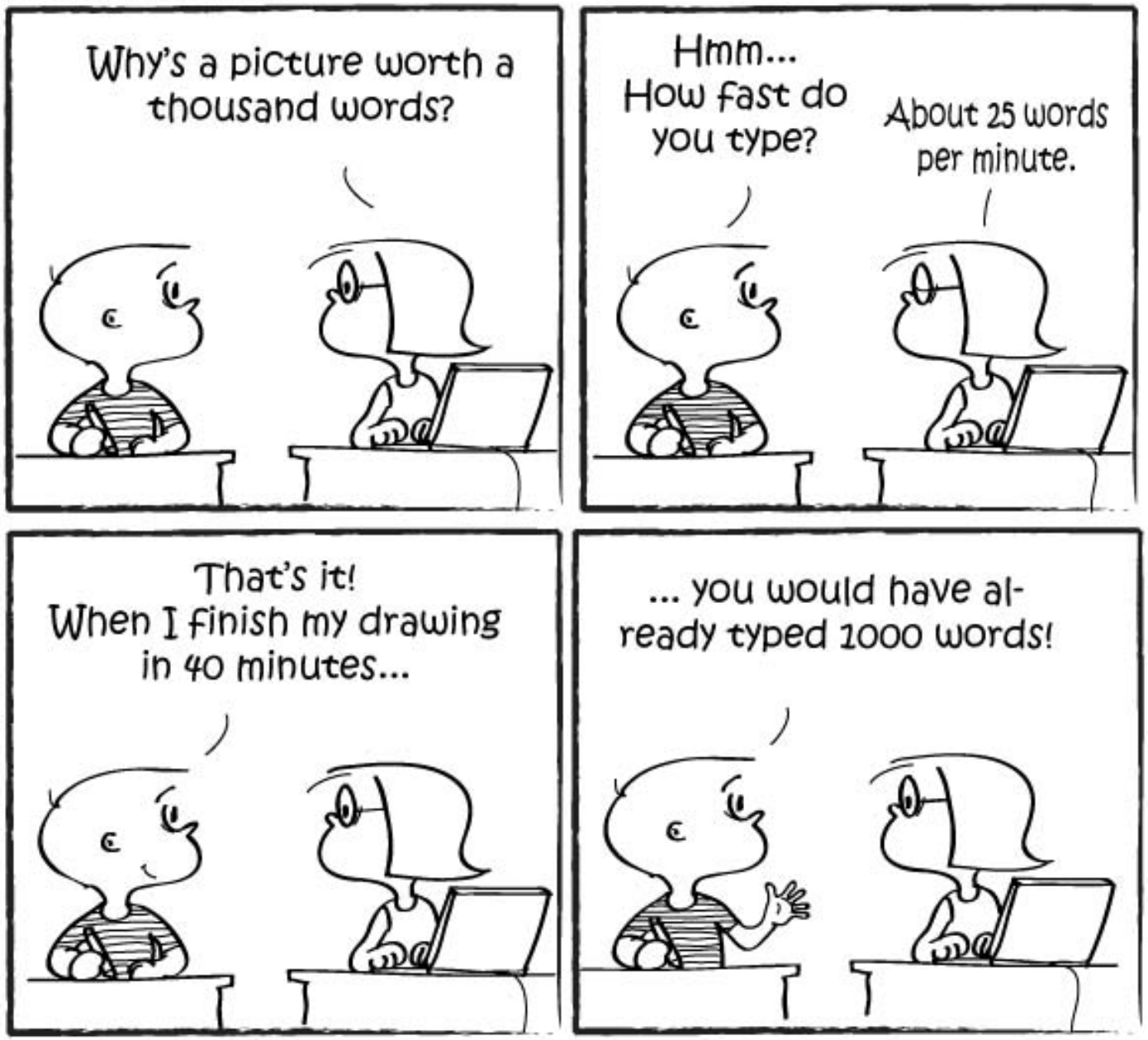
Transcriptome
Lecture 3

Outline

- Multiple Testing Procedures
- Data Visualization, Distance Measures
- Clustering
- Gene Annotation and Enrichment Analysis

