Advanced Gene Mapping Course

The Rockefeller University

January 27-31, 2025

Computer Exercises

Table of Contents

N/A exercise instructions are included in a Jupyter Notebook - no handout for this exercise.

Getting Started

You will have two ways to run the exercises: on the cloud or locally on your computer.

Cloud Access

Your username on the cloud follows this format:

https://statgenetics.github.io/statgen-courses/<firstname_lastname>

Please use the name from your registration form. For example, if your name is Jane Smith, your link would be:

https://statgenetics.github.io/statgen-courses/jane_smith

Local Installation

You can run the exercises locally on your computer. However, we do not recommend this for Mac M1, M2, or M3 chips since the exercises can run very slowly.

For detailed setup instructions, please visit:

https://github.com/statgenetics/statgen-courses/wiki/How-to-launch-course-tutorials#running-on-local-computer-with-software-installed

All software used, with the exception of ANNOVAR, are available on anaconda.org. To run locally, you need to set up the course software using **pixi**, a Conda package management tool. The link above provides instructions for setup and installation. Currently, all software used in the course are available on Linux OS. Most are also available on Mac and Windows. If you have a Mac with an Intel chip, you can install Docker on your computer then set up **pixi** in a Docker based Linux virtual machine — that will ensure all exercises work. For Mac M chips, while most exercises work, some will not function even with Docker installed. For those who want to experiment without installing into the operating system of your computer, Docker is also an option.

Docker Installation Guides

If you strongly prefer to install Docker on your computer before installing **pixi** inside Docker, please view these installation guides:

For MAC: https://www.youtube.com/watch?v=DRCDNBlxZ-w

For Windows PC: https://www.youtube.com/watch?v=sxv45NCSFMk

For Ubuntu Linux: https://www.youtube.com/watch?v=3K-sGzxsyK0

Launching Exercises

Please use the following commands to launch the exercises if you are running them locally on your computer. Note you must first set-up all the conda packages (via the pixi package manger) on your computer which was described on the previous page. For cloud computing please just click the appropriate folder to perform each exercise.

To perform the exercises on your computer please open a terminal and type the command `jupyter-lab` to open Jupyter lab. Next, copy the generated link and paste it into your browser. Depending on which exercise you wish to perform, enter one of the folders listed below.

- PLINK GWAS Analysis QC and Substructure
  - plink
- FASTLMM & GCTA*
  - fastlmm_gcta
- Epistasis (PINK & CASSI)*
  - epistatis
- Data Quality Control and Annotation of Sequence data
  - ngs_qc_annotation
- REGENIE
  - regenie
- Power and Sample Size Estimation*
  - Web-based
- TWAS
  - twas
- Fine-mapping (SuSiE)
  - finemapping
- Multivariate fine-mapping (mvSuSiE)
  - multivariate_finemapping
- Pleiotropy*
  - pleiotropy
- Mendelian Randomization Two-Sample*
  - Mendelian_randomization
- Polygenic Risk Score (LDpred2)
  - ldpred2

Run **get-data** command if you don’t see the data for the exercise already loaded.

*There is a handout available for this exercise in this booklet. For all other exercises the instructions to perform the exercise are available in the Jupyter Notebook.

Genome-wide Association Analysis - Data Quality Control

Introduction

In this exercise, you will learn how to perform data quality control (QC) by removing markers and samples that fail QC quality control criteria. You will also examine your samples for individuals that are related to each other and/or are duplicate samples. Each sample will also be tested for excess homozygosity and heterozygosity of genotype data. Each SNP will be tested for deviations from Hardy-Weinberg Equilibrium. These exercises will be carried out using PLINK1.9 and R.

1. Using PLINK

PLINK can upload data in different formats please see the PLINK documentation (https://www.cog-genomics.org/plink/1.9/input) for additional details. The data for this exercise is in PLINK/LINKAGE file format. There are two files: a pedfile (GWAS.ped) and a map file (GWAS.map). Please examine these files and the PLINK documentation. Please note the commands must be given in the directory where the data residues.

Navigate via the command prompt to the directory which contains the files for the exercise. Type **plink** in the command prompt and make note of the output. Next type:

```
plink --file GWAS
```

Note, that PLINK outputs a file called **plink.log** that contains the same output which you see on the screen. To see all options, type plink --help for more information. Determine how many samples there are in your data set and fill in Oval 1 of the flowchart below.

2. Data Quality Control

a. Removing Samples and SNPs with Missing Genotypes.

You will exclude samples that are missing more than 10% of their genotype calls. These samples are likely to have been generated using low quality DNA and can also have higher than average genotyping error rates.

```
plink --file GWAS --mind 0.10 --recode --out GWAS_clean_mind
```

Examine **GWAS_clean_mind.log** to see how many samples are excluded based on this criterion and fill in Box 1.

Create two versions of your dataset, one with SNPs with a minor allele frequencies (MAFs) ≥5% and the other with SNPs with a MAFs <5%.

You will now remove SNPs with MAFs≥5% that are missing >5% of their genotypes and then remove SNPs with MAFs<5% that are missing >1% of their genotypes. SNPs which are missing genotypes can have higher error rates than those SNP markers without missing data.

```
plink --file GWAS_clean_mind --maf 0.05 --recode --out MAF_greater_5
plink --file GWAS_clean_mind --exclude MAF_greater_5.map --recode --out MAF_less_5

plink --file MAF_greater_5 --geno 0.05 --recode --out MAF_greater_5_clean
```

Fill in Box 2a.

```
plink --file MAF_less_5 --geno 0.01 --recode --out MAF_less_5_clean
```

Fill in Box 2b.

Merge the two files.

```
plink --file MAF_greater_5_clean --merge MAF_less_5_clean.ped MAF_less_5_clean.map --
recode --out GWAS_MAF_clean
```

A more stringent criterion for missing data is used, samples missing >3% of their genotypes are removed.

```
plink --file GWAS_MAF_clean --mind 0.03 --recode --out GWAS_clean2
```

Fill in Box 3.

b. Checking Sex

Error of the reported sex of an individual can occur. Information from the SNP genotypes can be used to verify the sex of individuals, by examining homozygosity (F) on the X chromosome for every individual. F is expected to be <0.2 in females and >0.8 in males. To check sex run

```
plink --file GWAS_clean2 --check-sex --out GWAS_sex_checking
```

Use R to examine the GWAS_sex_checking.sexcheck file and determine if there are individuals whose recorded sex is inconsistent with genetic sex.

```
R
sexcheck = read.table("GWAS_sex_checking.sexcheck", header=T)
names(sexcheck)
sex_problem = sexcheck[which(sexcheck$STATUS=="PROBLEM"),]
sex_problem
q()
```

NA20530 and NA20506 were coded as a female (2) and from the genotypes appear to be males (1). In addition, 3 individuals (NA20766, NA20771 and NA20757) do not have enough information to determine if they are males or females and PLINK reports sex = 0 for the genotyped sex. Fill in the table below:

Table 1: Sex check

FID	IID	PEDSEX	SNPSEX	STATUS	F
NA20506	NA20506				
NA20530	NA20530				
NA20766	NA20766				
NA20771	NA20771				
NA20757	NA20757				

Reasons for these kinds of discrepancies, include the records are incorrect, incorrect data entry, sample swap, unreported Turner or Klinefelter syndromes. Additionally, if a sufficient number of SNPs have not been genotyped on the X chromosome it can be difficult to accurately predict the sex of an individual. In this dataset, there are only 194 X chromosomal SNPs. If you cannot validate the sex of the individual they should be removed. For this exercise, we are going to assume that when the sex was checked, we found it was incorrectly recorded (i.e. these samples were male). Therefore, this error could simply be corrected.

Question 1: Why do you expect the homozygosity rate to be higher on the X chromosome in males than females?

c. *Duplicate Samples*

The following PLINK command can be used to check for duplicate samples:

```
plink --file GWAS_clean2 --genome --out duplicates
```

Open the **duplicates.genome** file in R with the following command:

```
dups = read.table("duplicates.genome", header = T)
```

We are interested in the Pi-Hat (the estimated proportion IBD sharing) value. You may notice that there is more than one duplicate (Pi-Hat=~1). Also, examine the output for pairs of individuals with high Pi-Hat values which can indicate they are related. The amount of allele sharing [Z(0), Z(1) and Z(2)] across all SNPs provides information on the type of relative pair.

```
problem_pairs = dups[which(dups$PI_HAT > 0.4),]
problem_pairs
```

Table 2: Duplicate and Related Individuals

FID1	IID1	FID2	IID2	Z(0)	Z(1)	Z(2)	PI_HAT

F1D1- Family ID for 1st individual; ID1 - Individual ID for 1st individual; F1D2- Family ID for 2nd individual; ID2 - Individual ID for 2nd individual; Z(0)- P(IBD=0); Z(1)- P(IBD=1); Z(2)- P(IBD=2); PI_HAT-P(IBD=2)+0.5*P(IBD=1) (proportion IBD)

__Question 2:__ How many duplicate pairs do your find (**hint: Pi-Hat = ~1**)? Do pairs with a **Pi-Hat = ~1** have to be duplicate samples? What is another explanation? What proportion would you expect a parent/ child to share IBD? Can you find any such relationship?

Note: Pi-hat can be inflated and individuals appear to be related to each other if you have samples from different populations. This explains why we observe pairs of individuals with Pi-hat >0.05 since three distinct populations were analyzed. Additionally, this phenomenon can be observed if a subset(s) of samples have higher genotyping/sequencing error rates, which creates two or more "populations" and the individuals within these "populations" incorrectly appear to be related.

Using this R script please observe how many sample pairs have pi-hat >0.05:

```
problem_pairs = dups[which(dups$PI_HAT > 0.05),]
myvars = c("FID1", "IID1", "FID2", "IID2", "PI_HAT")
problem_pairs[myvars]
```

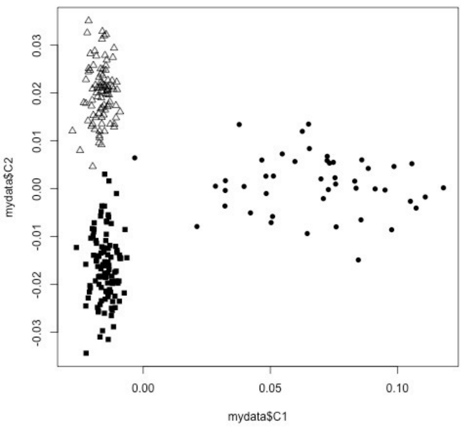
Create the following txt file:

1344 NA12057
1444 NA12739
M033 NA19774

name it 'IBS_excluded.txt' and save it to the folder with your PLINK data. Give the command:

```
plink --file GWAS_clean2 --remove IBS_excluded.txt --recode --out GWAS_clean3
```

Fill in Box 4 and Oval 3.



As part of QC usually the data is examined for outliers by plotting the first and second principal or multidimensional scaling (MDS) components. Using a subset of markers that have been trimmed to remove LD (r^2<0.5). Principal components analysis (PCA) and MDS will be performed in the second part of the exercise to detect outliers and control for populations substructure. Outlier can be due to study subjects coming from different populations e.g. European- and African-Americans or batch effects. If it is suspected that outliers are due to study subjects having been sampled from different populations than data from HapMap can be included to elucidate population membership, e.g. for a study of European-Americans if African-American study subjects are included they would cluster between the European and African HapMap samples. If you perform this type of analysis you should remove the HapMap samples and re-estimate the MDS or PC components before adjusting for population substructure or stratification. For this exercise data **is used** from HapMap Phase III which consists of CEU (Europeans from Utah), MEX (Mexicans from Los Angeles) and TSI (Tuscans from Italy). Three clusters can be observed that consist of the three data sets but no extreme outliers are observed. This data set is being used for demonstration purposes. Different populations should be analyzed separately and the results can be combined using meta-analysis. In part two of this exercise MDS and PC components will be constructed and analyzed.

d. Hardy-Weinberg Equilibrium (HWE):

To test for HWE we will test separately in each ancestry group and by case-control status. Therefore, we will need to use information on ancestry and cases-control status. Please note that this should be tested in the 3 different populations separately (CEU, MEX, TSI), but due to the small sample sizes, we tested it in the 3 populations together for example purposes. It should also be noted if the sample sizes are small it is difficult to detect a deviation from HWE.

```
plink --file GWAS_clean3 --pheno pheno.txt --pheno-name Aff --hardy
```

Using R examine the file **plink.hwe** and look for SNPs with p-values of $5x10^{-8}$ or smaller.

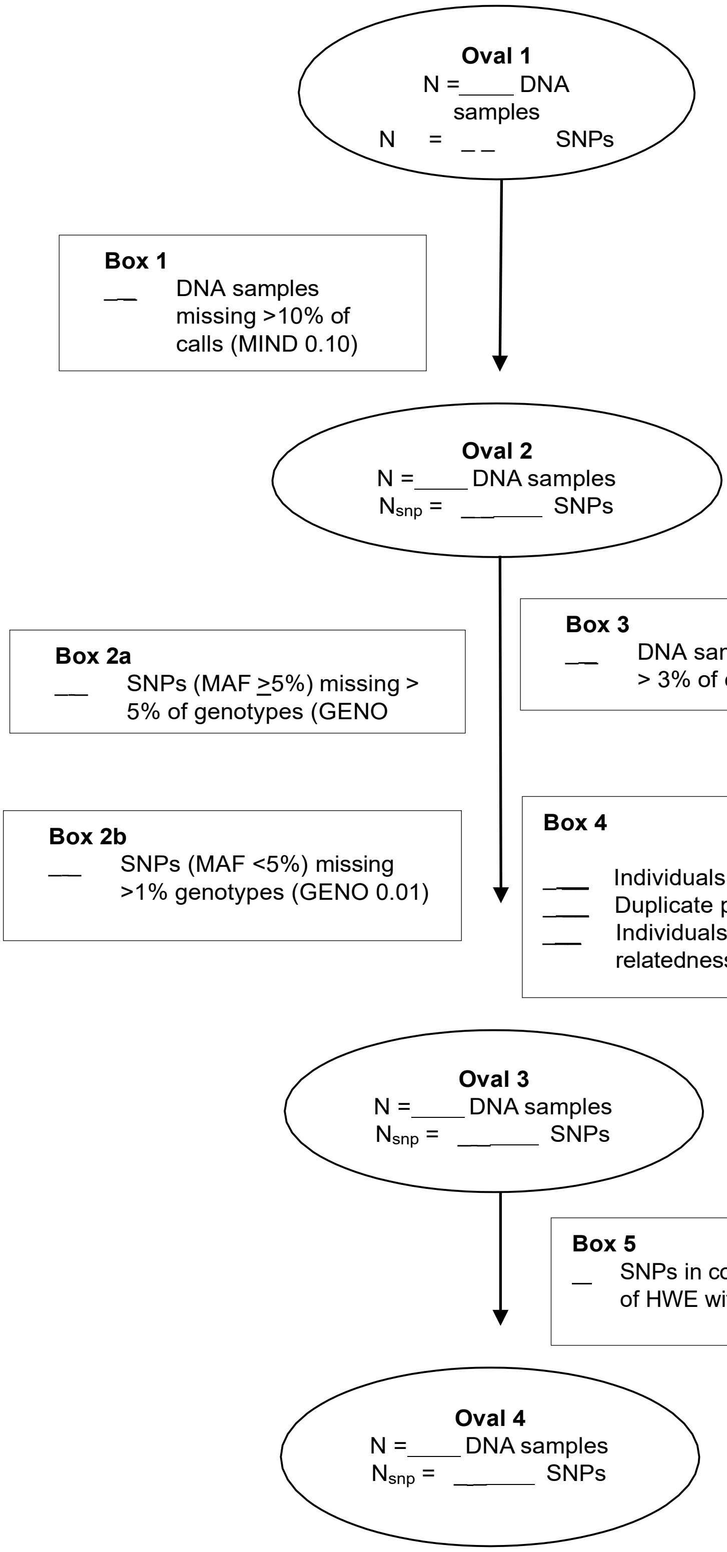
```
hardy = read.table("plink.hwe", header = T)
names(hardy)
hwe_prob = hardy[which(hardy$P < 0.00000005),]
hwe_prob
```

Using a criterion of p <$5x10^{-8}$ to reject the null hypothesis of HWE, how many SNPs fail HWE in the controls? Fill out Oval 5 and Box 4. Using the same criteria, how many SNPs fail HWE in the controls? Complete Table 2 with this information.

Table 3: Hardy-Weinberg Equilibrium

Cases			Controls		
SNP	Pvalue	Population(s)	SNP	Population(s)	Pvalue

There are a number of SNPs with HWE p-values in the range of 10^{-5} to 10^{-6} in the controls. Based on above criterion they will not be excluded however, if they reach genome-wide significance during association testing they SNPs should be further investigated to ensure there is no genotyping error. You can now fill in Box 5 and Oval 4.



Oval 1
N =____ DNA samples
N = _ _ SNPs

Box 1
___ DNA samples missing >10% of calls (MIND 0.10)

Oval 2
N =____ DNA samples
N_{snp} = _______ SNPs

Box 2a
___ SNPs (MAF ≥5%) missing > 5% of genotypes (GENO

Box 2b
___ SNPs (MAF <5%) missing >1% genotypes (GENO 0.01)

Box 3
___ DNA samples failed missing > 3% of calls (MIND 0.03)

Box 4
____ Individuals with inconsistent
____ Duplicate pairs
____ Individuals excluded due to relatedness

Oval 3
N =____ DNA samples
N_{snp} = _______ SNPs

Box 5
__ SNPs in controls out of HWE with $p<10^{-7}$

Oval 4
N =____ DNA samples
N_{snp} = _______ SNPs

Answers to Questions:

Oval 1 and 2 also and Box 1 information:

```
PLINK v1.90b4.9 64-bit (13 Oct 2017)
Options in effect:
  --file GWAS
  --mind 0.10
  --out GWAS_clean_mind
  --recode

Random number seed: 1515434515
16384 MB RAM detected; reserving 8192 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (6424 variants, 248 people) [Oval 1].
--file: GWAS_clean_mind-temporary.bed + GWAS_clean_mind-temporary.bim +
GWAS_clean_mind-temporary.fam written.
6424 variants loaded from .bim file.
248 people (125 males, 123 females) loaded from .fam.
1 person removed due to missing genotype data (--mind) [Box 1].
ID written to GWAS_clean_mind.irem .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 247 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 6 het. haploid genotypes present (see GWAS_clean_mind.hh ); many
commands treat these as missing.
Total genotyping rate in remaining samples is 0.996863.
6424 variants and 247 people pass filters and QC [Oval 2].
Note: No phenotypes present.
--recode ped to GWAS_clean_mind.ped + GWAS_clean_mind.map ... done.
```

Box 2a information:

```
PLINK v1.90b4.9 64-bit (13 Oct 2017)
Options in effect:
  --file MAF_greater_5
  --geno 0.05
  --out MAF_greater_5_clean
  --recode

Random number seed: 1515435189
16384 MB RAM detected; reserving 8192 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (5868 variants, 247 people).
--file: MAF_greater_5_clean-temporary.bed + MAF_greater_5_clean-temporary.bim +
MAF_greater_5_clean-temporary.fam written.
5868 variants loaded from .bim file.
247 people (125 males, 122 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 247 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 6 het. haploid genotypes present (see MAF_greater_5_clean.hh ); many
commands treat these as missing.
Total genotyping rate is 0.996858.
2 variants removed due to missing genotype data (--geno) [Box2a].
5866 variants and 247 people pass filters and QC.
Note: No phenotypes present.
--recode ped to MAF_greater_5_clean.ped + MAF_greater_5_clean.map ... done.
```

Box 2b information:

```
PLINK v1.90b4.9 64-bit (13 Oct 2017)
Options in effect:
  --file MAF_less_5
  --geno 0.01
  --out MAF_less_5_clean
  --recode

Random number seed: 1515435255
16384 MB RAM detected; reserving 8192 MB for main workspace.
Scanning .ped file... done.
```

```
Performing single-pass .bed write (556 variants, 247 people).
--file: MAF_less_5_clean-temporary.bed + MAF_less_5_clean-temporary.bim +
MAF_less_5_clean-temporary.fam written.
556 variants loaded from .bim file.
247 people (125 males, 122 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 247 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.996913.
59 variants removed due to missing genotype data (--geno) [Box2b].
497 variants and 247 people pass filters and QC.
Note: No phenotypes present.
--recode ped to MAF_less_5_clean.ped + MAF_less_5_clean.map ... done.
```

Box 3 information:

```
PLINK v1.90b4.9 64-bit (13 Oct 2017)
Options in effect:
  --file GWAS_MAF_clean
  --mind 0.03
  --out GWAS_clean2
  --recode

Random number seed: 1515435827
16384 MB RAM detected; reserving 8192 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (6363 variants, 247 people).
--file: GWAS_clean2-temporary.bed + GWAS_clean2-temporary.bim +
GWAS_clean2-temporary.fam written.
6363 variants loaded from .bim file.
247 people (125 males, 122 females) loaded from .fam.
0 people removed due to missing genotype data (--mind) [Box 3].
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 247 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 6 het. haploid genotypes present (see GWAS_clean2.hh ); many commands
treat these as missing.
Total genotyping rate is 0.99716.
6363 variants and 247 people pass filters and QC.
Note: No phenotypes present.
--recode ped to GWAS_clean2.ped + GWAS_clean2.map ... done.
```

Answer to Question 1: Why do you expect the homozygosity rate to be higher on the X chromosome in males than females?

Because males only have one allele for each SNP on the X chromosome they will appear homozygous.

