Independent Project - Abalone Gene Expression

Section 1: Technical Report

Introduction As the climate changes and heat wave events increase in frequency, a rising body of literature is directing efforts in improving the understanding of molecular signatures for thermal tolerance. Wild and cultured abalone are susceptible to high water temperatures and increased mortality rates are recorded in the summer months. A recent study by Shiel et al. (2020) investigated gene expression profiles of greenlip abalone (*Haliotis laevigata*) exposed to four temperature treatments in laboratory conditions in order to identify genes responsible for resilience to heat stress. Ancestral origin of the abalone used for the experiment was traced back to Farm Beach (F); Elliston (E); and Port Lincoln, South Australia (S). RNA was extracted from thirty-five abalone with an even number of susceptible (S) and unsusceptible (U) individuals selected from six tanks. In the present study this data was subset and only two temperature treatments (18°C and 21°C) were analysed, with the aim of evaluating gene expression levels in susceptible and resilient abalone. This work is especially significant for the aquaculture industry looking to selectively breed abalone stocks based on their ability to withstand heat stress.

Results Exploratory data analysis using PCA showed that location and tank might have an effect in differential gene expression (Fig. 1). In particular location S (Port Lincoln) seems clustered separately from other locations (Fig. 1B) and tank 5 is seen relatively separated from other tanks, therefore it is expected that differences between location and tank will affect gene expression.

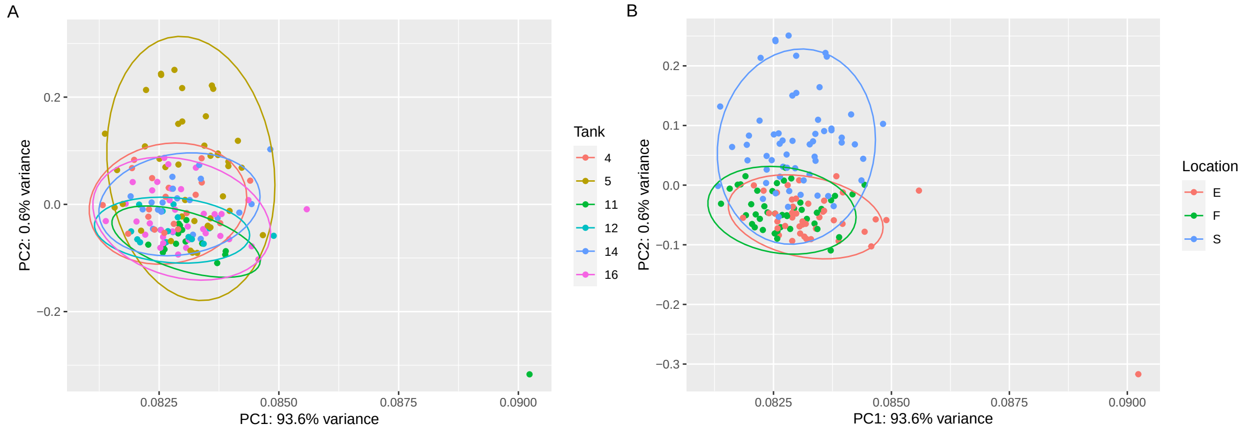


Figure 1: PCA plot shows the overall effect of the experimental factors (A) Tank and (B) Location by variation explained from two principal components

To account for the effect of tank and location in our statistical model, they are analysed as fixed effects along with condition after exposure to treatment. The effect of abalone genotype was also tested for association with gene expression using PCA visualization, but since no relationship was found this factor was not included in the model. Model formula:

$$\sim Location + Tank + Condition$$

42 and 64 genes were found to be differentially expressed in susceptible and resilient abalone sampled at 18 and 21 °C respectively. As shown in Fig. 2A, the most statistically significant gene was found in abalone sampled at 18 °C and this was an upregulated gene. In the lower temperature treatment (18°C) a higher

number of up-regulated genes were observed in resilient abalone compared to down-regulated genes, namely 25 and 17 respectively (Fig. 3). However the pattern was reversed in the higher temperature treatment (21°C), where the number of down regulated genes (n = 36) in resilient abalone exceeds that of up regulated genes (n = 28)(Fig.3).

