Genetic basis of caffeine consumption
BIO-373: Genetics Practical
24 November 2025

In this project, we will study if there is a genetic contribution to caffeine consumption based on genome-wide association studies (GWAS). This project is divided into six sections plus a bonus section. The report contains questions that sum up to 25 points. Individual questions are marked 1, 1a, 1ai, etc. Ensure your report refers to each section/question, is tidy, and uses clear language *(1 point).*

Resources description
You will find in the resource folder the following files:

- **genotypes.vcf** is a processed VCF file that contains genotypes for 284 individuals and 10878 SNPs. Columns starting from position ten correspond to individuals and each row corresponds to a SNP. Columns from 1-9 contain information for each SNP. The FORMAT column contains summary information about genotypes in our study cohort. Please note that given the format of the VCF file, you are provided additional information on top of the genotype (GT) – you will need to remove the rest (AD:DP:GQ:PL) to analyse your data. The genotype information is encoded as ./. (missing data), 0/0 (homozygous for the allele in the reference genome), 0/1 and 1/0 (heterozygous), and 1/1 (homozygous for the alternative allele).

- **annotations.txt** contains the individual ID, biological sex (isFemale), superpopulation (based on data from the 1000 Genomes Project: EUR = European, EAS = East Asian, etc) and the phenotype that we want to study, namely, caffeine consumption (average number of cups of coffee consumed per week) for 3,500 individuals. We will, however, study a subset of all individuals (284 individuals).

1 | Setting the environment
Import required packages. You will certainly need **numpy**, **pandas**, **matplotlib**, and you will likely use **seaborn**, **sklearn** (linear_model, decomposition), **scipy**, and **statsmodels.api**. Download the resources folder from Moodle with the data, and clean the .vcf file to keep just the genotype (GT) data.

1a. Perform exploratory data analysis: plot a histogram of caffeine consumption and a pie chart for the distribution of biological sex. (1 point)

2 | SNP-level filtering: call rate
The call rate for a given SNP is defined as the proportion of individuals in the study for which the corresponding SNP information is not missing. Missing information is denoted as "./.".

2a. Calculate the call rate for each SNP and plot the histogram of call rates. Then, keep variants whose call rate is equal to 100%. How many variants are removed? (1 point)

3 | SNP-level filtering: minor allele frequency (MAF)
Minor-allele frequency (MAF) denotes the proportion of the least common allele in our study cohort for each SNP. For GWAS, rare variants with a low MAF are excluded since we don't' have sufficient statistical power with our study cohort size to identify significant associations.

3a. For this purpose, calculate the MAF for each SNP and plot a histogram of VAFs for all SNPs. Remove all SNPs whose MAF is less than or equal to 5%. How many SNPs are removed? Provide an interpretation for the shape of the MAF histogram. (HINT: it will be easier to calculate the MAF if you convert genotypes (GT) to numeric values: 0/0 = 0, 0/1=1, 1/0=1, 1/1=2, ./.=NA) (2 points)

In GWAS data preprocessing, there are also other sample-level filtering steps. They include filtering on call-rate (similar to variants filtering), heterozygosity level, and relatedness. If you want to learn more about this (and GWAS in general), you can read this paper.

4 | GWAS on the full study cohort
Now we will try to understand if genetics factors associate with caffeine consumption.

4a. Explore the potential effect of biological sex on caffeine consumption (2 points)

i. Plot caffeine consumption by biological sex to visually explore if there is a difference between males and females. Plot a boxplot and density plot showing the distribution of caffeine consumption by gender.

ii. Use linear regression to test if there is a relationship between the caffeine consumption and biological sex (Look for the appropriate function in scikit-learn or statsmodels package). Look at the coefficients of your model and the Coefficient of Determination (R^2) to determine the impact of biological sex on the studied phenotype. Should we include biological sex as a covariate in our analysis? Why or why not?

4b. Population structure (2 points)

i. *Principal Components Analysis (PCA) is commonly used to discover and correct for population structure in GWAS. To do this, calculate the principal components (PCs) of the genotype matrix, and plot the first and second principal components.*
ii. *How many data clusters can you observe with PC1 and PC2? If more than one, what do the clusters represent? Should we correct for population structure? Why or why not?*

4c. Run a GWAS without correcting for covariates (2 points)

i. *We will use linear regression to test for association between SNPs and the caffeine consumption phenotype. Use linear regression to build your model to test for an association. Test each SNP independently (hint: write a for loop).*
ii. *Extract at each iteration of the loop the coefficient of association (β value from the regression model) between the SNP and the phenotype and the corresponding P-value. Here, you should use the linear regression implementation from the **statsmodels** package since it provides also P values.*

4d. Produce a Manhattan plot (3 points)

i. *Produce a scatterplot where each point is a variant with their linear position in the reference genome on each chromosome plotted on the x-axis and the significance of association on the y-axis. You should use a −log10 scale for the p-values (you can find it in the **math** package, or you can use **numpy.log10**). Alternate two colours between each chromosome, and label each chromosome on the x-axis. Detect significant SNPs by comparing against a nominal significance threshold (P=0.05), a commonly used genome-wide significance threshold that accounts for the total number of independent SNPs (haplotypes) in humans (P=0.05 / 1,000,000 SNPs=5e-8), and a P value threshold that provides suggestive evidence for statistical association based on the total number SNPs that we tested (P=0.05 / #SNPs). On the plot, lines in different colours that correspond to the nominal, suggestive, and genome-wide significance threshold. Colour-code SNPs that pass the suggestive significant threshold (HINT: for the Manhattan plot, positions of SNPs are given for each chromosome separately. You need to stitch them into consecutive numbers along the whole genome).*
ii. *How many SNPs pass the nominal, suggestive, and genome-wide significant P value threshold? Explain the results at each significance threshold?*

