Laura Caicedo-Quiroga
Ecological Genomics
08-03-2017
Homework #2

Differential gene expression analysis of healthy and sick *Pisaster ochraceus*

A snapshot of gene expression can be analyzed effectively using transcriptomes of non-model species obtained by RNA-Seq. We used RNA-seq data from sea-stars, *Pisaster ochraceus*, collected at the Intertidal and Subtidal zones in Monterrey Bay, California at various time points and disease states. In this study, we see the differences in levels of expression of *P. ochraceus* between healthy and sick individuals analyzed by three models that account for location of collection.

RNA was extracted for sea-stars on common garden conditions once every 3 days during 15 days. Ilumina reads for 77 samples were assessed for quality using FastQC 0.11.5, then trimmed off poor quality reads (Phred Quality scores <30) and adapters using Trimmomatic 0.36. Clean reads were assembled (paired and concatenated) to a transcriptome with Trinity 2.4.0. They were then mapped to the transcriptome using BWA, which outputs sequence alignment map files. DESeq2 was used to analyze gene expression by using count data, the number of sequence fragments for each transcript. Three models were used: only samples from the Intertidal (Model 1), only samples from the subtidal (Model 2) and all samples while controlling for location (Model 3).

The amounts and ratio of significantly expressed genes are summarized in Table 1. Of the 12,399 genes expressed for Model 1, 1.7% were significantly up-regulated, while 0.3% were significantly down-regulated. While for Model 2, 0.16% of the 12,392 genes expressed were significantly up-regulated and 0.91% were significantly down-regulated. The model that considers all samples (Model 3) has 12,947 genes expressed with 1.6% significantly up-regulated and 0.5% significantly down-regulated. Genes expressed significant and non-significantly were plotted in terms of $\log_2$fold change for all samples under the three models. Most significantly expressed genes were up-regulated, when analyzing all samples (Fig. 1). The most significantly expressed gene in the Intertidal and for all samples was the same: DN43080. This gene saw a $\log_2$fold change of 2.5, about four times up-regulated than the rest. DN42073, was the most significantly expressed gene for the Subtidal with a $\log_2$fold change of -5.7, about ten times more down-regulated than the rest. Principal Component Analyses were performed for all models grouping samples by healthy and sick individuals; Model 1 showed the most prominent grouping of samples by health status (Fig. 2).

The intertidal zone shows greater gene expression, which can be explained by a greater proportion of sick samples than observed for the subtidal (most individuals that were healthy stayed healthy during the experiment). The results of gene expression for Model 1 and 3 are similar as most samples come from the Intertidal zone, therefore it is no surprise that the both show more genes significantly up-regulated than significantly down-regulated; meanwhile Model 2 shows the opposite trend. Sick individuals may be expressing greater up-regulation than healthy individuals. Figure 2 shows a grouping of individuals by their health status, which explains that some factors are contributing to a difference in gene expression between healthy and sick sea-stars. A comparison of read counts for the gene most significantly expressed and health score for all samples under Model 3 shows that counts increased while health worsened for sick individuals (Fig. 3). There is greater expression of genes in sick individuals, which increases as disease progresses.

The function of these genes is yet to be identified, this will give us more insight into the type of genes that are experiencing greater regulation; we expect increased up-regulation of genes associated with stress and immune responses. It would be interesting to compare these results with studies on gene expression of

healthy sea-stars exposed to different environments to understand the differences in expression related to stressors other than sickness.

Model	Genes significantly Up-regulated	Percent of total expressed genes up-regulated	Genes significantly Down-regulated	Percent of total expressed genes down-regulated
Intertidal (Model 1)	205	1.7%	37	0.3%
Subtidal (Model 2)	20	0.16%	113	0.9%
All (Model 3)	209	1.6%	65	0.5%

Table 1: Summary of differentially expressed genes for the 3 models used.

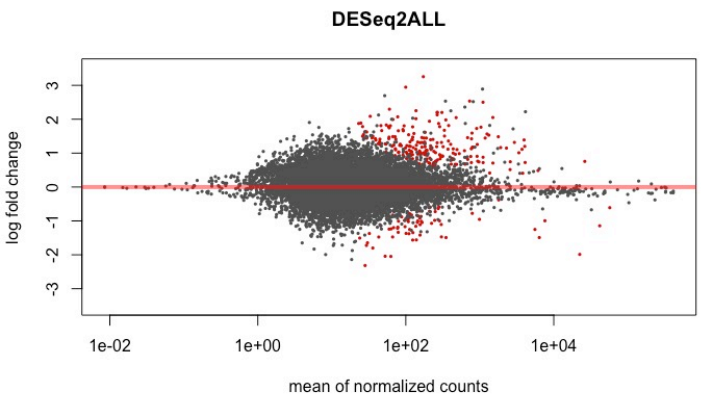

Figure 1: Differential gene expression seen for Model 3 seen as log_2fold change (y-axis) vs mean of normalized counts (x-axis). Each dot represents a gene, red dots represent significant expression, grey dots are genes not significantly expressed.

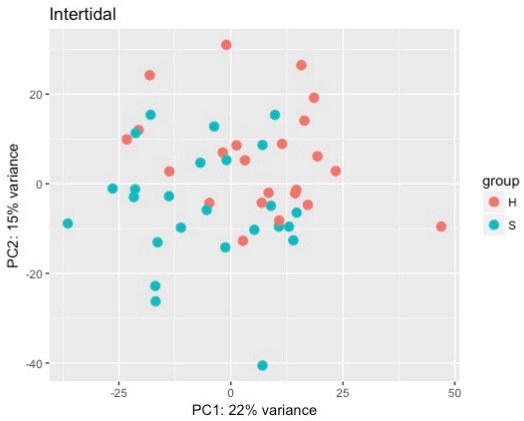

Figure 2: Principal Component Analysis plot for Model 1. PC1 and PC2 explain the difference in gene expression by healthy and sick individuals.

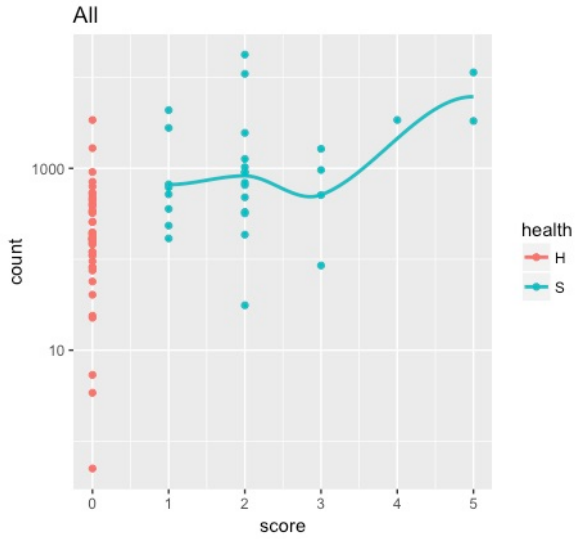

Figure 3: Plot of reads (counts) for gene DN43080 on the y-axis and health score on the x-axis of Healthy (H) and Sick (S) individuals (dots). A line of best fit is adjusted for Sick individuals.

Script at: https://github.com/lracaicedo/EcoGenomics/blob/master/Laura%20Caicedo_notebook.md#id-section15