Table 1: Sex check

FID	IID	PEDSEX	SNPSEX	STATUS	F
NA20506	NA20506	2	1	PROBLEM	1
NA20530	NA20530	2	1	PROBLEM	1
NA20766	NA20766	2	0	PROBLEM	0.2292
NA20771	NA20771	2	0	PROBLEM	0.2234
NA20757	NA20757	2	0	PROBLEM	0.2141

Table 2: Duplicate and Related Individuals

FID1	IID1	FID2	IID2	Z(0)	Z(1)	Z(2)	PI_HAT
M033	NA19774	M041	NA25000	0.0000	0.0000	1.0000	1.00
1344	NA12057	13291	NA25001	0.0000	0.0025	0.9975	1.00
1444	NA12739	1444	NA12749	0.0026	0.9807	0.0168	0.51
1444	NA12739	1444	NA12748	0.0026	0.9949	0.0025	0.50

F1D1- Family ID for 1st individual; ID1 - Individual ID for 1st individual; F1D2- Family ID for 2nd individual; ID2 - Individual ID for 2nd individual; Z(0)- P(IBD=0); Z(1)- P(IBD=1); Z(2)- P(IBD=2); PI_HAT-P(IBD=2)+0.5*P(IBD=1) (proportion IBD)

Question 2: How many duplicate pairs do your find (**hint: Pi-Hat = ~1)**? Do pairs with a **Pi-Hat = ~1** have to be duplicate samples? What is another explanation? What proportion would you expect a parent/ child to share IBD? Can you find any such relationship?

There are two duplicate pairs and also a trio (two parents and a child). Parent/child relationships will have a Pi_Hat value of ~0.5, but so will sibpairs. We can tell that this is a parent child relationship by examine Z(0), Z(1) and Z(2). We will retain only one sample from each duplicate pair and the parents NA12749 and NA12748. If you perform mixed-model analysis related individuals can be retained in the sample.

Oval 3 information

```
PLINK v1.90b4.9 64-bit (13 Oct 2017)
Options in effect:
  --file GWAS_clean2
  --out GWAS_clean3
  --recode
  --remove IBS_excluded.txt
Random number seed: 1515440989
16384 MB RAM detected; reserving 8192 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (6363 variants, 247 people).
--file: GWAS_clean3-temporary.bed + GWAS_clean3-temporary.bim +
GWAS_clean3-temporary.fam written.
6363 variants loaded from .bim file.
247 people (125 males, 122 females) loaded from .fam.
--remove: 244 people remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 244 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 6 het. haploid genotypes present (see GWAS_clean3.hh ); many commands
treat these as missing.
Total genotyping rate in remaining samples is 0.997225.
6363 variants and 244 people pass filters and QC [Oval 3].
Note: No phenotypes present.
--recode ped to GWAS_clean3.ped + GWAS_clean3.map ... done.
```

Table 3: Hardy Weinberg Equilibrium

Fail Cases		Fail Controls	
SNP	pvalue	SNP	pvalue
None			

```
PLINK v1.90b4.9 64-bit (13 Oct 2017)
Options in effect:
  --exclude HWE_out.txt
  --file GWAS_clean3
  --out GWAS_clean4
  --recode

Random number seed: 1515442367
16384 MB RAM detected; reserving 8192 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (6363 variants, 244 people).
--file: GWAS_clean4-temporary.bed + GWAS_clean4-temporary.bim +
GWAS_clean4-temporary.fam written.
6363 variants loaded from .bim file.
244 people (123 males, 121 females) loaded from .fam.
--exclude: 6362 variants remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 244 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 6 het. haploid genotypes present (see GWAS_clean4.hh ); many commands
treat these as missing.
Total genotyping rate is 0.997229.
6362 variants and 244 people pass filters and QC [Oval 4].
Note: No phenotypes present.
--recode ped to GWAS_clean4.ped + GWAS_clean4.map ... done
```

Genome-Wide Association Exercise
Association Analysis Controlling for Population Substructure

1. Population Stratification and Association Testing

The dataset from part I of this exercise which you performed data quality control (QC) on was obtained from HapMap Phase III data. It contains CEU founders (Caucasians from Utah), MEX founders (Mexicans from Los Angeles) and TSI (Tuscans from Italy). The CEU pedigree identifiers begin with only numbers e.g., 1347, the MEX pedigree identifies all start with M e.g., M017 and the TSI pedigree identifiers all start with NA e.g., NA0217. Before we start testing for association, we want to know if there are outliers. Even after removing the outliers when association analysis is performed population substructure and admixture may need to be controlled. If not, we risk observing an association, which is due to a difference in genotype frequencies in cases and controls, because of population substructure/admixture and not because of linkage disequilibrium (LD) between tagSNP(s) and the functional variant(s). We are going to use multidimensional scaling (MDS) and principal components analysis (PCA) within the PLINK software to generate 10 components. **Disclaimer: You usually should not analyze data from European-Americans, Mexican-Americans and Italians together even if you control for population stratification. They can be analyzed separately and the data combined using meta-analysis.**

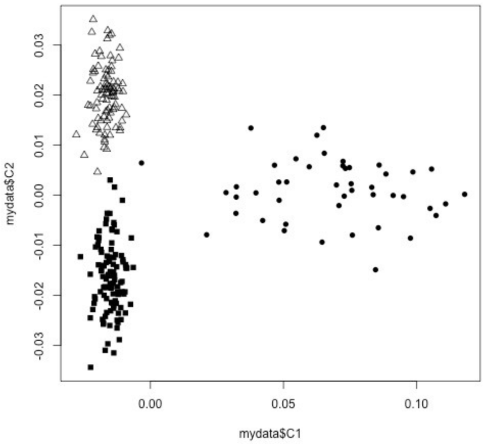
Note: For a GWAS study instead of this toy study, you will have a denser set of markers of which some will be in LD. You should first prune your SNPs to obtain a subset in linkage equilibrium/weak LD (R^2<0.5) prior to performing MDS or PCA analysis on the data. Although for association analysis is performed on the entire data set will be analyzed only this a subset of SNPs which are not in LD will be used to construct PCA and MDS components. For more information on how to do this in PLINK see https://www.cog-genomics.org/plink/1.9/ld.

```
plink --file GWAS_clean4 --genome --cluster --mds-plot 10
```

This command outputs the file **plink.mds** that contains the subject IDs and values for the 10 components we just generated. There is another file in your folder called mds_components.txt. This file is identical to your **plink.mds** file with the exception that a group column which codes CEU individuals as 1, MEX individuals as 2 and TSI individuals as 3. This is done so when we plot the MDS components in R you can see which group the points belong to and judge how well does the data cluster, e.g., are there outliers. The following commands will generate a jpeg image file containing the mds plot (filename=mds.jpeg) in your current working directory. Open R and use the following command:

```
mydata = read.table("mds_components.txt", header=T)


mydata$pch[mydata$Group==1 ] <--15
mydata$pch[mydata$Group==2 ] <--16
mydata$pch[mydata$Group==3 ] <--2

jpeg("mds.jpeg", height=500, width=500)
plot(mydata$C1, mydata$C2 ,pch=mydata$pch)
dev.off()
```

Visualizing population structure using MDS is useful for identifying subpopulations, population stratification and systematic genotyping or sequencing errors, and can also be used to detect individual outliers that may need to be removed, e.g. European-Americans included in a study of African-Americans. MDS coordinates help with visualizing genetic distances and population substructure. PLINK also offers another dimension reduction, --pca, for PCA, the PC components which can also be used for visualizing data to detect outliers in the same manner which was performed using MDS. Additionally, covariates either from either MDS or PCA can be used in a regression model to aid in correcting for population substructure and admixture.



We will now continue performing the analysis using PLINK but will use PCA instead of MDS. We will generate PCs and determine how many PC covariates should be included in the regression model. When SNPs are tested for an association with a trait analysis can be performed, first by including no PC components, then one PC component and then two PC components and so on. Please note that as each PC component is added all the SNPs are analyzed, e.g. a complete GWAS is performed. Examining λ can aid in determining how many PC components should be included in the analysis. If there is no population stratification or other biases, then λ should equal 1 or ~1. We will use λ to determine how many PC components from our analysis will be added to the logistic regression model. First, estimate λ without adjusting for any PC components:

```
plink --file GWAS_clean4 --pheno pheno.txt --pheno-name Aff --logistic --adjust --out
unadj
```

Generated the first 10 PCA values:

```
plink --file GWAS_clean4 --genome --cluster --pca 10 header
```

Eigenvectors are written to plink.eigenvec, and top eigenvalues are written to plink.eigenval. The 'header' modifier adds a header line to the .eigenvec file(s).

And then find out what λ is when we adjust for the first component:

```
plink --file GWAS_clean4 --pheno pheno.txt --pheno-name Aff --covar plink.eigenvec --
covar-name PC1 --logistic --adjust --out PC1
```

And the first and second components:

```
plink --file GWAS_clean4 --pheno pheno.txt --pheno-name Aff --covar plink.eigenvec --
covar-name PC1-PC2 --logistic --adjust --out PC1-PC2
```

and so forth for all 10 components in the .log file completing the table:

Table 1

	Un--adjusted	PC 1	PC 1--2	PC 1--3	PC 1--4	PC 1--5	PC 1--6	PC 1--7	PC 1--8	PC 1--9	PC1--10
λ											

The number closest to 1.0, with the least number of PC components, would be the best for adjusting without overfitting and introducing unnecessary noise. You can check your table against the one provided in the answers section.

Go to the **assoc.logistic file that corresponds to that number of components** and make a note of how you named the **.assoc.logistic** file for it and when you did not adjust for any components. Then go back to the R program to load the results and create a jpeg image file containing QQ plots for the adjusted and unadjusted results (using a modified script from http://www.broad.mit.edu/node/555) as follows:

```
broadqq <--function(pvals, title)
{
      observed <-- sort(pvals)
      lobs <-- --(log10(observed))

      expected <-- c(1:length(observed))
      lexp  <--  --(log10(expected  /  (length(expected)+1)))

      plot(c(0,7), c(0,7), col="red", lwd=3, type="l", xlab="Expected (--logP)", ylab="Observed (--logP)",
xlim=c(0,max(lobs)),  ylim=c(0,max(lobs)),  las=1,  xaxs="i",  yaxs="i",  bty="l",  main  =  title)
      points(lexp,  lobs,  pch=23,  cex=.4,  bg="black") }

jpeg("qqplot_compare.jpeg", height=1000, width=500)
par(mfrow=c(2,1))
aff_unadj<--read.table("unadj.assoc.logistic",     header=TRUE)
aff_unadj.add.p<--aff_unadj[aff_unadj$TEST==c("ADD"),]$P
broadqq(aff_unadj.add.p,"Some Trait Unadjusted")
aff_C1C2<--read.table("PC1--PC2.assoc.logistic", header=TRUE)
aff_C1C2.add.p<--aff_C1C2[aff_C1C2$TEST==c("ADD"),]$P
broadqq(aff_C1C2.add.p, "Some Trait Adjusted for PC1 and PC2")
dev.off()
```

Now look for SNPs with genome-wide significance using the following R connamds:

```
gws_unadj = aff_unadj[which(aff_unadj$P < 0.0000001),]
gws_unadj
gws_adjusted = aff_C1C2[which(aff_C1C2$P < 0.0000001),]
gws_adjusted
```

Note: These are the uncorrected p-values for multiple testing. The p-values which have been corrected using various multiple testing methods can be found in the .adjusted file.

A common question when you have a finding with genome-wide significance in a GWAS is "Is the SNP in a known gene?" One way to look this information up is annotate variants in batch (please look at the annotating exercise for more information). You can do this using the Ensembl Variant Predictor. Go to the website:

http://grch37.ensembl.org/Homo_sapiens/Tools/VEP (GRCh37 version)

Type the rs number(s) of the SNP(s) with genome-wide significance in "Either paste data", leave all options default and press run. In a few minutes you can view the results of your query.

Question 1: Did this study have a finding with genome-wide significance after adjusting for population substructure? Did you notice any difference in the p-values before and after adjustment for substructure? How many PC components should you include in the regression model. Please also, complete the tables below.

Table 2. SNPS with genome-wide significance unadjusted for substructure:

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P

Table 3. SNPs with genome-wide significance adjusted for components 1 and 2:

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P

Question 2: Why would you not want to include in your analysis individuals from different ethnic backgrounds even if you control for population substructure?

Question 3. Are any SNPs with genome-wide significance in known genes?

Answers and Output

Table 1

	Un--adjusted	PC1	PC1--2	PC1 --3	PC1 --4	PC1 --5	PC1 --6	PC1 --7	PC1 --8	PC1 --9	PC1--10
lambda	1.121	1.085	1.026	1.033	1.040	1.050	1.043	1.021	1.036	1.043	1.051

Answer to Question 1:

Question 1:

Did this study have a finding with genome-wide significance after adjusting for population substructure? How many PC components should you include in the regression model. Did you notice any difference in the p-values before and after adjustment for substructure?

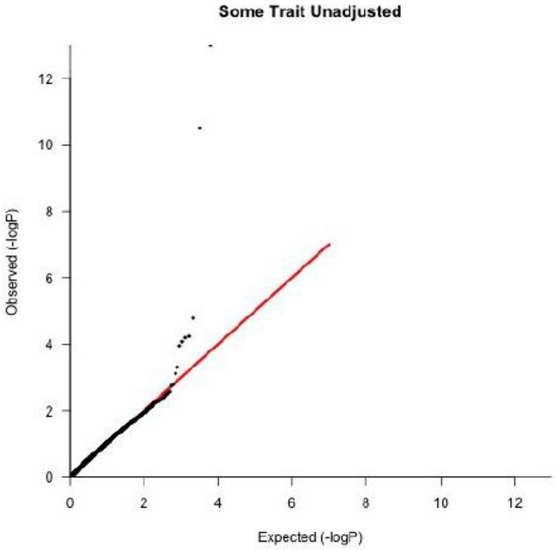
Yes, see tables below. It is best to include to two PC components in the analysis, however the lambda is still inflated. Since we are analyzing three unique populations inclusion of PCs did not adequately control for substructure. If you compare the QQ plots below you can see that for this dataset the most significant SNPs were changed minimally when we adjusted for substructure but some of the moderately significant SNPs became less significant after adjustment. However, in some situations the p-values can become smaller.

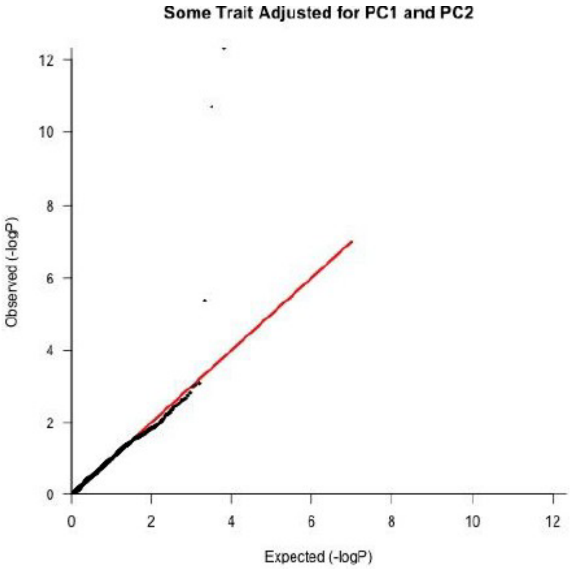
Table 2. SNPS with genome-wide significance unadjusted for substructure:

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
8	rs4571722	60326734	T	ADD	242	0.04126	--7.436	1.04E--13
4	rs10008252	179853616	G	ADD	244	0.1665	--6.639	3.16E--11

Table 3. SNPs with genome-wide significance adjusted for components 1 and 2:

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
8	rs4571722	60326734	T	ADD	242	0.04382	--7.237	4.59E--13
4	rs10008252	179853616	G	ADD	244	0.13070	--6.707	1.99E--11



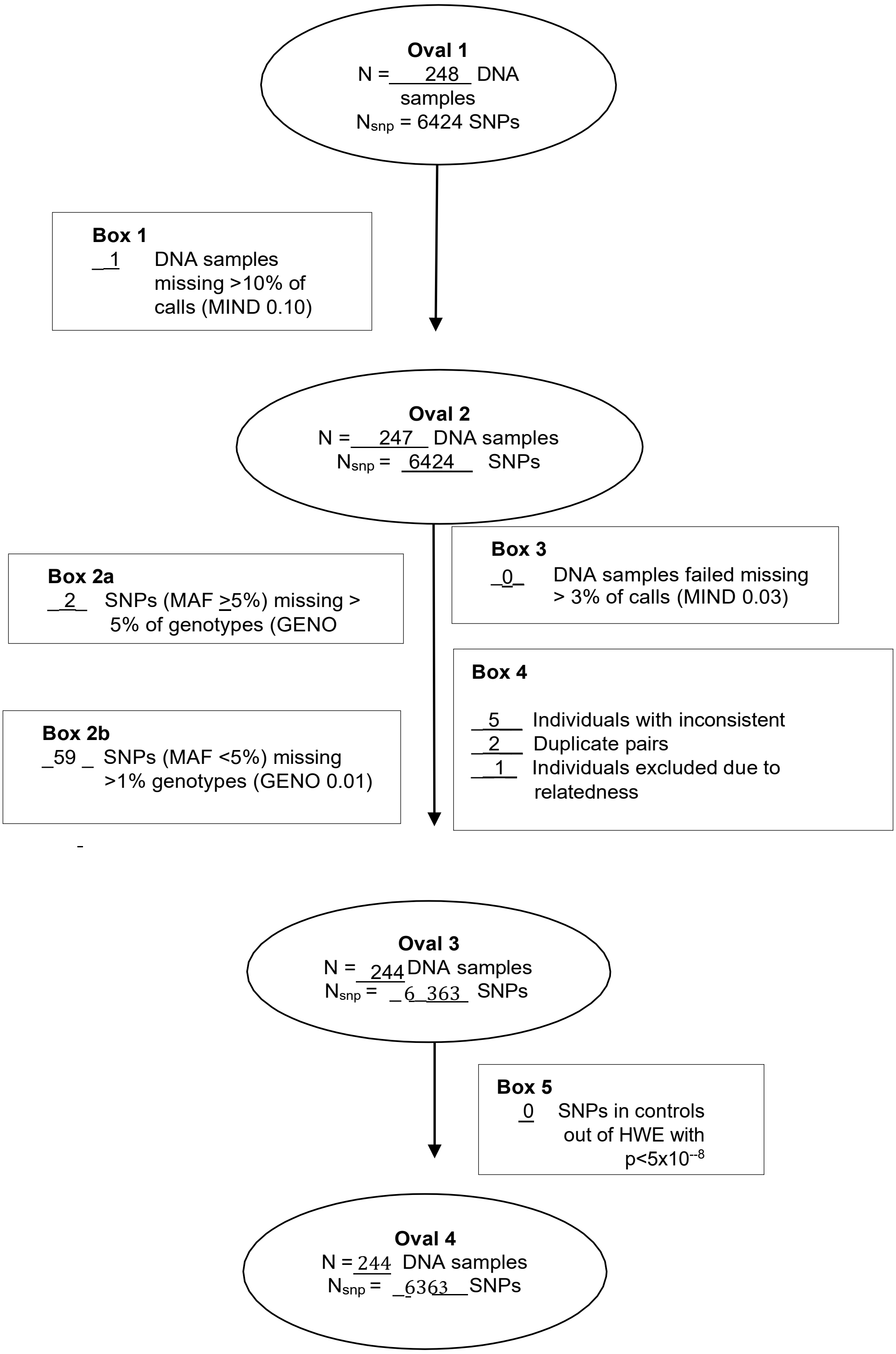


Question 2: Why would you not want to include in your analysis individuals from different ethnic backgrounds even if you control for population substructure?

Firstly, you may not be able to adequately control for population substructure. Secondly, even if within the different populations the same genes are involved, for common variants LD structure can vary between populations, e.g., the tagSNPs in the different populations can have different allele frequencies, therefore the functional variant will not be tagged equally well in all populations and power can be reduced. It is also possible that different variants are associated, but for common variants, which are very old, usually this is not the cause. If a study involves individuals of different ancestry analysis can be performed separately and the results can be combined via meta-analysis. Studying individuals of different ancestry can be highly beneficial to fine map loci.

Question 3. Are any SNPs with genome-wide significance in known genes?

No, both rs457122 and rs10008252 are intergenic/intronic.



Oval 1
N = 248 DNA samples
N_{snp} = 6424 SNPs

Box 1
1 DNA samples missing >10% of calls (MIND 0.10)

Oval 2
N = 247 DNA samples
N_{snp} = 6424 SNPs

Box 2a
2 SNPs (MAF ≥5%) missing > 5% of genotypes (GENO

Box 3
0 DNA samples failed missing > 3% of calls (MIND 0.03)

Box 2b
59 SNPs (MAF <5%) missing >1% genotypes (GENO 0.01)

Box 4
5 Individuals with inconsistent
2 Duplicate pairs
1 Individuals excluded due to relatedness

-

Oval 3
N = 244 DNA samples
N_{snp} = 6 363 SNPs

Box 5
0 SNPs in controls out of HWE with $p<5x10^{-8}$

Oval 4
N = 244 DNA samples
N_{snp} = 6363 SNPs

Computer Practical Exercise on Family-based Association using FaST-LMM, PLINK and R

Overview

Purpose

In this exercise you will be carrying out association analysis of data from a mini genome-wide association study. The data comes from families (related individuals) measured for a quantitative trait of interest. The purpose is detect which (if any) of the loci are associated with the quantitative trait.

Methodology

We will use the linear mixed model approach implemented in FaST-LMM and (for comparison) standard linear regression in PLINK.

Program documentation

PLINK documentation:

PLINK has an extensive set of docmentation including a pdf manual, a web-based tutorial and web-based documentation:

Original PLINK (1.07) (which has arguably clearer documentation):
http://zzz.bwh.harvard.edu/plink/

New PLINK (1.90) (which includes documentation on new additional features):
https://www.cog-genomics.org/plink2

R documentation:

The R website is at http://www.r-project.org/

From within R, one can obtain help on any command `xxxx` by typing `` `help(xxxx)' ``

FaST-LMM documentation:

Documentation can be downloaded together with the FaST-LMM program from

http://research.microsoft.com/en-us/downloads/aa90ccfb-b2a8-4872-ba00-32419913ca14/

Data overview

We will be using family data consisting of 498 individuals typed at 134,946 SNPs. All individuals have measurements of a quantitative trait of interest. You can assume that appropriate quality control (QC) checks on SNPs and individuals have been carried out prior to the current analysis i.e. the data set is already QC-ed.

