# Manipulation, Preprocessing, and Statistical Analysis of GWAS Genetics Dataset

Abtin Mogharabin

Working with genetic datasets is notoriously challenging due to their complex structure, which encompasses both biological and technical details. To effectively handle such data, a thorough understanding of both domains is essential. This document provides a comprehensive review of various techniques and methods for preprocessing and manipulating genetic GWAS datasets. While touching upon the key genetic information necessary available in other types of genetic data. We will explore a range of methods for preprocessing and analyzing these datasets, emphasizing the intuition and logic behind each method, along with some algorithmic details.

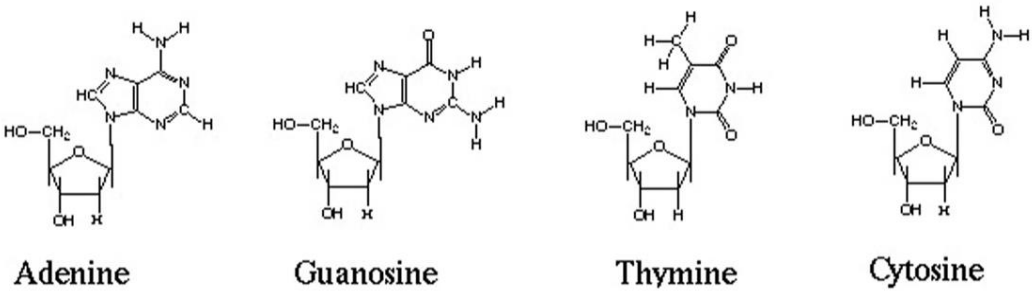
# Table of contents

## 1. Genome-Wide Association Studies (GWAS)

To summarize, Genome-wide association studies (GWAS) are a method of identifying genetic variants statistically associated with a biological trait of interest. Before we can understand what that means, we need to learn what is meant by 'genetic variants' and what is meant by 'traits':

### 1.1. Genetic Variation and Traits

Human DNA, the molecule that carries genetic information for the development and functioning of humans, consists of only four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). These bases are formed by the structures shown below, and they pair in specific shapes: A pairs with T, and G pairs with C.




To have an intuitive example of gene, assume (four) English alphabets to be nucleotides, then we can take each (three letter long) word generated by these alphabets to be a word and every sentence constructed by these words would be a protein. Then, each paragraph constructed by these sentences would be a gene. These paragraphs would then get together to generate a book chapter (genes create chromosomes), and book chapters would come together and result in a book, that delivers the instructions on biological processes to the body of each human.

To keep it short, the 4 A, C, G, & T bases combine in groups of three, constructing three nucleotide long structures for their corresponding amino acids and codons. Then, these amino acids combine and form a protein. And, multiple proteins stick to each other to generate the structure of a gene.

A gene can be understood as a sequence of DNA instructions that direct cells to produce proteins, the building blocks of the body. Genes consist of exons, which encode amino acids, and introns, non-coding regions of DNA. These genetic instructions are organized into chromosomes, with genes resembling paragraphs within these chromosomes. Surrounding the genes are extensive stretches of non-

coding DNA, containing sequences called transcription factors that regulate gene expression. Together, these genetic components form a complex system that governs various biological processes in the body.

* The human genome comprises 23 chromosomes and 3.3 billion letters.

### 1.2. Linkage Disequilibrium

To get to Linkage Disequilibrium, we first need to know what Genotyping is. To summarize, Genotyping is a process used in GWAS to identify specific nucleotides at various positions within an individual's genome. Unlike whole-genome sequencing, which deciphers every nucleotide in the genome, genotyping focuses on particular 'tag SNPs' to make the process more cost-effective. This approach is particularly useful in GWAS due to the importance of large sample sizes for detecting associations with complex traits. The rationale behind this method is that it provides significant information to impute, with high accuracy, the alleles at other genomic positions by using the principle of linkage disequilibrium.

