Lauren Ashlock
3/8/17

Assignment #2: RNA-seq for Gene Expression Analyses P/BIO 381

Introduction: The aim of this assignment was to use R package DESeq2 to identify differential patterns of gene expression among samples taken from healthy and sick *Pisaster ochraceus*[1,2]. DESeq2 uses the number of reads that map to each contig, in each sample, as a measure of gene expression. Genes with higher read counts are considered to be upregulated, while genes with lower read counts are downregulated. This analysis allows for the comparison of gene expression patterns in samples taken from healthy versus sick individuals. Through this analysis, it is also possible to identify specific genes that are highly differentially expressed. This information can be used to identify associated functions of orthologous genes, providing insight into the potential mechanisms of host response to sea star wasting disease.

Methods: We assessed the quality of raw RNA-seq data using fastqc, and cleaned our raw reads using Trimmomatic. After removing adapters and low quality bases from our file, we again evaluated the quality using fastqc. We then assembled a de novo transcriptome using fastq files from four sea stars (both sick and healthy). Clean reads were mapped to our reference transcriptome, producing sequence alignment files. Count data, for the number of reads that mapped to each contig, were extracted from our sequence alignment files. These count data were used for differential expression analysis in DESeq2. For my analysis of differential gene expression I compared gene expression among samples from healthy and sick intertidal individuals, healthy and sick subtidal individuals, and healthy and sick individuals, while controlling for location of collection.

Results: The analysis of differential expression in samples taken from healthy and sick intertidal individuals resulted in 205 significantly upregulated, and 37 significantly downregulated genes (Table 1). Analysis of samples taken from healthy and sick subtidal individuals yielded 20 significantly upregulated genes and 113 significantly downregulated genes. The final analysis, contrasting healthy and sick individuals while controlling for location, resulted in 209 significantly upregulated genes and 65 significantly downregulated genes. Principal Component Analyses (Figure 1) run for each of the models, demonstrate that there is not a clear clustering of samples taken from healthy and sick individuals in any of the models. Despite the lack of clear clustering, further examination of the most highly differentially expressed gene in each model demonstrates that there are differences in gene expression among samples taken from healthy and sick individuals (Figure 2).

	Intertidal	Subtidal	Full Model
Significantly Upregulated	205	20	209
Significantly Downregulated	37	113	65

Table 1. Summary of the number of genes significantly upregulated and significantly downregulated (adjusted $p < 0.1$) in each model.

Lauren Ashlock
3/8/17

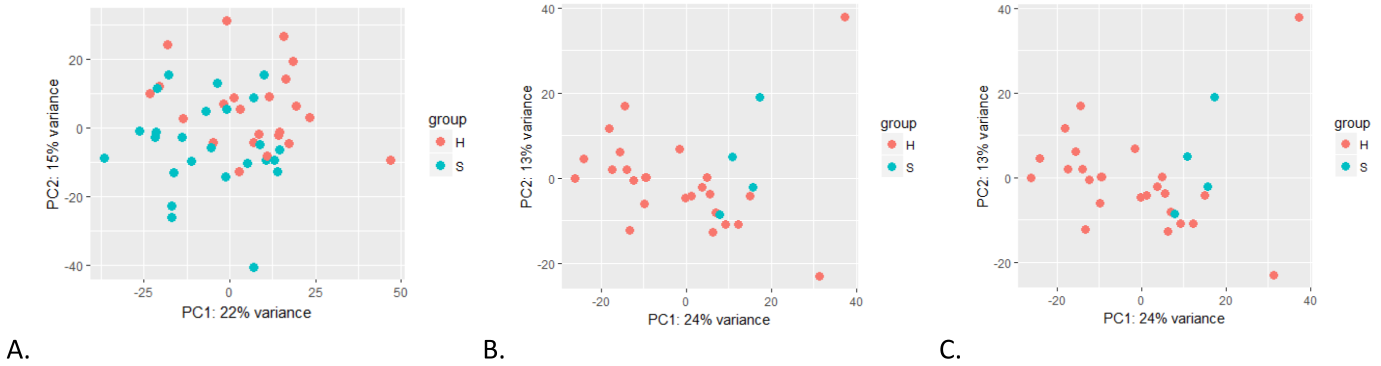

Figure 1. PCA plots comparing gene expression in healthy and sick individuals. A) Depicts healthy and sick intertidal individuals. B) Depicts healthy and sick subtidal individuals. C) Depicts healthy and sick individuals in all locations, while controlling for location.

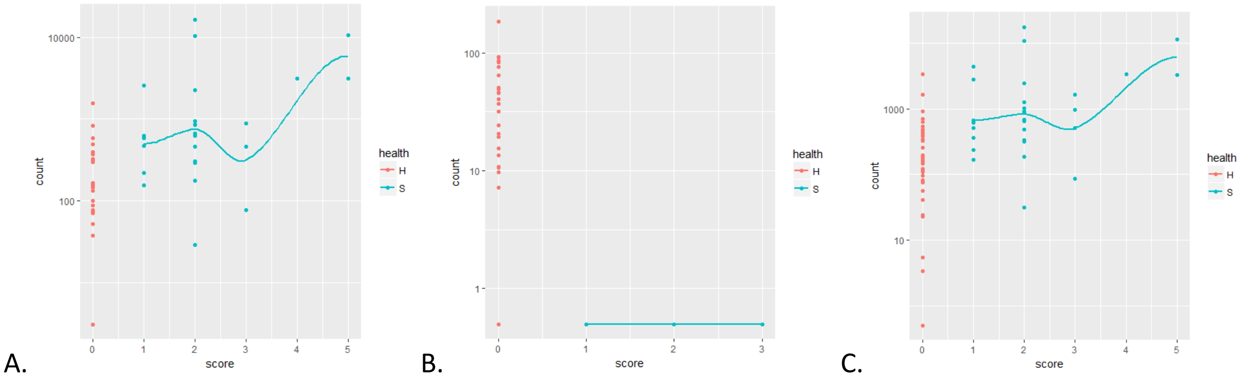

Figure 2. Depiction of the interaction between health, score, and read count for the top differentially expressed gene in each model. A) Differential expression of DN43080_c1_g1 in healthy and sick intertidal individuals. B) Differential expression of DN42073_c0_g1 in healthy and sick subtidal individuals. C) Differential expression of DN43080_c1_g1 in healthy and sick individuals.

Discussion: There is a higher level of upregulation of genes in samples taken from sick intertidal individuals as compared to samples taken from healthy intertidal individuals. Contrastingly, there is a higher level of downregulation of genes in samples taken from sick subtidal individuals as compared to samples taken from healthy subtidal individuals. While PCA plots demonstrate the lack of clear clustering of healthy with healthy and sick with sick samples, our top differentially expressed gene for each model show interesting patterns of differential gene expression. Panel A and C in Figure 2 are both looking at the interaction between read count, health, and score. The DN43080_c1_g1 gene shows an increased expression in the sick samples as compared to the healthy samples, while also increasing with increasing score. The results depicted in Figure 2B, however, show a striking lack of expression in gene DN42073_c0_g1 in subtidal sick samples as compared to subtidal healthy samples. It is interesting to see such differences in overall expression patterns in intertidal versus subtidal samples. Next steps in the analyses of these data would be to functionally annotate our reference transcriptome. With functional annotation, we could associate differentially expressed genes with their potential role in host response to sea star wasting disease.

Works Cited: 1. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2. Michael I Love, Wolfgang Huber and Simon Anders (2014): Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology