Ecological Genomics: Homework #1

Brendan Case

https://github.com/brendandaisy/ecological-genomics

25 February 2020

1 Background

One common source for genetic differentiation between population of the same species is habitat fragmentation, which can lead to founder effects and other consequences of population contraction due to a reduction in population size and gene flow [5]. Red spruce (*Picea rubens*) is a coniferous tree which has become highly isolated since the end of the Pleistocene at the southern end of its range, in the higher elevation Appalachian mountains of Maryland through North Carolina.

To study the extent of genetic differentiation between these southern "island" populations, we use whole genomic data collected from 340 mother trees at 65 locations across the tree's range, extracted using exome capture sequencing [2]. For the southern range of interest, there were 110 mother trees from 23 populations. For each population, this resulted in a forward and reverse strand read from several mother trees, stored as `fastq` files.

2 Bioinformatics Pipeline

To begin our analysis of the sequenced red spruce genomes, we performed quality control (QC) on the raw reads. Each nucleotide in a fastq file is accompanied by a Phred Quality score, which maps a character to the estimated probability the assigned base is incorrect. The characters are assigned in increasing order of their Unicode representation, so that comparing scores between bases is computationally convenient. In addition to visual inspection of the fastq files, we used the FastQC program to visualize read quality [1]. We then used the `trimmomatic` program to trim our paired reads and improve overall quality and aid alignment. Finally, we ran FastQC again on the trimmed reads.

Our next step in finding the extent of genetic diversity was to map the trimmed reads against a reference genome. We chose as reference a genome of the Norway spruce (*Picea abies*), which in 2013 was the first gymnosperm ever to be completely sequenced [4]. To increase computational efficiency, reduced the reference size by only choosing those scaffolds which compliment at least one bait from out exome capture. Mapping

was performed with `bwa`, while `sambaba` and `samtools` were used to convert the `.sam` files to binary, remove PCR duplicates, and sort alignments by leftmost coordinates. At this stage we also calculated the depth and some basic statistics for each alignment.

The final step in our analysis was calculate the genotype of each individual. Due to the large number of individuals and regions in this study, the average depth for each alignment was relatively low. Thus, we chose to use genotype likelihood methods with ANGSD to obtain a more disciplined picture of each individual's genotype [3]. Since the resulting site frequency spectrum (SFS) was bimodal, we folded the high-frequency SNPs since these genotypes were likely the true ancestral genotype. We then obtained our final SFS based on genotype likelihood, and calculated the observed average of pairwise differences π, the estimated global mutation rate θ, and Tajima's D test statistic [6],

$$D = \frac{\pi - \theta}{\mathrm{Var}(\pi - \theta)}. \tag{1}$$

We then compiled the mean of Tajima's D for each site in each population, to identify which populations showed evidence of undergoing recent contraction.

3 Results

Based on inspection of the FastQC output, the sequenced reads were of excellent quality. For the PRK population, after trimming was performed, all samples had a confidence interval of Phred quality scores well within the 28-40 interval for almost all base positions (as determined by the "Per base sequence quality" in the FastQC output) for both the forward and reverse strands.

Statistics for the sequence alignment against the *P. abies* genome can be found in Table 1. Each individual had an average read depth of around 3-3.8, indicating a relatively low depth and justifying our choice for employing genotype likelihood. Figure 1 shows the distribution of the estimated θ, average pairwise differences π, and Tajima's D for the PRK population, in addition to the folded SFS. The mean of these distributions then gave our global measures of genetic diversity, namely, the per site θ and π, along with Tajima's D. For PRK, these values were $\theta = 0.001882$, $\pi = 0.0038567$, and $D = 1.59265$.

4 Conclusion

Based on our positive value of Tajima's D in all population except one, we find there is a genetic signal matching the observed contraction in the red spruce's range. However, in spite of our use of genotype likelihoods, the relatively low read depth, along with the small number of samples per population, reduces the power of the Tajima's D test statistic. It is possible that further sampling will ultimately decrease the variance in diversity parameters, showing that Tajima's D is higher than observed here. Hence, future work should seek to corroborate our findings with further sampling, as well as explore measures of intra-population diversity such as GWAS.

