Ecological Genomics: Homework #2

Brendan Case

https://github.com/brendandaisy/ecological-genomics

27 March 2020

1 Background

Red spruce (*Picea rubens*) is a coniferous tree which has become highly isolated due to rising temperatures. While generally favoring cold, wet environments, as its range decreases it occasionally is forced to inhabit less ideal dry and warm environments. It is thus of conservation interest to understand the ability of this species to respond to stressful climates.

To study the effect of climate stress on these two ecotypes, we took seedlings from both groups and raised them in common garden conditions before dividing them into 3 treatment groups: control, heat (%50 increase in day/night temperature), and heat+drought (temperature treatment plus water withholding). We then performed RNA extraction from seedling tissue on days 0, 5, and 10 of exposure.

2 Bioinformatics Pipeline

Fastq files were inspected for quality using `fastqc`, then trimmed using the Java program `trimmomatic`. The trimmed reads were then mapped to a *P. abies* reference genome using Salmon's `index` command. We then obtained the complete counts matrix using Salmon's `quant` command on each sample, which were combined using `tximport` in R.

All analyses on the resulting counts matrix were performed using `DESeq2` in R. Briefly, log2 fold changes were found using maximum likelihood estimation of the following Negative Binomial GLM:

$$
\begin{aligned}
K_{ij} &\sim NB\left(\mu_{ij}, \alpha_i\right) \\
\mu_{ij} &= s_j q_{ij} \\
\log_2(q_{ij}) &= x_j \cdot \beta_i,
\end{aligned}
$$

where K_{ij} are the observed raw counts for gene i in sample j, μ_{ij} is the mean of the negative binomial distribution, α_i is a gene-wide dispersion parameter, s_j is a sample-specific size factor, and q_{ij} is the expected true proportion of gene i out of all expressed genes in sample j. The log of q_{ij} is given as the sum of fold changes β_{ir}, where r indexes

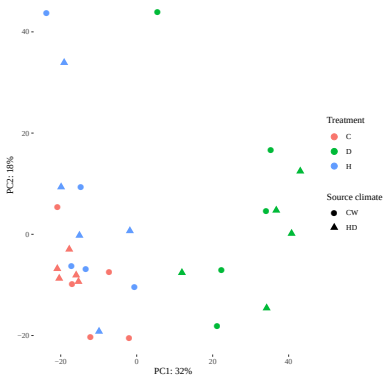

Figure 1: The first two components from a PCA of gene expression. The 3 experimental groups are given different colors, while the shape of replicates corresponds to source climate.

the experimental levels of sample j. Hence, β_{ir} gives the effect of treatment condition r on the expression of gene i [1]. In order to have a symmetric experimental design and maximize our ability to detect a $G \times E$ interaction, we used only the day 10 samples, and used `population + treatment + population:treatment` as our experimental factors.

3 Results

Among the day 10 samples, we detected expression of different 66,408 genes. After normalization, there were 7,277 reads per gene on average, and a median of 199 reads per gene, suggesting a high amount of skew in the level of expression between genes. After filtering out genes which on average had less than one read per sample, there were 17,487 genes with which to perform differential expression analysis.

We found notable differentiation in expression between treatment groups. In particular, the PCA shown in Figure 1 revealed different expression levels in the heat + drought treatment. However, the first two components did not pick up on differences between source climates.

The number of significantly expressed genes for various contrasts was computed (not shown). Notably, we found that 559 genes showed a significant "G by E" interaction; that is, these genes showed different responses to either the heat or heat+drought treatment, depending on their source climate. Of these genes, 231 were exclusively picked up in the climate $\times$ heat interaction, while 312 genes were exclusively picked up in the climate $\times$ heat + drought interaction. Thus, there were only 16 genes with a significant non-linear expression between climates and both stress treatments.

4 Conclusion

We found a number of DE genes between samples from different source climates and their response to climate stress. Our results indicate that not only are there widespread

differences in expression when saplings are exposed to both heat and drought stress, but that a number of genes may exhibit complex responses to these conditions depending on their climate/population of origin. Future work should investigate the functional ontology of these genes, as well as patterns of gene expression in response to stress over time.

References

[1] M. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.