THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Genome-wide Association Studies

Practical 3: Do the GWAS with relatives

Data Use Agreement

- To maximize your learning experience, we will be working with genuine human genetic data
- Access to this data requires agreement to the following in to comply with human genetic data ethics regulations
- Please email pctgadmin@imb.uq.edu.au to confirm that you agree with the following:
  - “I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

Objective

The objective of this practical is to run a GWAS using a sample where relatives are included in the analysis.

We will use the sparse-GRM implemented in GCTA to account for the covariance between 'close' relatives ($\pi > 0.05$)

Data

- Data for this practical is found in the directory:
  - /data/module1/7_relGWAS/

- Genotype & phenotype files:
  - data.bed → binary file containing genomic relationship matrix
  - data.bim → binary file with number of SNP markers used in GRM
  - data.fam → individual ID's corresponding to grm files
  - simData2.phen → phenotype file
- GRM files:
  - data2.grm.bin → binary file containing genomic relationship matrix
  - data2.grm.N.bin → binary file with number of SNP markers used in GRM
  - data2.grm.id → individual ID's corresponding to grm files

GCTA

- We will be using GCTA for this practical
  - Comprehensive website: https://yanglab.westlake.edu.cn/software/gcta/#Overview
  - ‘Citations’ section on how to cite and key papers
- Similar to PLINK, basic command:

```
– gcta64 --bfile <data prefix> --command
```

nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-019-0530-8

A resource-efficient tool for mixed model association analysis of large-scale data

Longda Jiang[1,4], Zhili Zheng[1,2,4], Ting Qi[1], Kathryn E. Kemper[1], Naomi R. Wray[1,3], Peter M. Visscher[1] and Jian Yang[1,2*]

The genome-wide association study (GWAS) has been widely used as an experimental design to detect associations between genetic variants and a phenotype. Two major confounding factors, population stratification and relatedness, could potentially lead to inflated GWAS test statistics and hence to spurious associations. Mixed linear model (MLM)-based approaches can be used to account for sample structure. However, genome-wide association (GWA) analyses in biobank samples such as the UK Biobank (UKB) often exceed the capability of most existing MLM-based tools especially if the number of traits is large. Here

Quick look to see what data you have

- *How many individuals & SNPs in the dataset?*

- *How many individuals are included in the GRM?*

Use the UNIX commands:

```
head file.txt
```

```
wc -l file.txt
```

[word count, count number of lines (only) as the flag]

(1) Identifying relatives & an unrelated set

- Use GCTA at the command line with the `--grm-singleton` flag, e.g.

```
gcta64 --grm data2 --grm-singleton 0.05 --out relatives
```

Three files produced:

- relatives.family.txt → all relative pairs and their relationship value
- relatives.singleton.txt → all ‘singletons’, no relatives in the dataset
- relatives.log → log file

(1) Identifying relatives & an unrelated set

We are running this command just to see what the data is and how many relatives we have in our dataset. Have a look at the output.

Q: How many individuals do you have in each set?

Q: How is the number of individuals in the `XX.singleton.txt` *related to those obtained with the* `--grm-cutoff` flag?

[please don't run this command, check the GCTA website if unsure]

Using fastGWA

We are going to use the `--fastGWA-lr and --grm-sparse` flags in GCTA to fit a <u>sparse</u> genomic relationship matrix (GRM) to model the covariance between closely related individuals

$$y = \alpha + x\beta + g + e$$

phenotypes

intercept

genotypes

SNP effect

(random) additive genetic effect

error

(2) Making a sparse matrix

- Use GCTA at the command line with the `--make-bK-sparse` flag, e.g.

```
gcta64 --grm data2 --make-bK-sparse 0.05 --out data2_sparse
```

Three files produced:

- data2_sparse.grm.sp → index and relationships over 0.05 from GRM
- data2_sparse.grm.id → corresponding ID file
- data2_sparse.grm.lod → log file

Use the R/unix to investigate your output.

Q: Why are the number of lines in the sparse GRM different from the families.txt file obtained previously?

