EG homework 1

P/BIO381 Ecological Genomics Spring 2020
Homework assignment #1

summarize and distill your technical work

- 2 pages (max) single spaced, including tables/figures (references on a separate 3rd page)

technical report
citations used when referring to methods or making factual assertions

Background (1-2 paragraphs)

A brief description providing context and **motivation of the problem** we're trying to address with these data

Brief background on the study species, biological samples, library prep, and sequencing strategy (look through early tutorials for info, plus your notes from class)

Study species

Picea rubens, commonly known as **red spruce**, is a species of spruce native to eastern North America,
on the red list, but least concern
The Central Appalachian Spruce Restoration Initiative (CASRI)[20] seeks to unite diverse partners with the goal of restoring historic red spruce ecosystems across the high-elevation landscapes of central Appalachians.

It thrives in the **cool, moist climates** of the high elevation mountains of the Apppalachians and northward along the coastal areas of Atlantic Canada.

Panylogical and paleoeco- logical studies have revealed that during the last glacial max- imum, around 18 000 yr BP (see Dyke and Prest, 1987), spe- cies from the genus *Picea* were confined to glacial refugia in the Great Plains (McLeod and MacDonald, 1997) and in the southeastern parts of the United States (Davis, 1983a, b). Sub- sequently, spruce species migrated northward.

Allopatric speciation promoted by geographic isolation is believed to be a dominant force shaping taxonomical diversity in the genus *Picea* (Wright, 1955).

On the other hand, long-term habitat fragmen- tation driven by large-scale climatic fluctuations, such as those likely experienced during the Pleistocene by boreal species, must have provided ideal conditions for allopatric speciation (Critchfield, 1984; Nowak et al., 1994).

plays a prominent role in montane communities throughout the Appalachians

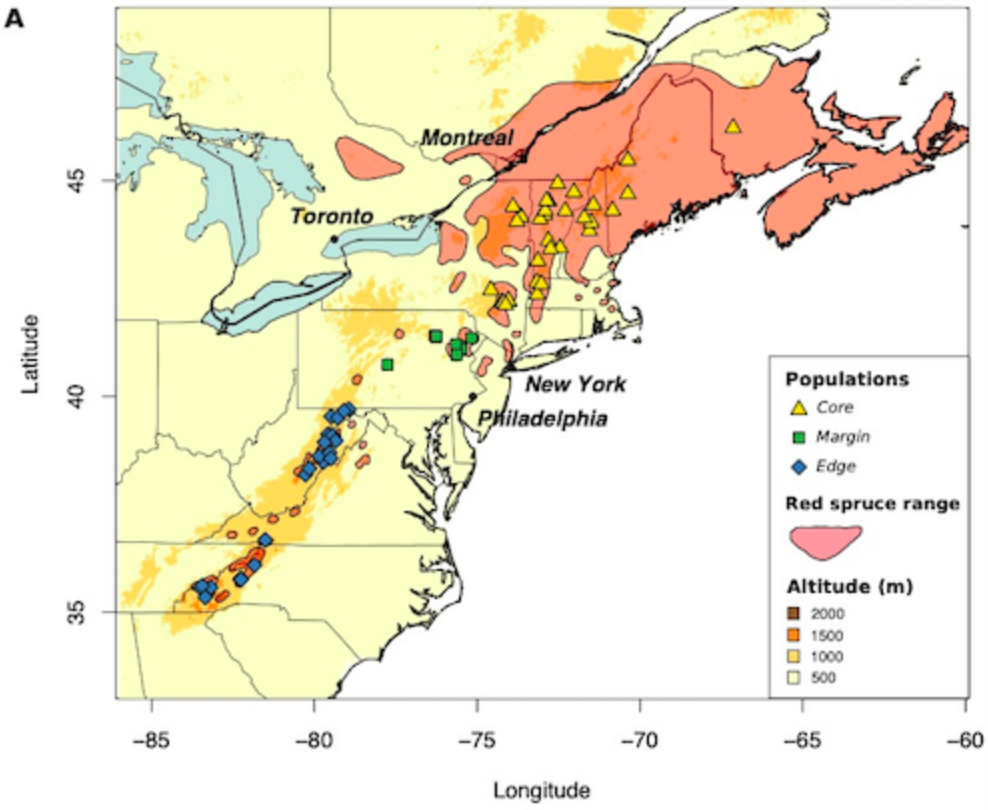
where populations are particular vulnerable to climate change is in the low-latitude trailing edge of the range, from Maryland to Tennessee, where populations are highly fragmented and isolated on mountaintops

"island" populations are remnants of spruce forests that covered the southern U.S. glaciers extended as far south as Long Island, NY. As the climate warmed at the end of the Pleistocene (~20K years ago), red spruce retreated upward in elevation to these mountaintop refugia, where they are now highly isolated from other such stands and from the core of the range further north.

