PCA using in-house code

# Required Software

* Repository (login required): <https://github.com/Metcela-Code/PCA>;  
  Download the master branch and unzip files in the folder used for analysis;
* R: current version 4.0.2 (2020-06-22) -- "Taking Off Again;"  
  Additional library: ggplot2 version 3.3.2;
* RStudio: current version 1.3.1073 "Giant Goldenrod."

# Basic principle

The PCA is executed using the R function "prcomp()." The documentation is available at <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp>.

Graphs are generated with the package ggplot2 using the output of "prcomp()." The documentation is available at <https://www.rdocumentation.org/packages/ggplot2/versions/3.3.2>.

The aim of each step is described as comments within the code.

# Loading of data

* Definition of the working directory  
  L5: define the path to the directory containing the data file.  
  Example: setwd("/home/informatics/Documents/Genomics")
* Definition of the data file  
  L8: define the file to be analyzed.  
  Example: data.matrix <- read.table("my\_analysis.txt", header = TRUE)  
  Attention: the data file is a tab-separated values text file. Data must be structured as follows. The first column doesn't have a header.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sample 1 | Sample 2 | Sample 3 | Sample 4 |
| Feature 1 | Value | Value | Value | Value |

* Scree plot  
  L27: define the title of the scree plot.  
  Example: ggtitle("Scree plot for my\_analysis")
* Explicit names of samples for the PCA plot  
  L35: define the names to be used in the legend of the PCA plot.  
  Example: pca.data["CellType"] <- c("Treatment 1", "Treatment 2", "Treatment 3", "Treatment 4")
* PCA plot  
  L46: define the title of the PCA plot.  
  Example: ggtitle("PCA plot for my\_analysis")
* Ranking of genes contributing most to PC1  
  L52: define the range of genes to be appended in the output ranking file.  
  Example (top 100): gene\_ranking <- names(gene\_score\_ranked[1:100])

# Saving of results

The code creates 3 files: a scree plot, a PCA plot, and a gene ranking list. Saving uses the standard export tools of RStudio.