This is the **published version** of the bachelor thesis:

Asensi Martínez, Elena; Casillas Viladerrams, Sònia, dir. Unraveling the roots of racial prejudice. 2023. (Grau en Genètica)

This version is available at https://ddd.uab.cat/record/299026

under the terms of the BY-SA license

Unraveling the roots of racial prejudice

Elena Asensi Martínez

Final degree project

Supervisor: Dra. Sònia Casillas Viladerrams

Universitat Autònoma de Barcelona

Barcelona, 31st May 2023

TABLE OF CONTENTS

1. INTRODUCTION

Throughout human history, categorizing populations into distinct groups has been controversial within society due to the social constructs and ethical implications underlying such classifications. The utilization of certain terms to categorize populations has been a subject of debate and controversy too, being “ancestry” or “ethnicity” more commonly used and widely accepted terminologies in comparison to “race”[1], which is frequently connected to discriminatory ideologies. Beyond the ethical and moral complexities of dividing civilization into different classes, science has played a role through the ages in determining the factors considered in establishing these classifications.

The concept of “race” has been and is still used in so many scientific fields to classify humans based on physical traits such as skin color, hair texture, and facial features. The categorization of populations based on “race” was a product of the *Enlightenment Movement,* which emerged in the late 17th century, and which saw the rise of many influential thinkers, writers, and philosophers who tried to understand the world and address social issues, challenging traditional religious and monarchical systems and having science as one of the guiding principles for societal progress.[2] These classifications varied among different thinkers, including the astrologer professor *James Bradley* who proposed a system dividing individuals into four groups based on their capillary type (“white people with beards” were Europeans, “white people without beards” were Amerindians or Indian-Americans, “black people with straight hair” were people of Abyssinia and “black people with curly hair” were the rest) or the naturalist *Lamarck* who considered six different groups in his book *“Philosophie zoologique”* (the Caucasian, the Mongolic, the Malaysian, the Hyperboric, the American, and the Ethiopian or black).[3]

All the categorizations among humans were based on the most visible morphological traits however, the first physical trait to be established as a reliable criterion was craniometry which was pioneered by the natural scientist *Samuel George Morton* who believed that variations in brain size among different social groups could be used to classify them.

In the 19th century, as genetics began to emerge, other justifications for classifying human populations into distinct groups surfaced. Numerous studies were conducted using various genetic markers, including repetitive DNA units and SNPs, that were linked to notable phenotypes (such as the *EDAR-V370A* allele for hair type or the *Duffy-null* allele for skin colour). The greater the number of markers used, the more subdivisions that were established. This highlighted the challenging problem of categorizing populations into distinct groups and further sparked debate on how science should approach this.

After a lengthy debate, there is now a widespread agreement on how to classify human populations based on their continents and origins. The consensus recognizes five major groups: Africans, Oriental Asians, Europeans, Native Americans, and Aboriginal Australians[4] (*Figure 1*).

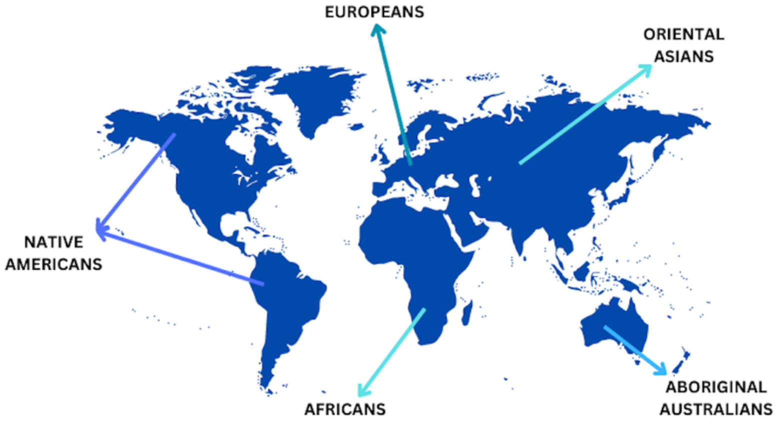


Figure 1: population classification consensus. (Own creation)

1.1 OBJECTIVES

The **objectives** of this final degree thesis are:

- **Objective 1:** to explain the historical and cultural origins of the term "race" by identifying the philosophers and scientists who have influenced its definition while discussing the various criteria used to categorize populations into distinct groups, and whether it has been more of a social construct than a scientific one.
- **Objective 2:** to examine how the introduction of genetics has altered the concept of "race" and impacted established categorizations:

a. To describe the significance of genetic variation in both inter-population and intra-population contexts, and analyze the genetic processes that have influenced the differentiation of distinct populations.
b. To conduct a PCA bioinformatic analysis to investigate the existence of genetic differences among populations and compare the findings to prior classifications.

- **Objective 3:** discuss the hazards of genetic determinism by illustrating how it has been employed to justify discriminatory practices and explore the ethical ramifications of categorizing populations into distinct groups. Additionally, explain the appropriate terminology to use when referring to various human populations.

2. POPULATION DIVISION BEFORE THE ADVENT OF GENETICS

Across the ages, the categorization of populations based on "racial ideology" has been a topic of interest for humanity. The study of human "races", also known as "**raciology**" emerged two centuries ago as an attempt by scientists to categorize each human population into distinct groups. However, despite the longevity of this scientific field, the objectivity of its classifications was never empirically demonstrated. Additionally, the criteria proposed for dividing populations into groups varied among different thinkers, philosophers, and scientists, and were often based on racist concepts rather than anything else.[3]

The issue with this classification system throughout history is that the initial attempts were often based on a **hierarchical system** rather than a descriptive one. This resulted in social repercussions and racial prejudice. In fact, "raciology" was viewed as a means for racism to be linked to **biological determinism**, which assumes that social inequalities are not a social construct, but rather an unchangeable biological one.[4]

2.1 HISTORICAL CONTEXT

We must understand that population division is an event that transcends the beginnings of human civilization. As early as 4000-5300 BC, **Egyptian drawings** highlighted these distinctions, as individuals from each village were distinguished by their clothing and physical characteristics, such as skin colour, nose shape, and hair texture.

A clear example of an established system of division is the **caste system** in India (1500 BCE) which was one of the first human attempts to categorize the population into different groups.[5] This was already a preamble to everything that society would seek throughout its history: a population division that would try to justify different treatment or consideration for different individuals based, mostly, on their physical appearance.

It is important to consider that during the **Middle Ages** (5th to 15th century), there was limited awareness of human diversity due to **geographical barriers** that prevented people from observing inter-population differences, resulting in a greater emphasis on intra-population variation. The limited understanding of human diversity started to diminish during the 15th and 16th centuries, with the onset of transoceanic travels covering thousands of kilometers. These journeys led to a gradual appreciation of the broader physical and cultural range of humans.

This discovery of human variation is an event that has taken a lot of time to accomplish and which is believed to have finished with the *Archbold* expedition organized by the *American Museum of Natural History* in 1938 where, in the river *Balim* of *New Guinea,* a population of 50000 Papua was discovered, and with the first contact with Andaman natives who live in *Sentinel Island* in the Indic Ocean in 1991.[3]

In 1492, Europeans were able to broaden their knowledge of the world's diversity when they discovered new lands in Asia, Africa, and America, and realized they were inhabited. However, it was not until the 17th century, during the *Enlightenment movement*, that the scientific method was introduced to provide a biological justification for racial prejudice, and human classification based on variability was approached in a supposedly "objective" manner. While the ***Enlightenment movement*** was influenced by egalitarian ideals, with many thinkers advocating for social equality and the abolition of differences such as slavery, others sought to promote their racist visions of social hierarchy.

During the 19th century, ***European Imperialism*** extended to various regions of the world, eroding any shred of equality and consolidating social hierarchies. The concept of "evolution" was also introduced to justify racist ideologies, using concepts such as "survival of the fittest" and the "fight for survival" to divide humanity. This led to the idea of innate racial superiority, which resulted in devastating events in human history, including World War II when Nazis used the guise of applied biology to justify their social policies.

2.2 CATEGORIZATIONS AND SOCIAL HIERARCHY

As discussed in the preceding chapter, the creation of a social hierarchy arose from the attempt to categorize populations into distinct groups, often referred to as "races" (although this terminology is evolving to avoid discriminatory language).

