P8139\_HW4\_association\_study

Ruixi Li

2024-03-28

# Data preparation

# read in data  
fms = read.csv("FMS\_data.csv")  
  
attach(fms)  
# The attach() function in R is used to modify the R search path by making the variables in a data frame available in the R environment as if they were variables in the global environment. This can simplify code when you're working with data frames because it allows you to refer to columns in the data frame directly by their names without having to use the data frame name each time.

# (1) Choose candidates genes

NamesEsr1Snps = names(fms)[substr(names(fms),1,4)=="esr1"]  
NamesEsr1Snps

## [1] "esr1\_rs1801132" "esr1\_rs1042717" "esr1\_rs2228480" "esr1\_rs2077647"  
## [5] "esr1\_rs9340799" "esr1\_rs2234693"

# The genotype matrix can now be defined by selecting the columns of fms that  
# correspond to the esr1 SNP names:  
fmsEsr1 = fms[,is.element(names(fms),NamesEsr1Snps)]

Although I can go through all SNPs in this dataset to find susceptible loci, I just tailored the candidate gene list to show the process more efficiently. I found some evidence showing that Estrogen receptor 1 gene polymorphisms are associated with metabolic syndrome in postmenopausal women in China. So I think (<https://doi.org/10.1186/s12902-018-0289-4>) it’s reasonable to use esr1 SNPs as my candidate genes.

# (1) Choose two traits

I am interested in determining whether there is an association between any of the SNPs within the esr1 gene and an indicator of Homeostatic model assessment(HOMA) in the FAMuSS study. HOMA is a method for assessing beta-cell function and insulin resistance (IR) from basal (fasting) glucose and insulin or C-peptide concentrations. I want to use HOMA as a measurement of metabolic syndrome. In Spanish population the threshold value of HOMA-IR drops from 3.46 using 90th percentile criteria to 2.05 taking into account of MetS components. So, I used two operationalization of HOMA with one as continuous and the other is cateforical(0=“<=2.05”, 1=“>2.05”).

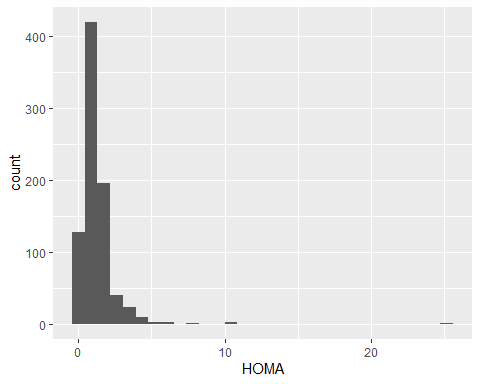
summary(HOMA)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.330 0.600 0.970 1.254 1.510 25.540 570

fms |> select(HOMA) |> ggplot (aes(x=HOMA))+ geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

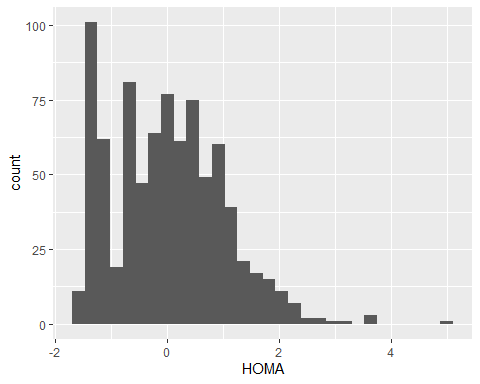
## Warning: Removed 570 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



fms\_log = fms |> mutate(HOMA=scale(log(HOMA)))  
fms\_log |> ggplot (aes(x=HOMA))+ geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 570 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



# We define our trait to be an indicator for whether HOMA > 2.05:  
Trait = as.numeric(HOMA>2.05)  
  
count = table(Trait)  
percentage = prop.table(count)  
cbind(count, percentage)

## count percentage  
## 0 728 0.8802902  
## 1 99 0.1197098

# EDA

# skimr::skim(fms)   
# I used this function to look at the datatype, missing stutas, duplicate and general distribution of all variables, including SNPs, outcomes and potential covariates. But to avoid too much output, I chose not to show it in word.