Linkage disequilibrium (LD) is a fundamental concept in genomics that refers to the non-random association of alleles at two or more loci. It plays a crucial role in genotyping by allowing researchers to predict the presence of certain genetic variants based on the known variants within the same haplotype block. LD arises due to the manner in which chromosomes are inherited and recombined through processes like crossing over during meiosis. Dexplaiuring meiosis, chromosomes can exchange DNA segments, leading to new combinations of alleles. However, this recombination is not entirely random, and certain segments (haplotype blocks) are more likely to be inherited together, maintaining their association over generations.

We should note that although the predictive power of LD facilitates the identification of genomic regions associated with certain traits, it also obscures the clarity of pinpointing the exact causal genetic variants. The dilemma arises as GWAS often identifies multiple SNPs in tight LD, blurring the lines in distinguishing the variant directly influencing the trait. This complexity necessitates a deeper dive into the biological mechanisms at play, guiding researchers to hypothesize and investigate the role of genes within these identified regions, thereby unraveling the intricate web of genetic influence on complex traits.

One main reason for such problems is Crossing Over. Crossing over is a critical genetic process that occurs during meiosis, the type of cell division responsible for producing gametes (sperm and egg cells).

During meiosis, homologous chromosomes—pairs of chromosomes, one inherited from each parent—align closely together. At this point, they can exchange equivalent segments of DNA in a process known as crossing over. This exchange results in chromosomes that are a mix of alleles (variations of a gene) from both parents, contributing significantly to genetic diversity in offspring. Unfortunately, although this process is vital for generating genetic diversity but also creates challenges in pinpointing the exact genetic variants responsible for specific traits. (Because it complicates the identification of the causal variants responsible for traits. Often, GWAS identifies multiple SNPs in LD with each other, making it difficult to determine which SNP is directly influencing the trait of interest.)

But despite the possible problems we just mentioned, genotyping and LD are still considered a critical aspect of genetics research. By identifying regions of the genome associated with specific traits, researchers can focus their efforts on understanding the biological processes involved. This might involve investigating the genes located within or near these regions and hypothesizing how variations in these genes might influence the traits, based on existing knowledge of gene functions.

Now, we cover some of the key statistical methods used in GWAS data processing and analysis:

### 1.3. P-values and Multiple Testing

**P-values**: P-values are a fundamental statistical measure used in GWAS to assess the strength of the evidence supporting an association between a genetic variant and a trait. A P-value quantifies the probability that an observed association (or one more extreme) would occur by chance if there were actually no true association between the variant and the trait. The closer the P-value is to 0, the stronger the evidence for a real association. Conversely, a P-value closer to 1 suggests weaker evidence. For instance, a P-value of 0.01 implies a 1% chance that the observed association could be a false positive, occurring purely by chance.

**Multiple Testing**: GWAS involves testing associations between potentially millions of genetic variants and traits, which introduces the multiple testing problem. With each additional test, the chance of encountering a false positive (a spurious association deemed significant purely by chance) increases. Historically, a P-value threshold of 0.05 was commonly accepted in statistical analyses, implying a 5% chance of a false positive. However, given the massive number of tests conducted in GWAS, this threshold would result in an unacceptably high number of false positives. So we need to deal with this

problem suitably. One common approach to address the multiple testing problem in GWAS is to employ a more stringent genome-wide significance threshold, typically set at $p < 5 \times 10^{-8}$. This threshold is determined by dividing the conventional significance level of 0.05 by the estimated number of independent common SNPs across the human genome, thus drastically reducing the likelihood of false positives.

However, it should be noted that although this adjustment increases the reliability of the results, it also has some disadvantages that we need to try dealing with them:

- This approach necessitates extremely large sample sizes to detect real, albeit small, genetic effects on complex traits. To summarize, correcting for multiple testing mitigates the risk of false positives but also affects the study's power to detect genuine associations, especially for complex traits where genetic effects are often subtle. Large sample sizes are crucial to overcome this limitation, enabling researchers to identify true genetic associations at the genome-wide significance level. A suitable data size usually contains thousands of patients.

