Practical 03 SG: Linkage Disequilibrium

Lovro Katalinić and Ivan Almer

Hand-in: 05/12/2020

Resolve the following exercise in groups of two students. Perform the computations and make the graphics that are asked for in the practical below. Take care to give each graph a title, and clearly label x and y axes, and to answer all questions asked. You can write your solution in a word or Latex document and generate a pdf file with your solution. Alternatively, you may generate a solution pdf file with Markdown. You can use R packages **genetics**, **HardyWeinberg** and **LDheatmap** for the computations. Take care to number your answer exactly as in this exercise. Upload your solution in **pdf format** to the web page of the course at raco.fib.upc.edu no later than the hand-in date.

1. The file FOXP2.zip contains genetic information of individuals of a Japanese population of unrelated individuals. The genotype information concerns SNPs of the Forkhead box protein P2 (FOXP2) gene region, located the long arm of chromosome number 7. This gene plays an important role in the development of speech and language. The `FOXP2.zip` file contains:

- `FOXP2.dat`: a text file with the genotype data which can be read in with R.
- `FOXP2.fam`: a PLINK file with data on the individuals (family id, individual id, ids of parents, sex and phenotype).
- `FOXP2.bed`: a PLINK file with binary genotype data.
- `FOXP2.bim`: a PLINK file with data on the genetic variants (chromosome, SNP identifier, basepair position along the chromosome and alleles).

```
## Loading required package: combinat

##
## Attaching package: 'combinat'

## The following object is masked from 'package:utils':
##
##     combn

## Loading required package: gdata

## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##
## 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 objects are masked from 'package:base':
##
##     %in%, as.factor, order
## 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
##
## Attaching package: 'data.table'
## The following objects are masked from 'package:gdata':
##
##     first, last
```

2. (1p) Load the `FOXP2.dat` file into the R environment. How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

```
dataset <- fread('./FOXP2.dat', data.table=FALSE)
#head(dataset)

cat(paste('Number of individuals:', nrow(dataset), '\n'))
```

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

```
cat(paste('Number of variants:', ncol(dataset) - 1, '\n'))
```

```
## Number of variants: 543
```

```
snp.dataset = dataset[,2:ncol(dataset)]
a = apply(snp.dataset, 2, unique)

list = c()
for(aa in a) {
  for(v in aa) {
    list = append(list, v)
  }
}

unique(list)
```

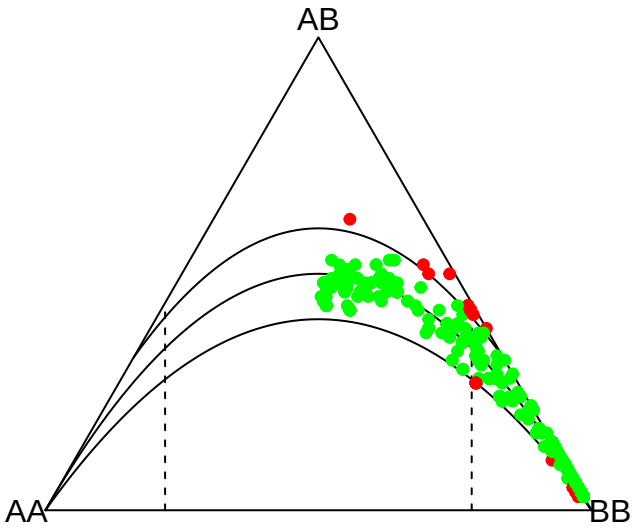
```
##  [1] "T/G" "G/G" "T/T" "C/T" "C/C" "G/A" "A/A" "A/T" "G/T" "A/G" "A/C" "T/C"
## [13] "C/G" "T/A" "C/A" "G/C"
```

```
cat(paste('There is no missing values!'))
```

```
## There is no missing values!
```

