1. We return to our RNA seq dataset of E. Coli genes from mice. The URL is here:

http://afodor.github.io/classes/stats2015/longitdunalRNASeqData.zip

Read and normalize the counts table ( “nc101\_scaff\_dataCounts.txt “ into R). For example:

setwd("C:\\somewhere")

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

# remove rare genes

myT <- myT[ apply( myT,1, median)> 5,]

myTNorm <- myT

for ( i in 1:ncol(myT))

{

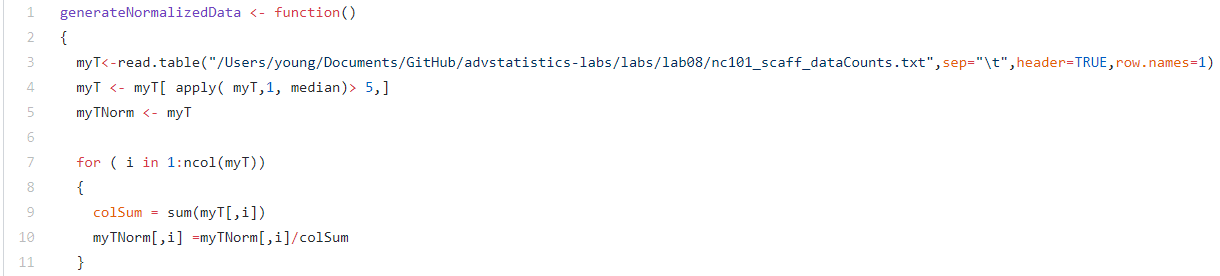
colSum = sum(myT[,i])

myTNorm[,i] =myTNorm[,i]/colSum

}

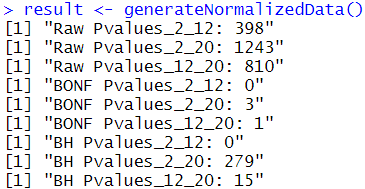
(The first 3 columns are “day 2”, the next 3 columns are “week 12” and the last 5 are “~~week 18~~” [“week 20”]). Remember, that day 2 is before the mice have inflammation symptoms, week 12 is associated with inflammation and ~~week 18~~ [week 20] is associated with cancer.

The above code for data normalization was implemented within the following code block (lines 3 to 11) at figure 1. The source file is located at [this link](https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab08/generateNormalizedData.R).

  
**Figure 1:** Implemented Data Normalization Code

1. For every row in the normalized spreadsheet, run three t-tests ( “day 2” vs. “week 12”, “day 2” vs. “week 18” and “week 12” vs. “week 18”. At a p < .05 threshold fill in the following table:

The function illustrated at figure 2 was utilized to calculate the nine answers for table 1. The source file is located at [this link](https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab08/generateNormalizedData.R).

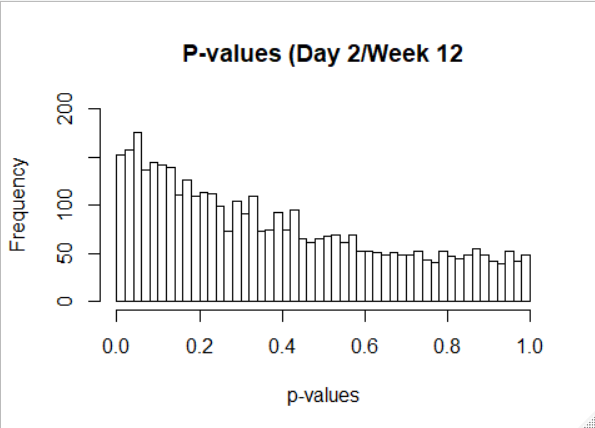
  
**Figure 2:** Output of R function [generateNormalizedData()](https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab08/generateNormalizedData.R)

|  |  |  |  |
| --- | --- | --- | --- |
|  | # of genes significant at p <0.05 uncorrected | # genes significant at p <0.05 BH FDR corrected | # genes significant at p <0.05 Bonferroni corrected |
| “day 2” vs. “week 12” | 398 | 0 | 0 |
| “day 2” vs. “week ~~18~~**(20)**” | 1243 | 3 | 279 |
| “week 12” vs. “week ~~18~~**(20)**” | 810 | 1 | 15 |