giggleBites



The Visualization

- MA plot
- Volcano plot
- Heatmap
- Dendrogram

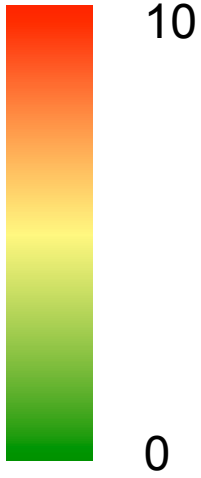
Heatmap

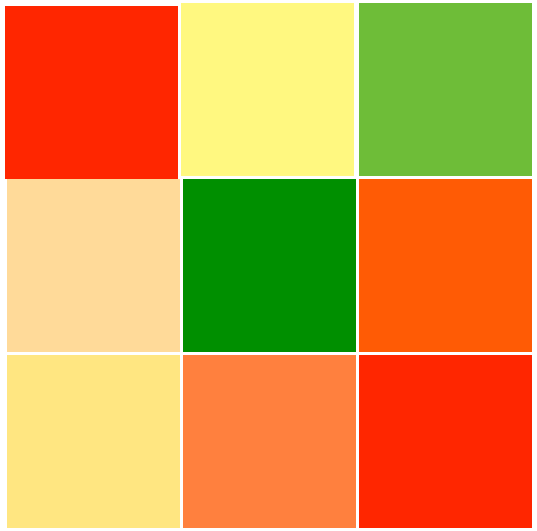
- A **heat map** is a graphical representation of data where the values taken by a variable in a two-dimensional map are represented as colors

- In Microarray/RNA-seq, it plots the level of expression of many genes (in y-axis) across a number of samples (in x-axis)
  - The data is in the form of matrix

Heatmap

	S1	S2	S3
g1	10	5	2
g2	7	0	9
g3	6	8	10





Heatmap: Example

- **ALL** data (Lymphoblastic leukemia study):
  - 12625 probes (genes) rows
  - 128 samples columns

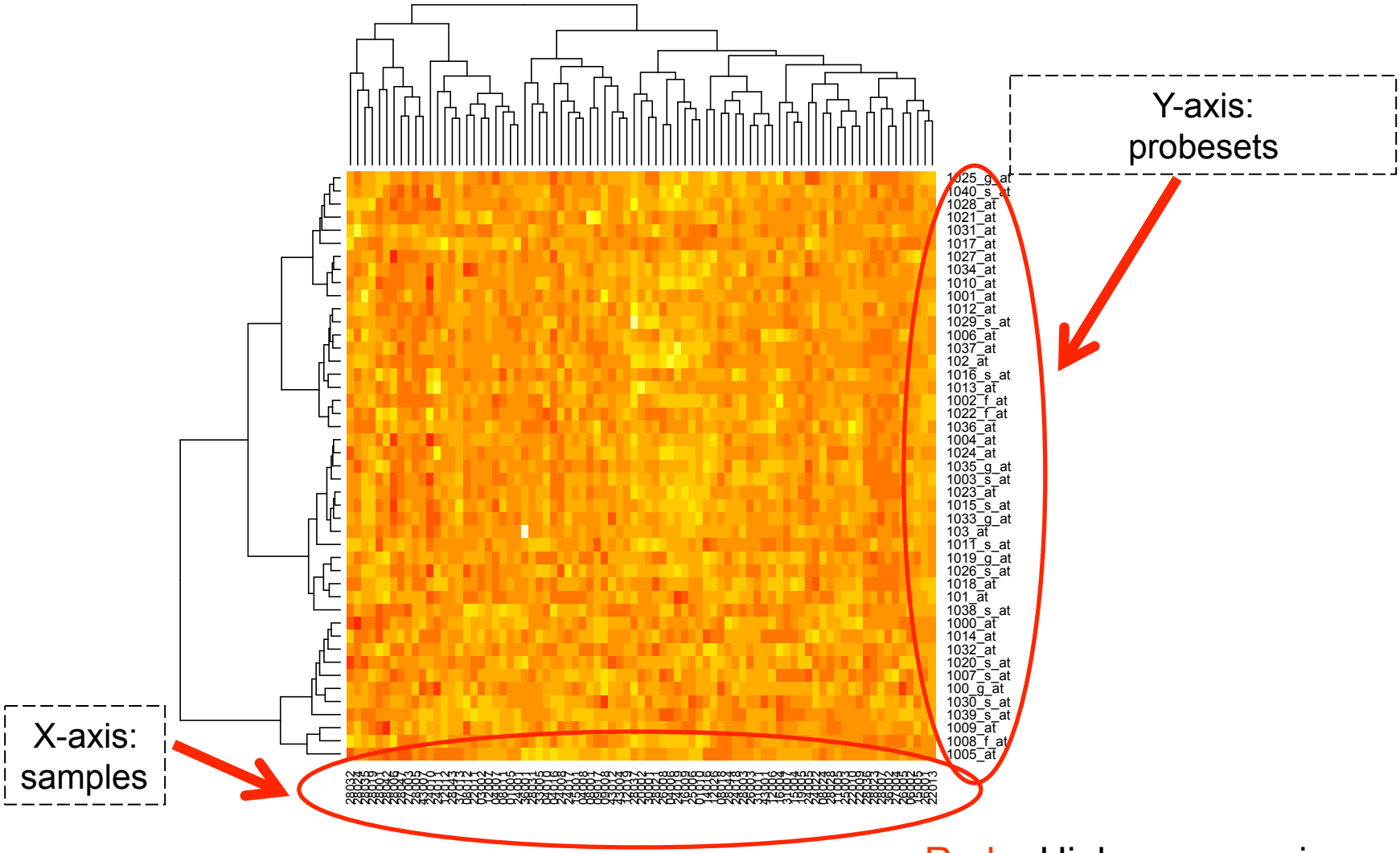
```
> exprs(ALLhm)
           ALL1/AF4 04006 E2A/PBX1 08018 ALL1/AF4 15004 ALL1/AF4 16004 ALL1/AF4 19005 AL
1007_s_at       6.816397       7.151422       6.822427       6.709222       6.798443
1044_s_at       4.570669       7.019295       4.892009       4.889920       4.339371
1065_at         8.475419       6.880097       9.939768       9.140339       9.579710
1081_at         8.631929      10.443100       8.487560       7.823037       9.879712
1134_at         7.854585       9.238699       7.559106       7.837794       7.864575
1140_at         8.039748       5.798014       6.791144       6.733774       7.276141
```

