Homework 2 | Csenge Petak | 03/25/2020

**Background** Red spruce (*Picea rubens*) is aspecies of conifers native to eastern North America that thrive in humid, cool climates [1]. According to panylogical and paleoecological studies, following the last glacial maximum around 18,000 years BF spruce species migrated northward from the Great Plains. As the temperatures rose and the ice melted south-eastern red spruce populations (like in West Virginia) experienced habitat fragmentation as they became more and more confined to high elevation mountains of the Appalachians [2]. It is believed that this geographic isolation caused many events of allopatric speciation in the genus *Picea,* for example, genetic data suggests that red sprucespeciated from the black spruce this way [3]. As populations experiencing habitat fragmentation and consequently reduced or absent gene flow become more vulnerable to sudden changes in environmental conditions and because these species prefer cool and moist climates, these red spruce populations could be endangered by current global climate change conditions.

This study aims at uncovering genetic resource represented by these fragmented edge populations and facilitating informed decision-making regarding conservation efforts given their hypothesised susceptibility to climate change and important roles in mountain communities [1]. The Keller Lab collected seeds and needle tissue samples from trees across the Appalachian Mountains, here I am going to discuss analysis of one population called XCV, one of the isolated edge populations mentioned above. XCV contained 5 samples. Extracted whole genomic DNA was used for exome capture sequencing, for which 80,000 120bp probes were designed using the closely related white spruce (*P. glauca*) transcriptome. Library preparation involved mechanical shearing of DNA (average resulting fragment size = 400 bp), ligation of barcoded adapters and PCR-amplification. A single run of an Illumina HiSeq X was used to generate paired-end 150-bp reads.

**Bioinformatics Pipeline**

**Trimming** To trim raw data and remove adapters the fast, multithreaded command line tool Trimmomatic (version 0.33) was used, which is a flexible trimming tool for Illumina next-generation sequencing (NGS) data [4]. First, adapter and other illumina-specific sequences were cut from the reads, then bases below the threshold quality = 20 were cut off from the start and the end of a read. Finally, reads were scanned with a 6-base wide sliding window and were cut when the average quality per base dropped below 20. Trimmed reads shorter than 35 were dropped. Before trimming the average number of sequences per sample (from both R1 and R2) was 3,088,690.8, which decreased to 2,909,506.8 based on FastQC analysis. Trimmed reads where then used for mapping (see Table 1).

**Mapping** Cleaned reads from each sample were mapped to the Norway spruce (*P. abies*) reference genome (full size = 19.6 Gb, N50 = 4869 bp)[5]. Based on BLAST search, the full reference genome was subsetted to only include contigs containing one or more probes used for the exon capture experiment in order to decrease the computational power needed for mapping. The reduced reference contained ~668 Mbp in 33,679 contigs. Sequence alignment files were generated using the Burrows-Wheeler Alignment Tool (BWA), and specifically the MEM algorithm, which was chosen because it has a split alignment feature and it is faster and more accurate than the other algorithms [6]. The tool was run with the option to allow reads to map to different contigs, as the state of the assembly is such that the location of different contigs on the chromosomes is not known, therefore reads mapped across two contigs could actually be close. Alignments containing unpaired reads were kept. The output SAM file was converted into a more efficient binary version (BAM) for further analysis using the sambamba command line tool (version 0.7.1) [7]. The same tool was then used on the BAM file to remove PCR duplicates. Mapping statistics were generated using samtools [9] and are included in Table 1.

**Analysis of Next Generation Sequence Data (ANGSD)** ANGSD is a multithreaded program suite that can perform various population genetic analyses either by using raw data directly or genotype likelihoods [8]. We used this program to estimate site frequency spectrum (SFS) and nucleotide diversities (Watterson’s estimate, , Tajima’s D) based on genotype likelihoods instead of “hard called” genotypes, because individuals seemingly homozygous to a certain SNP could actually be heterozygous if we didn’t have enough coverage at that site. By using genotype likelihoods this uncertainty is incorporated in the statistics. Parameters were set to exclude lower quality bases (min Phred = 20), reads that mapped poorly to the reference (min Phred = 20) and regions with not enough read depth (min = 3). Sites with more than 2 alleles were excluded and p-value threshold for SNPs was 1e-6. Since the ancestral state for SNPs couldn’t be determined with high confidence, folded SFS was calculated (where more frequent alleles are assumed to be the ancestral state).

**Results** When all data is considered, (climate+treatment+day) based on PCA results PC2 showed that daz 10 clusters but not climate and also drought clusters from the other 2, overall climate no effect, and drought treatment at day 10 is partitioned bz PC2) 16% variation. Thus we decided to separate data from daz10 and look at the effect of treatment and climate in that subset of the data.

Depite this decision, I decided to include an interesting observation from using all data with design ~climate + treatment + climate:treatment: comparing the effect of climate of origin across all treatments 1 gene was found to be significantlz differentially regulated, **Class VII chitinase (**MA\_10431378g0010, p = 4.97e-10). This gene’s expression was much higher in individuals from the hotdrz climate (15.96 log fold change). Interestingly, there was a significant interaction of the effect of climate and treatment on the expression of this gene: it was onlz upregulated in individuals from the hotdrz climate under the drz+hot treatment (p = 1.585e-8).

