Transcriptomic response of two climatically distinct populations of *Picea rubens* under heat and drought stress

Thomas O'Leary

Background

Transcriptomic multiple stressors experimental design

Indivduals of *Picea rubens*, a red spruce that thrives in cool and moist climates, are increasingly experiencing periods of environmental stress, including excessive heat and drought in some of the areas in its range (Ribbons). This may be due to microclimate differences, such as increased sun exposure and drier soil, but may also be due to larger scale changes in regional climate (Dumais and Prévost). Maternal families from two different source climates, Hot & Dry (N = 5) or Cool & Wet (N = 5)) were tested.[1] Individuals from these families were exposed to three experimental treatments: a control, heat stress, and combined heat and drought stress. Each treatment included a 16:8 hour light to dark photoperiod, with variable day and night time temperatures for a period of ten days. The control involved alternating temperatures of 23°C and 17°C. Temperatures were increased by 50% to 35°C and 26°C in the heat treatments. Individuals in the heat and drought treatment were additionally withheld water for the duration. The main aim of this study was to decribe the transcriptomic response of *P. rubens* under temperature and drought stress and note differences between indivduals from different source climates, potentially identifying candidate genes of functional consequence in response to environmental stress. For this write up, I will narrow my focus to the individuals sampled at day ten.

Sample collection, library preparation, and sequencing

Individuals from each group were germinated and raised to seedlings for ten weeks in common garden conditions, then transfered to their respective growth chambers and left under control condidtions for two weeks before the experimental treatments were engaged. At each collection time point, day zero, five, and ten, RNA was extracted from whole seedlings (*i.e.* root, stem, and needle tissue). After extraction, RNA quantity and quality was assessed on a Bioanalyzer. Samples (RNA concentration > 1 ng/μL) were sent to Cornell for library preparation and sequencing. Library preparation followed the Lexogen QuantSeq 3' mRNA-Seq protocols and they were sequenced in 86 bp reads on a NextSeq500.

Bioinformatics Pipeline

The quality of the raw reads were accessed using FastQC v0.11.3 (Andrews et al.). Low quality sequence was trimmed from the raw reads using Trimmomatic v0.33 (Bolger et al.) and the cleaned reads were vizualized again with FastQC. Those cleaned reads were then mapped using Salmon v1.1.0 (Patro et al.) to a Norway spruce, *Picea abies*, reference transcriptome (Congenie.org), which included all coding sequence and two kilobases downstream to account for the 3' tag sequencing technique. Transcript count matrices were imported into R v3.6.1 (R Core Team) with the package tximport v1.12.3 (Soneson et al.). Genes with an average of less than one transcript per sample were pre-filtered from the analysis.[2] Differential gene expression was called and an FDR of 0.05 using DESeq2 v1.24.0 (Love et al.). Further analysis and visualization was done using the packages ggplot2 and dplyr v1.2.1 (Wickham).

Results

Salmon mapped the transcript sequencing reads to the reference with a mean mapping rate of 52% (minimum 30%; maximum 71%).[3] Adding two kilobases downstream to the coding sequence, dramatically improved the mapping rates over only including the coding sequence, because of the 3' tag sequencing method that was used. There was a mean of 2.7 million reads mapped per sample (range: 1.0 million – 4.4 million). Principal component analysis of **all** expressed genes shows relatively overlapping clusters of samples based on treatment condition for the first two principal components, with hardly any separation based on source climate (Fig. 1 **A**).[4]

[1] The source climate was determined based on historical climate data (*e.g.* precipitation in the warmest quarter and the temperature of the hottest month).

[2] Although DESeq2 uses its own independent filtering to reduce the amount of genes with low counts, this pre-filtering was done to speed up the analysis and ensure that genes with very low transcript counts, and likely high-noise, were not included in results.

[3] These mapping rates are not atypically low for a non-model organism using a reference sequence from a closely related species.

[4] I tried the PCA with subsets of the most variable genes (as the default for `plotPCA()` allows). This did seem to qualitatively change the clusters – the fewer genes included, the more Heat and Heat & Drought cluster together away from the control.

