The purpose of the following analyses is to identify genes that are being significantly differentially expressed. Differential expression indicates that a gene is being more up- or down-regulated in the sample compared to the reference group. Differential gene expression (DGE) is of particular interest to our Sea Star Wasting Disease project because it may yield biologically informative results if there is DGE between healthy and sick individuals as it may indicate the organism’s response to its health status.

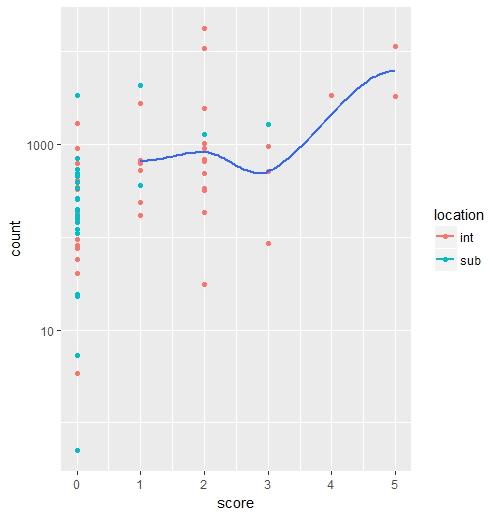
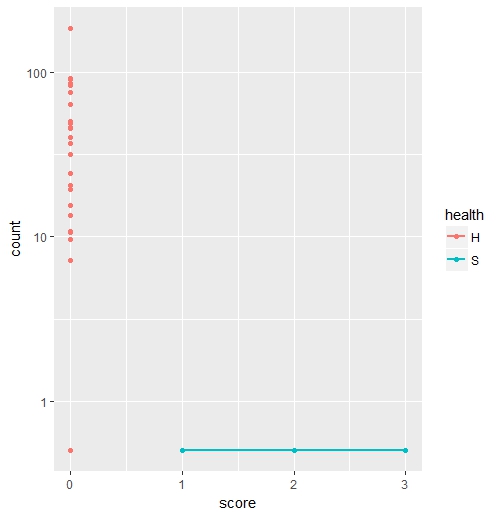
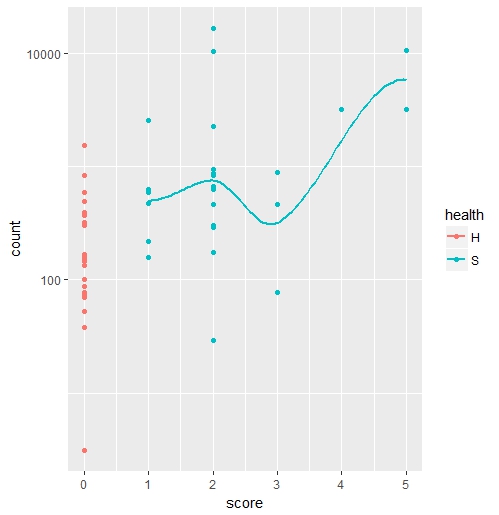
Tissue samples were collected every three days for a set of sea stars and RNA sequence data was obtained via Illumina sequencing which returned paired end 100 bp raw reads with counts (number of times a transcript was present) and sequences for the reads. These raw reads were then assessed for quality and then cleaned with FastQC and Trimmomatic which remove unwanted elements of the sequence data including adaptors, low quality bases, and short reads. Next, Trinity was used to perform a de novo assembly of the transcriptome and reads with open reading frames (ORFs) greater than 100 amino acids in length were identified and those with less than 100 amino acids were discarded using TransDecoder. Quality was reassessed by Blast-matching reads to currently available annotated genomes of closely related species. A custom python script then extracted the read counts and put the data into file types needed for differential gene analysis. Differential gene analysis was done using the R program DESeq2 which compares the number of counts that mapped to a gene between one group (sick) and the reference group (healthy) by using negative binomial linearized models and estimating log fold changes and dispersion. Three models were used to determine DGE, the first contained only individuals from the intertidal group, the second contained only individuals from the subtidal group, and the last model contained all individuals but controlled for location. The code for the DESeq2 analysis can be found on my online lab notebook under Entry 15.

Table 1 summarizes the number of up- and down-regulated genes for each model as well as the indicates the most differentially expressed gene in that model along with its p-value. Only a small percentage of the genes were differentially regulated between groups, and generally the majority of these genes were up-regulated in sick individuals compared to healthy individuals, although many more genes were down-regulated in sick individuals in the subtidal-only model. Figure 1 depicts scatter plots of the number of counts by the health status score for the most differentially expressed gene. Figures 1a and 1b represent the same gene but under two different models; however, although this gene was differentially expressed there is not a clear difference in number of counts. On the other hand, Figure 1c corresponds to the subtidal-only model and from this figure it is clear that healthy individuals expressed this particular gene with counts between 10 and 100, whereas sick individuals did not express this gene. Principle Component Analyses (PCA) were done on the data to determine if there is a true difference in transcriptome response between different groups within the data. No variable resulted in the defined groups (health, day, location, etc.) being tightly clustered and separated; however, Figure 2 show Principle Component Analysis (PCA) plots for each of the three models with the variable health status as the PCA separated the healthy and the sick individuals relatively more than other variables. The weak segregation of healthy and sick individuals indicates that there is no true difference in transcriptome response between healthy and sick individuals.

