

## Exome sequencing of edge populations of *Picea rubens* reveals targets for effective conservation.

### Introduction

While successful conservation efforts ranging from the Nature Conservancy to the US Fish and Wildlife service have proven to be effective in restoring populations that have been displaced, degraded, or destroyed, remnant populations of endangered species are often patchily distributed over their ancestral range. This presents a unique challenge, for while some large "hubs" of genetic diversity may persist in large patches of habitat, small remnant populations may be bottlenecked and, as a consequence of local adaptation, may therefore lack the genetic diversity required to adapt to future environmental changes. Or, crucially to a conservationist, they may lack the genetic diversity required to flourish in a *different* part of their ancestral range.

*Picea rubens* is a large, locally dominating coniferous species that has been trickling out of existence since the last glacial maximum. As the cool, wet conditions of the LGM receded, individuals of the previously ubiquitous *P. rubens* was stranded atop the tallest peaks in southern Appalachia as the hub population shrank to northern New England and southern eastern Canada. Human development and natural processes have eliminated the majority of red spruce refugia and replaced them with forests of pine, fir, or hardwood species in part due to the infrequency of natural *P. rubens* recolonization.<sup>1</sup> In order to guide restoration policy of red spruce populations on mountaintops in the southern edge of its native range (MD-NC), we sequenced the exomes of 110 mother trees from 23 populations of *P. rubens* from those states ("edge populations") to establish a living library of coding region diversity. This library can answer crucial questions about the ecological and evolutionary trajectories of these populations, whether they can withstand climate change, and what populations, if any, would be high-quality seed donors to future refugia.

Using two previously assembled *P. glauca* transcriptomes<sup>2 3</sup>, 80,000 120bp probes were designed to target 75,732 overlapping exomic regions and 4,268 intergenic regions, covering 38,570 *P. glauca* unigenes in total, each with 90bp or larger blast hits and >85% identity. Libraries consisted of 400bp fragments of random mechanically sheared *P. rubens* DNA that underwent end-repair reaction, ligation of adaptor-compatible residues, and PCR replication. SureSelect probes (Agilent Technologies: Santa Clara, CA) were used for solution-based targeted enrichment of pools of 16 libraries, following the SureSelectxt Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library protocol. Libraries were sequenced on a single run of a Illumina HiSeq X to generate paired-end 150bp reads.

### Bioinformatics

Raw paired-end reads were first visualized with FastQC to examine quality (Phred) scores. The Trimmomatic program<sup>4</sup> removed PCR and Illumina adaptors, then the cleaned paired-end reads were examined again with FastQC to verify that low-quality base calls were removed. Using the program bwa<sup>5</sup>, the forward and reverse reads were then mapped to a subset of scaffolds of the *Picea abies* genome that had BLAST matches with one or more probes from the exon capture procedure. The resulting .sam files were converted to .bam and indexed with samtools<sup>6</sup>. Sambamba's<sup>7</sup> markdup command was then used to remove PCR duplicates, then the .bam files were indexed again. Read depth and FLAG alignment counts were calculated with samtools' depth and flagstat commands, respectively.

To avoid incorrect variant calling, SNP genotype likelihoods were calculated with ANGSD<sup>8</sup> for .bam files filtered for depth, base, and mapping quality. Rough site frequency spectra estimates were made with the realSFS program, then ANGSD was used again to refine the SFS and obtain theta values with the doTheta command. Because a high number of "derived" alleles were present on the SFS, in a single script, ANGSD folded the SFS with the -fold flag, used the -pest flag to incorporate the previously generated SFS as a prior on the current allele frequency estimation, and included the -doThetas flag. The .sfs output files were analyzed in R to calculate values of Tajima's D, theta-w, theta-pi, and mean per site nucleotide diversity.

## Results

Average pairwise difference was consistently greater than mean number of segregating sites, indicating a scarcity of rare alleles. In fact, 22 of 23 populations sampled had Tajima's D values greater than 1.0 (mean = 1.2097).

