Homework 3
BIOS-7659/CPBS-7659
Due 10/13/2020 9AM

1. T-statistics

- From Canvas, download microarray data (hw3arraydata.txt) and gene names (hw3genenames.txt) in the hw3data directory. This is a gene expression study of apolipoprotein AI (apo AI). There are 8 mice in the control group (C57Bl/6 strain) and 8 mice with the apo AI gene knocked out. The data have already been pre-processed and log transformed. The gene names have been saved in a separate file since there are duplicate names (keep any eye on that). In all parts below, list the genes by name.
- You will need to install the following packages from Bioconductor

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")
BiocManager::install("impute")
BiocManager::install("limma")
```

- You will need to load these two packages (`impute` and `limma`), in addition to the `samr` package from Homework 2.
- HINT: Use `blank.lines.skip = FALSE` when reading in the gene names. Some genes have no annotation, if this option is TRUE, then those genes will be skipped.
- HINT: For problems 1 and 2, `apply()` may be handy to perform operations on each gene.

(a) For each gene, calculate the fold change between the knock-out and wildtype groups. List the top 10 genes that show the largest fold change (positive or negative).

(b) Obtain the p-values from a two sided t-test for differential expression. How many genes are significant at the 0.01 level? List the top 10 genes that have the largest t-statistics and their corresponding p-value.

(c) Alternative 't-statistics'

i. Calculate the 'modified' t-statistic and corresponding p-value using the `samr` package in R used in Homework 2. How many genes are significant at the 0.01 level? List the top 10 genes that have the largest 'penalized' t-statistics.

ii. Calculate the 'moderated' t-statistic and corresponding p-value using the `limma` package from BioConductor (Smyth, *Statistical Applications in Genetics and Molecular Biology*, 2004 3:1). To make these calculations, look at the users guide, Section 9.2: `http://www.bioconductor.org/packages/release/bioc/html/limma.html` How many genes are significant at the 0.01 level? List the top 10 genes that have the largest 'penalized' t-statistics.

(d) Compare and contrast the results for the four methods for ranking genes. Explain the differences in how the different t-statistics are calculated.

2. P-values and Multiple Testing

   - You will need to install the following package from Bioconductor

     `BiocManager::install("qvalue")`

     and `gtools` using `install.packages`.

   - Using the apo AI data from Problem #1 above:

   (a) Calculate p-values for the t-statistics using permutations (B=12870 possibilities). Now, how many genes are significant at the 0.01 level?

   HINT: To get all permutations use the "combinations" function in `gtools`. This can take up to 3 hours. Try a few permutations first to see if it is working. Do not use parallel computing methods in R.

   (b) Apply the following multiple testing adjustment methods to the original t-statistic (#1b) p-values and list the number of genes at a cutoff level of .01. Compare and contrast the different methods. Why are some more or less conservative than others?

   NOTE: Write your own functions to do these corrections, not canned functions or packages. Please turn in all your code.

      i. Bonferroni
      ii. Šidák
      iii. Holm step-down procedure
      iv. Benjamini-Hochberg procedure

   (c) Calculate q-values using the `qvalue` library. How many genes have a q-value less than 0.01. What is the π_0 parameter and what value is estimated using this package?