Bioinformatics and Statistical Genetics: Linkage Disequilibrium, Haplotype estimation

Sara Montese

```
library(genetics)
library(HardyWeinberg)
library(data.table)
library(haplo.stats)
library(tidyr)
```

1

1. Load the FOXP2.dat file into the R environment.

```
setwd("/Users/sara_/Desktop/ERASMUS/BSG/bio_lab/bio-lab-9/update")
filename <- "FOXP2.dat"
raw_data <- read.table(filename, sep = " ",header = T)
raw_data_df <- data.frame(raw_data)


SNPdata <- raw_data_df[,2:ncol(raw_data_df)]

n <- nrow(SNPdata)# individuals
p <- ncol(SNPdata) #SNPs
missing <- 100*sum(is.na(SNPdata))/(n*p)

cat("How many individuals and how many SNPs are there in the database?\n")

## How many individuals and how many SNPs are there in the database?

cat("There are ",n,"variants and ", p, "SNPs.\n")

## There are  104 variants and  543 SNPs.

cat("What percentage of the data is missing?\n")

## What percentage of the data is missing?

cat(missing)

## 0
```

2

2. Using the function LD from the genetics package, compute the LD statistic D for the SNPs rs34684677 and rs2894715 of the database. Is there significant association between the alleles of these two SNPs?

```
SNP1 <- genotype(SNPdata[,"rs34684677"], sep="/")
SNP2 <-genotype(SNPdata[,"rs2894715"], sep="/")
```

```
disequilibrium <- LD(SNP1,SNP2)
print(disequilibrium)
```

```
##
## Pairwise LD
## -----------
##                       D        D'       Corr
## Estimates: -0.05493703 0.9986536 -0.3144048
##
##                X^2     P-value   N
## LD Test: 20.56088 5.77645e-06 104
```

```
cat("D statistics:")
```

```
## D statistics:
```

```
D<-disequilibrium$D
cat(D)
```

```
## -0.05493703
```

A coefficient of linkage disequilibrium of -0.05 suggests a weak negative association between alleles at the two loci. This means that the occurrence of one allele at one locus is associated with the occurrence of a different allele at the other locus more often than would be expected by chance. We reject the null hypothesis; there is significant association between these two SNPs because the p-value is very small (below our threshold of 5 %).

3

```
haplo_data <- data.frame(SNP1, SNP2)
haplotypes <- apply(haplo_data, 1, paste, collapse = " - ")

#  ount occurrences of each haplotype
haplotype_counts <- table(haplotypes)

# calculate haplotype frequencies
haplotype_frequencies <- prop.table(haplotype_counts)

print(haplotype_frequencies)
```

```
## haplotypes
##  G/G - G/G  G/G - T/G  G/G - T/T  G/T - T/G  G/T - T/T  T/T - T/T
## 0.11538462 0.35576923 0.23076923 0.08653846 0.18269231 0.02884615
```

```
# print the most common haplotype
most_common_haplotype <-which.max(haplotype_frequencies)

cat("Most common haplotype: ", names(haplotype_frequencies)[most_common_haplotype]
, "\n")
```

```
## Most common haplotype:  G/G - T/G
```

4

4.Determine the genotype counts for each SNP. For how many variants do you reject Hardy-Weinberg equilibrium using an ordinary chi-square test without continuity correction? Is this what you would expect by chance? (hint: you can read the .bim in R in order to determine the alleles of each SNP, and use function MakeCounts from the HardyWeinberg package to create a matrix of genotype counts).