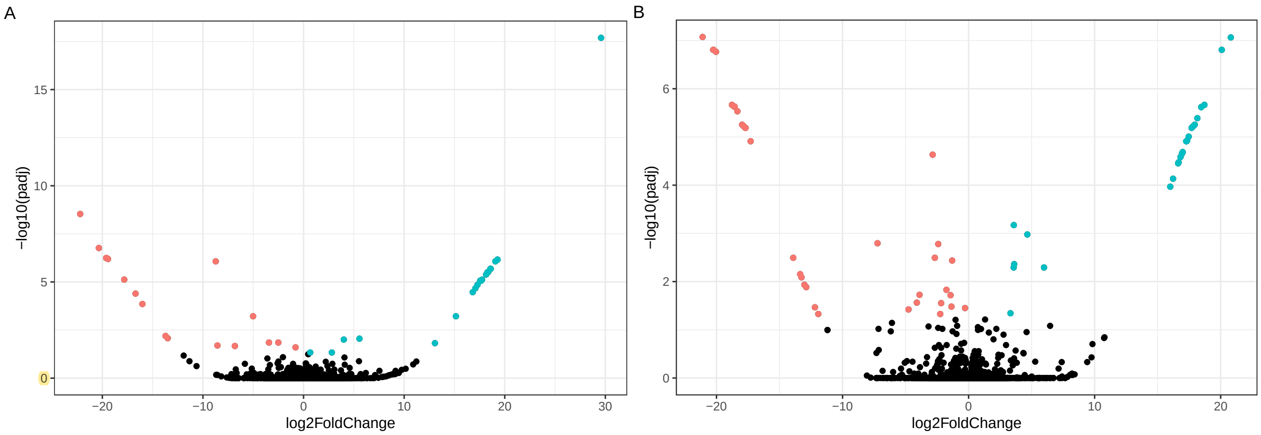


Figure 2: Volcano plots showing significantly differentially expressed genes (upregulated genes shown in light blue and downregulated genes in pink) found from the 18 °C temperature treatment (A) and from the 21 °C temperature treatment (B).

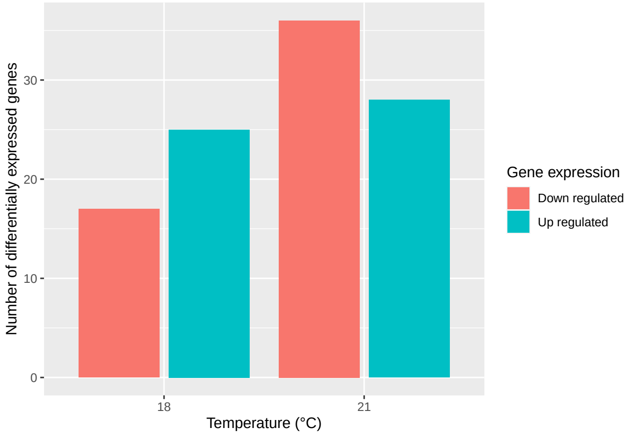


Figure 3: Bar plot demonstrates number of differentially expressed genes for temperature treatments 18 and 21 °C. Upregulated genes are shown in light blue and downregulated genes in pink.

Discussion

Abalone from Port Lincoln were selectively bred in captivity for enhanced growth rate (Shiel et al. 2020). This may explain why gene expression differed for abalone originated from this location compared to other locations. In *Acropora* sp. for example, a life history trade-off is observed between fast growth and heat tolerance (Gold & Palumbi 2018). Although in corals growth and heat stress tolerance is influenced by differences in *Symbiodiniacea* communities, it is possible that selectively bred abalone for fast growth have less heat resistance genes.

The original study found higher numbers of significantly differentially expressed genes in resilient abalone collected from 18°C compared to resilient abalone in 21°C (Shiel et al. 2020). The fact that the current study did not account for the effect of each individual abalone genotype in the statistical model may be the reason why opposing results were found. Additionally, the four different temperature treatments were not accounted for in the present model, which is another factor that can lead to different results. The subsampling method used for the current study reduced the sample size and some differentially expressed genes between temperature treatments may have been missed from the statistical analysis. Nevertheless, the values obtained from Shiel et al. (2020) are relatively similar to current results, thus providing a rough estimate of gene expression differences between resilient and susceptible abalone in low and high temperature treatments.

