Practical 01 SG: Descriptive analysis of genetic markers

Write here your names and surnames

Hand-in: 07/11/2022

Resolve the following exercise in groups of two students. 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 hand-in date.

You can make use of the R-package **genetics** (and other packages) to compute your answers, as you please. The first part of the practical is dedicated to the descriptive analysis of SNP data, whereas the second part is dedicated to the analysis of STR data. The datasets can be downloaded by clicking on their file names given below.

SNP dataset (10p)

1. The file TSICHR22RAW.raw contains a sample of SNPs on chromosome 22 of a sample of Italian individuals in Toscani in Italy. This data has been extracted from the 1000 genomes project at www.internationalgenome.org .
2. Load this data into the R environment, with the `read.table` instruction. The first six columns contain non-genetical information. Extract the variables individual ID (the second column IID) and the sex of the individual (the 5th column sex). 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.
3. (1p) How many variants are there in this database? What percentage of the data is missing? How many individuals in the database are males and how many are females?
4. (1p) 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?
5. (1p) Report the genotype counts and the minor allele count of polymorphism rs8138488_C, and calculate the MAF of this variant.
6. (2p) Compute the minor allele frequencies (MAF) for all markers, and make a histogram of it. Does the MAF follow a uniform distribution? What percentage of the markers have a MAF below 0.05? And below 0.01? Can you explain the observed pattern?
7. (2p) Calculate the minor allele frequency for males and for females and present a scatterplot of these variables. What do you observe? Calculate and report their correlation coefficient.
8. (1p) Calculate the observed heterozygosity (H_o), and make a histogram of it. What is, theoretically, the range of variation of this statistic?
9. (2p) Compute for each marker its expected heterozygosity (H_e), where the expected heterozygosity for a bi-allelic marker is defined as $1 - \sum_{i=1}^{k} p_i^2$, where p_i is the frequency of the ith allele. Make a

histogram of the expected heterozygosity. What is, theoretically, the range of variation of this statistic? What is the average of H_e for this database?

2 STR dataset

1. The object **NistSTRs** of the R package **HardyWeinberg** contains a set of STRs of individuals of Caucasian ancestry, which can be loaded with the instructions `library(HardyWeinberg)` and `data(NistSTRs)`. The rownames of the object consist of identifiers for each individual. Successive columns represent the two alleles of an individual for each STR. Note there exist *fractional alleles* (like 14.3) that indicate the particular STR sequence is repeated in-between a certain numbers of times. These fractional alleles are regarded as separate alleles (e.g. 14.3 is different from 14 and 15).
2. (1p) How many individuals and how many STRs contains the database?
3. (2p) Write a function that determines the number of alleles for a STR. Determine the number of alleles for each STR in the database. Compute basic descriptive statistics of the number of alleles (mean, standard deviation, median, minimum, maximum).
4. (2p) Make a table with the number of STRs for a given number of alleles and present a barplot of the number STRs in each category. What is the most common number of alleles for an STR?
5. (2p) Compute the expected heterozygosity for each STR. Make a histogram of the expected heterozygosity over all STRS. Compute the average expected heterozygosity over all STRs.
6. (1p) Calculate also the observed heterozygosity for each STR. Plot observed against expected heterozygosity, using all STRs. What do you observe?
7. (2p) Compare, overall, the results you obtained for the SNP database with those you obtained for the STR database. What differences do you observe between these two types of genetic markers?