Appropriate data

Appropriate data for this exercise is genome-wide genotype data for related and/or apparently unrelated individuals. Genome-wide data is required in order to estimate relationships between people and allow for relatedness in the analysis. The individuals should be phenotyped for either a dichotomous trait or a quantitative trait of interest.

Instructions

Data files

The data is in PLINK binary-file format. Check you have the required files by typing:

```
ls -l
```

You should find 3 PLINK binary-format files in your directory: `quantfamdata.bed`, `quantfamdata.bim` and `quantfamdata.fam`. The file `quantfamdata.bed` is the binary genotype file which will not be human readable. The file `quantfamdata.bim` is a map file. You can take a look at this (e.g. by typing `more quantfamdata.bim`). The file `quantfamdata.fam` gives the pedigree structure in a format that is compatible with the binary genotype file. You can take a look at this (e.g. by typing `more quantfamdata.fam`). Note this file is the same as the first six columns of a standard pedigree file, with the last column giving each individual's quantitative trait value.

Step-by-step instructions

1. Analysis in PLINK

To start with, we will use PLINK to perform a test equivalent to linear regression analysis, without worrying about the relatedness between individuals:

```
plink --bfile quantfamdata --assoc --out plinkresults
```

A copy of the screen output is saved in the file `plinkresults.log`. The association results are output to a file `plinkresults.qassoc`. Take a look at this file. Each line corresponds to the results for a particular SNP. Each line contains the following columns:

```
CHR      Chromosome number
SNP      SNP identifier
BP       Physical position (base-pair)
NMISS    Number of non-missing genotypes
BETA     Regression coefficient
SE       Standard error
R2       Regression r-squared
T        Wald test (based on t-distribtion)
P        Wald test asymptotic p-value
```

The most useful columns are T (the test statistic) and its p value (P).

To visualise these results properly we will use R. Open up a new terminal window, move to the directory where you performed this analysis, and start R (by typing `R`).

Now (within R) read in the data by typing:

```
res1<-read.table("plinkresults.qassoc", header=T)
```

This reads the results into a dataframe named "res1". To see the top few lines of this dataframe, type:

```
head(res1)
```

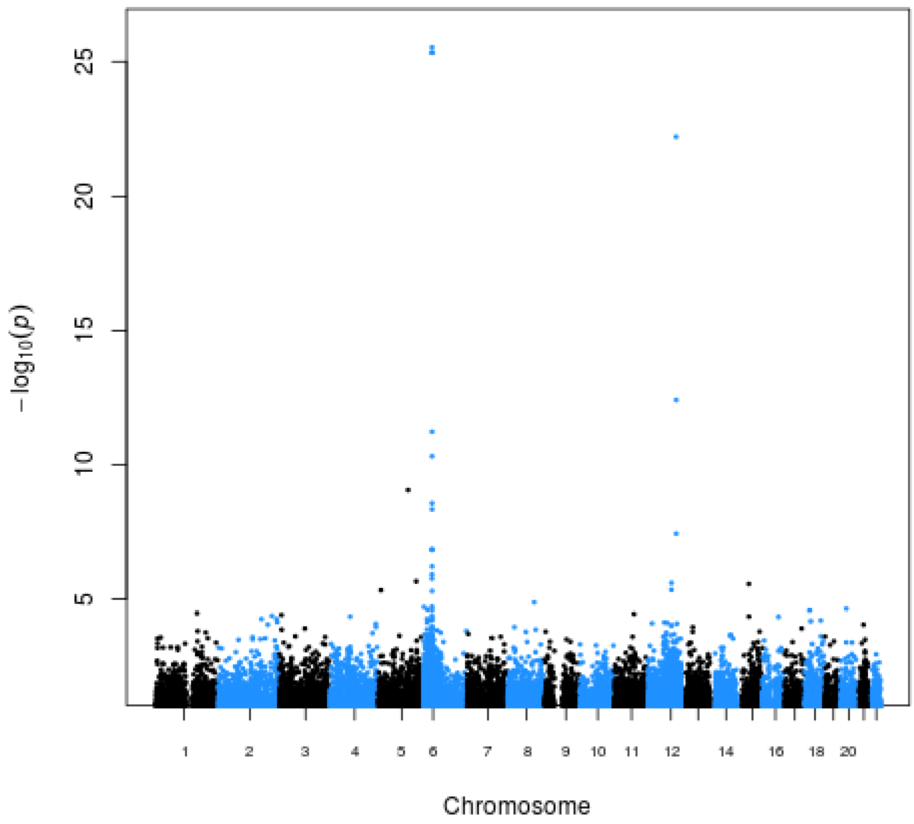
The data frame has 134,946 lines, one for each SNP. It would be very laborious to go through and look at each line by eye. Instead we will plot the results for all chromosomes, colouring each chromosome differently. To do this we need to first read in from an external file some special functions for creating such ``Manhattan'' plots:

```
source("qqmanHJCupdated.R")
```

Then we use the following command to actually make the plot, and save it in the file "mh1.png":

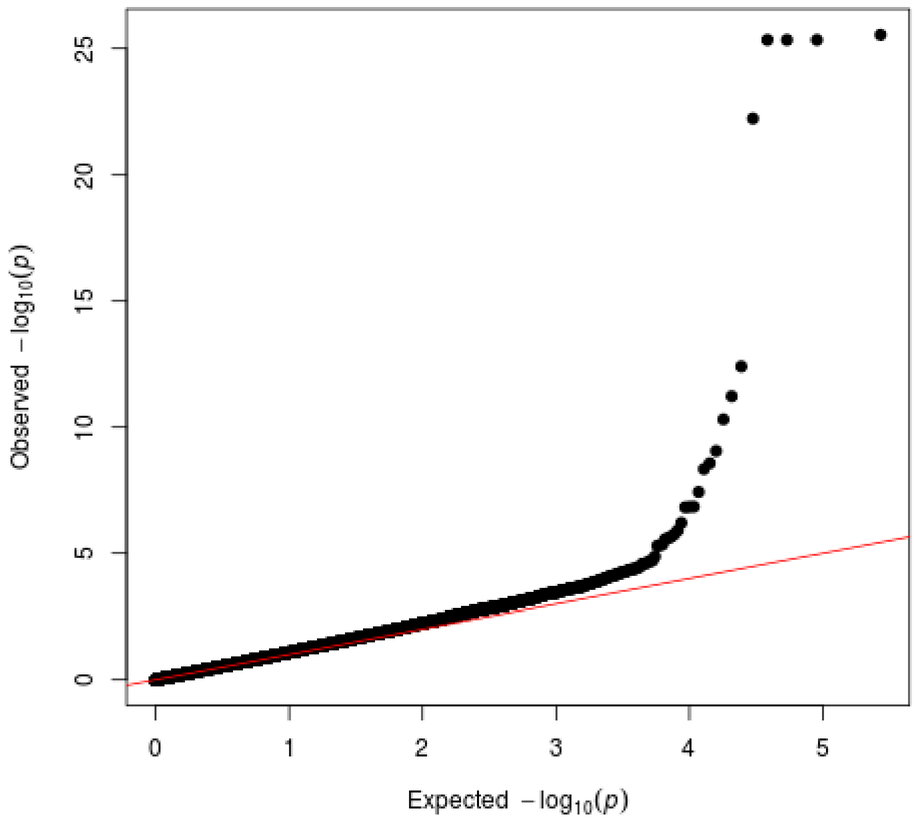
```
png("mh1.png")
manhattan(res1, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
```

Be warned, this may take some time to plot.



Visually it looks like there may be significant results on chromosomes 6 and 12, and possibly on chromosome 5 as well. One way to assess the significance of the results, in light of the large number of tests performed, is to use a Q-Q plot. To plot a Q-Q plot for these P values, and save it in the file "qq1.png", type:

```
png("qq1.png")
qq(res1$P)
dev.off()
```



What one would hope to see is most of the values lying along the straight line with gradient 1, indicating that most results are consistent with the null hypothesis of no association. However, one would also hope to see a few high values at the top that depart from the straight line, which are hopefully true associations.

Our results seem fairly consistent with this expectation, but there may be a little bit of inflation (i.e. a slope slightly bigger than 1) due to relatedness between individuals. To calculate the genomic control inflation factor, we first convert the P values to chi-squared test statistics on 1df, and then use the formula from Devlin and Roeder (1999):

```
chi<-(qchisq(1-res1$P,1))
lambda=median(chi)/0.456
lambda
```

You should find a slightly inflated value (lambda=1.10)

2. Analysis in FaST-LMM

Now we will try re-running the analysis using FaST-LMM, which estimates and accounts for the relatedness between individuals. Go back to the window where you ran PLINK and run FaST-LMM as follows:

```
fastlmmc -bfile quantfamdata -pheno quantfamdata.fam -mpheno 4 -bfileSim
quantfamdata -ML -out FLMMresults
```

Here we use the `-bfile quantfamdata` command to tell the program the name (stem) of the files with the input genotype data containing the SNPs to be tested for association, and the `-bfileSim quantfamdata` command to tell the program the name of the files containing the SNPs to be used for estimating relatedness. Here we just use the same files both times, but FaST-LMM would allow us to use different files for these two operations if we prefer.

The command `-pheno quantfamdata.fam -mpheno 4` tells FaST-LMM to read the phenotype data in from the file `quantfamdata.fam` , using the 4th phenotype column (not including the two first columns which give the family and person IDs). The `-ML` command tells FaST-LMM to use maximium likelihood estimation (in case you prefer this as opposed to the default restricted maximium likelihood (REML)). The command `-out FLMMresults` tells FaST-LMM the name to use for the output file.

Take a look at the results file. FaST-LMM automatically orders the results by significance.

Now go back to your R window and read the results into R:

```
res2<-read.table("FLMMresults", header=T)
```

Check the column names by typing:

```
head(res2)
```

The P value is in a column called ``Pvalue". Remember FaST-LMM has automatically ordered the results by significance, so these top few rows will show the most significant results.

First let us check the genomic control inflation factor. We convert the P values to chi-squared test statistics on 1df, and then use the formula from Devlin and Roeder (1999):

```
chi<-(qchisq(1-res2$Pvalue,1))
lambda=median(chi)/0.456
lambda
```

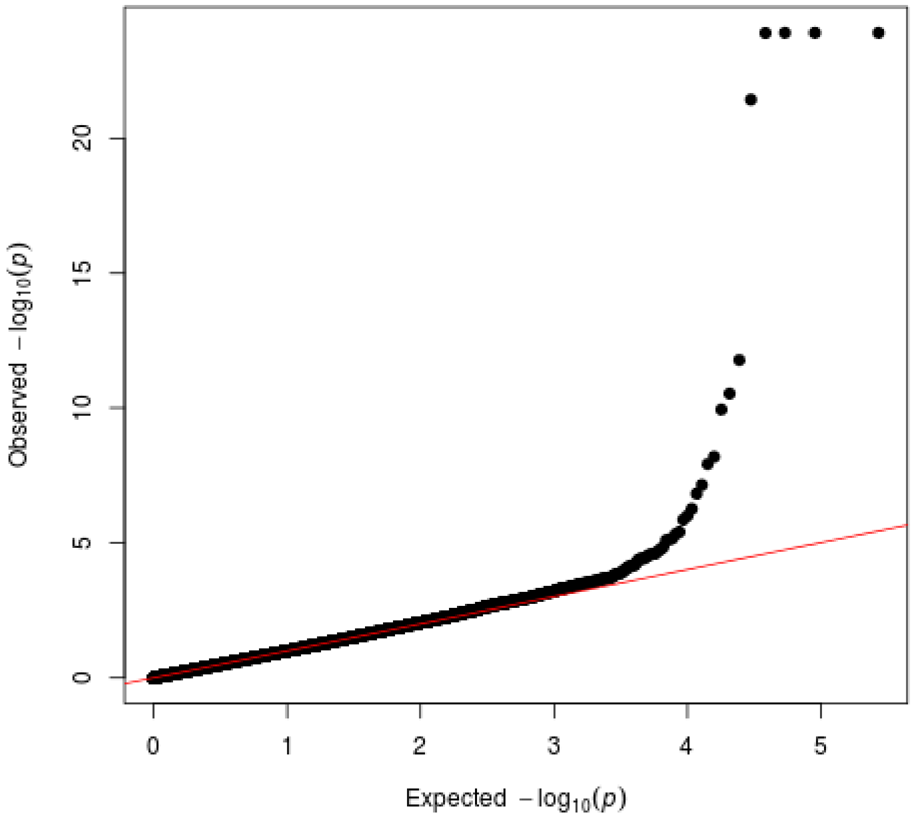
You should find a less inflated value (lambda=0.99) than we found previously with PLINK.

To plot Manhattan and Q-Q plots you can use similar commands to before, but the columns need to be named appropriately. The easiest thing is to make a new smaller dataframe containing the required data:

```
new<-data.frame(res2$SNP, res2$Chromosome, res2$Position, res2$Pvalue)
names(new)<-c("SNP", "CHR", "BP", "P")
head(new)
```

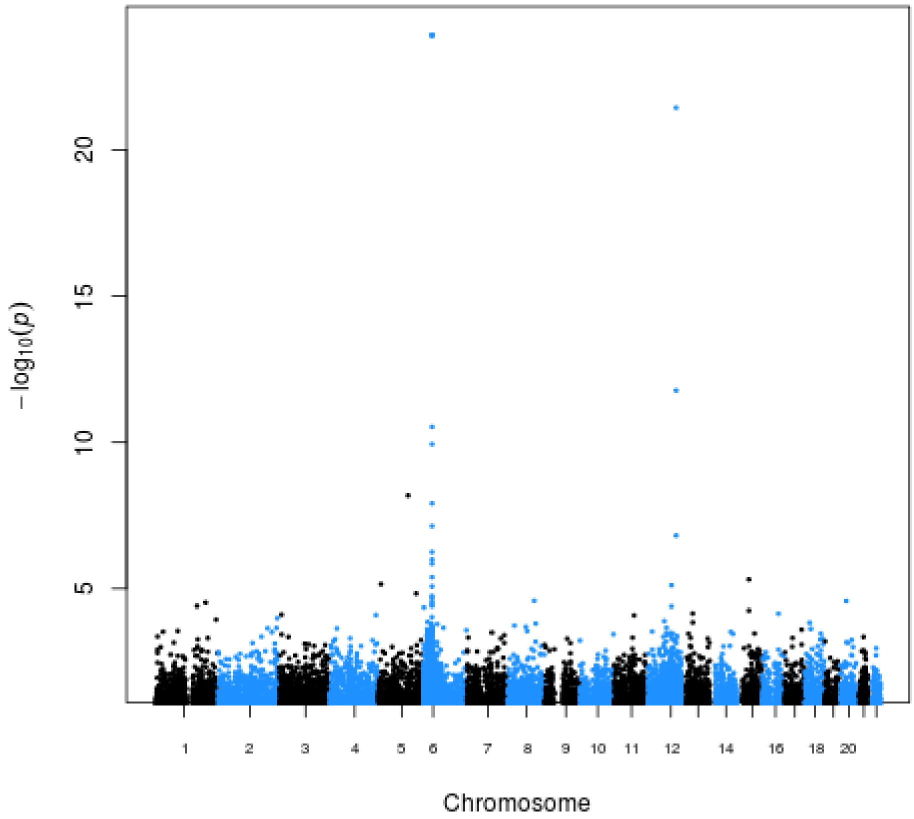
Now you can plot the Q-Q plot:

```
png("qq2.png")
qq(new$P)
dev.off()
```



And the Manhattan plot:

```
png("mh2.png")
manhattan(new, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
```



The significant effects on chromosomes 6 and 12 are still easily visible. In fact, this is simulated data, and these signals do correspond correctly to the positions of the underlying causal variants.

Answers

How to interpret the output

Interpretation of the output is described in the step-by-step instructions. In general, the output will consist of a likelihood-ratio or chi-squared test for whatever you are test you are performing, and regression coefficients or odds ratio estimates for the predictor variables in the current model. Please ask if you need help in understanding the output for any specific test.

Comments

Advantages/disadvantages

PLINK is useful for data management and analysis of genome-wide association data. FaST-LMM is more appropriate for analysis of related individuals, or for correcting for population stratification in apparently unrelated individuals.

Other packages

Other packages that can implement a similar analysis to FaST-LMM include EMMAX, GEMMA, MMM, GenABEL, Mendel.

References

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies Nat Methods 8(10):833-835.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81:559-575.

Exercises prepared by: Heather Cordell
Checked by:
Programs used: PLINK, R, FaST-LMM
Last updated: 01/17/2020 12:33:40

Computer Practical Exercise using GCTA (with R)

Overview

Purpose

This exercise repeats the linear mixed model analysis from the previous exercise using the program GCTA instead of FaST-LMM. In addition, we use GCTA to estimate the heritability accounted for by all genotyped SNPs, and by various subsets of SNPs.

Methodology

We will use the linear mixed model approach implemented in GCTA.

Program documentation

GCTA documentation:

Documentation can obtained together with the GCTA program from:

http://cnsgenomics.com/software/gcta/

Data overview

As a reminder, we are using family data consisting of 498 individuals typed at 134,946 SNPs. All individuals have measurements of a quantitative trait of interest.

Appropriate data

Appropriate data for this exercise is genome-wide genotype data for individuals who are phenotyped for either a dichotomous trait or a quantitative trait of interest. GCTA is really designed for the analysis of apparently unrelated individuals, but in this case we will apply it to a set of related individuals, in order to compare the results with those we obtained previously for these individuals.

Instructions

Data files

We will use the same PLINK binary-file format files quantfamdata.bed, quantfamdata.bim and quantfamdata.fam used previously. We will also use R to create an additional phenotype file required by GCTA.

Step-by-step instructions

1. Create phenotype file in R

To start with, we will use R to create the phenotype file required by GCTA. Start R (by typing R) and create a new phenotype file from the .fam file by typing the following commands:

```
fam<-read.table("quantfamdata.fam", header=F)
pheno=data.frame(fam[,1:2],fam[,6])
write.table(pheno,file="phenos.txt",col.names=F,row.names=F,quote=F)
```

Take a look at the file phenos.txt that you just created, to check you understand it.

2. GCTA Analysis

To use GCTA to perform association analysis while allowing for relatedness between individuals, type:

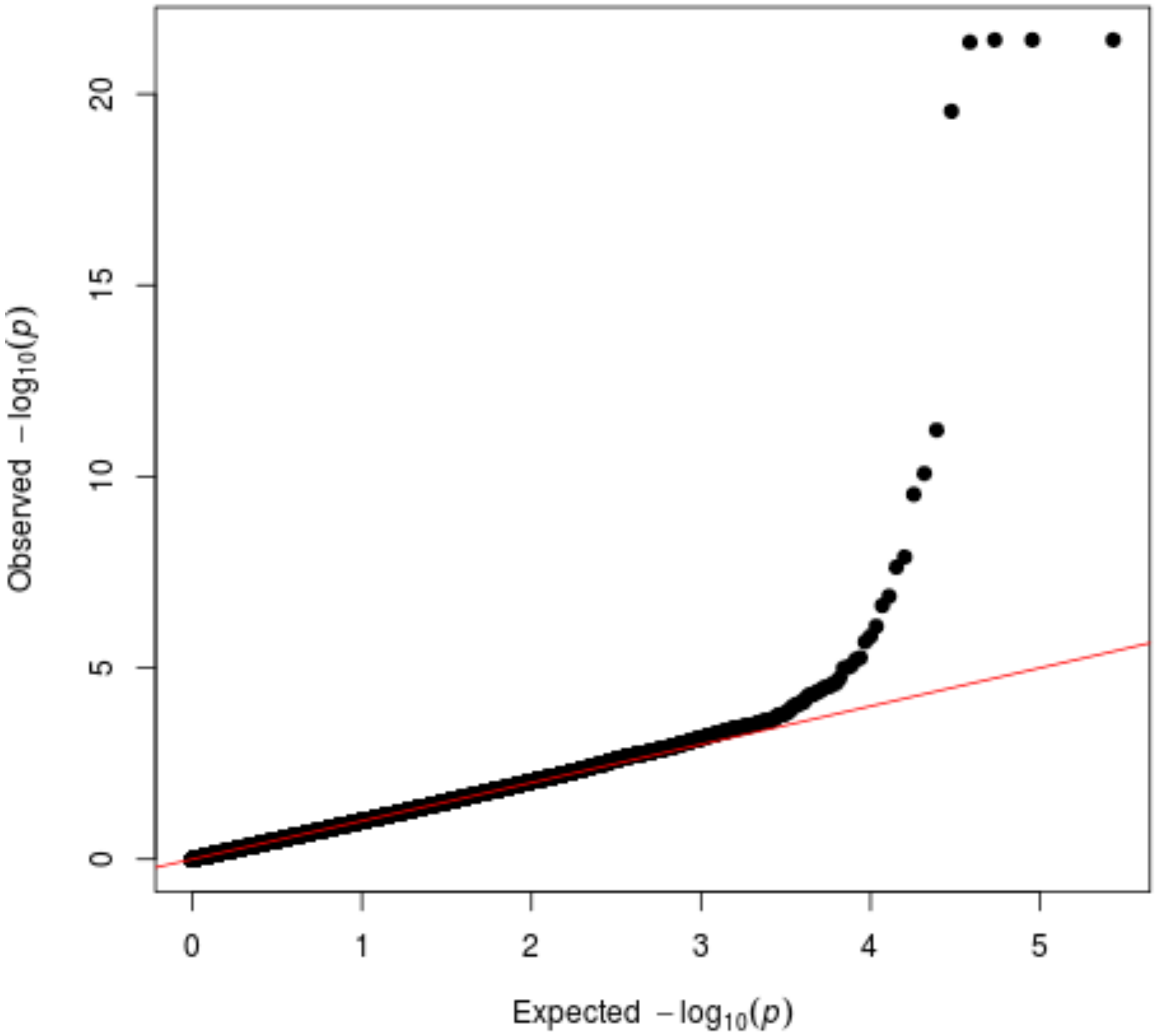
```
gcta64 --mlma --bfile quantfamdata --pheno phenos.txt --out GCTAresults
```

Here we use the --mlma option to tell GCTA to perform association analysis, we use the --bfile and --pheno options to tell GCTA which files to read in the genotype and phenotype data from, and we use GCTAresults as the stem name for the output files.

