Erika Maldonado-Rosado

13 April 2019

BINF 6203

Gene Expression Lab

**Introduction**

For this lad we did two different methods to calculate differential expression for the same RNASeq dataset. We wanted to find out which genes are the most significantly differentially expressed between two conditions. We are concerned with the magnitude of differential expression and the significance of differential expression. We only wanted to see genes with both high significance and large effect size change within a common threshold.

**Method**

For these two combined labs we were provided with the trimmed files. They used Trimmomatic to trim and FastQC to view their trimming. I then used bowties2 to do read mapping. I made the V. Vulnificus genomes my index file. The files were all provided by Dr. Gibas in her drop box. I was unable to use the larger files so I was forced to use the Sample files which was much smaller. I converted my files from fq to sam to bam.

After read mapping, I did read counting. I was unable to do this in my ubuntu terminal nor my Rstudio, so I went onto mamba and used the feature count command which covered my file to a .txt file. I used the V. vulnificus gtf as a reference file to do the feature count. I only used 4 out of my 8 files that were given to me. I did a combination of all 4 files with the same command and a different output file.

Next I moved my files from mamba to my terminal then moved it to my desktop. I used R do to the gene expression analysis. To do the gene expression analysis you need 5 different packages. You need DESeq2, pheatmap, ggplot2, genefilter, and calibrate. You read your file into an environment, so it can be called upon. The data is turned into a matrix. You then define the experimental relationship of your columns and define the data frame. You then create a DeSeq data set from your count matrix and data frame. After, you order and select your data then apply log transformation in order to create a pheatmap and a PCA plot. After generating the pheatmap and the PCA plot we then apply DESeq to test for statistically significant differential expression. We do this by getting ones results and ordering them by the p-value. You then calculate the number of true and false genes and put the results into a csv file. You then turn those results into a histogram. You then produce a volcano plot of significance vs. effect size and display a heatmap of significantly DE (true) genes.

Below is the code that I used and the results that I produced.

**Results**

**\*\*Bowtie and feature counts starts from the bottom up.**

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o CMCP6count.txt ASW1.bam ASW2.bam HS1.bam HS2.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o HS2-2count.txt HS2-2.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o HS2-1count.txt HS2-1.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o ASW2-2count.txt ASW2-2.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o ASW2-1count.txt ASW2-1.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o HS1-2count.txt HS1-2.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o HS1count.txt HS1.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o ASW1-2count.txt ASW1-2.bam

featureCounts -a Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23.gtf -o ASW1count.txt ASW1.bam

cp ASW1.bam ASW1-2.bam ASW1-2.bam HS1.bam HS1-2.bam ASW2-1.bam ASW2-2.bam HS2-1.bam HS2-2.bam /mnt/c/Users/erika/Onedrive/Desktop

samtools view -b HS2-2.sam > HS2-2.bam

bowtie2 -x VibrioDNA -U CMCP6\_2HSsample.2.fq -S HS2-2.sam

samtools view -b HS2-1.sam > HS2-1.bam

bowtie2 -x VibrioDNA -U CMCP6\_2HSsample.1.fq -S HS2-1.sam

samtools view -b ASW2-2.sam > ASW2-2.bam

bowtie2 -x VibrioDNA -U CMCP6\_2ASWsample.2.fq -S ASW2-2.sam

samtools view -b ASW2-1.sam > ASW2-1.bam

bowtie2 -x VibrioDNA -U CMCP6\_2ASWsample.1.fq -S ASW2-1.sam

samtools view -b HS1-2.sam > HS1-2.bam

bowtie2 -x VibrioDNA -U CMCP6\_1HSsample.2.fq -S HS1-2.sam

samtools view -b HS1.sam > HS1.bam

bowtie2 -x VibrioDNA -U CMCP6\_1HSsample.1.fq -S HS1.sam

samtools view -b ASW1-2.sam > ASW1-2.bam

bowtie2 -x VibrioDNA -U CMCP6\_1ASWsample.2.fq -S ASW1-2.sam

samtools view -b ASW1.sam > ASW1.bam

bowtie2 -x VibrioDNA -U CMCP6\_1ASWsample.1.fq -S ASW1.sam

bowtie2-build Vibrio\_vulnificus\_cmcp6.GCA\_000039765.1.23DNAgenome.fasta VibrioDNA

