HW1

Lily Shapiro

2/26/2020

Introduction

Anthropogenic climate change is predicted to drastically alter the distribution of various plant species as they adjust to variable environmental conditions (McKenney et al. (2007),Deb et al. (2017), Kelly and Goulden (2008)). The red spruce, *Picea rubens*, is a cool-adapted montane tree species distributed throughout the Appalachians and Atlantic Canada. Southern populations of *Picea rubens* are highly fragmented, as warming temperatures after the Last Glacial Maximum necessitated the movement of these populations towards higher elevations to maintain temperatures necessary for their survival. By extrapolating and analyzing data concerning the genetic diversity of these fragmented "edge" populations of *Picea rubens*, we will garner an increasingly robust picture of how to implement successful conservation strategies that will hopefully bolster the genetic architecture of these susceptible populations in the face of changing climate. Edge samples (seed and needle tissue) were collected from 110 mother trees in the edge region, from 23 distinct populations; whole genomic DNA was extracted from needles for exome-capture sequencing. 80,000 120 bp probes were created based on white spruce (*P. glauca*) transcriptomes assembled by Rigault et al. (2011) and Yeaman et al. (2014), with each probe representing a blast hit to the reference genome minimum 90 bp long and 85% identity, which covered a total of 38,570 unigenes. DNA was randomly mechanically sheared to a size of 400 bp followed by end-repair reaction. Barcoded adapters were added to the 3' end of fragments and then amplified using PCR. Libraries were sequenced using Illumina HiSeq, generating paired-end 150 bp reads.

QC Assessment + Trimming

Raw text sequence files in the format `.fastq` were quality assessed using the program FastQC (Andrews (2010)). Each read from the paired end run (R1 and R2) were used from 6 samples in the edge population XDS. The `.fastq` files were preliminarily assessed for quality by examining the Phred score (Ewing et al. (1998)) preceding the sequence in order to ensure sequencing accuracy and check that reads were indeed 150 bp long. In to gain more clarity in our quality assessment and more clearly interpret the quality of our reads, we edited a vim script to include a "for loop", allowing us to use FastQC to assess all of our samples at once and directing them as .html files as outputs for easy visualization, where FastQC graphs were then analyzed to determine sufficient quality. Reads were cleaned using Trimmomatic (Bolger, Lohse, and Usadel (2014)). Trimmomatic utilizes a sequence of consecutive trimming steps to increase read quality and preserve informative exomic data by removing Illumina adapters and removing bases within a read that are below a particular threshold quality. For our reads, the Trimmomatic script was similarly edited to the FASTQC vim script in order to carry out the cleaning process on all .fastqc files within our sample in one go. Paired clean reads were collected and saved, and then assessed again using FastQC to check cleaned quality.

Mapping

```
*Picea rubens* reads were mapped to a *Picea abies* (Norway spruce) reference genome for contig assembly
For further analysis, these `.sam` files were converted to their compressed binary counterparts in `.bam
```

Genotype Likelihood (GL) Analysis

We used the program `ANGSD` to use genotype likelihoods to estimate diversity statistics (Korneliussen, Albrechtsen, and Nielsen (2014)). These included Tajima's D (deviation in SFS from the null model, indicative

of selection), π (nucleotide diversity), and Watterson's θ (# segregating sites) among others. Other genomics studies may directly call SNP's within reads, but this ignores the probability of a variant nucleotide existing at that site and increases the chance that the wrong genotype is called, which would likely skew diversity analyses. Instead, we estimate genotype likelihoods that compare the probability of retrieving our sequencing data against the genotype of an individual at a particular site. The program `ANGSD` was preliminarily used to estimate these genotype likelihoods from our `.bam` files, refining a set with a minimum coverage depth and quality, among other markers. Monomorphic sites were retained. `ANGSD` was furthermore utilized to calculate a folded SFS for the population, necessary since we were not confident in the ancestral states at each SNP (based on the minor allele). This calculated SFS was used as prior to estimate θ diversity statistics, which were then calculated also using `ANGSD`. Diversity statistics for the population XDS were summarized in R.

