Inferring Population Structure in R

Kevin Keenan

22 April 2015

Introduction

Population structuring, the non-random distribution of genetic variation among individuals of a species, can be notoriously difficult to detect and characterise. In this practical we will attempt to characterise the genetic structuring among 444 brown trout, sampled from 25 seperate sites along three tributaries of a small river system. We will start with a standard PCA approach, making use only of genetic data, eventually building to sPCA and MEMGENE analysis, where we will use spatial data to refine any genetic structuring. Below is a map of our study system.

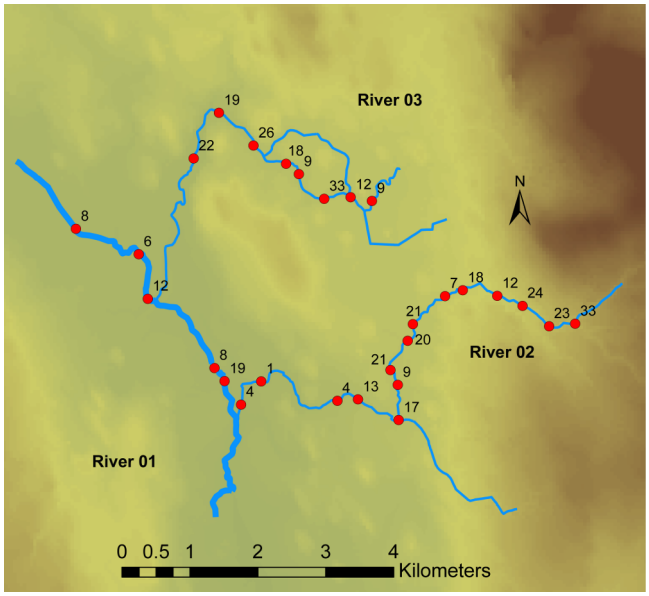


Figure 1: A map of our brown trout study system. Red circles show sampling locations, while adjescent numbers show sample sizes per site

You will notice that some sample sites have very small numbers, and overall our sampling is relatively high resolution. This is so because the brown trout, being a salmonid fish exhibits strong homing behaviour, meaning that population structure can arise over very small spatial scales. Therefore, to fully characterise the total genetic diversity and structuring of trout in this system, psudo-continuous sampling was employed.

This sampling design, while very good for understanding the total diversity within a given area, also presents a difficult analytical challenge. Clearly each site is not expected to represent a population sample, and in any case, some sample sizes per site would not be representative. As such, it would be meaningless to attempt to characterise populations based on the samples here (e.g. how relevent is allelic richness calculated per site sample?). The correct approach to analysing such data is to infer populations, and then characterise those populations (e.g. HWE, differentiation etc. . .).

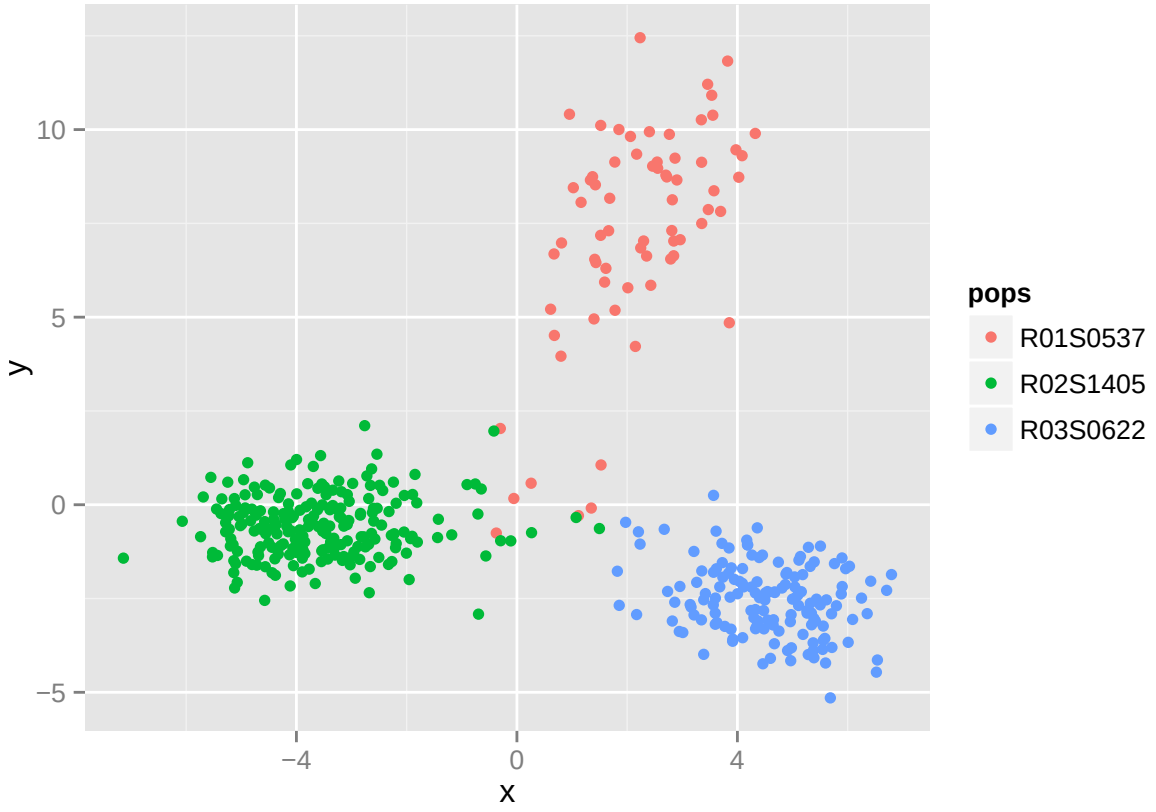
Practical 4: K-means clustering and DAPC

Introduction

As we have seen in the lecture, K-means is a complimentary method that makes use of our *dimentionally reduced* allele frequencies to cluster individuals into discrete groups. In the lecture we have seen how K-means can be interactively applied to the `genind` object generated by `read.genepop`. Following the assignment of individuals to discrete groups, we can use the **Discriminant Analysis of Principal Components** to maximise the between group differences to gain an understanding of their relative relationships. In this practical, using the *trout_riv.gen* data set again, we will go through the stages of clustering individuals using `find.clusters` and then take these clusters through to DAPC analysis, where we will attempt to validate our model of population structure.

1. Running K-means on our trout data
    a. Read *trout_riv.gen* and store it in the variable `trout_riv`.
    b. Scale the allele frequencies and replace missing data with allele means. Store these results in `scaled_riv`.
    c. Run `find.clusters` on the scaled data, coosing the appropriate number of PCs and then choosing the "best"" k value. Test for $k = 1 - 20$. Make sure the results are stored in the variable `riv_grps`.
    d. Can you think of a way to reproduce these two plots?

Original river locations



K-means cluster asignment

e. Using the `table.value` function does the accuracy of k-means groupings relative to an individuals river of origin look any clearer?

2. Running DAPC on our trout data

   a. Run `dapc` on your scaled data, using the k-means defined groupings. Choose a relatively high number of PCs (e.g. ~100) to retain initially and retain at least two discriminant functions. Save these results to the variable `temp_dapc`.
   b. Plot your dapc results.
   c. Now we need to get a better idea of how many PCs to retain. Run `optim.a.score` on `temp_dapc`.
   d. Using the recommended number of PCs run a validation test on our model. Retain 20 individuals from each inferred cluster as a supplementary data set.
   e. Plot the model performance as a scatter plot.
   f. Plot the model performance as a contingency plot.

Back to the lecture for now.