Ecological Genomics Tutorials: Population & Landscape Genomics 4

September 20, 2023

Learning Objectives for 09/20/23

1. Finish calculating mapping statistics to assess quality of the result
2. Introduce use of genotype-likelihoods for analyzing diversity in low coverage sequences
3. Use the 'ANGSD' program to calculate nucleotide diversity (thetas) and neutrality stats

1. Calculate mapping stats: How can we get a summary of how well our reads mapped to the reference?

- We can use the program samtools Written by Heng Li, the same person who wrote bwa. It is a powerful tool for manipulating sam/bam files.
- The samtools command `flagstat` gets us some basic info on how well the mapping worked
- We can also estimate depth of coverage (avg. number of reads/site) using the samtools command `depth`
- We'll use both of these commands in loops to assess the mapping stats on each sample in our population.
- We'll also use the `awk` tool to help format the output. (awk cheatsheet here)
- Here's a script to get us started:
  `/netfiles/ecogen/PopulationGenomics/scripts/bam_stats.sh`

2. Inference of population genomics from the aligned sequence data: should we call genotypes?

Many of the papers you'll read that do popgen on NGS data have a SNP calling step that results in a specific gneotype being called for each SNP site for each individual. For example,

SNP	Ind1	Ind 2
1	CC	CT

SNP	Ind1	Ind 2
2	AG	AA
3	GT	TT

But how do we know that Ind1 is homozygous at SNP-1 (CC) – couldn't it be CT and we just didn't have enough coverage to observe the second allele?

The basic problem is that read data are counts that produce a binomial distribution of allele calls at a given site, and if you have few reads, you might by chance not observe the true genotype. So, what's the right thing to do?

As with almost anything in statistics, the right thing to do is not throw away that uncertainty, but instead incorporate it into your analysis. That's what we're going to do...

Genotype-free population genetics using genotype likelihoods

A growing movement in popgen analysis of NGS data is embracing the use of genotype likelihoods to calculate stats based on each individual having a likelihood (probability) of being each genotype.

A genotype likelihood (GL) is essentially the probability of observing the sequencing data (reads containing a particular base), given the genotype of the individual at that site.

These probabilities are modeled explicitly in the calculation of population diversty stats like Theta-pi, Tajima's D, Fst, PCA, etc...; thus not throwing out any precious data, but also making fewer assumptions about the true (unknown) genotype at each locus

- We're going to use this approach with the program 'ANGSD', which stands for 'Analysis of Next Generation Sequence Data'
- This approach was pioneered by Rasmus Nielsen, published originally in Korneliussen et al. 2014.
- ANGSD has a user's manual (it's a work in progress...)

The basic work flow of ANGSD goes like this:

1. Create a list of bam files for the samples you want to analyze
2. Estimate genotype likelihoods (GL's) and allele frequencies after filtering to minimize noise
3. Use GL's to:
   a. estimate the site frequency spectrum (SFS)
   b. estimate nucleotide diversities and neutrality stats (Thetas, Tajima's D, ...)

1. In your `~/myscripts` folder, enter vim and create a file called `ANGSD.sh`

Create the header for your inputs and outputs

```
mkdir ~/myresults/ANGSD

INPUT=""

OUTPUT=~/myresults/ANGSD

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

MYPOP=""

ls ${INPUT}/${MYPOP}*sorted.rmdup.bam >${OUTPUT}/${MYPOP}_bam.list
```

Write (w) and quit(q) your file and try running it at the command line.

Check your output bamlist to see it was made properly!

- Where would you look for this file? (Hint, refer back to the `ls` command that makes it).
- How would you verify its contents? (hint: use `head` or `cat` or even `vim`)

2. Open your `ANGSD.sh` script back up in `vim`

Estimate your GL's and allele freqs, optionally filtering for base and mapping quality, sequencing depth, SNP probability, minor allele frequency, etc.

Add the following code chunk at the bottom of your script:

```
# File suffix to distinguish analysis choices
SUFFIX=""

# Estimating GL's and allele frequencies for all sites with ANGSD

######################

ANGSD -b ${OUTPUT}/${MYPOP}_bam.list \
-ref ${REF} -anc ${REF} \
-out ${OUTPUT}/${MYPOP}_${SUFFIX} \
-nThreads 1 \
-remove_bads 1 \
-C 50 \
-baq 1 \
-minMapQ 20 \
-minQ 20 \
-GL 1 \
-doSaf 1 \
##### below filters require `do-Counts`
```



```
#-doCounts 1 \
#-minInd 4 \
#-setMinDepthInd 1 \
#-setMaxDepthInd 40 \
#-setMinDepth 10 \
#-skipTriallelic 1 \
#-doMajorMinor 1 \
##### below filters require `doMaf`
#-doMaf 1 \
#-SNP_pval 1e-6 \
#-minMaf 0.01
```

What do all these options mean?

Option	Description
-nThreads 1	how many cpus to use – be conservative
-remove_bads 1	remove reads flagged as 'bad' by samtools
-C 50	enforce downgrading of map quality if contains excessive mismatches
-baq 1	estimates base alignment qualities for bases around indels
-minMapQ 20	threshold for minimum read mapping quality (Phred)
-minQ 20	threshold for minimum base quality (Phred)
-GL 1	calculate genotype likelihoods (GL) using the Samtools formula
-doSaf 1	output allele frequency likelihoods for each site
-doCounts 1	output allele counts for each site
-minInd 4	min number of individuals to keep a site (see also ext 2 filters)
-setMinDepthInd 1	min read depth for an individual to count towards a site
-setMaxDepthInd 40	max read depth for an individual to count towards a site
-setMinDepth 10	min read depth across ALL individual to keep a site
-skipTriallelic 1	don't use sites with >2 alleles

Option	Description
-doMajorMinor 1	fix major and minor alleles the same across all samples
-doMaf 1	calculate minor allele frequencies
-SNP_pval 1e-6	Keep only site highly likely to be polymorphic (SNPs)
-minMaf 0.01	Keep only sites with minor allele freq > some proportion.

NOTES

- If you want to restrict the estimation of the genotype likelihoods to a particular set of sites you're interested in, add the option `-sites selected_sites.txt` (tab delimited file with the position of the site in column 1 and the chromosome in column 2) or use `-rf selected_chromosome.chrs` (if listing just the unique "chromosomes" or contigs you want to anlayze)
- Some popgen stats you want to estimate only the polymorphic sites; for this you should include the `-SNP_pval 1e-6` option to eliminate monomorphic sites when calculating your GL's
- There are good reasons to do it BOTH ways, with and without the `-SNP_pval 1e-6` option. Keeping the monomorphic sites is essential for getting proper estimates of nucleotide diversity and Tajima's D. But other analyses such as PCA or GWAS want only the SNPs.

Write (q) and quit (q) your script. Then run at the command line

3a. 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, you can estimate the Site Frequency Spectrum (SFS), which is the precursor to many other analyses such as nucleotide diversities (as well as Fst, demographic history analysis, etc.)

```
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 \
        > ${OUT}/${MYPOP}_${SUFFIX}.sfs
```

3b. 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
```

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. We can use the `cut` command to get rid of it:

```
cut -f2- ${OUT}/${MYPOP}_${SUFFIX}.thetas.idx.pestPG > ${OUT}/${MYPOP}_${SUFF
```

This is now ready to bring into R to look at the mean and variability in nucleotide diversity for our pop. How does it compare to others?

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