Genome-Wide Association Studies

Data Processing and Imputation Pipeline

Beáta Éva Nagy

December 14, 2025

1 Introduction, Context

1.1 Human Genetics, Genome-Wide Association Studies

Every human somatic (non-sex) cell contains 46 **chromosomes**, which are organized into 23 pairs [1]. Each parent contributes one chromosome to each pair. Out of these 23 pairs, 22 are called autosomes: chromosomes that are present in both males and females and are responsible for controlling a wide range of traits such as eye colour, height, or susceptibility to diseases. The 23rd pair comprises the sex chromosomes, which consist of two X chromosomes in females and one X and one Y chromosome in males [1]. The Y chromosome therefore plays a central role in male sex determination.

These chromosomes carry approximately 20,000–22,000 **genes** [1]. Many of these genes are responsible for the production of proteins in the following way. Genes encode sequences of instructions for assembling amino acids (the building blocks of proteins) in a specific order, determining the three-dimensional structure and biological function of each protein. At a given gene location, individuals can carry different versions of a gene, called **alleles**. We inherit one allele from each parent. Within a population, the less frequent version of a genetic variant is referred to as the **minor allele**, while the more frequent version is called the **major allele.** [2]

Our DNA is composed of **base pairs**, which are formed by two complementary nucleotide bases binding together: adenine (A) always pairs with thymine (T), and guanine (G) always pairs with cytosine (C) [3]. A **base pair position** refers to the exact location (which is measured in base pairs) where a genetic variant is located on a given chromosome [2].

Genome-Wide Association Studies (GWAS) are a commonly used research approach, where we aim to identify genetic variants that are statistically associated with specific diseases or other characteristics, which are called **phenotypes** [4]. The genetic variants that are examined are called single-nucleotide polymorphisms (**SNPs**), where a single nucleotide (molecule that carries genetic information) differs between individuals. The main objective of GWAS is to detect associations between these variants and phenotypes, examined on large population samples [5].

More specifically, GWAS uses **statistical hypothesis testing** to assess if the observed associations between genetic variants and a phenotype have occurred by chance or reflect true biological relationships [4]. For each SNP tested, researchers perform association analyses with tests such as Pearson's chi-square test, Wald's test, or simpler, regression-based approaches. Logistic regression is generally applied for binary (categorical) phenotypes, while linear regression is used for continuous ones. These tests produce p-values that represent the probability of observing the test statistic under the null hypothesis that assumes that no true association exists between the genetic variant and the phenotype. In association testing, the significance threshold is generally set at $p < 0.05$. However, because GWAS simultaneously tests millions of SNPs, researchers use stricter thresholds. A commonly used approach is the Bonferroni correction, which divides the "conventional" significance threshold by the approximate number of independent tests (which is $\sim$1 million). As a result, the generally accepted GWAS significance threshold is $p < 5 \times 10^{-8}$. The increased strictness helps control

the rate of false-positive findings [6].

The first large-scale GWAS was conducted in 2005 [5]. Since then, GWAS has become a powerful and widely used tool for identifying genetic risk factors for diseases. Early studies successfully identified several genetic loci associated with conditions such as type 2 diabetes, Crohn's disease, and obesity [5]. These initial discoveries demonstrated that genome-wide association approaches are actually feasible, making GWAS an important and continuously evolving methodology in modern human genetics research [4].

1.2 Computational Tools for Genetic Data Analysis

To analyse genetic data, several specific computational tools have been designed [7]. These tools and software packages efficiently handle large-scale genomic datasets, supporting the different stages of genetic data processing and data analysis. Note, that some of these steps can also be performed with Python / R scripts, but the runtime and efficiency are better if we use the specialized tools.

- **PLINK** is a commonly used software tool for performing quality control steps and basic statistical analyses on large-scale genetic datasets [7]. For example, it allows researchers to filter samples and genetic variants based on specific criteria, such as minor allele frequency or the Hardy–Weinberg equilibrium (defined later). This is a very useful tool for supporting association testing methods that can be applied in GWAS.

- **BCFtools** is a software that helps the efficient manipulation of genetic data [8]. With the use of this tool, we can filter, index and analyze genetic variants, convert files to different formats efficiently, and create summary statistics as well.

- A genome reference assembly is a continuous genetic map that represents an organism's complete DNA sequence. It serves as a standard reference for research. Genome reference assemblies are periodically updated as sequencing technologies and annotation methods evolve (sometimes even daily), leading to more accurate genomic coordinates. As a result, genetic variant positions sometimes differ between assembly versions. It is important that we enable the efficient conversion between these versions, to ensure the consistency across different studies. **CrossMap** is a widely used solution for this, a software tool that does the conversion of genomic coordinates between different reference assemblies [9].