3. (1p) Determine the genotype counts for each SNP, and depict all SNPs simultaeneously in a ternary plot, and comment on your result. For how many variants do you reject Hardy-Weinberg equilibrium using an ordinary chi-square test without continuity correction? (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).

```
snp.alleles = fread('FOXP2.bim', data.table=FALSE)
marker.alleles = c()
for(i in 1:nrow(snp.alleles)) {
  al = paste(snp.alleles[i,5], snp.alleles[i,6], sep="/")
  marker.alleles = append(marker.alleles, al)
}

genotype.counts = MakeCounts(snp.dataset, marker.alleles, sep="/")
genotype.counts = genotype.counts[,1:3]

HWTernaryPlot(genotype.counts)
```




```
pvalues_chi <- HWChisqStats(genotype.counts, pvalues=TRUE)
length(pvalues_chi)
```

```
## [1] 543
```

```
reject.num = sum(pvalues_chi < 0.05)

cat(paste('We reject the H0 for', reject.num, 'variants.\n'))
```

```
## We reject the H0 for 33 variants.
```

4. (1p) 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 = snp.dataset$rs34684677
snp2 = snp.dataset$rs2894715
snp1.g = genotype(snp1, sep = "/")
snp2.g = genotype(snp2, sep = "/")
LD(snp1.g, snp2.g)
```

```
##
## 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 p-value is small enough so we reject the Equilibrium hypothesis, i.e. there exists some correlation between the 2 SNPs. 5. (2p) Also compute the LD statistic D for the SNPs rs34684677 and rs998302 of the database. Is there significant association between these two SNPs? Is there any reason why rs998302 could have stronger or weaker correlation than rs2894715?

```
snp1.g = genotype(snp.dataset$rs34684677, sep = "/")
snp2.g = genotype(snp.dataset$rs998302, sep = "/")
snp3.g = genotype(snp.dataset$rs2894715, sep = "/")

table(snp1.g, snp2.g)
```

```
##       snp2.g
## snp1.g G/G G/T
##    G/G  67   6
##    G/T  25   3
##    T/T   2   1
```

```
table(snp1.g, snp3.g)
```

```
##       snp3.g
## snp1.g G/G T/G T/T
##    G/G  12  37  24
##    G/T   0   9  19
##    T/T   0   0   3
```

```
out = LD(snp1.g, snp2.g)
out$D
```

```
## [1] 0.007208888
```

```
out$`P-value`
```

```
## [1] 0.1887601
```

```
out = LD(snp1.g, snp3.g)
out$D
```

```
## [1] -0.05493703
```

```
out$`P-value`
```

```
## [1] 5.77645e-06
```

There is no statistically significant association between the SNPs `rs34684677` and `rs998302` whereas there is a significant association between markers `rs34684677` and `rs2894715`.

6. (2p) Given your previous estimate of D for SNPs rs34684677 and rs2894715, infer the haplotype frequencies. Which haplotype is the most common?

```
snp1.g = genotype(snp.dataset$rs34684677, sep = "/")
snp3.g = genotype(snp.dataset$rs2894715, sep = "/")

out = LD(snp1.g, snp3.g)
out$D
```

```
## [1] -0.05493703
```

```
out$`P-value`
```