Heatmap: Example

- **ALL** data (Lymphoblastic leukemia study):
- source("http://www.bioconductor.org/biocLite.R")
- biocLite("ALL")

```
> library("ALL”)
> data(ALL)
> p=exprs(ALL)[1:45,1:81]
> heatmap(p)
```

Heatmap: Example



Red = Higher expression
Yellow = Lower Expression

The Visualization

- MA plot
- Volcano plot
- Heatmap
- Dendrogram

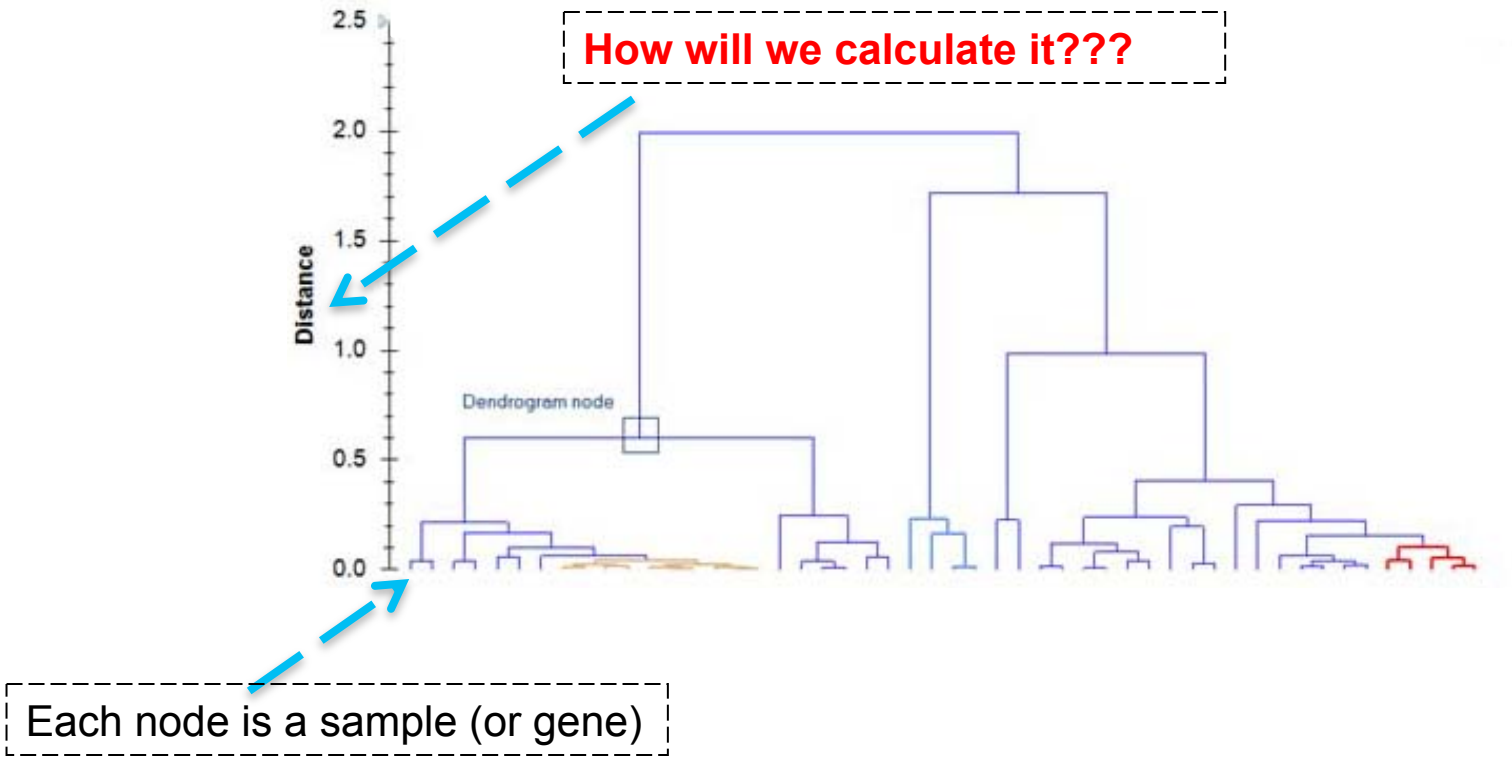
Dendrogram

- A **Dendrogram** is a tree diagram frequently used to illustrate the arrangement of the clusters produced by clustering.