Frontloading genes are known to be differentially expressed in stress resilient individuals, compared to susceptible individuals prior to heat stress exposure (Shiel et al. 2017). These genes have been found in heat resilient abalone across multiple populations and can potentially be used to predict susceptibility to heat stress (Shiel et al. 2020). Consistent with findings from Shiel et al. (2020), the present study found a higher number of upregulated genes in the low temperature treatment compared to downregulated genes, which may be some evidence of frontloading. The allocation of resources towards energy processes or cellular metabolism can be an example of "preparative defense" in resilient abalone (Shiel et al. 2020).

A greater number of upregulated genes were found in the higher temperature treatment compared to the lower temparature treatment, which may indicate immune and stress response gene signatures. For example, the original study found significantly higher expression of peroxidases, an antioxidant enzyme in the highest temperature treatment 21°C (Shiel et al. 2020). Interestingly, more genes were down regulated rather than upregulated in resilient abalone collected at the high temperature treatment, namely 36 and 28 genes respectively. These genes may indicate activation of transposable elements. Shiel and collaborators (2020) found transposase activity to be the most common molecular function of differentially expressed genes in their study and it was associated with the downregulation of comp106280_c0 gene in resilient abalone. Transposable elements are able to change gene function and have been proposed to increase host survival in the face of environmental stress (Levin & Moran 2011).

In conclusion, significant differences in gene expression identified between resilient and susceptible abalone provide clues to survival mechanisms during heat stress events. Some evidence suggests frontloading abilities in resilient abalone prior to heat exposure, which can be important information for future selective breeding programs aiming to improve abalone resilience to heat stress. However, future investigations are required to evaluate and confim gene functions in resilient abalone.

Section 2: Reproducible Analysis

Load the necessary libraries and read in the data

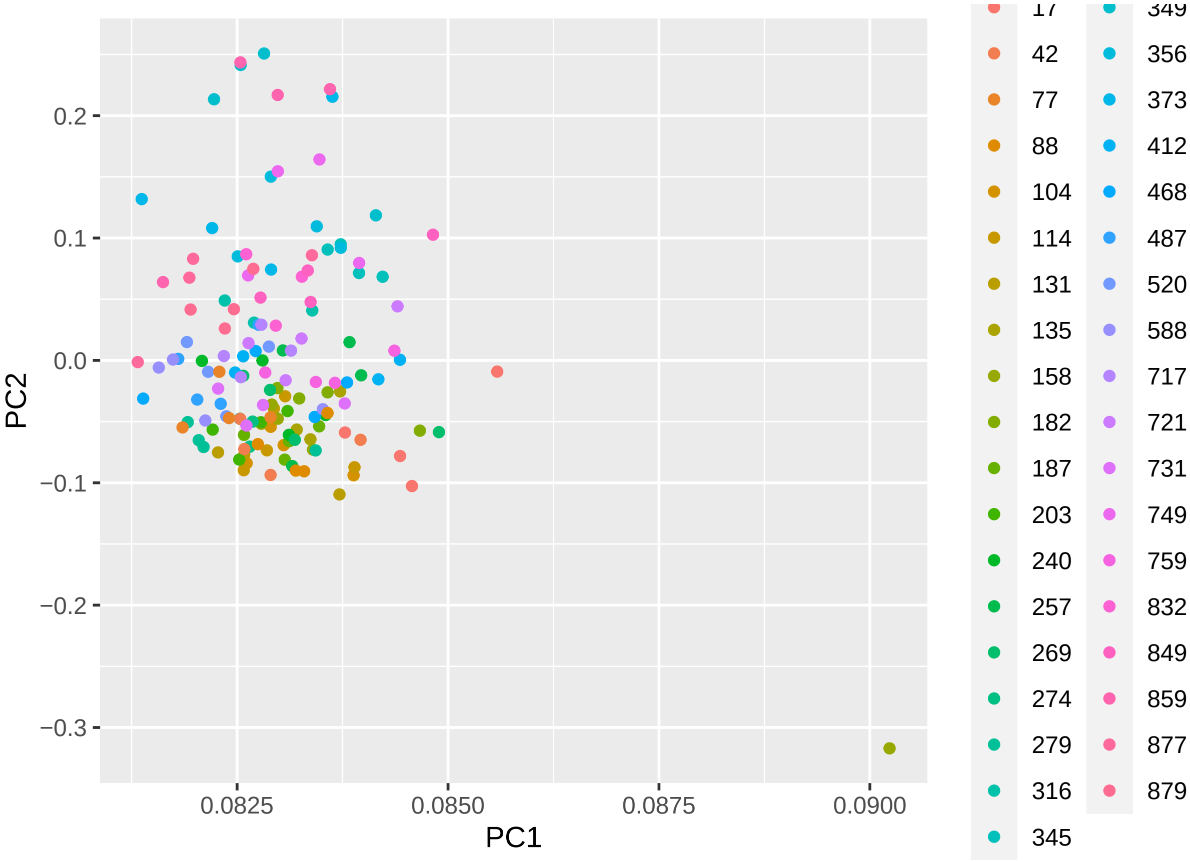
```
wget 'https://cloudstor.aarnet.edu.au/plus/s/P3AMORiB2ZrbywE/download' -O data.tgz
tar -zxvf data.tgz
```

```
counts <- read_tsv("raw_data/Abalone/count_table.tsv")
metadata <- read_tsv("raw_data/Abalone/metadata.tsv")
```

