Lab #7:

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 Mon., March 23rd.

(1) The following table shows the relationship between people with a disease state and the presence of an E. Coli with a particular gene:

|  |  |  |
| --- | --- | --- |
| Disease | Samples with the PKS gene | People without the PKS gene |
| Cancer (CRC) | 14 | 7 |
| Inflammation (IBD) | 14 | 21 |
| Control | 5 | 19 |

(data are from http://www.sciencemag.org/content/338/6103/120.full.pdf ; Fig. 3B)

Use the chi-square to test in R to evaluate a null hypothesis that the presence of the PKS gene is independent of the disease state. What is the p-value associated with that null hypothesis?

(2) Use the simulated data frame code from lab #6 to make a simulated RNA-seq dataset with 10 samples and 3,000 genes. Normalize the dataset by dividing each sample by the total number of reads in each sample. Apply a t-test to each one of these 3,000 genes (comparing the first 5 columns to the 2nd 5 columns). Is the resulting distribution of p-values uniform? Are any of these genes significantly different at a 10% false discovery rate.

(3) Consider this data set:

C:\Users\Anthony\git\afodor.github.io\classes\stats2015\caseControlData.txt

(described in this paper <http://www.nature.com/ismej/journal/v6/n10/pdf/ismej201243a.pdf> )

(This dataset is already normalized!)

For each OTU (column in the dataset), generate a t-test for a null hypothesis that the distribution of the OTU for case and control is the same.

(The command “grepl” might be helpful when applied to the “samples” column to tell case from control).

Are the p-values generated in this way uniformly distributed? How many OTUs are significantly different at a 10% false discovery rate ( the command, p.adjust(…, method="BH") might be useful).