Antonin Rosa

Hendrik Wellmann

Practical 3 Statistical Genetics

Linkage disequilibrium

```
library(haplo.stats)
```

Question 1

```
## Loading required package: arsenal
library(genetics)
```

```
## Loading required package: combinat
##
## Attaching package: 'combinat'
## The following object is masked from 'package:utils':
##
##     combn
## Loading required package: gdata
##
## Attaching package: 'gdata'
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
## Loading required package: gtools
## Loading required package: MASS
## Loading required package: mvtnorm
##
## NOTE: THIS PACKAGE IS NOW OBSOLETE.
##
##   The R-Genetics project has developed an set of enhanced genetics
##   packages to replace 'genetics'. Please visit the project homepage
##   at http://rgenetics.org for informtion.
##
##
## Attaching package: 'genetics'
```

```
## The following object is masked from 'package:haplo.stats':
##
##     locus
## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
```

```
library(HardyWeinberg)
```

```
## Loading required package: mice
##
## Attaching package: 'mice'
## The following object is masked from 'package:stats':
##
##     filter
## The following objects are masked from 'package:base':
##
##     cbind, rbind
## Loading required package: Rsolnp
## Loading required package: nnet
```

```
data <- read.table("FOXP2/FOXP2.dat", header = TRUE, sep = " ")
num_individuals <- nrow(data)
num_snps <- ncol(data)

cat("Number of individuals:", num_individuals, "\n")
```

```
## Number of individuals: 104
```

```
cat("Number of SNPs:", num_snps, "\n")
```

```
## Number of SNPs: 544
```

```
missing_percentage <- mean(is.na(data)) * 100
cat("Percentage of missing data:", missing_percentage, "%\n")
```

```
## Percentage of missing data: 0 %
```

```

g1 <- genotype(data$rs34684677)
g2 <- genotype(data$rs2894715)
LD(g1,g2)
```

Question 2

```
##
## 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
```

The SNPs rs34684677 and rs2894715 show a strong genetic linkage, meaning that variations in one SNP are highly likely to be associated with variations in the other. The high D' value (0.9986536) indicates a strong linkage. The moderate negative correlation (Corr = -0.3144048) suggests that variations in allele frequencies tend to occur in opposite directions. In summary, there's a significant and non-random association between alleles of rs34684677 and rs2894715, with a tendency for opposite variations.

Question 3 Low LD (D = -0.05493703) suggests weak association between SNPs rs34684677 and rs2894715. Haplotypes are likely formed independently. Estimate haplotype frequencies from genotype frequencies. The most common haplotype has the highest estimated frequency. Provide genotype frequencies for detailed calculation.

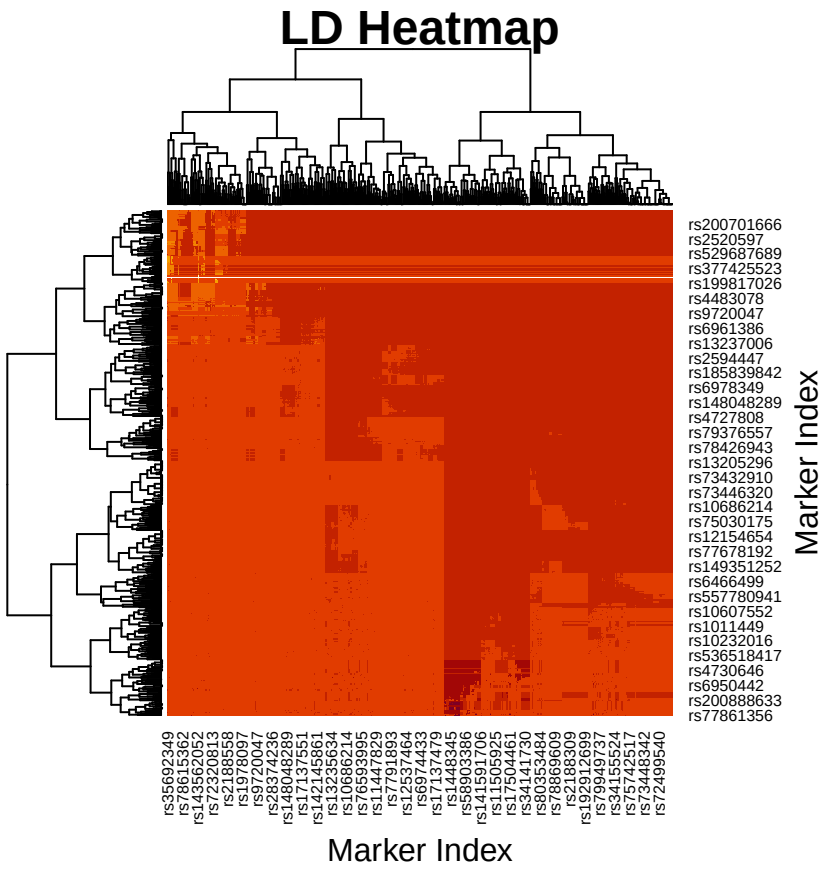
```
bim_file <- "FOXP2/FOXP2.bim"
bim_data <- read.table(bim_file, header = FALSE, stringsAsFactors = FALSE)

genotype_matrix <- data[,-1]
for (col in colnames(genotype_matrix)) {
  genotype_matrix[[col]] <- genotype(genotype_matrix[[col]])
}
calculate_MAF <- function(genotype_counts) {
  allele1 <- 2 * genotype_counts[[1]] + genotype_counts[[2]]
  allele2 <- 2 * genotype_counts[[3]] + genotype_counts[[2]]
  return (min(allele1, allele2) / (allele2 + allele1))
}

filter_index_MAF <- c()
res <- 0
for (i in 1:ncol(genotype_matrix)) {
  alleles <- c(paste(bim_data$V5[[i]], "/", bim_data$V6[[i]], sep = ""))
  count_gentoype <- MakeCounts(genotype_matrix[[i]], alleles = alleles, sep="/")
  MAF <- calculate_MAF(count_gentoype)
  if (MAF < 0.35) {
    filter_index_MAF <- c(filter_index_MAF, i)
  }
  p_value <- HWE.chisq(genotype_matrix[[i]])
  p_value <- p_value$p.value
  if (p_value < 0.05) {
    res <- res + 1
  }
}
print(res)
```