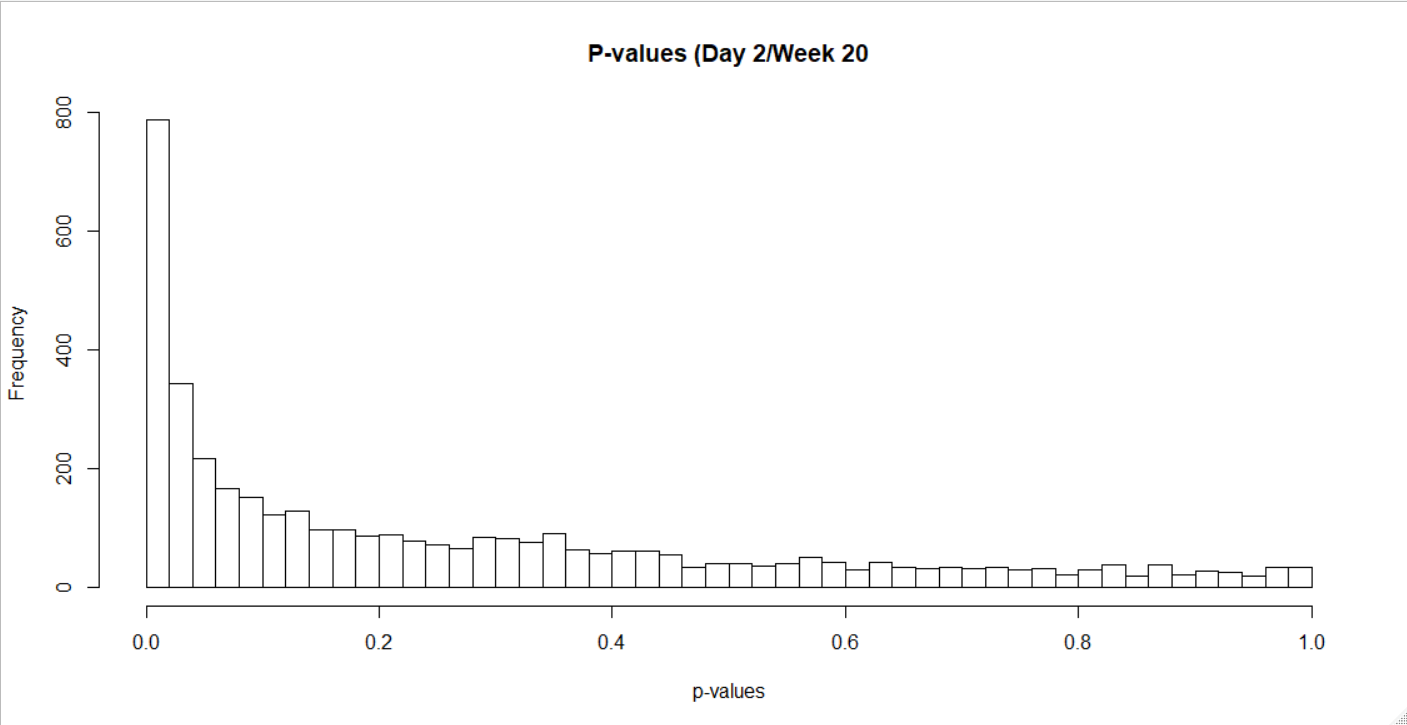
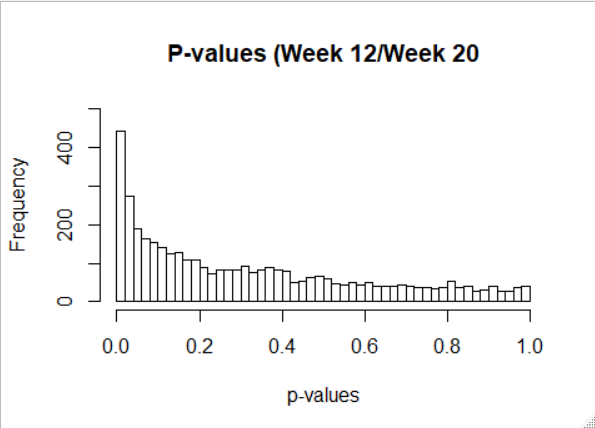
**Table 1:** Significant Gene Counts

1. Make histograms of all the uncorrected p-values for each of the three companions. Are any of the distributions uniform?

* All histograms (figures 3, 4 & 5) were generated from the previous R function located at [this link](https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab08/generateNormalizedData.R).
* Per figures 3, 4 & 5, none of the distributions were uniform.

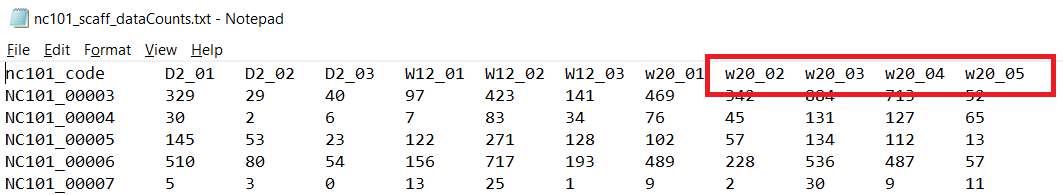


**Figure 3:** Uncorrected P-values for Day 2/Week 12

  
**Figure 4:** Uncorrected P-values for Day 2/Week 20  
  
  
**Figure 5:** Uncorrected P-values for Week 12/Week 20

1. Based on these data, when is the biggest shift in the transcriptome? Which samples are most different from one another?

* The biggest shift in the transcriptome occurred during week 20, the phase associated with cancer (third row of table 1). By the way, I’m using week 20 instead of week 18, because that week is labeled as *w20\_xx* within the source file (see figure 6).
* Samples day 2 and week 20 were the most different from one another, because the most significant gene count occurred between them (second row of table 1).

  
**Figure 6:** Source data file with week 20 labels (w20\_).