Priest Lab Interview Homework-Guyu (Tracy) Qin

**Part 1. Data exploration**

Please answer the following:

a)  How many individuals there are?

Answer: There are 44102 individuals.

b)  How many markers there are?

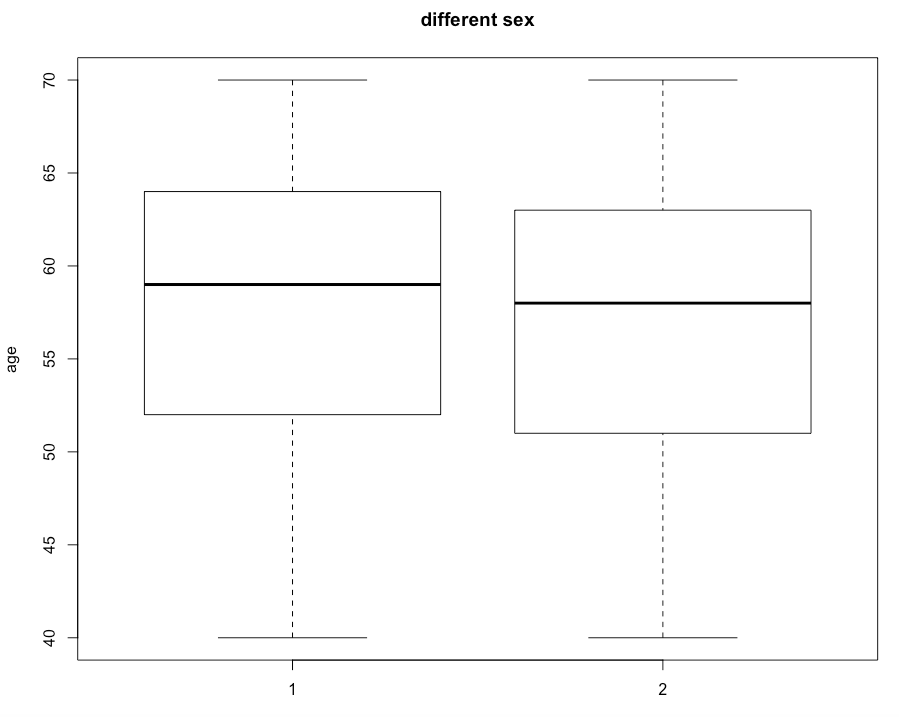
Answer: There are 81555 markers.

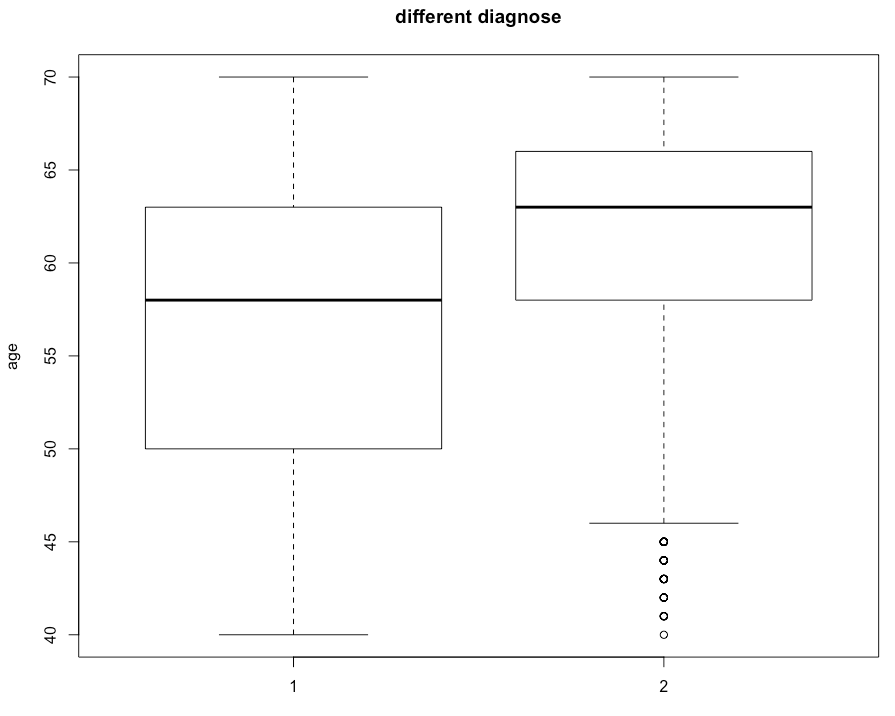
c)  How many people have a CAD diagnosis? How many healthy controls there are?

