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. In the case of *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, Bliss, and Vogelmann 1982). As glaciers began to melt approximately 20,000 years ago, *Picea 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. 1987). 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 *Picea rubens* along its current range, (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 only 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 1979; Watterson 1975; Tajima 1989).

Sample collection, library preparation, and sequencing

Whole genomic DNA was extracted from needle tissue that was collected from a total of 340 mother trees in 65 populations, of which 110 trees were from 23 edge populations. 80,000 120 bp probes were designed for exome capture based on the multiple developmental and tissue type transcriptomes from the related white spruce *Picea glauca* (Rigault et al. 2011; Yeaman et al. 2014). 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 90bp of 85% identity. 250ng to 1 μg of genomic DNA was mechanically sheared to an average length of 400 bp, exome fragments were enriched using 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. 2012). To remove low quality sequence data the raw fastq files were trimmed and paired using Trimmomatic v0.33 (Bolger, Lohse, and Usadel 2014) 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. 2013) using bwa v$0.7.12-r1039$ (H. Li and Durbin 2009). A reduced *Picea abies* reference genome was used because there is no available *Picea 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 the files were sorted using samtools v1.4 and sambamba v0.7.1 (Tarasov et al. 2015; H. Li et al. 2009). The effects of the lack of sequencing depth were compensated by calculating genotype likelihoods using ANGSD v$0.931-14$-gb9c8ddd (Korneliussen, Albrechtsen, and Nielsen 2014). This was done because there is some probability individuals are still heterozygous at loci where only one allele was sequenced due to the random chance of the other allele not appearing in the data. The genotype likelihoods were used to estimate the site frequency spectrums (SFS) and estimate genetic diversity metrics including nucleotide diversity (π), Wattserson's estimator (θ), and Tajima's D using ANGSD. SFS

Table 1: Table Caption!!!!!!!

Individual	Sequencing depth	Number of reads	Paired reads
BFA_01	3.56	2548769	1575782
BFA_02	4.37	3275145	2077886
BFA_03	3.29	2065376	1293090
BFA_04	3.35	2213653	1368740
BFA_05	3.25	1860172	1186482

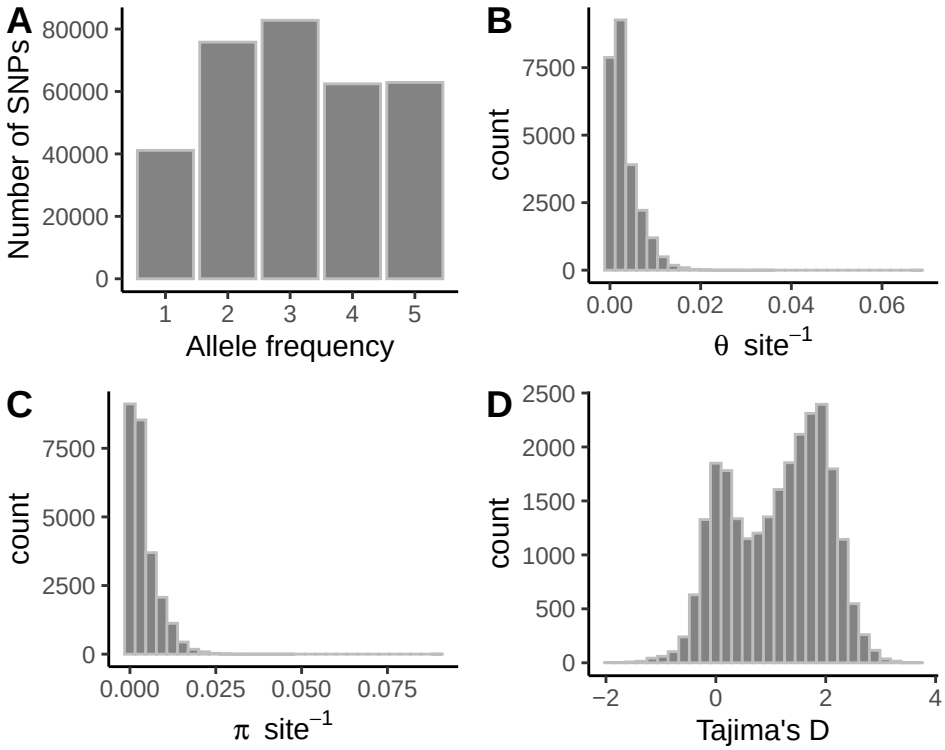

Figure 1: Figure caption goes here!!!!!

and distributions and means of per site π and θ were calculated and visualized in R v3.6.1 (R Core Team 2019) using ggplot and dplyr v1.2.1 (Wickham 2017).

Results

Among the five individuals in the BFA population, the mean per site sequence depth was 3.56 the average number of mapped reads was 2.39×10^6 and the mean percentage of paired reads was 91.3% (Table 1).

number of mapped reads

number of deplicates

persite average read depth

$\theta = 4N_e\mu$

Given that θ is the .... and the effective population size (N_e). The estimated per base per year mutation rate of *Picea* is $2.2\text{x}10^{-9}$ (Nystedt et al. 2013).

$$D = \frac{\pi}{S}$$

Conclusion

References

Andrews, Simon, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. 2012. "FastQC." Babraham, UK: Babraham Institute.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A flexible trimmer for Illumina sequence data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15 (1): 1–13. https://doi.org/10.1186/s12859-014-0356-4.

Li, Heng, and Richard Durbin. 2009. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25 (16): 2078–9. https://doi.org/10.1093/bioinformatics/btp352.

McLaughlin, S B, D J Downing, T J Blasing, E R Cook, and H S Adams. 1987. "An analysis of climate and competition as contributors to decline of red spruce in high elevation Appalachian forests of the Eastern United states." *Oecologia* 72 (4): 487–501. https://doi.org/10.1007/BF00378973.

Nei, M., and W. H. Li. 1979. "Mathematical model for studying genetic variation in terms of restriction endonucleases." *Proceedings of the National Academy of Sciences of the United States of America* 76 (10): 5269–73. https://doi.org/10.1073/pnas.76.10.5269.

Nystedt, Björn, Nathaniel R. Street, Anna Wetterbom, Andrea Zuccolo, Yao Cheng Lin, Douglas G. Scofield, Francesco Vezzi, et al. 2013. "The Norway spruce genome sequence and conifer genome evolution." *Nature* 497 (7451). Nature Publishing Group: 579–84. https://doi.org/10.1038/nature12211.

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

Rigault, Philippe, Brian Boyle, Pierre Lepage, Janice E.K. Cooke, Jean Bousquet, and John J. MacKay. 2011. "A white spruce gene catalog for conifer genome analyses." *Plant Physiology* 157 (1): 14–28. https://doi.org/10.1104/pp.111.179663.

Siccama, Thomas G., Margaret Bliss, and H. W. Vogelmann. 1982. "Decline of Red Spruce in the Green Mountains of Vermont" 109 (2): 162–68.

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

Tarasov, Artem, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. 2015. "Sambamba: Fast processing of NGS alignment formats." *Bioinformatics* 31 (12): 2032–4. https://doi.org/10.1093/bioinformatics/btv098.

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

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

Yeaman, Sam, Kathryn A. Hodgins, Haktan Suren, Kristin A. Nurkowski, Loren H. Rieseberg, Jason A. Holliday, and Sally N. Aitken. 2014. "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* 203 (2): 578–91. https://doi.org/10.1111/nph.12819.