

Unsupervised Classification

Máster en Bioinformática y Biología Computacional

Francisco García García, fgarcia@cipf.es





Outline

- 1. Introduction**
- 2. Clustering methods**
- 3. Distance parameters**
- 4. Exercises on Babelomics**

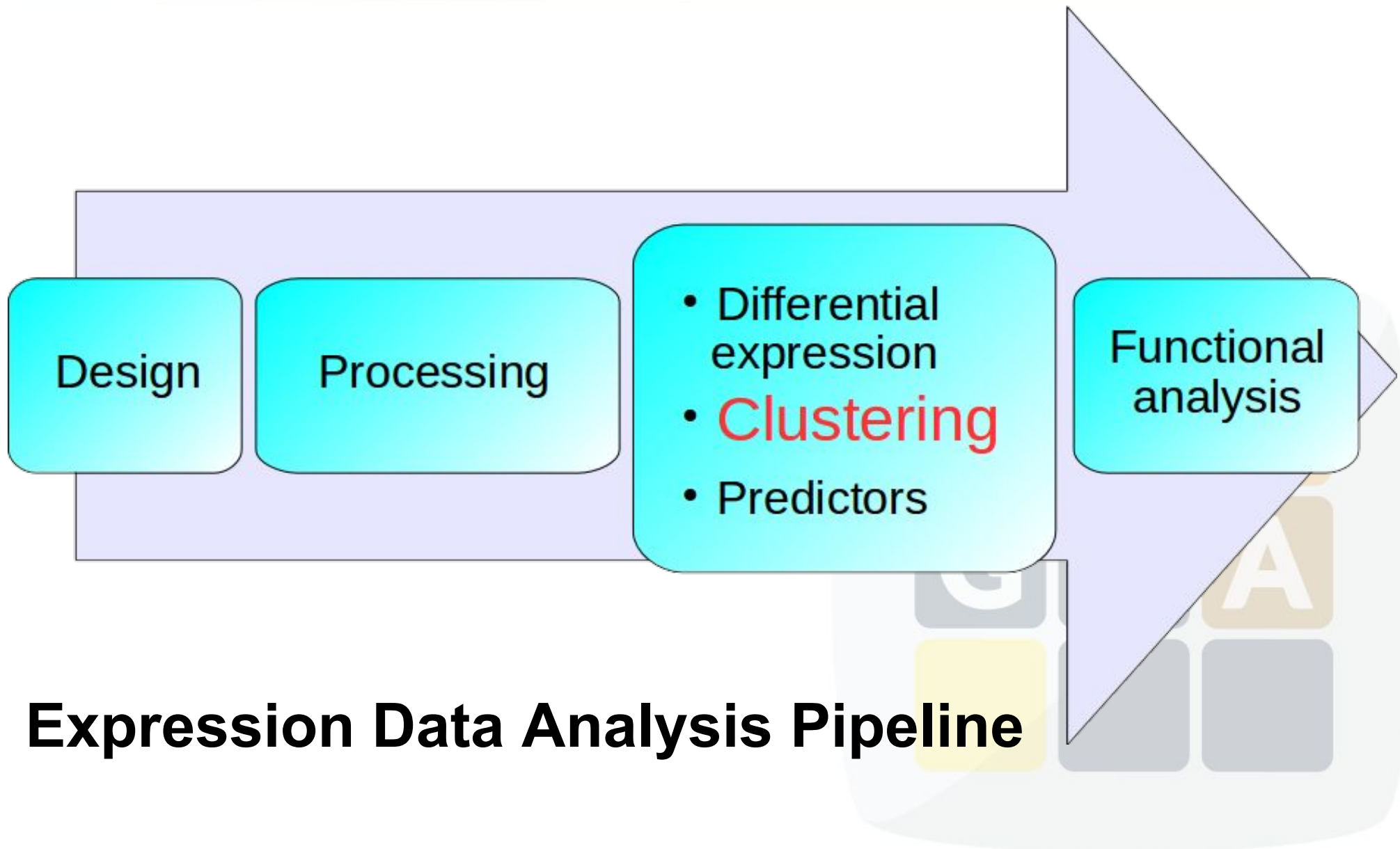


Outline

- 1. Introduction**
2. Clustering methods
3. Distance parameters
4. Exercises on Babelomics

1

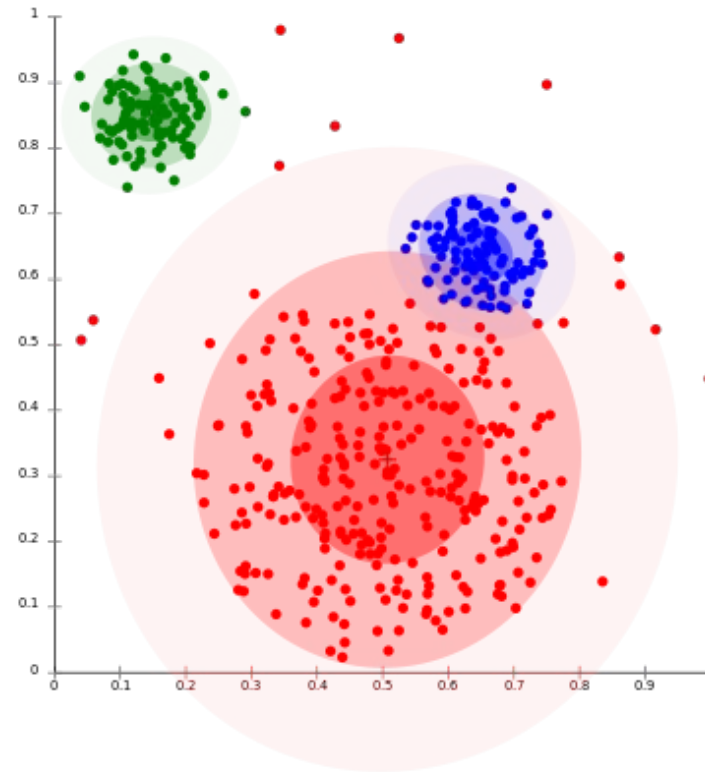
Introduction



1 What is clustering analysis?

A good clustering method will produce high quality clusters with:

