Population genomics of *Picea rubens*

Thomas O'Leary

Background

Picea rubens Ecology

In the context of a rapidly changing climate many organisms will be forced to either adapt, migrate, or go extinct. For *Picea rubens*, a red spruce that thrives in cool and moist climates, this may lead to futher range contraction or eventual extinction if the species is unable to adapt in time (Siccama et al.). After the last glacial maxima, *P. rubens* was forced to retreat to isolated populations along the mountain tops of the Appalacian range, isolated from the northen core of red spruce (McLaughlin et al.). As the climate continues to warm, these isolated populations may pop out of existence due to the increased environmental stress. However, because evolution can only act on standing genetic variation these isolated populations may represent an important genetic resource for the species at large, and may help inform conservation efforts. The ultimate aims of this study are to (i) describe the population structure and genetic diversiy of *P. rubens*, (ii) identify loci that show signs of positive selection and (iii) map the genetic basis of these adaptive phenotypes. In this write up, I will only begin to address the first aim in the context of one population (BFA), with five sequenced individuals, using a few genetic diversity metrics including nucleotide diversity (π), Wattserson's estimator (θ), and Tajima's D (Nei and Li; Watterson; Tajima).

Sample collection, library preparation, and sequencing

Whole genomic DNA was extracted from needle tissue that was collected from a total of 340 mother *P. rubens* in 65 populations, of which 110 trees were from 23 edge populations. 80,000 120 bp probes were designed for exome capture based on multiple developmental and tissue type transcriptomes from a related white spruce *Picea glauca* (Rigault et al.; Yeaman et al.). Approximately 95% of the probes were designed within exomic regions with the remaining 5% in intergenic regions, covering a total of 38,570 unigenes. The probes were blasted against the *P. glauca* reference genome to ensure at least 90 bp of 85% identity. 250 ng to 1 μg of genomic DNA was mechanically sheared to an average length of 400 bp, exome fragments were enriched using the designed probes, and following barcode adaptation, the libraries were pooled and paired-end 150 bp sequenced on an Illumina HiSeq X.

Bioinformatics Pipeline

The quality of the raw reads were accessed using FastQC v0.11.3 (Andrews et al.). To remove low quality sequence data the raw fastq files were trimmed and paired using Trimmomatic v0.33 (Bolger et al.) and the cleaned reads were vizualized again with FastQC. The paired-cleaned reads were mapped to a reduced verion of the Norway spruce *Picea abies* reference genome (Nystedt et al.) using bwa v0.7.12 $-$ *r*1039 (Li and Durbin). A reduced *P. abies* reference genome was used because there is no available *P. rubens* genome and the exome capture technique meant that only a small fraction of the genome near the designed probes would be sequenced.

The sequence aligned reads (sam files) were then converted to binary (bam) and the PCR duplicates were removed and files were sorted using samtools v1.4 and sambamba v0.7.1 (Tarasov et al.; Li et al.). The effects of the low sequencing depth were compensated by estimating genotype likelihoods using ANGSD v0.931 – 14-gb9c8ddd (Korneliussen et al.). This was done because there is some probability individuals are truly heterozygous at loci where only one allele was sequenced due to the random chance. The genotype likelihoods were used to estimate the site frequency spectrums (SFS) and genetic diversity metrics including nucleotide diversity (π), Wattserson's θ, and Tajima's D using ANGSD. A folded SFS was used because the reference genome for *P. abies* many not represent the true ancestral state at all loci so the minor allele frequency is being used to as a proxy for the derived allele. SFS, distributions and means of per site π and θ were calculated and visualized in R v3.6.1 (R Core Team) using ggplot2 and dplyr v1.2.1 (Wickham).

Table 1: Sequencing statistics for individuals in the BFA population.

Individual	Sequencing depth	Number of reads	Paired reads
BFA_01	3.56	2,548,769	1,575,782
BFA_02	4.37	3,275,145	2,077,886
BFA_03	3.29	2,065,376	1,293,090
BFA_04	3.35	2,213,653	1,368,740
BFA_05	3.25	1,860,172	1,186,482

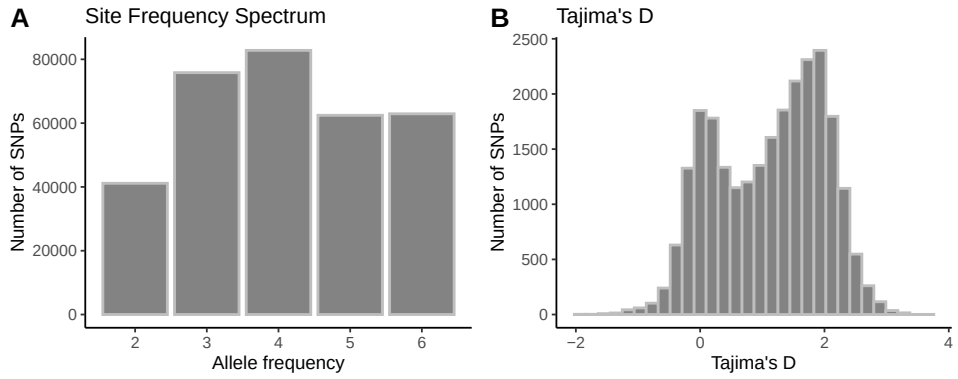

Figure 1: *Genomic diversity statistics for the BFA population.* **A**. Site frequency spectrum of only the polymorphic alleles. **B**. Frequency distribution of Tajima's D.

