Ecological Genomics Tutorials: Population & Landscape Genomics 5

Setpember 25, 2023

Learning Objectives for 09/25/23

1. Calculate diversity stats for our focal pops (SFS, theta-W, theta-Pi, Tajima's D)
2. Summarize the results in R and share to google doc
3. Introduce Fst in ANGSD using genotype probabilities

1. Calculate SFS and diversity stats

At the end of our last session, we used ANGSD to estimate genotype likelihoods for our red spruce populations. We wrote the script `ANGSD.sh` to work on these, which should have the following output files in your `~/myresults/ANGSD` directory:

```
-rw-r--r--. 1 kellrlab users 1417463962 Sep 20 11:44 9999_.saf.gz
-rw-r--r--. 1 kellrlab users    1086161 Sep 20 11:44 9999_.saf.idx
-rw-r--r--. 1 kellrlab users   79501025 Sep 20 11:44 9999_.saf.pos.gz
```

These "saf" files contain "site allele frequency" likelihoods, and are the info needed to estimate stats that depend on population allele frequencies, like nucleotide diversities, neutrality stats like Tajima's D, and population divergence stats like Fst. Each of these stats depends on the *SFS – the Site Frequency Spectrum* So, our workflow will be to use the .saf files to estimate the SFS, and then use the SFS to estimate our diversity stats and Fst.

In your `~/myscripts` folder, create `ANGSD_doTheta.sh` to estimate the SFS and nucleotide diversity stats for your pop

Based on the saf.idx files from ANGSD GL calls, first estimate the Site Frequency Spectrum (SFS)

```
REF="/netfiles/ecogen/PopulationGenomics/ref_genome/Pabies1.0-genome_reduced.

OUT=~/myresults/ANGSD

MYPOP=""

SUFFIX=""

#Estimation of the SFS for all sites using the FOLDED SFS
```



```
realSFS ${OUT}/${MYPOP}_${SUFFIX}.saf.idx \
        -maxIter 1000 \
        -tole 1e-6 \
        -P 1 \
        -fold 1 \
        > ${OUT}/${MYPOP}_${SUFFIX}.sfs
```

After the SFS, estimate the theta diversity stats:

Once you have the SFS, you can estimate the thetas and neutrality stats by adding the following code chunk at the end of your `ANGSD_doTheta.sh` script:

```
# Estimate thetas and stats using the SFS

realSFS saf2theta ${OUT}/${MYPOP}_${SUFFIX}.saf.idx \
        -sfs ${OUT}/${MYPOP}_${SUFFIX}.sfs \
        -outname ${OUT}/${MYPOP}_${SUFFIX}

thetaStat do_stat ${OUT}/${MYPOP}_${SUFFIX}.thetas.idx
```

If we wanted to analyze this on sliding windows, we could instead replace the above code chunk with the following:

```
# For sliding window analysis:

thetaStat do_stat ${OUT}/${MYPOP}_${SUFFIX}.thetas.idx \
-win 50000 \
-step 10000 \
-outnames ${OUT}/${MYPOP}_${SUFFIX}.thetasWindow.gz
```

An important distinction! **The unfolded vs. folded SFS**

The big difference here is whether we are confident in the ancestral state of each variable site (SNP) in our dataset

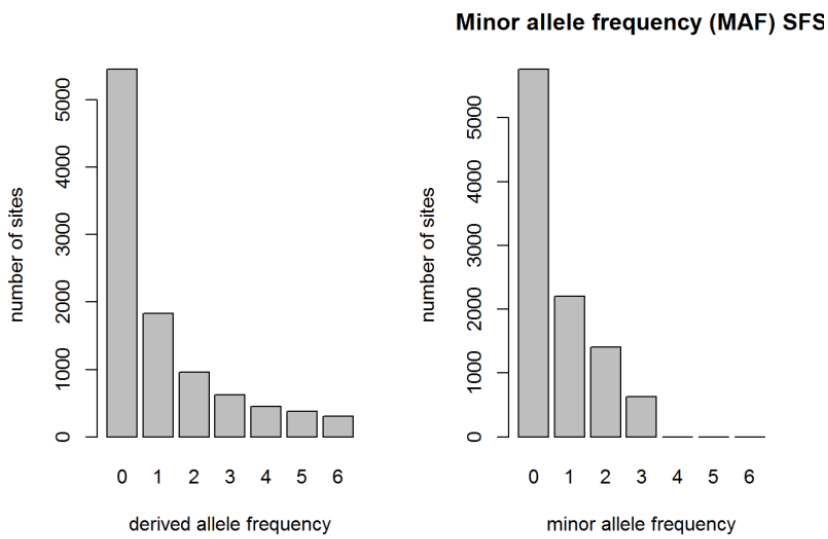
If we know the ancestral state, then the best info is contained in the **unfolded** SFS, which shows the frequency histogram of how many derived loci are rare vs. common

- bins in the **unfolded** SFS go from 0 to 2N – why?

