# Diversity and Divergence in Sticklebacks

**James O'Reilly**
Student Number: `r0773125`

## 1 Genetic Diversity

*Analyse the genetic diversity of all stickleback populations, separately and hierarchically, and interpret the data. A question you might ask is: why are they so different?*

### 1.1 Heterozygosity and number of alleles

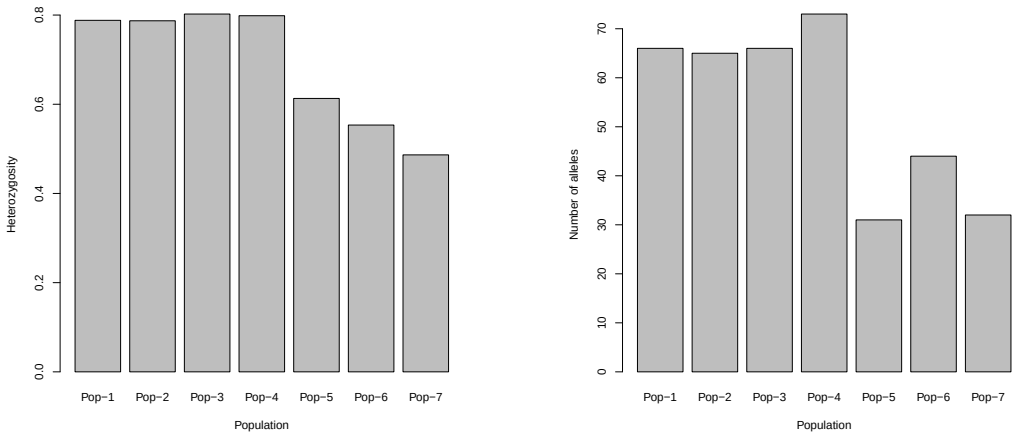
Expected heterozygosity is commonly used when measuring within population genetic diversity. Figure 1a shows the expected heterozygosity for each population. It is clear from the figure that the inland populations (4, 5 and 6) have less heterozygosity. Populations 3 and 4 have slightly higher heterozygosity but this is minimal.

Another measure of population diversity is the number of alleles present in the population (see Figure 1b. Again it is clear that the inland populations have fewer alleles than the coastal populations. It should be noted that this measure of allelic diversity is frequency independent, and we don't account for the different frequencies of alleles present. Naturally, there is a relationship between these two measurements – if there are more alleles in a population, then there going to be more possible heterozygous combinations. Just to illustrate this fact, say a population has $n$ alleles at a given locus, then there are $\binom{n}{x}$ possible ways of choosing two alleles, $n$ of which are homozygous combinations. This leaves $\binom{n}{x} - n$ heterozygous combinations.

Therefore, the ratio of heterozygous combinations to homozygous combinations is given by:

$$\frac{\binom{n}{x} - n}{n} \tag{1}$$

This ratio increases linearly as a function of $n$. Just looking at the maths, we see that as the number of alleles in a given population increases we should expect the heterozygosity to also increase. However, looking at the barplots in Figure 1, we can clearly see that there isn't a one-to-one mapping between the number of alleles and the expected heterozygosity. This is for a number of reasons, including the stochastic process through which alleles are chosen in a given population, differences in allele frequencies, and other macroevolutionary processes which affect how specific alleles are selected.




(a) (b)

Figure 1: (a) Expected heterogeneity for each population (b) Number of alleles for each population

## 1.2 Allelic Richness

Allelic richness is another measure of genetic diversity in populations. Assessing allelic richness in a set of populations requires that variations of sample size be taken into account. We use rarefaction to account for variations in sample size. Rarefaction is the process of subsampling a larger dataset in smaller chunks such that we can estimate diversity among groups using the same number of individuals.

Here we estimate genetic diversity by calculating the mean allelic richness across all loci for each population (see Figure 2). Again, looking at this figure it is clear that the coastal populations are more diverse, when using average allelic richness as our measure of diversity.