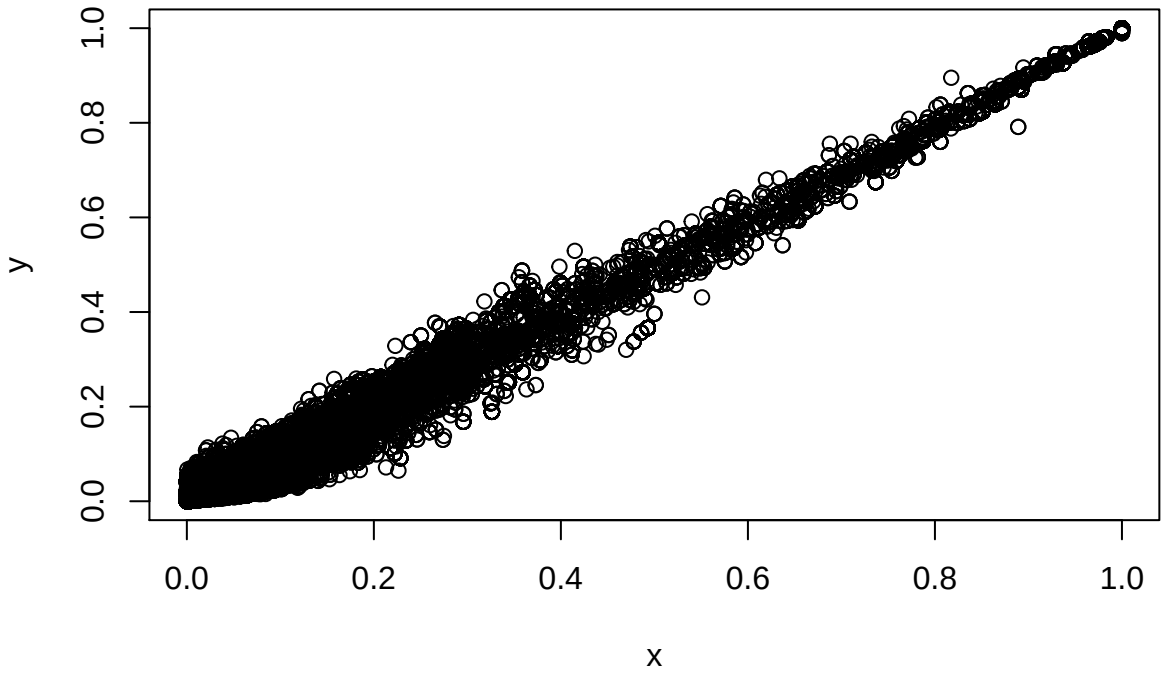
```
## [1] 5.77645e-06
```

7. (2p) Compute the LD statistics R^2 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. Also compute an alternative estimate of R^2 obtained by using the `PLINK` program. For this purpose you should:

- Download and install PLINK 1.90 from https://www.cog-genomics.org/plink2/
- Take care to store the files FOXP2.bim, FOXP2.fam and FOXP2.bed in a directory where PLINK can find them.
- Compute LD estimates with PLINK using `plink --bfile FOXP2 --r2 --matrix --out FOXP2`
- This creates a file with extension `FOXP2.ld` that contains a matrix with all R^2 statistics. Read this file into the R environment.
- Make a scatter plot for R's LD estimates against PLINK's LD estimates. Are they identical or do they at least correlate? What's the difference between these two estimators? Which estimator would your prefer and why?

```
RES <- data.frame(genotype(snp.dataset[,1],sep="/"))

for(i in 2:ncol(snp.dataset)) {
   snp <- genotype(snp.dataset[,i],sep="/")
   RES <- cbind(RES,snp)
}
```

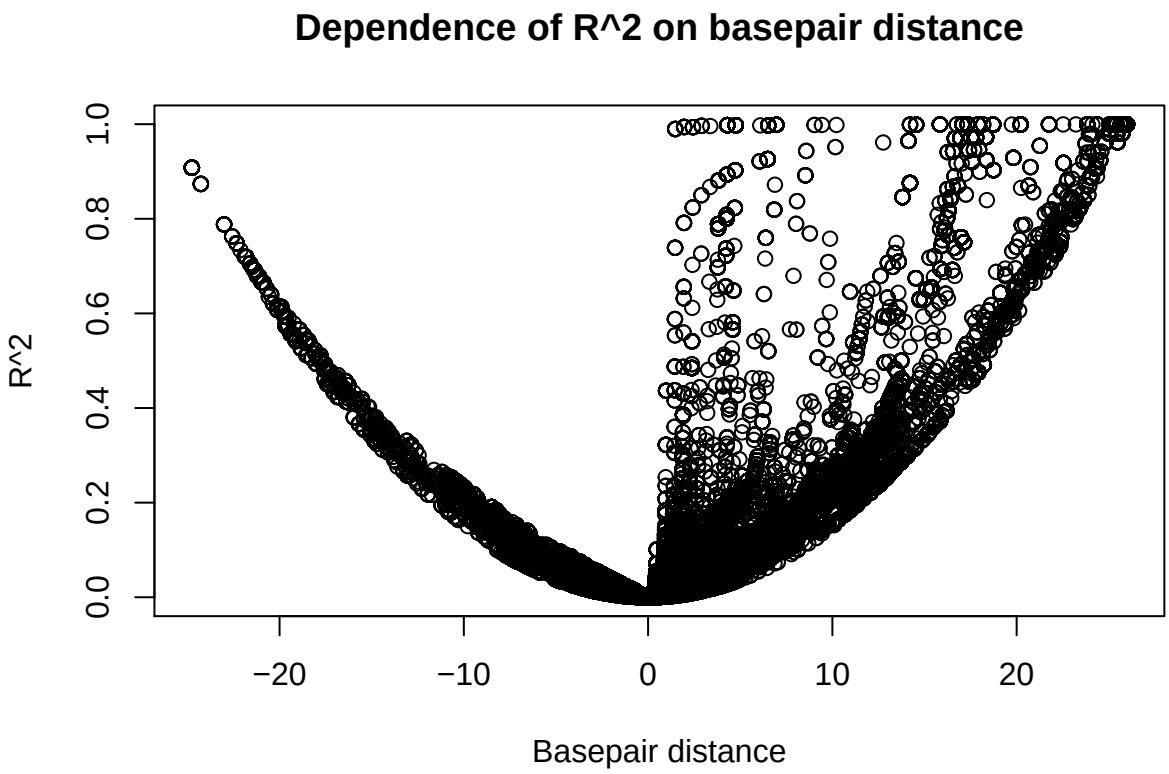
```
#output <- LD(RES)
#attributes(output)

#Dm <- output$D
#Dp <- output$"D'"
#R2 <- output$"R^2"
#X2 <- output$"X^2"

Dm = read.table("Dm.txt")
Dp = read.table("Dp.txt")
R2 = read.table("R2.txt")
X2 = read.table("X2.txt")

#write.table(Dm, file="Dm.txt", row.names=FALSE, col.names=FALSE)
#write.table(Dp, file="Dp.txt", row.names=FALSE, col.names=FALSE)
#write.table(R2, file="R2.txt", row.names=FALSE, col.names=FALSE)
#write.table(X2, file="X2.txt", row.names=FALSE, col.names=FALSE)

R2.generated = read.table("FOXP2.ld")
for(i in 1:ncol(R2.generated)) {
  for(j in 1:i) {
    R2.generated[i,j] = NA
  }
}

y = c(t(R2))
y = y[!is.na(y)]

x = c(t(R2.generated))
x = x[!is.na(x)]

plot(x,y)
```



