Practical 03 SG: Linkage Disequilibrium and Haplotype estimation

Quim Aguado, Marcel Cases

06-Dec-2021

Linkage Disequilibrium

Ex 2 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?

```
df <- read.table("FOXP2.dat", header=TRUE)
gendata <- df[,2:ncol(df)] # gendata does not contain monomorphics

n <- nrow(gendata) # number of individuals
p <- ncol(gendata) # number of variants
n
```

```
## [1] 104
```

```
p
```

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

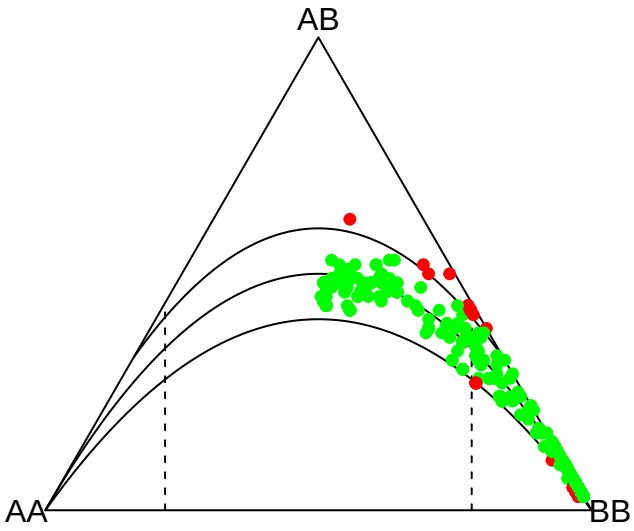
```
perc.mis <- 100*sum(is.na(df))/(n*p)
perc.mis
```

```
## [1] 0
```

There are 104 individuals and 543 SNP variants in the dataset. There is no missing data.

Ex 3 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?

```
bim <- read.table("FOXP2.bim", header=FALSE)
bim_alleles <- paste(bim[,5],bim[,6],sep = "/")
gendata.counts <- MakeCounts(gendata, bim_alleles, sep = "/")

HWTernaryPlot(gendata.counts[,1:3])
```



Most of the SNPs fall within the acceptability threshold.

```
gendata.chisq.stats <- HWChisqStats(gendata.counts,pvalues = FALSE)
gendata.chisq.pval <- HWChisqStats(gendata.counts,pvalues = TRUE)
gendata.chisq.pval.significant <- sum(gendata.chisq.pval<0.05) # number of significant SNPs
gendata.chisq.pval.significant
```

```
## [1] 33
```

```
#plot(density(gendata.chisq.stats))
```

We reject Hardy-Weinberg equilibrium for 33 variants.

Ex 4 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?

```
rs34684677 <- gendata[,c("rs34684677")]
rs2894715 <- gendata[,c("rs2894715")]
rs34684677.g <- genotype(rs34684677,sep="/")
rs2894715.g <- genotype(rs2894715,sep="/")
rs34684677.rs2894715.LD <- LD(rs34684677.g,rs2894715.g)
rs34684677.rs2894715.LD
```

```
##
## 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 estimated D is <0, which means that there is no significant association between the alleles of the two SNPs.

Ex 5 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?

```
rs998302 <- gendata[,c("rs998302")]
rs998302.g <- genotype(rs998302,sep="/")
rs34684677.rs998302.LD <- LD(rs34684677.g,rs998302.g)
rs34684677.rs998302.LD
```

```
##
## Pairwise LD
## -----------
##                       D        D'       Corr
## Estimates: 0.007208888 0.1792444 0.09112725
##
##               X^2   P-value   N
## LD Test: 1.727268 0.1887601 104
```

This time, D > 0, which means that the alleles of the two SNPs have some kind of correlation.

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

```
Geno <- cbind(substr(rs34684677,1,1), substr(rs34684677,3,3),
              substr(rs2894715,1,1), substr(rs2894715,3,3))
snpnames <- c("rs34684677","rs2894715")
HaploEM <- haplo.em(Geno,locus.label=snpnames,control=haplo.em.control(min.posterior=1e-4))
HaploEM
```