(3) running fastGWA

- Use GCTA at the command line with the `--fastGWA-mla` and `--grm-sparse` flag, e.g.

```
gcta64 --bfile data --fastGWA-mlm --grm-sparse data2_sparse
--pheno simData3.phen --out assocSparse
```

Plot results in R, examine the QQ-plot & calculate the genomic inflation factor

- Reminder: gif = qchisq(1-median(p),1)/qchisq(0.5,1)

(4) compare your results

- Compare your results to an analysis where you ignored close relatives, use either PLINK or GCTA to do a standard GWAS

```
plink --bfile data --assoc --pheno simData3.phen --out
assocPLINK
```

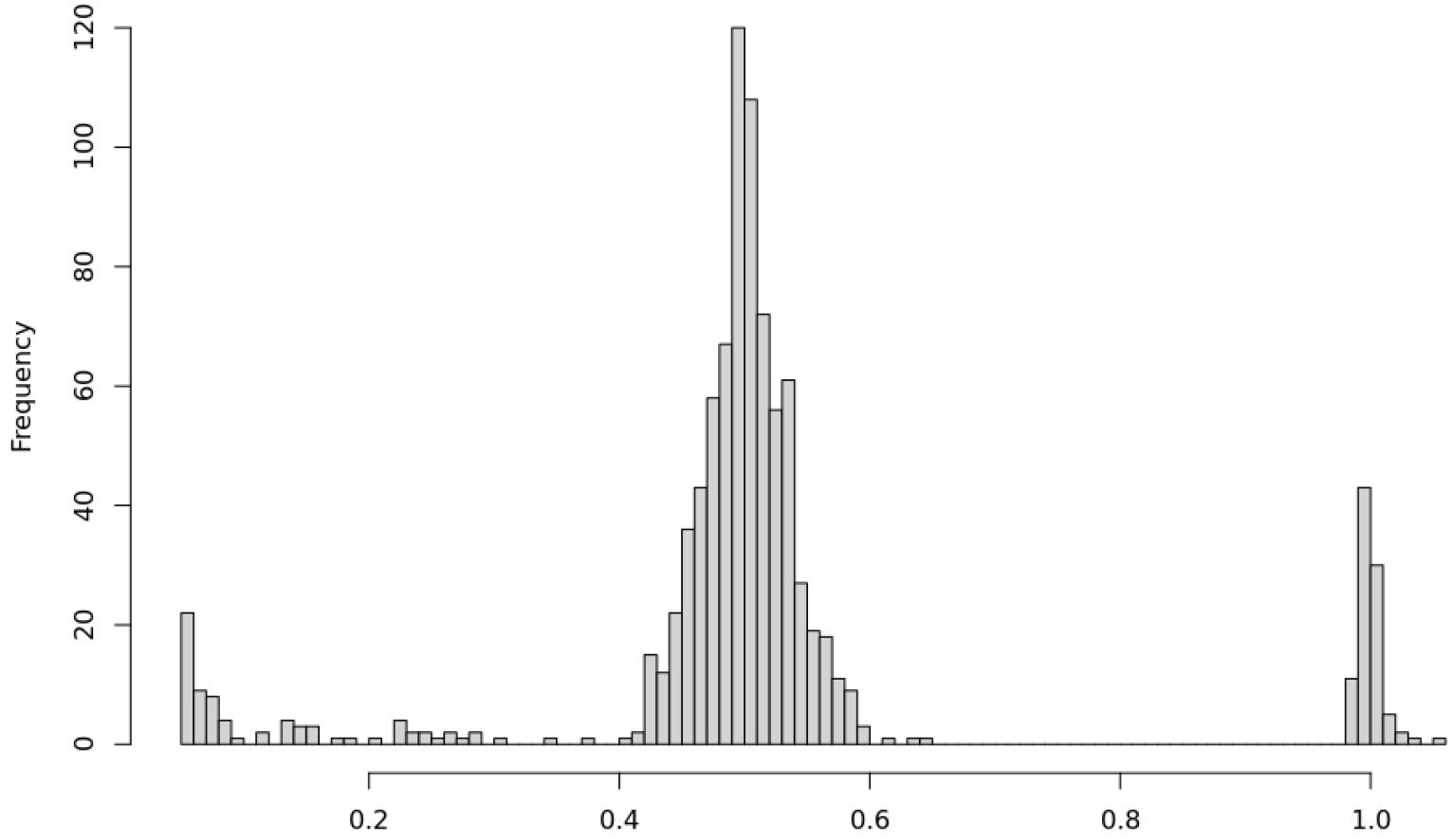
OR

```
gcta64 --bfile data --fastGWA-lr --pheno simData3.phen --out
assocGCTA
```