# (2) a/b : Association Analysis and multiple comparison

## Binary trait: Chi-squared test/Fisher’s Exact Test for association

# We write a function to record the p-values from applying the χ2-test to  
# the 2 × 3 contingency tables corresponding to each SNP and this trait, some cells of the frequency table < 5, so I used fisher's exact test rather than chi-square test:  
newFunction\_cate = function(Geno){  
 ObsTab = table(Trait,Geno)  
 return(fisher.test(ObsTab)$p.value)  
}  
  
# Apply this function to the columns of fmsEsr1:  
pvalues = apply(fmsEsr1,2,newFunction\_cate)  
pvalues

## esr1\_rs1801132 esr1\_rs1042717 esr1\_rs2228480 esr1\_rs2077647 esr1\_rs9340799   
## 0.5248122 0.8370360 0.9373894 0.4959759 0.5249899   
## esr1\_rs2234693   
## 0.8070826

# adjust for multiple testing  
p.adj = p.adjust(pvalues, method="BH")  
p.adj

## esr1\_rs1801132 esr1\_rs1042717 esr1\_rs2228480 esr1\_rs2077647 esr1\_rs9340799   
## 0.9373894 0.9373894 0.9373894 0.9373894 0.9373894   
## esr1\_rs2234693   
## 0.9373894

## Continuous trait: HOMA

# ANOVA  
newFunction\_cont = function(Geno){  
 anova\_result <- aov(HOMA ~ Geno, data = fmsEsr1)  
 p\_value <- summary(anova\_result)[[1]]["Geno", "Pr(>F)"]  
}  
  
# Apply this function to the columns of fmsEsr1:  
pvalues\_cont = apply(fmsEsr1,2,newFunction\_cont)  
pvalues\_cont

## esr1\_rs1801132 esr1\_rs1042717 esr1\_rs2228480 esr1\_rs2077647 esr1\_rs9340799   
## 0.3650536 0.8591891 0.4069314 0.4061698 0.4513844   
## esr1\_rs2234693   
## 0.5729941

# adjust for multiple testing  
p.adj\_cont = p.adjust(pvalues\_cont, method="BH")  
p.adj\_cont

## esr1\_rs1801132 esr1\_rs1042717 esr1\_rs2228480 esr1\_rs2077647 esr1\_rs9340799   
## 0.6770765 0.8591891 0.6770765 0.6770765 0.6770765   
## esr1\_rs2234693   
## 0.6875929

#Linear Regression  
newFunction\_cont = function(Geno){  
 lm\_model <- lm(HOMA ~ as.factor(Geno), data = fmsEsr1)  
 p\_value <- summary(lm\_model)$coefficients[2, 4] # Extracting the p-value for the Geno coefficient  
 return(p\_value)  
}  
  
# Apply this function to the columns of fmsEsr1:  
pvalues\_cont\_lm = apply(fmsEsr1,2,newFunction\_cont)  
pvalues\_cont\_lm

## esr1\_rs1801132 esr1\_rs1042717 esr1\_rs2228480 esr1\_rs2077647 esr1\_rs9340799   
## 0.1582841 0.7038117 0.9778607 0.2622762 0.2954338   
## esr1\_rs2234693   
## 0.2927388

# adjust for multiple testing  
p.adj\_cont\_lm = p.adjust(pvalues\_cont\_lm, method="BH")  
p.adj\_cont\_lm

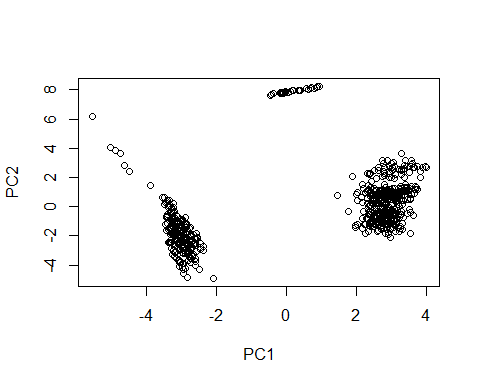
## esr1\_rs1801132 esr1\_rs1042717 esr1\_rs2228480 esr1\_rs2077647 esr1\_rs9340799   
## 0.4431507 0.8445741 0.9778607 0.4431507 0.4431507   
## esr1\_rs2234693   
## 0.4431507