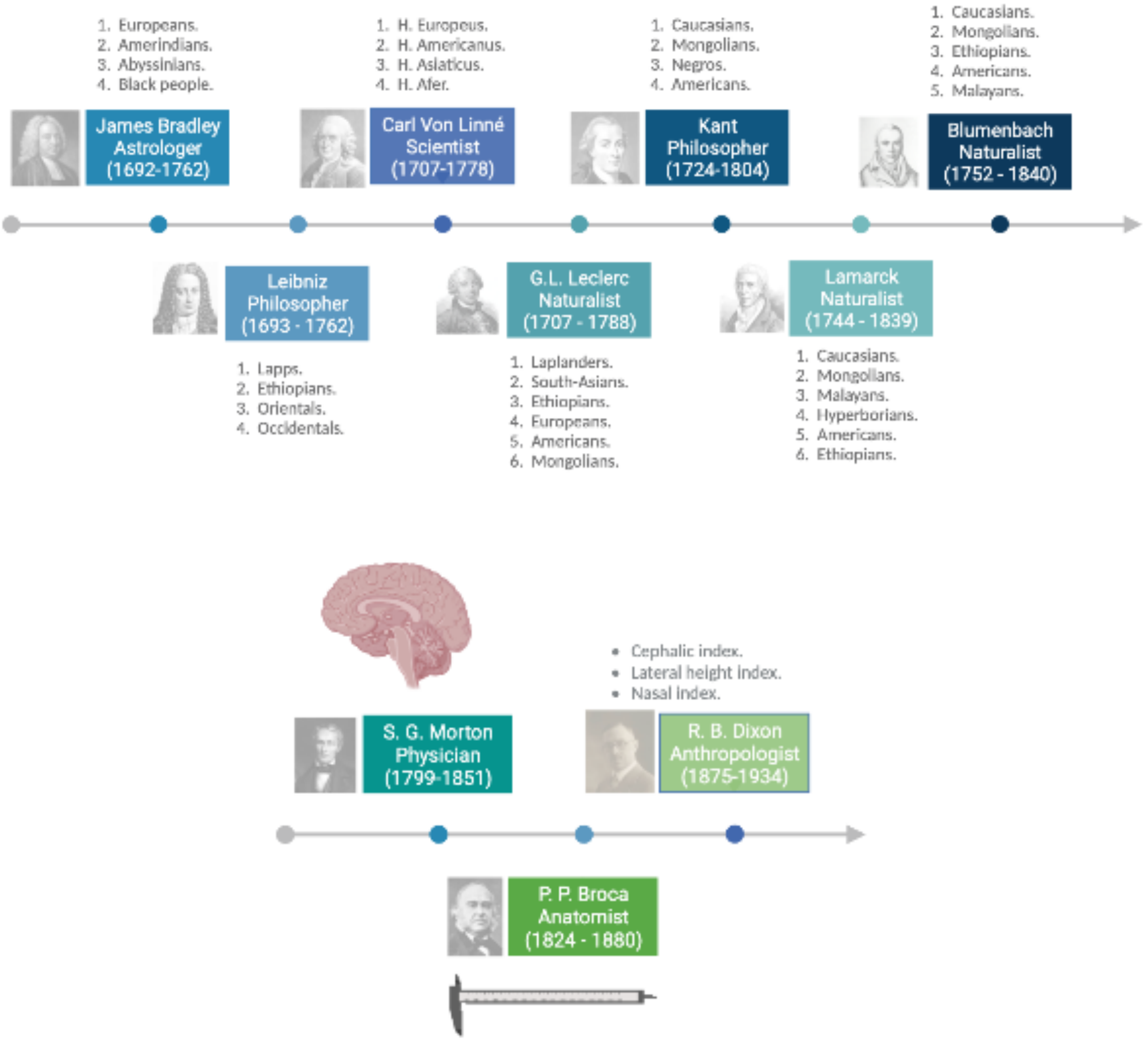


Figure 2: timeline reviewing some of the classifications of the human population before the introduction of genetics. Craniometry from Morton to Dixon. (Own creation)

The timeline presented in *Figure 2* shows an overview of how human populations were classified across different periods before the emergence of genetics. We can observe that there was never a unanimous agreement among scholars regarding these classifications and this issue has not been resolved yet.

It is relevant to highlight what *Carl Von Linné* and *Georges Louis Leclerc* brought to the subject. *Carl Von Linné* added the Latin nomenclature to all the species and included the humans in the primate species (*Homo Sapiens*) and based his classification on physical and character traits believing in the fixity of species (they remained constant and unchanging over time). A contemporary to *Linné*, *Georges Louis Leclerc*, who was also known as *Comte de Buffon*, disagreed with *Linné* and believed that species were subject to environmental influences that could lead to their modification over time, this is the reason why he proposed a different classification.

Moreover, *Haeckel*, a naturalist, developed the "**recapitulation theory**", which posited that the embryonic development of humans passed through all the stages, with each "race" representing a distinct stage. This theory suggested that the most marginalized groups in society were at an earlier stage of development and therefore, contributed to the social hierarchy. Some other scientists, such as the naturalist *Darwin*, at some point even believed that different human "races" were different species which lead to the debate between polygenists and monogenists[6]:

- **Polygenism:**
    - Monophyletic polygenism: humanity would have first appeared among several individuals, whose progeny gradually spread worldwide through emigration.
    - Polyphyletic polygenism: human species arose through separate evolutionary lines in several different places at different times.
- **Monogenism:** takes the evidence that the whole human "race" is descended from a single couple or single individual.

Before the introduction of genetics, the earliest racial classifications were exclusively based on morphological traits. However, the criterion which first saw a scientific consolidation was **craniometry**. This method was attractive to racial theorists because the cranium contains the brain, and they believed that the intellectuality of different “races” could be discerned by differences in their cranium and brain. The first craniometric studies were based on the observation of some cranial traits. However, as some scientists, like *Broca*, began to create craniometric tools, some metric measurements were established and the cranial variation was quantified. In *Table 1* we can see an example of one of the classifications performed using craniometry.

Table 1: example of population classification according to craniometric measures.

	Asians	**Europeans**	**Africans**
Skull shape	Wide	Medium	Large
Lateral profile of the skull	Tall and globose	Tall and rounded	Variable
Nasal shape	Medium	Narrow	Straight
Nasal profile	Concave	Straight	Straight/concave
Facial projection	Moderate	Low	High
Cheekbone shape	Projected	Not projected	Not projected
Chin projection	Moderate	Prominent	Low
Chin shape	Medium	Bilateral	Medium

3. POPULATION DIVISION UNDER THE PRISM OF GENETICS

At the end of the 19th and the beginning of the 20th century, with the arrival of genetics, the perception of human variation became more complex. It was believed that genetics would bring a solution to the subject nonetheless, it only contributed to generating more debate.

Before genetics, human differences were treated in terms of superficial physical traits like skin colour or hair shape however, this new science revealed that there was much more to human variation than meets the eye.

Genetics explains that differences among human groups rely on the different **allele frequencies,** that is to say, the percentages that each group has of each allele. To measure these allele frequencies tandem repeats[1] and SNPs[2] have been used. The investigator Anne Bowcock analyzed the CA repeats in 30 sites of the genome of individuals from 14 different populations and saw that different genetic human groups were formed according to different continents.[3] Moreover, Z. Li and M. Myers analyzed 1000 people from 51 different populations in 650000 different SNPs and concluded that human populations could be gathered into 5 different groups according to the continents: **Native Americans**, **Africans**, **Europeans**, **Oriental Asians,** and **Aboriginal Australians**.[3]

Although the studies mentioned previously considered multiple genome regions to analyze human diversity, the first genetic studies for this subject did not as genetic tools were not as developed as now. For instance, the first SNPs studies only included the ones related to the genetic markers **ABO** and **Rh** (blood groups). Later on, the ***EDAR-V370A*** allele for hair type or the ***Duffy-null*** allele for skin colour were used in different studies too. [3] The fewer genetic markers used, the less precise the analysis was leading to diminished subdivisions.

As genetics has evolved, more complex tools have been developed to analyze human variation to reach the more genetic markers the better. Nonetheless, the major topic nowadays is to understand and analyze the results obtained from genetic studies and figure out whether this population division exists or not and which genetic processes rely on it.

