STAT3306/7306 - Statistical Analysis of Genetic Data
Project Part I: Genome-wide association study analyses

Due: 2pm 04 October 2024

Overview

One of your new collaborators recently approached you with a project. His team received enough founding to genotype ~ 11,000 individuals and collect different quantitative and qualitative phenotypes on these individuals. The project appears simple: analyse the data and identify genomic region that are linked to the phenotype of interest. However, because it is the first time you work with this new collaborator and you want him to involve you in more of his projects, you want to have your work double checked by your supervisor. To that end, you write a detailed report summarising your
analyses and your conclusions.

The project consists in two part: ‘Part I: Genome-wide association study analyses’ and ‘Part II: Genetic architecture analyses’. Your results in Part I will be checked as Part II will build on them.

For this project, you will have access to the School of Mathematics and Physics (SMP) cluster. The data has been stored on that cluster, for ethics reasons you are forbidden to move it or copy it somewhere else. You can however, download the results of any analysis run on the cluster. See details in the ‘Technical aspect’ section below.

Data

The data test.bim, test.fam, test.bed consists in HapMap 3 genome-wide SNPs for ~11,000 people. Allele frequencies for the reference “1” allele are given in the file reference_allele_frequencies.txt. From this real data, each student will have access to covariates of gender and age and 3 simulated phenotypes (see 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 the remainder 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

The data consists in real genotype data from the UK Biobank study. In order to provide data sets of manageable size, we have simulated phenotypes onto these data spiking in effect sizes that are bigger than in real data; in this way associations are detected in a data set of this size.

Analysis

Analyse your data set and write a detailed report about all your analyses. This should include

- SNP QC
- Sample QC (*excluding checking ancestry – you can assume this has been done already*)
- Genome-wide association analysis of the three traits
- Describe the most associated region of the quantitative trait
- A comparison of the results for these different traits sets demonstrating an understanding of how the results from the three traits relate to each other.

Report

Write a report, of approximately 4 pages (excluding Supplementary data), in the format of a technical report: Intro, Methods, Results, Discussion and Supplementary material.

The Intro should be short, briefly summarising the project.

The Methods section of your report should include a description of the preparation of the data, the statistical models used in your analysis and of the statistical tests conducted. Provide details and explain every analysis. For example, any limitations in the data that affect your ability to perform standard QC should be described.

Results should only present important results of your analysis. This is usually written in a "stating" fashion, i.e. no judgements, just facts (we found, we observed, etc). Your results can be illustrated by important figures (usually 2-3, e.g. a Manhattan plot of your analysis). Non-important figures (e.g. a figure on missingness percentage) will go into Supplementary material.

Supplementary material: Include the relevant part of your code (please do NOT include log files). It should include your plink and R commands so they can be checked and reproduced if needed. Combine everything into a single document. This is the only part of your report that should include code. Extra figures should be included in this part.

Please combine everything into a single pdf and submit through Blackboard

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

Technical Aspects

Access to the cluster.

The School of Mathematics and Physics (SMP) cluster is accessible by ssh as follows, where you replace xxx by your UQ username (e.g. s41234567)

```
> ssh xxx@getafix.smp.uq.edu.au
UQ password.
DUO passcode (don’t forget to refresh).
```

You should now be in your home directory `/home/xxx`

Where is the data?

The genomic data is stored in `/data/STAT3306/Project/Data`
The phenotypes are stored in `/data/STAT3306/Project/Phenotypes`

The STAT3306 folder is secured and only STAT3306-7306 students can access it. All students have access to the same genomic data.

How do I run analyses?

Any computationally heavy jobs (e.g. running plink) should be submitted through the job server. Less computational tasks (e.g. testing an R script on a subset of the data) can be run in an interactive session.

When you submit a job to the server, this job is assigned a number and goes into the queue system. Depending on the resources you ask (time, memory) your job will allocated priority in the queue. When your number is up, your job starts. When your job is complete (either because it ran until completion or it crashed), two text files (*.e* and *.o*) are generated showing you the outputs of how your script ran.

Example job scripts are provided in the `/data/STAT3306/example` folder. You can see how to submit a job that runs R code or plink. It is strongly advised to copy/paste the examples in your home directory and to try it.

Useful commands:

```
> sbatch example_Rcode.slurm      # submit a job
> squeue -u <username>            # check your jobs
> scancel [jobid]                 # cancel a job
```

Using R on getafix
To use R on the SMP cluster, you firstly need to load the module:

```
> module load R
```

This is necessary in any job submissions scripts you submit.
To install a library in R (e.g. qqman), you need to run R from the login node and answer "yes" when asked use a personal library:

```
> install.packages("qqman")
```

Where do I save my results?
You can save your results in your `/home/<user>` directory. You're of course welcome to create directories as needed in your home directory.

If you need to subset the data (removing samples or removing SNPs), you should store the results in your data directory `/data/<user>`
You should not save **ANYTHING** in the `/data/STAT3306/` folder.

How do I transfer files from the clusters?
If you want to copy files from the cluster onto your computer (only results, not the data!), you can either use a command line program (scp or sftp), or use a software like Filezilla (Google will help you if you're not familiar with that sort of software)