Answer: There are 3993 people have a CAD diagnosis, including 38802 healthy controls.

d)  Can you plot the distribution of ages, stratified by gender and diagnose? (You can make a stratified plot, or draw multiple plots, as necessary.)

Answer: For age, 1 represents male, 2 represents female. For different diagnose, 1 represents control, 2 represents case.





e)  Can you show 2x2 contingency tables of CAD vs. sex? Can you repeat for the procedure for the remaining categorical covariates (CAD vs. HLD, CAD vs. HTN, CAD vs. T1D, CAD vs. T2D and CAD vs. smoking)? For each of those tables, run a test and determine whether there is a statistically significant relationship between pairs of variables.

Answer:

CAD vs. Sex

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| CAD Control | 18492 | 20310 |
| CAD Case | 2741 | 1252 |

The p-value is smaller than 2.2e-16. There is significant relationship between sex and CAD diagnoses.

CAD vs. HLD

|  |  |  |
| --- | --- | --- |
|  | HLD Control | HLD Case |
| CAD Control | 32526 | 6276 |
| CAD Case | 1283 | 2710 |

The p-value is smaller than 2.2e-16. There is significant relationship between HLD and CAD diagnoses.

CAD vs. HTN

|  |  |  |
| --- | --- | --- |
|  | HTN Control | HTN Case |
| CAD Control | 17216 | 21586 |
| CAD Case | 648 | 3345 |

The p-value is smaller than 2.2e-16. There is significant relationship between HTN and CAD diagnoses.

CAD vs. T1D

|  |  |  |
| --- | --- | --- |
|  | T1D Control | T1D Case |
| CAD Control | 38767 | 35 |
| CAD Case | 3990 | 3 |

The p-value is 0.7551. There is no significant relationship between T1D and CAD diagnoses.

CAD vs. T2D

|  |  |  |
| --- | --- | --- |
|  | T2D Control | T2D Case |
| CAD Control | 37760 | 1042 |
| CAD Case | 3628 | 365 |

The p-value is smaller than 2.2e-16. There is significant relationship between T2D and CAD diagnoses.

CAD vs. smoking

|  |  |  |
| --- | --- | --- |
|  | Smoking Control | Smoking Case |
| CAD Control | 20451 | 18351 |
| CAD Case | 1423 | 2570 |

The p-value is smaller than 2.2e-16. There is significant relationship between smoking and CAD diagnoses.

f)  Can you tell us something else you found while exploring the data?

Answer: PCA analysis have been done and PC1 and PC2 have been given in the .cov file. If set CAD as outcome for a GWAS, the PC1, PC2, sex, HLD, HTN, T2D and smoking are potential covariates.

R code：

##Priest Lab Interview Homework

##Part1. Data exploration

## import the anhwdata.bim

setwd("~/Desktop")

anhwdata <- read.delim("~/Desktop/GWAS homework /anhwdata.bim", header=FALSE, na.strings = "-9", sep="")

anhwdata <- data.frame(anhwdata)

## import the anhwdata.cov

anhwdata1 <- read.delim("~/Desktop/GWAS homework /anhwdata.cov", na.strings = "-9", sep="")

anhwdata1 <- data.frame(anhwdata1)

## check the dataset

str(anhwdata1)

##get all of the levels in the dataset

levels(as.factor(anhwdata1$age\_recruit))

