Bioinformatics and Statistical Genetics: Relatedness analysis

Write your names and surnames

21 de December, 2019

Hand/in before: 06/01/2020 **(15p)**

Introduction: In this practical we analyse a genetic dataset for the presence of possible cryptic relatedness. 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 may generate a solution .pdf file using this file with R Markdown. Alternatively, you can also write your solution in a word or Latex document and generate a .pdf file with your solution. You can use R packages **MASS**, **data.table** and others for the computations. Take care to number your answer exactly as in this exercise. Upload your solution to the web page of the course at raco.fib.upc.edu no later than the hand-in date.

Instructions:

1. The file CHD.zip contains genotype information, in the form of PLINK files **chd.fam**, **chd.bed** and **chd.bim**. The files contain genetic information of 109 presumably unrelated individuals of a sample of Chinese in Metropolitan Denver, CO, USA, and corresponds to the CHD sample of the 1,000 Genomes project (www.internationalgenome.org).
2. The **chd.bed** contains the genetic data in binary form. First convert the **.bed** file to a text file, **chd.raw**, with the data in (0, 1, 2) format, using the PLINK instruction:

   ```
   plink --bfile CHD --recodeA --out CHD
   ```

3. (1p) Read the genotype data in (0, 1, 2) format into the R environment. Consult the pedigree information. Are there any documented family relationships for this data set?
4. (2p) Compute the Manhattan distance between the inviduals on the basis of the genetic data. Use classical metric multidimensional scaling to obtain a map of the indivuals. Are the data homogeneous? Identify possible outliers.
5. (2p) Compute the average number of alleles shared between each pair of individuals over all genetic variants. Compute also the corresponding standard deviation. Plot the standard deviation against the mean. Do you think there are any pairs with a close family relationship? How many? Identify the corresponding individuals.
6. (1p) Make a plot of the percentage of variants sharing no alleles versus the percentage of variants sharing two alleles for all pairs of individuals. Do you think there are any pairs with a close family relationship? How many? Identify the corresponding individuals.
7. (1p) Can you identify any obvious family relationships between any pairs? Argue your answer.
8. (2p) Estimate the *Cotterman coefficients* for all pairs using PLINK. Read the coeffients into the R environment and plot the probability of sharing no IBD alleles against the probability of sharing one

IBD allele. Add the theoretical values of the Cotterman coefficients for standard relationships to your plot.

9. (2p) Make a table of pairs for which you suspect that they have a close family relationship, and list their Cotterman coefficients. State your final conclusions about what relationship these pairs probably have.

10. (2p) Is there any relationship between the MDS map you made and the relationships between the individuals? Report your findings.

11. (2p) Which of the three graphics (m, s), (p_0, p_2) or (k_0, k_1) do you like best for identifying relationships? Argue your answer.