### 1.4. Linear and Logistic Regression

Regression analysis in GWAS is a statistical method used to model the relationship between a dependent variable (phenotype or trait) and one or more independent variables (genetic variants, often SNPs). The goal is to quantify how much a change in the genetic variant is associated with a change in the trait. This involves constructing a statistical model that best explains the observed variability in the phenotype based on the genotype data.

Linear regression is applied to continuous traits, where the relationship between the genetic variant and the phenotype can be approximated by a straight line. Mathematically, the model can be expressed as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Here, Y represents the trait value, X is the number of minor alleles (0, 1, or 2), $\beta_0$ is the intercept (the expected mean value of Y when X=0), $\beta_1$ is the slope (indicating the effect size of the genetic variant), and $\epsilon$ is the error term (capturing all other factors affecting Y besides X). The slope $\beta_1$ is crucial as it quantifies the change in the trait for each additional copy of the minor allele.

The coefficient of determination, $R^2$, measures the proportion of variance in the dependent variable that is predictable from the independent variable. In the context of GWAS, it quantifies how much of the variation in the trait can be explained by the genetic variant. $R^2$ is calculated as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where $SS_{res}$ is the sum of squares of residuals (variation unexplained by the model), and $SS_{tot}$ is the total sum of squares (total variation in the dataset). A higher $R^2$ value indicates a better fit of the model to the data, meaning the genetic variant has a more substantial role in determining the trait. Then, p-values are calculated for the slope coefficients to test the null hypothesis that the coefficient is equal to zero (no effect). The P-value is obtained from the t-statistic (or F-statistic for multiple regression), which compares the estimated coefficient to its standard error.

For binary traits, logistic regression models the probability that a trait is present as a function of the genetic variants. The logistic model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Here, p is the probability of the trait being present, $\log(1-p / p)$ is the log-odds of the trait, and X represents the genetic variant. The coefficients $\beta_0$ and $\beta_1$ are estimated using maximum likelihood estimation. The effect size $\beta_1$ indicates how the log-odds of the trait change with each additional copy of the minor allele. Here, the coefficient $\beta_1$ quantifies the change in the log-odds of the trait for a one-unit change in the predictor (the genetic variant). To interpret the effect size in terms of odds ratio (OR), we exponentiate $\beta_1$:

$$OR = e^{\beta_1}$$

An OR > 1 indicates an increased likelihood of the trait with each additional minor allele, while OR < 1 indicates a decreased likelihood. $R^2$ in logistic regression can be defined in several ways, with one common approach being the Nagelkerke $R^2$, which adjusts the Cox & Snell $R^2$ to vary between 0 and 1, making it analogous to the $R^2$ in linear regression. P-values in logistic regression are also derived from the likelihood ratio test, comparing the likelihood of the model with and without the predictor.

### 1.5. Polygenic Scores

Polygenic scores (PGS) are a quantitative representation of an individual's genetic predisposition to a certain trait or disease. They are derived from the cumulative effect of multiple genetic variants across the genome, each contributing a small effect to the overall trait or disease risk. PGS are calculated by summing the product of the number of effects alleles an individual has and the effect size (usually the beta coefficient from GWAS) of each SNP, effectively providing a score that predicts an individual's genetic liability for complex traits or diseases.

The calculation of a polygenic score involves summing up the effects of numerous genetic variants across the genome. For each variant, the number of effect alleles (0, 1, or 2) an individual has is multiplied by the estimated effect size of that allele (often obtained from GWAS results) on the trait. The sum of these products gives the individual's polygenic score:

$$\text{PGS} = \sum_{i=1}^{n} (\beta_i \times \text{Allele Count}_i)$$

Here, $\beta_i$ is the effect size (beta coefficient) for the i-th SNP, and Allele Count$_i$ is the number of effects alleles the individual has for the i-th SNP.