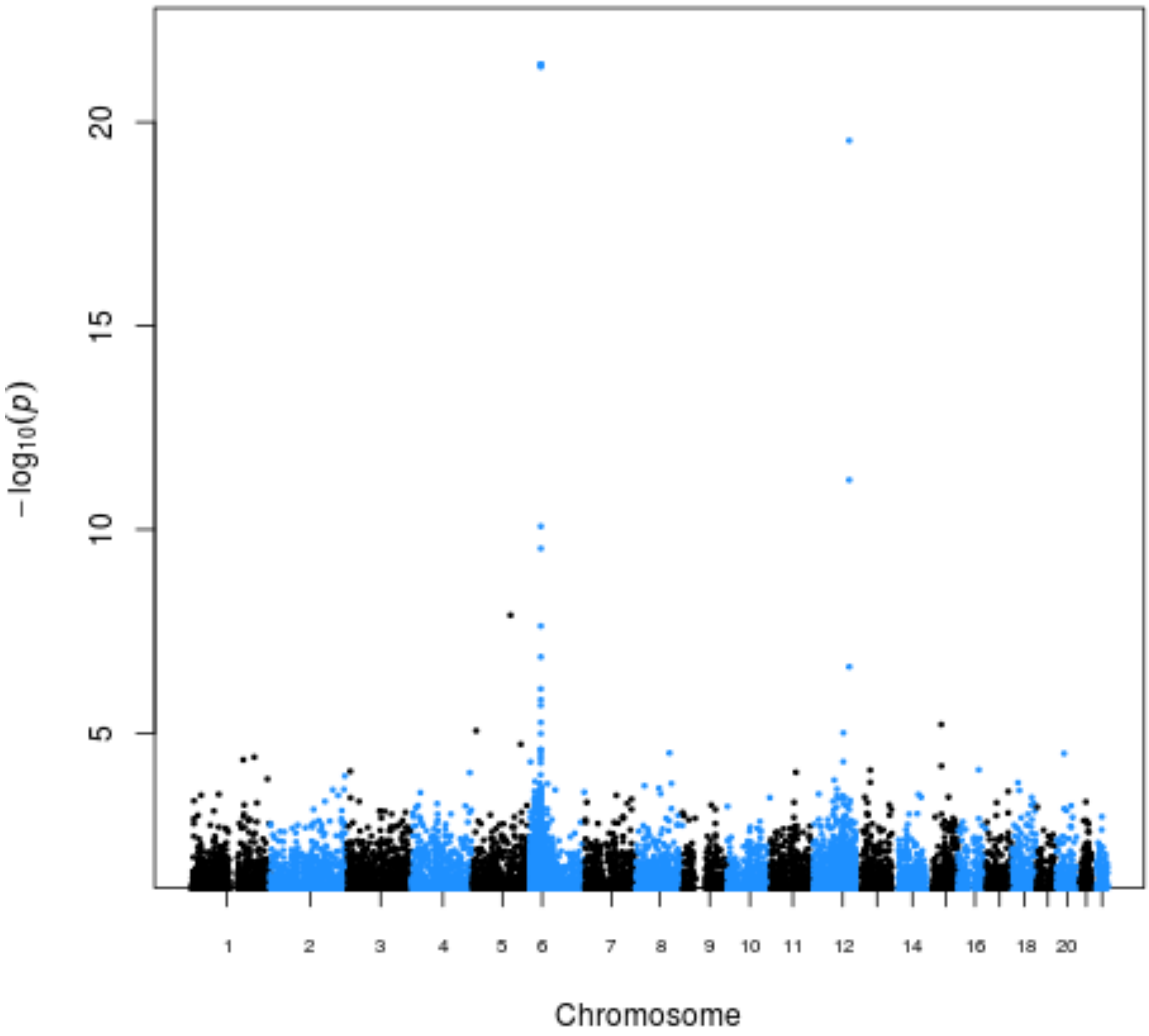
To calculate the genomic control inflation factor, and to produce QQ and Manhattan plots from the above analysis, you can use the following sequence of commands within R. (Make sure that you understand the commands - if not please ask an instructor).

```
source("qqmanHJCupdated.R")

res3<-read.table("GCTAresults.mlma", header=T)
head(res3)

chi<-(qchisq(1-res3$p,1))
lambda=median(chi)/0.456
lambda
```

```
new3<-data.frame(res3$SNP, res3$Chr, res3$bp, res3$p)
names(new3)<-c("SNP", "CHR", "BP", "P")
head(new3)

png("qq3.png")
qq(new3$P)
dev.off()
```



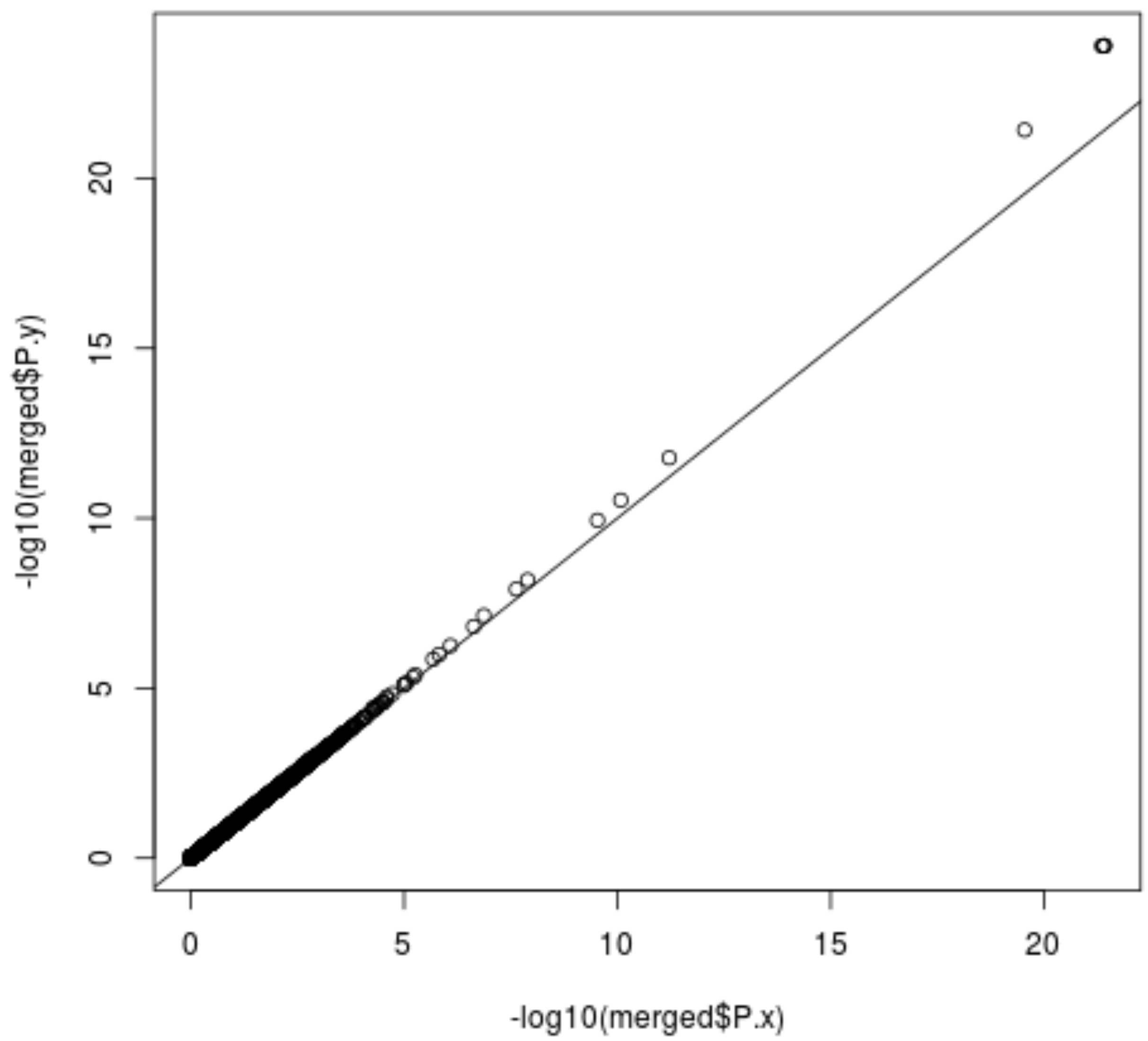
```
png("mh3.png")
manhattan(new3, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
```



You should find that the genomic control factor is close to 1.0, and the QQ and Manhattan plots are similar to those you obtained from FaST-LMM.

To compare the results (`res3`) with our previous FaST-LMM results (`res2`), use the following sequence of commands within R:

```
res2<-read.table("FLMMresults", header=T)
new2<-data.frame(res2$SNP, res2$Chromosome, res2$Position, res2$Pvalue)
names(new2)<-c("SNP", "CHR", "BP", "P")
merged=merge(new3,new2, by="SNP", sort=F)

head(res2)
head(new2)
head(new3)
head(merged)

png("compareGCTAFLMM.png")
plot(-log10(merged$P.x),-log10(merged$P.y))
abline(0,1)
dev.off()
```



You should find that the GCTA results (on the x axis) are very similar to the FaST-LMM results (on the y axis), although the -log10 P-values from FaST-LMM are consistently just a little bit higher than those from GCTA.

To use GCTA to estimate the heritability accounted for by all autosomal genome-wide SNPs, you need to first estimate the GRM, and then use the GRM to estimate the (SNP) heritability. This can be achieved using the following commands:

```
gcta64 --bfile quantfamdata --autosome --make-grm-bin --out GCTAgrm
gcta64 --reml --grm-bin GCTAgrm --pheno phenos.txt --out GCTAherit
```

The screen output estimates the SNP heritability V(G)/Vp to be 0.480590 or around 48%.

To estimate the heritabilty accounted for by SNPs on chromosomes 1, 2, 6 and 12 (for example), use the following commands:

```
gcta64 --bfile quantfamdata --chr 1 --make-grm-bin --out GCTAgrmchr1
gcta64 --reml --grm-bin GCTAgrmchr1 --pheno phenos.txt \
--out GCTAheritchr1

gcta64 --bfile quantfamdata --chr 2 --make-grm-bin --out GCTAgrmchr2
gcta64 --reml --grm-bin GCTAgrmchr2 --pheno phenos.txt \
--out GCTAheritchr2

gcta64 --bfile quantfamdata --chr 6 --make-grm-bin --out GCTAgrmchr6
gcta64 --reml --grm-bin GCTAgrmchr6 --pheno phenos.txt \
--out GCTAheritchr6

gcta64 --bfile quantfamdata --chr 12 --make-grm-bin --out GCTAgrmchr12
gcta64 --reml --grm-bin GCTAgrmchr12 --pheno phenos.txt \
--out GCTAheritchr12
```

You should find that the SNP heritabilities on chromosomes 1 and 2 do not look particularly significant (given the estimated standard errors), but the SNP heritabilities on chromosomes 6 and 12 are significant (as might be expected from the strong effects seen on these chromosomes).

The sum of the SNP heritabilities on these 4 chromosomes (0.206479+0.111512+0.368184+0.286570) adds up to more then the overall SNP heritability of 0.480589. This is due to the phenomenon that, in the presence of population substructure or close relatedness, chromosome-specific heritability estimates can be confounded by shared non-genetic effects (for examples shared environment) or corrrelations between SNPs on different chromosomes, leading to an over-estimate of the chromosome-specific heritability.

To correctly partition the overall heritability between the 22 autosomes, we need to first estimate chromosome-specific GRMs and then include them all simultaneously in the model.

We first calculate the GRMs for all additional chromosomes:

```
gcta64 --bfile quantfamdata --chr 3 --make-grm-bin --out GCTAgrmchr3
gcta64 --bfile quantfamdata --chr 4 --make-grm-bin --out GCTAgrmchr4
gcta64 --bfile quantfamdata --chr 5 --make-grm-bin --out GCTAgrmchr5
gcta64 --bfile quantfamdata --chr 7 --make-grm-bin --out GCTAgrmchr7
gcta64 --bfile quantfamdata --chr 8 --make-grm-bin --out GCTAgrmchr8
gcta64 --bfile quantfamdata --chr 9 --make-grm-bin --out GCTAgrmchr9
gcta64 --bfile quantfamdata --chr 10 --make-grm-bin --out GCTAgrmchr10
gcta64 --bfile quantfamdata --chr 11 --make-grm-bin --out GCTAgrmchr11
gcta64 --bfile quantfamdata --chr 13 --make-grm-bin --out GCTAgrmchr13
gcta64 --bfile quantfamdata --chr 14 --make-grm-bin --out GCTAgrmchr14
gcta64 --bfile quantfamdata --chr 15 --make-grm-bin --out GCTAgrmchr15
gcta64 --bfile quantfamdata --chr 16 --make-grm-bin --out GCTAgrmchr16
gcta64 --bfile quantfamdata --chr 17 --make-grm-bin --out GCTAgrmchr17
gcta64 --bfile quantfamdata --chr 18 --make-grm-bin --out GCTAgrmchr18
gcta64 --bfile quantfamdata --chr 19 --make-grm-bin --out GCTAgrmchr19
gcta64 --bfile quantfamdata --chr 20 --make-grm-bin --out GCTAgrmchr20
gcta64 --bfile quantfamdata --chr 21 --make-grm-bin --out GCTAgrmchr21
gcta64 --bfile quantfamdata --chr 22 --make-grm-bin --out GCTAgrmchr22
```

We then run the analysis:

```
gcta64 --reml --mgrm-bin multipleGRMs.txt --pheno phenos.txt \
--out GCTAherit22GRMs
```

Note this command makes use of a file `multipleGRMs.txt` which we created for you in advance, listing the stem names of the individual GRM files. Unfortunately, in this example the analysis fails to converge, probably because this type of analysis ideally requires a larger number of less closely related individuals.

To instead partition the heritability among two sets of SNPs, chromosome 6 and all other autosomes, we first join together the GRMs for all non-chromosome 6 chromosomes:

```
gcta64 --mgrm-bin multipleGRMsnot6.txt --make-grm --out GCTAgrmnot6
```

Note this command makes use of another file `multipleGRMsnot6.txt` which we created for you in advance, listing the stem names of the individual GRM files (excluding the one for chromosome 6).

We will run the analysis making use of another file `multipleGRMs6andnot6.txt` which we created for you in advance. Take a look at this file and check you understand it.

To run the analysis type:

```
gcta64 --reml --mgrm-bin multipleGRMs6andnot6.txt --pheno phenos.txt \
--out GCTAherit6andnot6
```

The results suggest that a total SNP heritability of 0.469171 can be partitioned as 0.294445 accounted for by chromosome 6, and 0.174726 accounted for by the other autosomes.

3. GCTA fastGWA Analysis

GCTA has an alternative method for performing mixed linear model (MLM)-based GWAS analysis that is partcularly designed for large biobank-scale datasets such as the UK Biobank. Here we will apply it to the (much smaller scale) dataset that we have already analysed.

First we have to make a sparse genetic relationship matrix (GRM) from the full-dense GRM e.g. using a cutoff value of 0.05 (so entries less than 0.05 are set to 0):

```
gcta64 --grm GCTAgrm --make-bK-sparse 0.05 --out GCTAsparsegrm
```

This creates a file GCTAsparsegrm.grm.sp containing the pairs of individuals whose entries in the GRM are greater than 0.05.

For real biobank-scale data, creating the full-dense GRM in the first place can be computationally challenging, and it can be advantageous to partition the GRM into m parts (by row), and compute the parts separately (before joining them back together). To compute i-th part in the current run, we use the syntax `--make-grm-part m i` . For example, use the following commands to re-calculate the full-dense GRM by partitioning the calculation into 3 parts (while also using 5 threads):

```
gcta64 --bfile quantfamdata --make-grm-part 3 1 --thread-num 5 \
--out test
gcta64 --bfile quantfamdata --make-grm-part 3 2 --thread-num 5 \
--out test
gcta64 --bfile quantfamdata --make-grm-part 3 3 --thread-num 5 \
--out test
```

Merge all the parts together:

```
cat test.part_3_*.grm.id > test.grm.id
cat test.part_3_*.grm.bin > test.grm.bin
cat test.part_3_*.grm.N.bin > test.grm.N.bin
```

Now we can create a sparse GRM from this re-calculated version of the full-dense GRM:

```
gcta64 --grm test --make-bK-sparse 0.05 --out newsparsegrm
```

Check that the top few lines of newsparsegrm.grm.sp seem to match GCTAsparsegrm.grm.sp:

```
head *.sp
```

To perform the association analysis, we would normally use the sparse GRM to model close relationships as random effects, while additionally including principal components as fixed effects. So let us first use GCTA to calculate the first 5 principal components:

```
gcta64 --grm-bin GCTAgrm --pca 5 --out pcs
```

To use original GCTAsparsegrm.grm.sp in a fastGWA analysis, type:

```
gcta64 --bfile quantfamdata --grm-sparse GCTAsparsegrm \
--fastGWA-mlm --pheno phenos.txt --qcovar pcs.eigenvec \
--out sparse_assoc
```

To use newsparsegrm.grm.sp in a fastGWA analysis, type:

```
gcta64 --bfile quantfamdata --grm-sparse newsparsegrm \
--fastGWA-mlm --pheno phenos.txt --qcovar pcs.eigenvec \
--out newsparse_assoc
```

Once we have included the principal components, it seems that the estimate of Vg is not statistically significant, and so fastGWA has automatically moved to using linear regression rather than a linear mixed model. This is not really what we wanted! Let us not include principal components - we would then expect the estimate of Vg to be significant, and so in this way we force fastGWA to use a linear mixed model:

```
gcta64 --bfile quantfamdata --grm-sparse GCTAsparsegrm \
--fastGWA-mlm --pheno phenos.txt --out LMMsparse_assoc

gcta64 --bfile quantfamdata --grm-sparse newsparsegrm \
--fastGWA-mlm --pheno phenos.txt --out LMMnewsparse_assoc
```

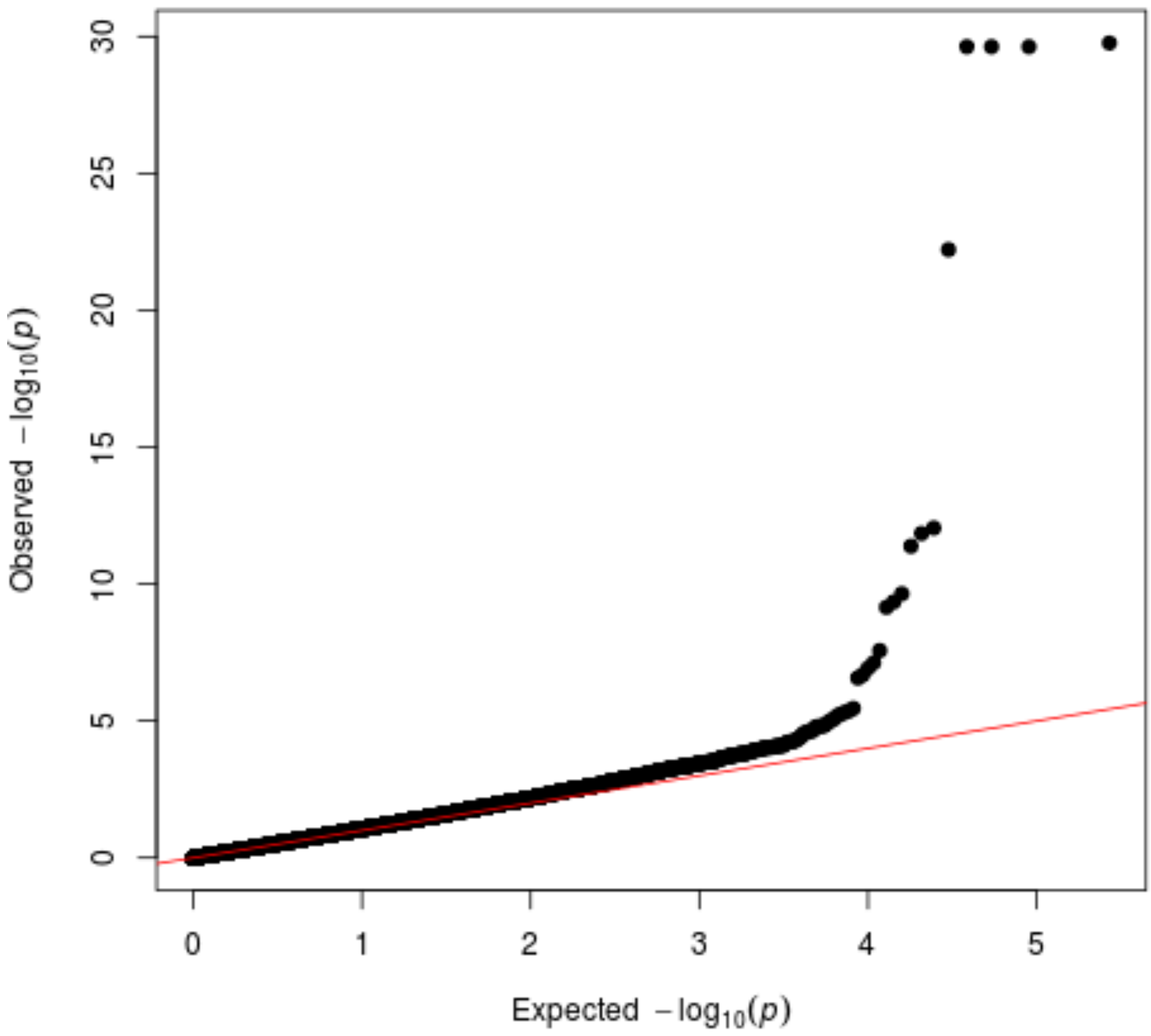
Let us now use R to check the QQ plots, Manhattan plots and Genomic control factors for these 4 sets of results. Start up R and use the following commands:

```
res4<-read.table("sparse_assoc.fastGWA", header=T)
res5<-read.table("newsparse_assoc.fastGWA", header=T)
res6<-read.table("LMMsparse_assoc.fastGWA", header=T)
res7<-read.table("LMMnewsparse_assoc.fastGWA", header=T)

head(res4)
head(res5)
head(res6)
head(res7)

chi<-(qchisq(1-res4$P,1))
lambda=median(chi)/0.456
lambda

chi<-(qchisq(1-res5$P,1))
lambda=median(chi)/0.456
lambda

chi<-(qchisq(1-res6$P,1))
lambda=median(chi)/0.456
lambda

chi<-(qchisq(1-res7$P,1))
lambda=median(chi)/0.456
lambda
```

