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

Solution - Analysing genetic diversity in outbreeding plant species

Roland Kölliker

16.04.2024

1) Load the datafile, convert the population variable to a factor and inspect the dataframe.

```
d.cent <- read.table("https://n.ethz.ch/~rolandko/download/centData.txt")
d.cent$pop <- as.factor(d.cent$pop)
str(d.cent)
```

```
'data.frame': 95 obs. of  269 variables:
 $ pop          : Factor w/ 5 levels "CH","HU","IT",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ E_AAC_M_CTA02: int  0 0 1 0 0 0 0 0 0 0 ...
 $ E_AAC_M_CTA03: int  1 1 1 1 1 1 1 1 1 1 ...
 $ E_AAC_M_CTA04: int  0 0 0 1 0 1 0 0 0 0 ...
 $ E_AAC_M_CTA07: int  0 0 0 1 0 0 1 1 0 0 ...
 $ E_AAC_M_CTA09: int  1 1 1 1 0 1 1 1 1 1 ...
 $ E_AAC_M_CTA14: int  1 1 0 1 1 1 1 1 1 0 ...
```

```
#display some samples and the first five columns
d.cent[c(1,20,39,58,77),1:5]
```

```
       pop E_AAC_M_CTA02 E_AAC_M_CTA03 E_AAC_M_CTA04 E_AAC_M_CTA07
CH1488  CH             0             1             0             0
IT1522  IT             0             1             0             0
NO1552  NO             0             1             1             1
SL1575  SL             0             1             0             1
HU1605  HU             1             1             0             1
```

2) Calculate pairwise genetic diversity among individual plants based on Euclidean distance Calculate mean, min, max values.

```
t.euc<-dist(d.cent[,-1], method="euclidean")
summary(t.euc)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
5.48    8.83    9.27    9.15    9.64   11.09
```

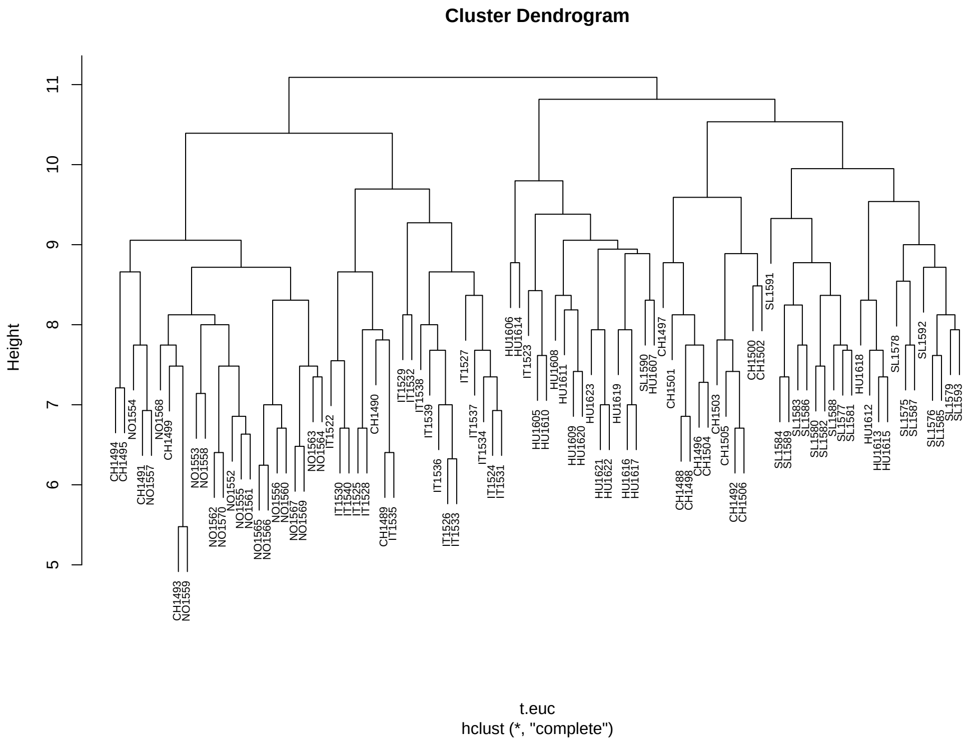
- Optional: Repeat calculations for each population separately. Which population is characterized by the lowest, which by the highest average Euclidean distance?

```
t.pop<-levels(d.cent$pop)
for(i in 1:5){
    t.eucpp<-dist(d.cent[d.cent$pop==t.pop[i],c(-1)], method="euclidean")
    print(t.pop[i])
    print(summary(t.eucpp))
}
```

```
[1] "CH"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.48    8.25    8.94    8.80    9.38   10.77
[1] "HU"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00    8.06    8.37    8.37    8.72    9.80
[1] "IT"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.33    7.94    8.48    8.38    8.92    9.70
[1] "NO"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.25    7.25    7.62    7.59    7.94    8.72
[1] "SL"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.35    8.19    8.66    8.60    8.94    9.64
```