**R Based Code**

```{r}

countdata<- read.table("/Users/erika/OneDrive/Desktop/LabCount/CMCP6count.txt",sep = '\t', header = TRUE, row.names = 1)

```

```{r}

countdata<- as.matrix(countdata)

```

```{r}

condition<-factor(c(rep("A",2),rep("B",2)))

```

```{r}

colData<- data.frame(row.names = colnames(countdata), condition)

```

```{r}

countdata<- as.matrix(countdata)

```

```{r}

condition<-factor(c(rep("A",2),rep("B",2)))

colData<- data.frame(row.names = colnames(countdata), condition)

```

```{r}

colData

```

```{r message=FALSE, warning=FALSE}

library(DESeq2)

dds<- DESeq2::DESeqDataSetFromMatrix(countData = countdata, colData = colData, design =~condition)

```

```{r}

select <- order(rowMeans(counts(dds,normalized=FALSE)),decreasing=TRUE)[1:25]

```

```{r}

select <- order(rowMeans(counts(dds,normalized = FALSE)), decreasing = TRUE)[1:25]

```

```{r}

df <- as.data.frame(colData(dds))

```

```{r}

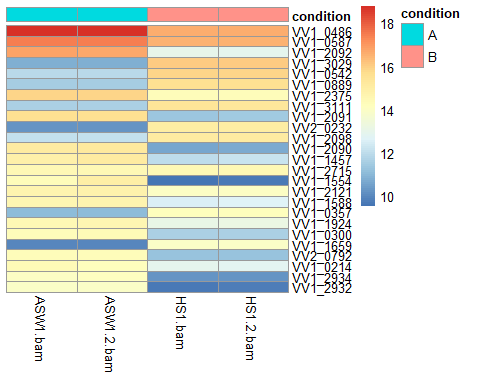
rld<- rlogTransformation(dds, fitType="mean")

