Midterm Report

2025.04.12

We used GWAS (Genome-Wide Association Study) to identify SNPs associated with height. In the analysis, gender and the first three principal components (PCs) were used for adjustment. The linear regression model is as follows:

$$\text{Height} = \beta_0 + \beta_1(SNP) + \beta_2(Gender) + \beta_3(PC1) + \beta_4(PC2) + \beta_5(PC3) + \varepsilon$$

The Height and gender information can be found in the file: med_data.xlsx

The report includes:

a. **Quality Control Flowchart**: Documenting the changes in the number of individuals and variants throughout the data cleaning process.
b. **List of SNPs significantly associated with height (p-value < 5×10^{-8})**, presented in table format. The table should include the following information for each SNP, and the results should be ordered by SNP location:
    1. SNP ID
    2. SNP Location
    3. Coding Allele
    4. Effect Size
    5. P-value
c. **Manhattan Plot**: Should include a horizontal line at 5×10^{-8} and label the IDs of significant SNPs (p-value < 5×10^{-8}).
d. Determine whether any of the significant SNPs identified in b. might represent the same signal by considering the Linkage Disequilibrium (LD) between SNPs.
e. Plink / R code