Question 4

```
## [1] 18
```

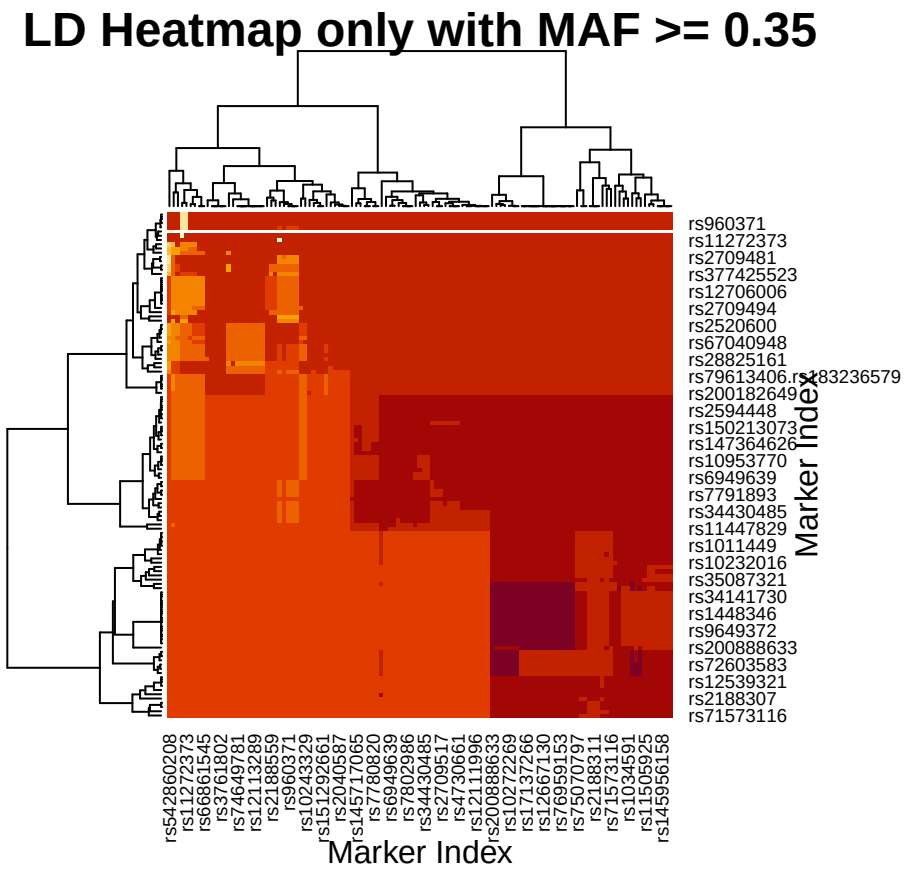
We can reject 18 variants.

```
ld_matrix <- LD(genotype_matrix)
r2_matrix <- ld_matrix$`R^2`
r2_matrix[is.na(r2_matrix)] <- 1
heatmap(r2_matrix, main = "LD Heatmap", xlab = "Marker Index", ylab = "Marker Index")
```



Question 5

```
r2_matrix <- r2_matrix[-filter_index_MAF, -filter_index_MAF]
heatmap(r2_matrix, main = "LD Heatmap only with MAF >= 0.35", xlab = "Marker Index", ylab = "Marker Inde
```



Question 6

Question 7

Haplotype estimation

```
data <- read.table("APOE/APOE.dat", header = TRUE, sep = " ")
cat("Number of individuals:", num_individuals, "\n")
```

Question 1

```
## Number of individuals: 104
```

```
cat("Number of SNPs:", num_snps, "\n")
```

```
## Number of SNPs: 544
```

```
missing_percentage <- mean(is.na(data)) * 100
cat("Percentage of missing data:", missing_percentage, "%\n")
```

```
## Percentage of missing data: 0 %
```

Question 2

Question 3