```

```{r}

pheatmap::pheatmap(assay(rld)[select, ], cluster\_rows = FALSE, show\_rownames = TRUE, cluster\_cols = FALSE, annotation\_col = df)

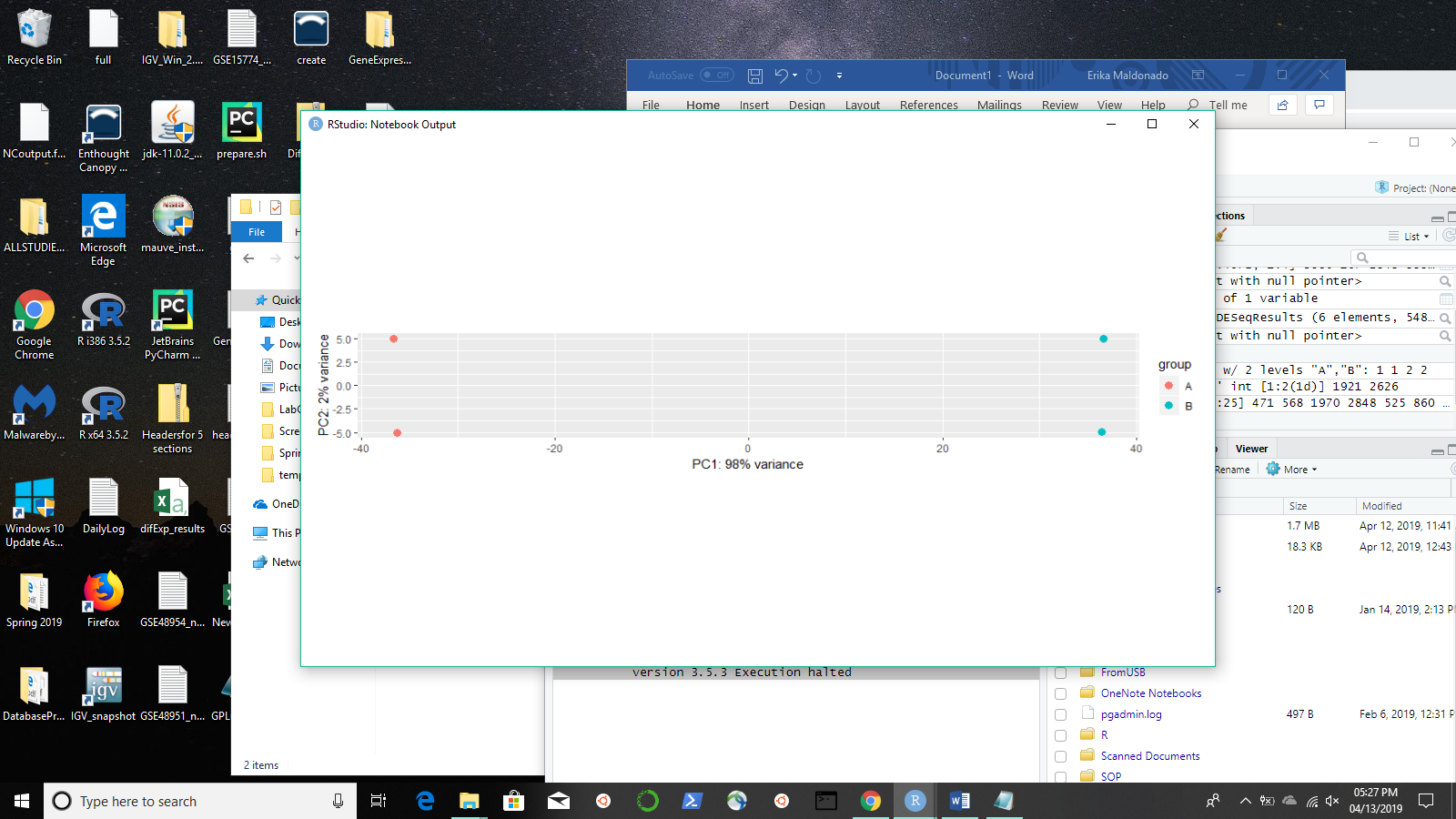
```



```{r}

plotPCA(rld, intgroup = "condition")

```



```{r}

dds <-DESeq(dds)

```

```{r}

res <- results(dds)

res <-res[order(res$padj), ]

```

```{r}

res

```

```{r}

deCount <- table(res$padj<0.05)