## import the anhwdata.fam

anhwdata2 <- read.delim("~/Desktop/GWAS homework /anhwdata.fam", header=FALSE, na.strings = "-9", sep="")

anhwdata2 <- data.frame(anhwdata2)

## add the column name

colnames(anhwdata2) <- c("FID","IID","PID","MID","sex\_2male.l2","P")

str(anhwdata2)

##checking if the name is the same to each other

anhwdata2$IID %in% anhwdata1$IID

##get all of the levels in the dataset

levels(as.factor(anhwdata2$P))

##count the total number of each phenotype

library(plyr)

count(anhwdata2$P)

## combine two data sets to one data

FAMCOV <- plyr::join(data.frame(anhwdata2),data.frame(anhwdata1),by="IID")

str(FAMCOV)

FAMCOV <- FAMCOV[,c(2,1,3:17)]

##plot the distribution of ages, stratified by gender and diagnose

head(FAMCOV)

boxplot(FAMCOV$age\_recruit~FAMCOV$sex\_2male.l2,ylab='age',main='different sex')

boxplot(FAMCOV$age\_recruit~FAMCOV$P,ylab='age',main='different diagnose')

##2x2 contigency tables

library(MASS)

View(FAMCOV)

table(FAMCOV$P,FAMCOV$sex\_2male.l2)

fisher.test(FAMCOV$P, FAMCOV$sex\_2male.l2)

table(FAMCOV$P,FAMCOV$HLDprev.l2)

fisher.test(FAMCOV$P, FAMCOV$HLDprev.l2)

table(FAMCOV$P,FAMCOV$HTNprev.l2)

fisher.test(FAMCOV$P, FAMCOV$HTNprev.l2)

table(FAMCOV$P,FAMCOV$T1Dprev.l2)

fisher.test(FAMCOV$P, FAMCOV$T1Dprev.l2)

table(FAMCOV$P,FAMCOV$T2Dprev.l2)

fisher.test(FAMCOV$P, FAMCOV$T2Dprev.l2)

table(FAMCOV$P,FAMCOV$binSmok.l2)

fisher.test(FAMCOV$P, FAMCOV$binSmok.l2)

**Part 2. Genetic association analysis**

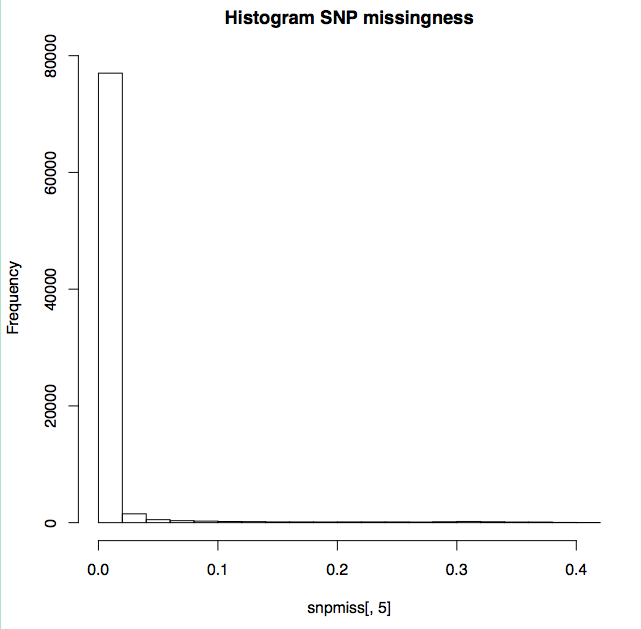
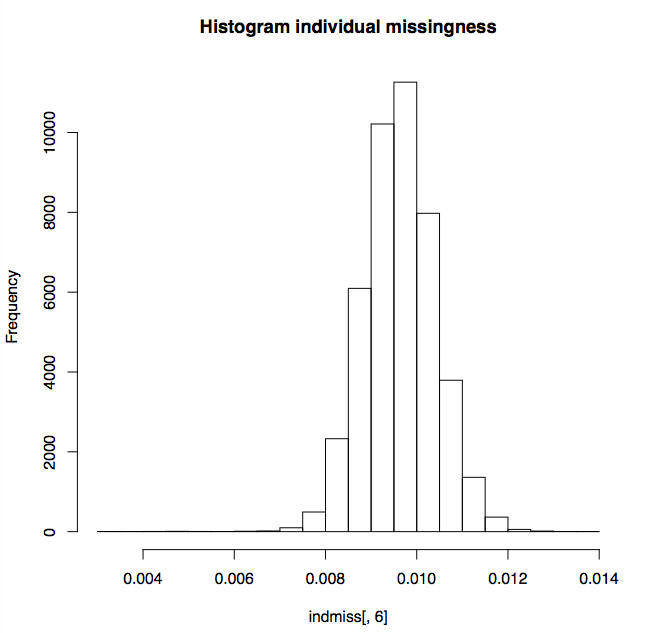
For this part, you will likely need to use Plink (or e.g., snpStats in R). Feel free to use any tool(s) you want.

