COMP 364 - Tools for the Life Sciences

Assignment #2: Differential Expression & Class Discovery
Due: March 17th, 2016
10% of total grade

The goal of the assignment is to learn more about R, to explore some statistical testing with differential expression in gene expression datasets, and to learn about class discovery, specifically clustering.

Question 1: [50% of assignment] Differential Expression in Real Data

a. Using only the `vanvliet` dataset, identify all probes that are differentially expressed via a t-test (unequal variances) between observed good and bad outcome patients (5 years), at p-value 0.001? Show your code.
b. Find the name of the genes that are differentially expressed from part (a). Remove any duplicate genes. Provide the name of the top 10 genes and show your code. (Hint: use the `intersect, union, setdiff` functions.)
c. Perform the same analysis as (a-b) but with a Wilcoxon test. Show your code and provide the name of the top 10.
d. For the top gene in the intersection of Parts b and c (significant in both analyses), write 3 lines describing what is known about this gene. Use the NCBI, ENSEMBL, GeneCards or other bioinformatics resource.
e. Plot (boxplot) the expression of the top 100 genes from part (b) in patients with good outcome and then in patients in bad outcome, so that they are side by side. Repeat this for part c.
f. Overall, do you believe that the t-test or Wilcoxon test gave better results. Justify your answer.

Question 2: [50% of assignment] Corrections for Multiple Testing: an Introduction

Consider that there are over 20,000 probes on the microarray platform for vanvliet (`length(huc$vanvliet$probe.info[,1]`).

In Question 1, you tested each of these probes for differential expression between good and bad outcome. When you test every probe this way, there is a chance of witnessing a difference just by chance, if the expression data in `vanvliet` is random and fluctuates according to some distribution (eg. a normal distribution). As a warm up, consider the following scenario:

a. Imagine you could afford to go to a Montreal Canadiens game in the Bell Centre. In total, let's assume there are 30,000 spectators. Imagine that every person at the game had a chance to win a vacation to Barbados if they guess the correct order of a sequence of Heads and Tails from 13 coin tosses (there are 2^13 = 8192 such orderings). Each person would write down their sequencing of 13 coin tosses on a piece of paper and drop it in a bin before the end of the second period. The coin toss would be done in secret and announced after the second period. How many people would you expect to win a vacation? Instead of 13, what number make it 50/50 that nobody would win?

Imagine that for any probe x, the gene expression of x in a patient is distributed according to N(0,1) *regardless* of patient outcome.

Suppose you have 10 good outcome patients, and 10 bad outcome patients. There is some chance that all 10 good outcome patients have an expression generated from N(0,1) that is negative, and all 10 bad outcome patients have a positive expression taken from N(0,1). If you were to use a t-test between good and bad outcome at a reasonable significance level (e.g. 0.01), it would likely report that probe x is differential between good and bad outcome. In reality, all that is happened is just luck of the draw: it might not happen very often, but sometimes all the good outcomes may have negative expression and the bad outcomes have positive expression. If you have many probes (e.g. 20,000 for vanvliet, >90,000 for miniTCGA, this kind of thing might happen a lot.

b. Make a new matrix that is a copy of the expression matrix for `vanvliet` (`huc$vanvliet$rand1.expr <- huc$vanvliet$expr`), however replace the expression values for every probe and every patient with a random variate generated from a normal distribution `N(0,3)`. Plot the histogram of expression over the entire matrix and show your code (but not the matrix because it's too large).

c. Like Question 1a, using `rand1.expr`, test each probe in `vanvliet` for differential expression via a t-test at p - value < 0.001. How many probes did you find to be significant? How many did you find to be significant in part 1a? Any idea what this might mean?

d. In general N(0,3) might not be a good way to model the background distribution of expression. Instead, plot the histogram of expression over the entire original

expression matrix `huc$vanvliet$expr`. It should look somewhat normal-like. Suggests a better normal distribution and repeat part (c). Are there more or less probes that are differential?

e. Often in bioinformatics, we perform *permutation tests* where we take the original data and scramble it up. For example, in huc$vanvliet$clinical$event.5, some number (call it b) patients have TRUE, and some number (call it g) have FALSE. Write R code to randomly re-assign huc$vanvliet$clinical$event.5 so that g patients have good outcome and b patients have bad outcome (try using the `sample` statement without replacement).

f. Compare the different background distributions from 2b-e against the results of question 1a. Where is the most and the least number seen? Do you see a pattern or can you suggest which approach is more or least suitable?

g. Finally, the simplest form of multiple testing corection is maybe the Bonferroni correction (http://en.wikipedia.org/wiki/Bonferroni_correction) where you divide your original p-value by the number of test eg. if the p-value is originally 0.001 (like above) and you have 10,000 probes to test, your adjusted p-value is 0.001 / 10,000 = 0.0000001. How many probes are significant in question 2a when you use the Bonferroni correction?