Practical 03 SG: LD and Haplotype estimation

Kathryn Weissman & Irene Fernández

2022-11-24

LD (15p.)

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).

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?

- There are 104 individuals in the database.
- There are 544 SNPs in the database.
- 0% of the data is missing.

```
file <- "~/Downloads/FOXP2/FOXP2.dat"
data <- fread(file, header = TRUE)
data[1:10, 1:10]
```

```
##          id rs34684677 rs1839115 rs4727804 rs4727805 rs200888633 rs12534908
##  1: NA18939        T/G       C/T       G/A       T/G         T/G        G/A
##  2: NA18940        G/G       T/T       A/A       G/G         T/G        A/A
##  3: NA18941        G/G       T/T       A/A       G/G         T/G        A/A
##  4: NA18942        G/G       T/T       A/A       G/G         T/T        A/A
##  5: NA18943        G/G       T/T       A/A       G/G         T/T        A/A
##  6: NA18944        T/T       C/C       G/G       T/G         G/G        G/G
##  7: NA18945        G/G       T/T       A/A       G/G         G/G        A/A
##  8: NA18946        T/G       C/T       G/A       G/G         G/G        G/A
##  9: NA18947        T/G       C/T       G/A       G/G         T/G        G/A
## 10: NA18948        G/G       T/T       A/A       G/G         G/G        A/A
##     rs12533049 rs77861356 rs6945561
##  1:        C/T        T/T       C/T
##  2:        T/T        T/T       T/T
```

```
##  3:        T/T        T/T       T/T
##  4:        T/T        T/T       T/T
##  5:        T/T        T/T       T/T
##  6:        C/C        T/T       C/C
##  7:        T/T        T/T       T/T
##  8:        C/T        T/T       C/T
##  9:        C/T        T/T       C/T
## 10:        T/T        A/T       T/T
```

```
n <- nrow(data); n # number of samples
```

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

```
p <- ncol(data); p # number of variables
```

```
## [1] 544
```

```
perc.mis <- 100*sum(is.na(data))/(n*p); perc.mis # percentage of missing data overall
```

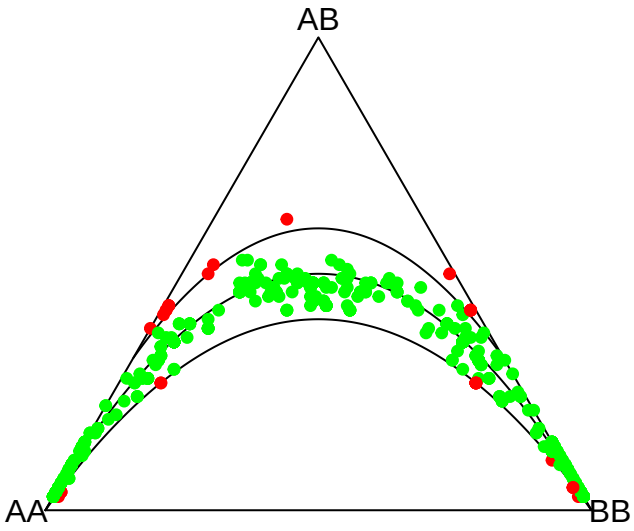
```
## [1] 0
```

3. (1p) Determine the genotype counts for each SNP, and depict all SNPs simultaneously 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).

- The green dots represent SNPs that are in equilibrium, and the red dots represent SNPs that are out of equilibrium. in the data set where the B allele frequency is greater than 0.5.
- Using an ordinary chi-square test we reject Hardy-Weinberg equilibrium for 33 variants.
- *Note:* In order to create the genotype counts matrix, we create a function similar to the one of the anterior practical.

```
genotype_matrix <- matrix(ncol = 3, nrow = 0, dimnames = list(NULL, c("AA", "AB", "BB")))
for(i in colnames(data[,-1])){
snp <- genotype(data[[i]], reorder = "ascii") # reorder is to use as specific order for the alleles
snp.counts <- c(sum(snp==paste(allele.names(snp)[1], allele.names(snp)[1], sep = "/")),
                sum(snp==paste(allele.names(snp)[1], allele.names(snp)[2], sep = "/")),
                sum(snp==paste(allele.names(snp)[2], allele.names(snp)[2], sep = "/")))
genotype_matrix <- rbind(genotype_matrix, snp.counts)
}
rownames(genotype_matrix) <- colnames(data[,-1])
```

```
ternaryPlot <- HWTernaryPlot(genotype_matrix)
```




```
results_ChiSq <- HWChisqStats(genotype_matrix, pvalues=TRUE)
sum(results_ChiSq <= .05)
```

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

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?

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?

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

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?

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.

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?

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?

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.

Haplotype estimation (10p.)

1. Apolipoprotein E (APOE) is a protein involved in Alzheimer's disease. The corresponding gene *APOE* has been mapped to chromosome 19. The file `APOE.dat` contains genotype information of unrelated individuals for a set of SNPs in this gene. Load this data into the R environment. `APOE.zip` contains the corresponding `.bim`, `.fam` and `.bed` files. You can use the `.bim` file to obtain information about the alleles of each polymorphism.

2. (1p) How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

3. (1p) Assuming all SNPs are bi-allelic, how many haplotypes can theoretically be found for this data set?

4. (2p) Estimate haplotype frequencies using the `haplo.stats` package (set the minimum posterior probability to 0.001). How many haplotypes do you find? List the estimated probabilities in decreasing order. Which haplotype number is the most common?

5. (2p) Is the haplotypic constitution of any of the individuals in the database ambiguous or uncertain? For how many? What is the most likely haplotypic constitution of individual NA20763? (identify the constitution by the corresponding haplotype numbers).

6. (1p) Suppose we would delete polymorphism rs374311741 from the database prior to haplotype estimation. Would this affect the results obtained? Justify your answer.

7. (1p) Remove all genetic variants that have a minor allele frequency below 0.10 from the database, and re-run `haplo.em`. How does this affect the number of haplotypes?

8. (2p) We could consider the newly created haplotypes in our last run of `haplo.em` as the alleles of a new superlocus. Which is, under the assumption of Hardy-Weinberg equilibrium, the most likely genotype at this new locus? What is the probability of this genotype? Which genotype is the second most likely, and what is its probability?