It should be noted that unlike monogenic traits (traits produced by the effect of a gene or an allele such as: the color of the animals, dwarfism, extreme muscularity, malformations, or severe health disturbances) that are influenced by a single gene, such as lactose tolerance or certain hereditary diseases, most complex traits like height, intelligence, or susceptibility to common diseases, are polygenic. Meaning they're influenced by many genetic variants each contributing a small effect. Polygenic scores aggregate these small effects across many SNPs to provide insights into an individual's genetic predisposition to these complex traits.

Here, the coefficient of determination ($R^2$) quantifies how well the polygenic score can predict the trait's variance in a population. It is essentially a measure of the predictive accuracy of the polygenic score. A higher $R^2$ indicates that the polygenic score explains a larger portion of the variance in the trait. For instance, an $R^2$ value of 0.11 in a polygenic score for IQ would imply that 11% of the variation in IQ across the population can be attributed to the genetic variants included in the score.

In general, PGS have various applications, including risk assessment for diseases, informing screening strategies, and understanding the biological basis of diseases with high comorbidity. They also

play an important role in personalized medicine, helping to predict drug response and tailoring treatments to individual genetic profiles. In psychological research, PGS can help disentangle the effects of genetics (nature) from environmental factors (nurture) in influencing complex behaviors and traits. Overall, it should be noted that polygenic scores can be used to predict both continuous traits (like height or IQ) and dichotomous traits (like the presence or absence of a disease). The interpretation of the score differs based on the trait type; for continuous traits, the score might predict the level of the trait, while for dichotomous traits, it might predict the likelihood or risk of the trait occurring.

But I also mention that there are also some issues with PGS. One significant issue is the "missing heritability" problem, where the genetic variance explained by current GWAS and PGS falls short of the heritability estimates derived from family and twin studies. Additionally, the predictive power of polygenic scores can vary widely across different populations, and their utility may be limited by the genetic architecture of the trait and the sample size of the GWAS from which the score is derived.

Now that we have covered the key aspects of GWAS data analysis, I will provide some details on SNPs:

**2. Single nucleotide polymorphism (SNP)**

SNPs represent the most common type of genetic variation among individuals. A SNP occurs when a single nucleotide (A, T, C, or G) in the genome differs between members of a species or paired chromosomes in an individual. These variations can influence how individuals respond to diseases, bacteria, viruses, chemicals, drugs, and other substances.

SNPs are found in approximately every 300 nucleotides along the human genome, which means there are about 10 million SNPs in the human genome. Most SNPs have no effect on health or development, but some can act as biological markers, helping scientists locate genes associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they can play a more direct role in disease by affecting the gene's function.

In general, SNP characteristics are 2:

- Bi-allelic Nature: Most SNPs are bi-allelic, meaning they have two possible nucleotide variations at a specific locus. This simplicity makes them highly amenable to genetic analysis and genotyping.

- Origin and Evolution: SNPs can arise due to errors in DNA replication or repair mechanisms, and once established in a population, they are subject to evolutionary forces such as mutation, drift, selection, and migration, which shape their distribution and frequency.

Due to their abundance and variability, SNPs serve as excellent markers for genetic fingerprinting, allowing researchers to differentiate between individuals or populations on a genetic level. This uniqueness in SNP patterns across individuals contributes to genetic diversity and can be used in various genetic studies, including population genetics, forensic analysis, and personal genomics.

### 2.1. SNP genotype data files

We cover two predominant data formats in genomics: the binary PLINK format (.bed, .bim, .fam files) and the classical plaintext format (.ped, .map files). Each format caters to different analytical needs, with the binary format being more efficient for computational analysis due to its compact nature.