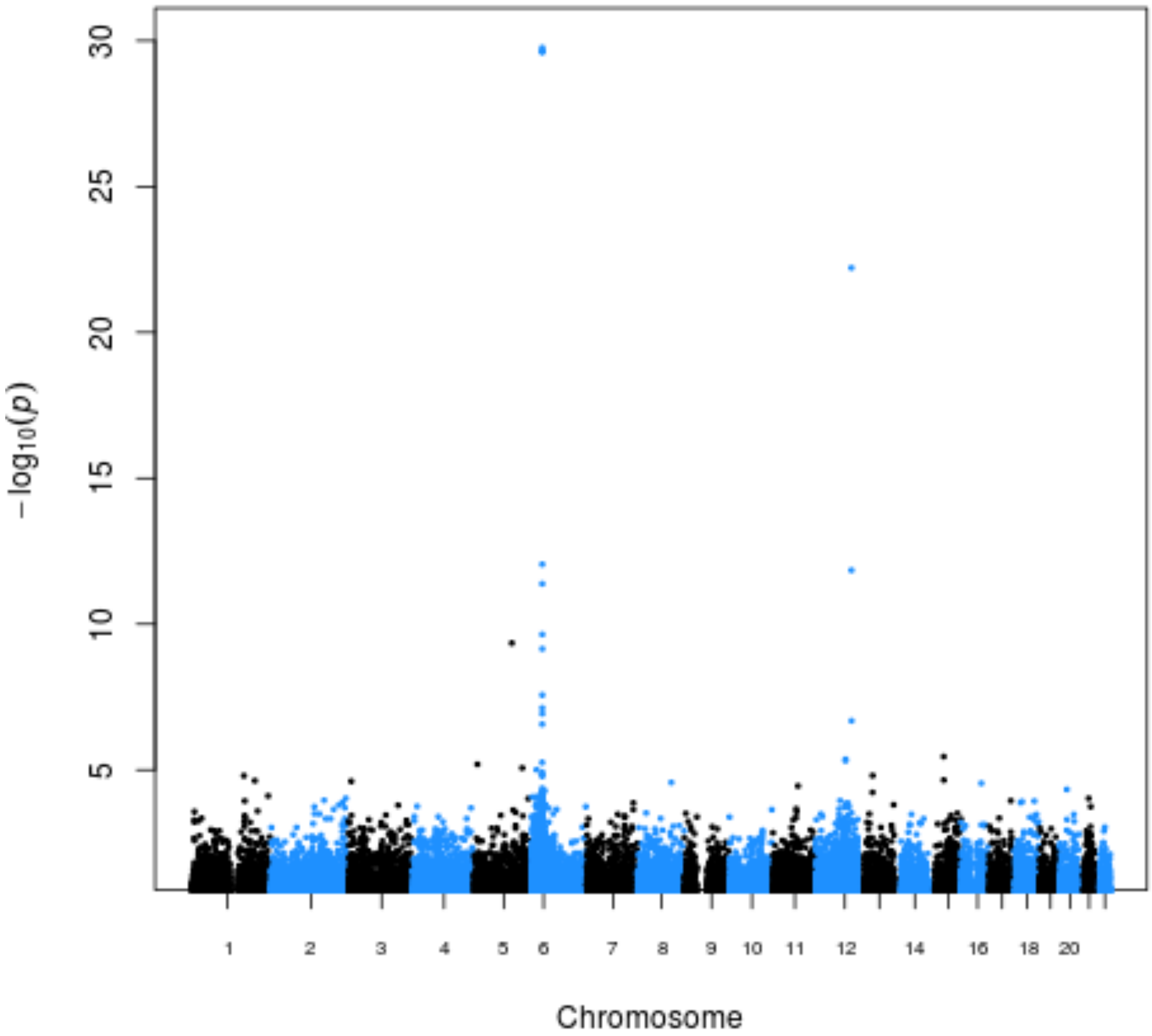
You should see that the linear mixed model (with a sparse GRM) does a better job at controlling for relatedness (giving lambda closer to 1.0) than just including 5 principal components.

Now continue in R to make the plots for these 4 analyses (which will be very similar to those you made previously):

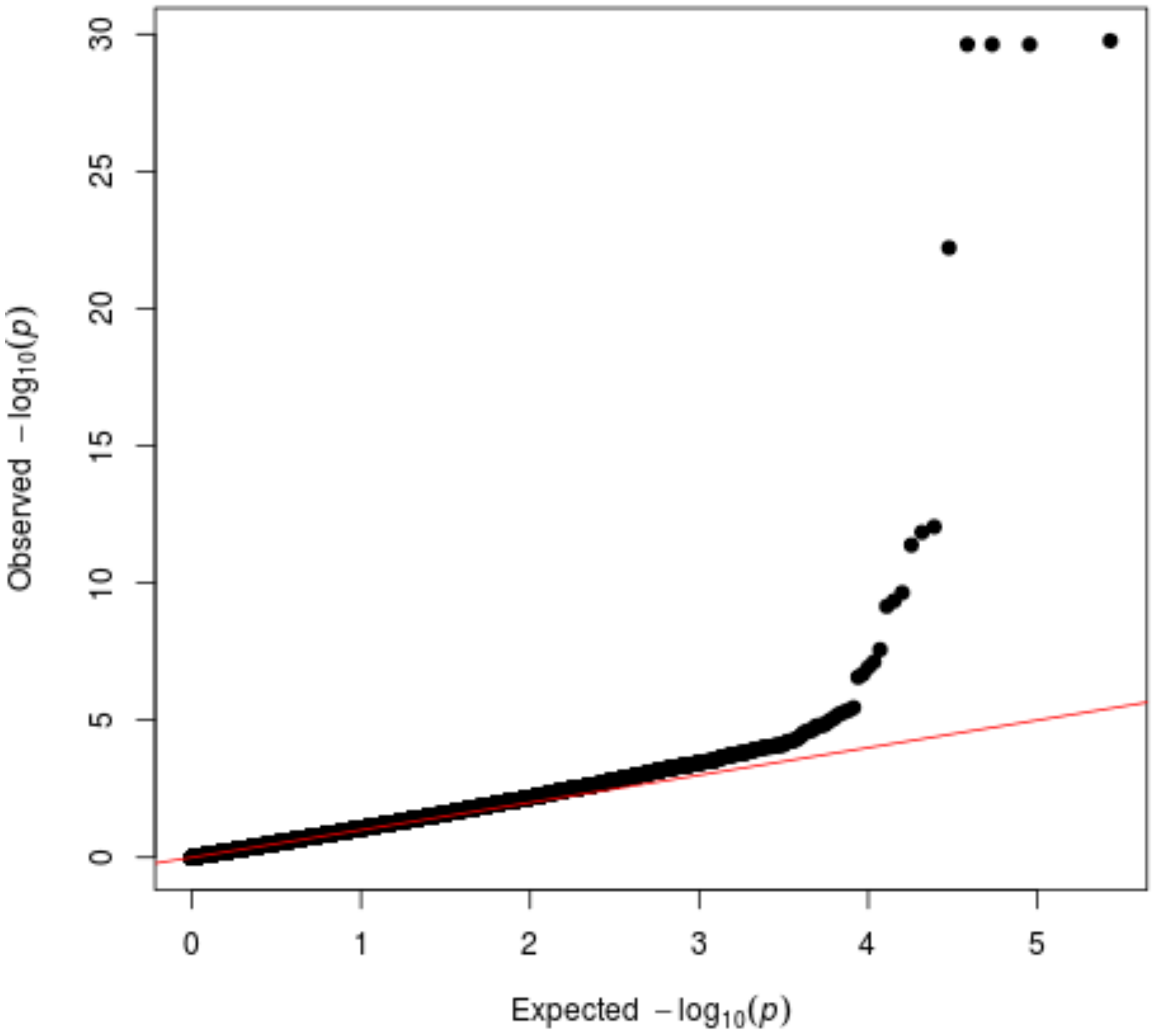
```
source("qqmanHJCupdated.R")

new4<-data.frame(res4$SNP, res4$CHR, res4$POS, res4$P)
names(new4)<-c("SNP", "CHR", "BP", "P")

png("qq4.png")
qq(new4$P)
dev.off()
```



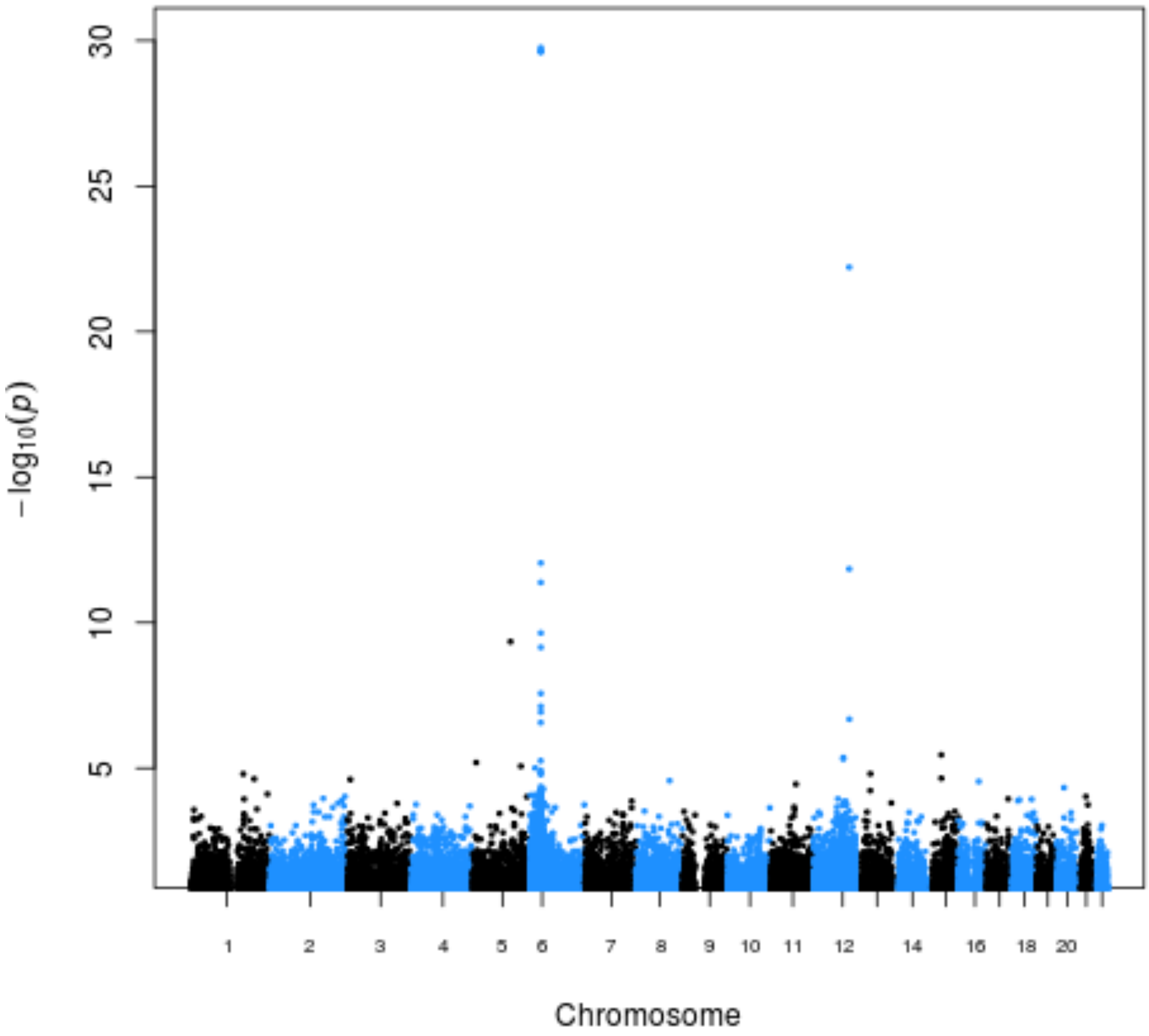
```
png("mh4.png")
manhattan(new4, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
```



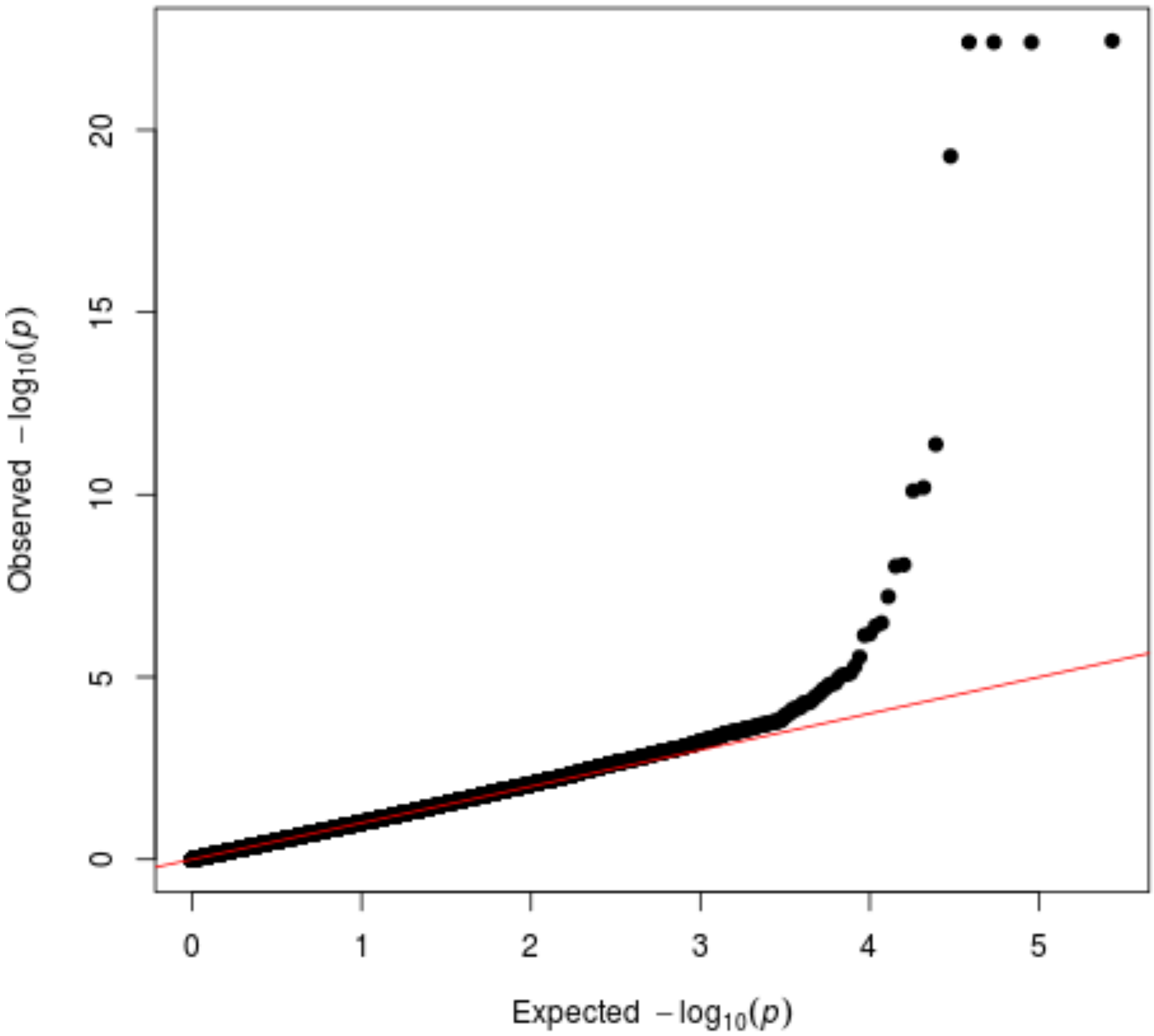
```
new5<-data.frame(res5$SNP, res5$CHR, res5$POS, res5$P)
names(new5)<-c("SNP", "CHR", "BP", "P")

png("qq5.png")
qq(new5$P)
dev.off()
```



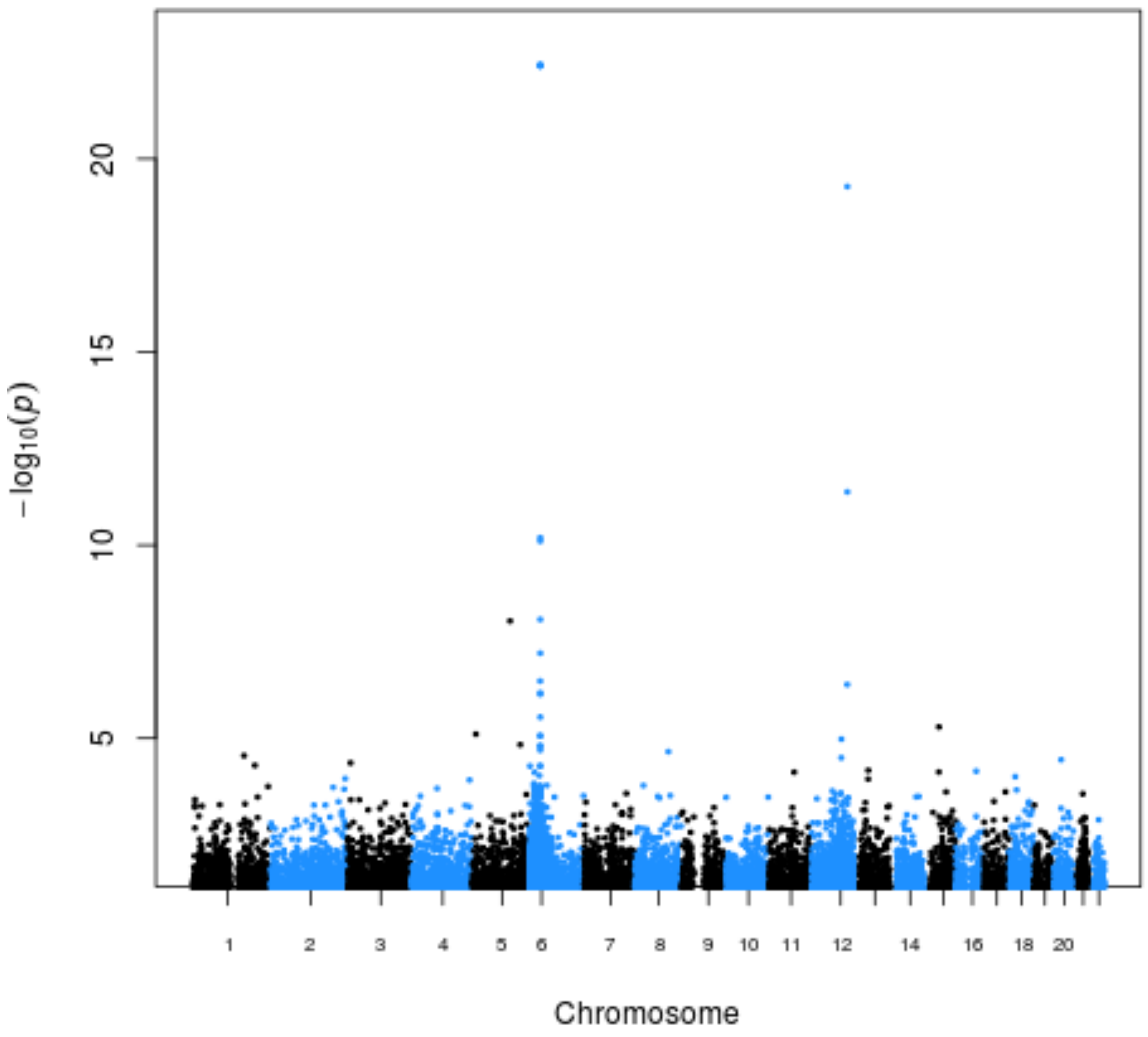
```
png("mh5.png")
manhattan(new5, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
```



