VCFs, Population Structure, and Filtering

Katie Wagner

2026-01-07

An introduction to exploring VCF files and population structure

By this point in the course, you understand basic VCF file structure and the data it contains. Let's dive into working with a VCF file, and thinking about how data filtering decisions can impact our downstream analyses. As we learned from reading the Hemstrom et al. 2024 paper this morning, "unexpected results from exploratory methods, such as principal component analysis (PCA), can be indicative of experimental or laboratory errors (for example, mislabelling), sequencing bias, sex-linked loci, selection or other phenomena". So, we will use PCA as a first-pass look at some VCF files, and then look at how filtering by locus for Hardy-Weinberg Equilibrium influences the structure that we observe.

We will be working with VCF files from RADseq datasets from two African sardines: *Stolothrissa tanganicae* and *Limnothrissa miodon*. First, let's load up the packages we will need for this work (you may need to install them first if you have not already).

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

#Importing your VCF files

You can find the data we will use at the course Github site. Download the folder for "sardines_tutorial" and navigate your R working directory to the "Data" subfolder.

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

Have a look at these objects you imported. They are of class "vcfR".

QUESTION: 1) How many variants are in these files? How many individuals?

In order to use them for analyses in the package 'adegenet' we will transform them from vcfR to genind objects.

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

Look at these genind objects.

QUESTION: 2) Is the information consistent with what you learned from looking at the vcfR object?

Principal components analysis

Principal components analysis, or PCA, is a common dimensionality reduction method to see how data cluster without assigning groups or number of groups a priori. We can apply this method to genetic as well as other types of data. We'll follow documentation from the adegenet package to run PCA on our data:

```
# calculate allele frequencies and replace NAs
genfreqstolo <- tab(stolovcfgen, freq=TRUE, NA.method="mean")
genfreqlimno <- tab(limnovcfgen, freq=TRUE, NA.method="mean")

# run the PCA for Stolothrissa to start
pca_res_stolo1 <- dudi.pca(genfreqstolo, center=TRUE, scale=FALSE)
```

This will ask you how many principal components you want to retain. The eigenvalues (height of the bars) for each of these compnents roughly corresponds to the amount of genetic variation explained by each principal component. General advice is to retain all components before they drop off sharply. We need to retain at least 2 to make the plots we want.

Make a simple plot for *Stolothrissa*:

```
plot(pca_res_stolo1$li, pch = 19, cex = 1, col = "blue")
```

If you know already how many components you want to retain, you can run this non-interactively by doing (this time let's retain 4 axes):

```
pca_res_stolo2 <- dudi.pca(df = genfreqstolo, center = TRUE, scale = FALSE, scannf = FALSE, nf = 4)
```

Make a plot of our first two axes (note how we can specify axes and and two from the dataframe of results using subsetting with [row,column]).

```
plot(pca_res_stolo2$li[,1:2], pch = 19, cex = 1, col = "blue")
```

Let's have a look at axes 2 and 3:

```
plot(pca_res_stolo2$li[,2:3], pch = 19, cex = 1, col = "blue")
```

QUESTION: 3) What do you see plotting axes 1 and 2? What do you see plotting axes 2 and 3?

Let's see if we see something similar with *Limnothrissa*!

```
pca_res_limno <- dudi.pca(genfreqlimno, center=TRUE, scale=FALSE, scannf = FALSE, nf = 4)

plot(pca_res_limno$li[,1:2], pch = 19, cex = 1, col = "red")

plot(pca_res_limno$li[,2:3], pch = 19, cex = 1, col = "red")
```

QUESTION: 4) What do you see in *Limnothrissa* in the PCAs, and how does this differ from what we have seen in *Stolothrissa*?

