Assignment 2

Maria Izabel Cavassim Alves

12/1/2021

# Assignment 2:

## Association testing using PLINK

Software: We will be using plink 1.9 which can be downloaded here: <https://www.cog-genomics.org/plink/1.9/>

We will also be using R and Rstudio to make plots and make simple calculations. R can be downloaded here: <http://mirrors.dotsrc.org/cran/>

Rstudio here: <https://www.rstudio.com/products/rstudio/download/>

# Data:

We will use a simulated date set of 47 cases and 41 controls. This data is found [here](https://www.dropbox.com/home/DataQCBio/Assignment2).

The following exercise uses the software PLINK and R. R is used to analize and visualize the results from PLINK. The primary data consists of 3 files:

1. [file].bed: contains the genotype information in binary format.
2. [file].bim: contains the chromosome, position and alleles of all the SNPs in the data set.
3. [file].fam: contains father ID, mother ID, gender and phenotype for each sample.

The other data are results of the analysis:

1. “gwa.assoc.fisher”: is the result of a simple association mapping analysis using allelic Fisher’s exact test.
2. “gwa.eigenval”: are the eigenvalues of the decomposition of the relationship matrix
3. “gwa.eigenvec”: are the eigenvectors of the decomposition of the relationship matrix
4. “plink.assoc.logistic”: is the logistic regression test of association while correcting for covariates.

# Exercise contents:

In this practical exercise, we will go through the steps of performing association tests using Plink while also adjusting for principle components (e.g. to correct for population structure).

Test for association with disease status using a Fisher’s exact test To test for association between SNPs and disease status using an allelic Fisher’s exact test, type the following command at the shell prompt:

# shell  
./plink --bfile gwa --fisher --out gwa

### Question 1

E2.1) Take a look at the output file “gwa.assoc.fisher”. What is the p-value and location of the most significant variant?

setwd('/Users/PM/Dropbox/Assignment1')  
  
results <- read.table('gwa.assoc.fisher', head=T)  
  
# look at the data frame by typing:  
View(results)  
# use the function min() to find the most significant gene

### Question 2

E2.2) Is the most significant variant significant if you do Bonferroni correction?

#### Make plots

We will use the R package “qqman” to make Manhattan plots and qq-plots. The package is available in CRAN so it can be installed using the “install.packages” commando. You can read in the association results and make a Manhattan plot by typing:

# install package  
install.packages("qqman")  
# load the package  
library(qqman)  
# plot  
manhattan(d)

### Question 3

E2.3) Are there other variants close to the most significant variant that are also associated with the disease?

# Position of the most sigficant SNP  
sig\_snp = results[which.min(results$P),]  
  
# Sorting by base pair  
sorted = results[with(results, order(CHR, BP)),]  
  
pos = as.numeric(which(sorted$BP == sig\_snp$BP))  
  
# SNPs around the significant one  
around = sorted[(pos-5):(pos+5),]  
  
# Using Manhattan plot to visualize the significants markers:  
manhattan(results)  
  
# Highlight SNPs of interest:  
manhattan(results, highlight = around$SNP)

### Question 4

E2.4) Is there a general inflation of the test statistic? Genomic Control. Hint: To make a QQ-plot you should use the “qq” function

# In order to look at the inflation of the test we are going to use Quantile-Quantile plot by looking at the distribution of the p-values  
  
qq(results$P)

The inflation factor, 𝝺 , can be calculated as the median chi-square value divided by 0.456. Given a p-value, p, the corresponding Chi-square can be calculated as:

inflation\_factor = median(qchisq(results$P, df=1, lower.tail = F))/0.456  
inflation\_factor

### Question 5

E2.5) What is the inflation factor?

To do genomic control (to adjust for inflated test statistic) you divide the chi-square values with the inflation factor. To turn a chi-square value, q, into a p-value you use the “pchisq” function: pchisq(q, df=1, lower.tail = F)

qs = qchisq(results$P, df=1, lower.tail = F)/inflation\_factor  
ps = pchisq(qs, df=1, lower.tail = F)  
  
qq(ps)

### Question 6

E2.6) What is the p-value of the most significant marker after genomic control?

## Principal Component Analyses (PCA)

As we saw during the QC workshop, it is best to perform the PCA on a LD-pruned set of SNPs:

./plink --bfile gwa --indep-pairwise 500kb 5 0.2 --out gwa

To use the pruned set of SNPs to calculate the relationship matrix and calculate the first 20 principle components (PCs) type:

./plink --bfile gwa --extract gwa.prune.in --pca 20 --out gwa

This calculates the eigenvalues and the eigenvectors, and stores them in two files (gwa.eigenval, gwa.eigenvec).

# install packages:  
# install.packages("ggplot2")  
# install.packages("gridExtra")  
library(ggplot2)  
library(gridExtra)  
  
eigen\_values = read.table('gwa.eigenval')  
eigen\_vectors = read.table('gwa.eigenvec', row.names = 1)  
  
plot1 = qplot(eigen\_vectors$V3, eigen\_vectors$V4, xlab="eigenvector 1", ylab="eigenvector 2", main ='Population Structure' )  
plot2 = qplot(eigen\_vectors$V4, eigen\_vectors$V5, xlab="eigenvector 2", ylab="eigenvector 3", main ='Population Structure')  
grid.arrange(plot1, plot2, ncol=2)

### Question 7

E2.7) Load gwa.eigenvec into R and make a plot with the first PC on the x-axis and the second PC on the y-axis. Does it look like there is population structure in the data? How many populations do you find?

The eigenvalues divided by the number of individuals should correspond approximately to the variance explained by the eigenvectors.

### Question 8

E2.8) How large a percentage of the variance does the first PC approximately explain?

# Variance explained :  
var\_expl = eigen\_values/nrow(eigen\_vectors)  
plot(c(1:nrow(var\_expl)), var\_expl[,1], main = 'Variance explained', xlab = "PC's", ylab = "Variance explained", type = "b")

### Adjusting for PCs

We can use a logistic regression test to perform an association test while correcting for covariates. To include the first PC as a covariate type:

./plink --bfile gwa --logistic --covar gwa.eigenvec --covar-number 1

The resulting file “plink.assoc.logistic” contains p-values for both the SNPs and the covariates. To get the p-values for the SNPs should look at the rows with the value “ADD” in the “TEST” column. It is possible to include more PCs. To include the first x covariates you can write “–covar-number 1-x”.

### Question 9

E2.9) Create Manhattan plot and QQ-plot for the new results. Does the QQ-plot look better?

### Question 10

E2.10) What is the inflation factor now? What is the inflaction factor if you include the first 10 PCs?