```
## ============================================================================
##                                  Haplotypes
## ============================================================================
##   rs34684677 rs2894715 hap.freq
## 1          G         G  0.33654
## 2          G         T  0.50000
## 3          T         T  0.16346
## ============================================================================
##                                   Details
## ============================================================================
## lnlike =  -164.8458
## lr stat for no LD =  7.63774 , df =  0 , p-val =  NA
```

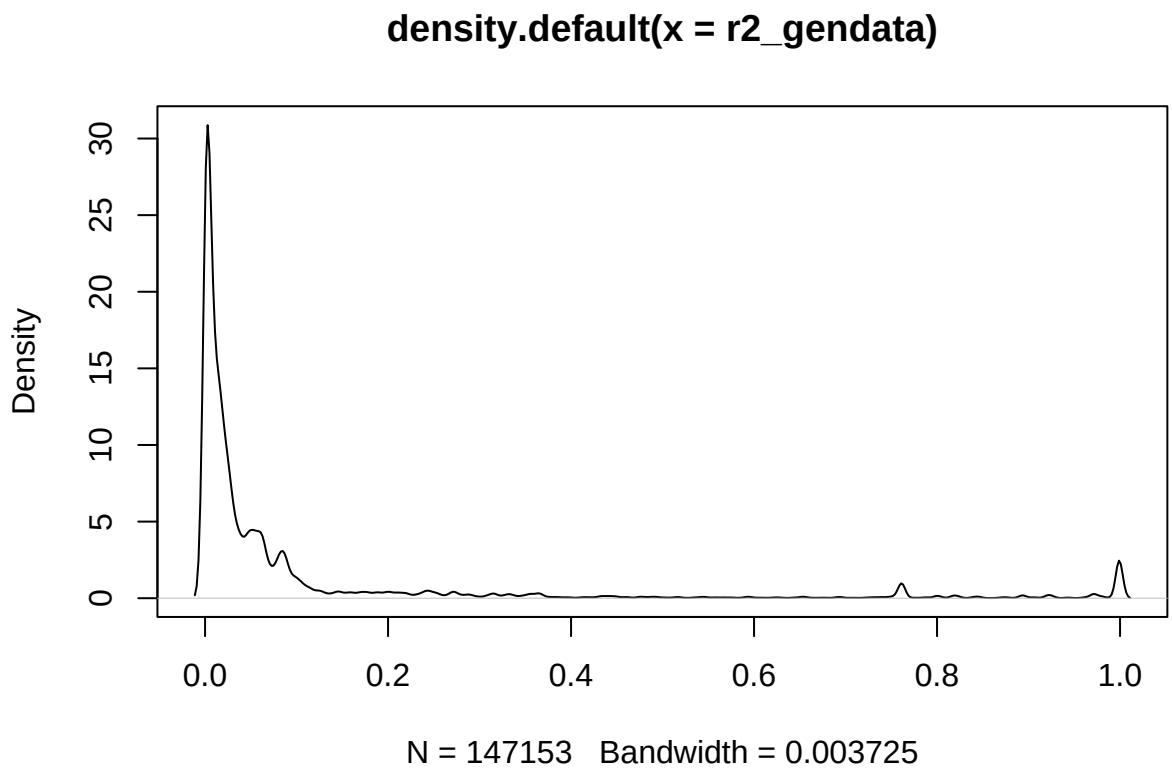
The most common haplotype between rs34684677 and rs2894715 SNPs is "G/T" (freq=0.5).

Ex 7 Compute the LD statistics R2 for all the marker pairs in this data base, using the LD function of the packages genetics.

Using R:

```
r2_gendata = c()

for(i in 2:p) { # generate all permutations
  for(j in 1:(i-1)) {
```

```
    a <- genotype(gendata[,i],sep="/")
    b <- genotype(gendata[,j],sep="/")
    r2_gendata <- c(r2_gendata, LD(a,b)$`R^2`)
  }
}

plot(density(r2_gendata))
```



Using PLINK 1.90:

We run the command:

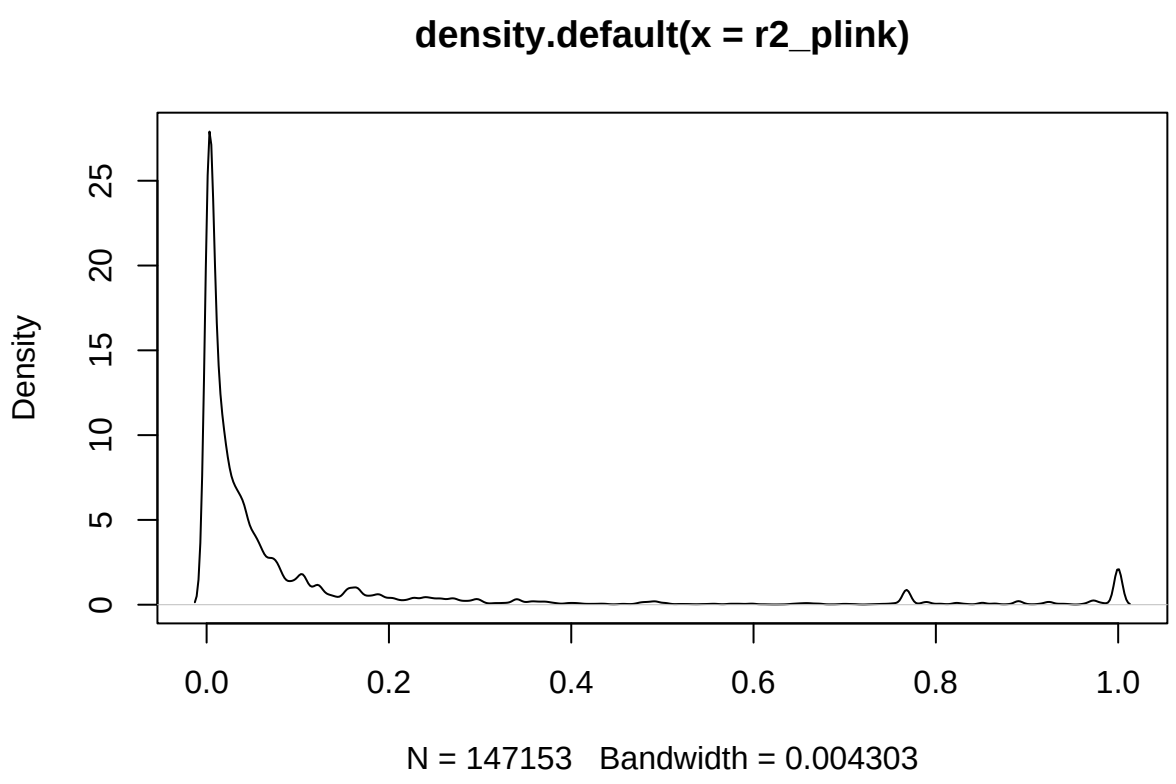
