Bioinformatics and Statistical Genetics	Hardy-Weinberg equilibrium (25p)
Jan Graffelman	Hand-in: 25/11/2019

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 `data.table` and `HardyWeinberg` 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 TSIChr22v4.raw contains genotype information of individuals from Tuscany in Italy, taken from the 1,000 Genomes project. The datafile contains all single nucleotide polymorphisms on chromosome 22 for which complete information is available. Load this data into the R environment. Use the `fread` instruction of the package `data.table`, which is more efficient for reading large datafiles. This data is in (0,1,2) format, where 0 and 2 represent the homozygotes AA and BB, and 1 represents the heterozygote AB. The first six leading columns of the data matrix can be ignored, as they do not contain any genetic information.

2. (1p) How many individuals does the database contain, and how many variants? What percentage of the variants is monomorphic? Remove all monomorphic SNPs from the database. How many variants remain in the database?

3. (3p) Extract polymorphism rs587756191_T from the datamatrix, and determine its genotype counts. Apply a chi-square test for Hardy-Weinberg equilibrium, with and without continuity correction. Also try an exact test, and a permutation test. You can use function `HWChisq`, `HWExact` and `HWPerm` for this purpose. Do you think this variant is in equilibrium? Argue your answer.

4. Determine the genotype counts for all these variants, and store them in a $p \times 3$ matrix.

5. (1p) Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. You can use `HWChisqStats` for this purpose. How many SNPs are significant (use $\alpha = 0.05$)?

6. (1p) How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone?

7. (2p) Which SNP is most significant according to the chi-square test results? Give it genotype counts. In which sense is this genotypic composition unusual?

8. (1p) Apply an Exact test for Hardy-Weinberg equilibrium to each SNP. You can use function `HWExactStats` for fast computation. How many SNPs are significant (use $\alpha = 0.05$). Is the result consistent with the chi-square test?

9. (2p) Which SNP is most significant according to the exact test results? Give its genotype counts. In which sense is this genotypic composition unusual?

10. (1p) Apply a likelihood ratio test for Hardy-Weinberg equilibrium to each SNP, using the `HWLratio` function. How many SNPs are significant (use $\alpha = 0.05$). Is the result consistent with the chi-square test?

11. (1p) Apply a permutation test for Hardy-Weinberg equilibrium to the first 10 SNPs, using the classical chi-square test (without continuity correction) as a test statistic. List the 10 p-values, together with the 10 p-values of the exact tests. Are the result consistent?

12. (1p) Depict all SNPs simultaeneously in a ternary plot with function `HWTernaryPlot` and comment on your result (because many genotype counts repeat, you may use `UniqueGenotypeCounts` to speed up the computations)

13. (1p) Can you explain why half of the ternary diagram is empty?

14. (2p) 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? 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?.

15. (1p) Imagine that for a particular marker the counts of the two homozygotes are accidentally interchanged. Would this affect the statistical tests for HWE? Try it on the computer if you want. Argue your answer.

16. (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 expect $\hat{f}$ to follow theoretically? Use a probability plot to confirm your idea.

17. (2p) Make a plot of the observed chi-square statistics against the inbreeding coefficient ($\hat{f}$). What do you observe? Can you give an equation that relates the two statistics?

18. (2p) We reconsider the exact test for HWE, using different significant levels. Report the number and percentage of significant variants using an exac test for HWE with $\alpha = 0.10, 0.05, 0.01$ and 0.001. State your conclusions.