Lab #5:

Please e-mail code, graphs and answers to questions to [bsmit269@uncc.edu](mailto:bsmit269@uncc.edu) and [afodor@uncc.edu](mailto:afodor@uncc.edu)

Please have lab submitted (whatever you have) before class on Wed., March 11th.

(1) Download the file http://afodor.github.io/classes/stats2015/longitdunalRNASeqData.zip

Read the counts table ( “nc101\_scaff\_dataCounts.txt “ into R).

Plot the log10(average +1) of the mean of the two week (columns 1-3) vs. 20 week data (columns 7-11) with an identity line in red.

(2) Calculate the sj scale factors as defined by Equation 5 here: http://genomebiology.com/2010/11/10/r106

(see slide 21 in lecture 10)

Plot the sj scale factors calculated in this way against the sum of each sample j. For this dataset, how closely related is the normalization scheme proposed by Eq. 5 to just defining the scale factor as the sum of all sequences in a sample?

(3) Calculate qhatip (equation 6) for the two weeks and twenty week samples (slide 23 of lecture 10 might help!).

Plot log10(qhatip+1) for two weeks vs. twenty weeks with a red identity line. Does the normalized data fall more closely on the identity line than the unnormalized data plot of (1)?

(4) Consider gene #8 at the two weeks point normalized by the scale factors; this can be expressed in R by:

myT[8,twoWeeks]/sjs[twoWeeks]

where myT is the count table read in by step 1 and sjs are the scale factors calculated by step 2.

Consider the mean and canonical variance of these numbers (expressible in R as):

apply(myT[8,twoWeeks]/sjs[twoWeeks],1,mean)

apply(myT[8,twoWeeks]/sjs[twoWeeks],1,var)

Calculate the equivalent r and p for the negative binomial distribution given that mean and variance (see slide 14 in lecture 9).

Plot the probability density function for the negative binomial distribution defined from this r and p for between 0 and 500 sequences.

In another set of samples, you observed an average of 120 reads (normalized to the common scale) from samples with the same variance as the two weeks samples, what is the p-value for a one sided evaluation of a null hypothesis that you could have seen 120 or fewer reads by chance in the probability density function defined by the r and p above?