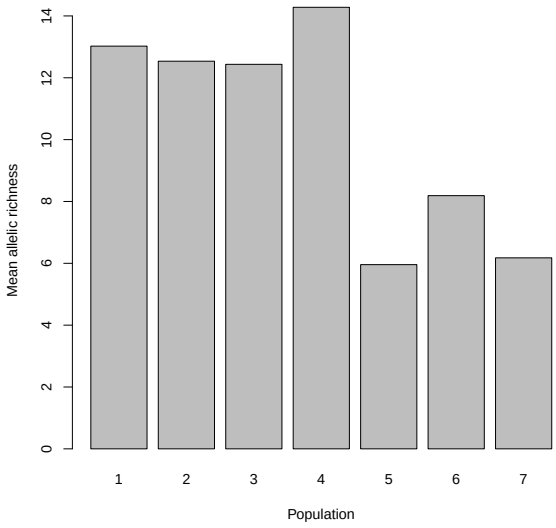



Figure 2: Mean allelic richness across loci for each population

## 1.3 What is the explanation for this difference?

The primary reason for looking at diversity is to perform some comparison, which provides some insights into the biological and/or demographic processes influencing your data. In our case, there is a clear geographical split between inland populations and coastal populations. There are a number of reasons that could explain why the inland populations are less diverse.

Firstly, the inland populations have fewer opportunities to outbreed, while the coastal populations might have a more diverse range of mates, resulting in an increased number of alleles in those populations. The inland populations are more likely to inbreed, leading to a smaller increase in the number of alleles present and an increased homozygosity.

Another possible explanation is that there is more selective pressure for the coastal populations. At a base level, genetic diversity is the fundamental component upon which evolution operates. Without diversity, there is no evolution and as such species cannot respond to selective pressure. Perhaps due to a more varied environment, or an increased number of predators in coastal regions, these populations are forced to adapt more quickly and therefore have increased diversity.

# 2 Genetic Divergence/Differentiation

## 2.1 Determining Genetic Structure

Genetic structure refers to any pattern in the genetic makeup of individuals within a population. Structure in populations influences the distribution (spatial, ecological, and temporal) of alleles. In order to investigate this structure, we need to be able to compare populations. Comparing populations naturally requires some quantification of *'different'*, and therefore some well-defined distance metric. Here we will use the Fixation Index ($F_{ST}$). Sewall Wright developed the Fixation Index as a way of measuring genetic differences between populations. This can be thought of as the fraction of total diversity that is not a consequence of the average diversity within subpopulations. There are different ways to calculate the Fixation Index, with different underlying distance metrics. Below we calculate the pairwise genetic differentiation between sample populations with a variant of ($F_{ST}$) which is based on the definition of genetic distance given in Nei's landmark text 'Molecular Evolutionary Genetics' written in 1987.[1]

Table 1 gives the pairwise $F_{ST}$ for the sample populations.

| | Pop-1 | Pop-2 | Pop-3 | Pop-4 | Pop-5 | Pop-6 | Pop-7 |
|---|---|---|---|---|---|---|---|
| Pop-1 | NA | – | – | – | – | – | – |
| Pop-2 | 0.0177 | NA | – | – | – | – | – |
| Pop-3 | 0.0272 | 0.0374 | NA | – | – | – | – |
| Pop-4 | 0.0638 | 0.0751 | 0.0709 | NA | – | – | – |
| Pop-5 | 0.1155 | 0.1231 | 0.1486 | 0.1745 | NA | – | – |
| Pop-6 | 0.1077 | 0.1229 | 0.1650 | 0.1877 | 0.1977 | NA | – |
| Pop-7 | 0.1800 | 0.1715 | 0.2226 | 0.2480 | 0.2889 | 0.2774 | NA |

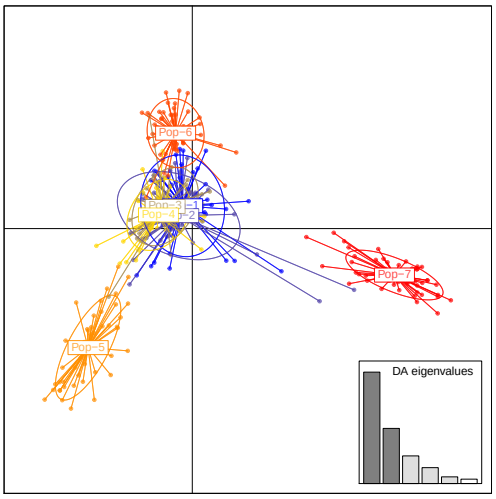
Table 1: Estimated pairwise $F_{ST}$ between the populations, calculated according to Nei (1987)

We can clearly see from Table 1 that there is most divergence between the inland populations. The pairwise divergence between coastal and inland populations is also notably higher than the coastal populations. How do explain this? I believe this structure can be explained using the geography of the region. The inland populations have very little opportunity to interact, as they are in separate river systems which flow downstream toward the coast. They are therefore the most divergent due these barriers to flow. The inland populations have more chances of interacting with the coastal populations as the rivers flow downstream, which explains the decreased divergence between coastal and inland populations. Lastly, the coastal populations are the least divergent as they share many aspects of their environment

