R packages for this computer lab

Module: Plant Genetic Resources (3502-470)

Karl Schmid, Katharina Böndel, University of Hohenheim

23 May 2025

Table of contents

For this computer lab the following R packages are required:

- `SNPRelate`
- `LEA`

Both packages need to be installed slightly differently to the standard `install.packages()`:

```
# install.packages("BiocManager") # run only once
# BiocManager::install("SNPRelate")
# BiocManager::install("LEA")
```

```
library("SNPRelate")
library("LEA")
```

Make sure that they are installed and loaded.

The data set

We continue to use our data set with teosinte individuals, maize landraces and maize improved varieties. For this computer lab we need the original vcf file `zea.vcf`.

Principal component analysis (PCA)

We use the R package `SNPRelate` to calculate the PCA. First, we need to read in the vcf, and then tranform it into a `genofile`:

```
vcf <- c("../data/zea.vcf") # read vcf, adapt the path to the vcf file.
snpgdsVCF2GDS(vcf, "zea.gds", method="biallelic.only") # create .gds file
snpgdsSummary("zea.gds")
genofile <- snpgdsOpen("zea.gds")
```

The screen output will give you some information about the data set.

Next, we calculate the PCA:

```
pca <- snpgdsPCA(genofile, autosome.only=FALSE)
```

We can also calculate the percentage of variation and display how much variation is accounted for by the principal components:

```
pc.percent <- pca$varprop*100
percentages <- round(pc.percent, 2)
percentages
```

We than save the first three principal components in a table. We also add the accession information to the table:

```
# vector with accession information
accession <- c(rep("landrace", 23),rep("teosinte", 19),rep("landrace",2),rep("variety",30))

table <- data.frame(sample.id = pca$sample.id, # make data frame
                    group = accession,         # column with accession information
                    PC1 = pca$eigenvect[,1],   # 1st principal component
                    PC2 = pca$eigenvect[,2],   # 2nd principal component
                    PC3 = pca$eigenvect[,3],   # 3rd principal component
                    stringsAsFactors = FALSE)
```

Finally, we plot the first two principal components:

```
plot(table$PC1,table$PC2,xlab="PC 1 (7.74 %)",ylab="PC 2 (5.60 %)",
     xlim=c(-0.6,0.4),
     ylim=c(-0.6,0.4),
     col="white", # this way we basically plot an "empty" graph to which we can then
                  # add the groups in different colours
     pch=1,las=1)
points(table$PC1[table$group=="teosinte"],table$PC2[table$group=="teosinte"],
       cex=1,pch=0,col="blue")
points(table$PC1[table$group=="landrace"],table$PC2[table$group=="landrace"],
       cex=1,pch=2,col="red")
points(table$PC1[table$group=="variety"],table$PC2[table$group=="variety"],
       cex=1,pch=5,col="orange")
legend("topright",legend=c("teosinte","landraces","varieties"),pch=c(0,2,5),
       col=c("blue","red","orange"),bty="n")
```

We can also plot the second and third component:

```
plot(table$PC2,table$PC3,xlab="PC 2 (5.60 %)",ylab="PC 3 (5.15 %)",
     xlim=c(-0.7,0.6),
     ylim=c(-0.7,0.6),
     col="white",
     pch=1,las=1)
points(table$PC2[table$group=="teosinte"],table$PC3[table$group=="teosinte"],
       cex=1,pch=0,col="blue")
points(table$PC2[table$group=="landrace"],table$PC3[table$group=="landrace"],
       cex=1,pch=2,col="red")
points(table$PC2[table$group=="variety"],table$PC3[table$group=="variety"],
       cex=1,pch=5,col="orange")
legend("topright",legend=c("teosinte","landraces","varieties"),pch=c(0,2,5),
       col=c("blue","red","orange"),bty="n")
```

> **Exercise**
>
> Can you identify clusters in the PCA? Does each group form a cluster? If not, how could the observed pattern be explained? Discuss!

Model based population structure analysis

Here, we use the R package LEA. First, we estimate individual admixture coefficients from a genotypic matrix that can be directly inferred from the vcf file. We assume up to ten different ancestral populations ($K = 10$) and do eight repetitions for each. We also estimate an entropy criterion that evaluates the quality of fit of the statistical model to the data using a cross-validation technique.

> **Warning**
>
> This step may run for quite some time.

```
ancestry_coeff<-snmf("../data/zea.vcf", K=1:10, CPU=12, entropy=TRUE, repetitions=8, project="new")
```

Next, we use the entropy criterion to choose the number of ancestral populations that explains the genotypic data best by plotting the entropy value against the different Ks:

```
plot(ancestry_coeff, col = "blue", pch = 19, cex = 1.2, las=2)
```

The K with the lowest entropy value is chosen as the best fitting one. If the pattern is not clear, choosing the “knee” is generally a good approach.

We can now plot the ancestry matrix:

> **Note**
>
> Please note that this code assumes that $K = 7$ best explains the data. Make sure to adjust the code if you obtain a different K.

```
best = which.min(cross.entropy(ancestry_coeff, K = 7)) # change the K if necessary
my.colors <- c("tomato","lightblue","olivedrab","gold","purple","navy","cyan4")
     # add more colours if you have K>7 (eg blue, brown, deeppink)
barchart(ancestry_coeff, K = 7, run = best, # change the K if necessary
         border = NA, space = 0, sort.by.Q = FALSE,
         col = my.colors,las=2,
         xlab = "Individuals",
         ylab = "Ancestry proportions",
         main = "Ancestry matrix") -> bp
axis(1, at = 1:length(bp$order),
     labels = accession, las=3, # accession was defined in the PCA part
                                # - make sure that it still exists in your environment
     cex.axis = .5)
```

Exercise

Discuss the outcome of this analysis.
Try to put the outcome also into context with the results of the phylogenetic reconstruction from the phylogenetics computer lab and with the results of the PCA.
What can you conclude for this data set of teosinte, maize landraces and varieties?