Exploratory data analysis To visualize the raw data and identify any trends before choosing the appropriate model, a principal component analysis is used with no design matrix set for DESeq. PCA makes the assumption that there is no unique variance and that the total variance is equal to common variance. For this reason, a variance stabilising transformation is performed.

```
# Change rownames to be transcript ID
counts <- counts %>% remove_rownames %>% column_to_rownames("transcript_id")

# Imports data to DESeq without setting a design matrix
dds <- DESeqDataSetFromMatrix(counts,metadata, design = ~ 1)

# Performs a variance stabilising transform to make data suitable for plotting with a PCA
vst <- varianceStabilizingTransformation(dds)

# Perform a PCA on the vst data
```

```
pcdata <- prcomp(assay(vst))

# Extract the rotation matrix from the PCA and join it with metadata
pcdata_meta <- pcdata$rotation %>% as.data.frame() %>%
  rownames_to_column("Sample") %>%
  left_join(metadata)

# Visualize PCA for each of the experimental factors (Tank, Location,
# Condition, Time and Abalone).
ggplot(pcdata_meta,aes(x=PC1,y=PC2)) +
  geom_point(aes(color=as.factor(Abalone)))
```



Here the experimental factor Tank is plotted using PCA. The PCA of the other factors requires a change in the argument of geom_point. Four more ggplots are produced with Tank, Location, Condition and Time being visualized with color instead of Abalone. An outlier was identified from the PCA's, with approximate coordinates 0.09 and -0.3 for PC1 and PC2 respectively. The sample name for this outlier (ES_11_158_4) was found in pcdata_meta using these coordinates. In the next step, data from abalone 158 will be removed.

Data wrangling

```
# Remove outlier abalone number 158
metadata <-subset(metadata, Abalone!=158) %>%
# Tranform all categories of the metadata into factors
  mutate(Location = factor(Location), Condition = factor(Condition), Tank = factor(Tank))

# Ensure that the column order in counts is the same as row order in metadata
counts <- counts[,metadata$Sample]
```

```
# Subset the data used for the first and second model
metadata1 <- metadata %>% filter(Time == 2)
counts1 <- counts[,metadata1$Sample]
metadata2 <- metadata %>% filter(Time == 4)
counts2 <- counts[,metadata2$Sample]
```

Model fitting

```
# Create a DESeq object for both models
ddsMF1 <- DESeqDataSetFromMatrix(counts1,colData = metadata1,
                                 design = ~ Location + Tank + Condition)
ddsMF2 <- DESeqDataSetFromMatrix(counts2,colData = metadata2,
                                 design = ~ Location + Tank + Condition)

# Perform normalization and model fitting
ddsMF1 <- DESeq(ddsMF1)
ddsMF2 <- DESeq(ddsMF2)
```