- Haplotype phasing is the process of determining which genetic variants are inherited together from each parent by identifying alleles that are physically linked on the same chromosome. **SHAPEIT4** is a commonly used phasing tool [10]. It determines which alleles come from the mother and which ones come from the father by comparing the observed genotypes to a reference haplotype panel.

- Reference panels are large datasets that contain fully sequenced genomes, capturing patterns of genetic variation within a population. These panels serve as a reference for genotype data imputation. It helps us predict the missing or unobserved genetic

variants based on known patterns [11]. Imputation increases the statistical power of GWAS, as it significantly increases the size of the dataset. **Minimac4** is a widely used software tool for performing genotype imputation using large reference panels [10, 11].

2 Project Overview

The input data for this study consisted of a multiethnic genetic database, divided into four distinct ethnic groups. Each contains genotype information for each individual, with one data entry (cell in the database) per SNP per person. In total, approximately 5,000 mother–child pairs were included in the analysis, and we had access to the genotype and associated phenotype information of each individual.

The main aim of the project was to construct two integrated datasets (one for mothers and one for children), containing filtered and imputed genotype data for each individual. These datasets form the basis of further GWAS analyses within each group.

The study participants came from diverse ethnic backgrounds, so we analyzed all the genetic data together rather than splitting it by ethnicity. This combined approach gives us more statistical power than analyzing each ethnic group separately. Also, ethnic differences can cause misleading findings if we analyse ethnicities one by one, thus, merging the databases gives us a more robust and reliable solution.

3 Execution Steps

3.1 Input Data

In GWAS analysis, genotyped data is stored in a standardized binary format, which consists of three files:

- The **BED** file (binary biallelic genotype table) is a compressed binary file, containing the genotype calls for each person at each SNP (which alleles are present at those particular positions) [7]. It is compressed in binary format to maximize efficiency and reduce file size; thus, it cannot be viewed as a plain text. Therefore, we need the above-mentioned softwares to be able to analyse the content of this file efficiently.

- The **BIM** file is a plain text file, containing one line per variant (SNP), with six columns [7]. As an example entry:

Chromosome	SNP ID	Position	Base Pair	Allele 1	Allele 2
1	rs3131962	0	756604	A	G

It contains which chromosome the given SNP is located in (1–22 for autosomes, 23 for X chromosome, 24 for Y chromosome), its unique identifier, the genetic map position (in centimorgans), the base pair position, and the two alleles at that locus (typically the minor allele, then the major allele).

- The **FAM** file (family file) is also a plain text file, with one line per individual, containing six columns of sample information [7]. An example of a mother–child pair:

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Phenotype
FAM001	IND001	0	IND005	1	-9
FAM001	IND005	0	0	1	-9

  Here, the identifiers of the family, the individual, the individual's father, mother (with 0 indicating missing values), the sex (coded as 1 = male, 2 = female), and phenotype data (disease status, 1 = control (unaffected), 2 = case (affected), -9/0/non-numeric = missing data) are listed.

  Generally, for mother-child pairs, mothers have missing paternal and maternal IDs, as that data is not needed for the project. The child has to be associated with the mother, so at least the maternal ID has to be present. In the example, the first individual is the child of the second individual (as the individual ID of the second person is the maternal ID of the first person).

In our database, the European population contains approximately 600,000 SNPs for each person, while the three other populations (Hispanic, Asian, and Afro-Caribbean) have approximately 1.0-1.2 million SNPs available, all of which are physically genotyped. We can see that these are huge databases with millions of cells; therefore, it is essential for us to maintain the efficiency and precision throughout the whole process.

It is also important to add that in our case, phenotype data is not stored in the FAM files, but separate phenotype files (which will be utilized later); therefore, the phenotype column contains -9 (missing information) everywhere.

3.2 Quality Control Pipeline

Quality control is the first important step in the process. The input data can be noisy sometimes, as both human errors and genotyping errors might occur. To ensure that our findings are precise, we need to eliminate unreliable data [12].

More precisely, it aims to remove low-quality samples, such as individuals with low genotyping rates (low percentage of SNPs successfully genotyped for a person), SNPs with missing details, and duplicate samples. This is an essential step, because these unreliable/incomplete genotypic data could add false positive findings. It could reduce the statistical power of the analysis; thus, we need to explicitly remove them. We used PLINK to perform these steps on each of the ethnicities separately, before merging them into one multi-ethnic database. The following steps were applied sequentially.