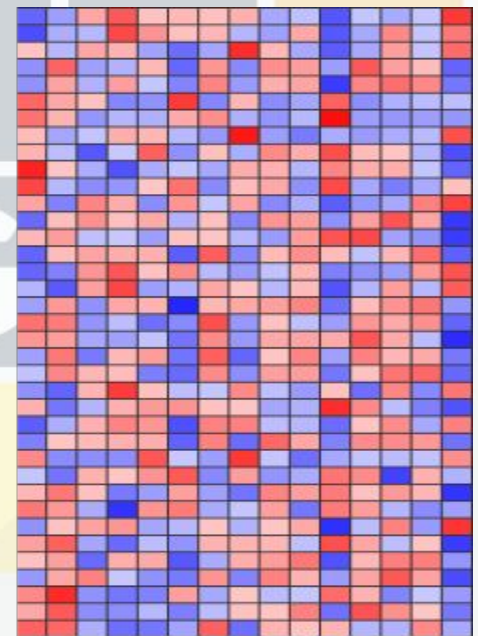
- High intra-class similarity (Green, Blue >> Red)
- Low inter-class similarity (Green vs Blue, Green vs Red >> Blue vs Red)



1 Input data

- Tab delimited file with numerical values (intensity)
- Genes in rows – samples in columns
- No class assigned to the samples (arrays)

gene1	10.23	9.98	10.41	10.55	10.65	9.69
gene2	10.51	9.74	10.65	10.63	10.43	10.35
gene3	9.89	10.02	9.89	11.03	10.21	10.77
gene4	10.25	10.83	8.94	10.16	10.49	10.46
gene...



1

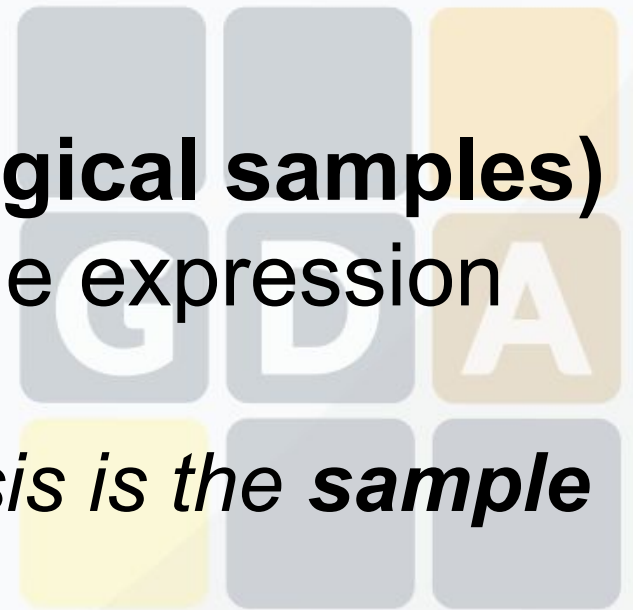
Goals

- Are there some **genes with a similar pattern** of gene expression across arrays?

*The unit (individual) of analysis is the **gene***

- Are there **some arrays (biological samples) with the same pattern** of gene expression across genes?

*The unit (individual) of analysis is the **sample***

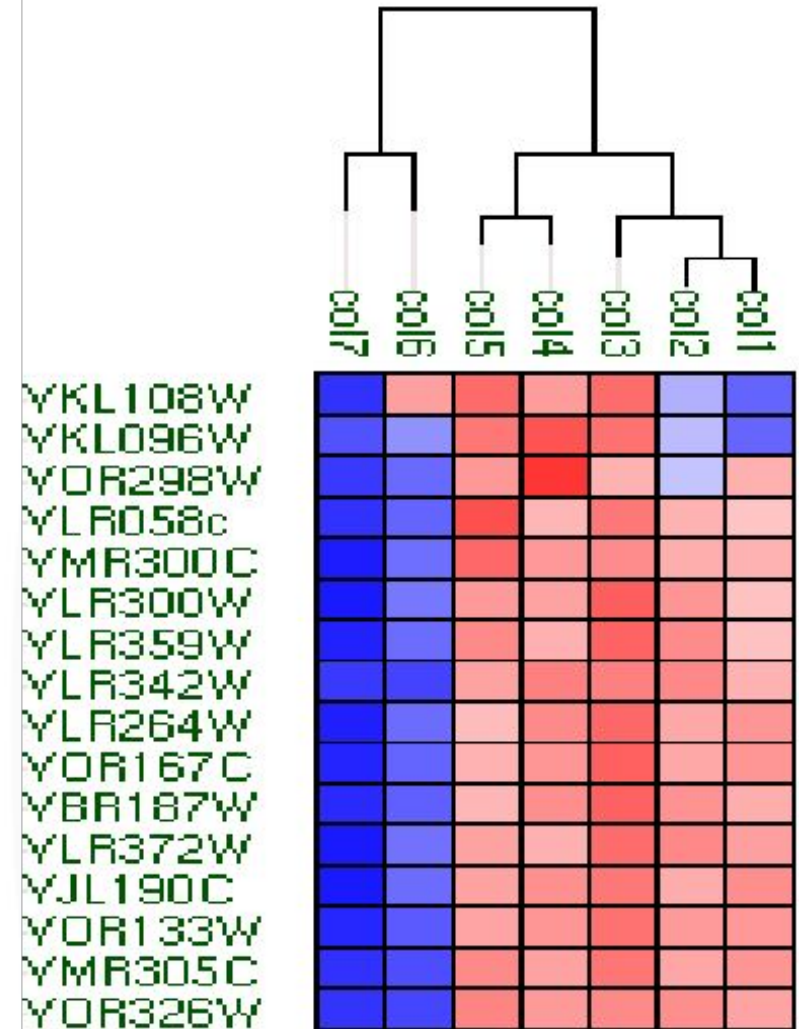
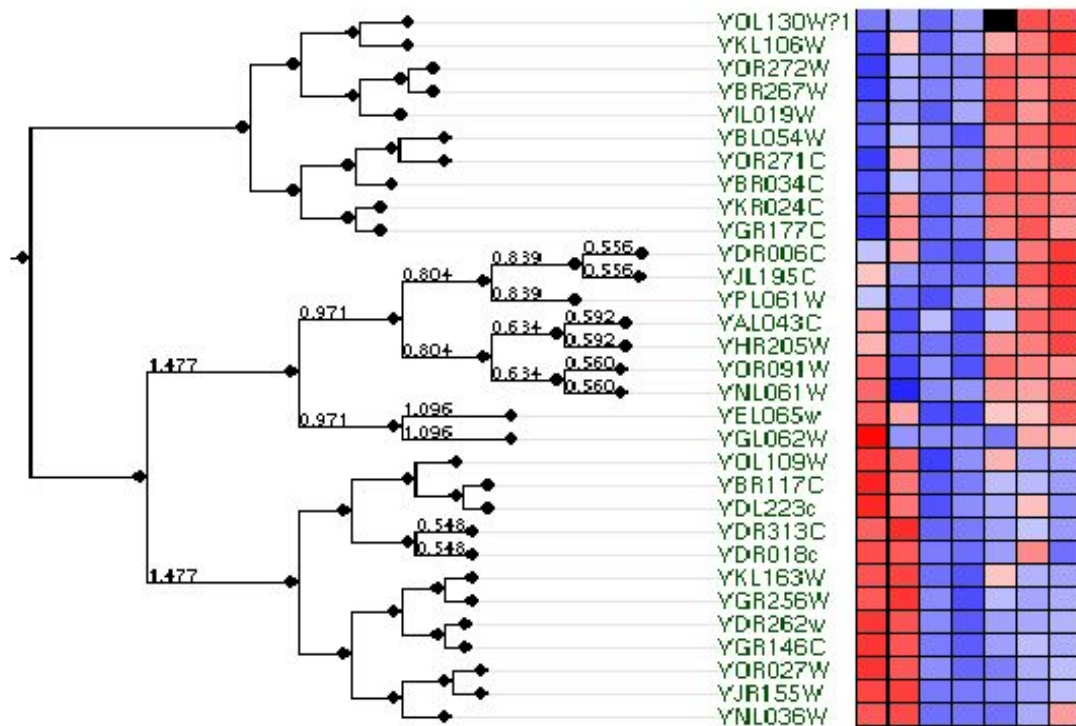