Results

FastQC determined the quality of reads to be good overall, with some wavering slightly in per base sequence quality. All sequences were 150 bp in length. Trimmed .fastq files showed an improvement in sequence quality, with little deviance in quality between samples. Trimmed cleaned pair reads were mapped to the reference genome using `bwa` and statistics on the quality of the mapping were then generated using `flagstats`. These statistics display that each sample of the population had an average of 2.27 million reads, with an average of 62% of reads mapped to the *P abies* reference for each sample. Depth of coverage for each sample derived from `samtools` ranged from 2.92 to 3.93, denoting the number of reads at each mapped position. SFS displays higher portions of SNP's found in medium frequency within the population. There appears to be less rare SNP's and low levels of common SNP's within this population, compared to SNP's appearing at mid-frequency. The percent of polymorphic loci compared to total sites was 1.05, and although high, might be skewed by the apparently low number of total sites at 435,746. Average Tajima's D for the XDS population was 1.0917. Mean Watterson's θ value was 0.00329, and mean π was equal to 0.004104.

Conclusion

Exome data collected across populations of *P. rubens* indicates there is low variance in diversity and low positive selection for any particular subset of the total sample set of populations. While there are obvious variations in data per population, the nucleotide diversity (π) values and Watterson's θ values were consistent across populations, indicating that no one population or group of populations contained significantly elevated levels of genetic diversity. In addition, positive and similar Tajima's D values across populations indicate that there was likely a bottleneck that occured before the fragmentation of these edge populations, and balancing selection is currently acting on the sampled portion of the species overall. Because of the lack of particular populations displaying increased directional selection, we can extrapolate that these fragmented populations do not appear to be showing signs of localized adaptation. To more accurately determine differences in genetic diversity across these populations of *P. rubens*, F_{st} would be calculated necessarily. In addition, diversity statistics were not calculated for a subset of the fragmented populations (10/25), the results of which might additionally inform conclusions on the overall diversity and demography of this species sampling. The small number of sites for the XDS population may have altered the calculation of percent polymorphic sites. Future studies will further use this data set to compare diversity across populations by analyzing F_{st} values. Additionally, studies incorporating environmental/climatic variables for each sample with corresponding SNP data (and modeling their correlations) will inform a possible genomic response to climate for this group.

Andrews, S. 2010. "FASTQC. A quality control tool for high throughput sequence data."

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30 (15). Oxford University Press: 2114–20.

Deb, Jiban C, Stuart Phinn, Nathalie Butt, and Clive A McAlpine. 2017. "The impact of climate change

on the distribution of two threatened Dipterocarp trees." *Ecology and Evolution* 7 (7): 2238–48. https://doi.org/10.1002/ece3.2846.

Ewing, Brent, LaDeana Hillier, Michael C Wendl, and Phil Green. 1998. "Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment." *Genome Research* 8 (3). Cold Spring Harbor Lab: 175–85.

Kelly, Anne E, and Michael L Goulden. 2008. "Rapid shifts in plant distribution with recent climate change." *Proceedings of the National Academy of Sciences* 105 (33). National Academy of Sciences: 11823–6. https://doi.org/10.1073/pnas.0802891105.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. "ANGSD: analysis of next generation sequencing data." *BMC Bioinformatics* 15 (1). BioMed Central: 356.

McKenney, Daniel W, John H Pedlar, Kevin Lawrence, Kathy Campbell, and Michael F Hutchinson. 2007. "Potential Impacts of Climate Change on the Distribution of North American Trees." *BioScience* 57 (11): 939–48. https://doi.org/10.1641/B571106.

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). American Society of Plant Biologists: 14–28. https://doi.org/10.1104/pp.111.179663.

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.