A goal of our study is to better understand the genetic resource represented by these fragemented edge populations, and to use that information to help inform conservation biologists working to restore red spruce in this region

Genetic data suggests that the *red spruce* peripatrically speciated from the black spruce during the *Pleistocene* due to glaciation

Experiment:
Keller Lab collected samples from trees across the Appalachian Mountains
seeds and needle tissue
exome capture sequencing



For our class work, we're going to be focusing on **just the Edge samples.**
edge region = 110 mother trees from 23 populations

Libraries were made by random mechanical shearing of DNA (250 ng -1ug) to an average size of 400 bp followed by end-repair reaction, ligation of an adenine residue to the 3'-end of the blunt-end fragments to allow the ligation of barcoded adapters, and PCR-amplification of the library.

Libraries were sequenced on a single run of a Illumina HiSeq X to generate paired-end 150-bp reads

Bioinformatics Pipeline (2-3 paragraphs)

Detailed description of the **various steps** you used for the analysis of the sequencing data.

Take it from QC assessment of the raw reads up to estimation of population genomic diversity.
Show understanding why each step was done

Include justification for using particular analysis approaches or choices as appropriate (e.g., Why did we map to a reduced ref instead of the entire *P. abies* reference genome?; Why did we use ANGSD instead of analyzing “hard called” genotypes?, etc...).

XCV population had 5 samples

Trimming

Visualize the quality of raw data = FastQC
Clean raw data = Trimmomatic
Visualize the quality of cleaned data = FastQC

-> why/how Trimmomatic: Trimmomatic performs the cleaning steps in the order they are presented. clip adapter early in the process and clean for length at the end uses both reads at the same time!
“fast, multithreaded command line tool that can be **used** to trim and crop Illumina (FASTQ) data as well as to remove adapters”
our settings:

```
Cut adapter and other illumina-specific sequences from the rea
d
Cut bases off the start of a read, if below a threshold qualit
y = 20
Cut bases off the end of a read, if below a threshold quality
= 20
Scan the read with a 4-base wide sliding window, cutting when
the average quality per base drops below 15 (SLIDINGWINDOW:4:1
5) = here 6:20
Drop the read if it is below a specified length = 35
```

-> what did we see

before trimming:

Total sequences	Sequence length	Sample	Times 2
1263747	150	1	2527494
1838400	150	2	3676800
1597569	150	3	3195138
1831965	150	4	3663930
1190046	150	5	2380092
1,544,345.4 on average	150 on average	beginning and ends lower quality	

3088690.8 on average total

after trimming:

Total sequences	Sequence length	Times 2
1209244	35-150	2418488
1757907		3515814
1488139		2976278
1704949		3409898
1113528		2227056
1,454,753.4 on average	no low quality bp on ends	

1,454,753.4 on average of cleaned, high quality reads going into mapping
x2 for everything because this is just data from R1!
2909506.8 total on average

Mapping

reference assembly: N50: 4869
Norway spruce (*P. abies*)

Rather than trying to map to the entire 19.6 Gbp reference (yikes!), we first subsetted the *P. abies* reference to include **just the contigs that contain one or more probes**

from our exon capture experiment. For this, we did a BLAST search of each probe against the *P. abies* reference genome, and then retained all scaffolds that had a best hit.

reduced reference contains:
668,091,227 bp (~668 Mbp) in 33,679 contigs
The mean (median) contig size is 10.5 (12.9) kbp
The N50 of the reduced reference is 101,375 bp

