Population and Landscape Genomics of Red Spruce

Background

Red spruce (*Picea rubens*) is a species of conifer found at high elevations in the eastern United States and southeastern Canada. It is most abundant in New England, Nova Scotia, and Eastern Quebec, but the range extends as far south as North Carolina, where populations can be found on "sky islands"—isolated patches of suitable habitat left behind on mountaintops in the Mid-Atlantic region in the wake of receding glaciers approximately 20,000 years ago (Capblancq *et al*., 2020). Red spruce's highly fragmented distribution and sensitivity to heat and drought suggest that the tree may be particularly vulnerable to climate change, and its status as a keystone species in eastern forest ecosystems makes it an important target for conservation and restoration efforts, including assisted migration. Ultimately, standing adaptive genetic variation, demographic history, and population structure across the range of red spruce will determine whether it will adapt to changing climatic conditions, migrate to track its preferred environmental niche, or become locally extirpated.

Our goal is to characterize genetic diversity across the range of red spruce in order to infer demographic history and population structure. We also identify outlier loci that may be under selection and test for associations between SNPs and bioclimatic variables. This study analyzes whole-exome sequencing data from 95 adult red spruce individuals sourced from 12 populations spanning the entire range of the species. We also include data from 18 black spruce individuals located far away from red spruce's range in order to explore possible introgression. Exomes were sequenced using baits based on the reference genome of the related species *P. glauca* (white spruce). Libraries were prepared by mechanically shearing extracted DNA, ligating barcoded adaptors, amplifying sequences with PCR, and using baits for targeted enrichment. All sequencing was performed on an Illumina HiSeq X, producing paired-end 150bp reads for downstream bioinformatics analysis.

Bioinformatics Pipeline

We visualized quality scores of the reads using 'FastQC' (Andrews, 2010). This showed good-quality sequence data for most of the reads, but the very beginning and end of the reads had lower Q-scores, so we used the program 'fastp' (Chen, 2023) to trim them. We then mapped the reads onto a subset of the *P. abies* reference genome that included only contigs matching the probes used for the exome sequencing in order to reduce the computational load of searching for matches across *P. abies*'s 19.6 Gbp genome. Reads were mapped to the reduced reference genome using 'bwa' (Li, 2013). We used 'sambamba' (Tarasov, *et al.*, 2015) to process and sort the resulting SAM files and to remove PCR duplicates, followed by 'samtools' (Danecek, *et al.*, 2021) to evaluate the depth of coverage and how well the reads mapped to the reference genome.

Because average read depth was low, we used the program 'ANGSD' (Korneliussen, *et al.*, 2014) to generate genotype likelihoods and allele frequencies for downstream analyses, rather than hard-calling genotypes that might be inaccurate due to insufficient sampling, and then calculated the site frequency spectrum based on those likelihoods. Comparing the SFS of paired populations of red spruce, or of a red spruce population and a black spruce population, allowed us to calculate F_{ST}. Population structure and admixture proportions were inferred using 'pcANGSD' (Meisner & Albrechtsen, 2018). We tested clustering schemes in which k=2, 3, and 4 to evaluate which best represented the structure of the data. 'pcANGSD' was also used to scan for F_{ST} outliers, loci that are vary considerably along one or more PC axes and therefore may be under selection (Galinsky, *et al.*, 2015) for k=2 and k=3. We identified and visualized in R contigs with p-values < 0.001 and < 0.0001, matched them to gene IDs from the *P. abies*

reference genome, and looked for gene annotations and functional enrichment in PlantGenIE (plantgenie.org). Finally, we extracted values for 19 bioclimatic variables from WorldClim (Fick & Hijmans, 2017) at a spatial resolution of 10 arc-minutes and conducted a genotype-environment association analysis (GEA) to test for correlations between environmental gradients and allele frequencies at the outlier loci.

