GWAS: the Influence of genetics on cholesterol level

Disclaimer: Make all your plots look nice - make sure they have the heading and axis labels and all labels are visible and readable.

1. Setting the environment

Import required packages. You will certainly need **numpy**, **pandas**, **matplotlib**. And maybe some others (you will find them later in the tutorial).

Download the resources folder from Moodle with the data.

2. Data preprocessing

1. Open the files and have a look at them.

In the resources folder, you should find the following:

- genotypes.vcf is a .vcf file of 319 samples for 25000 variants.
- phenotypes.txt contains the cholesterol level of the corresponding 319 samples.
- covariates.txt contains the gender of the corresponding 319 samples.

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 a dot (“.”). 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?

3. SNP-level filtering: minor allele frequency (MAF).

Minor-allele frequency (MAF) denotes the proportion of the least common allele for each SNP. For genome-wide association studies, rare variants with a low MAF are usually excluded. For this purpose, calculate MAF for each variant and plot its histogram. Then remove all the variants whose MAF is less than or equal to 1%. How many variants are removed?

In the GWAS data preprocessing, there are also sample-level filtering steps which we will skip. 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.

3. Genome-Wide Association Studies

Now we will try to understand if (and how) genetics influences cholesterol level.

1. Start your exploration by using statistical methods to look at the potential effect of gender on the phenotype.

First, plot cholesterol level by gender to visually explore if there is a difference. Plot boxplot and density plot showing the distribution of cholesterol level by gender. Explain what you see. Then use linear regression to see the relationship between your phenotype and the covariate (Look for the appropriate function in scikit-learn or statsmodels package). Look at the coefficients of your model and the Coefficient of Determination (R2) to know how much the covariates impact your phenotype. Based on the statistics, do you expect gender to impact the cholesterol level? Should we include gender as a covariate in our analysis?

2. Population structure

In association studies, Principal Components Analysis (PCA) is commonly used to correct for population structure. To do that, Calculate the principal components (PCs) of the genotype matrix, and plot the first and second principal components. How many clusters do you see? If more than one, what do the clusters represent? Should we correct for the population structure? Why?

3. Run a GWAS without correcting for covariates.

We will use linear regression to test for the association between the variants and the phenotype. Again, use linear regression to build your model to test the association. You should do it for each variant separately (with a loop). Extract at each iteration of the loop the coefficient of association (β) between the variant and the phenotype and the corresponding p-value. Here you should use **statmodels** package as **LinearRegression** from **sklearn** doesn't calculate the p-value. Alternatively, you can write your own function to calculate the p-value but **statmodels** is easier (you can read about the package for example, here, or use Google).

4. **Using the p-values from the previous step, produce a Manhattan plot**

You should produce a scatterplot where each point is a variant with their position in the 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 **math** package, or you can use **numpy.log10**). Alternate two colors between each chromosome, and label each chromosome on the x-axis. In order to detect significant p-values, you need to compare them with the significance threshold ($\alpha = 0.05$), but because you are performing multiple tests, you need to correct the significance threshold using the Bonferroni method by dividing the α by the number of tests. On the Manhattan plot, add a line corresponding to the Bonferroni corrected threshold (You should use the –log10 scale for the threshold). Also, try to plot the significant points in a different color.

HINT: positions of SNPs are given for each chromosome separately. You need to stitch them into consecutive numbers along the whole genome.

5. **Repeat the GWAS and Manhattan plot (steps 3 and 4), considering the top 10 principal components as covariates.**

HINT: You need to rerun PCA with 10 PCs.

6. **QQ plot**

A Quantile-Quantile plot (or QQ plot for short) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If two distributions are identical in shape, the Q-Q plot will display an x = y line.
Therefore, the dots on the Q-Q plot from a reliable GWAS analysis will align on an x=y line with only a few deviating values that are suggestive of association (significantly associated SNPs). Otherwise, if the line is shifted up or down from the diagonal, the GWAS analysis might be impaired by confounding factors. For this part, the goal is to make a QQ-plot for the observed p-values from GWAS with covariates (step 5).
First, generate the expected p-values. To do that, you need to identify what is the expected distribution of p-values (read this link, for example). To generate the expected distribution, you can use **scipy.stats** or **numpy** package.
Then sort both observed and expected p-values, also calculate -log10 for them. Finally, draw QQ plot (scatterplot of observed vs expected) and draw the x = y line. What can you say about the reliability of your GWAS results?