BSG-MDS practical 2 Statistical Genetics

Name Surname Name Surname

14/11/2023, submission deadline 21/11/2023

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

Hardy Weinberg Equilibrium (10p)

The file `TSIChr22v4.raw` contains genotype information of individuals from Tuscany in Italy, taken from the 1,000 Genomes project. The datafile contains all single nucleotide polymorphisms on chromosome 22 for which complete information is available.

Load this data into the R environment, with the `fread` instruction of the package `data.table`, which is more efficient for reading large datafiles. The first six columns contain non-genetical information. Create a dataframe that only contains the genetic information that is in and beyond the 7th column. Notice that the genetic variants are identifed by an "rs" identifier. The genetic data is coded in the (0, 1, 2) format with 0=AA, 1=AB, 2=BB.

1. (0.5p) How many variants are there in this database? What percentage of the data is missing?

2. (0.5p) Calculate the percentage of monomorphic variants. Exclude all monomorphics from the database for all posterior computations of the practical. How many variants do remain in your database?

3. (2p) Extract polymorphism rs587756191_T from the data, and determine its genotype counts. Apply a chi-square test for Hardy-Weinberg equilibrium, with and without continuity correction. Also try an exact test, and a permutation test. You can use thee functions `HWChisq`, `HWExact` and `HWPerm` for this purpose. Do you think this variant is in equilibrium? Argue your answer.

4. (1p) Determine the genotype counts for all polymorphic variants, and store them in a p × 3 matrix.

5. (1p) Apply an exact test for Hardy-Weinberg equilibrium to each SNP. You can use function `HWExactStats` for fast computation. What is the percentage of significant SNPs (use $\alpha = 0.05$)? Is this the number of markers that you would expect to be out of equilibrium by the effect of chance alone?

6. (1p) Which SNP is most significant according to the exact test results? Give its genotype counts. In which sense is this genotypic composition unusual?

7. (2p) Compute the inbreeding coefficient (f) for each SNP, and make a histogram of f. You can use function `HWf` for this purpose. Give descriptive statistics (mean, standard deviation, etc) of f calculated over the set of SNPs. What distribution do you expect f to follow theoretically? Use a probability plot to confirm your idea.

8. (2p) Apply the exact test for HWE to each SNP, using different significant levels. Report the number and percentage of significant variants using an exac test for HWE with $\alpha = 0.10$, 0.05, 0.01 and 0.001. State your conclusions.