Results

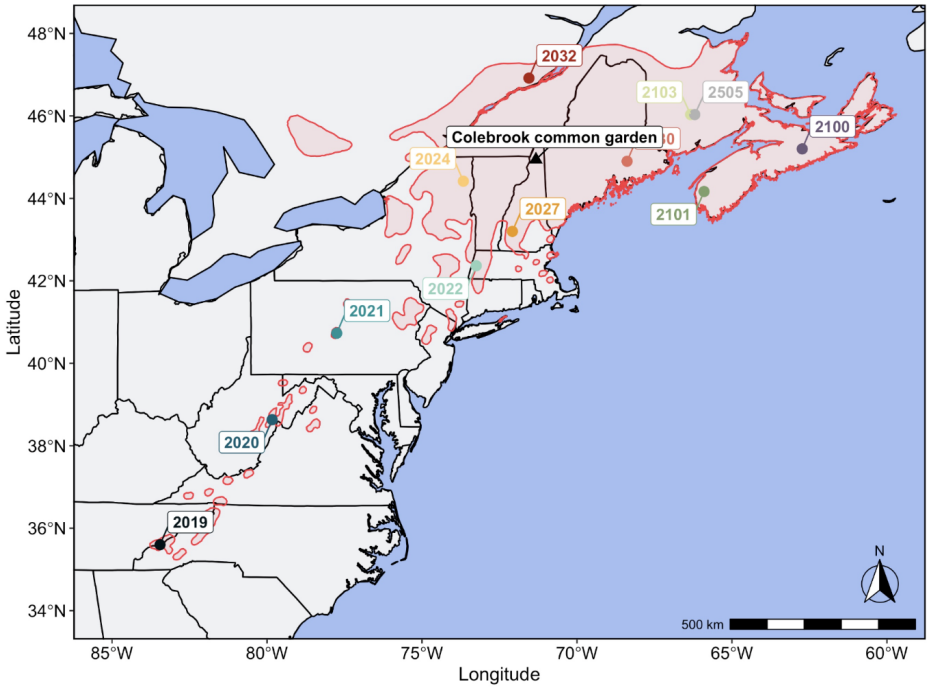
Overall, our analyses paint a picture of greater gene flow in the core of the range in New England and Canada, with the southernmost populations being more genetically distinct. Populations in the core of the northern range showed lower F_{ST} with black spruce, while the fragmented Mid-Atlantic populations showed the highest genetic differentiation, evidence of their relative isolation and lack of genetic connectivity with other spruce populations (Fig. 1). Population 2021 has a somewhat high F_{ST} given its position in central Pennsylvania; however, this area in particular has been identified as a zone of hybridization with black spruce (Capblancq, *et al*., 2020). In the admixture plot for k=2 (Fig. 2), we could interpret the blue ancestry proportion to be primarily black spruce hybridization, as it is highest for the northern populations (rightmost), declining as it moves south (left), with population 2021 standing out for its high proportion of blue ancestry. Testing k=3 and k=4 creates additional noise in the admixture plots, but does also highlight the genetic divergence of the northern and southern populations, some of which is likely due to historic range expansions and retractions during glaciation, not solely hybridization with black spruce.

Genetic PCAs for all three k-values produced three clusters: 1) populations in the main northern part of the range, 2) population 2021 in Pennsylvania, and 3) fragmented populations in the south. In all cases, latitude appears to be the main variable loading onto PC1 (~5.3% of variation), while PC2 (~1.6-3.5% of variation) could be driven by ancestry proportions. The selection scan for k=3 identified 675 outliers on PC1 and 1577 on PC2 for a cutoff of $p < 0.001$, and a more restricted set of 14 outliers on PC1 and 313 on PC2 for a cutoff of $p < 0.0001$. In the climate PCA, mean annual precipitation (bio12) loads most strongly on PC1 (52.2% of variation), and mean temperature of the warmest quarter (bio10) has the strongest correlation on PC2 (24.1%).

Conclusion

In this study, we have identified strong signatures of population structure and selection in red spruce, pointing to three genetically distinct populations in the north, south, and middle of the range. Hybridization with black spruce has evidently shaped the genetic variation found in these populations, as have range shifts following deglaciation and human impacts such as land use change, deforestation, and air pollution. A number of candidate loci have been identified as linked to environmental variation or under selection. Several limitations should be noted: first, PC1 amd PC2 explain very small percentages of the total variation observed; second, GEA is inherently a correlative analysis, obscuring possible mediating factors that could be responsible for these gene-environment relationships (e.g. pollinators or pests with distributions shaped by abiotic climate factors, the current gene-environment relationships have already been shifted from their optima by climate change). Our understanding of population and landscape genomics in red spruce and its potential conservation applications would be greatly improved by additional sequencing, transcriptomics, and common garden or growth chamber experiments that could elucidate the mechanistic relationship between the candidate loci identified here and the environmental stressors red spruce will face in the coming decades.