Regarding the alazsis of onlz data from daz 10, there were on average 2869038 reads per sample with an average number of counts per gene of 1296.096, where the median was 10. This shows that the number of counts per gene is not normallz distributed (right-skewed) and some gene had counts orders of magnitude higher. Finally, the average number of counts per gene per sample was 43.2. Using onlz data from daz 10, PCA showed the same pattern: climate of origin didn’t seem to have an effect while PC1 (32% variance) clustered based on drz+hot treatment. I found that when comparing climates of origin across treatments, again, 1 gene is significantlz differentially expressed: MA\_129323g0010, which was down regulated in hotdrz (p=0.011, foldchange = -21.55). Unfortunatelz, this gene is labelled as unkown on the congenie database and using NCBI BLAST of the CDS didn’t result in any hit. This gene was upregulated in both hot and hotanddrz treatments more in individuals coming from the coldwet climates. When comparing droughtplus heat treatment to control, 257 gene were significantly upregulated in drought plus heat treatment and 330 were significantly downregulated in drought plus heat treatment compared to control. Again, when looking at a heatmap, we can see that gene expression is highlz dependent drought treatment and less on climate or heat. When comparing heat treatment to control instead, there were onlz 6 genes that were differentiallz regulated – 4 up and 2 down compared to control. Remarkably, 3 out of the 4 gene upregulated in heat treatment were genes encoding heat shock proteins (two **KDa class I heat shock -like**, one **KDa class IV heat shock -like**). The upregulation of these genes were independent of climate of origin and fascinatingly onlz the class I heat shock proteins were also upregulated in drought+heat, making them the onlz two gene in the dataset that overlapped heat and heat plus drz, but not the class IV one. There were no significant interactions between climate and treatment, neither when the he specific effect of drought was considered in climate HD nor when the specific effect of heat was considered in climate HD.

Conclusion:

Finally, anazsis of both the cut down and full dataset showed that drought+hot had much higher of an effect on gene expression than climate or onlz hot treatment.

Found a gene more expressed from hotdrz, and another one more expressed in coldwet (when looking only at day 10)

**Conclusion** Our results showed a lack of rare alleles in the population. This could be caused by two processes: balancing selection or sudden population contraction. The latter is a much more likely explanation given the history of the red spruce species. The populations first went through contraction due to the melting of the glaciers ~20k years ago. Then, they were severely impacted by European settlers, who cut down many low elevation stands in the late 1800s, also making the forests more susceptible to wildfires [1]. While red spruce population sizes are starting to increase due to conservation efforts (see The Central Appalachian Spruce Restoration Initiative) and is in category “least concern” on the IUCN Red List, the current climate change could severely impact these edge populations. It is predicted the temperatures in West Virginia are going to increase by ~5 °F by 2050 and the frequency of extreme events is also likely to increase [10].

Based on this data only, the possibility of balancing selection can’t be excluded. Also, the analysis presented here was done on only 5 individuals (although the other 17 populations tested by others in the class all had positive Tajima’s D values, supporting these findings). Potential sources of error: reference genomes used for probe design and mapping were of different species, thus some sequences might have been missed or misaligned; and folding SFS can bias the results as sometimes the derived allele frequency is higher than the ancestral allele frequency. Further analysis should be done on what sites are under selection that could facilitate adaptation to warming climates to guide future conservation efforts.

![A screenshot of a cell phone

Description automatically generated]()

Figure 1. Folded SFS of the XCV edge population, only showing distribution of the ~1% polymorphic sites (0 column excluded). Derived alleles in most polymorphic sites are found in 2 or 3 individuals but sites with derived alleles in 4 or 5 individuals is also abundant, thus there is a lack of rare alleles.

**References**

1. Nowacki, Gregory, Robert Carr, and Michael Van Dyck. "The current status of red spruce in the eastern United States: distribution, population trends, and environmental drivers." (2010): 140-162.
2. Davis, Margaret B. "Quaternary history of deciduous forests of eastern North America and Europe." *Annals of the Missouri Botanical Garden* (1983): 550-563.
3. Jaramillo‐Correa, Juan P., and Jean Bousquet. "New evidence from mitochondrial DNA of a progenitor‐derivative species relationship between black spruce and red spruce (Pinaceae)." *American Journal of Botany* 90.12 (2003): 1801-1806.
4. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.
5. Nystedt, Björn, et al. "The Norway spruce genome sequence and conifer genome evolution." *Nature* 497.7451 (2013): 579-584.
6. Li, Heng. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." *arXiv preprint arXiv:1303.3997* (2013).
7. Tarasov, Artem, et al. "Sambamba: fast processing of NGS alignment formats." *Bioinformatics* 31.12 (2015): 2032-2034.
8. Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. "ANGSD: analysis of next generation sequencing data." *BMC bioinformatics* 15.1 (2014): 356.
9. Li, Heng, et al. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25.16 (2009): 2078-2079.
10. Byers, Elizabeth A., James P. Vanderhorst, and Brian P. Streets. "Classification and Conservation Assessment of Upland Red Spruce Communities in West Virginia." (2010).