Homework 6
BIOS-7659/CPBS-7659
Due 11/5/2020 9AM

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 2.6 in the User's Guide for help:
   `http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf`

   (d) Calculate TMM normalization factors using the `calcNormFactors` function (see Section 2.8.3). How do the effective library sizes compare to the original library size you calculated based on the sum of counts for each sample?

   (e) Use the `estimateDisp()` function to calculate the common, trended and tagwise dispersions. What is the common dispersion estimate? Plot the tagwise dispersion estimate for each gene vs. the average log counts per million. Add lines for the trended dispersion estimate and the common dispersion estimate. How does the common dispersion estimate compare to the dispersion estimates across genes? (see Section 2.10.1 in the User's Guide)

   (f) Fit the negative binomial model (see Section 2.10.2 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 object returned

from `estimateDisp()` from part d) with the dispersion option set to "common" , 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 the disperion = "tagwise" option in the `exactTest()` function.

How do the results change between the two approaches?

(g) 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: Remove Unwanted Variation.

- `edgeR` can fit models in a GLM framework, which allows us to control for covariates (see Section 2.11), such as factors of unwanted variation calculated using `RUVSeq` methods.
- Install the `RUVSeq` package from BioConductor.
- Review the `RUVSeq` documentation: `http://bioconductor.org/packages/release/bioc/vignettes/RUVSeq/inst/doc/RUVSeq.pdf`.

(a) Create a design matrix that includes an intercept and group indicator for the filtered Montgomery dataset you created in Question 1(c). Perform upper-quartile normalization using the `calcNormFactors` function and recalulate the common, trended and tagwise dispersions, providing the design matrix in the design option of the `estimateDisp` function (see Section 2.11.2 of the `edgeR` user's manual). 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?

(b) We will use `RUVSeq` to fit `edgeR` models that adjust for unwanted variation. Following section 2.1 of the `RUVSeq` documentation, create an expression set and perform upper-quartile normalization using the `betweenLaneNormalization` function. Plot the relative log expression of the upper-quartile normalized counts and create a PCA plot colored by group. Are there any issues you note in these plots that may benefit from additional normalization?

(c) Perform RUVg using negative empirical control genes. To define a set of control genes, take the 10,000 genes with the largest likelihood ratio test p-values from part (a). Follow the steps in Section 2.2 and 2.3 of the `RUVSeq` manual, using upper-quartile normalization and k=1. Create boxplots of RLE and a PCA of the normalized counts. Comment on these plots. How many genes had FDR adjusted p-values < 0.05 after controlling for 1 factor of unwanted variation?

(d) Repeat part (c) above using k=2. How many genes had FDR adjusted p-values < 0.05 after controlling for 2 factors of unwanted variation?

(e) Repeat part (d) using the RUVr method with k=2, following Section 4 of the `RUVSeq` manual. How many genes had FDR adjusted p-values < 0.05 after controlling for 2 factors of unwanted variation using RUVr?

(f) What are your thoughts about using RUV methods on this dataset? Do you have any concerns about these methods?

3. Next Generation Sequencing: Method Comparisons

- Install the `DESeq2` and `edgeR` packages from BioConductor. Familiarize yourself with these packages by looking at the User's Guide:

  `https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html` for DESeq2 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 DESeq2 (use `DESeqDataSetFromMatrix()` and edgeR (use `DGEList()`).

(b) Calculate the DESeq2 size factors (see section 1.3, 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?

(c) Calculate the DESeq2 dispersions using the "local" method (use `estimateDispersions()`). Calculate the edgeR "tagwise" dispersion for each gene (use `estimateDisp()`). Examine the histograms for the dispersions. Create a Bland-Altman plot to show the agreement between the log(dispersions) estimated using the two methods. How do the two sets of dispersions compare? How do you interpret the differences?

(d) 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?