Figures



Population	F_{ST}
2032	0.2810
2103	0.1820
2505	0.1982
2100	0.2816
2030	0.2470
2024	0.3191
2101	0.2419
2027	0.2956
2022	0.3168
2021	0.2189
2020	0.3395
2019	0.3381

Figure 1. Left: Map of red spruce distribution (red) with sampled populations marked (provided by SR Keller). Right: F_{ST} values measuring genetic divergence between each population and black spruce (ordered approx. north to south).

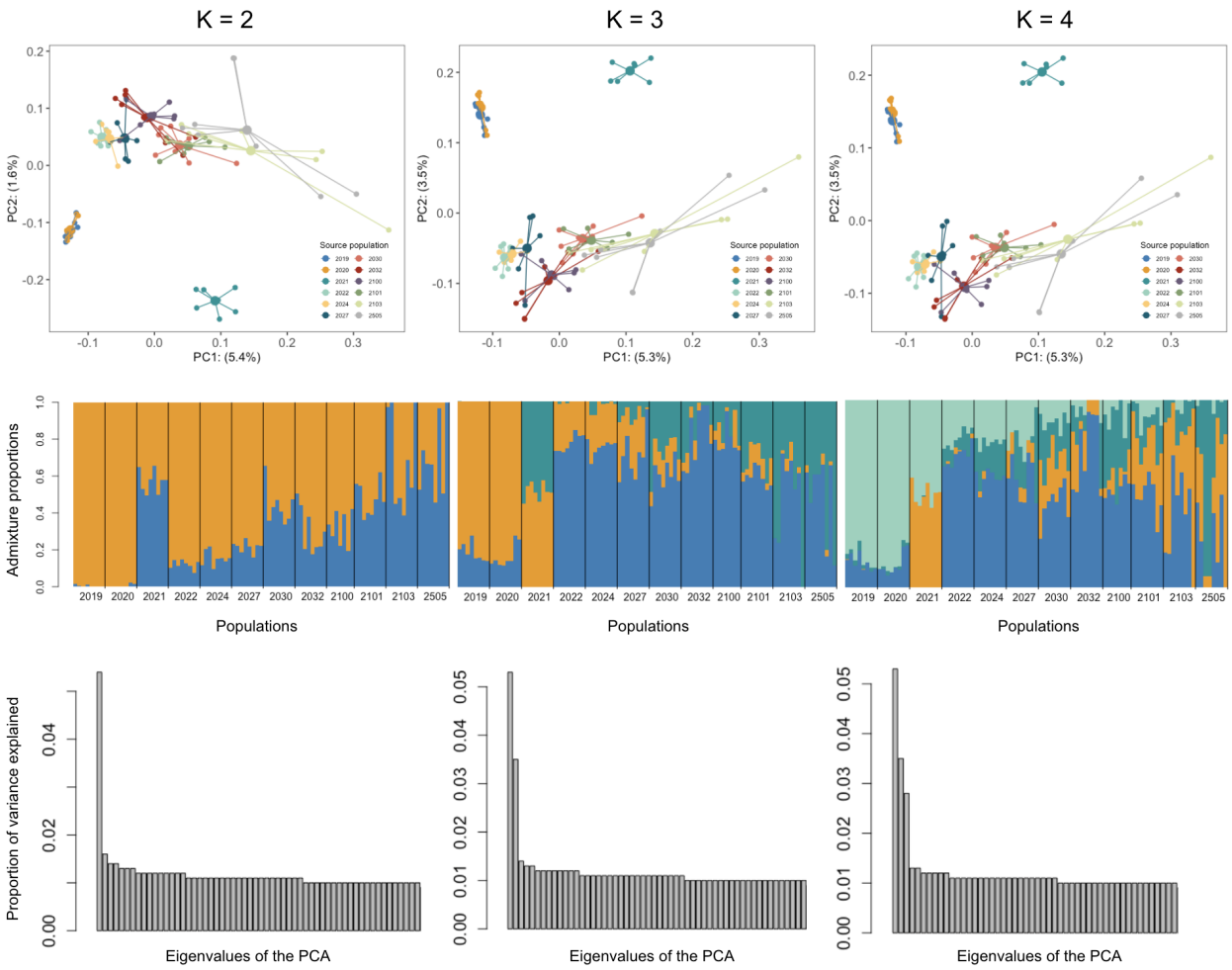


Figure 2. Top: genetic PCAs for 2, 3, and 4 possible clusters. Middle: admixture plots showing proportions of ancestry for each sample. Bottom: Scree plots of PC axes and the amount of variation each explains.

