Bioinformatics and Statistical Genetic

Linkage Disequilibrium

Leonardo Ortoleva & Egon Ferri

Contents


```
knitr::opts_chunk$set(echo = F)
```

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?

Table 1: Data

id	rs34684677	rs1839115	rs4727804	rs4727805	rs200888633	rs12534908
NA18939	T/G	C/T	G/A	T/G	T/G	G/A
NA18940	G/G	T/T	A/A	G/G	T/G	A/A
NA18941	G/G	T/T	A/A	G/G	T/G	A/A
NA18942	G/G	T/T	A/A	G/G	T/T	A/A
NA18943	G/G	T/T	A/A	G/G	T/T	A/A
NA18944	T/T	C/C	G/G	T/G	G/G	G/G
NA18945	G/G	T/T	A/A	G/G	G/G	A/A
NA18946	T/G	C/T	G/A	G/G	G/G	G/A
NA18947	T/G	C/T	G/A	G/G	T/G	G/A
NA18948	G/G	T/T	A/A	G/G	G/G	A/A

```
## [1] "The number of individuals is 104"
```

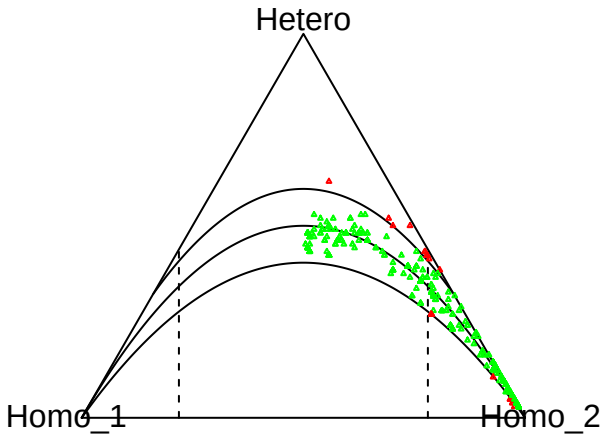
```
## [1] "The number of variants is 543"
```

```
## [1] "Percentage of the data is missing is 0%"
```

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?

Table 2: Genome counts

Homo_1	Hetero	Homo_2
3	28	73
8	28	68
8	28	68
0	15	89
17	46	41
8	28	68
8	28	68
0	3	101
8	28	68
12	46	46




```
## [1] "The number of variants SNPs for which we reject HWE is: 33"
```

```
## [1] "Percentage of out of equilibrium SNPs is: 6.077%"
```

From the ternary plot we see that almost every SNPs is in equilibrium, as we see from the χ^2 in the 94% of the cases we accept the null hypotesis of being in equilibrium. There is still a 6% of cases where we reject the null hypotesis, but as the plot shows they are still very near to the confidence bands.

Notes about notation: We defined as Homo_1 and Homo_2 the homozigotes genotypes (since we do not have always 'AA' and 'BB') and as Hetero the heterozygote. The graph is unbalanced to the right because (we guess) in the .bim file the rarest allele is displayed before.

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?

```
##
## 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
```

We see that we have a absolute deviation from the indipendence of 0.055, that once standardized become a D' of almost one, meaning that the two SNPs are coinherited almost always. So the p-value is very small (of the order of 10^{-6}) and we can strongly reject the null hypotesis of the two SNPs being indipendent.

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?

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

We see that we have a absolute deviation from the indipendence of 0.007, that once standardized become a D' of 0.179. The p-value is 0.189 and we can't reject the null hypotesis of the two SNPs being indipendent, even with a not-so-strict alpha of 0.1.

Table 3: Genotypes

rs2894715	rs34684677	rs998302
G/T	T/G	G/G
G/T	G/G	G/G
G/T	G/G	T/G
G/G	G/G	G/G
G/G	G/G	G/G
T/T	T/T	G/G

```
## [1] "positions of genomes rs34684677, rs2894715, rs998302:"

## [1] "columns in our dataset: 2, 11, 501"

## [1] "basepair: 114400288, 114402794, 114687416"
```

Genetic variants that are physically close on a chromosome typically have high correlations, so we can suppose this is why rs998302 has weaker correlation than rs2894715.