1

Goals

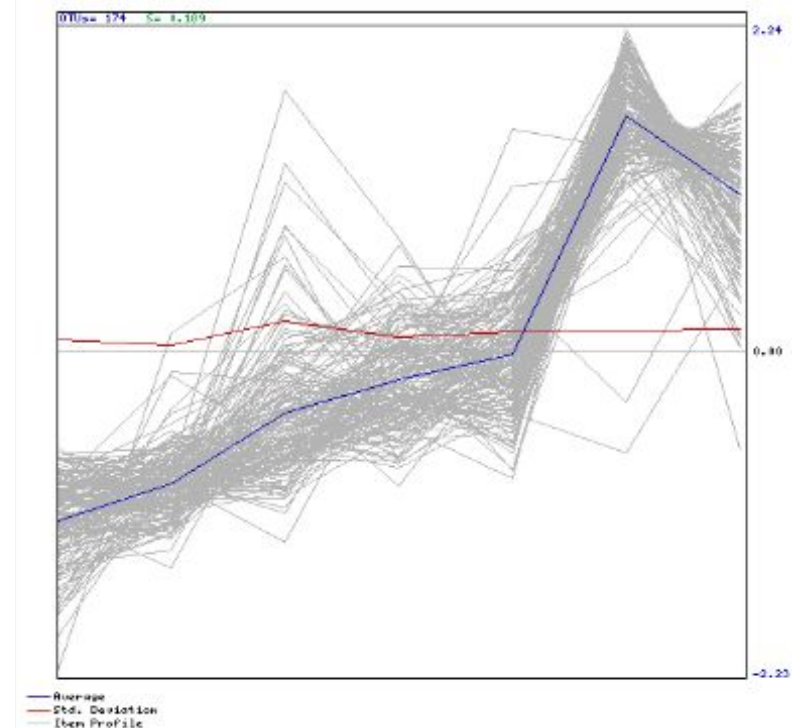
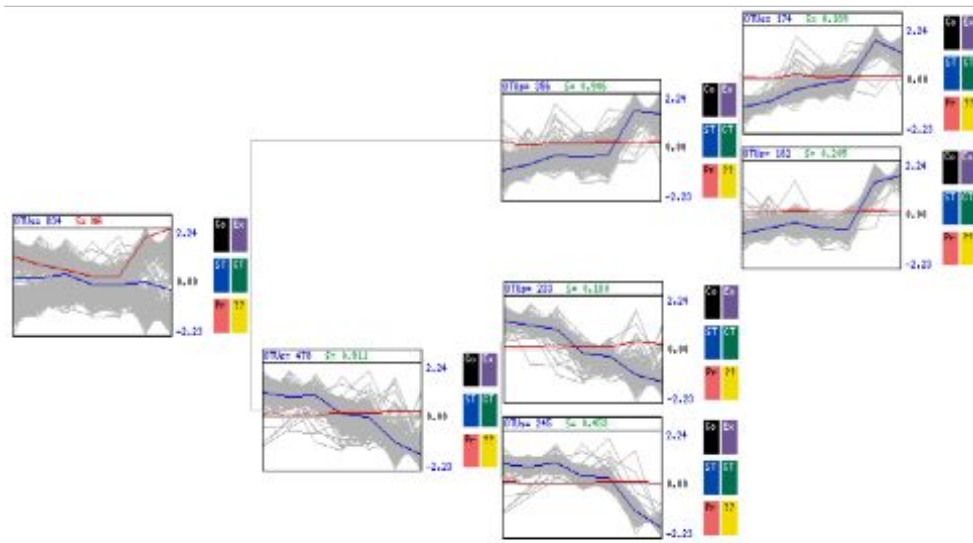
Array clustering

