Lab #4

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 Monday, Feb 16th.

We will, as we have been doing, go over the lab in class on Monday.

-----------

Download this dataset:

afodor.github.io/classes/stats2015/nc101\_scaff\_dataCounts.txt

(right-click in the browser and say “save as…”)

Each row in the spreadsheet represents a different gene in an RNA-seq experiment. The samples are E. Coli derived from two different mice under identical conditions (i.e. biological replicates).

(1) Read the dataset into R using commands something like...

setwd("C:\\classes\\Advanced\_Stats\_Spring2015\\Lab4\_HW1")

myT <- read.table("nc101\_scaff\_dataCounts.txt",header=TRUE,row.names=1)

(2) On a log-log scale, show a plot of the counts for the two samples “D2\_01” and “D2\_02”.

Qualitatively, do the biological replicates appear to have similar patterns of gene expression?

(3) Consider the first gene in the spreadsheet (e.g. NC101\_00003). Make a two by two contingency table:

|  |  |  |
| --- | --- | --- |
|  | Sequences in D2\_01 | Sequences in D2\_02 |
| Assigned to NC101\_00003 |  |  |
| Not assigned to NC101\_00003 |  |  |

use the two sided fisher.test to generate a p-value for the null hypothesis that the columns and rows of the contingency table are independent.

(4) Now generate a p-value for all the genes in the spreadsheet from the Fisher test. Plot out those p-values in a histogram. Are they uniformly distributed? Would you expect them to be? Are the p-values more significant, less significant or what we would expect under a uniform distribution? How does the p-value distribution change if you remove low abundance genes (with for example myT <- myT[ (myT$D2\_01 + myT$D2\_02 > 50),]

(5) Add 1 to every value in the table ( with something like myT = myT + 1 ). This is called adding a pseudo-count. Now consider the first gene (NC101\_00003 ) again. From the first experiment, calculate

expected frequency = p =

(# Assigned to NC101\_00003 in D2\_01)/total # of sequences in D2\_01.

Now use poisson.test to assign a p-value for the null hypothesis that value of p derived from D2\_01 could have produced the number of reads observed for this gene in D2\_02 .

(6) Repeat the calculation in (5) for every gene in the spreadsheet. Graph these p-values against the p-values produced in (4). Do they agree?

(Note: as we will see, neither of these methods are necessarily the best for analyzing these data! Also, with n=1 in both conditions, our sample size is small and we would want to be very careful about putting too much weight in any conclusions we draw from such a small sample size...)

(7) Why did we add a pseudo-count in step 5? What would happen if we didn't add the pseudo-count?