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

$$D = p_{AB} - p_A p_B$$

pA and pB can be estimated by the sample allele frequencies $\hat{p}_A$ and $\hat{p}_B$

$$p_{AB} = D + p_A p_B$$

So the most common haplotype is G T (G from $rs34684677$ and T from $rs2894715$).

rs34684677	rs2894715	
	T	G
G	0.5000741	0.3364644
T	0.1633875	0.0000741

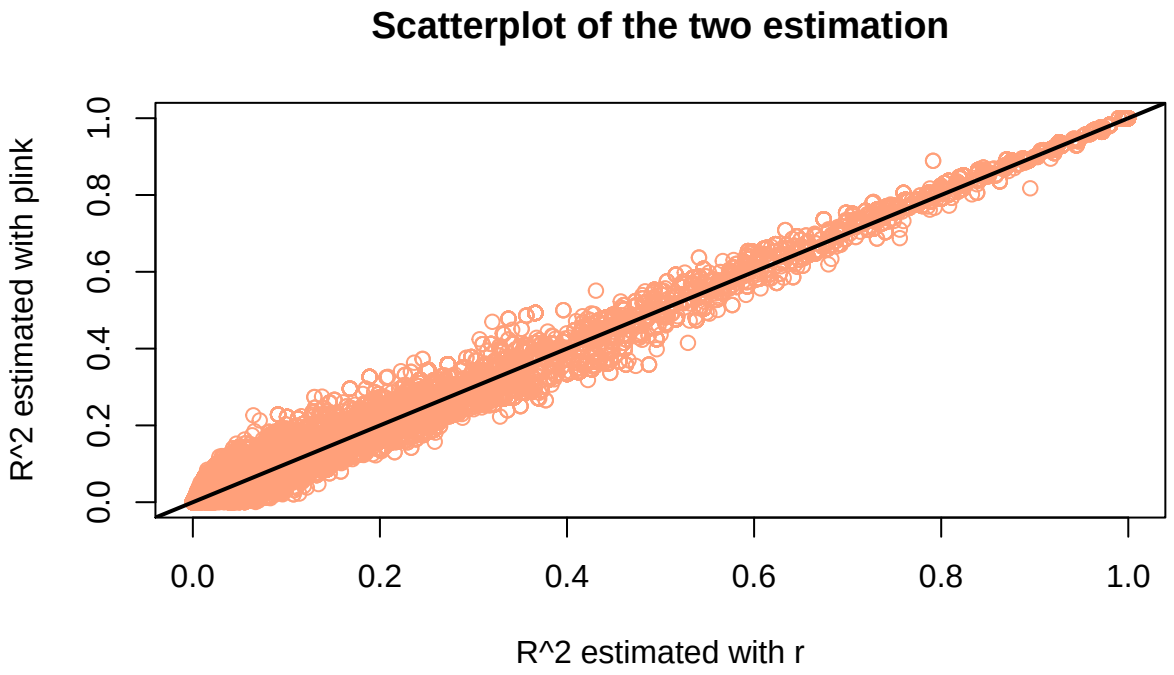
7 Compute the LD statistics R2 for all the marker pairs in this data base, using the LD function of the packages genetics. Also compute an alternative estimate of R2 obtained by using the PLINK program. Make a scatter plot for R's LD estimates against PLINK's LD estimates. Are they identical or do they at least correlate? What's the difference between these two estimators? Which estimator would your prefer and why?

```
## [1] "LD function R^2 estimates using package genetics"

##  [1] 0.72748082 0.72748082 0.99933883 0.39715349 0.28922765 0.28922765
##  [7] 0.12183886 0.00661790 0.00661790 0.04841817 0.72748082 0.99933883
## [13] 0.99933883 0.28922765 0.00661790 0.72748082 0.99933883 0.99933883
## [19] 0.28922765 0.00661790

## [1] "LD function R^2 estimates using plink"

##  [1] 0.7337180 0.7337180 1.0000000 0.3946210 0.3509320 0.3509320 0.1141820
##  [8] 0.0108172 0.0108172 0.0455501 0.7337180 1.0000000 1.0000000 0.3509320
## [15] 0.0108172 0.7337180 1.0000000 1.0000000 0.3509320 0.0108172
```

