BSG-MDS practical 3 Statistical Genetics

Name Surname Name Surname

21/11/2023, submission deadline 28/11/2023

Resolve the following exercise in groups of two students. Write the R scripts, 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, or generate a solution pdf file with R Markdown. Take care to number your answers 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 submission deadline.

You can make use of the R-package `genetics` and `HardyWeinberg` to compute your answers. The datasets can be downloaded from the web page of the course at raco.fib.upc.edu.

Linkage Disequilibrium (7p)

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 in 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 the `read.table` command in 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).

1. (0.5p) 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?
2. (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?
3. (1p) Given your previous estimate of D for SNPs rs34684677 and rs2894715, infer the haplotype frequencies. Which haplotype is the most common?
4. (1p) Determine the genotype counts for each SNP. For how many variants do you reject Hardy-Weinberg equilibrium using an ordinary chi-square test without continuity correction? Is this what you would expect by chance? (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).
5. (1p) Compute the LD 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. Extract the R^2 statistics and make an LD heatmap (hint: you can use the command image) using the R^2 statistic.
6. (1,5p) 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?

7. (1p) 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.

Haplotype estimation (3p)

Apolipoprotein E (Apo-E) 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. You will find the file `APOE.dat` in the zip folder called APOE.zip. In the zip foder you will find the corresponding `.bim`, `.fam` and `.bed` files for the APOE gene. Recall that you can use the `.bim` file to obtain information about the alleles of each polymorphism.

1. (0,5p) How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

2. (0,5p) Assuming that all SNPs are bi-allelic, how many haplotypes can theoretically be found for this data set?

3. (1p) Estimate haplotype frequencies using the `haplo.em` function that you will find in the `haplo.stats` package. How many haplotypes do you find? List the estimated probabilities in decreasing order. Which haplotype number is the most common?

4. (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? Comment on your results.