3.1. GENETIC VARIATION AT THE POPULATION LEVEL

The average proportion of nucleotide differences between a randomly chosen pair of humans (average nucleotide diversity or $\boldsymbol{\pi}$) is estimated to lie between 1 in 1000 and 1 in 1500. Indeed, since there are approximately three billion nucleotide base pairs in the haploid human genome, two randomly selected people, on average, differ by two to three million base pairs.[7]

[1] Tandem repeats:when one or more nucleotides are repeated and together in the genome.

[2] SNPs (single nucleotide polymorphisms): a change in one base of the DNA to a different one which can vary between populations (>1%).

However, what we might ask to ourselves is what proportion of this 0,1% of DNA that varies among individuals varies among main populations: it is crucial to determine the extent to which genetic variation is due to differences between populations (**inter-population** variability) versus differences within populations (**intra-population** variability).

To investigate this issue, some studies examined the world's population variability by dividing it into three major continents Africa, Asia, and Europe, and saw that approximately 85-90% of genetic variation was found within these continental groups and only an additional 10-15% of variation was found between them.[7] In other words, 90% of total human variation would be found among individuals of the same continent and only 10% among individuals of different continents.

Several investigations use the statistic $\mathbf{F_{ST}}$[3] to refer to the genetic variation related to differences between continental populations and it is considered to be consistent regardless of the used genetic markers. However, this statistic varies depending on how the population is being divided which shows the fragility of talking about genetic variation at a population level as a fixed criterion. In conclusion, regarding both of these statistics, $\boldsymbol{\pi}$ and $\mathbf{F_{ST}}$, human variation vary only slightly at the DNA level and only a small part of this variation relies on inter-continental variability.

The fact that, as explained in the previous chapter, inter-population variation (F_{ST}) is much lower than intra-population might lead us to not understand why, although, humans can be assigned to different groups according to their geographical origin. This is explained by the fact that F_{ST} detects the inter-population genetic differences even when they are smaller than the intra-population ones as it is a measure of the relative differences in genetic diversity rather than an absolute one. The genetic differences which accumulate between populations are due to various factors such as genetic drift, migration, and selection, and as long as they are consistent and systematic across populations compared, they can be detected by the F_{ST} analysis (based on allele frequency differences).

[3] F_{ST} (fixation index): is a measure of population differentiation relative to genetic structure. The formula is: $(H_T-H_S)/H_T$; where H_T is the total genetic diversity in the whole population and H_S is the avegare diversity within each one of the populations.

3.2. GENETIC VARIATION AT AN INDIVIDUAL LEVEL

When we genetically compare different populations, we are mightly making the mistake of distributing humans in pre-defined groups and therefore, possibly influencing the results of the study. Moreover, as commented before, these groups are usually arbitrarily made in so many ways. This is the reason why, by analyzing genetic variation at an individual level instead of a population one can overcome this issue.

Some studies, using only several dozen or fewer loci, have been performed and have not provided any evidence of clustering of the human population according to geographic origins. On the contrary, studies analyzing a lot of loci certainly showed that individuals clustered according to their ancestry or geographic origin.[7]

These analyses have revealed that genetic variation is not discrete but rather exists in a gradual manner which is known as “**clinal**”. If we examine the origin of the human population in Africa we can see that each time a group of individuals migrated from Africa, they randomly selected certain alleles, and each time a group divided, the number of alleles was reduced as the available alleles, and therefore the diversity, shortened. This caused a clinal genetic variation showing a gradient of genetic-variation-decrease, being Africa the most diverse continent.[4]

3.3. NATURAL SELECTION AND GENETIC DRIFT

Natural selection[4] and **genetic drift**[5] are the main biological forces that create differences among populations. Natural selection drives evolution in large societies, while genetic drift is more important in smaller populations.

[4] Natural selection: evolutive process in which organisms better adapted to their environment survive more than others.

[5] Genetic drift: variation in the relative frequency of genotype frequencies in populations due to the death or non-reproduction of certain individuals.

In the *HapMap* [6] project, Jonathan Pritchard searched for different genes under selection in three different populations: Africans, Oriental Asians, and Europeans. His discovery revealed that the genes under selection varied across the three distinct populations.[3] To know if a gene has been subjected to natural selection the genes around it reduce their variability. Each gene that has gone through selection explains some historical tension a population has suffered. An example is the allele *EDAR-V370A* which is the cause of thick hair in oriental Asians.

4. BIOINFORMATICS TOOLS FOR THE REPRESENTATIONS AND ANALYSIS OF POPULATION DIVISION: A USE CASE

The division of humans into different groups, even with the introduction of genetics, remains an ongoing matter without a definitive conclusion. Nowadays, the existence of separated genomic clusters around the world is a controversial topic among geneticists.

On the one hand, some studies argue that there should be a deconstruction of the relationship between genetics and "races" or ethnicities. They say that descriptors such as "race" or ethnicity capture only some of the ancestral information about biological and environmental factors that influence phenotypic characteristics and that the design of the study and how the groups are pre-defined vary the conclusions.[7]

On the other hand, many genomic investigations defend the position of classifying populations based on their genetic ancestry. Quantitative comparisons of the similarity between genes and geography on a worldwide scale have been performed, using a **PCA (Principal Component Analysis)** and multiple SNPs, and have found that components in PCA often produce a map that resembles the geographic distribution of sampling locations.[8]

[6] *Hap Map:* project to develop an haplotype map for human population in order to analyze genetic differences among individuals.

4.1. PCA ANALYSIS

To analyze if population stratification occurs, a PCA based on the protocol described by Shuai Cheng Li et al [9] will be performed. The **Principal component analysis** is a **statistical method** that has been used to identify structure in the distribution of genetic variation across geographical locations and ethnic backgrounds on many occasions.[10]

Overall, the PCA identifies the primary axes of variation in data and projects the sample onto these axes in a graphically appealing and easily understandable way. The underlying genealogical history of the samples is directly related to the principal components of the PCA which are the orthogonal axes, each of which is made up of a linear combination of allelic or genotypic values across SNPs or other types of variant, and which capture the most significant variation of data (**PC1**[7] and **PC2**[8]).[10]

For SNP data, the projection of samples onto the principal components can be obtained directly by considering the average coalescent times[9] between pairs of haploid genomes. The result provides a scheme for interpreting PCA projections always taking into consideration processes such as migration, geographical isolation, and admixture.[10]

4.2. SIMONS GENOME DIVERSITY PROJECT

The PCA analysis that I will perform will be based on the **Simons Genome Diversity Project (SGDP)**[11] which contains SNPs from 300 individuals from 142 populations: Africans, Native Americans, Central Asians or Siberians, East Asians, Oceanians, South Asians and West Eurasians (see specific populations in *Annex 1*).[11]

[7] PC1: direction of maximum variance in the dataset. It summarizes the biggest amount of variance among the data.