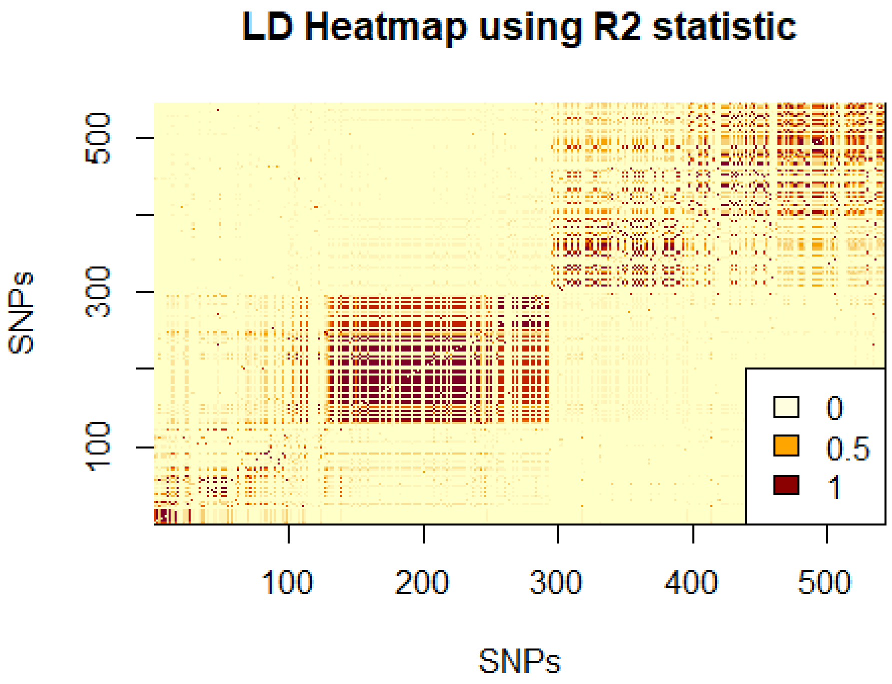
```
# read the .bim file
setwd("/Users/sara_/Desktop/ERASMUS/BSG/bio_lab/bio-lab-9/update")
bim_data <- read.table("FOXP2.bim", header = FALSE)

# extract the alleles
#alleles <- bim_data[,5:6]
alleles <-apply(bim_data[,5:6], 1, paste, collapse = "/")
alleles <- as.vector(alleles)

counts <- MakeCounts(SNPdata, alleles, sep="/")

p_values <- apply(counts[,1:3], 1, function(x) HWChisq(x,cc=0,verbose=TRUE)$pval)

# determine the number of variants that reject HWE
num_reject_hwe <- sum(p_values < 0.05)
cat("Number of rejections: ",num_reject_hwe)

## Number of rejections:  33

cat("\n")

cat("Percentage of rejections: ", num_reject_hwe*100/length(p_values), "%")

## Percentage of rejections:  6.077348 %
```

By setting the significance level to 0.05, we would expect 5% of the markers to show significant deviation from HWE by chance alone. Since the percentage of significant SNPs is more than 5%, we understand that there other factors, beyond chance.

5

Compute the LD for all the marker pairs in this data base, using the LD function of the packages genetics. Be prepared that this make take a few minutes. Extract the R2 statistics and make an LD heatmap (hint: you can use the command image) using the R2 statistic.

```
SNP_names <- colnames(SNPdata)
names<-list(SNP_names, SNP_names)
# matrix to store the disequilibrium results
r2_matrix <- matrix(0, nrow = ncol(SNPdata), ncol = ncol(SNPdata), dimnames = name
s)
# loop over each pair of SNPs
for (i in 1:(length(SNP_names)-1)) {
  for (j in (i+1):length(SNP_names)) {
    # genotypes for the pair of SNPs
    SNP1 <- genotype(SNPdata[,SNP_names[i]], sep="/")
    SNP2 <- genotype(SNPdata[,SNP_names[j]], sep="/")

    disequilibrium <- LD(SNP1, SNP2)

    r2_matrix[SNP_names[i],SNP_names[j]]<-disequilibrium$`R^2`
    r2_matrix[SNP_names[j],SNP_names[i]]<-disequilibrium$`R^2`

  }
}

# create the heatmap
#heatmap(r2_matrix, t(r2_matrix),xlab = "SNPs", ylab = "SNPs")
#title(main="LD Heatmap using R2 statistic", font.main=2)

#or also, using image
col <- c("light yellow", "orange","dark red") # set the color palette
image(1:ncol(r2_matrix), 1:nrow(r2_matrix), t(r2_matrix),xlab = "SNPs", ylab = "SN
Ps")
title(main="LD Heatmap using R2 statistic", font.main=2)

# add  legend
legend("bottomright", legend = c("0", "0.5", "1"), fill = col)
```