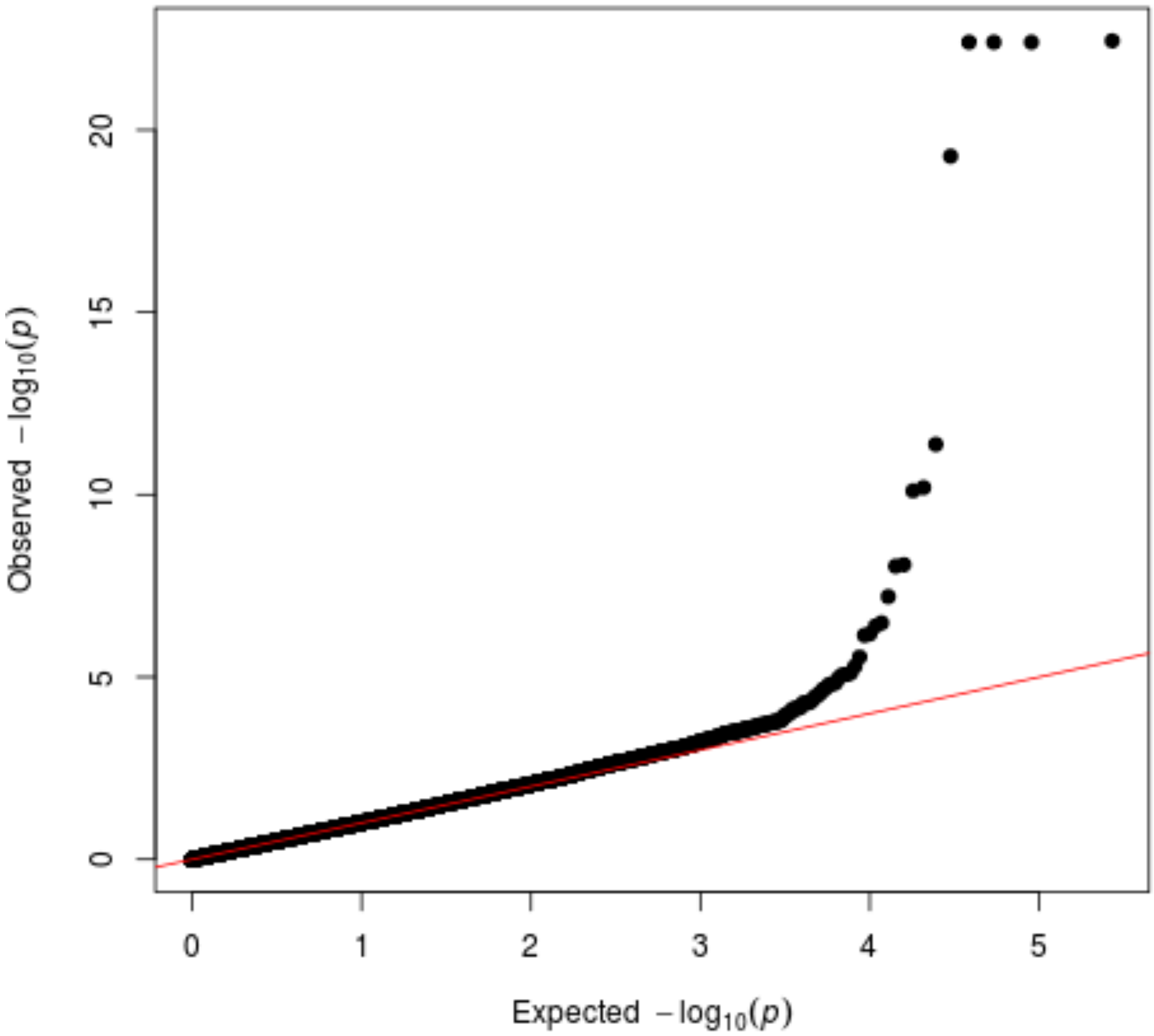
```
new6<-data.frame(res6$SNP, res6$CHR, res6$POS, res6$P)
names(new6)<-c("SNP", "CHR", "BP", "P")

png("qq6.png")
qq(new6$P)
dev.off()
```



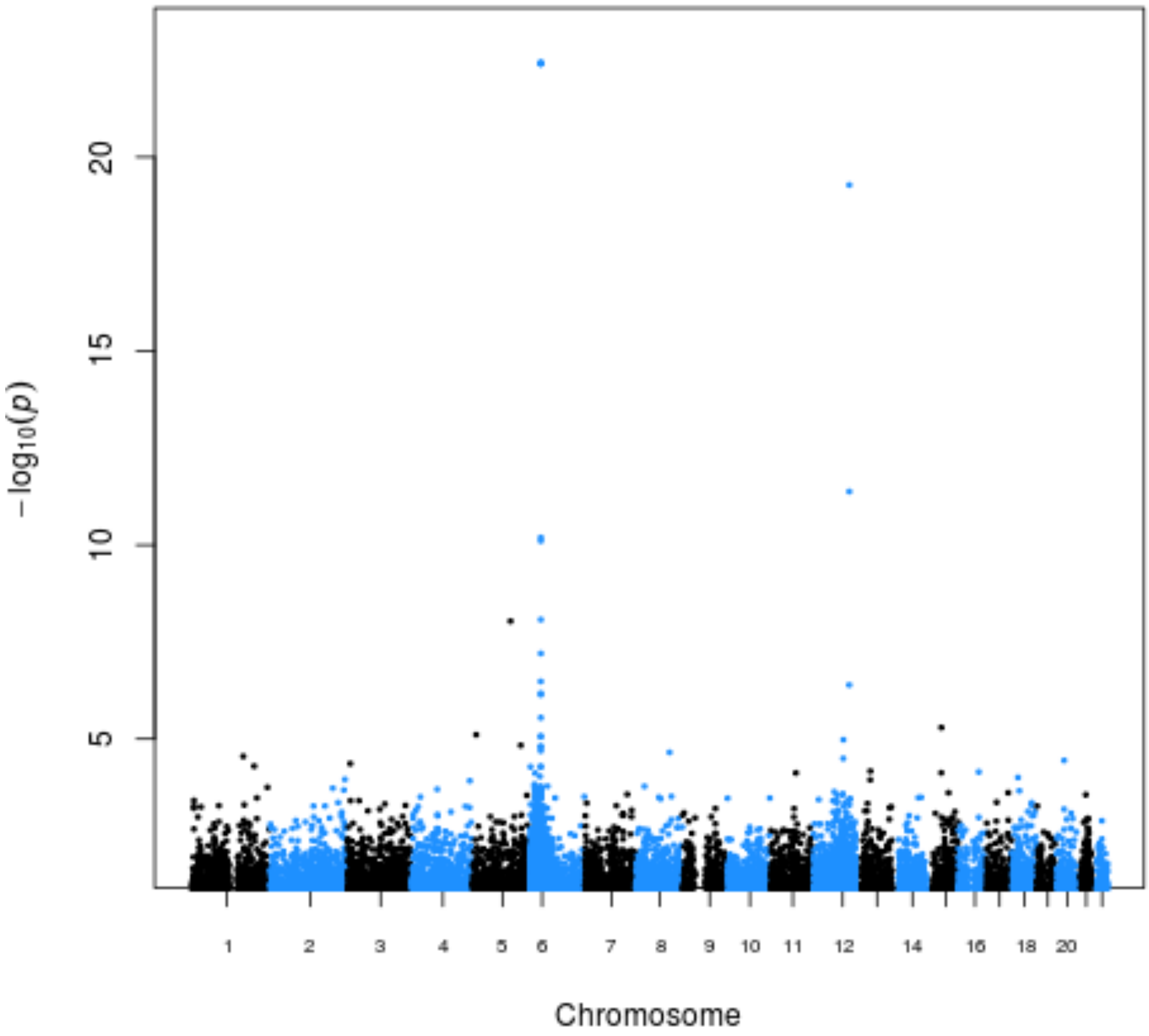
```
png("mh6.png")
manhattan(new6, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
```



```
new7<-data.frame(res7$SNP, res7$CHR, res7$POS, res7$P)
names(new7)<-c("SNP", "CHR", "BP", "P")

png("qq7.png")
qq(new7$P)
dev.off()
```

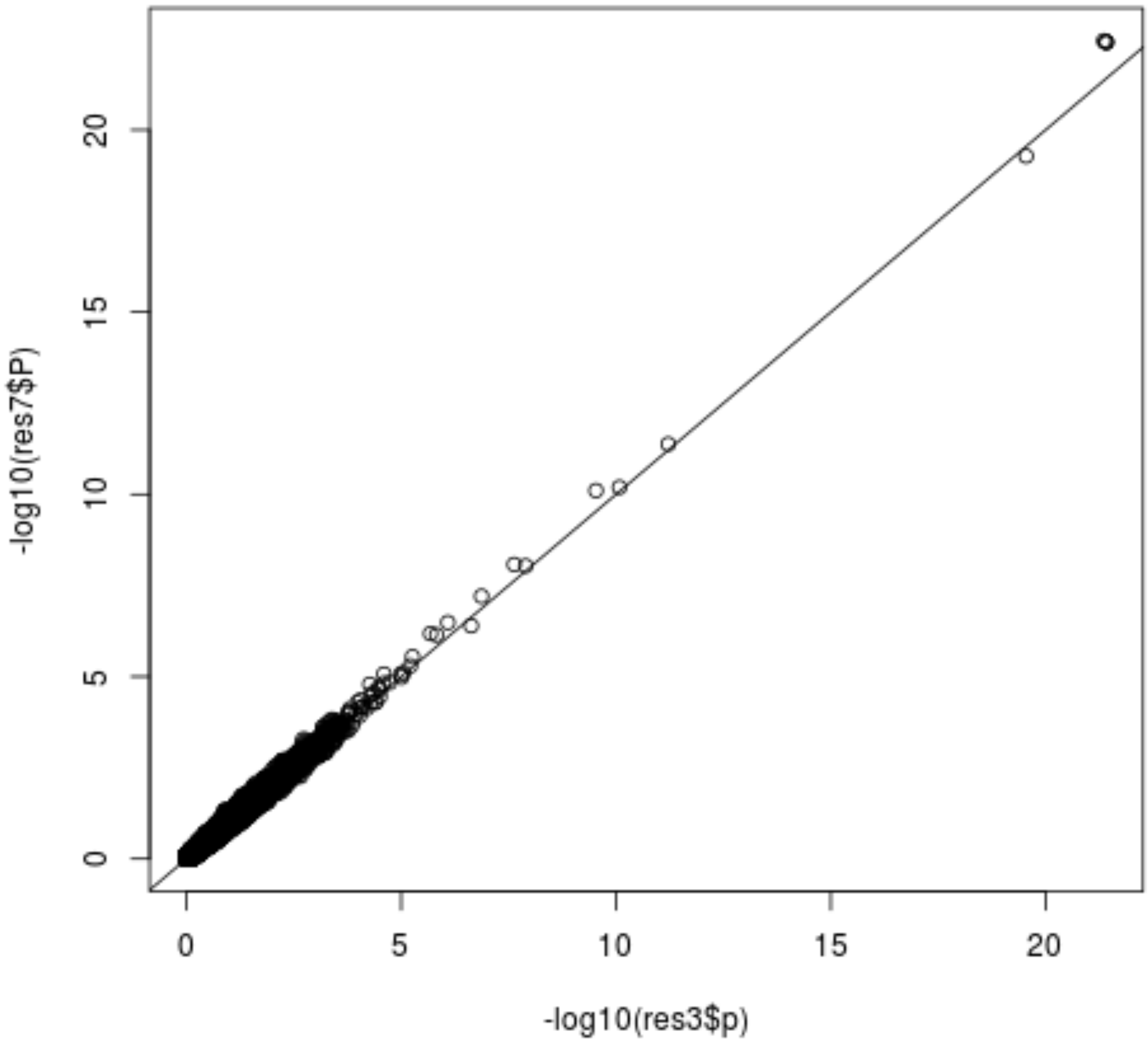


```
png("mh7.png")
manhattan(new7, pch=20, suggestiveline=F, genomewideline=F, ymin=2,
cex.x.axis=0.65, colors=c("black","dodgerblue"), cex=0.5)
dev.off()
```



Finally we will check how the results from the linear mixed model (with a sparse GRM) from fastGWA compares to the original GCTA `--mlma` results:

```
res3<-read.table("GCTAresults.mlma", header=T)
res7<-read.table("LMMnewsparse_assoc.fastGWA", header=T)
head(res3)
head(res7)

png("compareGCTAfastGWA.png")
plot(-log10(res3$p),-log10(res7$P))
abline(0,1)
dev.off()
```



As you can see, the results are extremely similar.

Answers

How to interpret the output

Interpretation of the output is described in the step-by-step instructions. Please ask if you need help in understanding the output.

Comments

Other packages

Another package that can implement a similar analysis to GCTA is DISSECT

References

Yang et al. (2011) GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics, 88:76-82.

Exercises prepared by: Heather Cordell
Checked by:
Programs used: R, GCTA
Last updated: 01/15/2021 14:53:5501/15/2021 14:53:36

Interaction analysis using PLINK and CASSI

Overview

Purpose

In this exercise you will be performing association analysis and testing for interaction effects using case/control data.

Methodology

The methodology used includes logistic regression in PLINK and CASSI, as well as some related alternative approaches.

Program documentation

PLINK documentation:

PLINK has an extensive set of docmentation including a pdf manual, a web-based tutorial and web-based documentation:

Original PLINK (1.07) (which has arguably clearer documentation):
http://zzz.bwh.harvard.edu/plink/

New PLINK (1.90) (which includes documentation on new additional features):
https://www.cog-genomics.org/plink2

CASSI documentation:

CASSI documentation is available from:

http://www.staff.ncl.ac.uk/richard.howey/cassi/downloads.html

Exercise

Data overview

The data consists of simulated genotype data at 100 SNP loci, typed in 2000 cases and 2000 controls. The data has been simulated in such a way that the first five SNPs have some relationship with disease, whereas the remaining 95 SNPs have no effect on disease outcome.

The complication with these data is that SNPs 1 and 2 have been simulated in such a way that they show no marginal association with the disease: their association will only be visible when you look at both SNPs in combination. SNPs 3-5 have been simulated to only have an effect on disease when an individual is homozygous at all three of these loci. Although potentially this could lead to marginal effects at the loci, formally this corresponds to a model of pure interaction, with no main effects, at these 3 SNPs.

Appropriate data

Appropriate data for this exercise is genotype data for a set of linked or unlinked loci typed in a group of unrelated affected individuals (cases) and in a group of unaffected or randomly chosen individuals from the same population (controls).

All the programs will deal with much larger numbers of loci than the 100 SNPs considered here. PLINK, in particular, was specifically designed for the analysis of large numbers of loci e.g. generated as part of a genome-wide association study.

Instructions

Data format

The data for the 100 SNPs simcasecon.ped is in standard linkage pedigree file format, with columns corresponding to family id, subject id (within family), father's id, mother's id, sex (1=m, 2=f), affection status (1=unaffected, 2=affected) and one column for each allele for each locus genotype. Note that since this is case/control rather than family data, there is only one individual per family and everyone's parents are coded as unknown.

PLINK requires an additional map file simcasecon.map describing the markers (in order) in the pedigree file. The PLINK-format map file contains exactly 4 columns:

```
chromosome (1-22, X, Y or 0 if unplaced)
rs number or snp identifier
Genetic distance (morgans) (not often used - so can be set to 0)
Base-pair position (bp units)
```

Take a look at the data files, and check that you understand how the data is coded. Then (if necessary) save the files as .txt files to the appropriate directory (folder) on your computer.

Step-by-step instructions

1. Analysis in PLINK

Move to the directory where you have saved the data files.

To carry out a basic association analysis in PLINK, type

```
plink --ped simcasecon.ped --map simcasecon.map --assoc
```

Here the `--ped xxxx` command tells PLINK that the name of the pedigree file is `xxxx` and the `--map yyyy` command tells PLINK that the name of the map file is `yyyy`. The `--assoc` command tells PLINK to perform a basic allele-based chisquared association test.

PLINK outputs some useful messages (you should always read these to make sure they match up with what you expect!) and outputs the results to a file `plink.assoc` .

Take a look at the file `plink.assoc` (e.g. by typing `more plink.assoc`). For each SNP the following columns of results are reported:

CHR	**Chromosome**
SNP	**SNP ID**
BP	**Physical position (base-pair)**
A1	**Minor allele name (based on whole sample)**
F_A	**Frequency of this allele in cases**
F_U	**Frequency of this allele in controls**
A2	**Major allele name**
CHISQ	**Basic allelic test chi-square (1df)**
P	**Asymptotic p-value for this test**
OR	**Estimated odds ratio (for A1, i.e. A2 is reference)**

Does there appear to be evidence of association at any of the five "true" loci? What about the 95 null loci?

Try performing a genotype-based (rather than an allele-based) analysis in PLINK and take a look at the results by typing the following 3 commands:

```
plink --ped simcasecon.ped --map simcasecon.map --model
head -1 plink.model
grep GENO plink.model
```

Again, does there appear to be evidence of association at any of the five "true" loci? What about the 95 null loci?

To test for pairwise epistasis in PLINK, the fastest option is to use the `--fast-epistasis` command:

```
plink --ped simcasecon.ped --map simcasecon.map --fast-epistasis
```

Formally, this tests whether the OR for association between two SNPs differs between cases and controls, which can be shown to appriximate a logistic regression based test of interaction between the SNPs. Results can be found in the file `plink.epi.cc`. Only pairwise interaction tests with p <= 0.0001 are reported (otherwise, for genome-wide studies, there would be too many results to report, given the large number of pairwise tests performed).

Take a look at the file `plink.epi.cc`. You should find a very significant interaction between SNPs 1 and 2, and a less significant iteraction between SNPs 15 and 77. Since this is simulated data, we know that this less significant result is a false positive.

A more powerful test for SNPs that are not in LD with one another (i.e. that are not too close to one another, in terms of their genomic location) is to additionally use the `--case-only` option:

```
plink --ped simcasecon.ped --map simcasecon.map --fast-epistasis --case-only
```

Results can be found in the file `plink.epi.co` . Again only pairwise interaction tests with p <= 0.0001 are reported. You should again find a very significant interaction between SNPs 1 and 2 (even more significant than previously, owing to the increased power with a case-only test).

A problem with the --fast-epistasis test is that it can be affected by LD between the SNPs (although only the case-only test is seriously affected). A more accurate test is to carry out logistic regresion by using the slower `--epistasis` command:

```
plink --ped simcasecon.ped --map simcasecon.map --epistasis
```

Results can again be found in the file `plink.epi.cc` (which will now have been overwritten). You can see that again the interaction between SNPs 1 and 2 remains highly significant (p=1.22E-63), together with just one other (false positive) interaction between SNPs 15 and 77.

Since the `--epistasis` option is slower, but most accurate, for genome-wide studies it might be sensible to first to screen for interactions using the `--fast-epistasis` command, but then confirm any findings using the `--epistasis` command on the smaller set of detected SNPs.

2. Analysis in CASSI

We will also compare our PLINK results with those obtained using the CASSI program, which implements a variety of tests including linear and logistic regression, and an improved Joint Effects (JE) test of pairwise interaction as described in Ueki and Cordell (2012). First we need to convert our data to PLINK binary format:

```
plink --ped simcasecon.ped --map simcasecon.map --make-bed --out simbinary
```

This should create PLINK binary format files `simbinary.bed`, `simbinary.bim` and `simbinary.fam`. Then we use the CASSI program with the input file `simbinary.bed` to perform pairwise interaction tests at all pairs of loci. (By default, only those pairs of SNPs showing interaction with a p-value < 0.0001 are output, though this can be changed if desired).

We start by using logistic regression. The logistic regression test in CASSI is essentially the same as the `--epistasis` test in PLINK, except that CASSI uses a likelihood ratio test rather than the asymptotically equivalent Wald (?) test used by PLINK. CASSI also has the advantage of allowing covariates into the analysis, if desired.

```
cassi -lr -i simbinary.bed
```

Take a look at the output file `cassi.out` The most important columns are the first 4 columns (listing the SNP numbers/names) and the last 4 columns listing the log odds ratio, its standard error, the likelihood ratio chi-squared test statistic and its p-value. It can be quite hard to work out which column is which, so we suggest you start up R by typing

```
R
```

and then read in and look at the results by typing

```
results<-read.table("cassi.out", header=T)
results
```

You can see that SNPs 1 and 2 show a very strong pairwise interaction (p=5.94E-72), which is actually a bit more significant than the result from PLINK (p=1.22E-63). We also still detect the false positive interaction between SNPs 15 and 77.

Now try using the Joint Effects (JE) test, telling CASSI to use the output filename `cassiJE.out`

```
cassi -je -o cassiJE.out -i simbinary.bed
```

Take a look at the output file `cassiJE.out`. The most important columns are the first 4 columns (listing the SNP numbers/names) and the last 4 columns listing the case/control and case-only interaction test chi-squareds and p-values. Again it can be quite hard to work out which column is which, so we suggest you read in and look at the results in R:

```
resultsJE<-read.table("cassiJE.out", header=T)
resultsJE
```

You can see that SNPs 1 and 2 show a very strong pairwise interaction (Case-Con test p-value JE_CC_P=1.67e-129; Case-Only test p-value JE_CO_P=1.71e-274). Interestingly we also detect, albeit at lower significance levels, the (true) pairwise interactions between SNPs 3 and 4 and between SNPs 4 and 5. We also detect two false positive interactions, between SNPs 15 and 77, and between SNPs 31 and 100.

Answers

Interpretation of output

Answers and interpretation of the output are described in the step-by-step instructions. Please ask if you need help in understanding the output for any specific test.

Comments

Advantages/disadvantages

PLINK and CASSI are designed for genome-wide studies, allowing the inclusion of many thousands of markers. Analysis in a standard statistical package does not generally allow so many markers, but may have some advantage of allowing a lot of extra flexibility with regards to the models and analyses performed e.g. it easy to include additional predictor variables such as environmental factors, gene-environment interactions etc. However, you are required to know or learn how to use the package in order to gain that extra flexibility, and to produce reliable results.

Study design issues

With case/control data it is relatively easy to obtain large enough sample sizes to detect small genetic effects. However, detection of interactions generally requires much larger sample sizes.

Other packages

Logistic regression analysis for detection of interactions can be performed in most statistical packages such as R, Stata, SAS, SPSS. Alternative Bayesian Epistasis mapping approaches are available in the BEAM (Zhang et al. 2007; Zhang 2011) or BIA software packages.