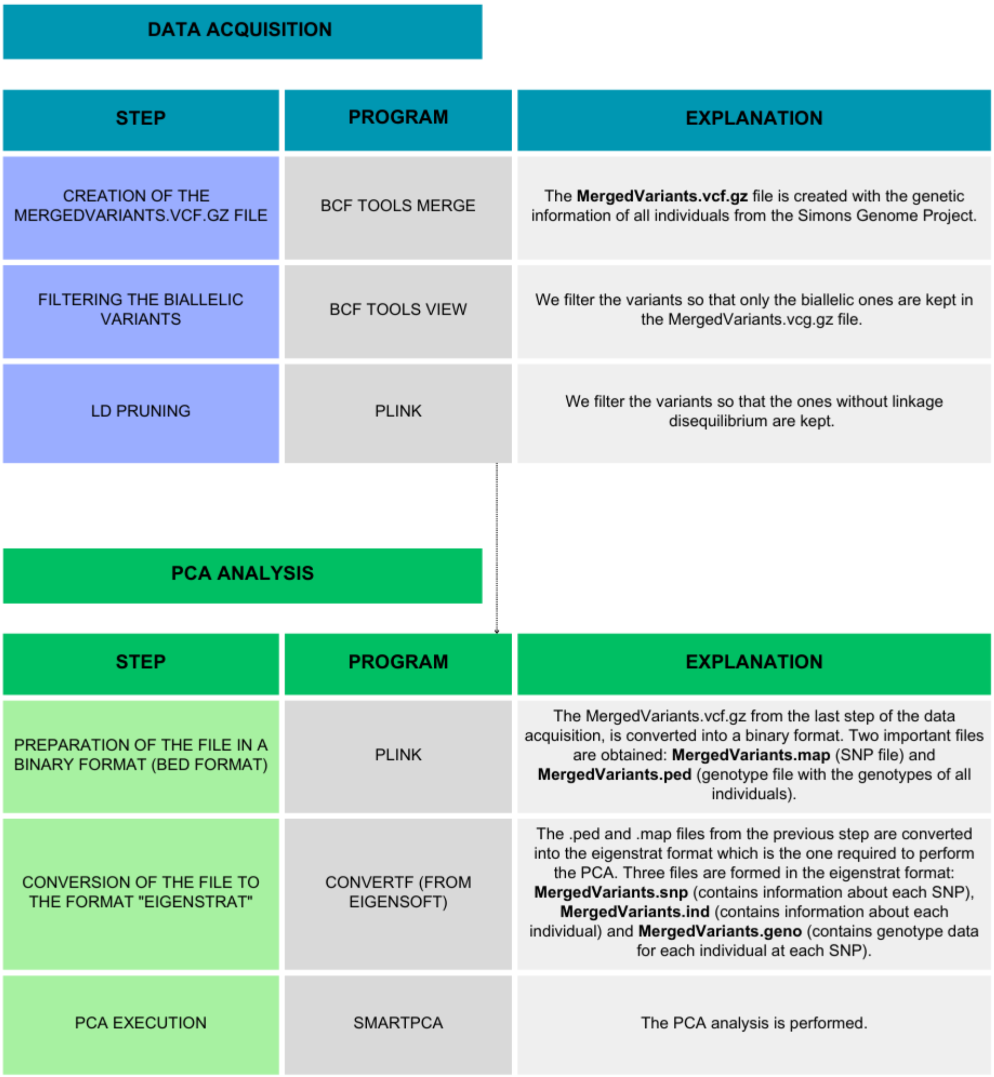
[8] PC2: orthogonal to PC1 and explains the second-biggest amount of variance but with a different direction of variation than PC1.

[9] Coalescent times: the amount of time that has gone by since the most recent ancestor of gene copies lived.

The genomes in this project were sequenced using the Illumina methodology and to at least 30x coverage. The data includes VCF files (Variant Call Formats Files) with genotype calls at every position in the genome.[11]

4.3. PROTOCOL

The protocol (*Figure 3*) was divided into obtaining the file to analyze and performing the PCA. The informatic code for the protocol can be seen in *Annex 2*.



DATA ACQUISITION

STEP	PROGRAM	EXPLANATION
CREATION OF THE MERGEDVARIANTS.VCF.GZ FILE	BCF TOOLS MERGE	The **MergedVariants.vcf.gz** file is created with the genetic information of all individuals from the Simons Genome Project.
FILTERING THE BIALLELIC VARIANTS	BCF TOOLS VIEW	We filter the variants so that only the biallelic ones are kept in the MergedVariants.vcg.gz file.
LD PRUNING	PLINK	We filter the variants so that the ones without linkage disequilibrium are kept.

PCA ANALYSIS

STEP	PROGRAM	EXPLANATION
PREPARATION OF THE FILE IN A BINARY FORMAT (BED FORMAT)	PLINK	The MergedVariants.vcf.gz from the last step of the data acquisition, is converted into a binary format. Two important files are obtained: **MergedVariants.map** (SNP file) and **MergedVariants.ped** (genotype file with the genotypes of all individuals).
CONVERSION OF THE FILE TO THE FORMAT "EIGENSTRAT"	CONVERTF (FROM EIGENSOFT)	The .ped and .map files from the previous step are converted into the eigenstrat format which is the one required to perform the PCA. Three files are formed in the eigenstrat format: **MergedVariants.snp** (contains information about each SNP), **MergedVariants.ind** (contains information about each individual) and **MergedVariants.geno** (contains genotype data for each individual at each SNP).
PCA EXECUTION	SMARTPCA	The PCA analysis is performed.

Figure 3: Protocol for the data acquisition and PCA analysis. (Own creation)

4.4. RESULTS

To perform the analysis the Simons Genome Project has been selected as it represents human variability more widely than other studies as it includes genomes from 142 populations (26 populations in the 1000 genomes).[12]

After following the protocol in the previous chapter, the graphics shown in *Figure 4* and *Figure 5* were obtained.

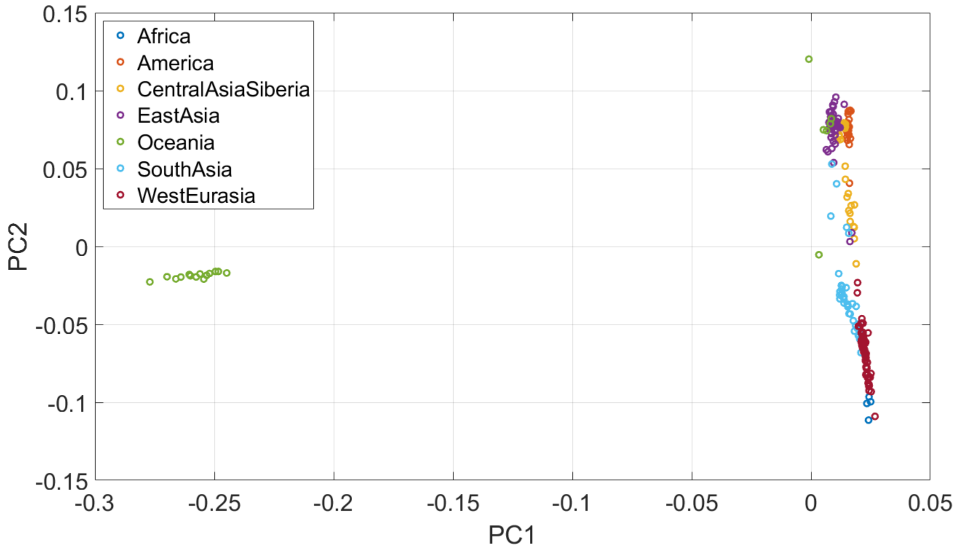


Figure 4: PCA result showing the big genetic separation of Oceania individuals among others. (Own creation)

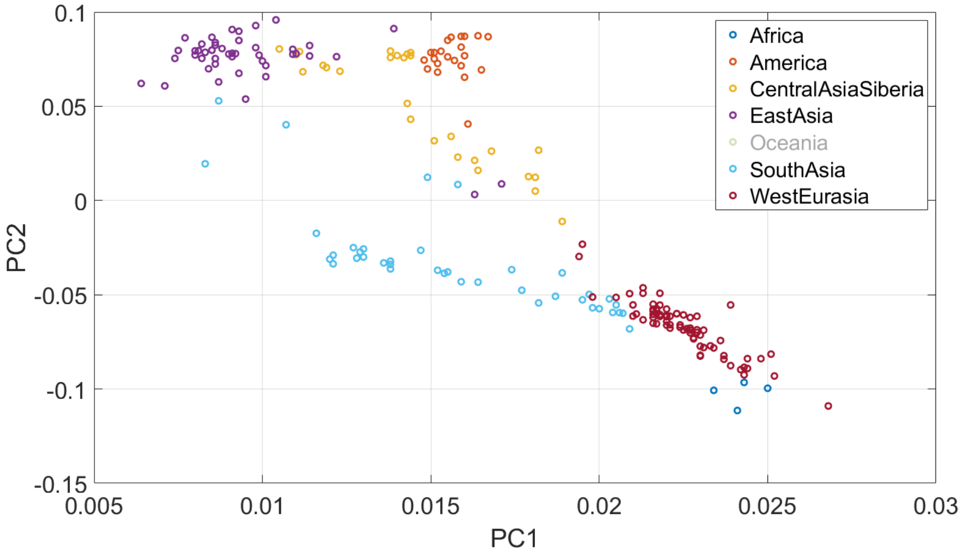


Figure 5: PCA result showing the genetic clusters of African, American, Central Asian Siberian, East Asian, South Asian, and West Eurasian populations, excluding the Oceanian one. (Own creation).

