Bioinformatics and Statistical Genetics (thaught by Jan Graffelman), HW3, Linkage Disequilibrium

Sten-Oliver Salumaa, Denys Kovalenko

8 December 2017

Task1

The file ABO-CHB.rda contains genotype information of individuals of a Chinese population of unrelated individuals. The genotype information concerns SNPs the ABO bloodgroup region, located on chromosome number 9. The file contains genotype information (Z, individuals in columns, SNPs in rows), the physical position of each SNP (pos) and the alleles for each SNP (alleles). Load this data into the R environment.

Data loaded.

Task2 (1p)

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

```
## [1] "Individual count:"
```

```
## [1] 45
```

```
## [1] "SNP count:"
```

```
## [1] 45
```

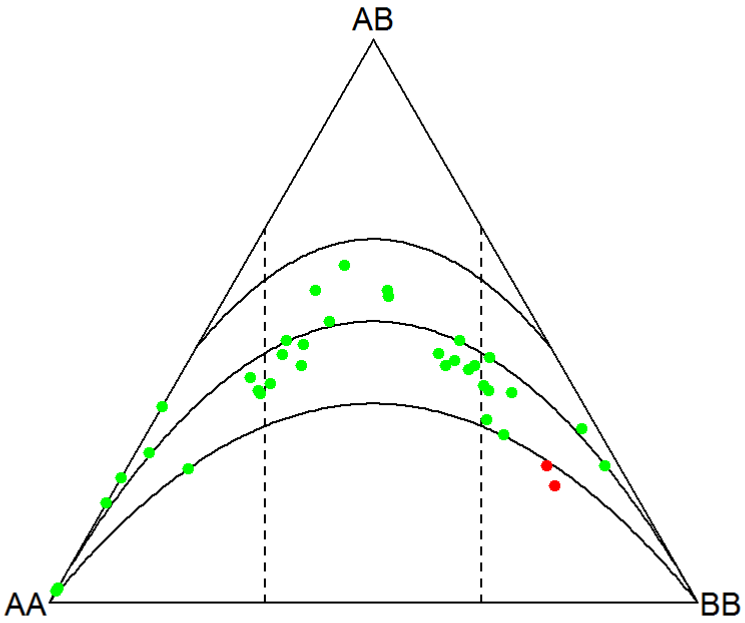
What percentage of the data is missing?

```
## [1] "Percentage of data missing:"
```

```
## [1] 3.012346
```

Task3 (1p)

Depict all SNPs simultaeneously in a ternary plot, and comment on your result. Do you believe Hardy-Weinberg equilibrium is tenable for the markers in this database?



Yes, it seems that HW equilibrium is tenable for this database since almost all of the data points reside in the equilibrium boundaries.

Task4 (1p)

Using the function LD from the genetics package, compute the LD statistic D for the first two SNPs in the database. Is there significant association between these two SNPs?

```
##
## Pairwise LD
## -----------
##                       D        D'       Corr
## Estimates: -0.01600862 0.9974604 -0.1449029
##
##               X^2  P-value  N
## LD Test: 1.889716 0.169234 45
```

```
## [1] "D statistics:"
```

```
## [1] -0.01600862
```

D' is close to 1, so first assumption is that we might have high LD. But we think there is no significant association between these two SNPs because the p-value is too high (above our threshold of 5 %)

Task5 (2p)

Given your previous estimate of D, and using the formulae from the lecture slides, compute the statistics D', Chi2, R2 and r by hand for the first pair of SNPs. Do your results coincide with those obtained by the LD function? Can you explain possible differences?

```
## [1] "D:"
```

```
## [1] -0.01600862
```

```
## [1] "D':"
```

```
## [1] -0.9804904
```

```
## [1] "r**2: "
```

```
## [1] 0.02063208
```

```
## [1] "X**2: "
```

```
## [1] 1.856887
```

```
## [1] "r: "
```

```
## [1] 0.1436387
```

Differences for D occur because we used a small sample for our calculations (we read that with small sample size D' can be inflated).