Results

The sequencing statistics indicate that among the five individuals in the BFA population, the mean per site sequence depth was 3.56 and the mean number of mapped reads was 2.39×10^6. The average percentage of paired reads was 91.3% (Table 1). The percent of single nucleotide polymorphisms across the geneome is 0.92%. The genetic diverity metrics calculated by ANGSD show a mean Watterson's θ of 0.00318 and the mean per site nucleotide diversity π of 0.00394. The mean of Tajima's D was calculated to be 1.11 (Figure 1: B). Given our calculated θ, the effective population size (N_e) can be calculated using the equation: $\theta = 4N_e\mu$, and the estimated *Picea spp.* mutation rate[1] of 1.1×10^{-10} base^{-1} gen^{-1} (Nystedt et al.). This calculation estimates an N_e of $7,221,523$ for the BFA population.

Conclusion

The conservation of *Picea rubens* along the eastern United States may depend on identifying vulnerable populations and their potential genetic resources in an effort to maintain genetic diverity, conserve, and restore the species. A quick look at the site frequency spectrum in Figure 1: A indicates a loss of rare alleles, with the rare dimorphic alleles being less frequent than other polymorphic alleles. Loss of rare alleles can happen during bottlenecking events as genetic drift begins to have a greater effect on small populations. Likewise, the positive mean ($\bar{D} = 1.11$) and distribution of Tajima's D (Figure 1: B) indicates that the BFA population is likely losing genetic diversity and undergoing a bottleneck event as the edge population continues to shrink. Although these numbers seem to agree that there is a genetic bottlenecking pattern within the BFA population, we must proceed with caution when making claims about the genetic diversity of all other edge populations and the species as a whole.

Future directions for this study must involve: incorporating the other sequenced populations and comparing genetic diversity across regions, as well as characterizing the population structure. This will help identify populations of particular concern. Incorporating phenotypic data, perhaps some metric of growth rate, would allow us to mine these data for loci that are associated with phenotypes of interest by means of a GWAS. Identifying loci under postive selection and begining to map theses adaptive phenotypes will help to understand the sort of adaptive challenges this species is facing.

[1]This estimate is based on the mutation rate of 2.2×10^{-9} base^{-1} year^{-1} reported in Nystedt et al. and assuming a generation time of 20 years — which is a wild guess, but it likely gets us at least within an order of magnitude of the true per generation mutation rate.

References[2]

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.

Korneliussen, Thorfinn Sand, et al. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics*, vol. 15, no. 1, 2014, pp. 1–13, doi:10.1186/s12859-014-0356-4.

Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics*, vol. 25, no. 14, 2009, pp. 1754–60, doi:10.1093/bioinformatics/btp324.

Li, Heng, et al. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics*, vol. 25, no. 16, 2009, pp. 2078–79, doi:10.1093/bioinformatics/btp352.

McLaughlin, S. B., et al. "An analysis of climate and competition as contributors to decline of red spruce in high elevation Appalachian forests of the Eastern United states." *Oecologia*, vol. 72, no. 4, 1987, pp. 487–501, doi:10.1007/BF00378973.

Nei, M., and W. H. Li. "Mathematical model for studying genetic variation in terms of restriction endonucleases." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, no. 10, 1979, pp. 5269–73, doi:10.1073/pnas.76.10.5269.

Nystedt, Björn, et al. "The Norway spruce genome sequence and conifer genome evolution." *Nature*, vol. 497, no. 7451, Nature Publishing Group, 2013, pp. 579–84, doi:10.1038/nature12211.

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

Rigault, Philippe, et al. "A white spruce gene catalog for conifer genome analyses." *Plant Physiology*, vol. 157, no. 1, 2011, pp. 14–28, doi:10.1104/pp.111.179663.

Siccama, Thomas G., et al. *Decline of Red Spruce in the Green Mountains of Vermont*. no. 2, 1982, pp. 162–68.

Tajima, Fumio. "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." *Genetics*, vol. 123, 1989, pp. 585–95.

Tarasov, Artem, et al. "Sambamba: Fast processing of NGS alignment formats." *Bioinformatics*, vol. 31, no. 12, 2015, pp. 2032–34, doi:10.1093/bioinformatics/btv098.

Watterson, G. A. "On the Number of Segregating Sites in Genetical Models without Recombination." *Theoretical Population Biology*, vol. 7, 1975, pp. 256–76, doi:10.1039/b316709g.

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

Yeaman, Sam, et al. "Conservation and divergence of gene expression plasticity following c. 140 million years of evolution in lodgepole pine (Pinus contorta) and interior spruce (Picea glauca × Picea engelmannii)." *New Phytologist*, vol. 203, no. 2, 2014, pp. 578–91, doi:10.1111/nph.12819.

[2] This does not exactly follow the MLA citation example you gave, but it was the best `.csl` I could find in time.