Now I'll give you some more information about these fish. Here is a data sheet including all sampling information for all individuals in both datasets, and some code to match and pair data from the info file with the genetic results. Make sure you look carefully at the code and understand how it is working.

```
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,]
```

There are specific sampling sites in this file as well as "General Location" (column 9). Let's color our samples by their general location to see if geography explains the genetic structure we are seeing.

```
#let's assign colors to the general regions
color_map <- c("Kigoma" = "black", "North Mahale" = "red", "South Mahale" = "green",
               "North DRC" = "blue")

#now append a color vector to our matched fish info subset dataframes
Sfullfishinfomatched$colorvec <- color_map[Sfullfishinfomatched$General.Location]
Lfullfishinfomatched$colorvec <- color_map[Lfullfishinfomatched$General.Location]

#now we can use that color vector that is the last column of our matched dataframe
plot(pca_res_stolo2$li[,1:2], pch = 19, cex = 1, col = Sfullfishinfomatched$colorvec)

plot(pca_res_limno$li[,1:2], pch = 19, cex = 1, col = Lfullfishinfomatched$colorvec)
```

QUESTION: 5) Do you think that geography explains the structure you are visualizing in the PCA? Look past PCs 1 and 2 to fully assess this, and discuss what you see.

Now let's filter for Hardy-Weinberg equilibrium and repeat the PCA process. We will go back to our VCF file to do this. Typically we would do this in vcftools, but we can also do it in R for simplicity in this test example.

```
#pull out genotypes and put them into a locus format needed for pegas
gt_stolo <- extract.gt(stolovcf, element = "GT")
gt_limno <- extract.gt(limnovcf, element = "GT")

loci_stolo <- vcfR2loci(stolovcf)
loci_limno <- vcfR2loci(limnovcf)

#we can test for HW equilibrium for each locus using the function hw.test() in the package
#pegas, where the argument B = the number of permutations. This takes a little bit.
hwe_stolo <- hw.test(loci_stolo, B = 1000)
hwe_limno <- hw.test(loci_limno, B = 1000)

# identify the loci that failed the HW test for each species.The p-values are in the 4th
#column of the hwe results dataframes.
pvals_stolo <- hwe_stolo[,4]
failed_stolo <- names(pvals_stolo[pvals_stolo < 0.05])

pvals_limno <- hwe_limno[,4]
failed_limno <- names(pvals_limno[pvals_limno < 0.05])
```

```
#now remove the loci that failed the HWE test from the VCFs. One note about the syntax
#of this code below: The vcfR object is an S4 object with "slot" names @meta, @fix, and
#@gt. We want to filter by locus as named in the HWE tests. The @fix slot in vcrR stands
#for Fixed VCF columns (CHROM, POS, REF, ALT, QUAL, FILTER, INFO). The locus names from
#pegas (failed_limno; failed_stolo) are formatted as CHROM + POS. If we look at the
#information in the @fix slot, you will see that this is split into two columns. So we
#have to create IDs that will match first in order to edit our VCF by pasting the
#CHROM + POS info in vcfR together.

# make a site ID vector corresponding to the order of loci in the @fix slot
site_ids_stolo <- paste(stolovcf@fix[,"CHROM"], stolovcf@fix[,"POS"], sep = "_")
site_ids_limno <- paste(limnovcf@fix[,"CHROM"], limnovcf@fix[,"POS"], sep = "_")

# check the first few to see if the format matches the names in our vectors of
#failed loci
head(site_ids)
head(failed_stolo)

#looks like it should work.

#now remove locus IDs corresponding to those that failed the HWE test

keep_stolo <- !(site_ids_stolo %in% failed_stolo)
vcf_filtered_stolo <- stolovcf[keep_stolo, ]

keep_limno <- !(site_ids_limno %in% failed_limno)
vcf_filtered_limno <- limnovcf[keep_limno, ]
```

Now repeat the PCA analyses on the datasets filtered for HWE, with those loci that failed the test removed.

QUESTION: 6) What do you find? How does filtering for HWE influence the PCA? Is this good? Is this bad? Explore and discuss!