Map (a.k.a. Align) cleaned reads from each sample to the reference assembly to generate **s**equence **a**lign**m**ent files = BWA
BWA-MEM = best for short sequences
-> very efficient and very well vetted read mapper

```
labels a read with a special flag if its mapping is split acro
ss >1 contig, -M -> allow reads to map on different contigs -
those contigs could be close by on the chromosome
keeps alignments involving unpaired reads
```

--> SAM file

Convert our sam files to the more efficient binary version (bam)

Get rid of any PCR duplicate sequences

both of these steps were done using sambamba version 0.7.1

-> mapping stats in results

Analysis of Next Generation Sequence Data (ANGSD)

Analysis of Next Generation Sequence Data (ANGSD) is a multithreaded program suite “can calculate various summary statistics, and perform association mapping and population genetic analyses by using genotype likelihoods”

We used ANGSD to estimate site frequency spectrum (SFS) and nucleotide diversities (Watterson’s theta, pi, Tajima’s D) based on genotype likelihoods instead

of “hard called” genotypes because

Ind1 is homoozygous at SNP-1 (CC) – couldn’t it be CT and we just didn’t have enough coverage to observe the second allele?

check settings for ANGSD:

various parameters to exclude lower quality bases, reads that poorly mapped to reference and not enough read depth (here 3)
don’t use sites with >2 alleles
estimate based on GL!
Keep only site highly likely to be polymorphic (SNPs) (pval 1e-6)
folded SFS

Results (1-2 paragraphs)

Report your findings from the different analysis steps. Use a combination of reporting results in-line in your text and summarizing more detailed information in tables and/or figures. You may use a max of 3 tables/figures total. Be sure each table/figure has a title, and a very brief legend describing its contents.

Include in your results section both “methodological” results (numbers of reads, mapping stats, depth of coverage, etc) as well as “biological” results (estimates of the genetic diversity metrics, etc).

calculate alignment statistics (% of reads mapping succesully, mapping quality scores, average depth of coverage per individual)

XCV.coverage.txt:
3.4351
4.03239
3.69699
4.05053
3.25873

-> how deep the coverage was, i.e. how many sequences we had for each reference region

XCV.flagstats.txt:

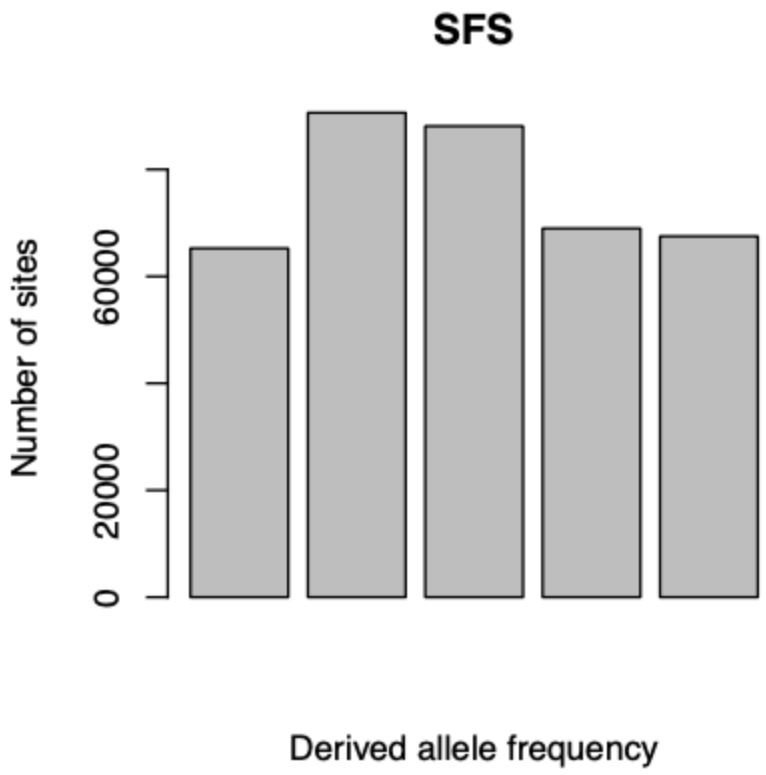
Num.reads	R1	R2	Paired	MateMapped	Singletons	MateMappedDiffChr	
2195061	1096682	1098379	1364782	2015494	56021	636926	313963
3143688	1570351	1573337	1916810	2865258	83851	928468	457713
2703977	1350702	1353275	1669360	2440168	80360	752840	375112
3079150	1537771	1541379	1921638	2785558	88094	842776	419362
2035599	1017017	1018582	1285144	1855002	56077	556484	276894

stats about BWA quality: how many total reads, R1, R2, how many paired mapped and how many cases only mapped one of the two reads (singleton), how many mapped to diff chromosomes... report in a table, including matmapped/total number of reads

