Genetic Diversity in Red Spruce (*Picea rubens*) - Zoe Portlas

Background

Red Spruce (*Picea rubens*) is a coniferous tree found in the Eastern United States and Canada. Its range includes a large, fairly continuous region in Eastern Quebec and New England, as well as trailing edge populations that exist in sky islands in the Appalachians from Pennsylvania to Tennessee. These trailing edge populations have limited gene flow between them and may possibly be much more sensitive to climate change than populations farther north in the more connected part of the species’ range. This data was collected to better inform conservation and restoration projects, with the specific aims that include understanding the genetic variation and structure present across the range of Red Spruce and identifying regions of the genome that show evidence of adaptation to climate.

In 2017, seeds and needle tissue were collected from across Red Spruce’s range. The edge region included 110 mother trees in 23 populations. Whole genomic DNA was extracted to use for exome capture sequencing, which captures only the protein coding regions of the DNA. The process of exome capture sequencing used baits that were designed using transcriptomes from White Spruce (Rigault, et al. 2011, Yeaman, et al., 2014). Whole exome sequencing is useful in this case because it is more cost effective and avoids problems with sequencing and assembling regions of the genome that are highly repetitive, in non-coding regions. The libraries were made by random mechanical fragmentation of the DNA, hybridization of the DNA fragments with probes, and PCR amplification, and were sequencing on Illumina HiSeq X to generate paired end 150 bp reads. This paper focuses on the DG population.

Bioinformatics pipeline

The quality of all of the reads in the DG population were assessed both before and after cleaning. The quality of both the raw and cleaned data was assessed using the program FastQC, which generates statistics and visualizations of the data for quality control. After checking the quality of the raw data, the cleaned data was generated using the program Trimmomatic. Cleaned data was then mapped to a Norway Spruce (*P. abies*) reference genome using the program bwa. The reference genome is a reduced reference because only contigs that contained at least one of the probes used in exome capture were used for mapping. Postprocessing was then performed on the data to remove PCR duplicates and index the files. PCR duplicates are removed to avoid artificial duplication due to the PCR process instead of the actual frequency of genes in the genome.

The program ANGSD was then used to generate a site frequency spectrum (SFS) and calculate the metrics of nucleotide diversity: π (a measure of pairwise nucleotide diversity), Watterson’s θ (a measure of the number of segregating sites), and Tajima’s D. Instead of assigning one genotype at loci based on the proportion of reads at each site, ANGSD uses a genotype likelihood framework to assign probabilities that individuals have different genotypes. This is done because the read depth for this data is relatively low and it avoids the risk of miscalling genotypes. Because the original SFS suggested there was a very high number of derived alleles, we assumed that this was due to incorrectly assigning the ancestral state of a portion of the alleles. Because we could not confirm the ancestral alleles with an outgroup, we fixed this by creating a site frequency spectrum of the minor allele frequency by folding the original SFS at the midpoint.

Results

Once the reduced reference genome, which only contains contigs containing 1 or more probes used in exon capture, was created, it contained 668,000,000 bp. As is typical for non-model organisms, the reference genome was not assembled into chromosomes. The N50 of the reference informs us about the size of the contigs assembled and for this reduced reference, was 101,375 bp. A larger number indicates that they are more similar to actual chromosome than a small number, allowing more spatial information to be inferred. The average depth of coverage for the DG population was ~3.5.

Using the genotype likelihood framework, a site frequency spectrum (Figure 1a) was generated and the measures of diversity π, θ, and Tajima’s D were calculated using ANGSD and scaled to the number of sites in the population (Table 1). The distribution of Watterson’s theta (θW), nucleotide diversity (π), and Tajima’s D was generated for the DG population (Figure 1b-d).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Population** | **Number of sites** | **Percent SNPs** | **θW** | **π** | **Tajima’s D** |
| DG | 36,720,648 | 0.881625 | 0.003067 | 0.003842 | 1.2006 |
| All | 31,465,329 | 0.841747 | 0.003037 | 0.003799 | 1.2097 |

Table 1. Number of sites, percent of sites that are SNPs, Watterson’s theta (θW), nucleotide diversity (π), and Tajima’s D for the DG population and averaged across all populations.

A picture containing screenshot

Description automatically generatedA picture containing screenshot

Description automatically generated

Figure 1. (a) Site frequency spectrum (SFS), distribution of (b) Watterson’s theta (θW), (c) nucleotide diversity (π), and (d) Tajima’s D for the DG population.

Conclusion

Because Red Spruce plays an important role in forest communities in the Appalachians, its susceptibility to climate change is a concern. Genomic data is an important contribution to better inform conservation and restoration as well as predict the species’ potential response to climate change. The measures of genomic diversity obtained for the DG population, one of 23 edge populations in the dataset, can give an indication of genomic resources in the population. Watterson’s theta for this population was 0.003067 and nucleotide diversity was 0.003842 (Table 1) and the distributions of both θW and π are very skewed towards 0 (Figure 1b, c), indicating that there is a low level of genetic diversity within the population. DG has similar levels of genetic diversity to the average across all populations, all of which have low genetic diversity. Tajima’s D for DG was 1.2006 and for all populations was 1.2097. Because it is above 0, it indicates that both the DG population and the populations in the edge region of Red Spruce’s range experienced a population contraction or bottleneck, which reflects the retreat of populations in this part of the range to small, isolated populations at high altitudes.

While using whole exome capture can be very cost effective and efficient, especially for a species like Red Spruce which has a large regions of repetitive sequences, only using data from the exome means that diversity in non-coding regions was not captured and measured in these analyses. While protein coding regions can offer an indication of the species’ response to climate change, non-coding regions may contain diversity in regulatory regions that may play a role in plasticity to warming climates, for example. Further work on this system could fill in these gaps by using transcriptomics and comparing plastic responses to temperature or water stress across populations.

References

Rigault, P., B. Boyle, P. Lepage, J.E.K. Cooke, J. Bousquet, J.J. MacKay. 2011. A White Spruce Gene Catalog for Conifer Genome Analyses. *Plant Physiology* 157(1)

Yeaman, S., K.A. Hodgins, H. Suren, K.A. Nurkowski, LH. Rieseberg, J.A. Holliday, S.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: 578-591.