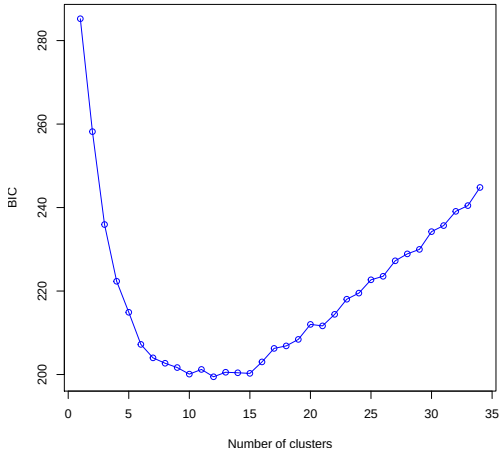
## 2.2 Discriminant Analysis of Principal Components

We can assess the variability between populations using a Discriminant Analysis of Principal Components (DAPC). DAPC optimises the variability between groups while minimising the variability within groups. DAPC is different to PCA in that it preserves local structure, allowing us to investigate variance between populations. Visualising the DAPC analysis in Figure 3a, we see that the inland populations have the most variability between each other. The inland populations have some variability with the coastal populations, and lastly the coastal populations show the least variability. This DPCA highlights the same genetic structure which the pairwise $F_{ST}$ revealed.

A clustering approach was used to determine the ideal number of clusters for this analysis. To identify the optimal number of clusters, $k$-means is run sequentially with increasing values of $k$, and different models are compared using Bayesian Information Criterion (BIC). BIC calculates the trade-off between model complexity and model performance. Figure 3b shows that there




(a) Visualising DAPC with 40 PCs and 5 discriminant functions




(b) Visualising BIC against number of clusters

is diminishing returns after $k > 7$. While there is some improvement in BIC score for $k > 7$, using more than seven clusters would make inference more tricky, as we have seven sample populations.

We can then visualise the number of individuals from a given population which were assigned to a certain cluster. Looking at Figure 4, we see that population 5, and 7 are assigned almost entirely to a single cluster. Population 6 is divided between two clusters, one of which also has instances from population 1. This fits with the data for pairwise $F_{ST}$ given in 1, which shows population 1 and 6 are not very divergent. In general, the inland populations are assigned to a small number of clusters, which fits with our previous findings that they are divergent from both each other and the coastal populations. The coastal populations in Figure 4 share assignments across a number of clusters, in particular cluster 6 and 7. Once more, this fits with our knowledge about the divergence between coastal populations. Overall, the genetic structure is consistent between these analyses.

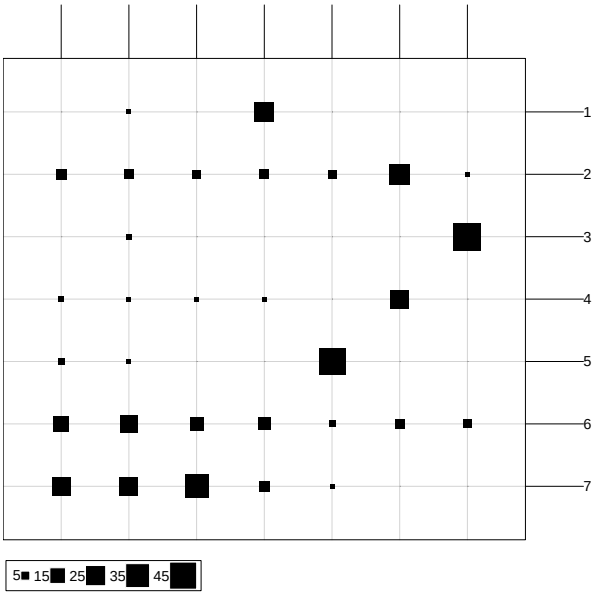



Figure 4: Assignment of individuals from sample populations to clusters

# References

[1] Nei, M. *Molecular evolutionary genetics*. Columbia university press, 1987.

[2] Raeymaekers, J., Maes, G., Audenaert, E., and Volckaert, F. Detecting holocene divergence in the anadromous–freshwater three-spined stickleback (gasterosteus aculeatus) system. *Molecular ecology 14*, 4 (2005), 1001–1014.