Sardine structure and filtering follow-up

Katie Wagner

2026-01-08

We learned on Thursday that our sardines *Limnothrissa miodon* and *Stolothrissa tanganicae* have strong population structure that does not correspond with spatial structure. We hypothesized that the structure on PC1 corresponds with sex – that the reason we have two groups is that one corresponds to a male group and one to a female group. We then hypothesized that the structure on PC2 for *Limnothrissa* could correspond with an inversion, where we have two groups homozygous for the inversion, and one group heterozygous for the inversion. Let's think about ways that we could characterize this genetic structure more, and test our hypotheses.

First, let's get back to where we were with VCF files loaded into R.

```r
require(adegenet)
require(vcfR)
require(pegas)
```

```r
stolovcf<-read.vcfR('stolo_pca.recode_noHWEfilter.vcf.gz')
limnovcf<-read.vcfR('limno_pca.recode_noHWEfilter.vcf.gz')
```

```r
stolovcfgen<-vcfR2genind(stolovcf)
limnovcfgen<-vcfR2genind(limnovcf)
```

Metadata associated with our samples:

```r
fullfishinfo<-read.csv('All_fish_info.csv')

#Stolothrissa
Spairedinfo<-match(row.names(stolovcfgen$tab), fullfishinfo[,1])
Sfullfishinfomatched<-fullfishinfo[Spairedinfo,]

#Limnothrissa
Lpairedinfo<-match(row.names(limnovcfgen$tab), fullfishinfo[,1])
Lfullfishinfomatched<-fullfishinfo[Lpairedinfo,]
```

As we discussed on Thursday, PCA summarizes genetic variation by finding axes (principal components) that maximize overall variance in the data. PCA is widely used to visualize population structure because it makes few assumptions and does not require prior group assignments. It is a great "first pass" at genetic data because it allows us to easily visualize the major axes of variation across a dataset.

Discriminant function analyses are complementary to PCA in that they allow us to assess divergence associated with a priori groupings. Rather than asking what the major axes of overall variation are, discriminant analyses explicitly maximize divergence among groups that you define. In genetics, Discriminant Analysis of Principal Components (DAPC) is a multivariate method designed specifically to identify and describe genetic clusters. The discriminant functions maximize between-group variation while minimizing within-group variation.

- PCA is exploratory and unsupervised; it reveals major patterns of genetic variation without using population labels.
- DAPC requires predefined groups (or groups inferred via some sort of clustering) and focuses on discrimination among those groups.
- PCA axes reflect total variance, whereas DAPC axes reflect differences among a priori defined groups

We can explore the variation associated with putative sex chromosomes on PC1 in *Limnothrissa* and *Stolothrissa* using DAPC.

First, let's see how a DAPC on "general location" looks to assess geographic variation in *Stolothrissa*. Note that when you run the dapc() function you will need to interactively choose the number of PCs to retain and the number of discriminant functions to retain. The possible number of discriminant functions is constrained by the data (i.e. in this case there can be maximum 4 with 4 locations).

```
### DAPC by general location
stolodapcloc<-dapc(stolovcfgen,pop= Sfullfishinfomatched$General.Location)

# plot results
scatter(stolodapcloc, scree.da=FALSE, bg="white", pch=20, cell=0, cstar=0, col=c('red','blue','green','
```

This looks largely consistent with what we saw on PC2 of the unfiltered vcf file for *Stolothrissa*, woud you agree?

Now let's use DAPC to look into the putative sex-association on PC1. The fullfishinfo file also includes a column for sex ("geneticSex"), which is drawn from the associations on PC1 that you saw previously. Note that since we only have two groups in this case, we will get a result that is bivariate, and not multivariate when we plot it.

```
#DAPC by genetic sex as determined by PC1
stolosexdapc<-dapc(stolovcfgen,pop= as.character(Sfullfishinfomatched$geneticSex))

#plot results
scatter(stolosexdapc, scree.da=FALSE, bg="white", pch=20, cell=0, cstar=0, col=c('red','blue','green','
```

The DAPC function gives us an output called $var.contr that provides the contributions of the original variables (alleles in the case of genetic data) to the principal components of DAPC. We can use this to pull out alleles associated with the putative sex divergence. Is the length of this allele contributions vector the expected size? (look back on questions 1 and 2 from Thursday to check!)

```
#first have a look at the allele contributions vector. How long is it? Is it the expected size?
length(stolosexdapc$var.contr)

#plot allele contributions
loadingplot(stolosexdapc$var.contr,cex.lab=.001)

#hard to see, let's look at a histogram
hist(stolosexdapc$var.contr)
```

