Lab # 12:

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, April 29.

(Note the Wed. due date for this problem set...)

-----------------------------------------------------------------------------------------------------

This week's dataset is the case/control colorectal adenomas data set, but this time starting with raw counts at the family level:

http://afodor.github.io/classes/stats2015/familyPivotedTaxaAsColumnsNotNormalized.txt

(1) Normalize the dataset so that the number of counts is (approximately) the same in each sample.

For example:

myT <- read.table("familyPivotedTaxaAsColumnsNotNormalized.txt",header=TRUE,row.names=1, sep="\t")

rowSums <- apply( myT, 1, sum)

avgPerSample <- mean(rowSums)

for( i in 1:nrow(myT))

{

rowSum = sum(myT[i,])

myT[i,] = avgPerSample \* myT[i,] / rowSum

}

myT <- round(myT)

(2) Consider only the bugs (columns) that have at least one zero count:

(A) Plot the log(p-values) for each bug from a null hypothesis that case/control is from the same distribution. On one axis of your plot put p-vales from a simple linear model (in which you have log transformed the data with for example log(myT+1)) and on the other axis put the p-values from the zero inflated negative binomial (ZINB) model in which both the binomial and negative binomial parts have a term for case/control.(You can get the p-values for the ZINB by "lrtest" in package "lmtest”)

Do the p-values from the two models seem broadly similar? At a 10% false discovery rate, which model produces more significantly different hits?

(B) Plot the AIC values for the two models built for each of the bugs. For most of the bugs, which model has the lowest AIC?