In both figures (*Figure 4 and 5*) a stratification of populations according to their continents of origin can be observed. Moreover, in *Figure 4* oceanian individuals are clustered far away from the other humans. This emphasizes the fact that Oceania is the most isolated continent which has made that genetic processes (such as natural selection) shape the particular genotypes of its individuals, making it more difficult to migrate from and to the continent throughout human history and therefore, creating a bigger differentiation.

4.5. RESULTS DISCUSSION

As far as I am concerned, the PCA results have shown a genetic stratification of the genetic groups clustered according to the continent from which they belonged being the clearer one the Oceanian aggrupation. However, these groups are not completely accurate with some individuals from different continents being slightly mixed.

Moreover, American individuals and African individuals are clearly separated and Asians and Europeans show more proximity.

Tracing back to the widespread consensus on humans classification mentioned in the introduction: Africans, Oriental Asians, Europeans, Native Americans, and Aboriginal Australians; I agree with the fact that, although fixed groups are not defined, individuals align better with individuals from their same ancestry than with others, always showing clinal variation patterns. Nonetheless, genetic variation should only be seen as a result of human history but never to justify racist behaviours. Moreover, this study has only been performed with the data obtained from the Simons Genome Project which includes only 300 genomes. That is to say that, in future studies, more data should be analyzed.

5. ETHICAL IMPLICATIONS

As seen in the previous chapters, genetics is used as a strong argument to divide populations into different groups. Nevertheless, genetics is also a tool for some scientists to justify their **genetic determinism**[10], leading to discriminatory practices.

While there is still no agreement on whether "races" exist or not, after apparent trials in declining beliefs in biological differences between "races" in the latest years, researchers are concerned about the "re-bioligisation" of "race" accompanying the genetic biotechnology revolution. This "**re-bioligisation**", is raised by the media which gives too much attention to the genetic differences among "racial" groups.[12]

The major concern is the deterministic messages that media reinforces which can lead to racial prejudices.

5.1. DETERMINISM IN THE HEALTH SYSTEM

Over the last few years, numerous studies have revealed profound racial disparities in disease.[13] This differential treatment of patients according to their belonged geographical group may have been the biggest impact genetics has had in perpetuating racial ideologies.

Nowadays, the messages that society receives related to genetically differentiated groups in the health system are mixed and unclear: some say that there is no scientific basis for the existence of these divisions while others use racial terms when describing research results (p.e. increased risk of breast cancer in Jews); all using the genetic concepts of population-specific markers, disease susceptibility or alleles.[14]

[10] Human traits, abilities and conditions are perceived as being determined by genetic factors (and not by environmental).[16]

Genetic predisposition can explain a little part of the variability in the presence and severity of disease, however, racial and ethnic groups (except very isolated ones) do not represent distinct gene pools then, and genetic justifications for different health treatments are weak.

A clear example of the dangerous determinism in the health system was seen with the polymorphism of the gene that encodes for the enzyme MOA (monoamine oxidase) which was associated with aggressivity. Then, the population was divided into Maoris[11] and non-Maoris, and the media reinforced the idea that Maoris were the carriers of this polymorphism and therefore, were aggressive.[13] This was one of the so many myths created regarding this theme.

5.2. ACCEPTED TERMINOLOGY

This thesis has used the term "**race**" as it was present constantly in literature however, it is important to highlight that in contemporary times, this term has **negative connotations** as it is closely associated with the concept of racism.

Ancestry and **ethnicity** are widely accepted terms in comparison to "race" and their use is being increased. Nonetheless, there is a lack of consensus in the field on how these terms should be addressed and which are their exact definitions.[1]

The important fact to take into consideration is that, as constantly mentioned in the thesis, population labels are not based on immutable criteria and can be influenced by social context.[15]

[11] Maori are the indigenous Polynesian people of New Zealand.

6. CONCLUSION

This thesis aimed to explore the historical and cultural origins of the term “race”, to examine how the introduction of genetics has altered the concept of it while performing a PCA to analyze genetic variability across the world and to explain some of the ethical impacts of racism.

It is clear from the research reviewed that there is not and has never been a consensus regarding how populations should be divided. Moreover, there is still disagreement among geneticists about whether populations are genetically clustered or not.

Considering the PCA results, a separation among individuals from different continents according to them is possible however, genetic determinism can never be justified in any circumstance as well as fixed genetic groups are not compatible with the actual idea of populations that mix and migrate.

As far as I am concerned, the term ethnicity should substitute the term “race” since it carries a more neutral connotation. However, the literature is full of contradictions for example the definitions of the terms “race” (*one of the main groups to which people are often considered to belong, based on physical characteristics that they are perceived to share*) and “ethnicity” (*a particular “race” of people*) from the Cambridge Dictionary. Therefore, more investigation is required in this area in order to abolish any racial prejudice still existing.

7. BIBLIOGRAPHY

1. Byeon, Y. J. J., Islamaj, R., Yeganova, L., Wilbur, W. J., Lu, Z., Brody, L. C., & Bonham, V. L. (2021). Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades. *The American Journal of Human Genetics*, *108*(12), 2215–2223. https://doi.org/10.1016/j.ajhg.2021.10.008

2. Contributors to Wikimedia projects. (2005a, 18 de abril). *Il·lustració - Viquipèdia, l'enciclopèdia lliure*. Viquipèdia. https://ca.wikipedia.org/wiki/Il·lustració

3. Fox, C. L. i. (2002). *Razas, racismo y diversidad*. Algar Editorial.

4. Wade, Nicholas (2015). *Una herencia incómoda*. (2015). Ariel.

5. BBC News. (2016, 25 de febrero). *What is India's caste system?*https://www.bbc.com/news/world-asia-india-35650616

6. *Monogenism and Polygenism | Encyclopedia.com*. (s.f.). Encyclopedia.com | Free Online Encyclopedia. https://www.encyclopedia.com/religion/encyclopedias-almanacs-transcripts-and-maps/monogenism-and-polygenism

7. Jorde, L. B., & Wooding, S. P. (2004). *Genetic variation, classification and "race." Nature Genetics, 36(11s), S28–S33.* doi:10.1038/ng1435

8. Wang, C., Zöllner, S., & Rosenberg, N. A. (2012). *A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. PLoS Genetics, 8(8), e1002886.*doi:10.1371/journal.pgen.1002886

9. Zhao, Z., Wang, Y., Zhang, Z., & Li, S. C. (2023). Protocol to analyze population structure and migration history based on human genome variation data. *STAR Protocols*, *4*(1), 101928. https://doi.org/10.1016/j.xpro.2022.101928

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

11. *Simons Genome Diversity Project*. (s.f.). Simons Foundation. https://www.simonsfoundation.org/simons-genome-diversity-project/

12. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N.,

Renaud, G., Erlich, Y., Willems, T., Gallo, C., ... Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

13. Kowal, E., & Frederic, G. (2012). *Race, genetic determinism and the media: An exploratory study of media coverage of genetics and Indigenous Australians. Genomics, Society and Policy, 8(2).*doi:10.1186/1746-5354-8-2-1

14. Foster, M. W. (2009). *Looking for race in all the wrong places: analyzing the lack of productivity in the ongoing debate about race and genetics. Human Genetics, 126(3), 355–362.*doi:10.1007/s00439-009-0674-1

15. Braun, L. (2002). *Race, Ethnicity, and Health: Can Genetics Explain Disparities? Perspectives in Biology and Medicine, 45(2), 159–174.*doi:10.1353/pbm.2002.0023

8. ANNEX

8.1. ANNEX 1

Table Annex 1: information about the Simons Genome Project and its individuals.