ANGSD

SFS sum = 37606347 = total number of sites considered
non-polymorhic sites = 37225909.15
number of polymorhic sites = 380437.9
thus **percentage polymorphic** = 1 - non-polymorhic sites/total number of sites = (1 - 37225909.15/37606347)*100 = 1.011632%

-> plot: x axis: number of individuals with derived allele, y-axis: number of sites
here tot number of individuals = 5
and folded spectra go from 0 to 1N!



Tajima's D is computed as the difference between two measures of genetic diversity: the mean number of pairwise differences and the number of segregating sites, each scaled so that they are expected to be the same in a neutrally evolving population of constant size. also, the expectation is that both estimators will be equal to theta, which is the population scaled mutation rate.

$\theta = 4N_e\mu$

pi
mean **theta Pi persite**: 0.004255
pi = average pairwise difference among individuals

and W
mean **theta W persite**: 0.003554
Watterson estimator
number of segregating sites

mean **Tajima's D:** 0.8910

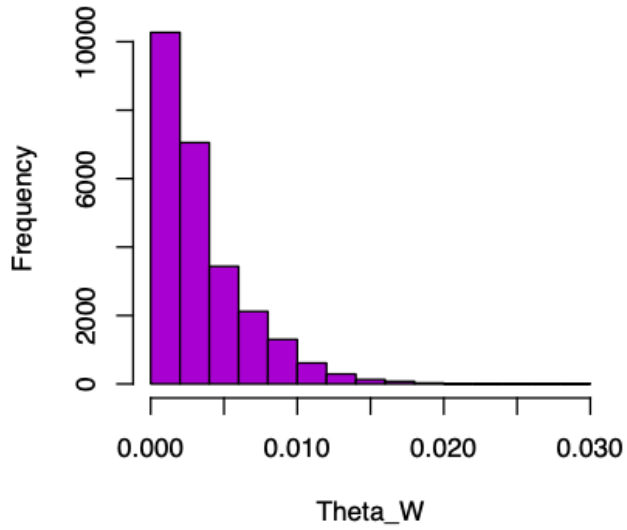
Tajima's D = pi - w
if >0 more diversity than expected ← balancing selection or pop contraction
if <0 less diversity than expected ← selection sweep or quick pop expansion

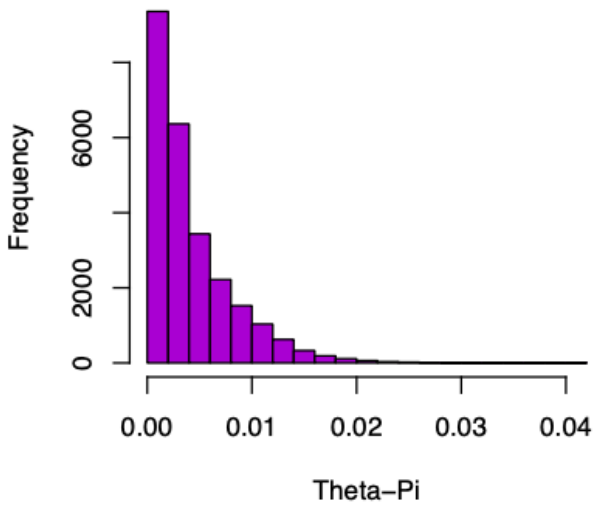
However, selective sweeps -> negative Tajima's D because # segragating sites > pi as most individuals will be the same but when mutations occur they are rare
and therefore pie underestimates theta compared to # segragating sites
pi - segrg site = negative as pi is smaller
or when population expansion after a recent bottleneck = same reason, most of variation is randomly missing after bottleneck but when mutations occur they are rare