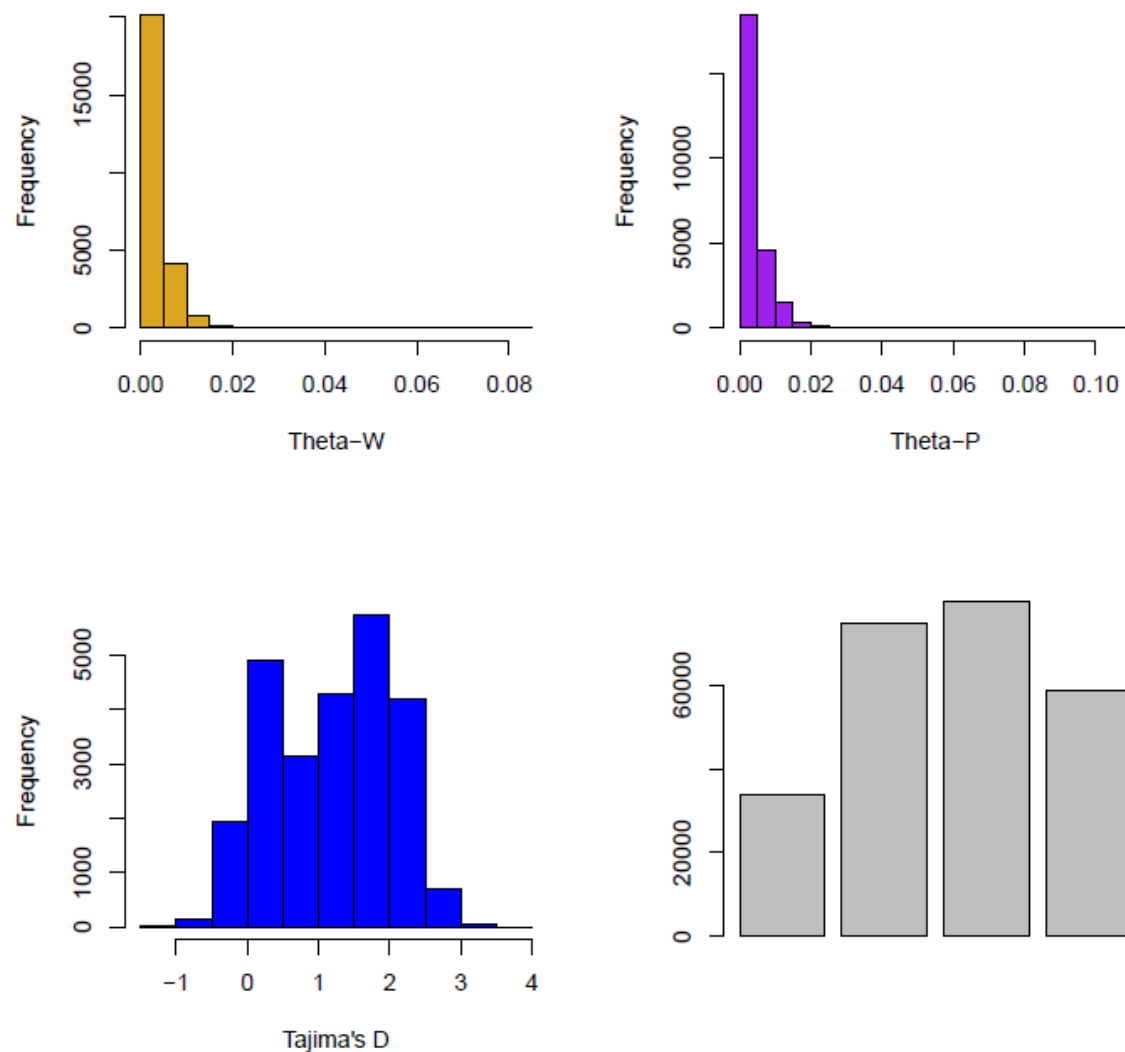


Figure 1. Frequency spectra of MRC population theta-W (top left), theta-pi (top right), Tajima's D (bottom left), and site frequency spectra (bottom right).

## Conclusion

Based on our findings, *Picea rubens* edge populations' genomes suggest the ancestral, contiguous population experienced a bottleneck, then as the populations fragmented and contracted, they experienced a relatively synchronized loss of rare alleles. The widespread Tajima's D values  $>1$  also suggest that these populations are shrinking, confirming the need for conservatory action. In order to maximize the efficacy of conservation efforts, the next question that needs answering is the extent to which genetic diversity needed to adapt to climate change exists in these edge populations. If not, sampling, sequencing, and analyzing the larger populations will potentially reveal a vast reserve of *P. rubens* available for transplantation to southern Appalachia.

- 
1. Rentch, James S., et al. "Red spruce stand dynamics, simulations, and restoration opportunities in the central Appalachians." *Restoration Ecology* 15.3 (2007): 440-452. APA [↔](#)
  2. Rigault, Philippe, et al. "A white spruce gene catalog for conifer genome analyses." *Plant Physiology* 157.1 (2011): 14-28. [↔](#)
  3. Yeaman, Sam, et al. "Conservation and divergence of gene expression plasticity following c. 140 million years of evolution in lodgepole pine (*Picea contorta*) and interior spruce (*Picea glauca* × *Picea engelmannii*)." *New Phytologist* 203.2 (2014): 578-591. [↔](#)
  4. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data, *Bioinformatics*, 30 (15): 2114-2120. # [↔](#)
  5. Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *bioinformatics* 25.14 (2009): 1754-1760. [↔](#)
  6. Li, Heng, et al. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25.16 (2009): 2078-2079. [↔](#)
  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. [↔](#)