Homework 6 - BIOS 7649

Dominic Adducci

Question 1: RNA-Seq: Differential Expression

```
##  [1] "C57BL_6J" "C57BL_6J" "C57BL_6J" "C57BL_6J" "C57BL_6J" "C57BL_6J"
##  [7] "C57BL_6J" "C57BL_6J" "C57BL_6J" "C57BL_6J" "DBA_2J"   "DBA_2J"
## [13] "DBA_2J"   "DBA_2J"   "DBA_2J"   "DBA_2J"   "DBA_2J"   "DBA_2J"
## [19] "DBA_2J"   "DBA_2J"   "DBA_2J"
```

```
##  [1] C57BL/6J C57BL/6J C57BL/6J C57BL/6J C57BL/6J C57BL/6J C57BL/6J C57BL/6J
##  [9] C57BL/6J C57BL/6J DBA/2J   DBA/2J   DBA/2J   DBA/2J   DBA/2J   DBA/2J
## [17] DBA/2J   DBA/2J   DBA/2J   DBA/2J   DBA/2J
## Levels: C57BL/6J DBA/2J
```

```
##  [1] "ENSMUSG00000000001" "ENSMUSG00000000003" "ENSMUSG00000000028"
##  [4] "ENSMUSG00000000031" "ENSMUSG00000000037" "ENSMUSG00000000049"
##  [7] "ENSMUSG00000000056" "ENSMUSG00000000058" "ENSMUSG00000000078"
## [10] "ENSMUSG00000000085"
```

```
## [1] 36536    21
```

```
## [1] "ENSMUSG00000000001" "ENSMUSG00000000003" "ENSMUSG00000000028"
## [4] "ENSMUSG00000000031" "ENSMUSG00000000037" "ENSMUSG00000000049"
```

Part 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 object for DESeq2 (use **DESeqDataSetFromMatrix()**) and the data object for edgeR (use **DGEList()**).

```
## Repeated column names found in count matrix
```

After creating a new data frame with genes that have at least 10 counts there are 11870 genes remaining.

Part B

Calculate the DESeq2 size factors (use **estimateSizeFactors()** and **sizeFactors()**). Calculate the edgeR size factors using the "TMM" method (use **calcNormFactors()**). What are size factors? How do the two sets of size factors compare?

Part C

Test for differences between the two strains using DESeq2 (use **nbinomWaldTest()**) and edgeR (use **glmFit()** and **glmLRT()**). 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?

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

The adjusted p-value from the DESeq2 method is calculated using the Benjamini-Hochberg method.

Question 2: RNA-Seq: Remove Unwanted Variation

Part A

Use the Montgomery data set from HW5. For this problem, the two groups are the first five subjects and the second group are the second five subjects. Test for differential expression between the groups using a standard likelihood ratio test (**glmFit()** and **glmLRT()**). See Section 2.11.3 of the **edgeR** user's manual. How many genes have FDR adjusted p-values < 0.05?