Molecular Plant Breeding
Roland Kölliker
roland.koelliker@usys.ethz.ch

Excercise - Analysing genetic diversity in outbreeding plant species

Roland Kölliker

16.04.24

Five populations of *Centaurea jacea* have been sampled from five countries (Switzerland; CH, Hungary; HU, Italy; IT, Norway; NO and Slovenia; SL). Populations consisted of 19 individual plants each and were analysed using amplified fragment length polymorphism (AFLP) markers. The file `"centData.txt"` contains the data of 268 markers for the 95 individual plants. Use multivariate analyses such as cluster analysis, principle components analysis (PCA) and analysis of molecular variance (AMOVA). You profit the most if you try to find the solutions yourself. However, don't hesitate to ask for assistance during the lecture. Also, a possible solution is given in `"https://n.ethz.ch/~rolandko/download/cent_fancy.R"`.

1) Load the datafile (`"centData.txt"`) and inspect the dataframe.
   `d.cent <-read.table("https://n.ethz.ch/~rolandko/download/centData.txt")`
   (make sure the population variable is a factor: `d.cent$pop <-as.factor(d.cent$pop)`)

2) Calculate pairwise genetic diversity among individual plants based on Euclidean distance (use `t.euc <-dist(d.cent[,-1],method="euclidean")`. Calculate mean, min, max values.
   - Optional: Repeat calculations for each population separately. Which population is characterized by the lowest, which by the highest average Euclidean distance?

3) Visualize the relationships among all 95 individual plants using cluster analysis (`hclust`). Are there indications for genetic structure in the dataset? Hint: `plot(hclust(...))`

4) Complex structures are sometimes easier to visualize using scatterplots.
   - Apply principle component analysis to the dataset using `r.pca <-prcomp(d.cent[,-1])`.
   - Inspect the results using the summary and the str function.
   - Plot the results using `biplot(r.pca)` and
     `plot(r.pca$x[,c(1,2)], main="PCA Centaurea",col=as.numeric(d.cent$pop), pch=16)`
   - Plot also PCA1 vs PCA3 and PCA2 vs PCA3
   - What can you say about the structure within and among populations?

5) Analyse the structure of genetic variation using the `adonis2` function of the `vegan` package:
   ```
   library(vegan)
   adonis2(d.cent[,-1]~d.cent$pop, method="euc", permutations=1000)
   ```

Molecular Plant Breeding
Roland Kölliker
roland.koelliker@usys.ethz.ch

6) In order to visualize relationships among populations, marker frequencies per population can be used. The file `"centFreq.txt"` contains marker frequencies for the 268 markers per population.

   - Inspect the dataset and calculate cluster analysis as in 2) and 3)
     `d.freq <-read.table("https://n.ethz.ch/~rolandko/download/centFreq.txt")`
   - Optional: Use `pvclust` to calculate bootstrap values for clusters:
     `library(pvclust)`
     `plot(pvclust(t(d.freq), nboot=100))`
     What can you say about the robustness of the clustering?
     (more information on `pvclust` can be found here
     `"https://n.ethz.ch/~rolandko/download/paperPvclust.pdf"`
     or here `"https://github.com/shimo-lab/pvclust"`)

7) The file `"centGeo.txt"` contains the geographical coordinates of the sampling sites.

   - Use cluster analysis to visualize the geographic relationships and compare it to the cluster analysis based on marker data:
     `d.geo <-read.table("https://n.ethz.ch/~rolandko/download/centGeo.txt")`
   - Calculate the correlation between the distance matrix based on AFLP data and the distance matrix based on geographical distances. Hint: `mantel(x,y)`

8) Compare your results with the solution on the moodle platform (available from 11:00h).

9) Download and complete the graded test from the moodle platform.