RCSI UNIVERSITY OF MEDICINE AND HEALTH SCIENCES

GWAS Lab 1 Assignement

ATT Year 3

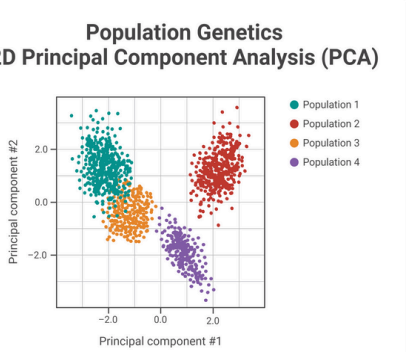
Advanced Genetics and Cell Biology

Background

Genome-wide association studies (GWAS) are used to identify associations between single nucleotide polymorphisms (SNPs) and phenotypic traits.

In order for the results of a GWAS to be reliable, proper quality control (QC) of the data must be performed. This is including but is not limited to, missingness filters (individual- and SNP-based), minor allele frequency (MAF) thresholding, autosomal SNP selection, removal of SNPs deviating from Hardy-Weinberg Equilibrium (HWE), and the removal of genetic outliers.

In order to identify genetic outliers, one must perform a principal component analysis (PCA). The PCA also allows the generation of the principal components as covariates for a GWAS regression.



Resources

- Lecture material (Years 2&3)
- Practical Lab Manual
- Human Molecular Genetics (Chapter 18)
- Quality Control Procedures for Genome Wide Association Studies (Advanced)
- Genome-Wide Association Studies

RCSI UNIVERSITY OF MEDICINE AND HEALTH SCIENCES

Assignment

Using **RStudio**, complete the same QC steps we followed in the GWAS lab session. Please complete all QC steps from the flow chart at the end of this document. Calculate the first two principal components of variation to exclude any genetic outliers. Please also determine what population supergroup the samples correspond to in PC space using the 1000G data.

Images can be saved in RStudio using **Export** in the right-hand panel.

All text files (e.g. genotype files) can be saved out using the command: *write.table(object, "filename", quote = FALSE, row.names=FALSE)*

The input files you will need are available here:

- Target genotype matrix (with ID and Phenotype information) -> **assignment_pop.raw**
- Target Bim file -> **assignment_pop.bim**
- Population Codes -> **population_file_lab_1.txt**
- 1000G genotype matrix -> **lab_1_1000G_cleaned.raw**

The QC Steps should include:

- Remove individuals with genotype missingness of > 1%
- Remove SNPs with missingness of >1%
- Select autosomal SNPs only
- Remove SNPs with MAF <2.5%
- Apply a first HWE filter (1e-5) for the control SNPs across the dataset.
- Apply a second HWE filter (1e-10) for all SNPs across the dataset.
- Remove genetic outliers (+- 6 SD from mean PC1 or PC2 score)

Write a **short** report with following:

- Introduction
  - E.g. What is GWAS, why perform quality control, what is the purpose of PCA?
- Methods
  - E.g. What QC steps did you take?
- Results
  - E.g. The graphs from the analysis with figure legends.
  - Please also upload your final post-QC genotype matrix and the computed PC Scores.
- Discussion
  - E.g. What do you learn from the steps taken? Were you happy with the end data?
- Conclusions
  - E.g. Is it ok to proceed with the actual GWAS regression now?

Please reference your work correctly and thoroughly if consulting outside material (Vancouver Style).

RCSI UNIVERSITY OF MEDICINE AND HEALTH SCIENCES

Deadline

Your report is due **Sunday December 1st at 5PM.**

Word Count

There is no specific word count but the report should contain the following graphs:

- Histogram of individual missingness
- Histogram of SNP missingness
- Histogram of MAF distribution
- PCA Plot (Target Population only)
- PCA Plot (Target Population with coloured 1000G Data)
- You should also upload your computed target-only PC scores and your final post-QC genotype matrix (target-only, before 1000G merge).

Referencing

Please reference using Vancouver style (guide here).

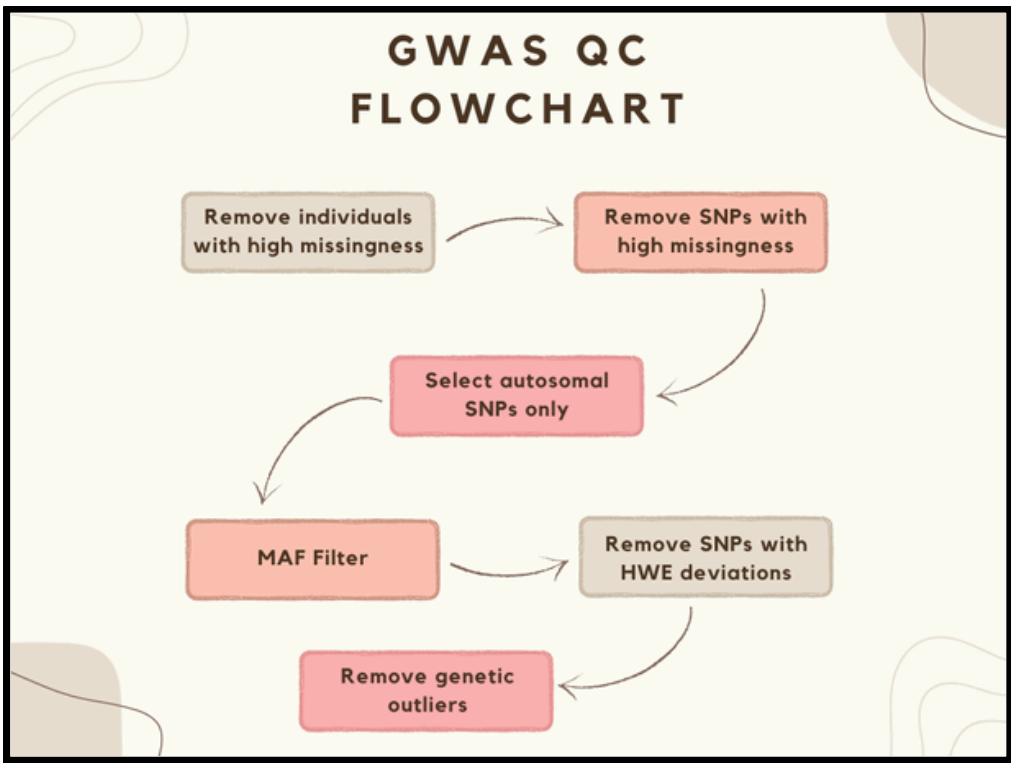
Grading

Your assignment will be graded according to the following rubric:

Evaluation Criteria	Marks
Implementation of data of QC and PCA calculation	20
Sufficient description of the graphs and QC results	30
Critical evaluation of the graphs and QC results	30
Referencing	10
Clarity and presentation	10

Contact

Please email ciarankelly@rcsi.ie with the subject title “ATT GWAS lab report” if you have any questions.


GWAS QC
FLOWCHART
Remove individuals with high missingness
Remove SNPs with high missingness
Select autosomal SNPs only
MAF Filter
Remove SNPs with HWE deviations
Remove genetic outliers