a)  Can you make a histogram showing the distribution of missing values per individual? And a histogram with number of missing values per marker?

Answer:

Figure A indicated the histogram showing the distribution of missing values per individual.

Figure B indicated the histogram with number of missing values per marker.



A

B

b)  In the dataset, some individuals come from the same families. Find which individuals come from the same family, and determine which subjects need to be excluded. (Hint: you can estimate identity-by-descent and filter subjects based on a pi\_hat threshold.)

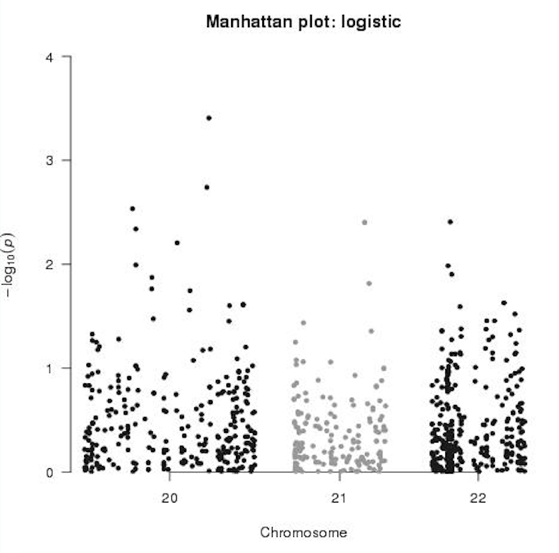
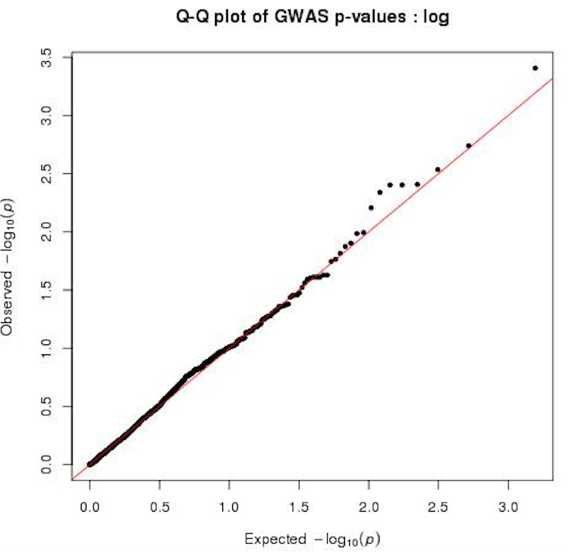
Answer: The subjects need to be excluded were recorded in the het-fail-qc.txt file.

c)  Run a genetic association test (a.k.a. GWAS) using CAD as outcome. Justify which covariates you include in the analysis (if any), and apply filtering criteria to samples and markers (e.g., missingness, Hardy-Weinberg equilibrium).