```
./plink --bfile FOXP2 --r2 --matrix --out FOXP2
```

The run log is:

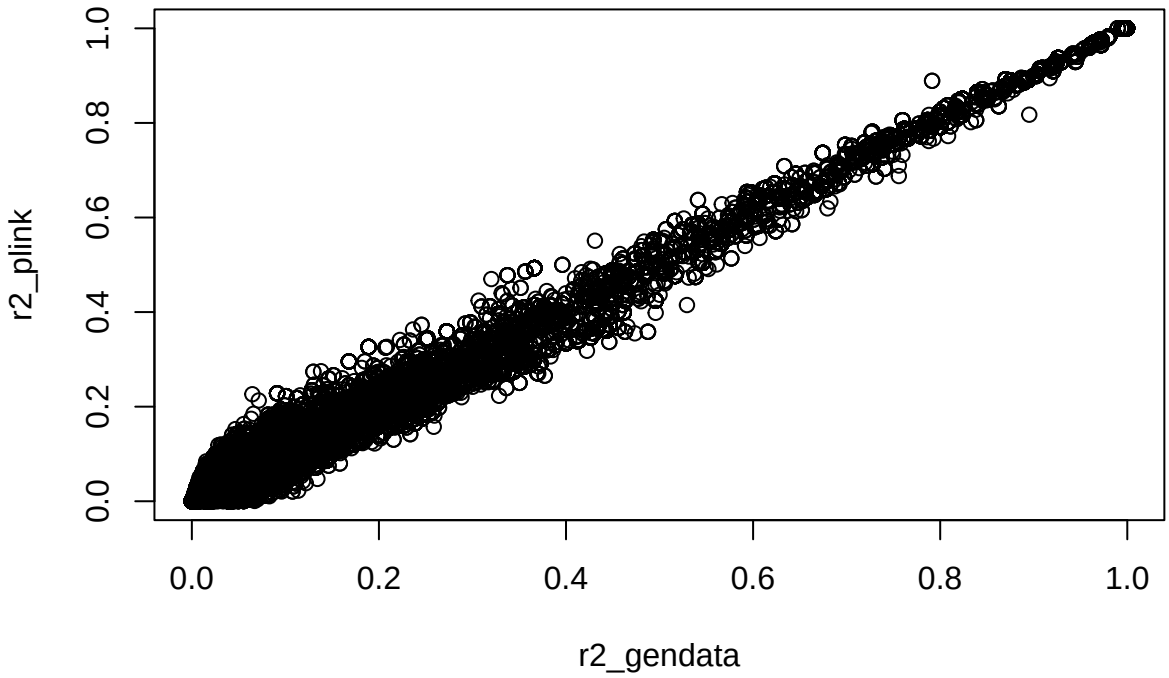
```
PLINK v1.90b6.24 64-bit (6 Jun 2021)           www.cog-genomics.org/plink/1.9/
(C) 2005-2021 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to FOXP2.log.
Options in effect:
  --bfile FOXP2
  --matrix
  --out FOXP2
  --r2

