An Interdisciplinary Summer School on Mining of Biological Data for Master and PhD students,

Norwegian University of Life Sciences, As, Norway.

Tutorial for module 8: “Clustering”,
rooms U223 and U226, 15:30-16:45, Monday, 13 of August, 2018.

Mentor: Petr V. Nazarov, Ph.D., Researcher of Proteome and Genome Research Unit, Department of Oncology, Luxembourg Institute of Health, Strassen, Luxembourg.

Online materials: http://edu.sablab.net/nmbu2018/

Data for tutorial: http://edu.sablab.net/data/txt/mRNAIFNg.txt
It contains annotated genes in rows and samples in columns, values are log2 transformed expressions.

NOTE: before starting, you may go through the code used for the lecture:
http://edu.sablab.net/nmbu2018/lecture.html

NOTE: depending on the task you can consider samples as objects (gene expression are features then) or genes as objects (expression in samples are features).

Tasks

1. **Install required packages** (follow the online materials)
2. **Import the data** (follow the online materials)
   a. Prepare **matrix X** with gene expressions removing lowly expressed and non-annotated features (GeneSymbol is “”)
   b. Create standardized gene expression **matrix Z**, so that all genes have mean = 0, st.dev. = 1
   c. Perform and plot PCA of samples and genes (use both X and Z)
3. **Cluster the samples** (expected outputs are presented in online materials)
   a. Use pheatmap() to make bi-cluster of genes and samples (for X and Z)
   b. Cluster the samples using k-means or PAM and define the reasonable number of clusters. Visualize in PCA plot. Any difference when using X or Z matrices?
4. **Cluster the genes** (expected outputs are presented in online materials)
   a. Use k-means or PAM methods to cluster the genes from standardized Z matrix. Visualize them on corresponding PCA plot (genes as objects, samples as features).