	97.70
298	European	72.0	European	75.30
299	European	86.2	European	99.90