Gene clustering



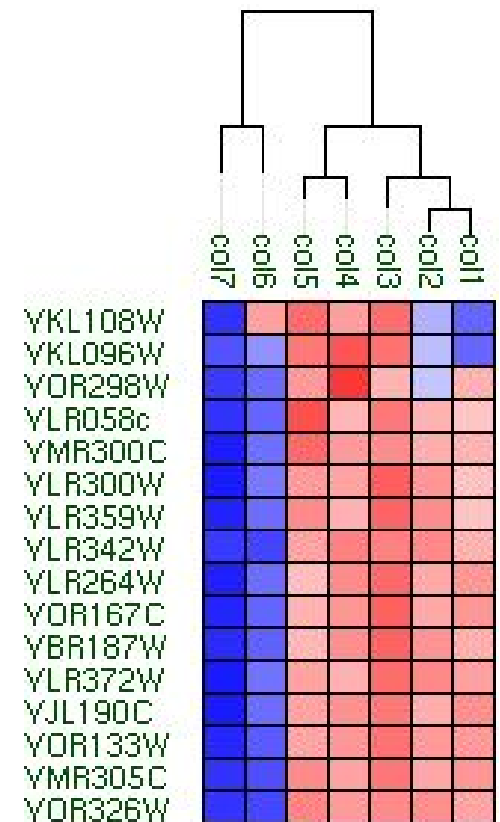
1 Gene cluster utility

- Find genes that behave the same across patients
- Indicate possible gene functionality
- Find temporal patterns of gene expression



1 Array clustering utility

- Discover new subgroups in a set of patients of the same disease
- Descriptive analysis
- Perform quality control checking:
 - Outlier (array) detection
 - Batch effect assessment



1 Steps for clustering analysis

1. Hierarchical or non hierarchical?
2. Choose a clustering method.
3. Choose a distance.





Outline

1. Introduction
- 2. Clustering methods**
3. Distance parameters
4. Exercises on Babelomics

2 Clustering methods

- **Hierarchical Methods**

- Aggregative
- Divisive

Provide a tree

- **Non Hierarchical – Partitioning - Methods**

- Usually need the number of clusters to be set
- Do not provide a tree

2 Aggregative hierarchical clustering

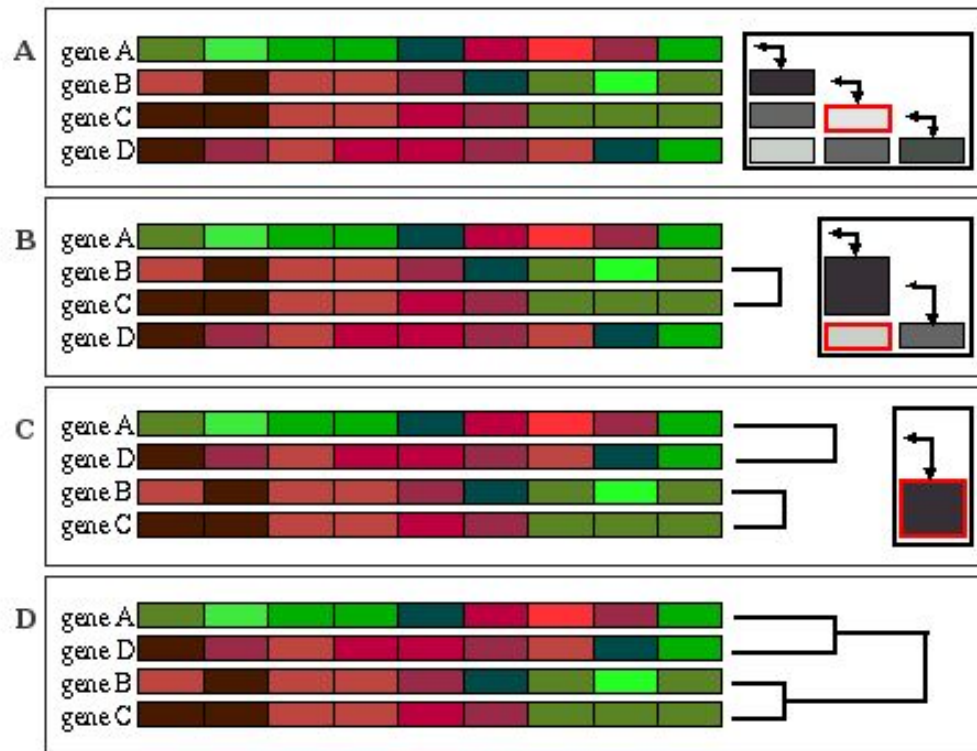
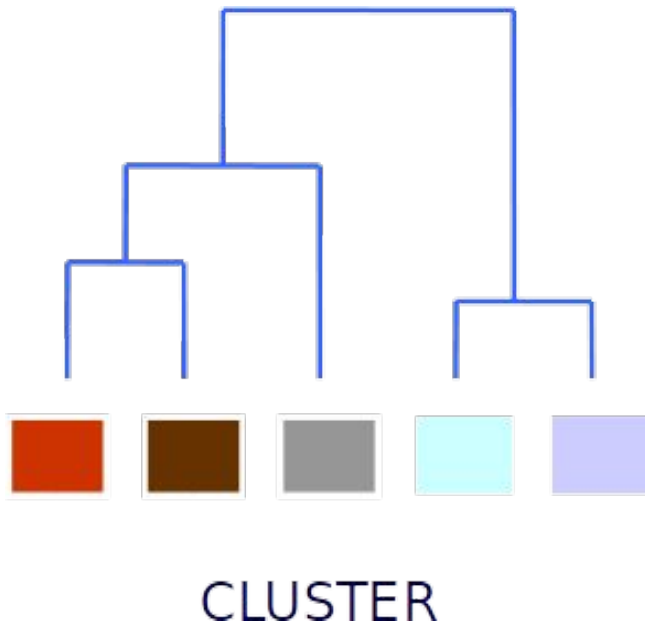


Fig. 1.5.1 UPGMA: the all-to-all distance matrix (black box) is calculated and two closest elements (red box) are merged. **(A)** The two closest elements are genes B and C. **(B)** The genes B and C are merged and the all-to-all distance matrix is calculated again using the new cluster instead of the genes B and C. Now, the two closest elements are genes A and D. **(C)** The genes A and D are also merged. The elements must be reordered to fit the topology of the tree. The all-to-all distance matrix is calculated again with the two remaining elements. **(D)** The process ends when all the complete dendrogram is built.