Ind.	Sample ID (SGDP)	Sex	Population ID	Region	Country
1	B_Dinka-3	XY	Dinka	Africa	Sudan
2	B_Ju_hoan_North-4	XY	Ju_hoan_North	Africa	Namibia
3	B_Mandenka-3	XY	Mandenka	Africa	Senegal
4	B_Mbuti-4	XY	Mbuti	Africa	Congo
5	B_Yoruba-3	XY	Yoruba	Africa	Nigeria
6	B_Karitiana-3	XY	Karitiana	America	Brazil
7	B_Mixe-1	XX	Mixe	America	Mexico
8	B_Dai-4	XY	Dai	EastAsia	China
9	B_Han-3	XY	Han	EastAsia	China
10	B_Australian-3	XX	Australian	Oceania	Australia
11	B_Australian-4	XY	Australian	Oceania	Australia
12	B_Papuan-15	XY	Papuan	Oceania	PapuaNewGuinea
13	BR_Kashmiri_Pandit-1	XY	Kashmiri_Pandit	SouthAsia	India
14	BR_Kharia-1	XY	Kharia	SouthAsia	India
15	BR_Kurumba-1	XY	Kurumba	SouthAsia	India
16	BR_Mala-1	XX	Mala	SouthAsia	India
17	BR_Onge-1	XX	Onge	SouthAsia	India
18	BR_Onge-2	XX	Onge	SouthAsia	India
19	B_Crete-1	XX	Crete	WestEurasia	Greece
20	B_Crete-2	XY	Crete	WestEurasia	Greece
21	B_French-3	XY	French	WestEurasia	France
22	B_Sardinian-3	XY	Sardinian	WestEurasia	Italy
23	S_BantuHerero-1	XY	BantuHerero	Africa	BotswanaOrNamibia
24	S_BantuHerero-2	XY	BantuHerero	Africa	BotswanaOrNamibia
25	S_BantuKenya-1	XY	BantuKenya	Africa	Kenya
26	S_BantuKenya-2	XX	BantuKenya	Africa	Kenya

27	S_BantuTswana-1	XY	BantuTswana	Africa	BotswanaOrNamibia
28	S_BantuTswana-2	XY	BantuTswana	Africa	BotswanaOrNamibia
29	S_Biaka-1	XY	Biaka	Africa	Central African Republic
30	S_Biaka-2	XY	Biaka	Africa	Central African Republic
31	S_Dinka-1	XY	Dinka	Africa	Sudan
32	S_Dinka-2	XY	Dinka	Africa	Sudan
33	S_Esan-1	XY	Esan	Africa	Nigeria
34	S_Esan-2	XX	Esan	Africa	Nigeria
35	S_Gambian-1	XY	Gambian	Africa	Gambia
36	S_Gambian-2	XX	Gambian	Africa	Gambia
37	S_Igbo-1	XY	Igbo	Africa	Nigeria
38	S_Igbo-2	XX	Igbo	Africa	Nigeria
39	S_Ju_hoan_North-1	XY	Ju_hoan_North	Africa	Namibia
40	S_Ju_hoan_North-2	XY	Ju_hoan_North	Africa	Namibia
41	S_Ju_hoan_North-3	XY	Ju_hoan_North	Africa	Namibia
42	S_Khomani_San-1	XX	Khomani_San	Africa	SouthAfrica
43	S_Khomani_San-2	XX	Khomani_San	Africa	SouthAfrica
44	S_Kongo-2	XX	Kongo	Africa	Congo
45	S_Lemande-1	XX	Lemande	Africa	Cameroon
46	S_Lemande-2	XY	Lemande	Africa	Cameroon
47	S_Luhya-1	XX	Luhya	Africa	Kenya
48	S_Luhya-2	XY	Luhya	Africa	Kenya
49	S_Luo-1	XY	Luo	Africa	Kenya
50	S_Luo-2	XX	Luo	Africa	Kenya
51	S_Mandenka-1	XY	Mandenka	Africa	Senegal
52	S_Mandenka-2	XX	Mandenka	Africa	Senegal
53	S_Masai-1	XY	Masai	Africa	Kenya
54	S_Masai-2	XY	Masai	Africa	Kenya
55	S_Mbuti-1	XY	Mbuti	Africa	Congo
56	S_Mbuti-2	XX	Mbuti	Africa	Congo
57	S_Mbuti-3	XY	Mbuti	Africa	Congo
58	S_Mende-1	XY	Mende	Africa	SierraLeone
59	S_Mende-2	XX	Mende	Africa	SierraLeone
60	S_Mozabite-1	XY	Mozabite	Africa	Algeria

61	S_Mozabite-2	XX	Mozabite	Africa	Algeria
62	S_Saharawi-1	XY	Saharawi	Africa	Western Sahara (Morocco)
63	S_Saharawi-2	XY	Saharawi	Africa	Western Sahara (Morocco)
64	S_Somali-1	XX	Somali	Africa	Kenya
65	S_Yoruba-1	XX	Yoruba	Africa	Nigeria
66	S_Yoruba-2	XY	Yoruba	Africa	Nigeria
67	S_Chane-1	XY	Chane	America	Argentina
68	S_Chipewyan-1	XX	Chipewyan	America	Canada
69	S_Chipewyan-2	XY	Chipewyan	America	Canada
70	S_Cree-1	XY	Cree	America	Canada
71	S_Cree-2	XX	Cree	America	Canada
72	S_Karitiana-1	XY	Karitiana	America	Brazil
73	S_Karitiana-2	XX	Karitiana	America	Brazil
74	S_Mayan-1	XX	Mayan	America	Mexico
75	S_Mayan-2	XX	Mayan	America	Mexico
76	S_Mixe-2	XX	Mixe	America	Mexico
77	S_Mixe-3	XX	Mixe	America	Mexico
78	S_Mixtec-1	XY	Mixtec	America	Mexico
79	S_Mixtec-2	XX	Mixtec	America	Mexico
80	S_Nahua-1	XY	Nahua	America	Mexico
81	S_Nahua-2	XY	Nahua	America	Mexico
82	S_Piapoco-1	XX	Piapoco	America	Colombia
83	S_Piapoco-2	XX	Piapoco	America	Colombia
84	S_Pima-1	XY	Pima	America	Mexico
85	S_Pima-2	XX	Pima	America	Mexico
86	S_Quechua-1	XX	Quechua	America	Peru
87	S_Quechua-2	XY	Quechua	America	Peru
88	S_Quechua-3	XX	Quechua	America	Peru
89	S_Surui-1	XX	Surui	America	Brazil
90	S_Surui-2	XX	Surui	America	Brazil
91	S_Zapotec-1	XY	Zapotec	America	Mexico
92	S_Zapotec-2	XY	Zapotec	America	Mexico
93	S_Aleut-1	XY	Aleut	CentralAsiaSiberia	Russia

94	S_Aleut-2	XX	Aleut	CentralAsiaSiberia	Russia
95	S_Altaian-1	XY	Altaian	CentralAsiaSiberia	Russia
96	S_Chukchi-1	XY	Chukchi	CentralAsiaSiberia	Russia
97	S_Eskimo_Chaplin-1	XY	Eskimo_Chaplin	CentralAsiaSiberia	Russia
98	S_Eskimo_Naukan-1	XX	Eskimo_Naukan	CentralAsiaSiberia	Russia
99	S_Eskimo_Naukan-2	XX	Eskimo_Naukan	CentralAsiaSiberia	Russia
100	S_Eskimo_Sireniki-1	XY	Eskimo_Sireniki	CentralAsiaSiberia	Russia
101	S_Eskimo_Sireniki-2	XX	Eskimo_Sireniki	CentralAsiaSiberia	Russia
102	S_Even-1	XX	Even	CentralAsiaSiberia	Russia
103	S_Even-2	XY	Even	CentralAsiaSiberia	Russia
104	S_Even-3	XX	Even	CentralAsiaSiberia	Russia
105	S_Itelman-1	XX	Itelman	CentralAsiaSiberia	Russia
106	S_Kyrgyz-1	XY	Kyrgyz	CentralAsiaSiberia	Kyrgyzystan
107	S_Kyrgyz-2	XX	Kyrgyz	CentralAsiaSiberia	Kyrgyzystan
108	S_Mansi-1	XY	Mansi	CentralAsiaSiberia	Russia
109	S_Mansi-2	XX	Mansi	CentralAsiaSiberia	Russia
110	S_Mongola-1	XY	Mongola	CentralAsiaSiberia	China
111	S_Mongola-2	XX	Mongola	CentralAsiaSiberia	China
112	S_Tlingit-1	XY	Tlingit	CentralAsiaSiberia	Russia
113	S_Tlingit-2	XX	Tlingit	CentralAsiaSiberia	Russia
114	S_Tubalar-1	XX	Tubalar	CentralAsiaSiberia	Russia
115	S_Tubalar-2	XX	Tubalar	CentralAsiaSiberia	Russia
116	S_Ulchi-1	XX	Ulchi	CentralAsiaSiberia	Russia