```

```{r}

write.csv(res, file = "difExp\_results.csv")

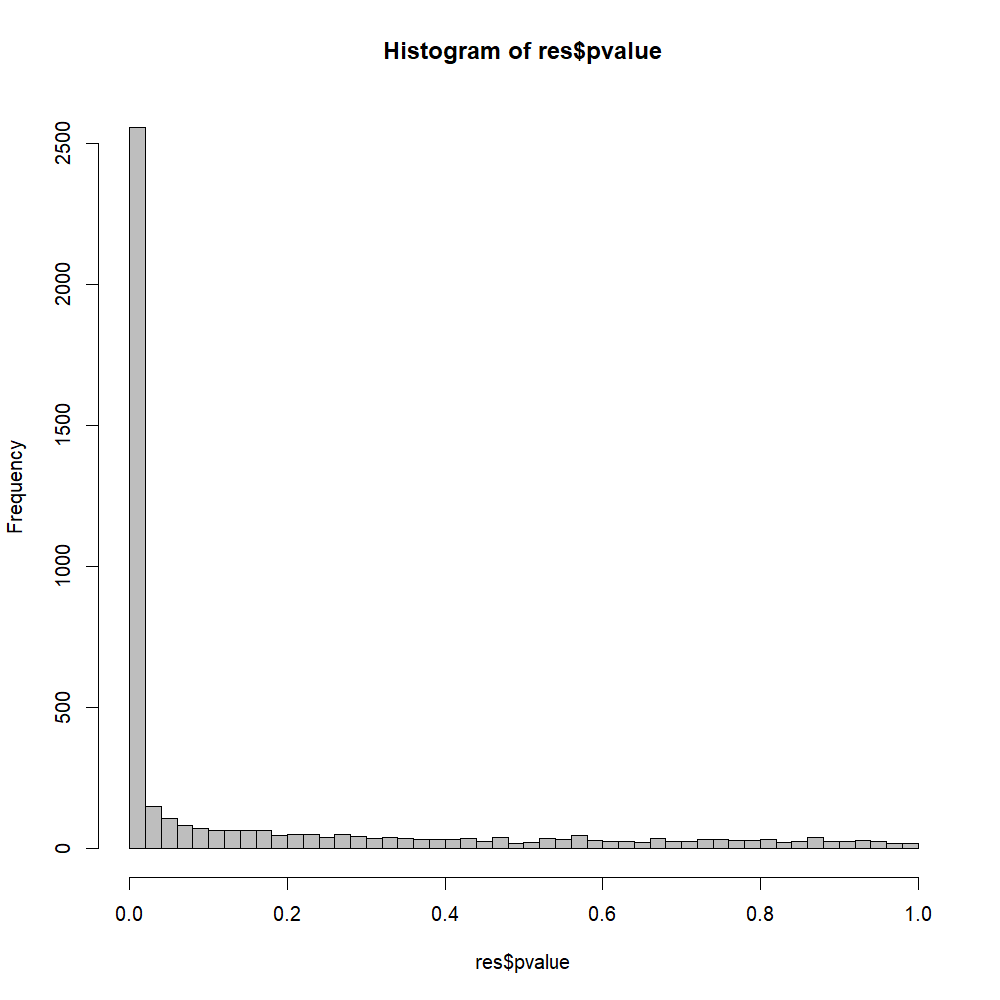
```

```{r}

png("DE\_pvals.png", 1000, 1000, pointsize = 20)

hist(res$pvalue, breaks = 50, col = "grey")

