Homework 1 | Csenge Petak | 02/27/2020

Background Red spruce (*Picea rubens*) is a species 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 spruce speciated 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, π, 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 In total, there were 3,7606,347 sites considered in the ANGSD analysis, of which 1.011632% was polymorphic (SNP) in the population. Visual inspection of the folded SFS indicates that there isn't an abundance of rare alleles, as in that case we would expect to see a lot of sites where only few individuals has the derived allele (i.e. first few bars much taller than the others). Instead, here we see that the number of sites that have the derived allele in one individual is similar to the number of sites that have the derived allele in four or five individuals (See Figure 1).

Tajima's D is computed as the difference between π, the average pairwise difference among individuals, and the Watterson estimator, the number of segregating sites. These two measures of genetic diversity are scaled so that $\pi = W = \theta$ in a neutrally evolving population of constant size, where θ is the effective population size scaled mutation rate. Here, mean per-site $\theta\pi$ was 0.004255, mean per-site θW was 0.003554 and mean Tajima's D was 0.8910. Tajima's D being positive and so $\theta\pi > \theta W$ mean that there is a lack of rare alleles, consistent with the SFS results.

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.

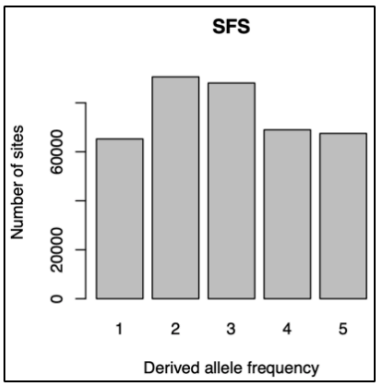

Table 1. Trimming and alignment statistics. Trimming only reduced the total number of sequences by 5.8% of which 90.4% was mapped to the reference genome. Overall, more than half of the reads were sucsessfully mapped in pairs and the depth of coverage was sufficient.

Sample	Total sequences before trimming	Total sequences after trimming	Total number of sequences mapped	Paired	Proportion of reads mapped correctly in pairs	Depth of coverage
1	2527494	2418488	2195061	1364782	0.622	3.435
2	3676800	3515814	3143688	1916810	0.610	4.032
3	3195138	2976278	2703977	1669360	0.617	3.697
4	3663930	3409898	3079150	1921638	0.624	4.051
5	2380092	2227056	2035599	1285144	0.631	3.259
Average	**3,088,690.80**	**2,909,506.80**	**2,631,495**	**1631546.8**	**0.620**	**3.695**

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).