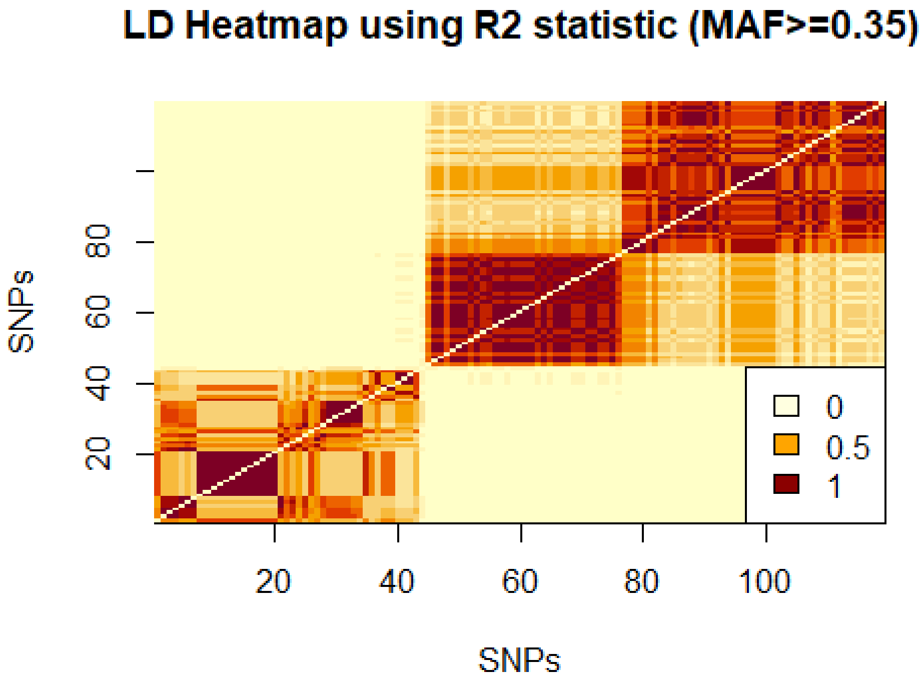


6

Make another heatmap obtained by filtering out all variants with a MAF below 0.35, and redoing the computations to obtain the R2 statistics in R. Can you explain any differences observed between the two heatmaps?

```r
# function to calculate Minor Allele Frequency (MAF) for a single marker
calculate_maf <- function(marker) {
  sum <- summary(genotype(marker))#, alleles = c("A", "C", "G", "T")))
  maf <- min(sum$allele.freq, na.rm = TRUE)
  return(maf)
}


# function to filter variants based on MAF
filter_by_maf <- function(data, maf_threshold) {
  maf_values <- sapply(data, calculate_maf)
  selected_columns <- maf_values >= maf_threshold
  filtered_data <- data[, selected_columns, drop = FALSE]
  return(filtered_data)
}

#MAF threshold
maf_threshold <- 0.35

# filter variants
filtered_SNPdata <- filter_by_maf(SNPdata, maf_threshold)

# print the filtered dataset
#print(filtered_SNPdata)

SNP_names <- colnames(filtered_SNPdata)
names<-list(SNP_names, SNP_names)

# matrix to store the disequilibrium results
r2_matrix_filtered <- matrix(0, nrow = ncol(filtered_SNPdata), ncol = ncol(filtere
d_SNPdata), dimnames = names)
# loop over each pair of SNPs
for (i in 1:(length(SNP_names)-1)) {
  for (j in (i+1):length(SNP_names)) {
    # Get the genotypes for the pair of SNPs
    SNP1 <- genotype(filtered_SNPdata[,SNP_names[i]], sep="/")
    SNP2 <- genotype(filtered_SNPdata[,SNP_names[j]], sep="/")

    disequilibrium <- LD(SNP1, SNP2)

    r2_matrix_filtered[SNP_names[i],SNP_names[j]]<-disequilibrium$`R^2`
    r2_matrix_filtered[SNP_names[j],SNP_names[i]]<-disequilibrium$`R^2`

  }
}
```

```
image(1:ncol(r2_matrix_filtered), 1:nrow(r2_matrix_filtered), t(r2_matrix_filtered
),xlab = "SNPs", ylab = "SNPs")

title(main="LD Heatmap using R2 statistic (MAF>=0.35)", font.main=2)
legend("bottomright", legend = c("0", "0.5", "1"), fill = col)
```