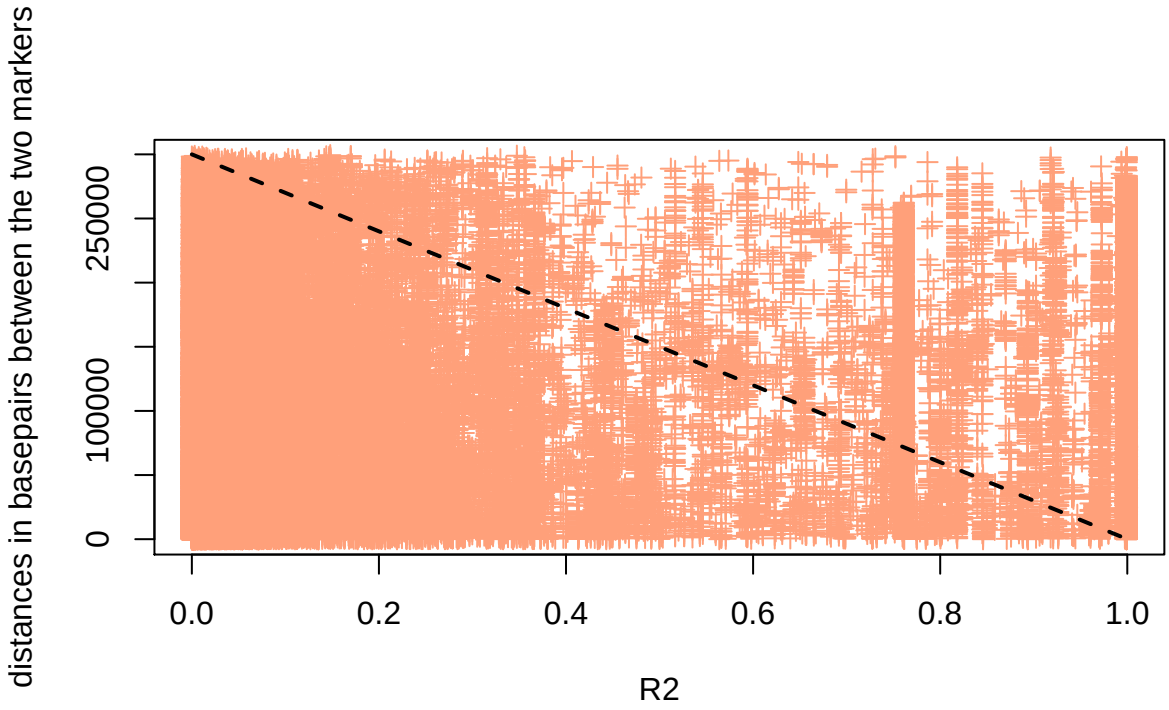


```
## [1] "The correlation coefficient is: 0.99"
```

They are not identical but actually they do correlate almost perfectly.

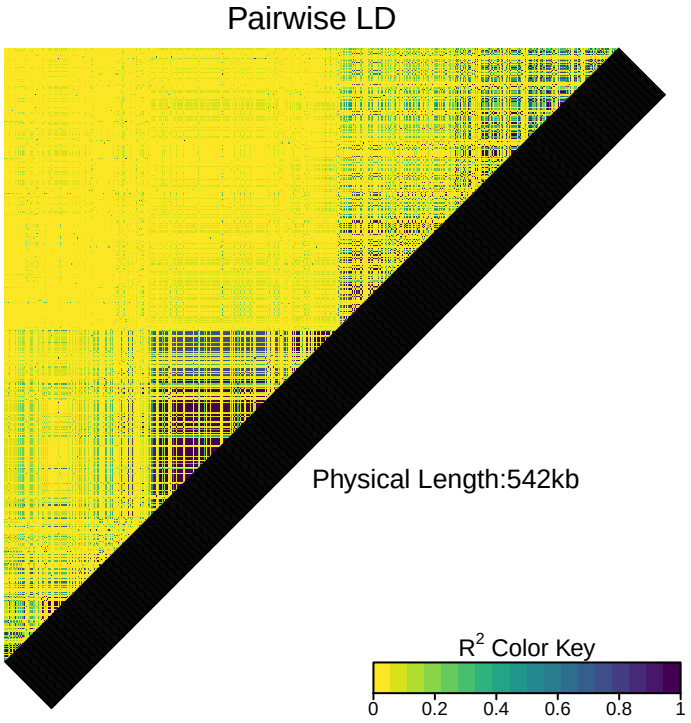
The first estimator evaluates precisely the LD statistic of each pair, but takes a lot of time. The plink estimator is almost immediate so we can suppose that there is some optimization that allows us to reduce the calculation time, but as we can see the cost is almost 0 because the statistics are almost the same, due to this, we can say that we prefer the second estimator.