NumReads	R1	R2	Paired	MateMapped	Singletons	MateMappedDiff	AvgDepth
2051979	1024826	1027153	1283918	1869112	56137	572314	3.27
1989570	994197	995373	1290946	1835388	46968	531000	3.30
2408441	1203148	1205293	1495276	2200526	64433	689130	3.54
2771356	1384400	1386956	1701572	2537188	71621	817970	3.81
1823911	911112	912799	1131078	1661130	51672	518236	3.05

Table 1: Results from running samtools `flagstat` and `depth` on each sample from the PRK population. Column definitions: NumReads indicates total number of reads. R1/R2 is the number of forward/reverse reads, respectively. Paired is the number of reads which are paired during sequencing. MateMapped is the number of reads which were mapped as a proper pair. Singletons are the number of singleton reads. MateMappedDiff is the number of reads which had a mate mapped to a different chromosome. AvgDepth is the average depth of each read.

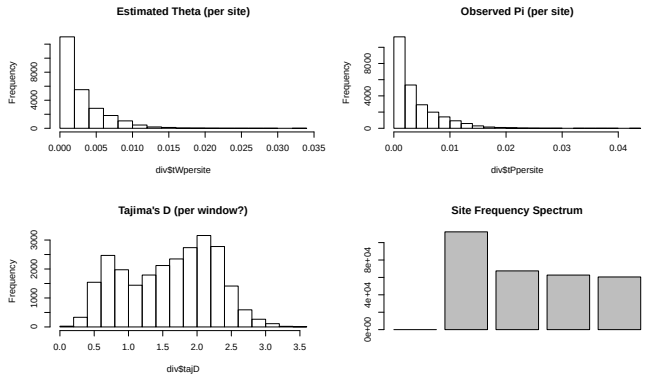

Figure 1: Population diversity statistics for the PRK population, calculated with ANGSD. The per site values for θ and π were assumed to be the corresponding value for each sequence window, divided the window length. Tajima's D was calculated using the values for θ and π for each window. Bottom right: the folded site frequency spectrum of SNPs for PRK. The first bin of the histogram was removed to allow for comparison for the other much smaller bins.

References

[1] Fastqc. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. Accessed: 2-25-2020.

[2] M. Jones and J. Good. Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, 25(1):185–202, 2016.

[3] T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(1):1–13, 2014.

[4] B. Nystedt, N. Street, A. Wetterbom, A. Zuccolo, Y. C. Lin, D. Scofield, F. Vezzi, N. Delhomme, S. Giacomello, A. Alexeyenko, R. Vicedomini, K. Sahlin, E. Sherwood, M. Elfstrand, L. Gramzow, K. Holmberg, J. Hällman, O. Keech, L. Klasson, M. Koriabine, M. Kucukoglu, M. Käller, J. Luthman, F. Lysholm, T. Niittylä, Å. Olson, N. Rilakovic, C. Ritland, J. Rosselló, J. Sena, T. Svensson, C. Talavera-López, G. Theißen, H. Tuominen, K. Vanneste, Z. Q. Wu, B. Zhang, P. Zerbe, L. Arvestad, R. Bhalerao, J. Bohlmann, J. Bousquet, R. Garcia Gil, T. Hvidsten, P. De Jong, J. MacKay, M. Morgante, K. Ritland, B. Sundberg, S. L. Thompson, Y. Van De Peer, B. Andersson, O. Nilsson, P. Ingvarsson, J. Lundeberg, and S. Jansson. The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451):579–584, 2013.

[5] W. Provine. Ernst Mayr: Genetics and speciation. *Genetics*, 167(3):1041–1046, 2004.

[6] F. Tajima. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123:585–595, 1988.