There is not a significant difference between R's estimate and PLINK's estimate. We can see the high linear correlation (1-1 correlation) between the 2 estimators. If we had a structured dataset I would probably prefer to use PLINK because of its speed, but if the data is complex, R would be a better option although it is slower.

8. (2p) 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 R^2 statistics against the distance (expressed as the number of basepairs) between the markers. Comment on your results.

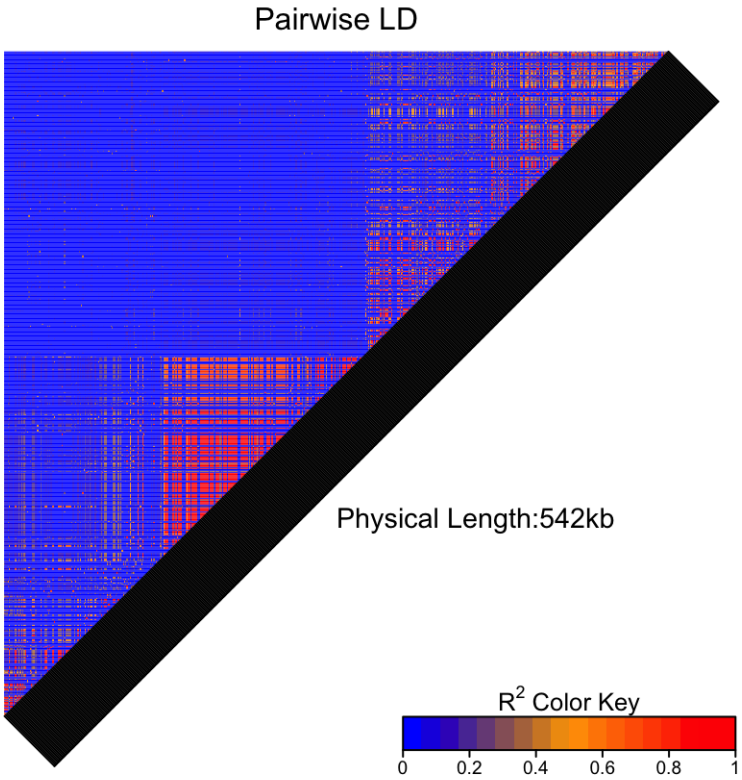
```
D.basepairs = nrow(snp.dataset) * Dm

x = c(t(D.basepairs))
x = x[!is.na(x)]

y = c(t(R2))
y = y[!is.na(y)]

plot(x,y, xlab = 'Basepair distance', ylab = 'R^2', main = 'Dependence of R^2 on basepair distance')
```



We see the quadratic behavior of R^2 with the minimum value when distance between base pairs reaches 0.0. This is an expected result because the R^2 really has a quadratic relationship with D.