3.2.1 Remove SNPs with >2% missingness

SNP missingness for a given SNP means the missing rate of that column in the database across all individuals [12]. If the SNP is missing for more than 2% of the individuals, it is likely that some technical issue occurred; therefore, we decided to remove these variants from our database.

3.2.2 Remove individuals with >2% missingness

It is also possible that technical issues occurred for specific individuals rather than for specific SNPs. We also calculated the missing rate of SNPs, but now on the level of individuals. We removed all individuals with greater than 2% missing rate (more than 2% of their SNPs are missing), as these samples are not reliable enough for the GWAS analysis later on [12].

3.2.3 Keep only A/C/G/T SNPs

Genetic variants can take many different forms, for example, they can appear in the form of substituting one DNA base (A/C/G/T) with another one, inserting extra DNA sequences, or removing them from the genome. We excluded the more complex ones, like insertions or deletions, and we only kept genetic positions where exactly two versions of SNPs exist (like A or G). These are called **biallelic** SNPs [13]. The complex versions are harder for genotyping software to identify, leading to possible errors, which is the reason why we excluded them.

3.2.4 HWE filter

In the final step of the quality control process, we introduced the **Hardy-Weinberg (HW) Equilibrium** [12] filter with the threshold 1×10^{-6} . For a biallelic SNP with two alleles: allele A with frequency p, and allele B with frequency q (where $p + q = 1$), the Hardy-Weinberg equilibrium states that genotype frequencies should follow the equation:

$$p^2 + 2pq + q^2 = 1$$

For testing the Hardy-Weinberg condition, a chi-squared goodness-of-fit test is used, which compares the observed genotype counts at each SNP with the counts predicted by the HW equation. The test statistic is calculated as follows:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

We calculated this value, compared it to the threshold p-value, and excluded the SNPs which were below the threshold (with observed frequencies significantly deviating from the HW predictions).

There can be several reasons behind the difference in SNP frequencies. As an example, it is possible that a certain subset of people was chosen unknowingly with a bias, and it does not reflect an average population accurately [12]. We removed them to ensure that a population is represented with its full diversity.

3.3 Strand Alignment

Before performing the data imputation, we also needed to make sure that all genetic variants are oriented consistently. DNA has two complementary strands, so any given SNP can be represented in two different ways: read from the **forward strand** and the reverse strand [14]. For example, A/T on one orientation, and T/A on the other. It is possible that datasets are

not coordinated and the orientations are inconsistent, causing errors and unexpected results in the imputation process.

Our input data uses Illumina's TOP/BOT system, but the imputation software requires using a different notation system (forward/reverse) [14]. Using an Illumina reference file, we identified the SNPs which were oriented inconsistently and flipped their labels to match the forward strand orientation. This way, our data became compatible with the imputation process.

3.4 Genome Build Conversion from Build 36 to Build 37

Genomic knowledge and sequencing technologies continuously improve, leading to more accurate chromosomal coordinates over time. As a result, the human reference genome is periodically updated by correcting errors and making more precise estimations of chromosomal coordinates. These updated versions are referred to as **genome builds**.

Our genotype data were originally aligned to genome build 36, but we needed build 37 for imputation because our reference panel and imputation software use that version. With an Illumina reference file, we updated these positions to build 37 coordinates (based on matching the SNP ID-s). After the conversion, the new coordinates might not be ordered, so we also reordered the variants based on their new positions.

3.5 Conversion to VCF format + chromosome filtering

To prepare the data for imputation, we converted genotype files from PLINK binary format (BED) to Variant Call Format (VCF), which is the standard input format that the imputation software requires [15].

In addition, we removed chromosomes that are not autosomal (not chromosomes 1–22), for example, sex chromosomes or the mitochondrial chromosome. Excluding them is a standard step for GWAS, as they have different inheritance patterns and characteristics that would make imputation more complicated.

We used PLINK to perform all steps up to this point.

3.6 Sorting and indexing the VCF file

For successful imputation, variants must be ordered sequentially by chromosome number and base pair position. Therefore, we sorted the VCF files based on the SNP coordinates.

Afterwards, for each VCF file, we generated index files. As we work with large databases, enabling indexing makes the imputation algorithm more efficient. This is because indexing allows us to access specific parts of the data without having to read through the entire file.

We used BCFtools to perform these two steps.