**PLINK Binary Format:** This format consists of three files (.bed, .bim, .fam) that work together to store genotype data efficiently. It's designed for high-performance computational analyses, enabling rapid processing of large-scale genomic datasets.

- .bed File: Contains the binary representation of genotype data, indicating the presence of specific alleles at SNP positions across individuals. Its binary nature makes it compact and fast to process but not human-readable.
- .bim File: Acts as a detailed map for SNPs included in the .bed file. It lists each SNP's chromosome, identifier, genetic distance, and physical position, providing essential context for interpreting genotype data.
- .fam File: Provides metadata about samples in the dataset, including individual and family identifiers, parental information, sex, and phenotype. This file is crucial for understanding the demographics and traits of the sampled population.

**Classical Plaintext Format:** The .ped and .map files store similar information to the binary format but in a human-readable form that allows manual checking and editing.

- .ped File: Combines sample metadata with genotype data, listing individuals' familial relationships, phenotypes, and genotypes for each SNP. This format allows for a comprehensive view of the dataset but can be large and slow to process for extensive datasets.
- .map File: Corresponds to the .bim file in the binary format, detailing the genomic positions of SNPs. It provides the foundation for locating each SNP within the genome, crucial for linkage and association studies.

### 2.2. SNP data quality control and filtering

When dealing with genetics data, Quality control (QC) of SNP data is a critical preliminary step in genomic analysis to ensure the reliability and accuracy of your results. This process involves several key steps, each designed to identify and eliminate problematic data.

- **Missingness Per SNP and Per Individual:** This is done to remove SNPs and individuals with a high proportion of missing genotype data. High missingness can indicate poor DNA quality, issues during genotyping, or errors in data handling. In order to do this, we calculate the proportion of missing genotypes for each SNP and each individual. SNPs or individuals exceeding a predefined threshold (often 5-10%) are excluded from further analysis. This step helps maintain the integrity of statistical analyses by ensuring that conclusions are drawn from robust data.
- **Minor Allele Frequency (MAF):** This is done to exclude SNPs with very low MAFs, as they are less informative for association studies and more susceptible to genotyping errors. Rare alleles may also not be accurately represented in the sample, leading to biased estimates. In order to do this, we calculate the frequency of the minor allele for each SNP across all individuals. SNPs with MAF below a certain threshold (commonly 1-5%) are removed. Retaining only SNPs with higher MAF ensures that the remaining genetic variants are more likely to be of biological and statistical significance.
- **Hardy-Weinberg Equilibrium (HWE):** This is done to check for deviations from HWE, which can indicate inbreeding, population stratification, or genotyping errors. HWE states that genetic variant frequencies should remain constant from one generation to the next in the absence of evolutionary influences. To do this, we perform a chi-squared test to compare observed genotype frequencies with those expected under HWE for each SNP in the control group or the entire

sample for population-based studies. SNPs with significant deviations (e.g., p-value < 1e-6) are typically excluded, except for disease-associated SNPs in case-control studies, where deviations might be expected.

- **Sex Chromosome Anomalies:** This is done to identify discrepancies in sex chromosome data that may indicate sample contamination, mislabeling, or chromosomal abnormalities. In this step, we compare reported sex with sex inferred from sex chromosome genotypes. Discrepancies might lead to the exclusion of those samples or necessitate a review of the sample metadata.

- **Batch Effects:** This is done to detect and correct for variations in data that result from different sample handling, processing times, or genotyping platforms, rather than underlying genetic differences. In this step, principal component analysis or other clustering methods are used to visualize batch effects. Correction methods might include normalizing data across batches or including batch as a covariate in subsequent analyses.

- **Relatedness and Duplicates:** This is done to identify and remove closely related individuals or duplicate samples, as their inclusion can inflate the type I error rate in association studies. In this step, we estimate pairwise relatedness between individuals using identity-by-descent or identity-by-state metrics. Pairs exceeding a certain relatedness threshold may be pruned by excluding one individual from each related pair or duplicate set.

