Bioinformatics and Statistical Genetics	Haplotype estimation
Jan Graffelman	November 25, 2018

Name: .....

Name: .....

You can make use of the R-packages **haplo.stats, genetics, HardyWeinberg** (and other packages) to compute your answers. Prepare a .pdf file with all your answers and figures. Send your work by email to the course instructor (`jan.graffelman@upc.edu`) no later than Wednesday 5^{th} of December 2018.

1. Myoglobin is an oxygen-binding protein found in muscle tissue. The protein is encoded by the MB gene, which resides on the long arm of chromosome 22. The file MB.rda contains genotype information of unrelated individuals for a set of SNPs in the MB gene. The file contains genotype information in object `Y`. Load this data into the R environment.

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. (1p) Estimate haplotype frequencies using the haplo.stats package (set the minimum posterior probability to 0.001). How many haplotypes do you find? List the haplotypes and their estimated probabilities. Which haplotype is the most common? .....

   .....

5. (2p) Is the haplotypic constitution of any of the individuals in the database ambiguous or uncertain? If so, for which individuals? What is, in case, the most likely haplotypic constitution of any possibly uncertain individuals? .....

   .....

6. (1p) Suppose we would delete SNP rs5999890 from the database prior to haplotype estimation. Would this affect the results obtained? Justify your answer. Delete this SNP from the database and estimate again the haplotype frequencies. List the haplotypes and their estimated frequencies.
   .....

7. (1p) Individual NA18525 has missing values for several variants. Does this mean that that `haplo.em` cannot estimate the constitution of this individual? Investigate the haplotypic constitution of this individual, and what implication the estimation has for his/her missing values .....

   .....

8. (2p) We could consider the newly created haplotypes as the alleles of a new locus. 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? .....

.....

9. (1p) Simulate a set of independent markers using the the multinomial distribution (R function `rmultinom`) that mimicks the Myoglobin data in terms of sample size, number of SNPs and minor allele frequencies, assuming HardyWeinberg equilibrium (that is, simulate the markers with multinomial probabilities $p^2, 2pq$ and q^2, where p is the observed minor allele frequency) Create haplotypes on the basis of the simulated data. Do you find the same number of haplotypes? Can you explain the difference? .....

.....