Task 6 (2p)

Given your previous estimate of D, infer the haplotype frequencies. Which haplotype is the most common?

```
## [1] "GG frequency: "
```

```
## [1] 0.744763
```

```
## [1] "AA frequency: "
```

```
## [1] 0.0003185363
```

```
## [1] "GA frequency: "
```

```
## [1] 0.1274592
```

```
## [1] "AG frequency: "
```

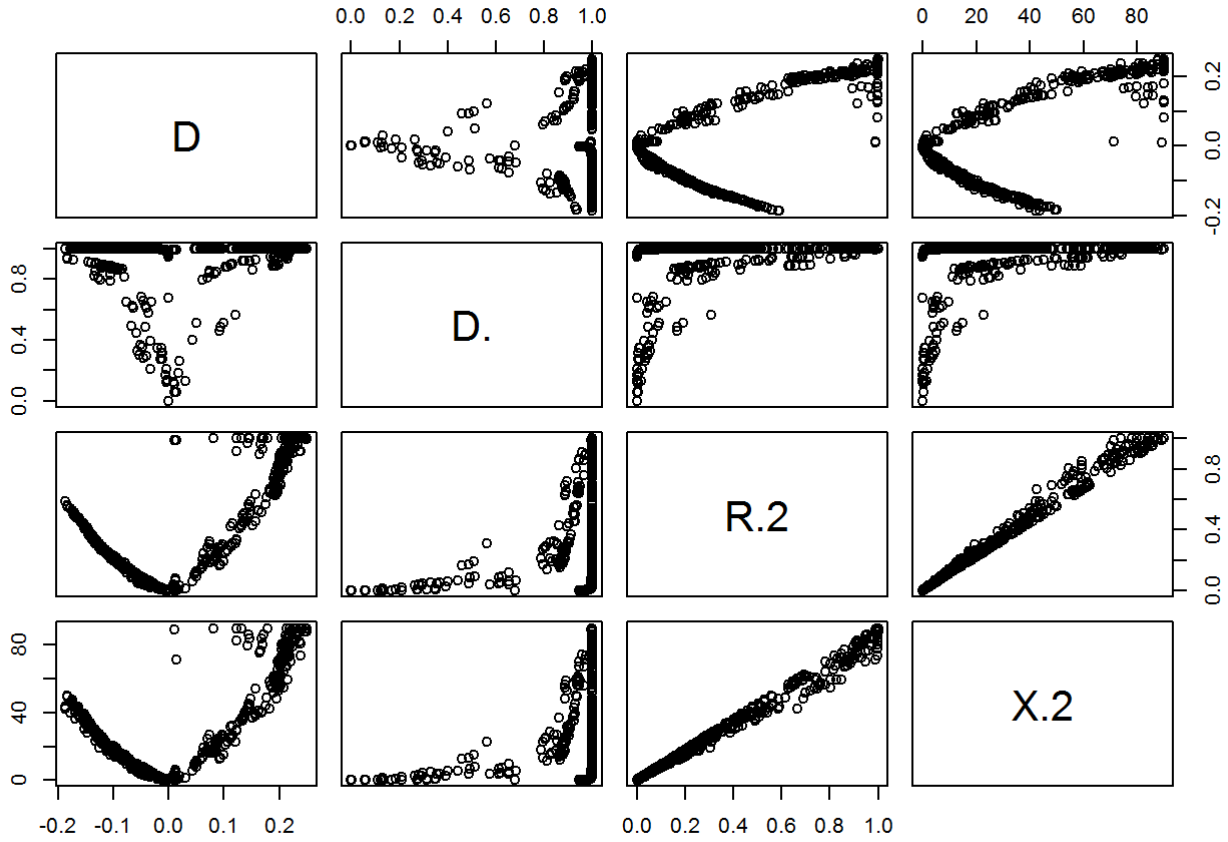
```
## [1] 0.1274592
```

```
## [1] 0.744763
```

Most common is GG haplotype with probability of 0.744763.

Task 7 (2p)

Compute 4 LD statistics for all the marker pairs in this data base (D, D', ï¿½2 and R2). Make a scatterplot matrix of these. Is there an exact linear relationship between ï¿½2 and R2 ? Why (not) so?