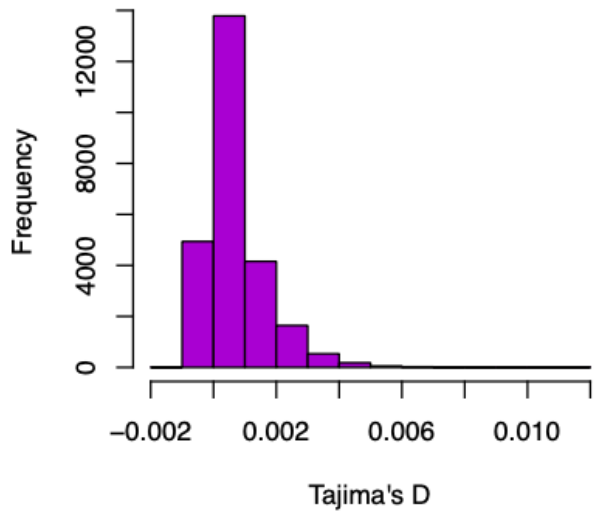
if Tajima's D is positive: More haplotypes (more average heterozygosity) than # of segregating sites. -> lack of rare alleles -> Balancing selection, sudden population contraction
because a) polymorphism is maintained in the population
b) remove "inbetween" randomly, lot of pairwise differences
how to distinshush between the two?
based on apriori knowledge probs not a) because strong selection for trees on upper regions by settlers

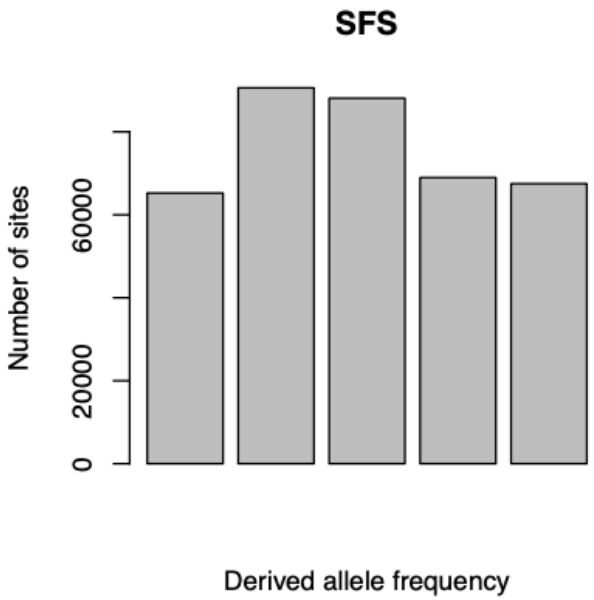
SFS also shows a lot of sites with high derived allele frequency (however could be skewed)

it is more than 0 -> based on what we know about red spruce contraction is a likely explanation, random individuals are removed → lots of pairwise difference









Conclusion (1-2 paragraphs)

Give your biological conclusion so far from the data:
What have we learned about the diversity and demographic history in this set of populations?
Relate back to your motivation given in the Background section.

Discuss any caveats or uncertainties that should be considered when interpreting the biological conclusions.
Discuss any methodological challenges encountered along the way that are relevant to your results and their interpretation.
Discuss opportunities for future directions.

3 components: glaciers -> human activity cutting lower ones -> further climate change
other species like this?

coming back (current expansion of red spruce into portions of its former range) -> least concern cathegory - however climate change my take away their remaining little habitat

West Virginia is predicted to warm approximately 5 degrees Fahrenheit by mid-century under medium emissions scenarios (Gervitz et al. 2009). Precipitation is predicted to increase from 5% to 8% in the same time period. The increased precipitation is not, however, enough to offset increased evapotranspiration as habitats warm. Overall, habitats are predicted to experience net drying throughout the state, especially during summer and early fall (Gervitz et al. 2009, Young et al. 2010).
Extreme events, including floods, droughts, and severe storms are expected to increase as well (Pachauri and Reisinger 2007). poorly understood impacts on species and communities

References (listed on a separate page)

- Cite papers in MLA format. Example:
- Nystedt, Björn, et al. "The Norway spruce genome sequence and conifer genome evolution." *Nature* 497.7451 (2013): 579-584.

(these are already in MLA)
Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. "ANGSD: analysis of next generation sequencing data." *BMC bioinformatics* 15.1 (2014): 356.

Github:

Have your github lab notebook up to date, and any scripts used for your analysis available in your github “myscripts” folder.