3) Visualize the relationships among all 95 individual plants using cluster analysis. Are there indications for genetic structure in the dataset?

```
plot(hclust(t.euc), cex=0.8)
```



4) Complex structures are sometimes easier to visualize using scatterplots.

- Apply principle component analysis to the dataset, inspect and plot the results

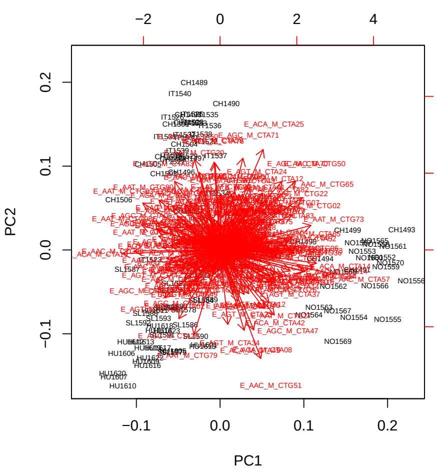
```
r.pca<-prcomp(d.cent[,c(-1)])
summary(r.pca)
```

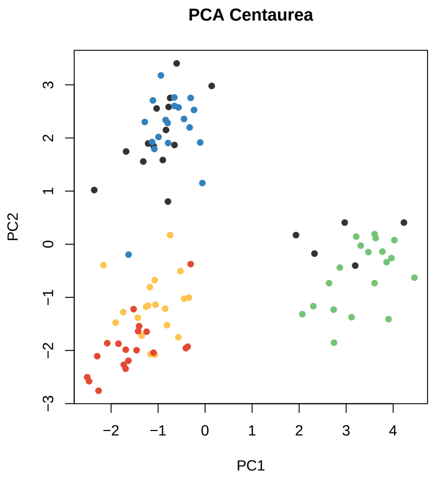
```
Importance of components:
                          PC1    PC2   PC3    PC4    PC5    PC6    PC7
Standard deviation     1.9999 1.7471 1.466 1.2581 1.1031 1.0720 0.9959
Proportion of Variance 0.0949 0.0724 0.051 0.0376 0.0289 0.0273 0.0235
Cumulative Proportion  0.0949 0.1673 0.218 0.2559 0.2847 0.3120 0.3355
```

```
str(r.pca)
```

```
List of 5
 $ sdev     : num [1:95] 2 1.75 1.47 1.26 1.1 ...
 $ rotation: num [1:268, 1:95] -0.04219 -0.00855 0.09449 0.11732 0.00563 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:268] "E_AAC_M_CTA02" "E_AAC_M_CTA03" "E_AAC_M_CTA04" ...
  .. ..$ : chr [1:95] "PC1" "PC2" "PC3" "PC4" ...
 $ center  : Named num [1:268] 0.147 0.989 0.126 0.242 0.979 ...
  ..- attr(*, "names")= chr [1:268] "E_AAC_M_CTA02" "E_AAC_M_CTA03" ...
 $ scale   : logi FALSE
 $ x       : num [1:95, 1:95] -1.094 -0.602 0.144 3.193 -0.773 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:95] "CH1488" "CH1489" "CH1490" "CH1491" ...
  .. ..$ : chr [1:95] "PC1" "PC2" "PC3" "PC4" ...
 - attr(*, "class")= chr "prcomp"
```

```
biplot(r.pca, cex=0.5)
plot(r.pca$x[,c(1,2)], main="PCA Centaurea",col=as.numeric(d.cent$pop), pch=16)
```