117	S_Ulchi-2	XX	Ulchi	CentralAsiaSiberia	Russia
118	S_Yakut-1	XX	Yakut	CentralAsiaSiberia	Russia
119	S_Yakut-2	XY	Yakut	CentralAsiaSiberia	Russia
120	S_Ami-1	XY	Ami	EastAsia	Taiwan
121	S_Ami-2	XY	Ami	EastAsia	Taiwan
122	S_Atayal-1	XY	Atayal	EastAsia	Taiwan
123	S_Burmese-1	XY	Burmese	EastAsia	Myanmar
124	S_Burmese-2	XY	Burmese	EastAsia	Myanmar
125	S_Cambodian-1	XY	Cambodian	EastAsia	Cambodia
126	S_Cambodian-2	XX	Cambodian	EastAsia	Cambodia
127	S_Dai-1	XX	Dai	EastAsia	China
128	S_Dai-2	XY	Dai	EastAsia	China
129	S_Dai-3	XX	Dai	EastAsia	China
130	S_Daur-2	XX	Daur	EastAsia	China
131	S_Han-1	XX	Han	EastAsia	China
132	S_Han-2	XY	Han	EastAsia	China
133	S_Hezhen-1	XY	Hezhen	EastAsia	China
134	S_Hezhen-2	XX	Hezhen	EastAsia	China
135	S_Japanese-1	XY	Japanese	EastAsia	Japan
136	S_Japanese-2	XX	Japanese	EastAsia	Japan
137	S_Japanese-3	XY	Japanese	EastAsia	Japan
138	S_Kinh-1	XX	Kinh	EastAsia	Vietnam
139	S_Kinh-2	XY	Kinh	EastAsia	Vietnam
140	S_Korean-1	XY	Korean	EastAsia	Korea
141	S_Korean-2	XX	Korean	EastAsia	Korea
142	S_Lahu-1	XX	Lahu	EastAsia	China
143	S_Lahu-2	XY	Lahu	EastAsia	China
144	S_Miao-1	XY	Miao	EastAsia	China
145	S_Miao-2	XX	Miao	EastAsia	China
146	S_Naxi-1	XY	Naxi	EastAsia	China
147	S_Naxi-2	XY	Naxi	EastAsia	China
148	S_Naxi-3	XX	Naxi	EastAsia	China
149	S_Oroqen-1	XY	Oroqen	EastAsia	China

150	S_Oroqen-2	XX	Oroqen	EastAsia	China
151	S_She-1	XX	She	EastAsia	China
152	S_She-2	XY	She	EastAsia	China
153	S_Thai-1	XY	Thai	EastAsia	Thailand
154	S_Thai-2	XX	Thai	EastAsia	Thailand
155	S_Tu-1	XY	Tu	EastAsia	China
156	S_Tu-2	XX	Tu	EastAsia	China
157	S_Tujia-1	XY	Tujia	EastAsia	China
158	S_Tujia-2	XX	Tujia	EastAsia	China
159	S_Uygur-1	XX	Uygur	EastAsia	China
160	S_Uygur-2	XY	Uygur	EastAsia	China
161	S_Xibo-1	XY	Xibo	EastAsia	China
162	S_Xibo-2	XY	Xibo	EastAsia	China
163	S_Yi-1	XY	Yi	EastAsia	China
164	S_Yi-2	XX	Yi	EastAsia	China
165	S_Lezgin-1	Not Assigned	Lezgin	WestEurasia	Russia
166	S_Bougainville-1	XX	Bougainville	Oceania	PapuaNewGuinea
167	S_Bougainville-2	XX	Bougainville	Oceania	PapuaNewGuinea
168	S_Dusun-1	XX	Dusun	Oceania	Brunei
169	S_Dusun-2	XX	Dusun	Oceania	Brunei
170	S_Hawaiian-1	XY	Hawaiian	Oceania	USA
171	S_Igorot-1	XX	Igorot	Oceania	Philippines
172	S_Igorot-2	XY	Igorot	Oceania	Philippines
173	S_Maori-1	XY	Maori	Oceania	New Zealand
174	S_Papuan-1	XX	Papuan	Oceania	PapuaNewGuinea
175	S_Papuan-10	XY	Papuan	Oceania	PapuaNewGuinea
176	S_Papuan-11	XY	Papuan	Oceania	PapuaNewGuinea
177	S_Papuan-12	XY	Papuan	Oceania	PapuaNewGuinea
178	S_Papuan-13	XX	Papuan	Oceania	PapuaNewGuinea
179	S_Papuan-14	XX	Papuan	Oceania	PapuaNewGuinea
180	S_Papuan-2	XY	Papuan	Oceania	PapuaNewGuinea
181	S_Papuan-3	XY	Papuan	Oceania	PapuaNewGuinea
182	S_Papuan-4	XY	Papuan	Oceania	PapuaNewGuinea
183	S_Papuan-5	XY	Papuan	Oceania	PapuaNewGuinea

184	S_Papuan-6	XY	Papuan	Oceania	PapuaNewGuinea
185	S_Papuan-7	XY	Papuan	Oceania	PapuaNewGuinea
186	S_Papuan-8	XY	Papuan	Oceania	PapuaNewGuinea
187	S_Papuan-9	XY	Papuan	Oceania	PapuaNewGuinea
188	S_Balochi-1	XY	Balochi	SouthAsia	Pakistan
189	S_Balochi-2	XY	Balochi	SouthAsia	Pakistan
190	S_Bengali-1	XY	Bengali	SouthAsia	Bangladesh
191	S_Bengali-2	XX	Bengali	SouthAsia	Bangladesh
192	S_Brahmin-1	XY	Brahmin	SouthAsia	India
193	S_Brahmin-2	XY	Brahmin	SouthAsia	India
194	S_Brahui-1	XY	Brahui	SouthAsia	Pakistan
195	S_Brahui-2	XY	Brahui	SouthAsia	Pakistan
196	S_Burusho-1	XY	Burusho	SouthAsia	Pakistan
197	S_Burusho-2	XX	Burusho	SouthAsia	Pakistan
198	S_Hazara-1	XY	Hazara	SouthAsia	Pakistan
199	S_Hazara-2	XY	Hazara	SouthAsia	Pakistan
200	S_Irula-1	XY	Irula	SouthAsia	India
201	S_Irula-2	XY	Irula	SouthAsia	India
202	S_Kalash-1	XY	Kalash	SouthAsia	Pakistan
203	S_Kalash-2	XX	Kalash	SouthAsia	Pakistan
204	S_Kapu-1	XY	Kapu	SouthAsia	India
205	S_Kapu-2	XY	Kapu	SouthAsia	India
206	S_Khonda_Dora-1	XY	Khonda_Dora	SouthAsia	India
207	S_Kusunda-1	XY	Kusunda	SouthAsia	Nepal
208	S_Kusunda-2	XY	Kusunda	SouthAsia	Nepal
209	S_Madiga-1	XY	Madiga	SouthAsia	India
210	S_Madiga-2	XY	Madiga	SouthAsia	India
211	S_Makrani-1	XY	Makrani	SouthAsia	Pakistan
212	S_Makrani-2	XX	Makrani	SouthAsia	Pakistan
213	S_Mala-2	XY	Mala	SouthAsia	India
214	S_Mala-3	XY	Mala	SouthAsia	India
215	S_Pathan-1	XY	Pathan	SouthAsia	Pakistan
216	S_Pathan-2	XX	Pathan	SouthAsia	Pakistan
217	S_Punjabi-1	XY	Punjabi	SouthAsia	Pakistan