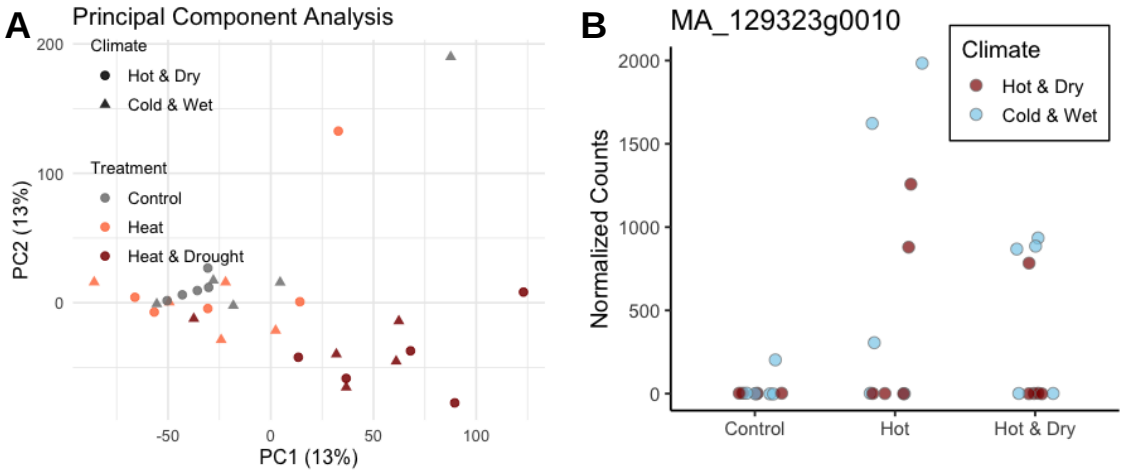

Figure 1: **(A)** Principal component analysis of all expressed genes. **(B)** Normalized transcript counts of the gene MA_129323*g*0010.

When comparing the two source climates, across all three treatments, there was only one differentially expressed gene, MA_129323*g*0010 (uncharacterized protein function), which was down-regulated in the Hot & Dry source climate when compared to the Cool & Wet source climate (Fig 2 **B**; L2FC = −20.64; FDR = 0.04). Additionally, there were no genes found to have an interaction between the two factors of climate and treatment. However, there were more genes differenitally expressed when comparing the stress treatments to the control. In heat stress, there were 18 up-reulated genes and 2 down-regulated. In the case where water was also withheld, the number of differentially expressed genes increased dramatically, with 776 genes up-regulated and 585 down-regulated on day ten in response to heat and drought stress, when compared with the day ten controls.

Together these results would seem to suggest that the transcriptomic response of *P. rubens* is largely driven by the environmental conditions and that their demographic background (source climate) contributes little to any differenital expression. However, after collapsing treatment and source climate factors and comparing among source climates each stress treatment to its own control, and interesting pattern of differential expression emerges between source climates (Fig. 2). There are many more differentially expressed genes in Hot & Dry source climate individuals than in the Cool & Wet group. One of the genes differentially expressed in response to heat and drought only in the Hot & Dry source climate is MA_135205*g*0010 (L2FC = 4.13, FDR = $1.7\text{x}10^{-9}$), a gene which produces a 70 kDa peptidyl-prolyl isomerase.

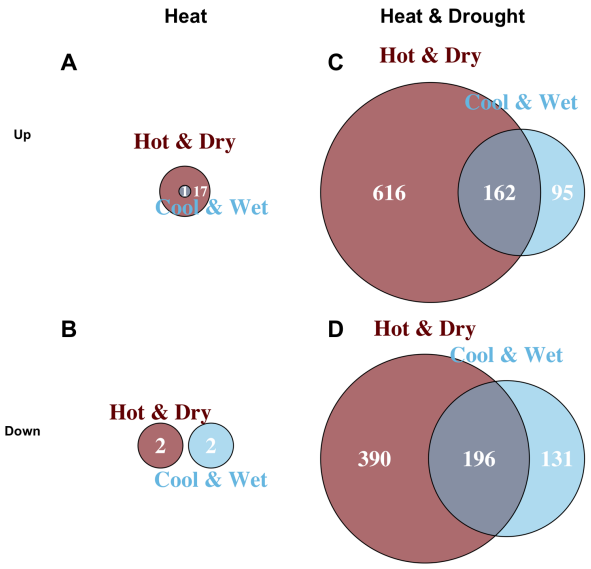

Figure 2: Venn diagram of genes that are up-regulated **(A)** and down-regulated **(A)** under heat stress, and genes that are differentially up-regulated **(C)** and down-regulated **(D)** in responsed to combined heat and drought stress (FDR < 0.05).