Note: --matrix flag deprecated.  Migrate to "--distance ibs flat-missing",
"--r2 square", etc.
12714 MB RAM detected; reserving 6357 MB for main workspace.
```

```
543 variants loaded from .bim file.
104 people (0 males, 0 females, 104 ambiguous) loaded from .fam.
Ambiguous sex IDs written to FOXP2.nosex .
Using up to 8 threads (change this with --threads).
Before main variant filters, 104 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is exactly 1.
543 variants and 104 people pass filters and QC.
Note: No phenotypes present.
--r2 square to FOXP2.ld ... done.
```

```
plink_raw <- read.table("FOXP2.ld", header=FALSE)

r2_plink = c()
for(i in 2:ncol(plink_raw)) {
  for(j in 1:(i-1)) {
    r2_plink <- c(r2_plink, plink_raw[i,j])
  }
}

plot(density(r2_plink))
```



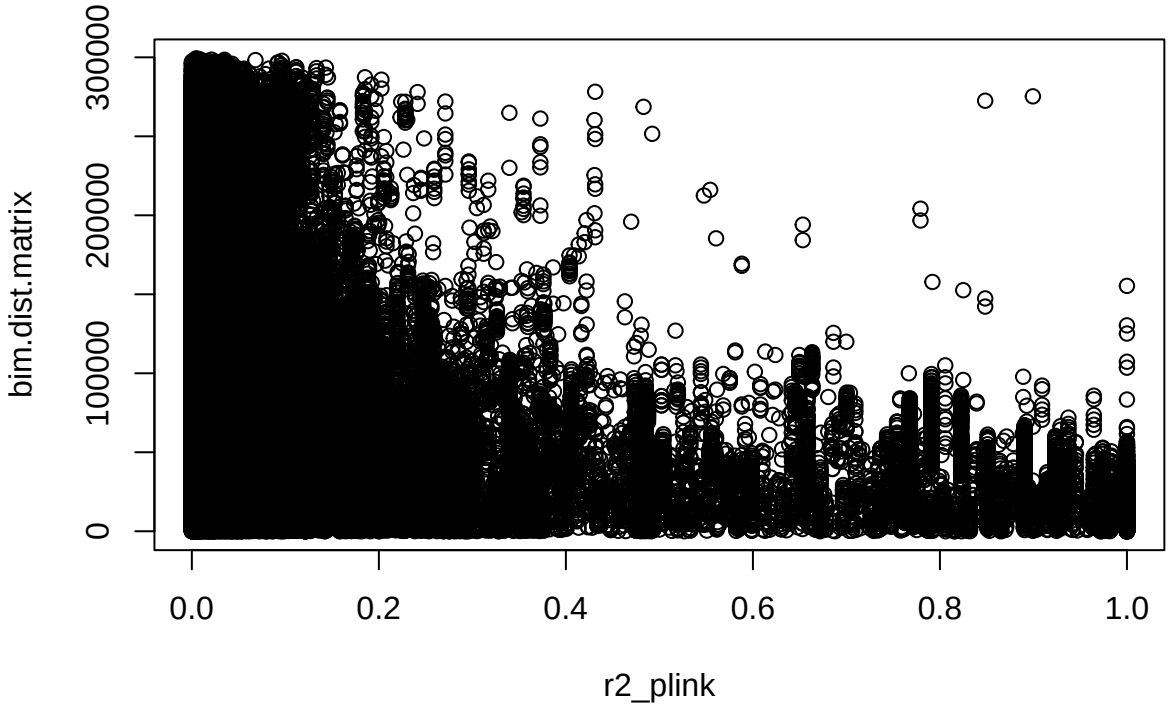

```
plot(r2_gendata,r2_plink)
```



The scatter plot of r2_gendata and r2_plink is not equal but closely correlated. We would clearly choose PLINK because it is optimized for efficient computing speed.

Ex 8 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.

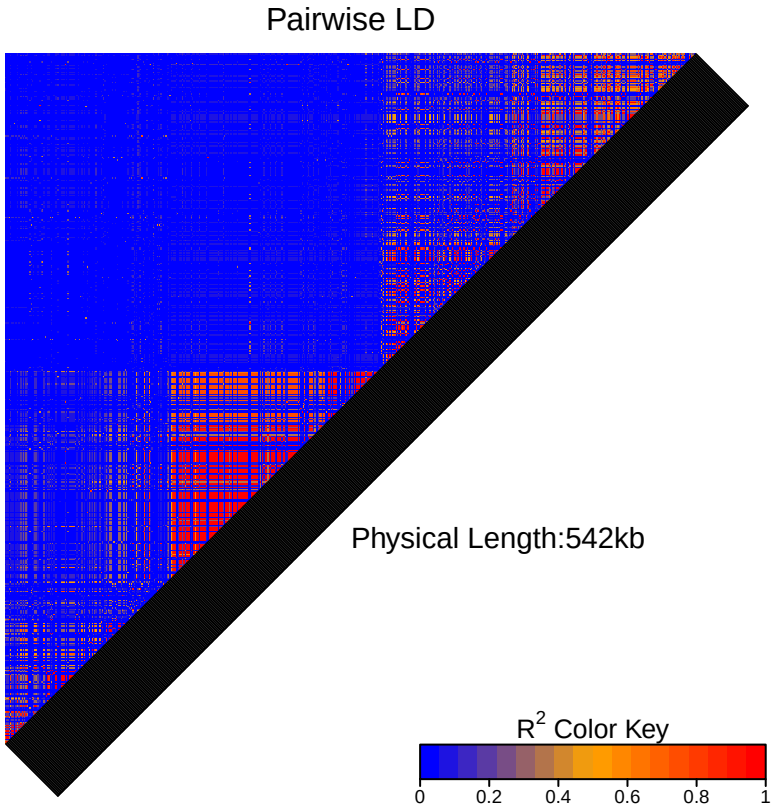
```
bim.dist <- bim[,4]
bim.dist.matrix <- c() # expressed as a vector

