BSG-MDS practical 5 Statistical Genetics

Name Surname Name Surname

12/12/2023, submission deadline 19/12/2023

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 other packages) to compute your answers, as you please. The datasets can be downloaded from the web page of the course at raco.fib.upc.edu.

Relatedness analysis (10p)

The file `YRI06.raw` contains SNPs of a Yoruba population consisting of parent-offspring trios (2 parents and 1 child). We wish to investigate if the individual's genotype is consistent with the specified family relationships.

Load the data into the R environment, with the `fread` instruction of the package `data.table`, which is more efficient for reading large data files. Note that the genetic information starts on the 7th column. The column 3 and 4 contain information about the family relationship of the participants. The genetic variants are identified by an "rs" identifier and the genotypes are coded in the (0, 1, 2) format with 0=AA, 1=AB, 2=BB.

1. (0.5p) Load the `YRI06.raw` 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) Compute, for each pair of individuals (and report the first 5), the mean m of the number of alleles shared and the standard deviation s of the number of alleles shared.

3. (1.5p) Compute, for each pair of individuals (and report the first 5), the fraction of variants for which the individuals share 0 alleles (p_0), and the fraction of variants for which the individuals share 2 alleles (p_2). Check if $m = 1 - p_0 + p_2$ holds.

4. (2p) Plot m against s and plot p_0 against p_2. Comment on the results.

5. (2p) Plot m against s and use the pedigree information of the `YRI06.raw` file to label the data points in the scatterplot. Recall that column 3 and 4 from the `YRI06.raw` contain information about the family relationship of the participants. Create two labels: one for individuals that have a parent-offspring relationship and another one for unrelated individuals. Comment on the results.

6. (2p) Use the package `SNPRelate` to estimate the IBD probabilities, and plot the probabilities of sharing 0 and 1 IBD alleles (k0 and k1) for all pairs of individuals. Use the pedigree information of the `YRI06.raw` file to label the data points in the scatterplot (same as before, one colour for parent-offspring relationship and another colour for unrelated individuals).

7. (1p) Do you think the family relationships between all individuals were correctly specified?