*4e. Repeat the GWAS and Manhattan plot and correct for covariates based on the top 10 principal components from the genotype matrix. (**HINT**: You need to re-run the PCA with 10 PCs) (1 point)*

i. *How does correcting for the top 10 PCs change the results? Describe the plot and results.*
ii. *Report the nominal P value and beta coefficient for the most significant SNP.*

4f. Repeat the GWAS and Manhattan plot, but instead of correcting for the top 10 PCs, use the assigned SuperPopulation code as a covariate. (1 point)

i. *Describe what you see. How does using assigned super population labels instead of the top 10 PCs affect the results?*
ii. *Which approach would you generally prefer to use to control for population stratification and why?*

5 | GWAS meta-analysis

A GWAS across human populations is useful to explore potentially significant SNPs, yet most studies focus on single populations to better account for population biases (eg, random genetic drift, population bottlenecks). An alternative way to correct for population structures is group individuals by population, perform a GWAS for each population, and perform a meta-analysis that combines P values from all studied populations. For this, you will first obtain a P-value for each SNP and each population, and then combine P values. We will use the Fisher's method to combine P values.

5a. Split your genotype data obtained after call rate filtering by super population. Then for each super population filter out rare variants (MAF threshold: 0.05). How many variants remain in each super population? Why do we get different number of SNPs in each population? (2 points).

5b. Perform a GWAS analysis for each population using the top 10 PCs as a covariate. Please use the PCA decomposition function from sklearn with svd_solver="full", random_state=0 for reproducible results. Plot a Manhattan plot for each population and highlight significant SNPs at a suggestive significance threshold of P=1e-5. (1 point)

5c. With SNPs common to all populations (after MAF filtering), you will use the Fisher's method to combine p-values from all populations, and discover SNPs that drive caffeine consumption based on a meta-analysis. (3 points)

i. *Isolate SNPs and sum over individual population p-values (ie, -2 * sum(ln(p-value))*
ii. *Use the values from (i) to compute the population-wide p-value for each SNP. Assume these values follow a Chi2 distribution with 2 * n_population as the degree of freedom.*
iii. *Plot a Manhattan plot using the population-wide p-values from 5cii and a suggestive significance threshold of P=1e-5.*

5d. How do significant SNPs and p-values based on our meta-analysis differ compared to our analysis in Section 4e? (1 point)
HINT: Filter for SNPs that passed MAF filtering in all populations at the beginning. Make sure that your SNPs are aligned when combining p-values.

6 | Functional Analysis

Previous GWAS with a large cohort (>100,000 individuals) for coffee consumption identified chr15:74735539 (C/T) (GRCh38 reference genome build) as a genome-wide significant SNP. Look up the genomic coordinates in the gnomAD database (https://gnomad.broadinstitute.org/) and find the corresponding dbSNP **rsid**. Then, look up the rsid on Open Targets Genetics (https://platform.opentargets.org/) to determine the most likely gene(s) that is (are) associated with coffee consumption.

6a. What is the rsid and what is (are) the top associated gene(s) for this SNP? (0.5 point)

6b. What evidence was used to determine the SNP-to-gene association(s)? (0.5 point)

Optional | Mathematical intuition for the Fisher's method
This is an optional extension to this project. If you receive full points from previous sections, you will not gain extra points completing this section, but if you do not receive full points, completing this section will contribute to your final score.

You observed that genotype frequencies are significanlty influenced by population structure (ie, a SNP is more or less frequent in a population due to factors such as random genetic drift, population bottlenecks, etc.). Hence, we need to account for these effects in our GWAS. One solution is to use Fisher's method to perform a meta-analysis across study cohorts (eg, population-specific GWAS). Here we will develop a mathematical intuition to how Fisher's method solves this problem. *(4 points)*

I. Under the null hypotheses, p-values are uniformly distributed between 0 and 1. Show through a density plot that values drawn from a uniform distribution U=U(0,1) and transformed with -2 * ln(U) follow a Chi2 distribution with 2 degrees of freedom. *(1 point)*

II. In (1), you showed that a uniform distribution can be transformed into a Chi2 distribution. We will now extend it to a group-based population. One property of the Chi2 distribution is that the sum of n Chi2 random variables results in a Chi2 distribution varying only by the degree of freedom (df). Show in a density pot, that summing values drawn from five uniform distributions U(0,1) and summed as "-2*(ln(p1) + ln(p2) + ln(p3) + ln(p4) + ln(p5))" follow a Chi2(df=2*5). *(1 point)*

III. The null hypothesis of a group-based Chi2 distribution is that there is no significant effect in any groups. The alternative is that there is at least one group that shows an effect. Find a combination of 5 p-values that, once combined with Fisher's method, results in a final p-value below the threshold you used throughout this project. *(1 point)*
    Case 1: Fix four p-values to 0.5 and vary the last one.
    Case 2: All p-values are equal and vary them together.

IV. Always be skeptical! The Fisher's method assume to have independent random variables, can you list three reasons why this assumption might not hold in a GWAS study where groups are different populations. *(1 point)*