There is linear relationship between X^2 and R^2 (line in form y = multiplier*x). By definition, X2 and R2 are tied by equation R2=X2/2n (number of chromosomes)

Task 8. (2p)

Compute a distance matrix with the distance in base pairs between all possible pairs of SNPs. Make a plot of the R2 statistics against the distance between the markers. Comment on your results

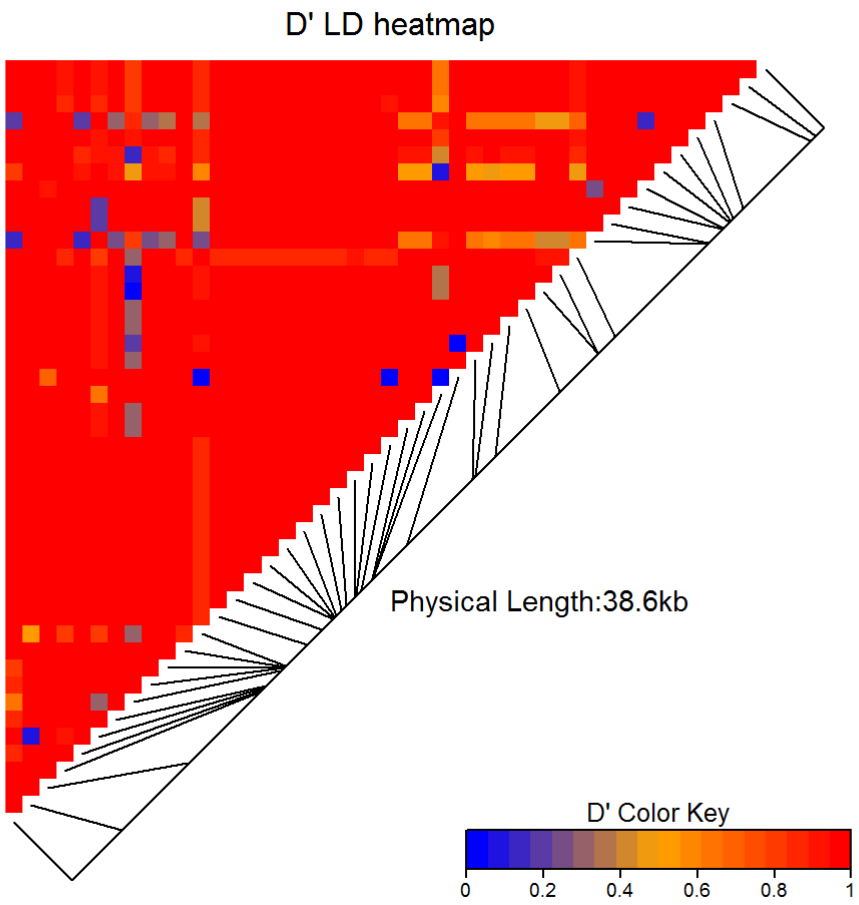
```
##  [1] 0.020996843 0.001604921 0.082446540 0.080404806 0.108551812
##  [6] 0.008845405 0.323189809 0.045851091 0.003663188 0.175521582
```