STOP here

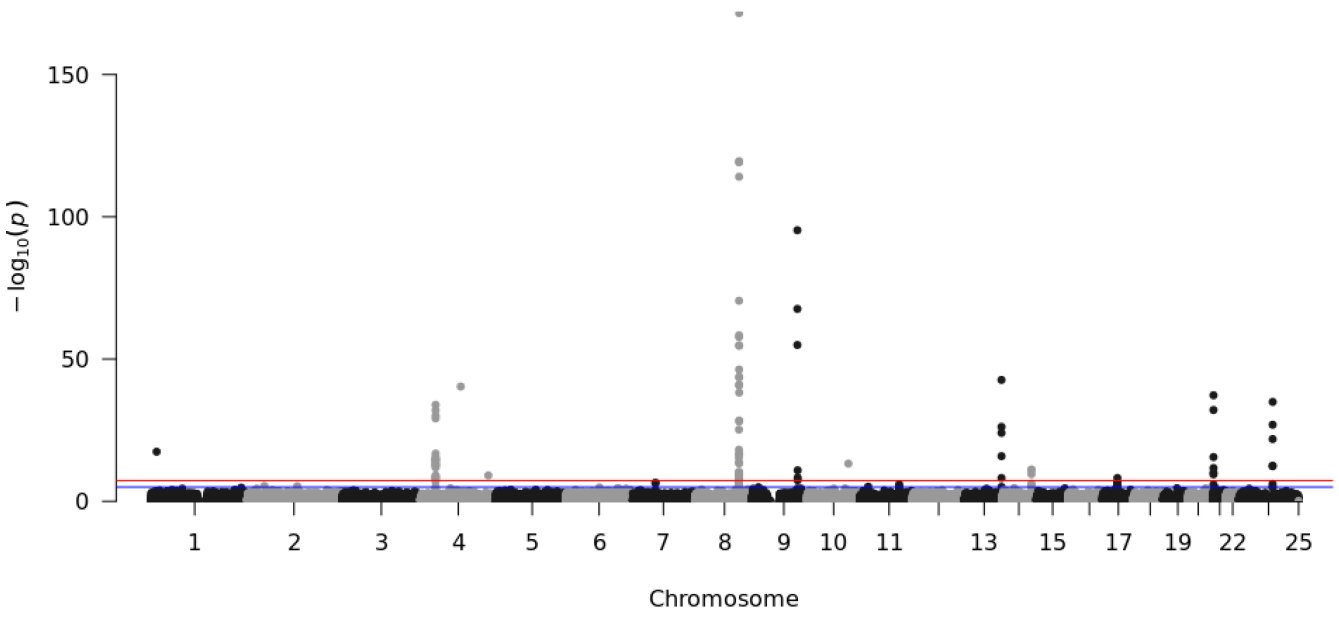
(2) Making a sparse matrix

A histogram of the elements in the sparse matrix