Several packages are available for implementing different data-mining and machine-learning approaches for detecting interactions or detecting association allowing for interaction. See Cordell (2009) and other references below for more details.

References

Cordell HJ (2009) Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10(6):392-404.

Y Chung and S Y Lee and R C Elston and T Park (2007) Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. Bioinformatics 23:71-76.

L W Hahn and M D Ritchie and J H Moore (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions Bioinformatics 19:376-382.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81:559-575.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF and Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138-147.

Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genetics 8(4):e1002625.

Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39:1167-1173.

Zhang Y (2011) A novel Bayesian graphical model for genome-wide multi-SNP association mapping. Genet Epidemiol 36: 36-47.

Exercises prepared by: Heather Cordell
Checked by:
Programs used: PLINK, CASSI
Last updated: 01/17/2020 12:35:48

Sample Size Calculations - Cochran-Armitage Test for Trend

Webpage for the exercises:
~~http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html~~
http://zzz.bwh.harvard.edu/gpc/cc2.html

Question 1
For a complex disease study, you plan to collect 35,000 cases and 70,000 controls and wish to know if this is a sufficient sample size to detect associations with disease susceptibility loci. The disease has a population prevalence of 5%. You wish to estimate the power for a genotypic relative risk of 1.2 and a disease allele frequency of 0.02. What is the power for $\alpha=5\text{x}10^{-8}$ under a under a multiplicative model ($\gamma_2 = \gamma_1^2$) a.)_________and dominant model ($\gamma_2 = \gamma_1$) b.) _____________?

Question 2
For your study, you hypothesize that you will try to replicate associations for 100 variants that are in linkage equilibrium and you want to reject the null hypothesis using a p-value of 0.05. What is the Bonferroni correction you should use a.)__________. Determine what your power would be if you used a Bonferroni correction to control for the Family Wise Error Rate (FWER) for testing 100 variants. Using the parameters provided in question 1 but for a sample size of 20,000 cases and 20,000 controls what is the power under the multiplicative model b.)_______________ and under a dominant model c.)___________________?

Question 3
You determine that you can ascertain 50,000 cases and 50,000 controls what is the power using the same parameters as described in question 1 for the multiplicative model _______________ and dominant model______________________?

Question 4
The power of the Cochran-Armitage test for trend is dependent on the underlying genetic model. Using the parameters from question 1 which of the following underlying genetic models: multiplicative ($\gamma_2 = \gamma_1^2$), additive($\gamma_2 = 2\gamma_1 - 1$), dominant ($\gamma_2 = \gamma_1$) or recessive ($\gamma_1 = 1$) would you predict to be the most powerful a.)______________ and least powerful b.)____________________?

Question 5
For study design with equal numbers of cases and controls a genotype relative risk of 1.5 under a recessive model for a disease with a population prevalence of 0.05 and disease allele frequency of 0.1. How many cases a.)______ and controls b.) ________should you ascertain for $\alpha=5.0$ x 10^{-8} and $1-\beta=0.80$? *Use power2 or Genetic Power Calculator, GAS power cannot calculate for more than 100,000 cases.

Question 6
You are performing a rare variant association study and you assume that that cumulative frequency of the causal variants in your gene region is 0.01 with every variant having an effect size of 1.4. The disease you are studying has a prevalence of 5%. For a study with 0.8 power and an $\alpha=2.5$ x 10^{-6} under a dominant model for equal numbers of cases and controls what is the total sample size a.) __________ do you need to ascertain. What is the total sample size b.)_________you need to ascertain if the cumulative frequency of causal variants is only 0.005?

Question 7

You are performing a study using the UK Biobank and for your phenotype of interest you have 50,000 cases and 100,000 controls. For a disease with 10% prevalence, disease allele frequency of 0.01, where each variant has an effect size of 1.2 under a dominant model what would be the power for an aggregate test where the cumulative allele frequency is 0.01 _________and a single variant test ___________? Clue use the appropriate alpha for each test.

Question 8

Using have a replication sample of 50,000 cases and 50,000 controls and you plan to try to replicate 15 genes and 100 variants. Using the same parameters as in question 7 what would be your power to replicate a.)________________? Note for alpha use a Bonferroni correction.

Question 9

For the above power calculations, you have been using the relative risk which only approximates the odds ratio when a.) ______________________? You are performing a power calculation for a case control study for a disease/variant frequency of 0.01. You use a dominant model and a gamma of 1.2 for a disease with a prevalence for 0.2. What is the odds ratio for which the power calculations are being performed b.) _________________? *Use Genetic Power Calculator – information not provided by GAS or Power2.

ANWSERS

1. a.) 0.702 b.) 0.654
2. a.) 5.0x10^{-4} b.) 0.690 c.) 0.657
3. a.) 0.798 b.) 0.755
4. a.) multiplicative b.) recessive
5. a.) 170,910 b.) 170,910
6. a.) ~43,000 b.) ~84,300
7. a.) 0.73 b.) 0.45 Hint: use α=5x10^{-8} for single variant test and α=2.5x10^{-6} for the aggregate test
8. a.) 0.87 (Hint: use α=4.3x10^{-4})
9. a.) only for disease with low prevalence does the relative risk does not estimate the odds ratio b.) 1.26

Advanced Gene Mapping Course: Pleiotropy Exercise

Andrew DeWan, PhD, MPH

This exercise was designed to give you practical experience identifying cross phenotype associations using both univariate and multivariate methods and then dissecting these cross-phenotype associations to determine if they show evidence of biological and/or mediated pleiotropy.

A population-based dataset with 3000 subjects and two quantitative traits (Trait 1 and Trait 2) along with 2000 SNPs on one chromosome were simulated. Let's assume that Trait 1 was measured 20 years prior to Trait 2 (i.e. Trait 1 will act as the mediator in our mediation analysis). The two quantitative traits are correlated and there are markers associated with one or both phenotypes as well as unassociated.

The dataset has been QC'd. The files for the initial analyses are:

pleiotropy_exercise.bed, .bim, .fam and pleiotropy_exercise_phenotypes.txt

I have included a summary table that you will want to fill out as you are working through this exercise. This will help keep track of the SNPs you select for the mediation analysis as well as the interpretation of the results at the end of the exercise

<u>Univariate analyses</u>

a. Conduct a univariate analysis (using --linear) in PLINK for both datasets and both traits
Note: You will need to use the --pheno/--pheno-name commands to specify the phenotype file and phenotype name.

```
plink\
 --bfile pleiotropy_exercise\
 --pheno pleiotropy_exercise_phenotypes.txt\
 --pheno-name Trait1\
 --sex\
 --linear\
 --out Trait1
```

For use in several downstream steps, let's create files with only the header and SNP results for each of the univariate analyses:

```
grep 'TEST' Trait1.assoc.linear > Trait1_snp.assoc.linear
grep 'ADD' Trait1.assoc.linear >> Trait1_snp.assoc.linear
grep 'TEST' Trait2.assoc.linear > Trait2_snp.assoc.linear
grep 'ADD' Trait2.assoc.linear >> Trait2_snp.assoc.linear
```

b. Try visualizing the data by creating a Hudson plot in R. This will give you some sense of the overlapping signals between the two association analyses.

```
library(hudson)
dat1<-read.table("Trait1_snp.assoc.linear",header=T)
dat2<-read.table("Trait2_snp.assoc.linear",header=T)
names(dat1_snps)<-c("CHR","SNP", "POS", "A1", "TEST", "NMISS", "BETA", "STAT", "pvalue")
names(dat2_snps)<-(names(dat1_snps)
gmirror(top=dat1_snps, bottom=dat2_snps, tline=5e-08, bline=5e-08,
+ toptitle="Trait11", bottomtitle = "Trait2",
+ highlight_p = c(0.00000005,0.00000005), highlighter="green",
+ file = 'pleiotropy_hudson', res = 300, type = 'pdf')
```

c. Now Identify genome-wide significant SNPs (p<5x10^{-8}) that overlap for both traits. This can be done using some simple R code:

```
Trait1 <- read.table(“Trait1_snp.assoc.linear”, header = T)
Trait2 <- read.table(“Trait2_snp.assoc.linear”, header = T)
SigTrait1 <- subset(Trait1, P<0.00000005)
SigTrait2 <- subset(Trait2, P<0.00000005)
intersect(SigTrait1$SNP, SigTrait2$SNP)
```

d. As you can see, there are some genome-wide significant SNPs that are adjacent or close to each other. To explore whether or not these are independent associations, let’s perform some simple LD clumping. You will want to carry through the index SNP identified for each clumped region. You will also want to carry through any SNPs from 1c above that were not part of a clumped region. plink\

```
--bfile pleiotropy_exercise\
--clump Trait1_snp.assoc.linear,Trait2_snp.assoc.linear\
--clump-kb 250\
--clump-p1 5e-8\
--clump-p2 5e-8\
--clump-r2 0.2\
--clump-replicate\
--clump-verbose\
--out Trait1_Trait2_clump
```

Multivariate analysis

a. Before moving on to dissecting the cross-phenotype associations, let's see if we can include a few additional SNPs/regions to explore by using multivariate analysis. But let's only consider additional regions that are genome-wide suggestive for both phenotypes.

   First run a multivariate analysis on Traits 1 and 2.

```
plink.multivariate\
 --noweb\
 --bfile pleiotropy_exercise\
 --mult-pheno pleiotropy_exercise_phenotypes.txt\
 --sex\
 --mqfam\
 --out Trait1_Trait2
```

   Please note: You should use the --noweb flag due to this program being built on an old version of PLINK.

b. Now let's identify the intersection of SNPs that are genome-wide significant in the multivariate analysis and at least suggestive for each trait in the univariate analysis, i.e. we want to make sure that both traits are contributing to the multivariate signal.

```
Trait1<-read.table("Trait1_snp.assoc.linear", header=T)
Trait2<-read.table("Trait2_snp.assoc.linear", header=T)
multi<-read.table("Trait1_Trait2.mqfam.total", header=T)
sigMulti<-subset(multi, P<0.00000005)
suggTrait1<-subset(Trait1, P<0.000005)
suggTrait2<-subset(Trait2, P<0.000005)
Reduce(intersect, list(suggTrait1$SNP, suggTrait2$SNP, sigMulti$SNP))
```

   Select the additional SNPs that are identified from the intersection of the multivariate analysis and genome-wide suggestive lists for both traits that were not in your original list.

c. You may want to re-run the LD clumping with a suggestive threshold to see if these additional SNPs clump with your existing clumps or are new potential regions to explore.

```
plink\
 --bfile pleiotropy_exercise\
 --clump Trait1_snp.assoc.linear,Trait2_snp.assoc.linear\
 --clump-p1 0.000005\
 --clump-p2 0.000005\
 --clump-r2 0.2\
 --clump-replicate\
 --clump-verbose\
 --out Trait1_Trait2_clump_suggestive
```

Mediation analyses

a. For each SNPs that you have identified as a cross phenotype association (evidence of overlapping association signals as well as incorporating results from LD clumping and multivariate association) you will need to extract this data from the original plink files and create a genotype file that is coded as 0|1|2 for the genotypes. This can be done in PLINK using the --recodeA command and the --extract command by providing a file with the list of snps. This will give you a .raw genotype file with only the snps that you will be using in the mediation analysis.

b. Conduct a mediation analysis in R using the *mediation* R library. Sample code for this is below (Note: replace <SNP> with the variable name for the SNP you are investigating. You will need to repeat this for each SNP that you have selected):

```
library(mediation)
genotypes <- read.table("snps_for_mediation.raw", header=T)
phenotypes<-read.table("pleiotropy_exercise_phenotypes.txt", header=T)
combined<-merge(genotypes,phenotypes)
med.fit<-lm(Trait1~rs125_0, data=combined)
out.fit<-lm(Trait2~Trait1+rs125_0, data=combined)
med.out<-mediate(med.fit,out.fit,treat="rs125_0", mediator="Trait1", boot=TRUE,
+boot.ci.type="bca", sims=1000)
summary(med.out)
```

This will print out a summary of the mediation analysis.

Please note: The more simulations (sims) you specific in the med.out step the more the CI and p-value estimates will be, however, this can also be time-consuming. If this step is taking a substantial amount of time (>20 minutes) you may want to reduce the number of simulations for the purposes of completing the exercise.

Questions:

1) Which of the SNPs have genome-wide significant ($p<5x10^{-8}$) associations for both traits?

2) Did the multivariate analyses result in additional SNPs that had genome-wide significant cross phenotype associations? Which SNP(s)?

3) For each SNP analyzed in the mediation analysis, determine if there is a significant direct effect which is indicative of some level of biological pleiotropy. Do any of the SNPs exhibit complete mediation?

4) Why do some of the SNPs have negative values for the proportion mediated?

Summary table of pleiotropy results

SNP	Beta (Trait 1)	P (Trait 1)	Beta (Trait 2)	P (Trait 2)	MV (P)	MV Loading (Trait 1)	MV Loading (Trait 2)	ADE	ADE (P)	ACME	ACME (P)	Total Effect	Total Effect (P)	Prop Mediated	Prop Mediated (P)

Pleiotropy Exercise - Answers

Andrew DeWan, PhD, MPH

This exercise was designed to give you practical experience identifying cross phenotype associations using both univariate and multivariate methods and then dissecting these cross phenotype associations to determine if they show evidence of biological and/or mediated pleiotropy.

A population-based dataset with 3000 subjects and two quantitative traits (Trait 1 and Trait 2) along with 2000 SNPs on one chromosome were simulated. Let's assume that Trait 1 was measured 20 years prior to Trait 2 (i.e. Trait 1 will act as the mediator in our mediation analysis). The two quantitative traits are correlated and there are markers associated with one or both phenotypes as well as unassociated.

The dataset has been QC'd. The files for the initial analyses are:

pleiotropy_exercise.bed, .bim, .fam and pleiotropy_exercise_phenotypes.txt

I have included a summary table that you will want to fill out as you are working through this exercise. This will help keep track of the SNPs you select for the mediation analysis as well as the interpretation of the results at the end of the exercise.

Univariate analyses

a. Conduct a univariate analysis (using --linear) in PLINK for both datasets and both traits
   Note: You will need to use the --pheno/--pheno-name commands to specify the phenotype file and phenotype name.

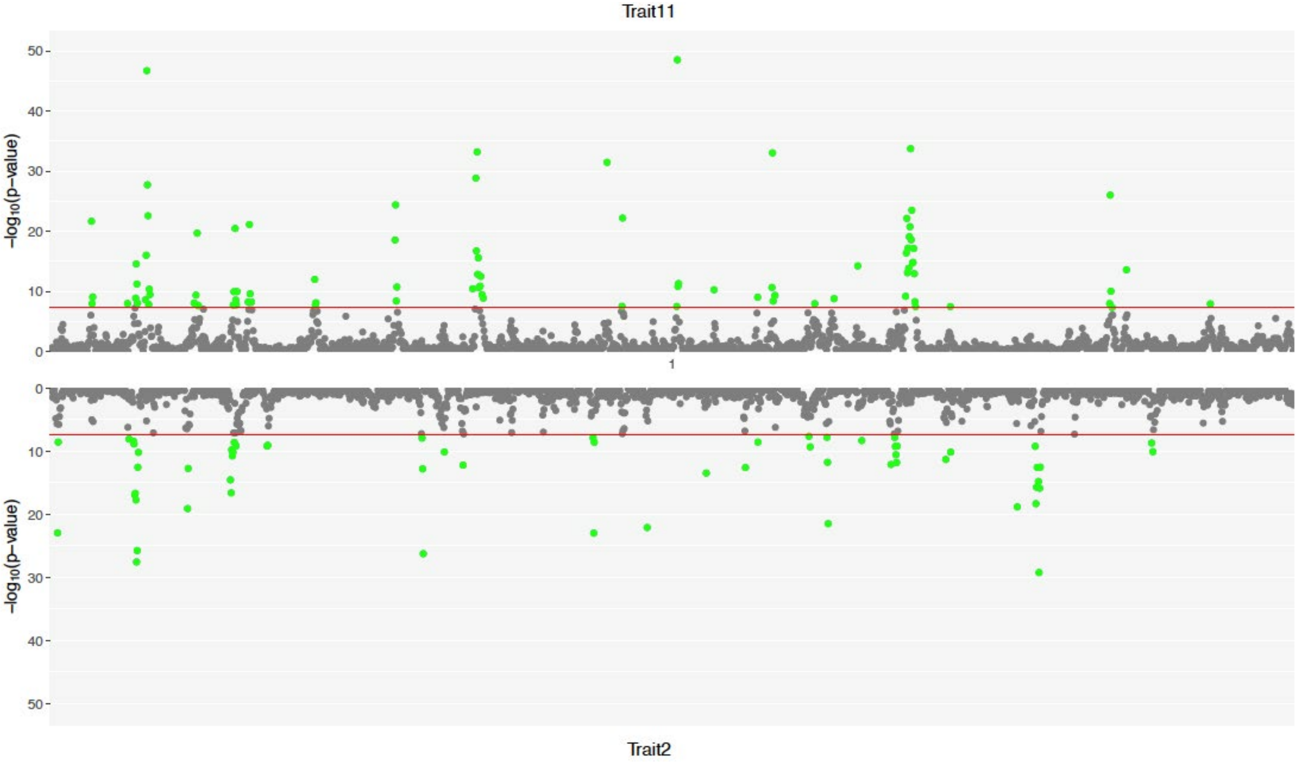
```
plink\
 --bfile pleiotropy_exercise\
 --pheno pleiotropy_exercise_phenotypes.txt\
 --pheno-name Trait1\
 --sex\
 --linear\
 --out Trait1
```

   For use in several downstream steps, let's create files with only the header and SNP results for each of the univariate analyses:

```
grep 'TEST' Trait1.assoc.linear > Trait1_snp.assoc.linear
grep 'ADD' Trait1.assoc.linear >> Trait1_snp.assoc.linear
grep 'TEST' Trait2.assoc.linear > Trait2_snp.assoc.linear
grep 'ADD' Trait2.assoc.linear >> Trait2_snp.assoc.linear
```

b. Try visualizing the data by creating a Hudson plot in R. This will give you some sense of the overlapping signals between the two association analyses.

```
devtools::install_github('anastasia-lucas/hudson')
library(hudson)
dat1<-read.table("Trait1_snp.assoc.linear",header=T)
dat2<-read.table("Trait2_snp.assoc.linear",header=T)
names(dat1_snps)<-c("CHR",  "SNP",  "POS",  "A1",  "TEST",  "NMISS",  "BETA",
+"STAT", "pvalue")
names(dat2_snps)<-(names(dat1_snps)
gmirror(top=dat1_snps, bottom=dat2_snps, tline=5e-08, bline=5e-08,
+ toptitle="Trait11", bottomtitle = "Trait2",
+ highlight_p = c(0.00000005,0.00000005), highlighter="green",
+ file = 'pleiotropy_hudson', res = 300, type = 'pdf')
```



c. Now Identify genome-wide significant SNPs ($p<5x10^{-8}$) that overlap for both traits. This can be done using some simple R code:

```
Trait1 <- read.table(“Trait1_snp.assoc.linear”, header = T)
Trait2 <- read.table(“Trait2_snp.assoc.linear”, header = T)
SigTrait1 <- subset(Trait1, P<0.00000005)
SigTrait2 <- subset(Trait2, P<0.00000005)
intersect(SigTrait1$SNP, SigTrait2$SNP)
```

The overlapping genome-wide significant variants are: rs138, rs139, rs140, rs141, rs296, rs299, rs1138, rs1448

d. As you can see, there are some genome-wide significant SNPs that are adjacent or close to each other. To explore whether or not these are independent associations, let’s perform some simple LD clumping. You will want to carry through the index SNP identified for each clumped region. You will also want to carry through any SNPs from 1c above that were not part of a clumped region.

```
plink\
 --bfile pleiotropy_exercise\
 --clump Trait1_snp.assoc.linear,Trait2_snp.assoc.linear\
 --clump-kb 250\
 --clump-p1 5e-8\
 --clump-p2 5e-8\
 --clump-r2 0.2\
 --clump-replicate\
 --clump-verbose\
 --out Trait1_Trait2_clump
```

OUTPUT:

```
CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    2   rs139     139000   2.86e-28        9      0      0      0
0      9

                          KB      RSQ  ALLELES    F            P
  (INDEX)   rs139          0    1.000        0    2     2.86e-28

            rs137         -2    0.247    00/00    2     2.17e-17
            rs138         -1    0.399    00/00    1     1.34e-09
            rs138         -1    0.399    00/00    2     1.91e-18
            rs139          0        1    00/00    1     2.77e-15
            rs140          1    0.229    00/00    1     6.05e-12
            rs140          1    0.229    00/00    2     1.85e-26
            rs141          2    0.235    00/00    1      9.9e-09
            rs141          2    0.235    00/00    2     2.98e-13

          RANGE: chr1:137000..141000
           SPAN: 4kb

------------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    1   rs296     296000   1.15e-10        5      0      0      0
0      5

                          KB      RSQ  ALLELES    F            P
  (INDEX)   rs296          0    1.000        0    1     1.15e-10

            rs295         -1    0.429    00/00    1     2.01e-08
            rs296          0        1    00/00    2      2.6e-09
            rs299          3    0.267    00/00    1     2.77e-09
            rs299          3    0.267    00/00    2     7.29e-10

          RANGE: chr1:295000..299000
           SPAN: 4kb
```

From clump 1, let’s choose rs139 (and not rs138, rs140 and rs141) and from clump 2 let’s choose rs296 (and not rs295 and rs299) to carry forward to our mediation analysis. We will also carry forward rs1138 and rs1448 since these two SNPs are not part of any other clumps but are genome-wide significant for both traits.

Multivariate analysis

a. Before moving on to dissecting the cross phenotype associations, let's see if we can include a few additional SNPs/regions to explore by using multivariate analysis. But let's only consider additional regions that are genome-wide suggestive for both phenotypes.