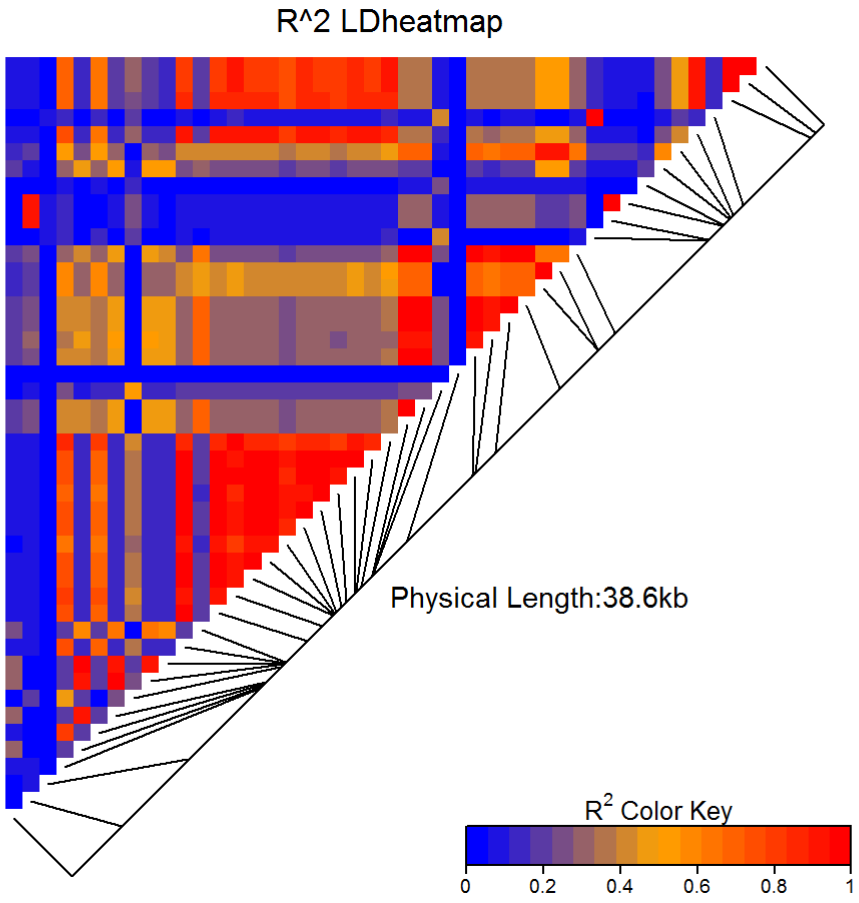
```
##  [1] -2564 -5975 -3411 -9663 -7099 -3688 -9689 -7125 -3714   -26
```

There is correlations between distance and R^2 - the lower the distance, the higher defree of correlation between markers, which agrees with intuition and theory

Task 9. (2p)

Make two LD heatmaps of the markers in this database, one using the R2 statistic and one using the D’ statistic, and use the positional information on the markers. Are the results consistent?