218	S_Punjabi-2	XY	Punjabi	SouthAsia	Pakistan
219	S_Punjabi-3	XX	Punjabi	SouthAsia	Pakistan
220	S_Punjabi-4	XX	Punjabi	SouthAsia	Pakistan
221	S_Relli-1	XY	Relli	SouthAsia	India
222	S_Relli-2	XY	Relli	SouthAsia	India
223	S_Sindhi-1	XY	Sindhi	SouthAsia	Pakistan
224	S_Sindhi-2	XX	Sindhi	SouthAsia	Pakistan
225	S_Yadava-1	XY	Yadava	SouthAsia	India
226	S_Yadava-2	XY	Yadava	SouthAsia	India
227	S_Abkhasian-1	XY	Abkhasian	WestEurasia	Abkhazia
228	S_Abkhasian-2	XY	Abkhasian	WestEurasia	Russia
229	S_Adygei-1	XY	Adygei	WestEurasia	Russia(Caucasus)
230	S_Adygei-2	XX	Adygei	WestEurasia	Russia(Caucasus)
231	S_Albanian-1	XX	Albanian	WestEurasia	Albania
232	S_Armenian-1	XY	Armenian	WestEurasia	Armenia
233	S_Armenian-2	XY	Armenian	WestEurasia	Armenia
234	S_Basque-1	XY	Basque	WestEurasia	France
235	S_Basque-2	XX	Basque	WestEurasia	France
236	S_BedouinB-1	XY	BedouinB	WestEurasia	Israel(Negev)
237	S_BedouinB-2	XX	BedouinB	WestEurasia	Israel(Negev)
238	S_Bergamo-1	XY	Bergamo	WestEurasia	Italy(Bergamo)
239	S_Bergamo-2	XX	Bergamo	WestEurasia	Italy(Bergamo)
240	S_Bulgarian-1	XY	Bulgarian	WestEurasia	Bulgaria
241	S_Bulgarian-2	XY	Bulgarian	WestEurasia	Bulgaria
242	S_Chechen-1	XY	Chechen	WestEurasia	Russia
243	S_Czech-2	XY	Czech	WestEurasia	Czechoslovia(pre1989)
244	S_Druze-1	XX	Druze	WestEurasia	Israel(Carmel)
245	S_Druze-2	XY	Druze	WestEurasia	Israel(Carmel)
246	S_English-1	XY	English	WestEurasia	England
247	S_English-2	XX	English	WestEurasia	England
248	S_Estonian-1	XY	Estonian	WestEurasia	Estonia
249	S_Estonian-2	XY	Estonian	WestEurasia	Estonia
250	S_Finnish-1	XX	Finnish	WestEurasia	Finland
251	S_Finnish-2	XY	Finnish	WestEurasia	Finland

252	S_Finnish-3	XY	Finnish	WestEurasia	Finland
253	S_French-1	XY	French	WestEurasia	France
254	S_French-2	XX	French	WestEurasia	France
255	S_Georgian-1	XY	Georgian	WestEurasia	Georgia
256	S_Georgian-2	XY	Georgian	WestEurasia	Georgia
257	S_Greek-1	XY	Greek	WestEurasia	Greece
258	S_Greek-2	XY	Greek	WestEurasia	Greece
259	S_Hungarian-1	XX	Hungarian	WestEurasia	Hungary
260	S_Hungarian-2	XY	Hungarian	WestEurasia	Hungary
261	S_Icelandic-1	XX	Icelandic	WestEurasia	Iceland
262	S_Icelandic-2	XX	Icelandic	WestEurasia	Iceland
263	S_Iranian-1	XY	Iranian	WestEurasia	Iran
264	S_Iranian-2	XY	Iranian	WestEurasia	Iran
265	S_Iraqi_Jew-1	XX	Iraqi_Jew	WestEurasia	Iraq
266	S_Iraqi_Jew-2	XY	Iraqi_Jew	WestEurasia	Iraq
267	S_Jordanian-1	XY	Jordanian	WestEurasia	Jordan
268	S_Jordanian-2	XY	Jordanian	WestEurasia	Jordan
269	S_Jordanian-3	XY	Jordanian	WestEurasia	Jordan
270	S_Lezgin-2	XY	Lezgin	WestEurasia	Russia
271	S_North_Ossetian-1	XY	North_Ossetian	WestEurasia	Russia
272	S_North_Ossetian-2	XY	North_Ossetian	WestEurasia	Russia
273	S_Norwegian-1	XX	Norwegian	WestEurasia	Norway
274	S_Orcadian-1	XY	Orcadian	WestEurasia	OrkneyIslands
275	S_Orcadian-2	XX	Orcadian	WestEurasia	OrkneyIslands
276	S_Palestinian-1	XY	Palestinian	WestEurasia	Israel(Central)
277	S_Palestinian-2	XX	Palestinian	WestEurasia	Israel(Central)
278	S_Palestinian-3	XX	Palestinian	WestEurasia	Israel(Central)
279	S_Polish-1	XY	Polish	WestEurasia	Poland
280	S_Russian-1	XY	Russian	WestEurasia	Russia
281	S_Russian-2	XX	Russian	WestEurasia	Russia
282	S_Saami-1	XX	Saami	WestEurasia	Finland
283	S_Saami-2	XY	Saami	WestEurasia	Finland
284	S_Samaritan-1	XY	Samaritan	WestEurasia	Israel
285	S_Sardinian-1	XY	Sardinian	WestEurasia	Italy(Sardinia)

286	S_Sardinian-2	XX	Sardinian	WestEurasia	Italy(Sardinia)
287	S_Spanish-1	XY	Spanish	WestEurasia	Spain
288	S_Spanish-2	XX	Spanish	WestEurasia	Spain
289	S_Tajik-1	XY	Tajik	WestEurasia	Tajikistan
290	S_Tajik-2	XY	Tajik	WestEurasia	Tajikistan
291	S_Turkish-1	XY	Turkish	WestEurasia	Turkey
292	S_Turkish-2	XX	Turkish	WestEurasia	Turkey
293	S_Tuscan-1	XX	Tuscan	WestEurasia	Italy(Tuscany)
294	S_Tuscan-2	XY	Tuscan	WestEurasia	Italy(Tuscany)
295	S_Yemenite_Jew-1	XX	Yemenite_Jew	WestEurasia	Yemen
296	S_Yemenite_Jew-2	XY	Yemenite_Jew	WestEurasia	Yemen
297	T_Sherpa-2	XX	Sherpa	SouthAsia	Nepal
298	T_Tibetan-1	XX	Tibetan	SouthAsia	Tibet
299	T_Tibetan-2	XX	Tibetan	SouthAsia	Tibet
300	T_Sherpa-1	XY	Sherpa	SouthAsia	Nepal

8.2. ANNEX 2

Data acquisition:

#merge de vcf data

```
Ls DATA/*.gz > sdgp.lst
```

```
bcftools merge -0 -O < -l sdgp.lst -o MergedVariants.vcf.gz && tabix
MergedVariants.vcf.gz
```

Input: sdgp.lst

Output: MergedVariants.vcf.gz

#get snp variants with only two alts

```
bcftools view -m2 -M2 -v snps -O z -o MergedVariants.snp.bi.vcf.gz
MergedVariants.vcf.gz && tabix MergedVariants.snp.bi.vcf.gz
```

```
mv MergedVariants.snp.bi.vcf.gz MergedVariants.vcf.gz
```

Input: MergedVariants.vcf.gz

Output: MergedVariants.snp.bi.vcf.gz (renamed MergedVariants.vcf.gz)

#LD pruning

```
../plink-1.9/plink -vcf MergedVariants.vcf.gz --indep-pairwise 500 50 0.2 -our
MergedVariants -const-fid
```

Input: MergedVariants.vcf.gz

Output: MergedVariants.nosex, MergedVariants.prune.in, MergedVariants.prune.out

```
../plink-1.9/plink –vcf MergedVariants.vcf.gz --extract MergedVariants.prune.in –
recode vcf –out MergedVariants.prune –const-fid
```

Input: MergedVariants.vcf.gz, MergedVariants.prune.in

Output: MergedVariants.prune.vcf

```
mv MergedVariants.prune.vcf MergedVariants.vcf && bgzip MergedVariants.vcf && tabix
MergedVariants.vcf.gz
```

Input: MergedVariants.prune.vcf

Output: MergedVariants.vcf.gz

PCA analysis:

#Prepare the variant in “bed” format for downstream analysis

```
plink --vcf MergedVariants.vcf.gz --double-id --recode –out MergedVariants
```

Input: MergedVariants.vcf.gz

Output: MergedVariants.nosex, MergedVariants.map (SNP file), MergedVariants.ped (genotype file, binary file with all genotypes)

#Convert the files from .ped and .map to eigenstrat format adequate for the PCA

```
convertf -p convert.par –-out-missing-phenotype
```

Input: MergedVariants.map (SNP file), MergedVariants.ped (genotype file, binary file with all genotypes)

Output: MergedVariants.geno, MergedVariants.snp (= .map), MergeVariants.ind

#PCA execution

```
smartpca -p pca.car
```

Input: MergedVariants.geno, MergedVariants.snp, MergedVariants.Ind

Output: **MergedVariants.pca.evec**, MergedVariants.pca.eval