   First run a multivariate analysis on Traits 1 and 2.

```
plink.multivariate\
 --noweb\
 --bfile pleiotropy_exercise\
 --mult-pheno pleiotropy_exercise_phenotypes.txt\
 --sex\
 --mqfam\
 --out Trait1_Trait2
```

Please note: You should use the --noweb flag due to this program being built on an old version of PLINK.

b. Now let's identify the intersection of SNPs that are genome-wide significant in the multivariate analysis and at least suggestive for each trait in the univariate analysis, i.e. we want to make sure that both traits are contributing to the multivariate signal.

```
Trait1<-read.table("Trait1_snp.assoc.linear", header=T)
Trait2<-read.table("Trait2_snp.assoc.linear", header=T)
multi<-read.table("Trait1_Trait2.mqfam.total", header=T)
sigMulti<-subset(multi, P<0.00000005)
suggTrait1<-subset(Trait1, P<0.000005)
suggTrait2<-subset(Trait2, P<0.000005)
Reduce(intersect, list(suggTrait1$SNP, suggTrait2$SNP, sigMulti$SNP))
```

   Select the additional SNPs that are identified from the intersection of the multivariate analysis and genome-wide suggestive lists for both traits that were not in your original list.

   We identify the following overlapping SNPs: rs125, rs135, rs137, rs138, rs139, rs140, rs141, rs295, rs296, rs298, rs299, rs300, rs920, rs921, rs923, rs1138, rs1166, rs1361, rs1448. Of course, this list includes the original set of 8 variants that were genome-wide significant for both Traits 1 and 2.

c. You may want to re-run the LD clumping with a suggestive threshold to see if these additional SNPs clump with your existing clumps or are new potential regions to explore.

```
plink\
 --bfile pleiotropy_exercise\
 --clump Trait1_snp.assoc.linear,Trait2_snp.assoc.linear\
 --clump-p1 0.000005\
 --clump-p2 0.000005\
 --clump-r2 0.2\
 --clump-replicate\
 --clump-verbose\
 --out Trait1_Trait2_clump_suggestive
```

OUTPUT:

```
 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    2   rs139     139000   2.86e-28        9      0      0      0
0      9

                          KB      RSQ  ALLELES    F            P
  (INDEX)   rs139          0    1.000        0    2     2.86e-28

            rs137         -2    0.247    00/00    1     6.05e-08
            rs137         -2    0.247    00/00    2     2.17e-17
            rs138         -1    0.399    00/00    1     1.34e-09
            rs138         -1    0.399    00/00    2     1.91e-18
            rs139          0        1    00/00    1     2.77e-15
            rs140          1    0.229    00/00    1     6.05e-12
            rs140          1    0.229    00/00    2     1.85e-26
            rs141          2    0.235    00/00    1      9.9e-09
            rs141          2    0.235    00/00    2     2.98e-13

          RANGE: chr1:137000..141000
           SPAN: 4kb

------------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    1   rs921     921000   6.29e-23        5      0      0      0
1      4

                          KB      RSQ  ALLELES    F            P
  (INDEX)   rs921          0    1.000        0    1     6.29e-23

            rs920         -1    0.224    00/00    1     3.11e-08
            rs920         -1    0.224    00/00    2     6.25e-08
            rs921          0        1    00/00    2     1.94e-07
            rs922          1    0.202    00/00    1     4.52e-07

          RANGE: chr1:920000..922000
           SPAN: 2kb

------------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    2   rs136     136000    1.3e-17        5      0      1      0
0      4

                          KB      RSQ  ALLELES    F            P
  (INDEX)   rs136          0    1.000        0    2      1.3e-17

            rs134         -2    0.229    00/00    2     3.97e-09
            rs135         -1    0.379    00/00    1     1.41e-06
            rs135         -1    0.379    00/00    2     1.47e-09

          RANGE: chr1:134000..136000
           SPAN: 2kb

------------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    2  rs1361    1361000   1.68e-12        9      0      1      2
1      5

                          KB      RSQ  ALLELES    F            P
  (INDEX)  rs1361          0    1.000        0    2     1.68e-12

           rs1359         -2    0.238    00/00    2     5.98e-10
           rs1360         -1    0.281    00/00    2      2.8e-11
           rs1361          0        1    00/00    1     2.65e-07
           rs1362          1    0.271    00/00    2     6.54e-10
           rs1363          2    0.204    00/00    2     1.64e-07

          RANGE: chr1:1359000..1363000
           SPAN: 4kb

------------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    1   rs296     296000   1.15e-10        5      0      0      0
0      5

                          KB      RSQ  ALLELES    F            P
  (INDEX)   rs296          0    1.000        0    1     1.15e-10

            rs295         -1    0.429    00/00    1     2.01e-08
            rs295         -1    0.429    00/00    2     8.62e-08
            rs296          0        1    00/00    2      2.6e-09
            rs299          3    0.267    00/00    1     2.77e-09
            rs299          3    0.267    00/00    2     7.29e-10

          RANGE: chr1:295000..299000
           SPAN: 4kb

------------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    1  rs1138    1138000   9.58e-10        3      0      1      0
0      2

                          KB      RSQ  ALLELES    F            P
  (INDEX)  rs1138          0    1.000        0    1     9.58e-10
```

```
           rs1137         -1    0.315    00/00    1     4.09e-07
           rs1138          0        1    00/00    2      2.9e-09

          RANGE: chr1:1137000..1138000
           SPAN: 1kb

------------------------------------------------------------------
```

Based on the multivariate analysis and additional clumping, you should add the following SNPs to your list of SNPs for mediation: rs125, rs135, rs300, rs921, rs923, rs1166, rs1361.

The final list of SNPs that were selected to carry through to the mediation analysis are:

rs125, rs135, rs139, rs296, rs300, rs921, rs923, rs1138, rs1166, rs1361, rs1448

Mediation analyses

a. For each SNPs that you have identified as a cross phenotype association (evidence of overlapping association signals as well as incorporating results from LD clumping and multivariate association) you will need to extract this data from the original plink files and create a genotype file that is coded as 0|1|2 for the genotypes. This can be done in PLINK using the --recodeA command and the --extract command by providing a file with the list of snps. This will give you a .raw genotype file with only the snps that you will be using in the mediation analysis.

b. Conduct a mediation analysis in R using the *mediation* R library. Sample code for this is below (Note: replace <SNP> with the variable name for the SNP you are investigating. You will need to repeat this for each SNP that you have selected):

```
library(mediation)
genotypes <- read.table("snps_for_mediation.raw", header=T)
phenotypes<-read.table("pleiotropy_exercise_phenotypes.txt", header=T)
combined<-merge(genotypes,phenotypes)
med.fit<-lm(Trait1~rs125_0, data=combined)
out.fit<-lm(Trait2~Trait1+rs125_0, data=combined)
med.out<-mediate(med.fit,out.fit,treat="rs125_0", mediator="Trait1", boot=TRUE,
+boot.ci.type="bca", sims=1000)
summary(med.out)
```

This will print out a summary of the mediation analysis.

Please note: The more simulations (sims) you specific in the med.out step the more the CI and p-value estimates will be, however, this can also be time-consuming. If this step is taking a substantial amount of time (>20 minutes) you may want to reduce the number of simulations for the purposes of completing the exercise.

Questions:

1) Which of the SNPs have genome-wide significant ($p<5x10^{-8}$) associations for both traits?

   rs138, rs139, rs140, rs141, rs296, rs299, rs1138, rs1448

2) Did the multivariate analyses result in additional SNPs that had genome-wide significant cross phenotype associations but that also had genome-wide suggestive (p<5x10-6) univariate association for each trait? Which SNP(s)?

   rs125, rs135, rs137, rs295, rs298, rs300, rs920, rs921, rs923, rs1166, rs1361

   Instead of running mediation analysis on all 19 SNPs, I suggested that you perform LD clumping to reduce this number of SNPs and only focus on the index SNP for each clump (or if the index SNP was not associated with both traits to choose another SNP from among the clumped SNPs). This reduced the set of SNPs to 11.

3) For each SNP analyzed in the mediation analysis, determine if there is a significant direct effect which is indicative of some level of biological pleiotropy. Do any of the SNPs exhibit complete mediation?

   All SNPs show a significant direct effect on Trait 2 indicating some level of biological pleiotropy. rs923 has an ADE p-value of 0.002 but this is still less than the Bonferroni corrected p-value of 0.0045, adjusting for the 11 SNPs. No SNP shows an association with Trait 2 that is completely mediated through its association with Trait 1, i.e. an ACME estimate that is equal to (or close to) the total effect. The strongest mediated effect is for rs921 in which the mediated effect accounts for ~40% of the total effect of the SNP on Trait 2.

4) Why do some of the SNPs have negative values for the proportion mediated?

   The estimate of the proportion mediated is not the best way to interpret the mediation results, despite its seemingly obvious interpretability. In reality this proportion does not range from 0-1 but can rather be less than 0 and greater than 1. The negative proportion mediated values that we see for many of the SNPs we have analyzed is due to the fact that these SNPs have an effect estimate for the total effect and mediated effect that are in opposite directions, i.e. the effect of the SNPs on Trait 1 and Trait 2 is in opposite directions. Depending on your study question, you may want to limit your selection of SNPs to only those with effects on the two traits in the same direction. We did not see this among our SNPs, but a proportion mediated > 1 can happen when the strength of the association with the mediator (Trait 1) is much higher than the strength of the association with the outcome (Trait 2). This is often why it is recommended that the direct, indirect and total effects be used in the interpretation rather than the proportion mediated.

Summary table of pleiotropy results

SNP	Beta (Trait 1)	P (Trait 1)	Beta (Trait 2)	P (Trait 2)	MV[1] P	MV[1] Loading (Trait 1)	MV[1] Loading (Trait 2)	ADE	ADE (P)	ACME	ACME (P)	Total Effect	Total Effect (P)	Prop Mediated	Prop Mediated (P)
rs125	0.072	1.08E-08	0.062	6.45E-07	1.80E-10	0.8397	0.7096	0.045	<2e-16	0.015	<2e-16	0.0596	<2e-16	0.2516	<2e-16
rs135	-0.040	1.41E-06	0.050	1.47E-09	2.82E-17	-0.5457	0.7024	0.059	<2e-16	-0.008	<2e-16	0.0509	<2e-16	-0.3317	<2e-16
rs139	-0.065	2.77E-15	0.090	2.86E-28	2.26E-50	-0.5249	0.7196	0.103	<2e-16	-0.014	<2e-16	0.0891	<2e-16	-0.1580	<2e-16
rs296	-0.056	1.15E-10	-0.051	2.60E-09	1.78E-14	0.8100	0.7456	-0.039	<2e-16	-0.012	<2e-16	-0.0512	<2e-16	0.2306	<2e-16
rs300	-0.046	1.85E-08	-0.039	2.09E-06	1.34E-10	-0.8386	-0.7110	-0.029	<2e-16	-0.010	<2e-16	-0.0392	<2e-16	0.2507	<2e-16
rs921	0.109	6.29E-23	0.057	1.95E-07	1.16E-24	-0.9475	-0.5144	0.036	<2e-16	0.023	<2e-16	0.0595	<2e-16	0.3908	<2e-16
rs923	0.041	1.48E-06	0.042	4.04E-07	2.35E-09	0.7615	0.7957	0.034	0.002	0.009	<2e-16	0.0421	<2e-16	0.2035	<2e-16
rs1138	-0.050	9.58E-10	0.048	2.90E-09	2.44E-20	0.6511	-0.6027	0.058	<2e-16	-0.011	<2e-16	0.0468	<2e-16	-0.2319	<2e-16
rs1166	-0.051	4.77E-10	0.041	6.30E-07	1.02E-17	-0.7175	0.5275	0.049	<2e-16	-0.011	<2e-16	0.0382	<2e-16	-0.2918	<2e-16
rs1361	-0.056	2.65E-07	0.076	1.68E-12	3.52E-21	-0.5349	0.7114	0.087	<2e-16	-0.012	<2e-16	0.0751	<2e-16	-0.1614	<2e-16
rs1448	-0.079	3.46E-08	0.092	1.51E-11	2.21E-20	-0.5847	0.6679	0.108	<2e-16	-0.017	<2e-16	0.0908	<2e-16	-0.1879	<2e-16

[1] Multivariate

Advanced Gene Mapping Course: Mendelian Randomization

Exercise Andrew DeWan, PhD, MPH

This exercise is designed to give you practical experience conducting a two-sample Mendelian randomization study using the accompanying R code to run TwoSampleMR.

There are several online sources to search for summary statistics including: https://www.mrbase.org/ and https://gwas.mrcieu.ac.uk/

Part I:

You will be conducting an analysis to investigate the causal relationship between low density lipoprotein (LDL) and coronary heart disease (CHD) based on summary statistics from previously published GWAS data.

Exposure: Fasting LDL measurements from in 173,082 subjects and 2,437,752 genetic variants. Subjects are of European, East and South Asian and African ancestry.

Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013 Nov;45(11):1274-1283. doi: 10.1038/ng.2797. Epub 2013 Oct 6. PMID: 24097068; PMCID: PMC3838666. GWAS ID: ieu-a-300

Outcome: CHD (e.g. myocardial infarction (MI), acute coronary syndrome, chronic stable angina, or coronary stenosis >50%) in 184,305 subjects (60,801 cases and 123,504 controls) and 9,455,779 genetic variants. Subjects are of European, East and South Asian, Hispanic and African ancestry.

Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015 Oct;47(10):1121-1130. doi: 10.1038/ng.3396. Epub 2015 Sep 7. PMID: 26343387; PMCID: PMC4589895. GWAS ID: ieu-a-7

1) 1) Conduct an MR analysis of LDL and CHD. Please note the following:

   A. For the exposure for this publication, you will be using the larger set of subjects for this first analysis (N=173,082)
   B. For the exposure, you're using a p-value threshold of 5e10-8, LD Rsq = 0.001 and clumping distance of 10000kb. Also make sure "Perform Clumping" is checked if you're using MR-base
   C. For the outcome for this publication, you're using the trait denoted "Coronary heart disease"
   D. We have selected the options: allowing LD proxies to be selected for the outcome using a minimum Rsq of 0.8; allowing for palindromic SNPs with a

MAF threshold of 0.3; “Allele harmonization” to “Attempt to align strands for palindromic SNPs.” In TwoSampleMR, LD proxy parameters are set in the extract_outcome_data function and the allele harmonization option is set in the harmonise_data function with “action = 2”.

E. You are running the following methods:

a. Inverse variance weighted (NOTE: this is a random effects model)
b. MR Egger
c. Weighted Median

Questions:

1. How many variants are included in your genetic instrument for the exposure and how many are included in the outcome analysis? Of these, how many are proxies?

2. Based on the descriptions above, is the study used to define the IV appropriate for the outcome population?

3. Is there evidence of an association between LDL and CHD?

4. Is there evidence of heterogeneity in the genetic effects?

5. Is there evidence of pleiotropy?

6. How would you interpret the results of the three analyses together (i.e. IVW, MR Egger and Weighted Median)?

2) Re-run the analysis but for myocardial infarction (MI) using outcome data from the same publication (ieu-a-798).

Questions:

1. Is there evidence of an association between LDL and MI?

2. Can the association between LDL and CHD be explained by MI?

3) Feel free to explore associations with additional exposures such as HDL, BMI (you can use the Yengo et al. SNPs) or other exposures/outcomes of interest to you.

(NOTE: The code to run the analyses in Parts II and III are commented our in the R code due to the excessive run times. However, the answers to this section are provided in the answer key.)

Part II:

Let's now see if we can validate the finding of an association between LDL and CHD by using different exposure data source and potentially dissect this signal to see if we pinpoint the features of LDL that might be driving this signal. We will use metabolomics data that was generated in a sample of 24,925 individuals.

Kettunen, J., Demirkan, A., Würtz, P. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of *LPA*. *Nat Commun* **7,** 11122 (2016). https://doi.org/10.1038/ncomms11122.

Exposures: LDL.C, LDL.D, S.LDL.C, S.LDL.L, S.LDL.P, M.LDL.C, M.LDL.CE, M.LDL.L, M.LDL.P, M.LDL.PL, L.LDL.C, L.LDL.CE, L.LDL.FC, L.LDL.L, L.LDL.P, L.LDL.PL (16 metabolites)

Where S. = small, M. = medium and L. = large; .C = total cholesterol, .D = diameter, .L = total lipids, .P = concentration, .CE = cholesterol esters, .PL = phospholipids, .FC = free cholesterol

In TwoSampleMR these are brought in using data(metab_qtls) and then specifying the specific LDL variables above.

You are using the same CHD outcome as you did for Part I (Nikpay PMID: 26343387, GWAS ID: ieu-a-7) using the full set of cases and controls (N=184,305).

1) Conduct an MR analysis as you did previously (NOTE: the TwoSampleMR code only specifies the IVW test for ease of summarizing the results but feel free to add the MR Egger and Weighted Median tests).

   Questions:

   1. For LDL.C, does the association between LDL and CHD validate the previous findings?

   2. Considering all the associations, do these results differentiate between the different characteristics of LDL (Please note: you will want to take into account the 16 association tests)?

3. What might be one explanation for the similarity between results for the different LDL characteristics?

4. Are there any concerns about heterogeneity or pleiotropy?

Part III:

Let's now look at the associations with VLDL metabolites using the same exposure and outcome data sources.

Exposures: 33 VLDL metabolites (Please note additional abbreviations: XS. = very small, XL. = very large, XXL. = extremely large; .TG = triglycerides)

1) Conduct an MR analysis as you did previously (NOTE: the TwoSampleMR code only specifies the IVW test for ease of summarizing the results but feel free to add the MR Egger and Weighted Median tests).

   Question:

   1. Considering all of the associations, are there any obvious trends in the results (Please note: you will want to take into account the 33 association tests)?

Answers

Questions:

1. How many variants are included in your genetic instrument for the exposure and how many are included in the outcome analysis? Of these, how many are proxies?

   There are 79 variants that surpass the p<5e-8 threshold for LDL in the exposure GWAS. Of these 77 are identified in the CHD outcome GWAS, 1 of which is a proxy.

2. Based on the descriptions above, is the study used to define the IV appropriate for the outcome population?

   They are generally well matched in terms of the population ancestries in the two studies, however, the outcome GWAS has subjects of Hispanic ancestry which could be a minor issue. This would be something to mention in the Discussion section of a manuscript. There is a subset of only European subjects for LDL but not for CHD, however, if you had access to the original data you could subset the subjects by ancestry to better match the exposure and outcome groups.

3. Is there evidence of an association between LDL and CHD?

   Yes, the IVW yields a beta = 0.4114 (p=1.626e-15) which corresponds to an OR of 1.51 (95% CI: 1.36 – 1.67) per SD increase in LDL.

4. Is there evidence of heterogeneity in the genetic effects?

   Yes, there is significant heterogeneity across effects of each SNP on CHD (p=2.822e-40) indicating that the random effects model is appropriate.

5. Is there evidence of pleiotropy?

   From the MR Egger regression there is no significant evidence of pleiotropy as the regression intercept is not significantly different from zero (p=0.118).

6. How would you interpret the results of the three analyses together (i.e. IVW, MR Egger and Weighted Median)?

   The IVW method (OR = 1.51, 95% CI: 1.36 – 1.67, p=1.626e-16), MR Egger (OR = 1.66, 95% CI: 1.42 – 1.93, p=1.086e-8) and Weighed Median (OR = 1.49, 95% CI: 1.36 – 1.63, p=1.962e-19) are relatively consistent meaning the causal effect estimate is likely to be between 1.49 and 1.66. There is no evidence that this estimate is influence by horizontal pleiotropy as the MR Egger intercept is not significant.

Questions:

1. Is there evidence of an association between LDL and MI?

Yes, IVW method provides significant evidence of an association between LDL and MI (OR = 1.48, 95% CI: 1.33 – 1.66, p=1.42e-12). The other MR measures of association are consistent with this estimate and there is again no evidence of horizontal pleiotropy.

2. Can the association between LDL and CHD be explained by MI?

We would need to test the other traits included in the CHD definition to see if they were associated with LDL or not and test for heterogeneity of the effects. However it is reassuring that the effect estimates are consistent between the larger CHD group and the smaller subgroup of subjects with MI.

3 Feel free to explore associations with additional exposures such as HDL, BM (you can use the Yengo et al. SNPs) or other exposures/outcomes of interest to you.

I'm more than happy to discuss additional results one-on-one or when we discuss the answers to this exercise.

Part II:

Questions:

1. For LDL.C, does the association between LDL and CHD validate the previous findings?

Yes, although the magnitude of the effect is slightly attenuated. The IVW yields a beta = 0.3665 (p=1.19e-07) which corresponds to an OR of 1.44. The previous OR estimate was 1.51 and the 95% CIs overlap.

2. Considering all the associations, do these results differentiate between the different characteristics of LDL (Please note: you will want to take into account the 16 association tests)?

All the associations yield statistically significant results, except for LDL diameter which doesn't meet the corrected significance threshold (p<0.003125). The remaining metabolites have statistically significant betas ranging from 0.3665 (LDL diameter) to 0.4709 (concentration of small LDL).

3. What might be one explanation for the similarity between results for the different LDL characteristics?

There is a high degree of overlap between the variants contained in the instrument variable for each of the metabolites.

4. Are there any concerns about heterogeneity or pleiotropy?

There is significant heterogeneity across all the metabolites, except for the diameter of LDL. There is some evidence of pleiotropy among the large LDL metabolites, but not among any of the others.

Part III:

Question:

1. Considering all the associations, are there any obvious trends in the results (Please note: you will want to consider the 33 association tests)?

The results tend to be significant ($p<0.00015$) for the small and very small VLDL metabolites (except for the small VLDL triglycerides, $p=0.00019$), whereas the large, very large and extremely large LDL metabolites are non-significant (except for the cholesterol esters in large VLDL).