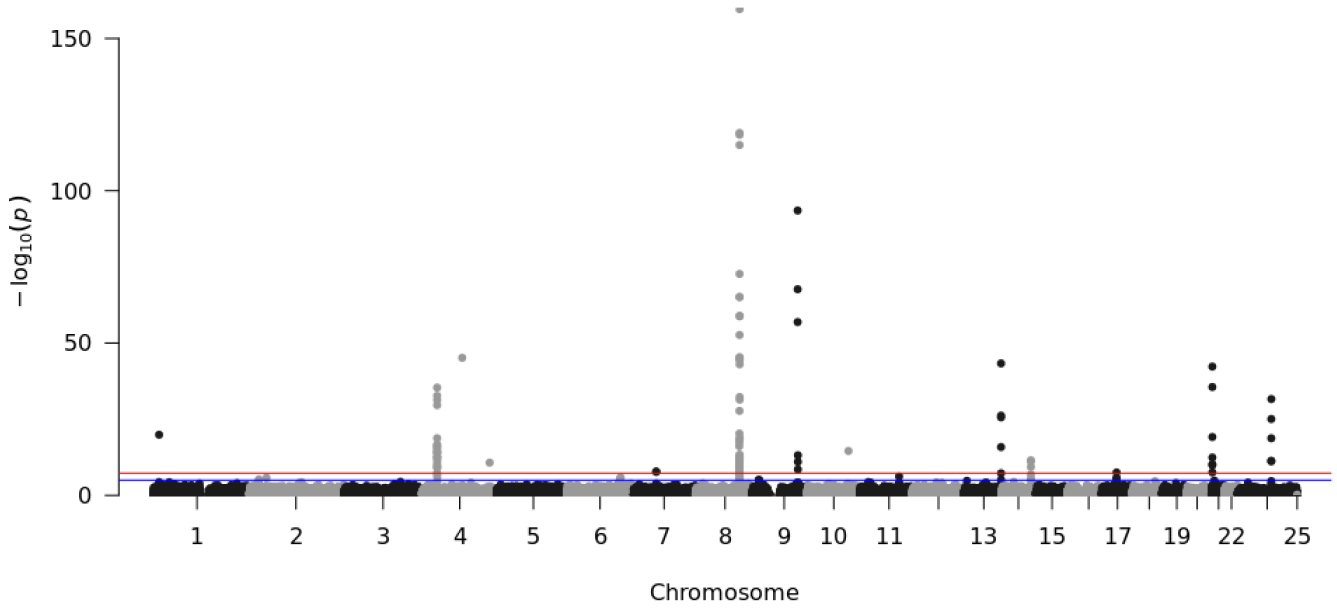


(4) compare your results

- GCTA-lr:

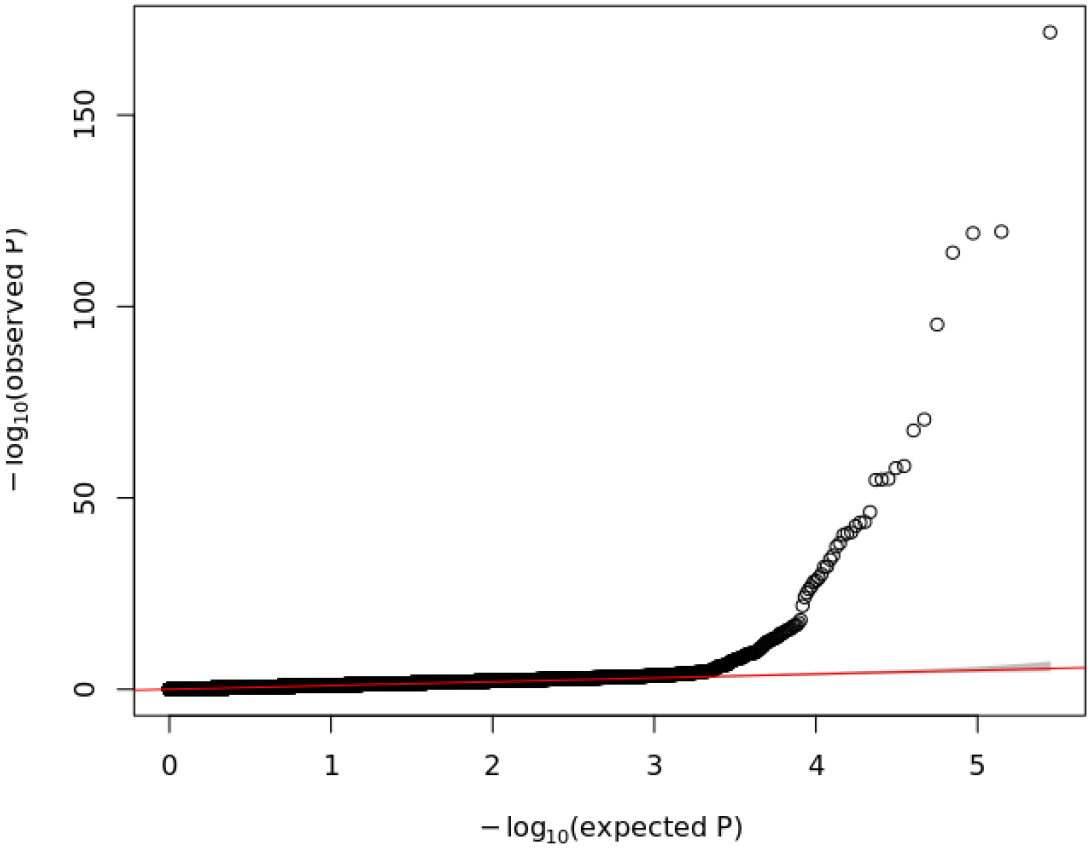


- GCTAsparse:

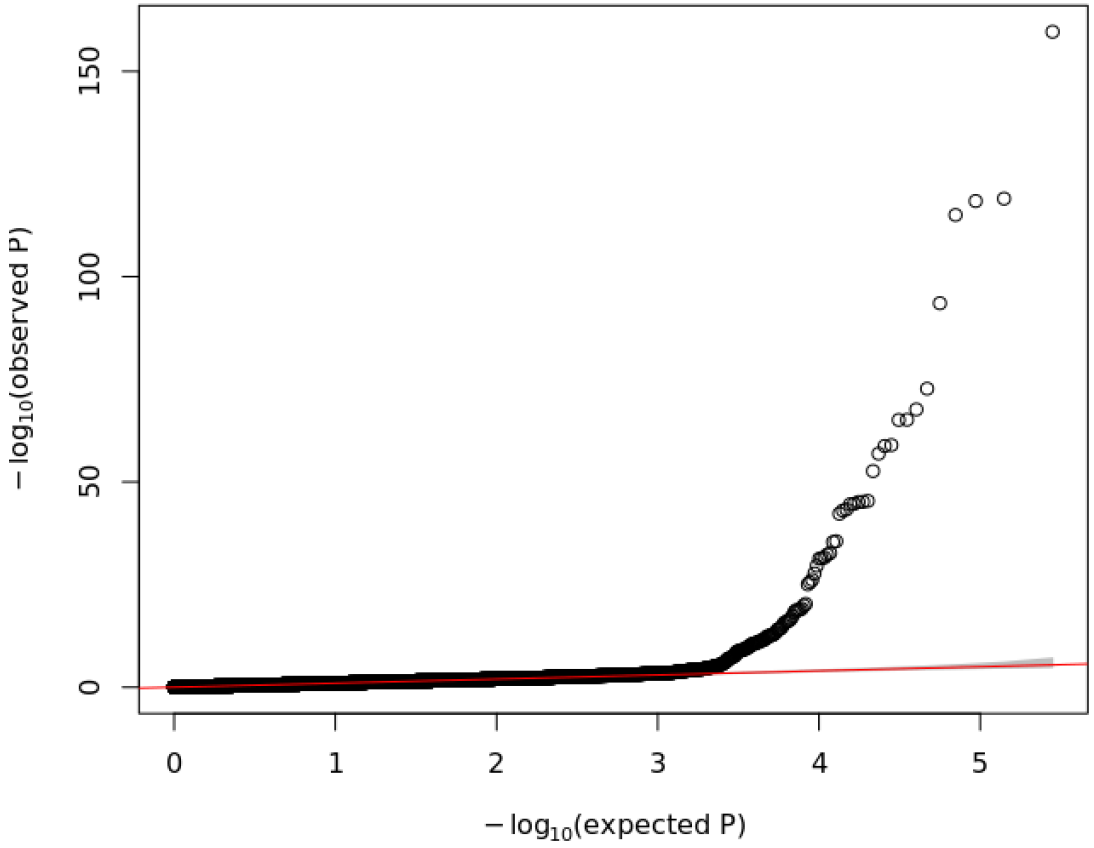


(4) compare your results

- PLINK:



- GCTAsparse:



(4) compare your results

- PLINK, GIF = 1.109886
- GCTAsparse, GIF = 1.007401

- The effect of fitting the sparse GRM is subtle in these results

  -> most obvious in the GIF

- This prac used a <u>simulated</u> phenotype where there was a small common environmental effect between relative pairs

Q: Do we expect inflation of the test statistic in the absence of a common environment?

e.g. imagine that you were analysing data from IVF siblings raised independently