Answer: I included 7 covariates, PC1,PC2, sex, HLD, HTN, T2D and smoking. I applied geno 0.2, mind 0.2, maf 0.05, pihat min0.2 and hwe 1e-10 filtering criteria to samples and markers.

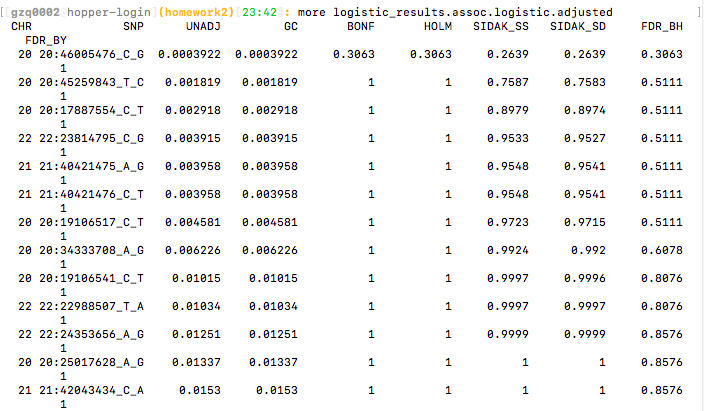
d)  Make manhattan and qq plots for your results. Is there genomic inflation? What else can you tell us from these plots?

Amswer: The common genome-wide significance threshold of 5.0e -8 was used. I am not sure about the genomic inflation. I used Bonferroni corrected the multiple testing *p*-value, along with FDR.



e)  What are the top signals that you see? Does any of them meet genome-wide significance criteria? Is there any limitation or problem with the top signals?

Answer: The top signals are as follows. But none of them meet the genome-wide significance threshold of 5.0e -8. The top two signals at chromosome 20 could be false positive SNPs.



f)  Think about limitations of the current analysis and results and tell us how, in an ideal world, you would improve this GWAS (e.g., different study design conditions, additional covariates, etc.).

Answer: In this study, 80K SNP markers were utilized for the analysis. Ideally, the density of the SNP markers should be higher, the sex, BMI, age, HLD, HTN, T2D and smoking should be added as covariates into genome association analysis one by one or in different combination to find potential markers related to CAD.

**Part 3. Writing: summary of results**

Imagine you need to write an abstract from the analysis you just made. Please provide a summary of results based on sections 1 and 2 above. There is no strong restriction on the length, but ideally it should be ~1⁄2 page, and definitely no more than 1 page (no figures).

Abstract:

Coronary artery disease (CAD) is the most common of cardiovascular diseases, which is also the most common cause of death globally. In this study, 44102 individuals (22001 males and 22101 females) and 81554 markers were utilized to conduct a standard genome-wide association study using PLINK and R to detect the markers associated with CAD diagnosis. Three files of PLINK binary data and one file of covariates were applied. The five steps of quality control were conducted, including filtering the missingness, minor allele frequency, Hardy-Weinberg equilibrium, heterozygosity and relatedness. The population structure was assessed by the principle component analysis (PCA) with the nucleotide polymorphism (SNP) markers. A logistic regression analysis including seven covariates were performed to test the association for the CAD binary trait. Bonferroni and false discovery rate (FDR) were applied for the multiple testing correction. The Manhattan plot didn’t show any markers which display significant associations with CAD based on the genome-wide significance threshold of 5.0e -8. It may because of the low density of SNP markers. In the future study, a higher density of markers should be applied, predicted gene annotation and function examination should be performed, and a pathway analysis should be conducted.

**Part 4. Creative work sample**

For this (optional) section, it would be great if you could share a sample of previous creative work -- preferably not an academic paper. Examples include: a GitHub repository, a piece of writing, something you built, a presentation, or just something you would like to show.

Answer:

1. I would like to share one report of a longitudinal data analysis, which is often used in clinical trials.

2. I would also like to share one homework solution made by myself in the STAT7840 Multivariate Statistical Analysis course, which was given good comments and shared with other classmates as a key sample.