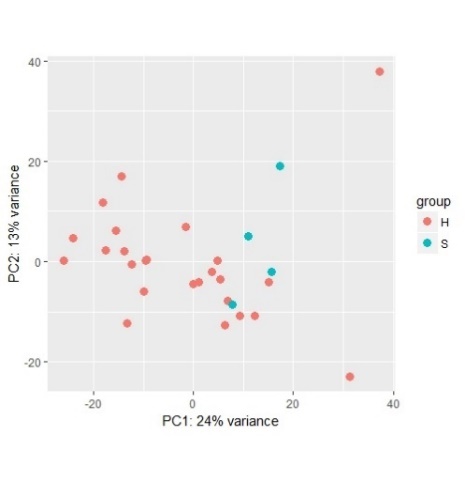
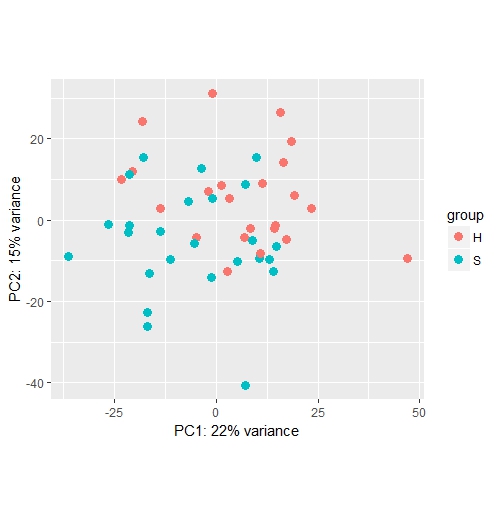
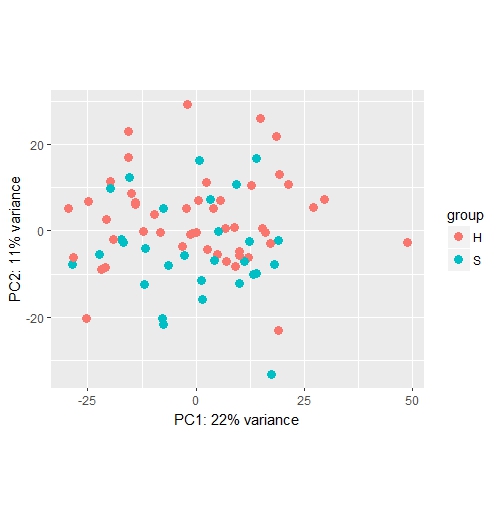
While more than 274 genes were differentially expressed between healthy and sick individuals, the PCA plots indicate that there is no difference in transcriptome response; however, the PCA includes all differentially expressed genes, of which, not all may be relevant to the health status of the individual. Therefore, it is necessary to determine the function of the differentially expressed genes and create a subset of candidate genes that may be relevant to the health status of individuals. To do this, the corresponding gene in the reference genome should be referred to if it has been annotated and if the reference genome is incompletely annotated, a Blast search in NCBI may be necessary to identify orthologues to determine gene function. Once a suite of candidate genes has been selected, I would rerun the PCA analysis using only the differentially expressed candidate genes to evaluate whether a certain family of genes can be predictors of health status. In addition to identify candidate genes, another interesting analysis would be to compare the gene expression of individuals just before and just after the become sick to identify genes that may be differentially expressed due to the onset of illness.

**Table 1:** Summary table listing, for each model, the numbers of up- and down-regulated genes, outliers, and samples with counts under 100 that were not included in the analysis. Genes were considered differentially expressed when p<0.1 and the most differentially expressed gene with its p-value are also reported for each model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **# Up Genes** | **# Down Genes** | **# Outliers** | **# Low Counts** | **Most Differentially Expressed Gene** | **P-value of DGE** |
| **Intertidal-only** | 205 (1.7%) | 37 (0.3%) | 0 (0%) | 9381 (76%) | DN43080\_c1 | 1.84e-06 |
| **Subtidal-only** | 20 (0.16%) | 113 (0.91%) | 647 (5.2%) | 4289 (35%) | DN42073\_c0 | 2.38e-06 |
| **Both Locations** | 209 (1.6%) | 65 (0.5%) | 400 (3.1%) | 7679 (59%) | DN43080\_c1 | 3.28e-07 |

  a. b. c.

**Figure 1:** Scatter plots of counts vs. health score of the most significantly differentially expressed gene for each model: a) Full data set controlling for location; most differentially expressed gene: DN43080\_c1 b) Intertidal-only data set; most differentially expressed gene: DN43080\_c1 c) Subtidal-only data set; most differentially expressed gene: DN42073\_c0.



1. b. c.

**Figure 2.** PCA plots depicting health-status groups for each of the three models with the percentage of the variance in the total data explained by each principle components labeled on the axes: a) Full data set controlling for location b) Intertidal-only data set c) Subtidal-only data set.