2 Aggregative hierarchical clustering

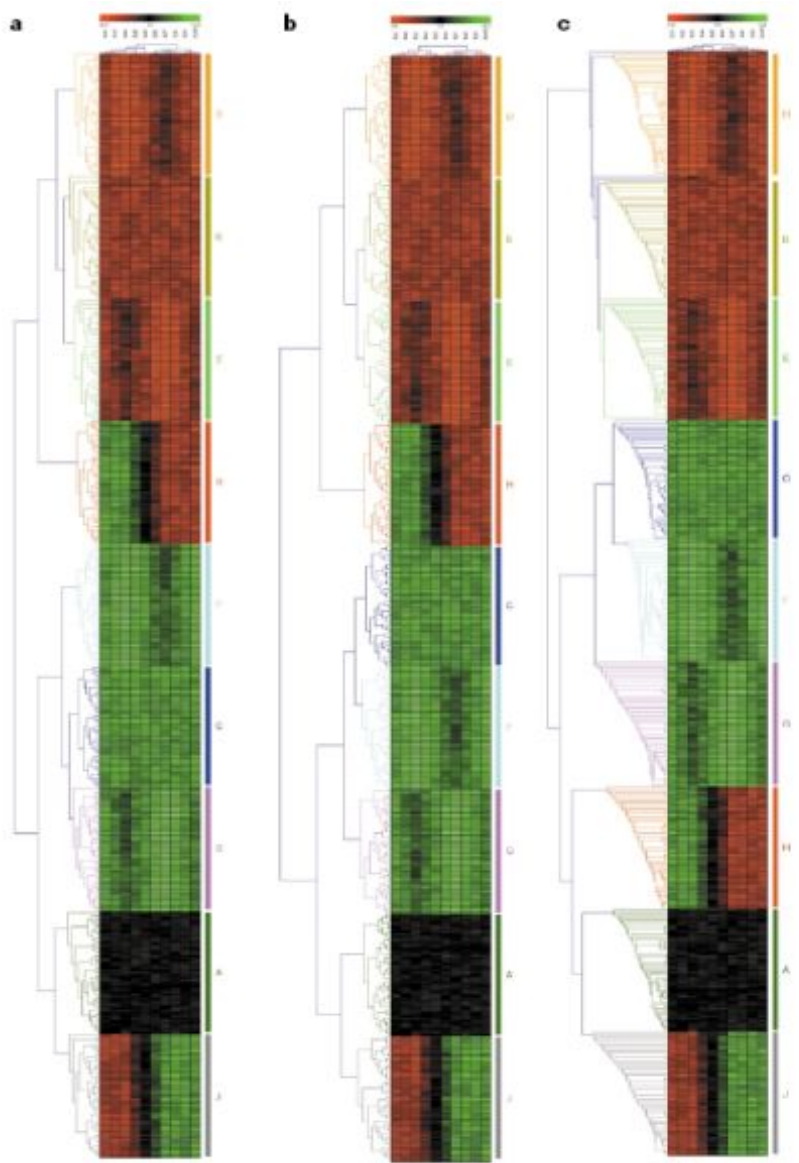


Relationships among profiles are represented by branch lengths.

The closest pair of profiles are recursively linked until the complete hierarchy is reconstructed

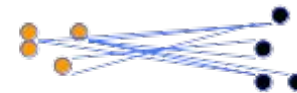
Allows to explore the relationship among groups of related genes at higher levels.

2 Different aggregative criteria

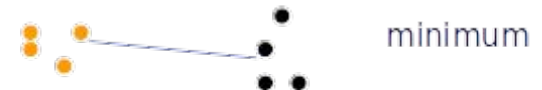


Different point to set
distance definition

a) Average linkage



b) Single linkage



c) Complete linkage



2 Clustering methods in Babelomics

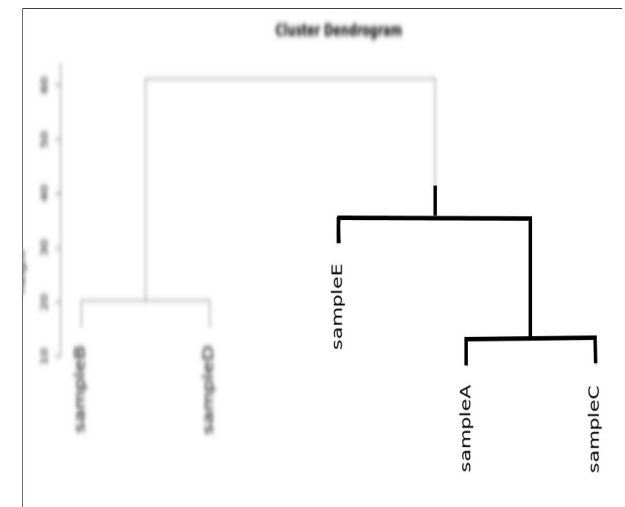
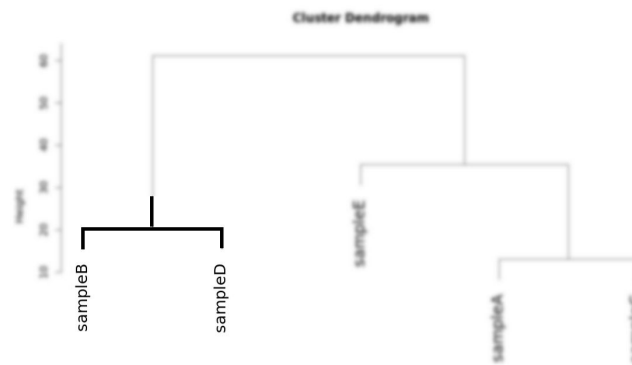
1. Unweighted Pair Group Method with Arithmetic Mean (**UPGMA**)
2. Self-Organizing Tree Algorithm (**SOTA**)
3. **K-Means**



