GWAS Tool Development

Partners:
Chloe Keggen, A17081400
Ieva Sereiva, A16764544

Our goal is to build a command-line GWAS tool that analyzes genotype-phenotype associations across a genome. The tool will be developed as a Python package, implementing statistical models to identify these associations It will incorporate logistic and linear regression models for analysis, include clumping to identify independent genetic signals, and provide a comprehensive set of statistics and visualizations with Manhattan and QQ plots.

Our tool will replicate the approach in Lab 3 by running linear GWAS using the –linear option and using clumping to identify significant SNPs in independent genetic signals. Our tool will accept the same input as:
plink –vcf <vcf> –maf <maf> -pheno <phen> –linear –out <out file>

We have two input files, one path to the phenotype files containing the trait data, and the other to the VCF containing genotype data/SNPs.

To benchmark our tool against PLINK and the GWAS process in Lab 3, we will compare speed, memory usage, and output results with PLINK, focusing on the linear regression model. We will analyze the similarities and differences in the p-values for SNPs between the two tools, and test both with different dataset sizes.

For this GWAS implementation, we will use sample genomes from the 1000 Genomes Project. We chose to explore the phenotype of body height. Using NCBI's PhenGenI, we found the SNPs associated with body height (PheGenI: Phenotype-Genotype Integrator - National Center for Biotechnology Information (nih.gov)).
If there are issues with permissions for human GWAS data, we will use mouse GWAS data from the following paper: Systems genetics in diversity outbred mice inform BMD GWAS and identify determinants of bone strength | Nature Communications, where we would be examining SNPs associated with bone strength. Furthermore, to ensure data quality, we will perform QC filtering on the selected dataset. Since most mouse GWAS data are typically in RCTL format, we will convert such files into VCF format to facilitate the benchmarking process.