Bioinformatics and Statistical Genetics	Hardy-Weinberg equilibrium (20p)
Iván Galván-Femenía & Jan Graffelman	November 10, 2016

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 (`ivan.galvan@upc.edu`) before the 18^{th} of November 2016.

1. The file CHBChr3-2000.rda contains genotype information (2000 SNPs) of individuals from a Chinese population of unrelated individuals. Load this data into the R environment. The file contains three data objects, `X` (genotype information), `Als` (the possible alleles for each marker) and `Pos` (the position of each marker in base pairs)

2. (2p) Remove all SNPs that consist of missing values only from the database. Also remove all monomorphic SNPs from the data bases. Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. How many SNPs remain? How many SNPs are significant (use $\alpha = 0.05$)? .....

   .....

3. (1p) Do you think the monomorphic markers are in Hardy-Weinberg equilibrium? Argue your answer .....

   .....

4. (1p) How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone? .....

   .....

5. (1p) Apply an Exact test for Hardy-Weinberg equilibrium to each SNP. Use the `pvaluetype="selome"` option. How many SNPs are significant (use $\alpha = 0.05$). Is the result consistent with the chi-square test? .....

   .....

6. (1p) Apply a likelihood ratio test for Hardy-Weinberg equilibrium to each SNP. How many SNPs are significant (use $\alpha = 0.05$). Is the result consistent with the chi-square test? .....

   .....

7. (1p) Apply a permutation test for Hardy-Weinberg equilibrium to each SNP, using the classical chi-square test (without continuity correction) as a test statistic. Reduce the number of permutations

in order to keep computation time within reasonable limits (argument `nperm`). How many SNPs are significant (use $\alpha = 0.05$). Are the result consistent with the chi-square or exact test? .....

.....

8. (1p) Depict all SNPs simultaeneously in a ternary plot, and comment on your result (because many genotype counts repeat, you may use `UniqueGenotypeCounts` to speed up the computations) ....

.....

9. (3p) Make a histogram of the p-values obtained in the chi-square test. What distribution would you expect if HWE would hold for the data set? What distribution do you observe? Also make a Q-Q plot of the p values obtained in the chi-square test against the quantiles of the distribution that you consider relevant. What is your conclusion?. .....

.....

.....

.....

10. (1p) Imagine that for a particular marker the counts of the two homozygotes are accidentally interchanged. Would this affect the statistical tests for HWE? Argue your answer. .....

.....

11. (3p) Compute the inbreeding coefficient ($\hat{f}$) for each SNP, and make a histogram of $\hat{f}$. You can use function `HWf` for this purpose. Give descriptive statistics (mean, standard deviation, etc) of $\hat{f}$ calculated over the set of SNPs. What distribution do you think $\hat{f}$ follows? Use a probability plot to confirm your idea. .....

.....

12. (2p) Make a plot of the observed chi-square statistics agains the inbreeding coefficient ($\hat{f}$). What do you observe? Can you give an equation that relates the two statistics? .....

.....

13. (1p) Make a chi-square probability plot of the observed chi-square statistics against their theoretical quantiles. Does the sample statistic follow a chi-square distribution? .....

14. (1p) Simulate SNPs under the assumption of Hardy-Weinberg equilibrium. Simulate the SNPs of this database, and take care to match each of the SNPs in your database with a simulated SNP that has the same sample size and allele frequency. You can use function `HWData` of the HardyWeinberg package for this purpose. Compare the distribution of the observed chi-square statistics with the distribution of the chi-square statistics of the simulated SNPs by making a Q-Q plot. What do you observe? State your conclusions .....

.....

15. (1p) Do you think genotyping error is a problem for the database you just studied? Explain your opinion. .....

.....