2

UPGMA

- **UPGMA** is a simple agglomerative (bottom-up) hierarchical clustering method.
- This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- It is not the more accurate among the methods but is really extensively used especially for gene expression data. Provides a tree



2

SOTA

- SOTA starts the classification with a binary topology composed of a root node with two leaves.
- A divisive (top down) method.
- The self-organizing process splits the data (e.g. samples) into two clusters.
- After reaching convergence at this level, the network is inspected.
- If the level of variability in one, or more, terminal nodes is over a given threshold, then, the tree grows by expanding these terminal nodes. **Provides a tree.**

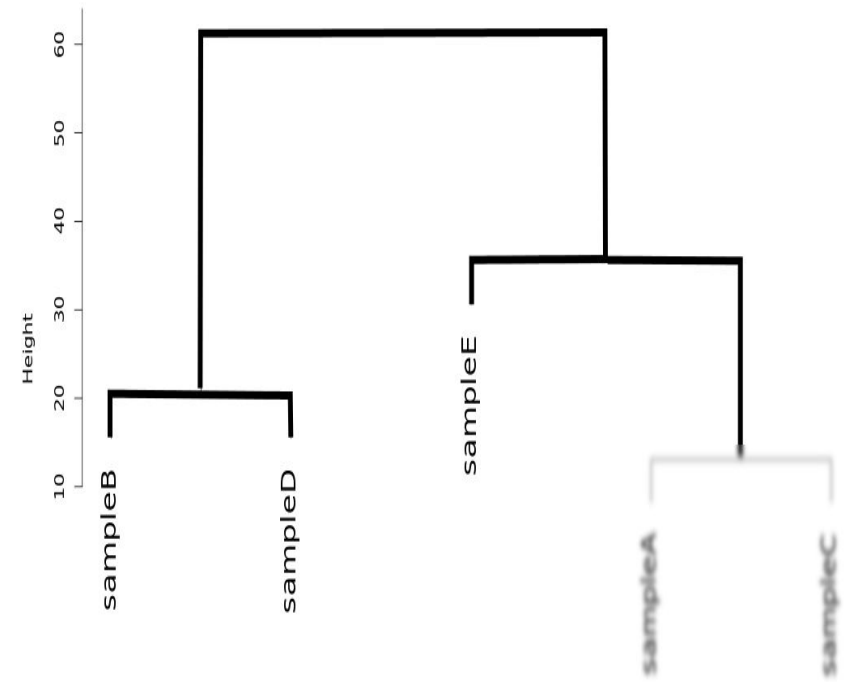
2

SOTA

Cluster Dendrogram



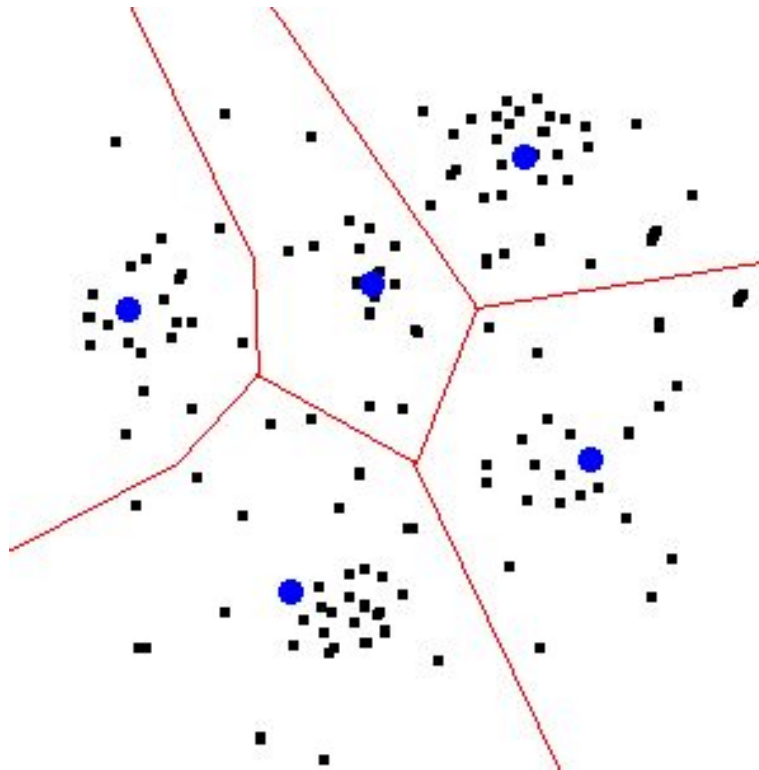
Cluster Dendrogram



2

K-Means

- K-means aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.
- Do not provide a tree.
- Usually need the number of cluster to be set.
- Its result is very sensitive to the initialization step: choosing initial cluster centers.