Based on the histogram it looks there is a bivariate distrbution of allele contributions with loadings greater than 4e-04 being the "highly associated" alleles. Let's pull out those alleles with high associations to look at some more.

```r
#how many loci are highly associated with "sex"?
length(which(stolosexdapc$var.contr>4e-04))

#save these loci to look at in other ways
stolohighload<-which(stolosexdapc$var.contr>4e-04)
```

We have specific hypotheses about heterozygosity for sex-associated loci, namely that for heterogametic sex determination we expect individuals to be either homozygous for one chromosome, or heterozygous (under the assumption that the sex chromosomes are not so divergent that they do not map to one another). We also have tools for looking at heterozygosity of loci, right in the adegenet "genlight" object we already have made.

```r
## summary of dataset, includes calculation of locus-specific heterozygosity
stolodiv<-summary(stolovcfgen)

# plot heterozygosity at all loci in the dataset
plot(stolodiv$Hobs, xlab='Locus number',ylab='Observed heterozygosity')
```

Plotting heterozygosity of all loci gives a lot of variance (as we'd expect from highly variable allele frequencies across loci). But what if we look at only the putative sex-associated loci? To do this we need to figure out how the loci from the genlight object correspond with the alleles that have high contributions to the sex DAPC. This is a little bit involved because we need to make sure the names are consistently formatted.

```r
## look at heterozygosity of "sex" loci

### here are the loci with the high loadings
row.names(stolosexdapc$var.contr)[stolohighload]

## go from each allele to a single locus name
Sloci1<-unlist(strsplit(row.names(stolosexdapc$var.contr)[stolohighload],split='.',fixed=TRUE))

#pull out odd indices to get locus names
odd <- function(x) x%%2 != 0
take<-odd(1:length(Sloci1))

Ssexlocinames<-Sloci1[take]
take2<-odd(1:length(Ssexlocinames))
Ssexlocinames2<-Ssexlocinames[take2]

#match sex loci names with Heterozygosity locus names
names(stolodiv$Hobs)
Ssexlociintotal<-match(Ssexlocinames2,names(stolodiv$Hobs))
stolodiv$Hobs[Ssexlociintotal]
```

After we've finished that formatting stuff, we can do the fun stuff! Let's look at heterozygosity of the sex loci:

```r
## plot heterozygosity of sex loci only
plot(stolodiv$Hobs[Ssexlociintotal], xlab='Locus number',ylab='Observed heterozygosity', main="Stolothr

#let's look at the two "sexes" individually and see how they look"
StolodivsexAa<-summary(stolovcfgen[Sfullfishinfomatched$geneticSex=="Stolo_Sex_A"])
```

```
StolodivsexBb<-summary(stolovcfgen[Sfullfishinfomatched$geneticSex=="Stolo_Sex_B"])

#plot heterozygosity of individual "sexes"
# sex A
plot(StolodivsexAa$Hobs[Ssexlociintotal], xlab='Locus number',ylab='Observed heterozygosity', main="Sto
#sex B
plot(StolodivsexBb$Hobs[Ssexlociintotal], xlab='Locus number',ylab='Observed heterozygosity', main="Sto
```

QUESTIONS:

1) How does heterozygosity for the “sex” loci differ from heterozygosity for loci across the whole dataset?
2) Which sex appears to be the heterogametic sex?
3) Repeat this for Limnothrissa (this might seem hard, but it isn’t really…just replace the *Limnothrissa* objects for where we have *Stolothrissa* throughout the code you already have. The one change you will need to make is the threshold for determining which loci have high association with “sex”, because these values are specific to the analysis. Look at the histogram and determine a suitable threshold). Does *Limnothrissa* appear to have fewer or more sex-associated loci than *Stolothrissa*? For *Limnothrissa* which sex is heterogametic?
4) Think about our use of PCA and DAPC in these analyses. How are these analyses complementary? What do we gain by using them together?

You are welcome to explore the datasets in any other directions you are curious about. To see where we went from here to turn these data into a publication, you have have a look at the paper we published in Molecular Ecology in 2020.