Midterm Report Assignment: Genome-Wide Association Study (GWAS) on Height

2025.04.12

Objective:

Conduct a GWAS to identify single-nucleotide polymorphisms (SNPs) significantly associated with human height. This project will involve performing quality control, running a regression-based GWAS, generating visualizations, and interpreting results, with adjustments for confounding factors.

Provided Files:

med_data.xlsx: Contains phenotype and covariate data, including Height and Gender.

Linear Regression Model:

In the following analysis, gender and the first three principal components (PCs) are used for adjustment. The linear regression model is as follows:

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

Report Requirements:

The report should include:

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, and the results should be ordered by SNP location:

   1. SNP ID
   2. SNP Location (Chromosome & Location)
   3. Coding Allele
   4. Effect Size of SNP
   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. Code (Plink / R/ …)