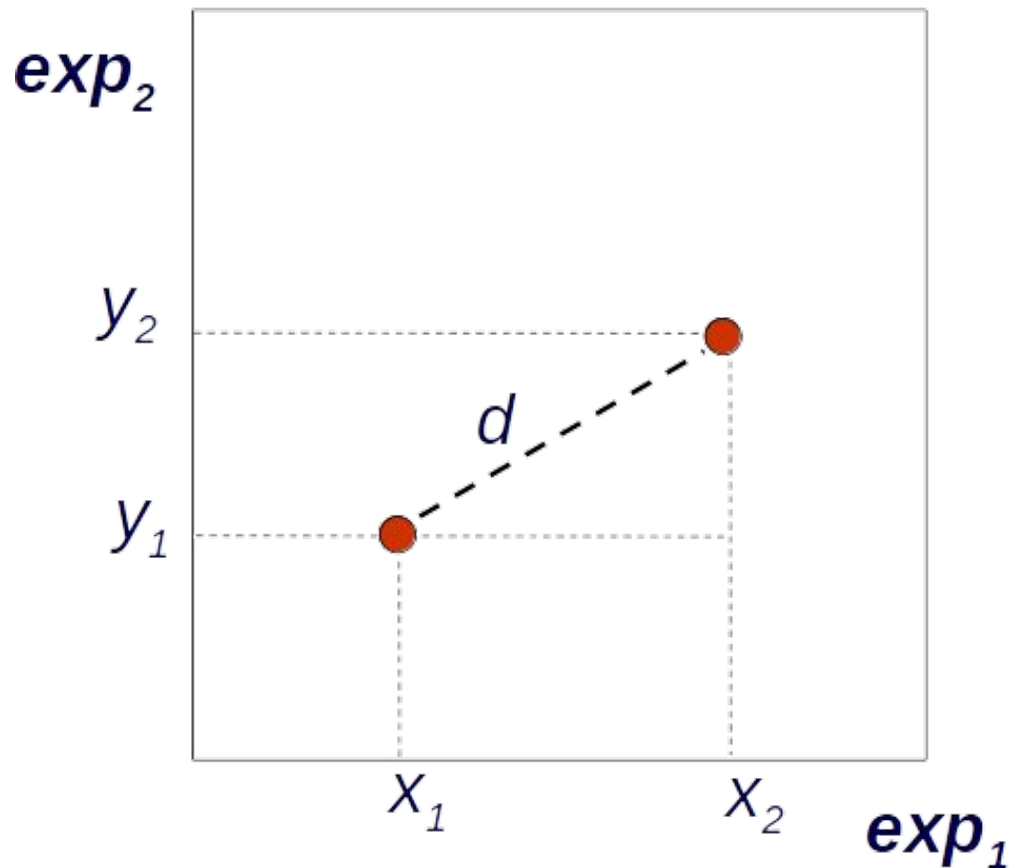




Outline

1. Introduction
2. Clustering methods
- 3. Distance parameters**
4. Exercises on Babelomics

3 Euclidean distance



	exp_1	exp_2
gene		
A	x_1	y_1
B	x_2	y_2
C	x_3	y_3

$$d_{x,y} = \sqrt{\sum (x_i - y_i)^2}$$

3 Correlation distance

Based in correlation coefficients.
Looks for similar patterns across individuals.

The correlation coefficient between n pairs of observations, whose values are (x_i, y_i) is:

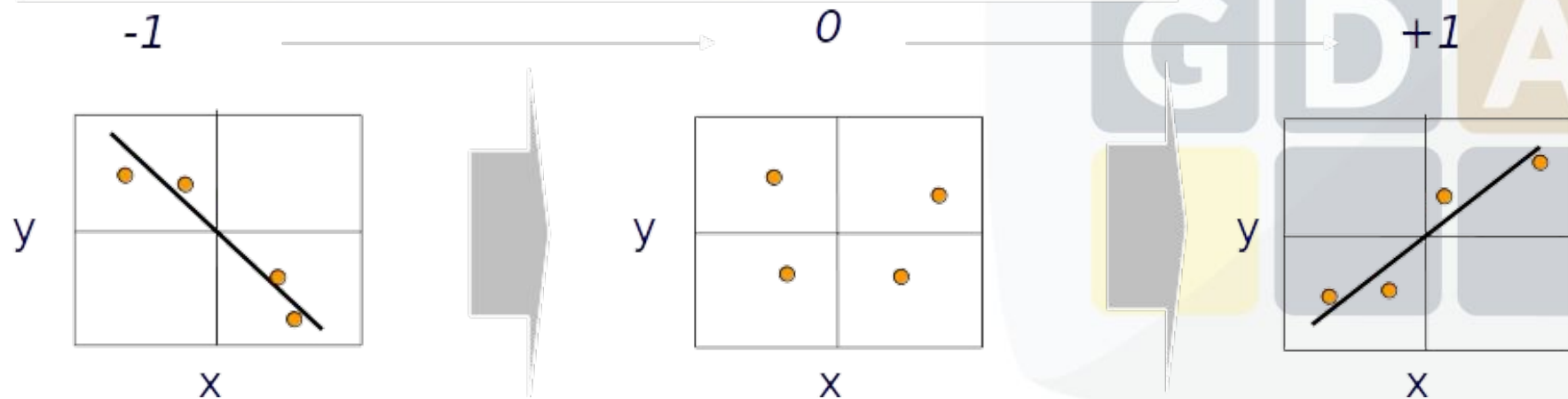
- **Pearson Correlation Coefficient**

S_x = Standard deviation of x

S_y = Standard deviation of y

$$\frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

The linear correlation coefficient measures the strength of the linear relationships between the paired x and y values in a sample.



3 Differences between distances

Differences (euclidean)