```



```{r}

library(calibrate)

library(DESeq2)

volcanoplot1 <- function(res, lfcthresh = 2, sigthresh =0.05, main = "Volcano Plot", legendpos = "bottomright", labelsig = TRUE, textcx = 1, ...){

with(res, plot(log2FoldChange, -log10(pvalue), pch = 20, main = main,...))

with(subset(res, padj<sigthresh ), points(log2FoldChange, - log10(pvalue), pch = 20, col = "red",...))

with(subset(res,abs(log2FoldChange)>lfcthresh), points(log2FoldChange, -log10(pvalue), pch = 20, col = "orange",...))

with(subset(res,padj<sigthresh & abs(log2FoldChange)>lfcthresh), points(log2FoldChange, -log10(pvalue),pch = 20, col = "green",...))

if(labelsig){

require(calibrate)

with(subset(res,padj<sigthresh & abs(log2FoldChange)>lfcthresh),textxy(log2FoldChange, -log10(pvalue), labs = Gene, cex = textcx,...))

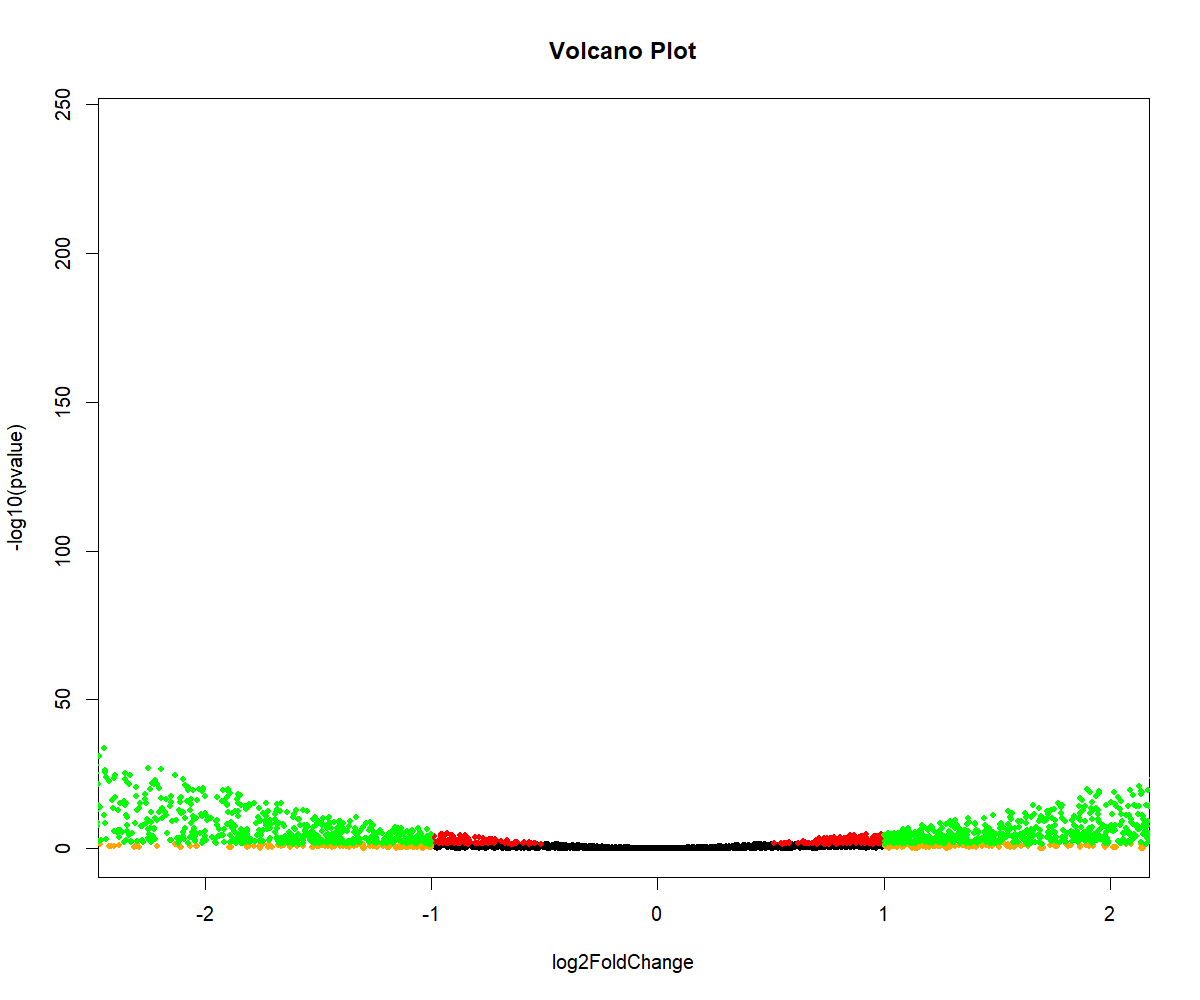
}

legend(legendpos, xjust = 1, yjust = 1, legend = c(paste("FDR<", sigthresh, sep = ""), paste("|LogFC|>",lfcthres,sep = ""),"both"),pch = 20, col = c("red","orange","green"))

}

png("volcano1.png",1200,1000, pointsize = 20)

volcanoplot1(res,lfcthresh = 1, sigthresh = 0.05, textcx = .8, xlim = c(-2.3,2))



```{r}

library("genefilter")

library("pheatmap")

topVarGenes <- head(order(rowVars(assay(rld)),decreasing=TRUE),20)

mat <- assay(rld)[ topVarGenes, ]

mat <- mat - rowMeans(mat)

df <- as.data.frame(colData(rld))

pheatmap::pheatmap(mat, annotation\_col=df)

```

