Genetics Project.

Impact of genetics on cholesterol levels.

Fall 2021

1 Setting the environment

ggplot2 and **data.table:** please make sure they are installed in your environment.
resources: download the folder from Moodle.

2 Data preprocessing

On the Moodle you will find:
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.

1. Load the data into Rstudio using **fread** function. Pay attention to the available arguments inside the **fread** especially *data.table*, *stringsAsFactors*, and *na.strings*.

   For the following steps (steps 2 and 3), you might use **colSums()**, **rowSums()**, **apply()**, **is.na()**, or for/while loop.

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. Calculate call rate for each SNP and plot the histogram of call rates using the ggplot function **geom_histogram()**. 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 using **geom_histogram**. Then remove all the variant whose MAF is less than or equal 1%. How many variants are removed?

3 Genome Wide Association Studies

1. Start your exploration by using statistical methods to look at the potential effect of gender on the phenotype. For that, you can use linear regression function **lm()** to see the relationship between your phenotype and the covariate. Use the function **lm()** to fit a model with formula $Cholesterol \sim gender$. The function **lm()** returns a fitted model object, use it in the function **summary()** to get more statistics about your regression. Look at the Coefficient of Determination (R^2) to know how much the covariates impact your phenotype. Based on the statistics, do you expect the gender to impact the cholesterol level?

2. Plot a boxplot for the cholesterol level in different genders. You can use **geom_boxplot()**. Separately, plot the distribution of cholesterol for different genders using **geom_density()**. Is gender a covariate for cholesterol level? Why?

3. 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?

4. Run a GWAS without correcting for covariates. You should use a for loop (or one of R functional counterpart such as **apply()**) as a control structure to iterate on every variant. We will use linear regression to test for the association between the variants and the phenotype. Again, use the function **lm()** to build your model and the **summary()** function to get more statistics. Extract at each iteration of the loop the coefficient of association (β) between the variant and the phenotype , and the corresponding pvalue.

5. Using the pvalues from the previous step, produce a Manhattan plot (You should use a $-log10$ scale for the pvalues). You can use the ggplot function, **geom_point()**. Using the data in file the variant_info.txt, alternate two colors between each chromosome. In order to detect significant pvalues, 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 Bonferroni method by dividing the significant threshold 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). You can use the ggplot function, **geom_hline()**. **Bonus:** plot the significant points in a different color.

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

7. Compare the results with and without covariates.

8. **Bonus: QQ plot**. 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 a $x = y$ line. Therefore, the Q-Q plot from a reliable GWAS analysis will display a $x = y$ line with only few deviating values that are suggestive of association. Otherwise, if the line is shifted up or down the GWAS analysis might be impaired by confounding factors. For this part, the goal is to draw a QQ-plot for the observed pvalues from GWAS with covariates (step 6). First, generate the expected pvalues using **ppoints()** function. Then sort and scale both observed and expected pvalues using **sort()** and $-log10$. Finally, draw QQ plot using **geom_points()** and draw $x = y$ using **geom_abline()**