When you don't know the ancestral state confidently, you can make the SFS based on the minor allele (the less frequent allele; always < 0.5 in the population).

- bins in the **folded** SFS go from 0 to 1N – why?

Essentially, the folded spectra wraps the SFS around such that high frequency "derived" alleles are put in the small bins (low minor allele freq).



Now we have some diversity results!

For either of the results files above, the first column of the results file (*.thetas.idx.pestPG) is formatted a bit funny and we don't really need it. There are also a bunch of other neutrality stats that we don't need right now. We can use the `cut` command to get rid of these extra columns. The below `cut` command retains columns 2-5, 9 and 14

```
cut -f2-5,9,14 ${OUT}/${MYPOP}_${SUFFIX}.thetas.idx.pestPG > ${OUT}/${MYPOP}_
```

The columns correspond to the following stats:

Col	Statistic	Description
2	Chr	The chromosome or (more appropriately) contig being analyzed
3	WinCenter	The center of the contig, in bp
4	tW	Watterson's Theta – an estimate of nucleotide diversity based on segregating sites
5	tP	Theta Pi – estimate of nucleotide diversity based on pairwise divergence
9	Tajima	Tajima's D – a neutrality stat that tests for the difference in tW-tP

Col	Statistic	Description
14	nSites	The number of base pairs being analyzed along this stretch of cont.

2. Summarize diversity stats in R

We're now ready to use `Filezilla` to download these 2 files:

1. .thetas.idx.pestPG
2. .sfs

Save these to your `results` folder in your githuib repo so we can import into R to look at the mean and variability in nucleotide diversity for our pop. Here's some basic R code to help you along:

```
setwd("") # set your path to your results folder in your repo where you saved

list.files() # list out the files in this folder to make sure you're in the r

# First let's read in the diversity stats
theta <- read.table("_.thetas",sep="\t",header=T)

theta$tWsite = theta$tW/theta$nSites #scales the theta-W by the number of sit
theta$tPsite = theta$tP/theta$nSites #scales the theta-Pi by the number of si

summary(theta)

# You can order the contig list to show you the contigs with the highest valu

head(theta[order(theta$Tajima, decreasing = TRUE),]) # top 10 Tajima's D valu

head(theta[order(theta$Tajima, decreasing = FALSE),]) # bottom 10 Tajima's D

#You can also look for contigs that have combinations of high Tajima's D and
#theta[which(theta$Tajima>1.5 & theta$tPsite<0.001),]


sfs<-scan('9999_.sfs')
sfs<-sfs[-c(1,which(sfs==0))]
sfs<-sfs/sum(sfs)

# Be sure to replace "9999" with your pop code in the "main" legend below
barplot(sfs,xlab="Chromosomes",
        names=1:length(sfs),
        ylab="Proportions",
        main="Pop 9999 Site Frequency Spectrum",
        col='blue')
```

```
# Put the nucleotide diversities, Tajima's D, and SFS into a 4-panel figure
par(mfrow=c(2,2))
hist(theta$tWsite, xlab="theta-W", main="Watterson's theta")
hist(theta$tPsite, xlab="theta-Pi", main="Pairwise Nucleotide Diversity")
hist(theta$Tajima, xlab="D", main="Tajima's D")
barplot(sfs,names=1:length(sfs),main='Site Frequency Spectrum')


# To reset the panel plotting, execute the line below:
dev.off()
```

We can compare the diversity in our different pops by entering your diversity stats in this google doc:

https://docs.google.com/spreadsheets/d/1y3GMvnGP65fYfBoJsdrixGLkGCe_TcDlo_7GFyPe1QM/edit?usp=sharing

3. Use ANGSD and the SFS for multiple pops to calculate genetic divergence between pops (Fst)

We can calculate Fst between any pair of populations by comparing their SFS to each other. For this, we'll need to estimate the SFS for pairs of populations; we can each contribute to the overall analysis by looking at how our focal pop is divergent from the others.

- For this analysis, let's calculate Fst between our focal red spruce population (MYPOP) and the black spruce samples...this could tell us which of our pops might be hybridizing. What would we expect for Fst in this case?
- Let's write a bash script called ANGSD_Fst.sh that includes the following code:

```
# Start with the usual bash script header

# Give yourself some notes

# Path to Black Spruce (BS) input saf.idx data:

BLKSPR="/netfiles/ecogen/PopulationGenomics/fastq/black_spruce/cleanreads/bam

#Path to save your output:

OUTPUT=

MYPOP=""

SUFFIX=""

# Estimate Fst between my red spruce pop and black spruce:

realSFS ${MYPOP}_.saf.idx ${BLKSPR}/BS_all.saf.idx -P 1 >${MYPOP}_BS.sfs
```



```
realSFS fst index ${MYPOP}_.saf.idx ${BLKSPR}/BS_all.saf.idx -sfs ${MYPOP}_BS
realSFS fst stats ${MYPOP}_BS.fst.idx
```