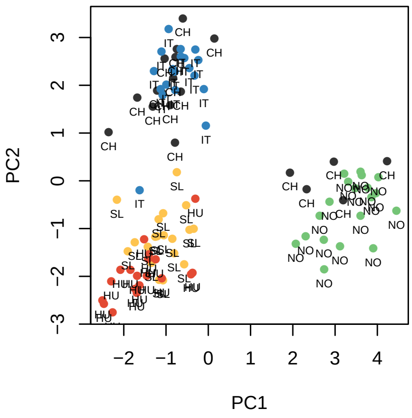


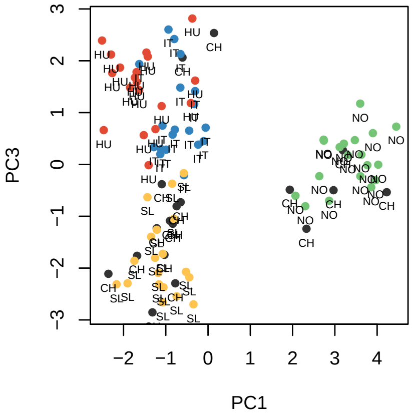
- Plot also PCA1 vs PCA3 and PCA1 vs PCA2

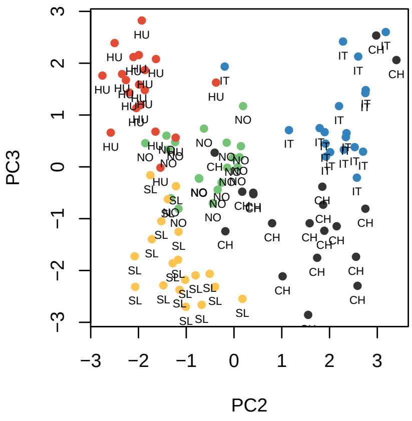
```
plot(r.pca$x[,c(1,2)], col=as.numeric(d.cent$pop), pch=16)
text(r.pca$x[,c(1,2)], labels=d.cent$pop,cex=0.6, pos=1)

plot(r.pca$x[,c(1,3)], col=as.numeric(d.cent$pop), pch=16)
text(r.pca$x[,c(1,3)], labels=d.cent$pop,cex=0.6, pos=1)

plot(r.pca$x[,c(2,3)], col=as.numeric(d.cent$pop), pch=16)
text(r.pca$x[,c(2,3)], labels=d.cent$pop,cex=0.6, pos=1)
```







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)
```

```
           Df SumsOfSqs MeanSqs F.Model    R2 Pr(>F)    
d.cent$pop  4       801   200.3     5.7 0.202  0.001 ***
Residuals  90      3161    35.1         0.798           
Total      94      3962                 1.000           
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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)

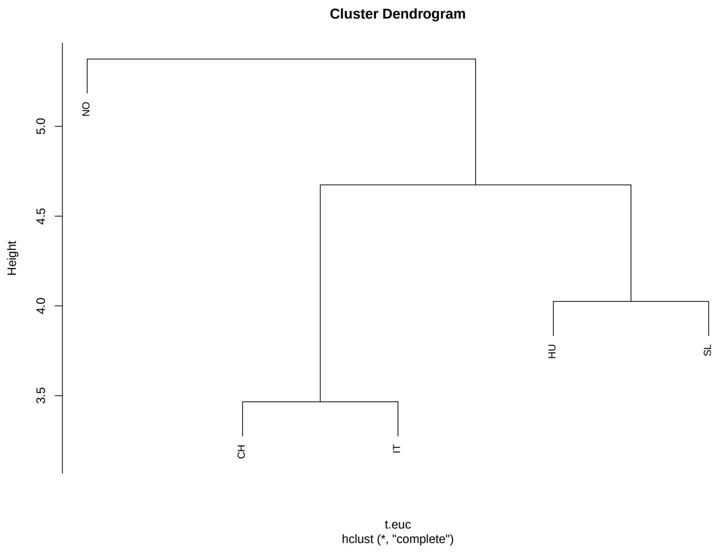
```
d.freq <- read.table("https://n.ethz.ch/~rolandko/download/centFreq.txt")
str(d.freq)
```

```
’data.frame’: 5 obs. of  268 variables:
 $ E_AAC_M_CTA02: num  0.105 0.474 0 0 0.158
 $ E_AAC_M_CTA03: num  1 1 1 0.947 1
 $ E_AAC_M_CTA04: num  0.1579 0.0526 0 0.4211 0
 $ E_AAC_M_CTA07: num  0.2105 0.1579 0.1053 0.6842 0.0526
 $ E_AAC_M_CTA09: num  0.947 1 1 1 0.947
 $ E_AAC_M_CTA14: num  0.789 0.474 0.158 0.947 0.737
```

```
#display the first four markers
d.freq[,1:4]
```

