Population genomics noteook

Author: Gwen Ellis

Affiliation: University of Vermont, Dept. of Biology

E-mail contact: gwen.ellis@uvm.edu

Start Date: 2023.09.11

End Date: 2023.XX.XX

Project Descriptions:

This notebook will be used to document my workflow during the population genomics module of the 2023 Ecological Genomics course.

Table of Contents:

Entry 1: 2023-09-11.

- Reviewed red spruce (*Picea rubens*) study system and exome data
- set up working directories
- Fastqc of single sample in *P. rubens* data and analyzed quality

Entry 2: 2023-09-13.

- Reviewed fastqc output
- We saw good quality data for most of the read length, though the first 5bp had some variable base frequencies and the very end reads had slightly lower Q scores
- Using the our "fastp.sh" script, we mapped sequencing data from 2019 *P. rubens* population to genome using fastp
- comparison of pre and post trimming html files looked good! Removed the low quality bases at head and tail of sequence
- mapped sequences to Norway spruce (*P. abies*) reference exome using bwa program in "mapping.sh" script

Entry 3: 2023-09-18.

- Visualized .sam files and checked sam FLAGs
- Using sambamba and samtools, converted .sam files to .bam files, removed duplicate reads, and indexed files with our "process_bam.sh" script

Entry 4: 2023-09-20.

- calculating mapping stats using samtools with our population specific "bam_stats.sh" bash script
- output script 2019.stats.txt is organized in rows by each sample in the population with the following columns: Num_reads, Num_R1, Num_R2, Num_Paired, Num_MateMapped, Num_Singletons, Num_MateMappedDiffChr, Coverage_depth
- Coverage depth of this population is ~4x on average, which is low coverage but works for what we need since we're not going to be calling exact genotypes and instead will be estimating genotype likelihoods
- using ANGSD (Analyzing Next Generation Sequencing Data) program to calculate genotype likelihoods and allele frequencies for 2019 population with "ANGSD.sh" script with the following parameters:
- -nThreads 1 -remove_bads 1 -C 50 -baq 1 -minMapQ 20 -minQ 20 -GL 1 -doSaf 1
- output is site allele frequency (saf) likelihoods

Entry 5: 2023-09-25.

- estimating the site frequency spectrum (sfs) using saf likelihoods and our ANGSD_doTheta.sh, then using this to estimate nucleotide diversity and F_ST_ : mypop vs. black spruce
  - theta_w_ = segragating sites
  - theta_pi_ = pairwise diversity
  - Tajima's D = (theta_pi_ - theta_w_)/SD
- using an unfolded sfs since we don't know if the reference allele is the ancestral allele or not.
- See Summary_diversity.R for diversity metric summary plots

Entry 6: 2023-09-27.

- calculating Fst between 2019 population and black spruce
- unweighted: 0.094249

- weighted: 0.338065
- Within species Fst for trees is usually around 0.05, but between species is closer to 0.1 or 0.2
- pcANGSD for population structure analysis (PCA) across all collected red spruce populations
- ancestry plot for K=2

Entry 7: 2023-10-02.

- identifying outlier loci based on 2 Eigen vectors
- using Norway spruce (*P. abies*) annotated genome to identify genes that outlier loci SNPs are in
- Interestingly, PC1 has fewer outlier loci than those separated by PC2 (13 vs. 212)
- Used PlantGenIE to see what possible functions our identified genes have
- PFAM enrichment for PC1 outliers reveals 9 protein families with significant p-value

Entry 8: 2023-10-04.

- genotype-environment associations
  - outlier list (PC1) -> genetic PC1 and PC2 as covariants -> use bioClim variables for E of GEA
- bio12 and bio10 were most correlated with PC1 and PC2, respectively