- **Population Stratification:** This is to control for spurious associations that arise when allele frequencies differ across subpopulations due to ancestry rather than a genuine association with the trait. Here, methods like principal component analysis are used to detect and adjust for population structure. Subsequent analyses can include these components as covariates to account for stratification.

### 2.3. Principal component analysis (PCA) based on SNP data

PCA is utilized in genomics to discern patterns of genetic diversity and structure within and across populations. It reduces the dimensionality of genetic data, transforming it into principal components (PCs) that capture the majority of the variation in the data. PSA is commonly applied in genomic diversity studies and exploratory data analysis to visualize relationships and clustering among individuals or populations based on genetic similarity.

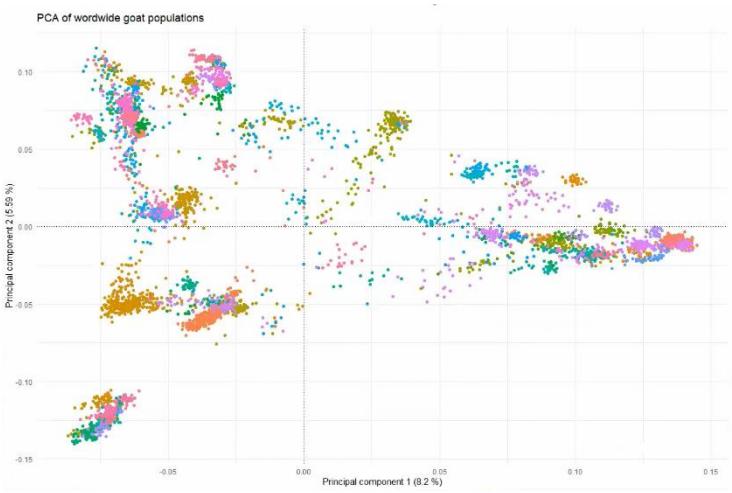
First note that as we described in the last section, before PCA, data must undergo rigorous quality control. This is to ensure accurate results. By filtering SNPs for missing data, minor allele frequency, and possibly Hardy-Weinberg equilibrium, among other metrics, we ensure that only high-quality and informative SNPs are retained. This reduces noise and potential biases in the PCA. Then, we continue with the following steps:

- **Computing Genetic Distances:** PCA requires a matrix of genetic distances or similarities between individuals, which reflects the genetic dissimilarity based on SNP genotypes. The genetic distance between two individuals can be calculated using various metrics, one common method being the pairwise identity by state (IBS) distance. IBS distance considers how often two individuals share the same alleles at a set of SNPs. For instance, if individuals share both alleles at an SNP, the IBS score might be 2 (completely identical), if they share one allele, the score might be 1 (half identical), and if they share no alleles, the score is 0. Then, we compile these pairwise distances into a distance matrix, where each cell (i, j) represents the genetic distance between individual i and individual j. Tools like PLINK calculate this matrix using SNP genotype data, considering each SNP's contribution to the overall genetic similarity or dissimilarity between every pair of individuals in the study.

- **Dimensionality Reduction:** PCA transforms the high-dimensional genetic data into a lower-dimensional space represented by the principal components. The first few PCs usually capture the most significant patterns of variation. This is done by mathematically decomposing the genotype matrix into principal components (PCs) through eigen decomposition of the covariance matrix of SNP data or singular value decomposition (SVD) of the data matrix itself. Each PC corresponds to an eigenvector of the covariance matrix, and the variance captured by each PC corresponds to its eigenvalue. In this process, the genotype data for each SNP is projected onto these PCs to transform the high-dimensional SNP data into a lower-dimensional space. The projection is a linear combination of the original SNP genotypes weighted by their loadings on the PCs, with the loadings indicating the correlation between each SNP and the PC.