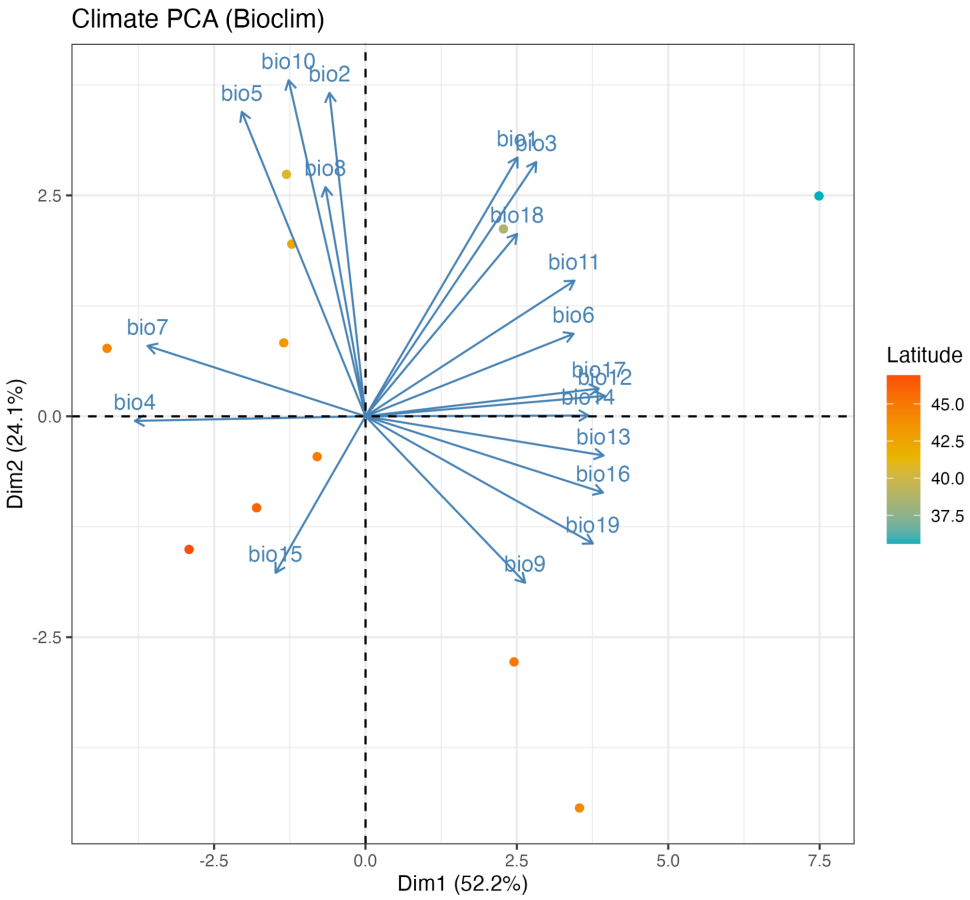


Figure 3. PCA of 19 bioclimatic variables from WorldClim. Arrows represent the bioclimatic variables, and the red spruce populations are plotted in their PCA space as points colored by latitude.

References

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Capblancq, T., Butnor, J.R., Deyoung, S., Thibault, E., Munson, H., Nelson, D. M., Fitzpatrick, M.C., & Keller, S.R. (2020). Whole‐exome sequencing reveals a long‐term decline in effective population size of red spruce (Picea rubens). *Evolutionary Applications, 13*(9), 2190-2205. https://doi-org.ezproxy.uvm.edu/10.1111/eva.12985.

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta 2*(e107). https://doi.org/10.1002/imt2.107.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., & Li, H. (2008). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). https://doi.org/10.1093/gigascience/giab008.

Fick, S.E. & Hijmans, R.J. (2017). WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. *International Journal of Climatology, 37*(12), 4302-4315. https://doi.org/10.1002/joc.5086.

Galinsky, K.J., Bhatia, G., Loh, P., Georgiev, S., Mukherjee, S., Patterson, N.J., & Price, A.L. (2015). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics, 98*(3), 456-472. https://doi.org/10.1016/j.ajhg.2015.12.022.

Korneliussen, T.S., Albrechtsen, A. & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics, 15*(356). https://doi.org/10.1186/s12859-014-0356-4.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*. https://arxiv.org/abs/1303.3997.

Meisner, J. & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics, 210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics, 31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Supplemental

Data, scripts, and results available at:
https://github.com/noraheaphy/ecological_genomics_2023/tree/main