3.7 Variant Normalization

Variant **normalization** ensures all genetic variants are represented consistently [16]. We used a reference genome (for build 37) for comparison. Without this step, the same variant can be recorded in different ways, and the database could contain duplicate entries, leading to inaccurate results (for example, inaccurate frequency calculations or associations).

We also applied **left-alignment**, which standardizes how insertions and deletions are positioned [17]. It shifts them to their leftmost possible location to maintain consistency.

We used BCFtools to perform these two steps.

3.8 Splitting the VCF into chromosomes

At this stage, each ethnicity was represented by a single VCF file containing variants from all 22 autosomes. We split them into 22 separate VCF files, one for each chromosome (containing all the SNPs of the given chromosome). This step is necessary because the imputation software processes chromosomes independently. It allows the software to run many processes in parallel, reducing the overall runtime of the imputation process.

We used PLINK to perform this step.

3.9 Minor Allele Count Filter

Extremely rare variants do not provide enough statistical power for reliable association testing. For this reason, we decided to exclude them. We applied a **minor allele count (MAC)** filter, to only keep SNPs where the minor allele was observed at least five times across all individuals within each dataset (MAC ≥ 5).

We used BCFtools to remove these variants.

3.10 Phasing

A **haplotype** is a set of alleles that are inherited together on a single chromosome copy. Because humans have two copies of each chromosome (one from each parent), standard genotype data does not tell us which alleles come from which parent. **Phasing** is the process of inferring this information [10].

Phasing uses the **linkage disequilibrium (LD)**, which means alleles located close together on a chromosome tend to be inherited together more frequently than expected by chance [2]. Phasing algorithms analyze nearby variants, looking for known LD patterns, and determine the most likely arrangement of alleles into maternal and paternal haplotypes.

Imputation uses haplotype matching: it matches the given person's haplotype to a reference haplotype. Phasing is necessary to perform because it is computationally infeasible to analyze all possible allele combinations – we need to determine which is the most likely one [10].

We performed phasing using SHAPEIT4.

3.11 Imputation

We performed genotype imputation with the 1000 Genomes Project Phase 3 reference panel (build 37), which contains haplotype data for millions of variants from thousands of individuals across multiple ancestries [11]. This reference panel is suitable for our dataset, as we need to impute data on a multi-ethnic set of individuals.

We performed imputation locally using Minimac4 to protect patient privacy, rather than uploading sensitive data to public servers. The process used 16 computational threads for efficiency.

To assess the quality of the performed imputation, we used the R^2 **metric,** which estimates the squared correlation between the imputed genotypes and true genotypes [11]. Higher values indicate more reliable imputation. We decided to exclude SNPs with R^2 value below 0.3 (using PLINK), which is a common threshold of acceptance [11].

3.12 Data Postprocessing

After the imputation, we merged the chromosome-specific files into a single dataset for each ethnicity. Then, we split these datasets into two separate genotype files: one containing the genotypes of mothers and one containing the genotypes of children. This step allows us to analyze maternal and child genetic effects separately.

We used PLINK to perform this step.

3.13 Multiethnic Database Building

We aim to create a multiethnic dataset for both mothers and children. To be able to combine the ethnic groups, we removed any SNPs that weren't present in all groups. After this filtering, we merged everything into two datasets—one for mothers and one for children.

We used PLINK to perform this step.

3.14 MAF >0.025 filter

From the combined multiethnic dataset, we removed rare variants using a **minor allele frequency (MAF)** filter in PLINK. We excluded SNPs where the less common version appeared in fewer than 2.5% of people [12]. This step removes variants that might be present more frequently in one ethnic group, but are rare overall; therefore, they are not reliable enough for statistical analysis.

4 Results

The table below summarizes the SNP count of each ethnicity before and after imputation:

	European	**Hispanic**	**Thai**	**Afro-Caribbean**
Initial	0.6 million	1.0 million	1.1 million	1.1 million
$R^2 = 0.3$	29.0 million	32.6 million	26.2 million	38.6 million
$R^2 = 0.8$	11.0 million	18.0 million	9.4 million	23.6 million

Besides using the R^2 filter > 0.3, we also included the results after filtering for $R^2 > 0.8$. Both results imply good quality imputation; $R^2 > 0.8$ provides even better quality data, with the trade-off of lower SNP count. We can adaptively use these two imputed datasets later on, and choose the one that suits our goal more: maximizing quality or statistical power.

The initial data consists of strategically selected variants that capture common genetic variation. We can see that the initial SNP count significantly increased, which shows that it successfully included millions of predicted variants across the genome. More data provides greater statistical power, which increases the chances of finding true associations.

