Bioinformatics and Statistical Genetics	Haplotype estimation (10p)
Jan Graffelman	Hand-in: 10/12/2019

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. Alternatively, you may generate a solution pdf file with Markdown. You can use R packages `genetics, LDheatmap, haplo.stats` and others for the computations. Take care to number your answers exactly as in this exercise, preferably by copying each requested item into your solution. Upload your solution to the web page of the course at raco.fib.upc.edu no later than the hand-in date.

1. Apolipoprotein E (APOE) 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. APOE.zip contains the corresponding `.bim`, `.fam` and `.bed` files. You can use the `.bim` file to obtain information about the alleles of each polymorphism.

2. (1p) How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

3. (1p) Assuming all SNPs are bi-allelic, how many haplotypes can theoretically be found for this data set?

4. (2p) Estimate haplotype frequencies using the haplo.stats package (set the minimum posterior probability to 0.001). How many haplotypes do you find? List the estimated probabilities in decreasing order. Which haplotype number is the most common?

5. (2p) Is the haplotypic constitution of any of the individuals in the database ambiguous or uncertain? For how many? What is the most likely haplotypic constitution of individual NA20763? (identify the constitution by the corresponding haplotype numbers).

6. (1p) Suppose we would delete polymorphism rs374311741 from the database prior to haplotype estimation. Would this affect the results obtained? Justify your answer.

7. (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?

8. (2p) We could consider the newly created haplotypes in our last run of `haplo.em` as the alleles of a new superlocus. Which is, under the assumption of Hardy-Weinberg equilibrium, the most likely genotype at this new locus? What is the probability of this genotype? Which genotype is the second most likely, and what is its probability?