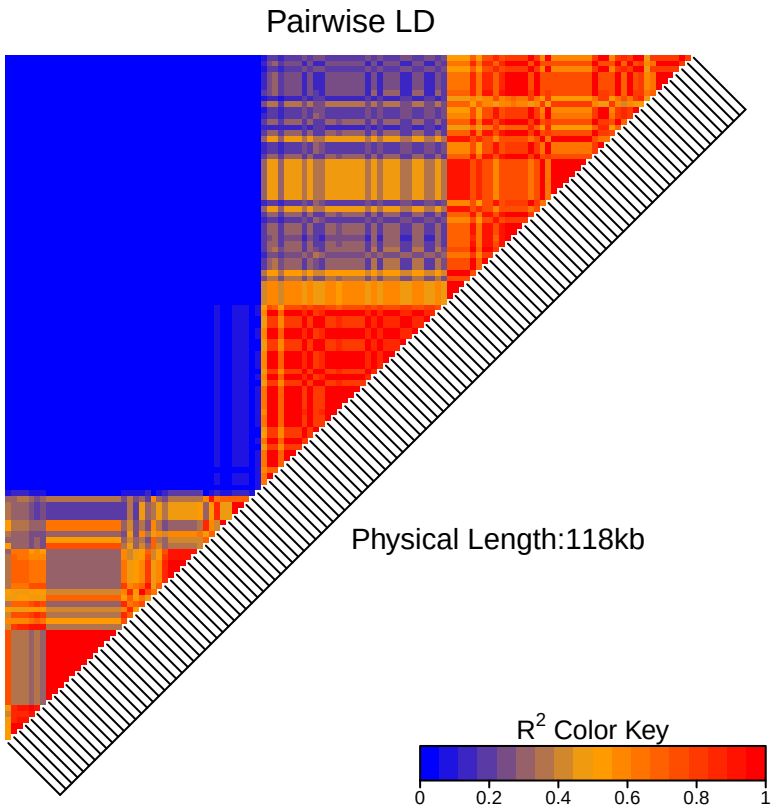
9. (2p) Make an LD heatmap of the markers in this database, using the R^2 statistic with the `LD` function. Make another heatmap obtained by filtering out all variants with a MAF below 0.35, and redoing the computations to obtain the R^2 statistics in R. Can you explain any differences observed between the two heatmaps?

```
rgb.palette <- colorRampPalette(rev(c("blue", "orange", "red")), space = "rgb")

#LDheatmap(RES,LDmeasure="r",color=rgb.palette(18))
# We generated the image once but the waiting time was around 10 minutes so we saved the image and pres
# display image
include_graphics("./res_heatmap.png", dpi = 200)
```




```
maf <- function(x){
  x <- genotype(x,sep="/")
  out <- summary(x)
  af1 <- min(out$allele.freq[,2],na.rm=TRUE)
  af1[af1==1] <- 0
  return(af1)
}

mafs = apply(snp.dataset, 2, maf)
mask = mafs > 0.35

snp.dataset.filtered = snp.dataset[,mask]

RES.filtered <- data.frame(genotype(snp.dataset.filtered[,1],sep="/"))

for(i in 2:ncol(snp.dataset.filtered)) {
   snp <- genotype(snp.dataset.filtered[,i],sep="/")
   RES.filtered <- cbind(RES.filtered,snp)
}

output.filtered <- LD(RES.filtered)

Dm <- output.filtered$D
Dp <- output.filtered$"D'"
R2 <- output.filtered$"R^2"
X2 <- output.filtered$"X^2"
```

```
LDheatmap(RES.filtered,LDmeasure="r",color=rgb.palette(18))
```



We can see that the uncorrelated pairs prevail in the case of the whole dataset, compared to the case when we filter out the variants with maf less than 0.35. The argument for that could be that the variants with maf < 0.35 don't provide enough "variation" in data to show correlation with other variants.

