Homework 6
BIOS-7659/CPBS-7659
Due 11/1 in class

1. Next Generation Sequencing: Differential Expression

   - Install the `cqn` and `edgeR` packages from BioConductor. Familiarize yourself with these packages by looking at the manuals.
   - Use the `data()` function to load the `montgomery.subset` data set from Homework 5 For this problem, the two groups are the first 1-5 subjects and then the second five subjects 6-10. Use `data()` to load `uCovar`, which contains the GC content and length of the genes in this data set.
   - The genes are listed by their Ensembl identifier. To investigate your gene lists, go to the Ensembl genome browser: `http://www.ensembl.org/index.html`

   (a) Calculate the RPKM for each gene in `montgomery.subset` using your own code (no need to call any other functions in the packages). Perform a t-test to find genes that are differentially expressed between the two groups. What are the top genes? (Output the t-statistics and p-value)

   (b) It is good practice to plot the histogram of p-values. What shape would be expected? Plot the histogram of p-values from part a). What do you see?

   Extra Credit: What explains the odd pattern that you find?

   (c) How many genes have at least 10 counts across subjects (i.e., total sum across the gene ≥ 10)? Create a new data frame with only those genes. Then, create an edgeR object using the `DGEList` function including the variable for `lib.size` with the total reads per subject to include as an offset. See Section 1.4 in the User's Guide for help:
   `http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf`

   (d) Using the `estimateCommonDisp()` function, what is the common dispersion estimate? Using the `estimateTagwiseDisp()` function, plot a histogram of the dispersion estimate for each gene. How does the common dispersion estimate compare to the distribution of dispersion estimate across genes? (see Section 2.9.1 in the User's Guide)

   (e) Fit the negative binomial model (see Section 2.9.2 and 3.2 in the User's Guide) and test for differential expression using the common dispersion estimate and report the final results for the top 10 genes. Note: Use `exactTest()` on the objects returned from `estimateCommonDisp()` from part d), followed by `topTags()`.

   Now test for differential expression using the genewise (or tagwise) dispersion estimate and report the final results for the top 10 genes. Note: Use `exactTest()` on the objects returned from `estimateTagwiseDisp()`, followed by `topTags()`.

   How do the results change between the two approaches?

(f) For the top 10 genes based on the common dispersion, extract the raw counts (counts are contained in the `counts` value in the `DGEList` you created). What counts do you observe across the subjects for these genes? Using Ensembl what type of genes are in the top list? Repeat now with the top 10 genes using the tagwise dispersion estimates. What pattern of counts and genes do you observe? How are they different than the genes using the common dispersion?

2. Next Generation Sequencing: Method Comparisons

- Install the `DESeq` and `edgeR` packages from BioConductor. Familiarize yourself with these packages by looking at the User's Guide:
  `http://master.bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf` for DESeq and
  `http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf` for edgeR.
- A RNA-seq study of brain striatum expression from two mouse strains (C57BL/6J, DBA/2J) in Bottomly (2011) *PLoS One* 6(3):e17820 can be downloaded from the repository Recount: `http://bowtie-bio.sourceforge.net/recount/`
- See `http://bowtie-bio.sourceforge.net/recount/make_esets.r` for information on how to access data from Recount. You can use this code:

```
load(url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/
bottomly_eset.RData"))
library(Biobase)
phenoData(bottomly.eset)  #gives information about the table
phenoData(bottomly.eset)@data  #outputs the table
phenoData(bottomly.eset)\$strain  #gives mouse strain variable as vector
featureNames(bottomly.eset)[1:10]  # gives first 10 genes in count table
bottomly.count.table <- exprs(bottomly.eset) #creates count table
dim(bottomly.count.table)  #36536x21
head(row.names(bottomly.count.table)) #names of genes
```

(a) Create a new data frame with genes that have at least 10 counts (summed across samples). How many genes are kept? Create the data objects for DESeq (see section 1.1-1.2 in the User's Guide, use `newCountDataSet()` and edgeR (see section 1.4 and 4.1.2 in the User's Guide, use `DGEList()`).

(b) Calculate the DESeq size factors (see section 1.3, use `estimateSizeFactors()` and `sizeFactors()`). Calculate the edgeR size factors using the "TMM" method (see section 4.1.4, use `calcNormFactors()`). What are size factors? How do the two sets of size factors compare?

(c) Calculate the DESeq dispersions using the"local" method (see section 2, use `estimateDispersions()`). Calculate the edgeR "tagwise" dispersion for each

gene (see section 2.9.1, use `estimateTagwiseDisp()`). Examine the histograms for the dispersions. How do the two sets of dispersions compare? How do you interpret the differences?

(d) Test for differences between the two strains using DESeq (see section 3.1, use `nbinomTest()`) and edgeR (see section 2.9.2 and 3.2, use `exactTest()`). Note that the two methods do not return the same amount of details for the results. Using adjusted p-values with the Benjamini-Hochberg method (Note: check what the functions provide or if you need to do this yourself), how many genes are found in each method to be differentially expressed? What is the overlap between the methods?

Check the results for one example gene that is significant in one method but not the other. Compare the methods based on the estimate of the fold change and p-value for the example. What do you conclude about the differences between the two methods?