# (2) c: Investigate population stratification using Principal Component Analysis (PCA)

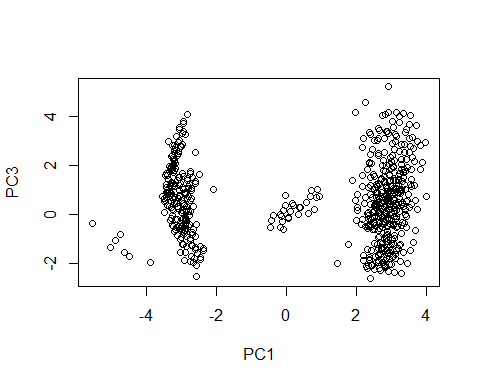
# Examine population substructure using all 24 SNPs within the akt1 gene.  
NamesAkt1Snps = names(fms)[substr(names(fms),1,4)=="akt1"]  
NamesAkt1Snps

## [1] "akt1\_t22932c" "akt1\_g15129a" "akt1\_g14803t"   
## [4] "akt1\_c10744t\_c12886t" "akt1\_t10726c\_t12868c" "akt1\_t10598a\_t12740a"  
## [7] "akt1\_c9756a\_c11898t" "akt1\_t8407g" "akt1\_a7699g"   
## [10] "akt1\_c6148t\_c8290t" "akt1\_c6024t\_c8166t" "akt1\_c5854t\_c7996t"   
## [13] "akt1\_c832g\_c3359g" "akt1\_g288c" "akt1\_g1780a\_g363a"   
## [16] "akt1\_g2347t\_g205t" "akt1\_g2375a\_g233a" "akt1\_g4362c"   
## [19] "akt1\_c15676t" "akt1\_a15756t" "akt1\_g20703a"   
## [22] "akt1\_g22187a" "akt1\_a22889g" "akt1\_g23477a"

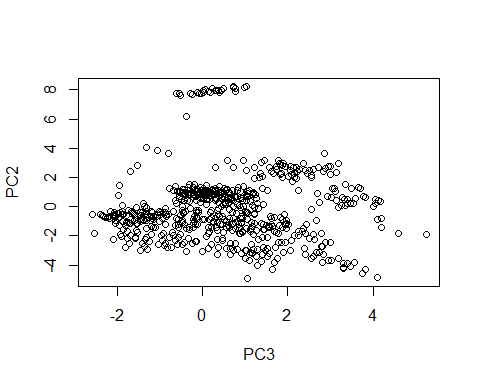
# Convert the genotype data from factor variables to numeric variables using data.matrix()  
# Note that we additionally assign the missing data a number  
  
FMSgeno = fms[,is.element(names(fms),NamesAkt1Snps)]  
FMSgenoNum = data.matrix(FMSgeno)  
FMSgenoNum[is.na(FMSgenoNum)] = 4  
  
PC.FMS = prcomp(FMSgenoNum,retx=TRUE, center=TRUE, scale=TRUE)  
plot(PC.FMS$"x"[,1],PC.FMS$"x"[,2],xlab="PC1",ylab="PC2")



plot(PC.FMS$"x"[,1],PC.FMS$"x"[,3],xlab="PC1",ylab="PC3")



plot(PC.FMS$"x"[,3],PC.FMS$"x"[,2],xlab="PC3",ylab="PC2")



# (2) d: Summary

I am interested in determining whether there is an association between any of the SNPs within the 6 esr1 gene and an indicator of Homeostatic model assessment (HOMA) in the FAMuSS study. HOMA is a method for assessing beta-cell function and insulin resistance (IR) from basal (fasting) glucose and insulin or C-peptide concentrations. I want to use HOMA as a measurement of metabolic syndrome. In Spanish population the threshold value of HOMA-IR drops from 3.46 using 90th percentile criteria to 2.05 taking into account of MetS components. So, I used two operationalizations of HOMA with one as continuous (log-transformed and z-score scaled to make it normalized) and the other is categorical (0="<=2.05", 1=">2.05"). HOMA-indicated susceptible population to metabolic syndrome accounted for 11.97% whereas the non-susceptible population accounted for 88.03%. All SNPs were categorical variables with three unique values. Since I didn’t assume any genetic models(additive/dominant/recessive), the SNPs were kept with three levels. Therefore, I used one-way ANOVA for the association analysis for continuous trait. For binary trait, fisher’s exact test was applied due to small cell frequency (<5). No association was found between SNPs among the 6 “ers1” SNPs and HOMA (both continuous and binary) after BH adjustment. PCA was conducted using all 24 SNPs within the akt1 gene to investigate population stratification in this association study. There is a clear separation along PC1, indicating that it captures a significant amount of the genetic variation that differentiates between two main groups. In contrast, I didn’t see clear pattern in PC2 and PC3. For further analysis, I might consider using these components as covariates in genetic association models to correct for population stratification.