Model investigation / hypothesis testing After model fitting we investigate the results from the model to find significantly differentially expressed genes from two temperature treatments. The *results* function will perform independent filtering based on the mean of normalized counts for each gene and optimizing the number of genes which will have an adjusted p-value below 0.05. In other words, the probability of these genes to be incorrectly identified as differentially expressed is below 5%.

```
# Access results of DESeq2 analysis for model 1 and 2.
resMF1 <- results(ddsMF1)
resMF2 <- results(ddsMF2)

# Calculate significantly differentially expressed genes from model 1
signif_genes1 <- resMF1 %>%
  as.data.frame() %>%
  rownames_to_column("transcript_id") %>%
  arrange(padj) %>%
 filter(padj<0.05)

signif_genes1 %>%
  group_by(transcript_id) %>%
# Count number of significantly differentially expressed genes
# from temperature treatment 18 degrees
  n_distinct()
```

```
## [1] 42
```

```
# Calculate significantly differentially expressed genes from model 2
signif_genes2 <- resMF2 %>%
  as.data.frame() %>%
  rownames_to_column("transcript_id") %>%
  arrange(padj) %>%
 filter(padj<0.05)
signif_genes2 %>%
  group_by(transcript_id) %>%
# Count number of significantly differentially expressed genes
# from temperature treatment 21 degrees.
  n_distinct()
```

```
## [1] 64
```

Next we are interested to count the number of upregulated (log2FoldChange>0) or downregulated (log2FoldChange<0) genes in both temperature treatments to finally produce a barplot as a summary figure.

```
# Calculate number of upregulated and downregulated genes between
# heat stress resilient abalone relative to susceptible abalone
# collected at 18 degrees celsius.
signif_genes1 %>%
  group_by(transcript_id) %>%
  filter(log2FoldChange>0) %>%
  n_distinct()
```

```
## [1] 25
```

```
signif_genes1 %>%
  group_by(transcript_id) %>%
  filter(log2FoldChange<0) %>%
  n_distinct()
```

```
## [1] 17
```

```
# Calculate number of upregulated and downregulated genes between
# heat stress resilient abalone relative to susceptible abalone
# collected at 21 degrees celsius.
signif_genes2 %>%
  group_by(transcript_id) %>%
  filter(log2FoldChange>0) %>%
   n_distinct()
```

```
## [1] 28
```

```
signif_genes2 %>%
  group_by(transcript_id) %>%
  filter(log2FoldChange<0) %>%
  n_distinct()
```

```
## [1] 36
```

Summary figures The dataframes obtained from the *results* function are used next to generate a volcano plot (Fig. 2). This is an important step used to visualize (1) which model produced the most statistically significant differentially expressed genes and (2) if these genes were upregulated or down regulated respective to each temperature treatment.

Lastly, the number of upregulated and down regulated genes in resilient abalone for both temperature treatments as calculated above are summarized in a bar plot (Fig. 3).

References

Gold, Z. & Palumbi, S.R. 2018, “Long-term growth rates and effects of bleaching in Acropora hyacinthus”, *Coral reefs*, vol. 37, no. 1, pp. 267-277.

Levin, H.L. & Moran, J.V. 2011, “Dynamic interactions between transposable elements and their hosts”, *Nature reviews. Genetics*, vol. 12, no. 9, pp. 615-627.

Shiel, B.P., Hall, N.E., Cooke, I.R., Robinson, N.A. & Strugnell, J.M. 2017, “Epipodial Tentacle Gene Expression and Predetermined Resilience to Summer Mortality in the Commercially Important Greenlip Abalone, Haliotis laevigata”, *Marine biotechnology (New York, N.Y.)*, vol. 19, no. 2, pp. 191-205.

Shiel, B.P., Cooke, I.R., Hall, N.E., Robinson, N.A. & Strugnell, J.M. 2020, “Gene expression differences between abalone that are susceptible and resilient to a simulated heat wave event”, *Aquaculture*, vol. 526, pp. 735317.