- **Visualization and Interpretation:** In the PCA plot, each individual is represented as a point in the space defined by the first two or three PCs. The proximity between points reflects genetic similarity. This visualization can reveal clusters that often correspond to known population structures, genetic backgrounds, or even unexpected outliers that may warrant further

investigation. In addition, the axes of a PCA plot correspond to the PCs, which are orthogonal to each other, ensuring that each PC captures unique variance in the data. The percentage of total variance explained by each PC is often annotated on the axes. This assists in analyzing the data's dimensionality and the informativeness of the PCs. There are some sample PCA plots shown below:



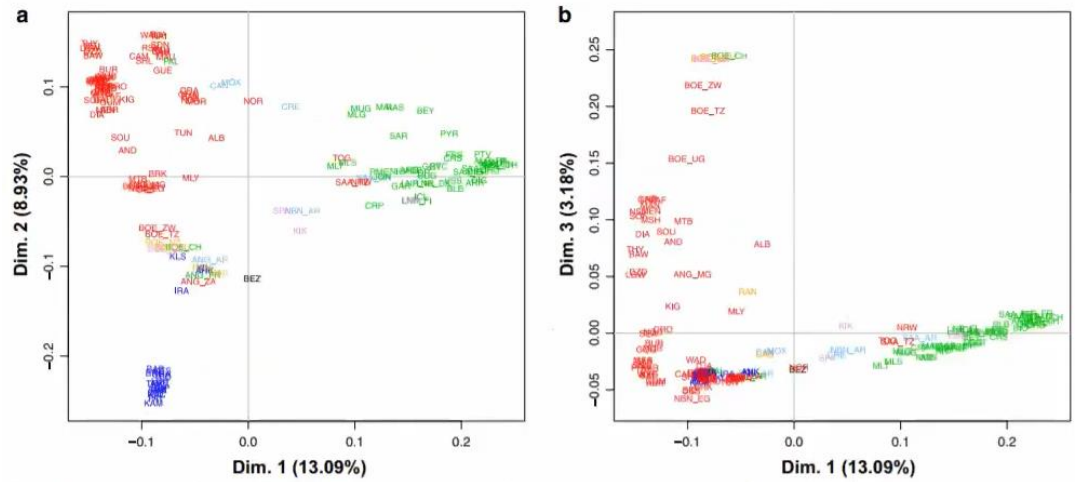




**Fig. 4** Multi-dimensional scaling plot. Dimension 1 versus 2 (**a**) and dimension 1 versus 3 (**b**). The population labels are coloured according to the continent of origin as follows: red = Africa, green = Europe, blue = West Asia, pink = North America, light blue = South America, orange = Oceania, black = wild goats. To increase readability, the country codes are omitted from the population labels, with the exception of breeds sampled in multiple countries

Sources available here and here

- **Eigenvectors and Eigenvalues:** Eigenvectors determine the direction of the PCs in the multidimensional space of the original data. The loading of a SNP on a PC (its weight in the eigenvector) quantifies the SNP's contribution to that PC. SNPs with high loadings significantly

influence the variance captured by that PC. In addition, eigenvalues quantify the amount of variance in the original data explained by each PC. The ratio of a PC's eigenvalue to the sum of all eigenvalues gives the proportion of total variance captured, guiding the selection of how many PCs to consider for a comprehensive yet concise representation of the data.

- **Advanced Visualization Techniques:** Also note that by incorporating metadata (e.g., population labels, geographical origins) as point annotations or colors in PCA plots, we can visually check and look for possible relations between genetic variation and external variables, potentially finding genetic patterns related to geography, phenotypes, or other factors.

A very good series on data wrangling with PLINK could be found here.

# References

Most of the information in this tutorial comes from my accumulated knowledge after working with such datasets over the last few months, so there are no specific references. However, I suggest taking a look at the excellent genetics boot camp tutorial from Gábor Mészáros , which includes many good explanations and visualizations.