5 Next Steps

We will continue this project by running a GWAS on the constructed database. Then, to analyse the results, we create a Manhattan-plot. It displays genomic coordinates along the x-axis, and the negative logarithm of association p-values on the y-axis, thus, each point represents one SNP. SNPs above the significance threshold ($p > 5 \times 10^{-8}$) show regions that are associated with the given trait we are interested in. We will record these significant SNPs for further analysis. We will analyse them using linear or logistic regression, depending on the types of the variables.

During the next semester, we plan to run an interaction GWAS analysis as well, based on a more complex interaction database.

References

[1] Isha Pathak and Bruno Bordoni. *Genetics, Chromosomes*. StatPearls Publishing, Apr. 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK557784/.

[2] *Understanding Human Genetic Variation*. NCBI Bookshelf / NIH Curriculum Supplement Series, 1998. URL: https://www.ncbi.nlm.nih.gov/books/NBK20363/.

[3] *Heredity, Genes, and DNA - The Cell*. NCBI Bookshelf. URL: https://www.ncbi.nlm.nih.gov/books/NBK9944/.

[4] Elise Uffelmann et al. "Genome-wide association studies". In: *Nature Reviews Methods Primers* 1.59 (2021). DOI: 10.1038/s43586-021-00056-9. URL: https://www.nature.com/articles/s43586-021-00056-9.

[5] Patrick F. Sullivan, Benjamin M. Neale, and Kenneth S. Kendler. "Genomewide association studies: History, rationale and prospects for psychiatric disorders". In: *American Journal of Psychiatry* 166.5 (Apr. 2009), pp. 540–556. DOI: 10.1176/appi.ajp.2008.08091354. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3894622/.

[6] Andrew T. Goodwin. "Statistical methods for genome-wide association studies". In: *Seminars in Reproductive Medicine* (Apr. 2019). URL: https://www.sciencedirect.com/science/article/abs/pii/S1044579X1730278X.

[7] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *American Journal of Human Genetics* 81.3 (2007), pp. 559–575. DOI: 10.1086/519795. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC1950838/.

[8] Christopher C. Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1 (2015), p. 7. DOI: 10.1186/s13742-015-0047-8. URL: https://arxiv.org/pdf/1410.4803.pdf.

[9] Haibo Zhao et al. "CrossMap: a versatile tool for coordinate conversion between genome assemblies". In: *Bioinformatics* 30.7 (2014), pp. 1006–1007. DOI: 10.1093/bioinformatics/btt730. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3967108/.

[10] Antonio De Marino et al. "A comparative analysis of current phasing and imputation software: impacts of rare variants and monophyletic populations". In: *PLoS ONE* 17.10 (2022), e0260177. DOI: 10.1371/journal.pone.0260177. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9581364/.

[11] Tomoko Naito, Marlene Dallmann-Sauer, and Rounak Dey. "Genotype imputation methods for whole and complex exome sequence data". In: *Journal of Human Genetics* 69.3 (2024), pp. 131–141. DOI: 10.1038/s10038-023-01213-6. URL: https://www.nature.com/articles/s10038-023-01213-6.

[12] Jonathan RI Coleman et al. "Quality control, imputation and analysis of genome-wide genotyping data". In: *Briefings in Functional Genomics* 15.4 (2016), pp. 298–307. DOI: 10.1093/bfg/elv037. URL: https://academic.oup.com/bfg/article/15/4/298/2412127.

[13] Broad Institute. *Biallelic vs Multiallelic sites*. GATK Documentation. Accessed: December 2025. URL: https://gatk.broadinstitute.org/hc/en-us/articles/360035890771-Biallelic-vs-Multiallelic-sites.

[14] Sarah C Nelson et al. "Is 'forward' the same as 'plus'? A clarification of strand asymmetries in variant databases". In: *Nature Genetics* 44.7 (2012), pp. 707–708. DOI: 10.1038/ng.2326. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6099125/.

[15] Petr Danecek et al. "The variant call format and VCFtools". In: *Bioinformatics* 27.15 (2011), pp. 2156–2158. DOI: 10.1093/bioinformatics/btr330. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3137218/.

[16] Databricks. *Streamlining Variant Normalization on Large Genomic Datasets with Glow*. Accessed: December 2025. Dec. 2019. URL: https://www.databricks.com/blog/2019/12/05/streamlining-variant-normalization-on-large-genomic-datasets-with-glow.html.

[17] Genomics England. *AggV2 Variant Normalisation and Representation*. Accessed: December 2025. URL: https://re-docs.genomicsengland.co.uk/variant_normalisation/.