STAT3306/7306 - Statistical Analysis of Genetic Data
Project Part 2: Genetic architecture analyses

Due: 2pm 25 October 2024

Overview

Following 'Part I: Genome-wide association study analyses', you are now asked further analyses on your data. 'Part II: Genetic architecture analyses'.

Data

To put everyone on the same page, a set of cleaned data is available on the cluster:
/data/STAT3306/Project/Data_QC
/data/STAT3306/Project/Phenotypes_QC

As with Part I, the data testFiltered.bim, testFiltered.fam, testFiltered.bed consists in HapMap 3 genome-wide SNPs for 11, 000 people. You are also supplied a covariate file, with age, sex and the first 5 PCs. You will use the same 3 simulated phenotypes as in Part I (see a reminder details of which phenotype you have been assigned below):

1. a quantitative trait
2. a binary trait in which those scoring in the top 20% of the phenotype are scored 1 = case and 0 = control
3. a binary trait in which those scoring in the top 20% of the phenotype are scored 1 = case and those scoring in the bottom 30% of the phenotype are scored 0 = control

You are also provided a file with SNPs annotated into two groups, labelled "1" (being the annotation class you are interested in) and "0" (the rest of the genome).

Analysis

You will need to perform the following analyses:

• Whole genome SNP-heritability estimates for the three traits

• SNP-heritability estimated based on the provided genomic annotation that requires 2 or more genomic relationship matrices ("Partitioning heritability")

NOTE: **Calculations to be performed will be fairly intensive, especially the calculation of GRMs. It is essential you do not use a login node to run jobs!**

An example slurm job for GTCA with sufficient memory for GRM calculations is given in the **/data/STAT3306/example** directory.

Report

Write a short report in the format of a technical report: Intro, Methods, Results, Discussion and Supplementary material.

In your report, you should pay particular attention to:

• Explanation of models, analyses, assumptions, the statistical models used and the statistical tests conducted.

• Making comparisons of the results for these different data sets demonstrating an understanding of how the results from the three traits relate to each other.

As with Part I, provide all code etc in a Supplementary Materials section.

Please combine everything into a single pdf and submit through Blackboard

GCTA Help

The GCTA page has good examples of running a GREML:
https://cnsgenomics.com/software/gcta/#GREML

Data Assignment
Students have been assigned one of ten phenotypes for analysis as detailed below:

s4697683	BMI
s4698512	Fasting Glucose
s4698086	Fasting Insulin
s4663625	Ferritin
s4786571	Height
s4895814	Neuroticism
s4585713	Sleep duration
s4646995	Smoking (Pack Years)
s4809085	Systolic blood pressure
s4745210	Waist to Hip ratio
s4781897	BMI
s4810563	Fasting Glucose
s4903824	Fasting Insulin
s4585711	Ferritin
s4824469	Height
s4813931	Neuroticism
s4810122	Sleep duration
s4877617	Smoking (Pack Years)
s4688638	Systolic blood pressure
s4783467	Waist to Hip ratio
s4850648	BMI
s4790034	Fasting Glucose
s4765132	Fasting Insulin
s4761542	Ferritin
s4815491	Height
s4834587	Neuroticism
s4785145	Sleep duration
s4697848	Smoking (Pack Years)
s4867658	Systolic blood pressure
s4698653	Waist to Hip ratio
s4849428	BMI
s4789404	Fasting Glucose
s4900511	Fasting Insulin
s4704475	Ferritin
s4703815	Height
s4699103	Neuroticism
s4782709	Sleep duration
s4799891	Smoking (Pack Years)
s4510576	Systolic blood pressure

s4643024	Waist to Hip ratio
s4744037	BMI
s4856179	Fasting Glucose
s4818905	Fasting Insulin
s4638729	Ferritin
s4743939	Height
s4857153	Neuroticism
s4727337	Sleep duration
s4722208	Smoking (Pack Years)
s4805153	Systolic blood pressure