Can you explain any differences observed between the two heatmaps?

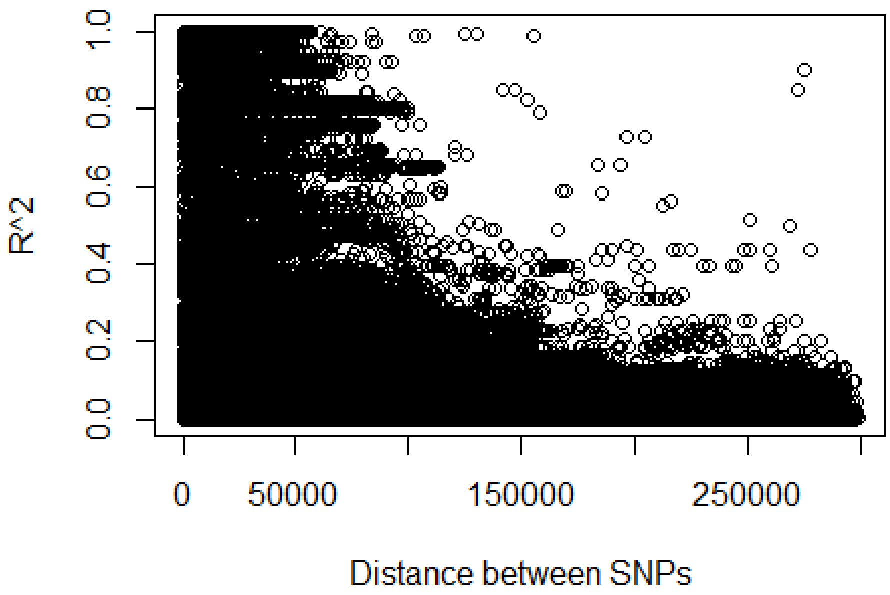
By filtering out variants with a MAF <=0.35, the proportion of variants that have a R2 close to 1 increases. This means that by removing variants, we are removing rare alleles from the data, and the remaining variants are more correlated.

7

Compute a distance matrix with the distance in base pairs between all possible pairs of SNPs, using the basepair position of each SNP given in the .bim file. Make a plot of R's R2 statistics against the distance (expressed as the number of basepairs) between the markers. Comment on your results.

```
setwd("/Users/sara_/Desktop/ERASMUS/BSG/bio_lab/bio-lab-9/update")

# calculate differences between base pair positions
distance_matrix <- outer(bim_data$V4, bim_data$V4, "-")


r2 <- r2_matrix[upper.tri(r2_matrix)]
dist <- distance_matrix[upper.tri(distance_matrix)]
```

```
plot(-dist, r2, ylab="R^2", xlab = "Distance between SNPs")
```



The above plot highlights the relationship between R2 and the distance between markers. We can notice there are high values of R2 at short distances; this could mean that When markers are very close to each other, there's a higher chance that they will be in strong LD. This is because, over short distances, recombination events are less likely to have occurred, and linked markers tend to be inherited together. Then, we notice a reduction of R2 with increasing distance between SNPs. As distance between markers increases, the probability of recombination events between them also increases.