for(i in 2:length(bim.dist)) {
  for(j in 1:(i-1)) {
    bim.dist.matrix <- c(bim.dist.matrix, bim.dist[i]-bim.dist[j])
  }
}

plot(r2_plink,bim.dist.matrix)
```



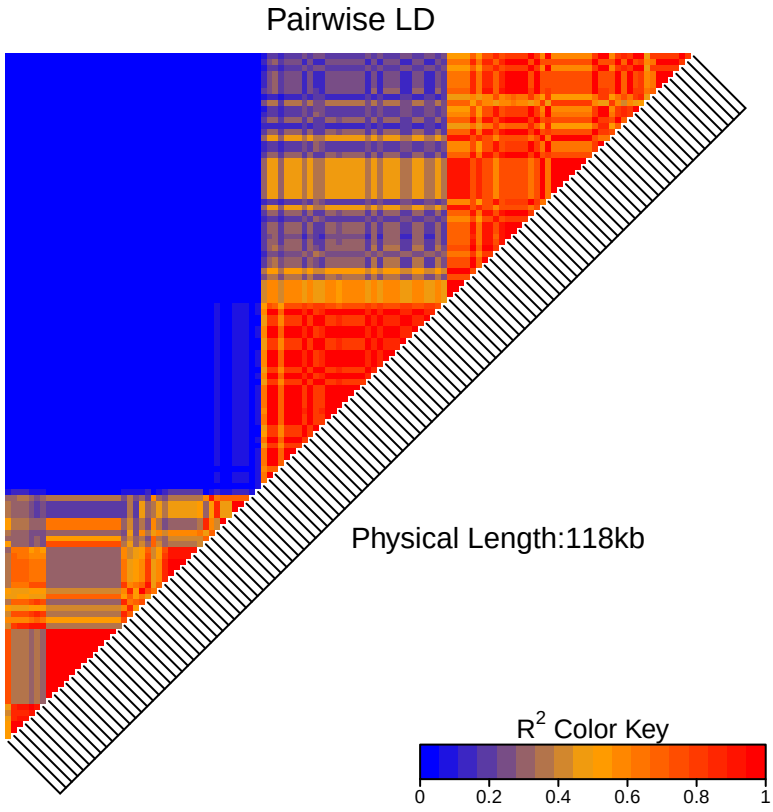
There is no correlation between the distance and R^2 of each pair of SNP.

Ex 9 Make an LD heatmap of the markers in this database, using the R2 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 R2 statistics in R. Can you explain any differences observed between the two heatmaps?

```
gendata.genotype <- data.frame(genotype(gendata[,1],sep="/"))
for(i in 2:ncol(gendata)) {
   snp <- genotype(gendata[,i],sep="/")
   gendata.genotype <- cbind(gendata.genotype,snp)
}
rgb.palette <- colorRampPalette(rev(c("blue", "orange", "red")), space = "rgb")
LDheatmap(gendata.genotype,LDmeasure="r",color=rgb.palette(18))
```




