Bioinformatics and Statistical Genetics	Linkage Disequilibrium (16p)
Jan Graffelman	November 26, 2017

Name: .....

Name: .....

You can make use of the R-packages **HardyWeinberg** and **genetics** (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`) before 7/12/2017.

1. The file `ABO-CHB.rda` contains genotype information of individuals of a Chinese population of unrelated individuals. The genotype information concerns SNPs the ABO bloodgroup region, located on chromosome number 9. The file contains genotype information (`Z`, individuals in columns, SNPs in rows), the physical position of each SNP (`pos`) and the alleles for each SNP (`alleles`). 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) Depict all SNPs simultaeneously in a ternary plot, and comment on your result. Do you believe Hardy-Weinberg equilibrium is tenable for the markers in this database? .....

4. (1p) Using the function `LD` from the genetics package, compute the LD statistic D for the first two SNPs in the database. Is there significant association between these two SNPs? .....

5. (2p) Given your previous estimate of D, and using the formulae from the lecture slides, compute the statistics D', χ^2, R^2 and r by hand for the first pair of SNPs. Do your results coincide with those obtained by the `LD` function? Can you explain possible differences? .....

6. (2p) Given your previous estimate of D, infer the haplotype frequencies. Which haplotype is the most common? .....

7. (2p) Compute 4 LD statistics for all the marker pairs in this data base (D, D', χ^2 and R^2). Make a scatterplot matrix of these. Is there an exact linear relationship between χ^2 and R^2? Why (not) so? .....

8. (2p) Compute a distance matrix with the distance in base pairs between all possible pairs of SNPs. Make a plot of the R^2 statistics against the distance between the markers. Comment on your results. .....

9. (2p) Make two LD heatmaps of the markers in this database, one using the R^2 statistic and one using the D' statistic, and use the positional information on the markers. Are the results consistent? ..

10. (2p) Simulate 45 independent SNPs under the assumption of Hardy-Weinberg equilibrium. Simulate as many SNPs as you have in your database, and take care to match each SNP in your database with a simulated SNP that has the same sample size and allele frequency. Make two LD heatmaps of the simulated SNPs, one using R^2 and one using D'. Compare these to the LD heatmap of the ABO region. What do you observe? State your conclusions .....

11. (1p) Do you think there is strong or weak LD for the ABO region you just studied? Explain your opinion. .....