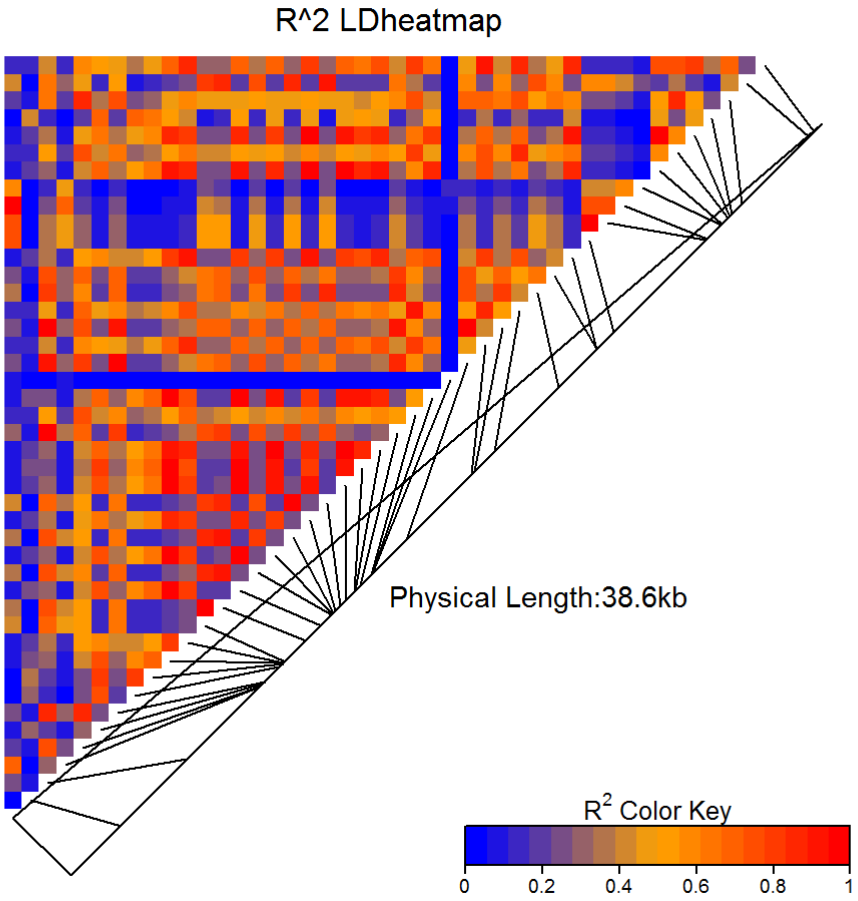


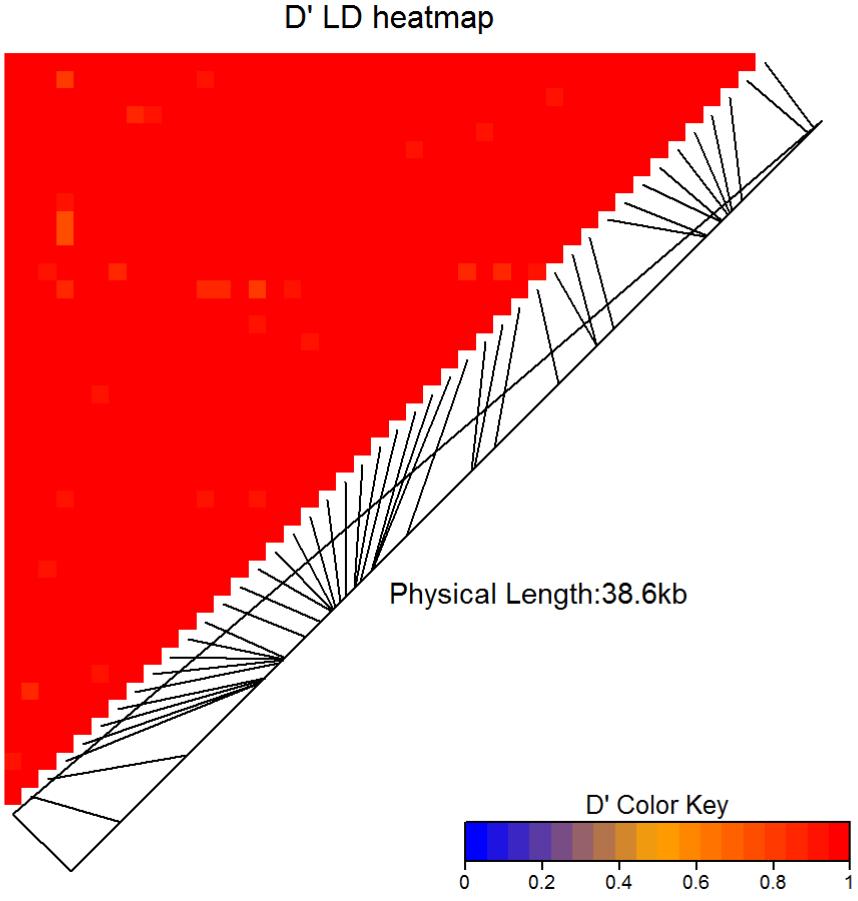


Results are incosistent - with D' LDHeatMap we can see that D's are uniformly distributed and are close to 1 for most of SNP pairs, while with R^2 we see several clusters of closely positioned SNPS to have high R^2. As D' is highly dependent on allele frequencies, R^2 shows regions of LD more clearly

Task 10. (2p)

Simulate 45 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 two LD heatmaps of the simulated SNPs, one using R^2 and one using D'. Compare these to the LD heatmap of the ABO region. What do you observe? State your conclusions





The first LD heatmap using measure of D' look similar to the AB0 region data. The differences step in on the second heatmap using R^2 data. The AB0 region SNPs seem to have clusters on the heatmap - regions that clearly influence each other. But generated data doesn't show any clear regions. So we don't see LD in

generated data, as expected.

Task 11 (1p)

Do you think there is strong or weak LD for the ABO region you just studied? Explain your opinion

There is a strong LD for the AB0 region. We see high D’ values which means that (almost) all SNPs of this region are strongly associated. Also, we see some low R^2 values which doesn’t directly mean that there is weak LD but might also mean that in some loci - although having strong association - are represented with low quantities. But we can see regions even in R^2 LD heatmap that have high values and thus showing high LD (same region on D’ heatmap is high-valued as well).