```
n_poly <- nrow(gendata)
nmis <- function(x) {
  y <- sum(is.na(x))
  return(y)
}
nmis.per.snp <- apply(gendata,2,nmis)
pmis.per.snp <- 100*nmis.per.snp/n_poly
Y2 <- gendata[,nmis.per.snp < nrow(gendata)]
maf <- function(x){ # minor allele frequency
  x <- genotype(x,sep="/")
  out <- summary(x)
  af1 <- min(out$allele.freq[,2],na.rm=TRUE)
  af1[af1==1] <- 0
  return(af1)
}
maf.per.snp <- apply(Y2,2,maf)
#hist(maf.per.snp)
gendata.genotype.maf_0.35 <- gendata.genotype[maf.per.snp>=0.35]
LDheatmap(gendata.genotype.maf_0.35,LDmeasure="r",color=rgb.palette(18))
```



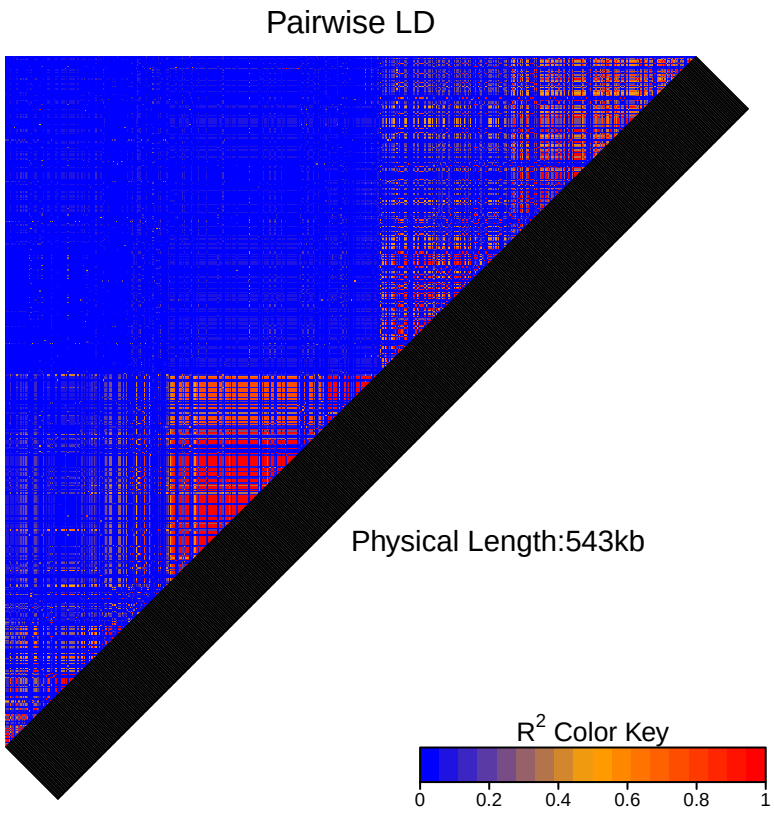
Ex 10 Can you distinguish blocks of correlated markers in the area of the FOXP2 gene? How many blocks do you think that at least seem to exist?

We can observe at least 5 different blocks of correlated SNPs.

Ex 11 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^2,2pq,qq))). 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 R2 as your statistic. Compare the results with the LD heatmap of the FOXP2 region. What do you observe? State your conclusions.

```
gendata.sim <- c()
for(i in 1:ncol(gendata)) {
  snpi <- gendata[,i]
  snpi.g <- genotype(snpi,sep="/")
  freqA <- summary(snpi.g)$allele.freq[1,2]
  freqB <- summary(snpi.g)$allele.freq[2,2]
  nameA <- summary(snpi.g)$allele.names[1]
  nameB <- summary(snpi.g)$allele.names[2]
  hom1 <- paste(nameA,"/",nameA,sep="")
  hom2 <- paste(nameB,"/",nameB,sep="")
  het <- paste(nameA,"/",nameB,sep="")
  samplei <- sample(c(hom1,het,hom2),nrow(gendata),replace=TRUE,prob=c(freqA*freqA,2*freqA*freqB,freqB*
  gendata.sim <- cbind(gendata.sim,samplei)
}
#gendata.sim[1:6,1:6]
```

```
gendata.sim.genotype <- data.frame(genotype(gendata.sim[,1],sep="/"))
for(i in 2:ncol(gendata.sim)) {
   snp <- genotype(gendata.sim[,i],sep="/")
   gendata.sim.genotype <- cbind(gendata.genotype,snp)
}
LDheatmap(gendata.sim.genotype,LDmeasure="r",color=rgb.palette(18))
```



We observe that the R^2 heatmap of the original dataset is very similar to the one generated with artificial data from size and frequency of the alleles from the original dataset: we see the same correlated markers area/blocks.