Genetic diversity and Intersexual Fst in *Syngnathus floridae*

Coley Tosto

2023-11-08

```
#This is a cohesive list of all the libraries used in this document
library(DESeq2)
```

```
#Tajima's D data calculated with ANGSD
tajD <- read.delim("data/floridae.thetas.idx.pestPG", header = TRUE)

#Intersexual Fst calculated with ANGSD
fst <- read.delim("data/floridae_fm_fst.txt", header = FALSE)

#The abundance matrix generated via salmon and tximport to be used for the DE analysis
txi.salmon <- readRDS("data/txi.salmon.floride.RDS")

#The samples file generated for tximport to be used for DE analysis
samples <- read.table("FL_samples.txt", header = TRUE)

#Make sure the conditions are in the samples file as a factor
samples$Sex <- as.factor(samples$Sex)
samples$Organ <- as.factor(samples$Organ)
```

Generating BAM files

`Bowtie2` was used to map the processed reads back to the reference transcriptome, generating SAM files that will be converted to BAM files and coordinate sorted with `SAMtools` and then used by ANGSD to calculate measurements of Tajima's D and intersexual F_{st}.

`Bowtie2` was installed inside of a conda environment on the RCC named `trinity`. Bowtie2 v2.5.0 was used. `SAMtools` was installed in the same conda environment, v1.18 was used. The following script was then used to generate the index, map the reads back to it to create the SAM files, then convert those SAM files to BAM files, and finally coordinate sort the BAM files.

```
#!/bin/bash

#Create the arguments
ref_trans=$1 #Transcriptome .fasta file
index_dir=$2 #dir/basename the bt2 files will be written to
read_dir=$3 #location of the processed reads that will be aligned to index
sam_dir=$4 #Where you want the output SAM files to be stored
bam_dir=$5 #Where you want the output BAM files to be stored

#Build an index based on the transcriptome
```

```
echo "Indexing the refrence transcriptome"
time bowtie2-build --threads 16 $ref_trans $index_dir


#Map the reads back to the index
echo "Now beginning alignments ..."
for fq in $read_dir*_R1.fq.gz
    do

    #Extract sample name from the file
    sample=$(basename $fq _R1.fq.gz)

    #Echo the name of the sample that is currently running
    echo "Aligning reads for $sample"

    #Align this pair of reads
    time bowtie2 -x $index_dir \
        -1 $read_dir${sample}_R1.fq.gz \
        -2 $read_dir${sample}_R2.fq.gz \
        -S $sam_dir${sample}.sam --threads 16

    #Convert the SAM files to BAM files and coordinate sort
    echo "Converting sam file to bam file for ${sample}"
    time samtools view -T $ref_trans -b $sam_dir${sample}.sam > $bam_dir${sample}.bam
    echo "Sorting bam file for ${sample}"
    time samtools sort -@ 16 -o $bam_dir${sample}_sorted.bam $bam_dir${sample}.bam

done
```

The script was run as `nohup bash bash_scripts/bowtie2_alignment.sh trinity_supertran_floridae.fasta bowtie2_index/floridae floridae_kmer_corrected/ floridae_SAM/ floridae_BAM/ > bt2.log 2>&1 &`.

Running ANGSD

ANGSD will be used for the calculation of both Tajima's D and also Intersexual F_{st}. It was installed in the `shared/` folder on the RCC following the instructions given on their website as:

```
wget http://popgen.dk/software/download/angsd/angsd0.940.tar.gz
tar xf angsd0.940.tar.gz

cd htslib;make;cd ..

cd angsd
make HTSSRC=../htslib
cd ..
```

ANGSD version 0.940-dirty was used for the analysis.

The following script was then used to filter the BAM files, estimate site frequency spectrum (SFS), calculate the thetas and the calculate Tajima's D when `$ALLSFS=true and $THETAS=true`. It was then also used to filter BAM files and estimate SFS **separately** for males and females, combine them together and then calculate intersexual F_{st} when `$INDSFS=true and $MFFST=true`.

```
#!/bin/bash

#Choose what to run
ALLSFS=false
THETAS=false
INDSFS=true
MFFST=true

#Creating arguments
bam_dir=$1 #Folder where all of the desired BAM files are stored
angsd_dir=$2 #Path to the angsd executable
ref_trans=$3 #Reference transcriptome file (.fasta)
sfs_all_out=$4 #Desired name for the output SFS across all samples
stat_out=$5 #Desired name for the output of the Theta and/or Fst calculation

#Create the list of all SORTED bam files, just the female bam files and just the male bam files
ls ${bam_dir}*_sorted* > all_bams.txt
ls ${bam_dir}*M*_sorted* > male_bams.txt
ls ${bam_dir}FL*F*_sorted* > fem_bams.txt

if [ $ALLSFS = true ]; then
    #Estimate the allele frequencies from the BAM files
    echo "Estimating the site frequency spectrum across ALL samples"
    ${angsd_dir}angsd -b all_bams.txt \
            -doSaf 1 \
            -anc ${ref_trans} \
            -minMapQ 20 -remove_bads 1 -uniqueOnly 1 -only_proper_pairs 1 \
            -minQ 13 -minInd 4 \
            -GL 1 -P 16 -out ${sfs_all_out}

    ${angsd_dir}misc/realSFS ${sfs_all_out}.saf.idx -fold 1 -maxIter 100 -P 16 > ${sfs_all_out}.sfs

fi


if [ $THETAS = true ]; then
    #Calculating Thetas
    echo "Calculating thetas ..."
    ${angsd_dir}misc/realSFS saf2theta ${sfs_all_out}.saf.idx \
            -sfs ${sfs_all_out}.sfs \
            -outname ${theta_out}

    #Calculating Tajima's D
    echo "Calculating Tajima's D..."
    ${angsd_dir}misc/thetaStat do_stat ${stat_out}.thetas.idx

fi


if [ $INDSFS = true ]; then
    #Calculate the site frequency spectrum for males and females seperately
    echo "Estimating seperate SFS..."
    #First calculate per pop saf for each population (in our cases the different sexes)
    ${angsd_dir}angsd -b fem_bams.txt -doSaf 1 -anc ${ref_trans} \
            -minMapQ 20 -remove_bads 1 -uniqueOnly 1 -only_proper_pairs 1 \
            -minQ 13 -minInd 7 \
            -GL 1 -P 12 -out ${sfs_all_out}_fem

    ${angsd_dir}angsd -b male_bams.txt -doSaf 1 -anc ${ref_trans} \
            -minMapQ 20 -remove_bads 1 -uniqueOnly 1 -only_proper_pairs 1 \
            -minQ 13 -minInd 7 \
            -GL 1 -P 12 -out ${sfs_all_out}_mal

    #Calculate the 2dsfs prior
    ${angsd_dir}misc/realSFS ${sfs_all_out}_fem.saf.idx ${sfs_all_out}_mal.saf.idx -fold 1 > fem.mal.ml

fi

if [ $MFFST = true ]; then
    #Calculate M-F Fsts
    echo "Calculating male-female Fsts..."

    #Prepare the Fst for easy window analysis etc.
    ${angsd_dir}misc/realSFS fst index ${sfs_all_out}_fem.saf.idx ${sfs_all_out}_mal.saf.idx -sfs fem.ma

    #Get the global estimate
    ${angsd_dir}misc/realSFS fst stats ${stat_out}_fm_fst.fst.idx

fi
```

When $ALLSFS=true and $THETAS=true Following the ANGSD website we can see the first step for calculate Tajima's D is to **filter and estimate SFS**:

- To get the estimates of SFS you first generate a `.saf` file (site allele frequency likelihood) followed by an optimization of the `.saf` file which will then estimate the site frequency spectrum.
    - `doSaf 1` was used to calculate saf based on individual genotype likelihoods assuming HWE.
    - `GL 1` was used since SAMtools was used to generate the BAM files.
    - Because we don't have the ancestral state, we estimated the folded SFS by giving `-anc` the reference transcriptome and applying `-fold 1` to `realSFS`.
- For filtering, many options were used:
    - `minMapQ`: set the minimum mapping quality (20 was used here)
    - `remove_bads`: removes reads with a flag above 255 (set to 1 for remove)
    - `uniqueOnly`: when set to 1, removes reads that have multiple best hits
    - `only_proper_pairs`: when set to 1, includes only pairs of reads where both mates mapped correctly
    - `minQ`: Minimum base quality score (set to 13 here)
    - `minInd`: Remove if there was data in less than X individuals (4 here).

The second step is to then **calculate the thetas for each site**: - This is done using the `.sfa.idx` and the `.sfs` files from the step before

Lastly, with the output from the thetas calculation (`.thetas.idx`) we can **estimate Tajima's** D.

When $INDSFS=true and $MFFST=true Following the ANGSD website we can see the first step for calculating F_{st} is similar to Tajima's D, we have to **filter and estimate SFS**. The difference here is that we calculate the SAF **SEPARATELY** for the different populations and then put them together with the `realSFS` function. Because we want to look at differences between males and females our different "populations" are the two sexes.

- `-fold 1` was given to `realSFS` again because we do not have the ancestral state.
- The same filtering that was used for Tajima's D was applied here, however this time the RNA-seq data was filtered to only included bases where we had data in more than half of the individuals for male and females separately.
  - This was done by setting `minInd` to 7

After getting the male-female SFSs and combining them, intersexual F_{st} can be calculated with `realSFS fst`.

Tajima's D

After running ANGSD we can start to look through some of the results. To start with I am plotting the site frequency spectrum that was generated by ANGSD.

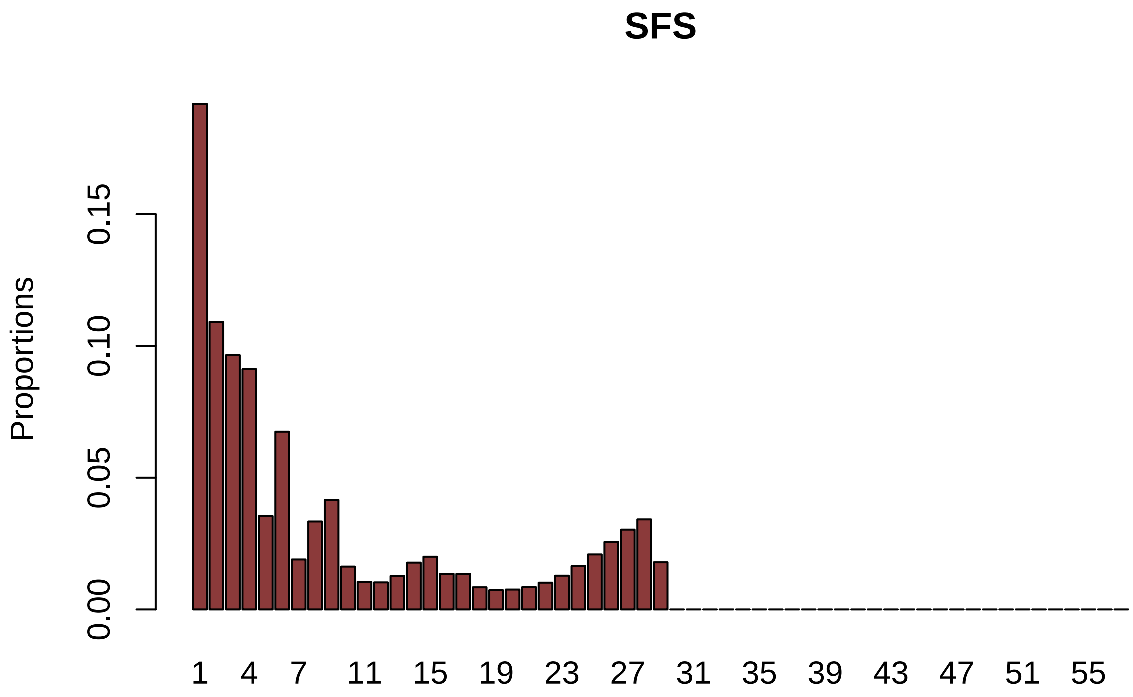

Figure 1: Site frequency spectrum calculated by ANGSD.

I then want to look at the overall distribution of Tajima's *D* that was calculated. Additionally, I want to see if each row of the Tajima's *D* dataset corresponds to an individual gene. There are 873 rows in the Tajima's *D* dataset. If each row was one gene than I would have originally expected to see 268936 rows as that is how many Trinity genes we have in our assembly. If we look as the unique IDs for our "chromosome" column in the Tajima's *D* dataset we can see that there are 873 unique IDs. This does tell us that each row is likely to be an individual Trinity gene and it is possible that the low amount of rows here compared to the number of genes in out transcriptome could be due to the filtering restrictions we applied above.

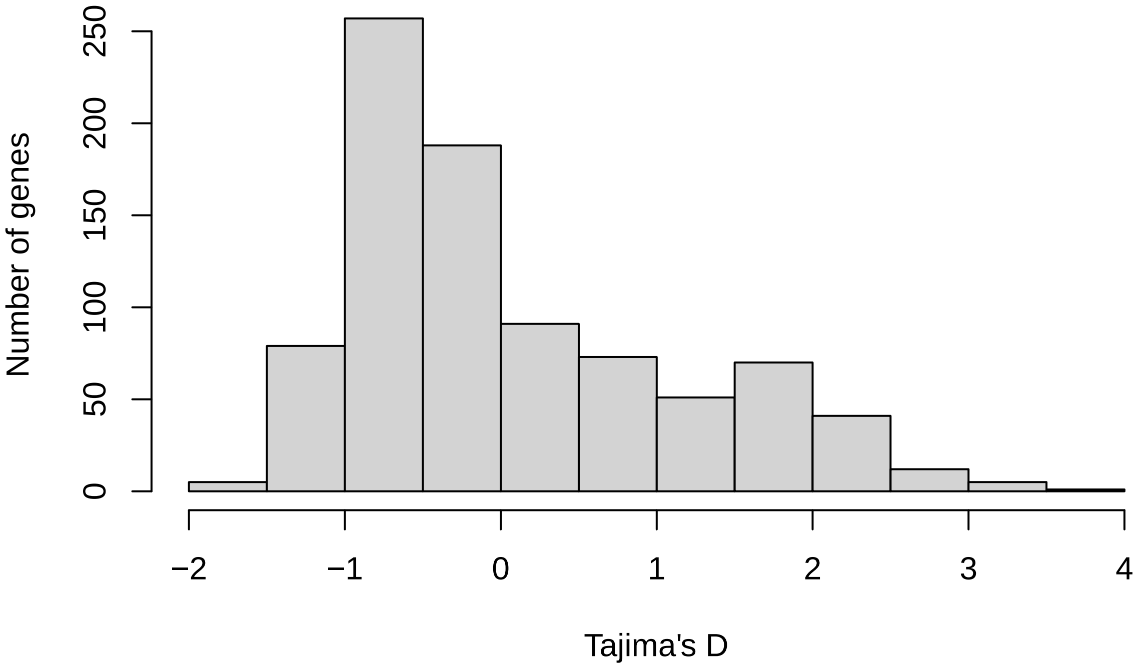

Figure 2: Histogram showing the ditribution of Tajima's D values.