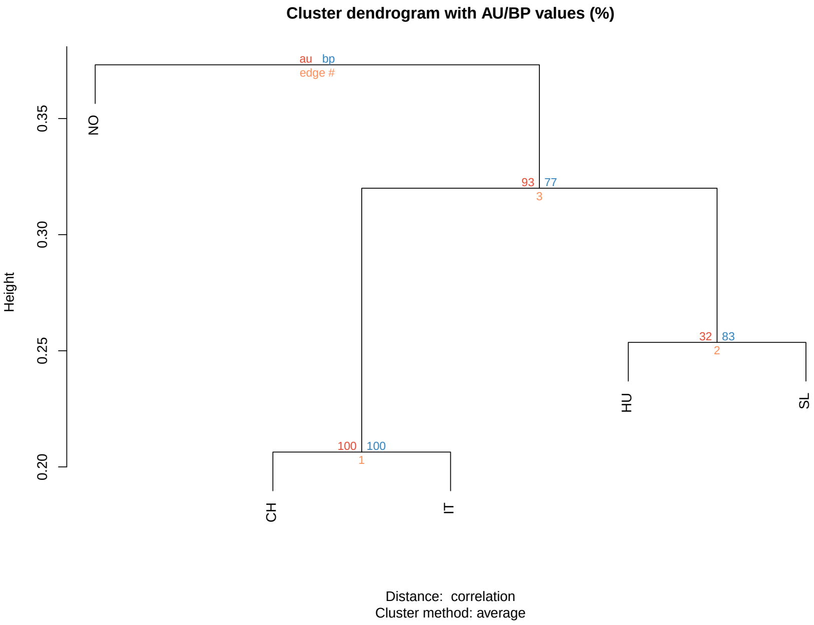
```
   E_AAC_M_CTA02 E_AAC_M_CTA03 E_AAC_M_CTA04 E_AAC_M_CTA07
CH        0.1053        1.0000       0.15789       0.21053
HU        0.4737        1.0000       0.05263       0.15789
IT        0.0000        1.0000       0.00000       0.10526
NO        0.0000        0.9474       0.42105       0.68421
SL        0.1579        1.0000       0.00000       0.05263
```

```
t.euc<-dist(d.freq, method="euclidean")
plot(hclust(t.euc), cex=0.8)
```



- Optional: Calculate bootstrap values for clusters. What can you say about the robustness of the clustering?

```
library(pvclust)
plot(pvclust(t(d.freq), nboot=10))
```

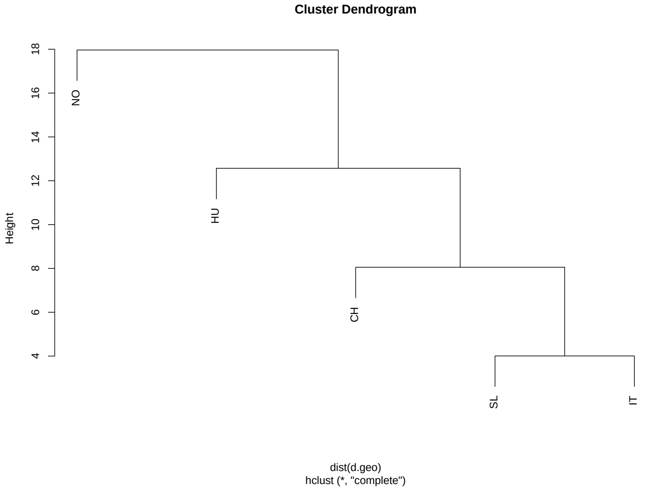


7) The file "centGeo.txt" contains the geographical coordinates of the sampling sites.

- 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")
d.geo
plot(hclust(dist(d.geo)))
```

```
       N     E
SL 45.49 14.49
IT 43.19 11.21
HU 47.43 19.14
NO 61.12 10.09
CH 47.00  6.58
```



- Calculate the correlation between the distance matrix based on AFLP data and the distance matrix based on geographical distances.

```
library(vegan)
mantel(dist(d.freq), dist(d.geo))
```

```
Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = dist(d.freq), ydis = dist(d.geo))

Mantel statistic r: 0.728
      Significance: 0.033

Upper quantiles of permutations (null model):
  90%   95% 97.5%   99%
0.586 0.672 0.719 0.773
Permutation: free
Number of permutations: 119
```

8) `"https://n.ethz.ch/~rolandko/download/cent_fancy.R"` contains some additional R code. Repeat the analysis using this file.