10. (1p) Can you distinguish blocks of correlated markers in the area of the FOXP2 gene? How many blocks do you think that *at least* seem to exist?

The blocks of correlated markers are more to the red end of the color spectrum.

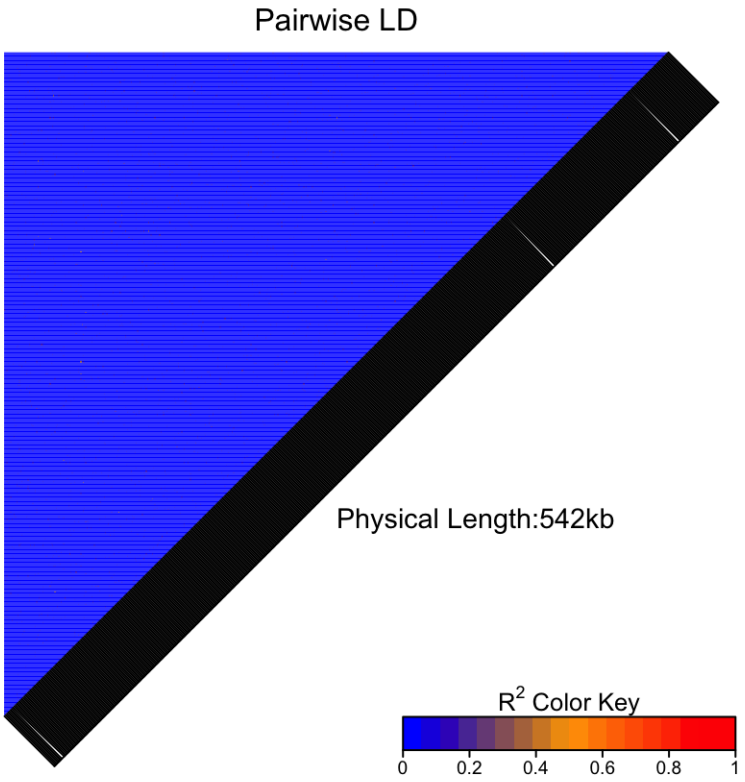
For filtered variants there should be at least 2400 pairs which are correlated (e.g. R > 0.5). The red triangles divide the height in roughly 3 parts, and we have 3 halves of a square with side length of 119/3.

For the original dataset it is hard to get a feeling about the number of correlated pairs, but a good guess would be to use the same number as above, i.e. 2400 pairs.

11. (1p) Simulate independent SNPs under the assumption of Hardy-Weinberg equilibrium, using R's `sample` instruction `sample(c("AA","AB","BB"),n,replace=TRUE,prob=c(p*p,2*p*q,q*q))`. Simulate as many SNPs as you have in your database, and take care to match each SNP in your database with a simulated SNP that has the same sample size and allele frequency. Make an LD heatmap of the simulated SNPs, using R^2 as your statistic. Compare the results with the LD heatmap of the FOXP2 region. What do you observe? State your conclusions.

```
n = nrow(snp.dataset)
sample.snp <- function(x) {
  x.g = genotype(x, sep = "/")
  out = summary(x.g)
  p = out$allele.freq[1,2]
```

```
  q = out$allele.freq[2,2]
  A = out$allele.names[1]
  B = out$allele.names[2]
  AA = paste(A,A, sep = "/")
  AB = paste(A,B, sep = "/")
  BB = paste(B,B, sep = "/")

  return(sample(c(AA,AB,BB),n,replace=TRUE,prob=c(p*p,2*p*q,q*q)))
}

sample.dataset = data.frame(apply(snp.dataset, 2, sample.snp))
#head(snp.dataset)
#head(sample.dataset)

RES.sample <- data.frame(genotype(sample.dataset[,1],sep="/"))

for(i in 2:ncol(sample.dataset)) {
   snp <- genotype(sample.dataset[,i],sep="/")
   RES.sample <- cbind(RES.sample,snp)
}

#LDheatmap(RES.sample,LDmeasure="r",color=rgb.palette(18))

# We generated the image once but the waiting time was around 10 minutes so we saved the image and pres
# display image
include_graphics("./res_sample_heatmap.png", dpi = 200)
```



We can see that the generated pairs show almost no correlation when generated under HW equilibrium assumption, which is very interesting to see :)