$$d_{x,y} = \sqrt{\sum (x_i - y_i)^2}$$

$B \sim C$

Correlation

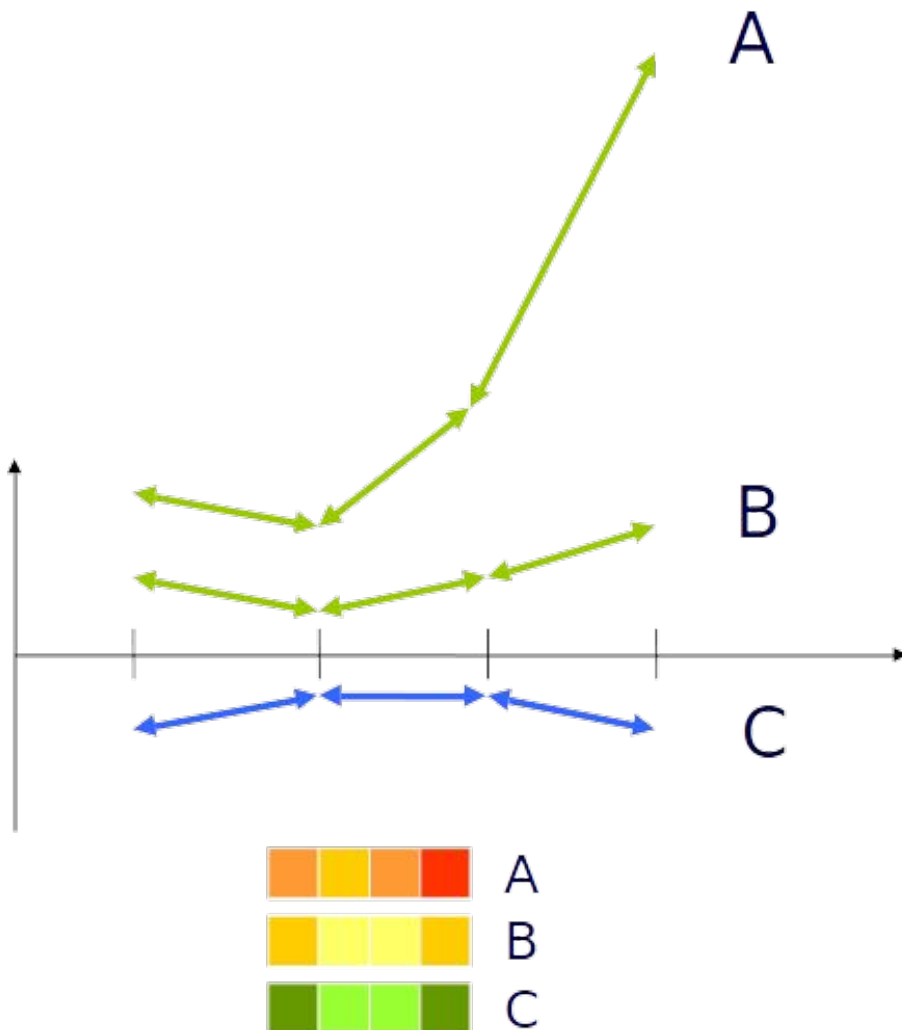
- Pearson Correlation Coefficient

S_x = Standard deviation of x

S_y = Standard deviation of y

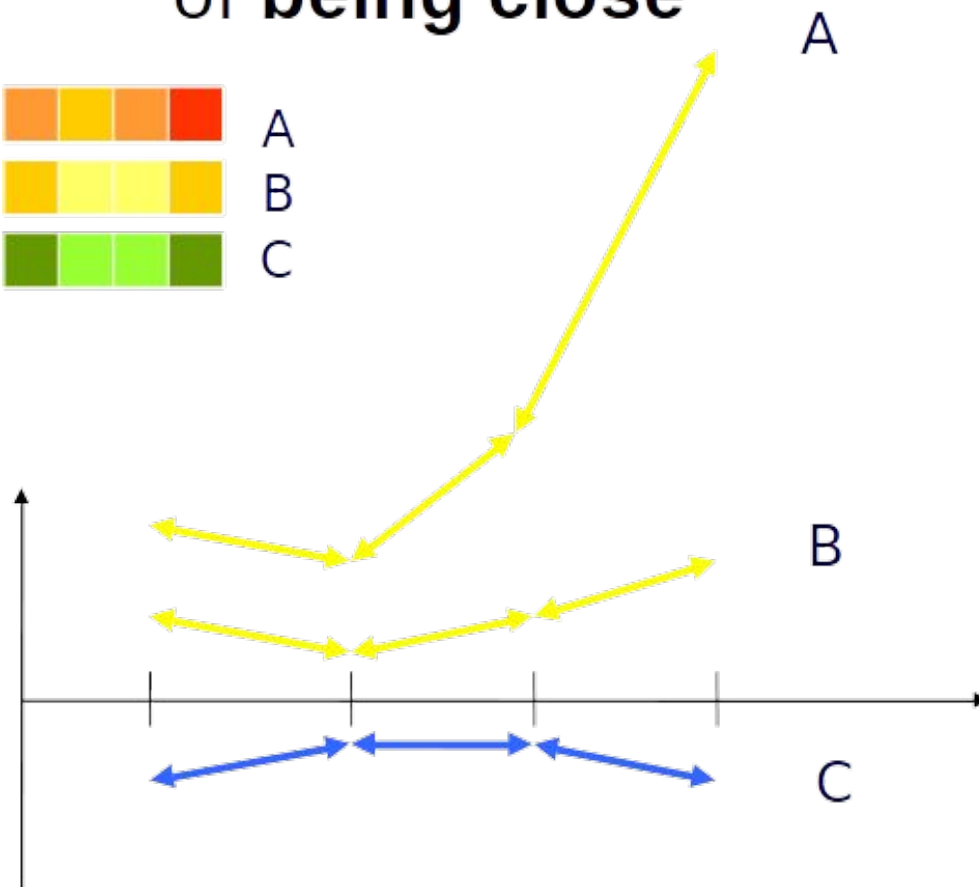
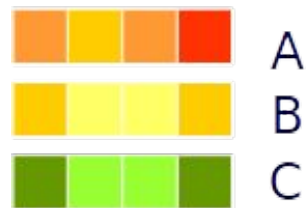
$$\frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

$A \sim B$

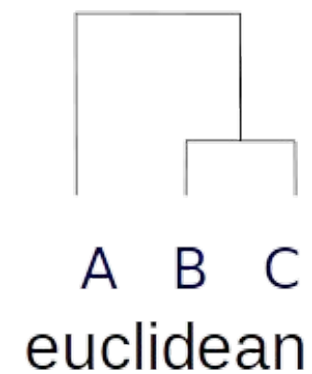
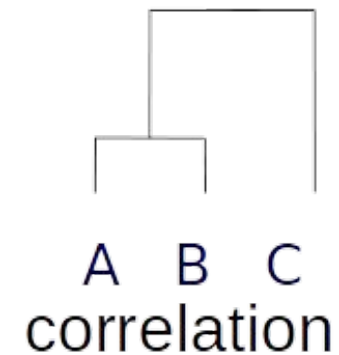


3 Differences between distances

Different definitions
of **being close**



Correlation: tendencies
Euclidean: global similarity



3 Distances in Babelomics

Different distances account for different properties:

1. Euclidean

- Normal
- Squared

2. Correlation coefficient

- Spearman
- Pearson



3 Any questions?





Outline

1. Introduction
2. Clustering methods
3. Distance parameters
- 4. Exercises on Babelomics**

4 Exercises on Babelomics

RPKM

TMM

UPLOAD
DATA

NORMALIZE
DATA

EDIT
DATA

CLUSTERING

PREDICTORS

#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FOR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	100	84	94	86	354	362	373	369	326
NFYA	59	61	53	56	59	59	66	63	66	62
STPQ1	34	43	41	31	46	6	7	7	8	7