HAPLOTYPE ESTIMATION

1

How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

```
setwd("/Users/sara_/Desktop/ERASMUS/BSG/bio_lab/bio-lab-9/update")
filename <- "APOE.dat"
raw_data_apoe <- read.table(filename, sep = " ",header = T)
raw_data_apoe_df <- data.frame(raw_data_apoe)


APOEdata <- raw_data_apoe_df[,2:ncol(raw_data_apoe_df)]

n <- nrow(APOEdata)# individuals
p <- ncol(APOEdata) #SNPs
```

```
missing <- 100*sum(is.na(APOEdata))/(n*p)

cat("How many individuals and how many SNPs are there in the database?\n")

## How many individuals and how many SNPs are there in the database?

cat("There are ",n,"variants and ", p, "SNPs.\n")

## There are  107 variants and  162 SNPs.

cat("What percentage of the data is missing?\n")

## What percentage of the data is missing?

cat(missing)

## 0
```

2

Assuming that all SNPs are bi-allelic, how many haplotypes can theoretically be found for this data set?

The number of possible haplotypes for a set of bi-allelic SNPs (Single Nucleotide Polymorphisms) can be calculated as 2^n, where n is the number of SNPs. This is because each SNP can have two possible alleles, and the total number of combinations is therefore 2^n. So the theoretical number is 5.846007e+48.

```
2**(p)

## [1] 5.846007e+48
```

3

Estimate haplotype frequencies using the haplo.em function that you will find in the haplo.stats package. How many haplotypes do you find? List the estimated probabilities in decreasing order. Which haplotype number is the most common?

```
separate_cols <- function(data) {
  separated_data <- data
  for (col_name in colnames(data)) {
    separated_data <- separate(separated_data, col_name, into = c(paste0(col_name,
"_1"), paste0(col_name, "_2")), sep = "/")
  }
  return(separated_data)
}

splitted_APOEdata <- separate_cols(APOEdata)



#locus_labels <- colnames(APOEdata)
haplo_em <- haplo.em(geno = splitted_APOEdata) #, locus.labels = locus_labels)
```

```
probabilities <-haplo_em$hap.prob
cat("Number of Estimated haplotypes: ", length(probabilities),"\n")

## Number of Estimated haplotypes:  34

sorted_probs <- sort(probabilities, decreasing = TRUE)
most_commont_haplotype <- which.max(probabilities)
cat("The number of the most common haplotype is ", most_commont_haplotype, "\n")

## The number of the most common haplotype is  29

print(sorted_probs)

##  [1] 4.027353e-01 1.308411e-01 7.166460e-02 6.821004e-02 5.020990e-02
##  [6] 4.547779e-02 3.572712e-02 3.499574e-02 2.252312e-02 1.907203e-02
## [11] 1.869159e-02 1.579814e-02 8.704437e-03 7.228961e-03 4.672897e-03
## [16] 4.672897e-03 4.672897e-03 4.672897e-03 4.672897e-03 4.672897e-03
## [21] 4.672897e-03 4.672897e-03 4.672897e-03 4.672897e-03 4.044694e-03
## [26] 3.489384e-03 3.386394e-03 3.015442e-03 2.225364e-03 1.860058e-03
## [31] 1.277019e-03 1.251185e-03 8.413670e-04 1.955155e-07
```

4

Remove all genetic variants that have a minor allele frequency below 0.10 from the database, and re-run haplo.em. How does this affect the number of haplotypes? Comment on your results.

```
maf_threshold <- 0.10

# Filter variants
filtered_apoe_data <- filter_by_maf(APOEdata, maf_threshold)

splitted_APOEdata_filtered <- separate_cols(filtered_apoe_data)


#locus_labels <- colnames(filtered_apoe_data)
haplo_em <- haplo.em(geno = splitted_APOEdata_filtered)#, locus.labels = locus_lab
els)

probabilities <-haplo_em$hap.prob
cat("Number of Estimated haplotypes: ", length(probabilities),"\n")

## Number of Estimated haplotypes:  9
```

As we can see from the two previous results, the number of estimated haplotypes would likely decrease if remove all genetic variants with a MAF below 0.1. This is because haplotypes are combinations of alleles at multiple loci that are transmitted together on the same chromosome. By removing variants, we are removing rare alleles from the data. Since these rare alleles contribute to the diversity of haplotypes, their removal would likely result in a reduced number of estimated haplotypes.