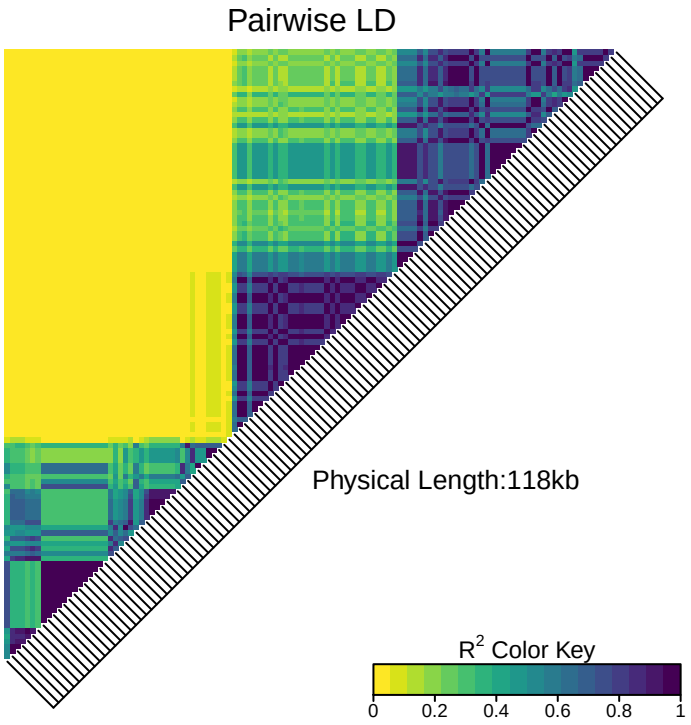
8 Compute a distance matrix with the distance in base pairs between all possible pairs of SNPs, using the basepair position of each SNP given in the .bim file. Make a plot of R's R2 statistics against the distance (expressed as the number of basepairs) between the markers. Comment on your results.



Genetic variants that are physically close on a chromosome typically have high correlations (as we can see from the plot) but this is not a strong correlation since we have a lot of values out of our expectations; especially we have a lot of close pairs with weak correlation. As we will see in next paragraphs, pruning the SNPs with small minor allele frequencies can help in reducing this behaviour.

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?



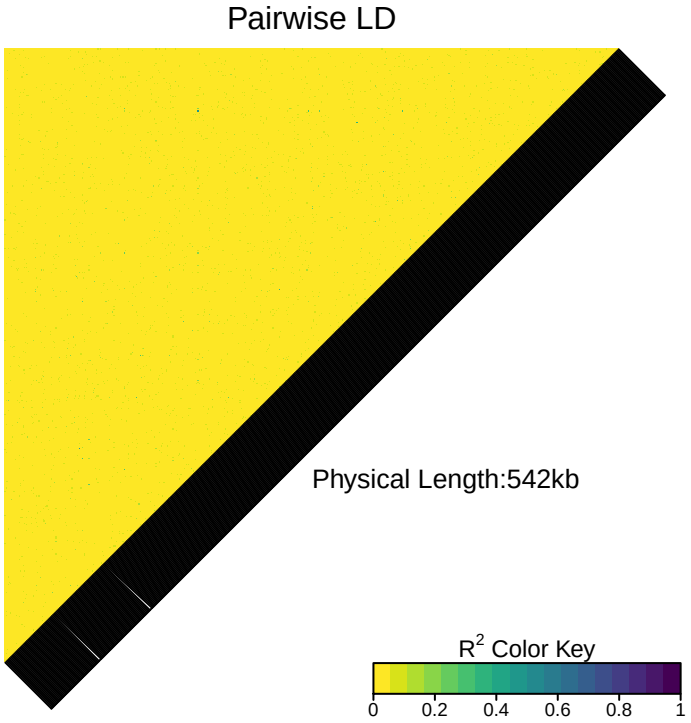


The difference is clear: flitering out with small MAFs the blocks that we can barely see in the first plot become more visible since the correlation are a lot stronger.

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 see two enormous blocks with an r of minimum 0.4 that cut the data in two clusters. In the first cluster (up-right) we see two really strong blocks. In the other big cluster we can see three smaller strong blocks.

11 Simulate independent SNPs under the assumption of Hardy-Weinberg equilibrium. 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



As we could expect, simulating this way we don't have any correlation between variables, since we for construcion we are simulating each SNPs for each individual independently.