Statistics 5570/6570 Spring 2014 Homework 4

Expected Completion of Content: 21 Mar
Assignment Due (by 11:59 P.M.): 2 Apr

Directions: (Same as with previous homeworks, with 5 points for neatness and format, including code in appendix; there are 65 points possible.) The main purpose of this assignment is to help students gain familiarity with gene set testing tools in Bioconductor, and to discuss some of the statistical issues involved. For Exercises 2-4, refer to the GSE36149 data (including the README file) available on the course website (same data as for previous homework), and focus on the comparison between all six cell line / dosage level combinations (see `trt` in starter code below). Do not filter the data for this homework.

1. (a) (10 points) In your own words, explain the general goal of gene set testing.
   (b) (8 points) Explain briefly how the global test of a gene set is different from the t-statistic-based tests of differential expression we discussed in Unit 3.

2. Perform the global test on the set of all genes.
   (a) (8 points) Report the p-value and conclusion in the context of this set of all genes.
   (b) (4 points) Why might this test be of interest?

3. Perform the global test on all biological process (BP) gene ontology (GO) terms containing between 5 and 200 genes.
   (a) (8 points) Report the histogram of resulting p-values (there should be nearly 1000 of them).
   (b) (4 points) What can you conclude based on this histogram?
   (c) (2 points) What is the most significant BP GO term here? (Report the identifier and alias or short description.)

4. Create a graph visualization of the ten most significant BP GO terms from your result in Exercise 3, within the overall BP ontology. Make the graph interactive by using the `interactive.graph` function from slide 20 of Notes 4.4.
   (a) (8 points) Report the graph showing in the legend the single most significant BP GO term from Exercise 3c (i.e., click on it).
   (b) (8 points) In your own words, explain why such a figure can be useful.

```
######### Code to load necessary packages and get started ##########
library(affy); library(hgu95av2.db); library(globaltest); library(GOstats)
library(GO.db); library(Rgraphviz); library(annotate)
data <- ReadAffy(...)
eset <- exprs(rma(data))
trt <- c('CR','CR','CS','CS','HR','HR','HS','HS','LR','LR','LS','LS')
```