Conclusion

Vitally important to the conservation of *Picea rubens* along the eastern United States is understanding the ability of different populations to respond to environmental challenge. These preliminary results may suggest that individuals from Hot & Dry climates have a larger transcriptomic response to heat and drought stress than individuals from Cool & Wet climates, although the effect of this transriptomic response on the whole organism phenotype remains unexplored.

Future directions with this data set should involve exploring how the transcriptomic response changes over the course of the treatment and diving into the physiology of the specific candidate genes. Overrepresentation analysis on the differentially expressed gene sets may reveal functional categories that are enriched in the plastically responding genes and potentially identify pathways invovled in the greater responding Hot & Dry source climates. Additionally, further experimentation could explore the same familes under these stressful conditions for a longer period of time and making phenotypic measurements (*e.g.* growth, budburst) and performing a genome-wide association of the maternal tree lines. Incorporating these two data sets, would potentially allow insight into the genomic architure underlying this possible signal of source climate differentiation.

References[5]

Andrews, Simon, et al. *FastQC*. Babraham Institute, Jan. 2012, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bolger, Anthony M., et al. "Trimmomatic: A flexible trimmer for Illumina sequence data." *Bioinformatics*, vol. 30, no. 15, 2014, pp. 2114–20, doi:10.1093/bioinformatics/btu170.

Dumais, Daniel, and Marcel Prévost. "Germination and establishment of natural red spruce (Picea rubens) seedlings in silvicultural gaps of different sizes." *Forestry Chronicle*, vol. 92, no. 1, 2016, pp. 90–100, doi:10.5558/tfc2016-021.

Love, M. I., et al. *Differential analysis of count data - the DESeq2 package*. 2014, p. 550, doi:110.1186/s13059-014-0550-8.

Patro, Rob, et al. *Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference*. no. 4, 2017, pp. 417–19, doi:10.1038/nmeth.4197.Salmon.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2019, https://www.r-project.org/.

Ribbons, Relena Rose. "Disturbance and climatic effects on red spruce community dynamics at its southern continuous range margin." *PeerJ*, vol. 2014, no. 1, 2014, doi:10.7717/peerj.293.

Soneson, Charlotte, et al. "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." *F1000Research*, vol. 4, no. 1521, 2015, doi:10.12688/f1000research.7563.1.

Wickham, Hadley. *tidyverse: Easily Install and Load the 'Tidyverse'*. 2017, https://cran.r-project.org/package=tidyverse.

[5]General musings that I want to get down while I am thinking about them. It is probably worth talking to someone about this: I wonder what is the real difference between these two DESeq design methods (*i.e.* **(i)** exp ~ climate + treatment + climate:treatment, and **(ii)** exp ~ group, where group is the collapsed together climate and treatment factors). I guess I get conceptually how they are different, but it seems to me that upon analysis of the data these two approaches produce somewhat competing ideas of what is going on with these samples. The first says that there is only one gene different between the two source climates (Fig. 1 **B**). And the second seems to suggest that there is really quite a bit of difference in plastic response to stress between the two source climates (Fig. 2). So is the first approach testing for DE in source climate regardless of treatment, and then treatment regardless of climate, and then the interactions? It seems to me like collapsing the factors gets at the questions we are asking more? Because collapsing the factors allows you to ask more pairwise questions (*e.g.* What are the DEGs under Heat & Drought in the Hot & Dry source climate samples, compared to *their* control? And does that differ from the DEGs identified in the Cool & Wet group under the same stressors?). It seems like comparing each stress group to its own control is a better approach. Maybe the first experimental design is doing that under the hood, but I don't think so? I don't know. Just a bit confused the more I think about it… Took a walk. I guess if you were to run the exact same experiment, but measure some sort of macro-phenotype (*e.g.* trunk diameter) then you would do a two-way ANOVA on the two factors, climate and treatment, which is probably the equivalent of ~ climate + treatment + climate:treatment design. So I guess what I am doing in exp ~ group is more like doing a bunch of pairwise t-tests, which is maybe not the best approach and may lead to false positives if you don't correct the p-value for multiple comparisons like in Tukey's ad hoc. So are all those hundreds of genes that are identified really just false-positives? Food for thought. Munch munch. Maybe a more strict FDR would be appropriate. I would be interested to hear what you think? From a few Stackoverflow answers, it seems like Michael Love is a big fan of collapsing the factors and simplifying the design, but now I don't know what I think about it now.