Bioinformatics and Statistical Genetics	Linkage Disequilibrium (15p)
Jan Graffelman	Hand-in: 03/12/2019

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, HardyWeinberg, LDheatmap` for the computations. Take care to number your answer 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. The file `FOXP2.zip` contains genetic information of individuals of a Japanese population of unrelated individuals. The genotype information concerns SNPs of the Forkhead box protein P2 (FOXP2) gene region, located the long arm of chromosome number 7. This gene plays an important role in the development of speech and language. The `FOXP2.zip` file contains:

   - `FOXP2.dat`: a text file with the genotype data which can be read in with R.
   - `FOXP2.fam`: a PLINK file with data on the individuals (family id, individual id, ids of parents, sex and phenotype).
   - `FOXP2.bed`: a PLINK file with binary genotype data.
   - `FOXP2.bim`: a PLINK file with data on the genetic variants (chromosome, SNP identifier, basepair position along the chromosome and alleles).

2. (1p) Load the `FOXP2.dat` file into the R environment. How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

3. (1p) Determine the genotype counts for each SNP, and depict all SNPs simultaeneously in a ternary plot, and comment on your result. For how many variants do you reject Hardy-Weinberg equilibrium using an ordinary chi-square test without continuity correction? (hint: you can read the `.bim` in R in order to determine the alleles of each SNP, and use function `MakeCounts` from the HardyWeinberg package to create a matrix of genotype counts).

4. (1p) Using the function `LD` from the genetics package, compute the LD statistic D for the SNPs rs34684677 and rs2894715 of the database. Is there significant association between the alleles of these two SNPs?

5. (2p) Also compute the LD statistic D for the SNPs rs34684677 and rs998302 of the database. Is there significant association between these two SNPs? Is there any reason why rs998302 could have stronger or weaker correlation than rs2894715?

6. (2p) Given your previous estimate of D for SNPs rs34684677 and rs2894715, infer the haplotype frequencies. Which haplotype is the most common?

7. (2p) Compute the LD statistics R^2 for all the marker pairs in this data base, using the `LD` function of the packages `genetics`. Be prepared that this make take a few minutes. Also compute an alternative estimate of R^2 obtained by using the `PLINK` program. For this purpose you should:

   - Download and install PLINK 1.90 from https://www.cog-genomics.org/plink2/
   - Take care to store the files FOXP2.bim, FOXP2.fam and FOXP2.bed in a directory where PLINK can find them.
   - Compute LD estimates with PLINK using:

     ```
     plink -bfile FOXP2 -r2 -matrix -out FOXP2
     ```

     This creates a file with extension `FOXP2.ld` that contains a matrix with all R^2 statistics. Read this file into the R environment.

   Make a scatter plot for R's LD estimates against PLINK's LD estimates. Are they identical or do they at least correlate? What's the difference between these two estimators? Which estimator would your prefer and why?

8. (2p) Compute a distance matrix with the distance in base pairs between all possible pairs of SNPs, using the basepair position of each SNP given in the `.bim` file. Make a plot of R's R^2 statistics against the distance (expressed as the number of basepairs) between the markers. Comment on your results.

9. (2p) Make an LD heatmap of the markers in this database, using the R^2 statistic with the `LD` function. Make another heatmap obtained by filtering out all variants with a MAF below 0.35, and redoing the computations to obtain the R^2 statistics in R. Can you explain any differences observed between the two heatmaps?

10. (1p) Can you distinguish blocks of correlated markers in the area of the FOXP2 gene? How many blocks do you think that *at least* seem to exist?

11. (1p) Simulate independent SNPs under the assumption of Hardy-Weinberg equilibrium, using R's `sample` instruction (`sample(c("AA","AB","BB"),n,replace=TRUE,prob=c(p*p,2*p*q,q*q))`). 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 an LD heatmap of the simulated SNPs, using R^2 as your statistic. Compare the results with the LD heatmap of the FOXP2 region. What do you observe? State your conclusions.