- In Microarray/RNA-seq, it can represent the distance between a number of samples (or genes)

Dendrogram

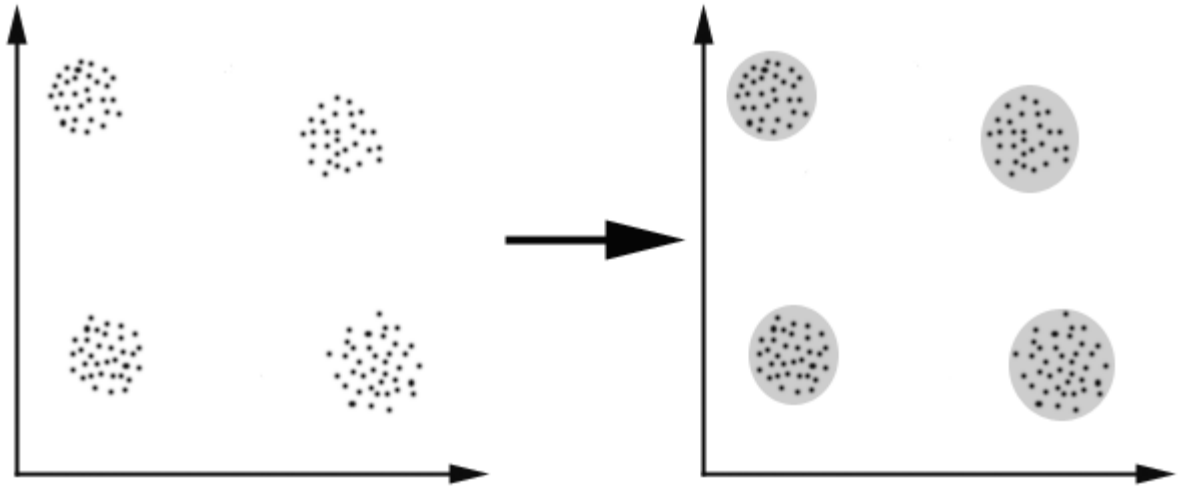
- In Microarray/RNA-seq, a dendrogram can represent the distance between a number of samples (or genes)



Outline

- Multiple Testing Procedures
- Data Visualization, Distance Measures
- Clustering
- Gene Annotation and Enrichment Analysis

Clustering: what is it?

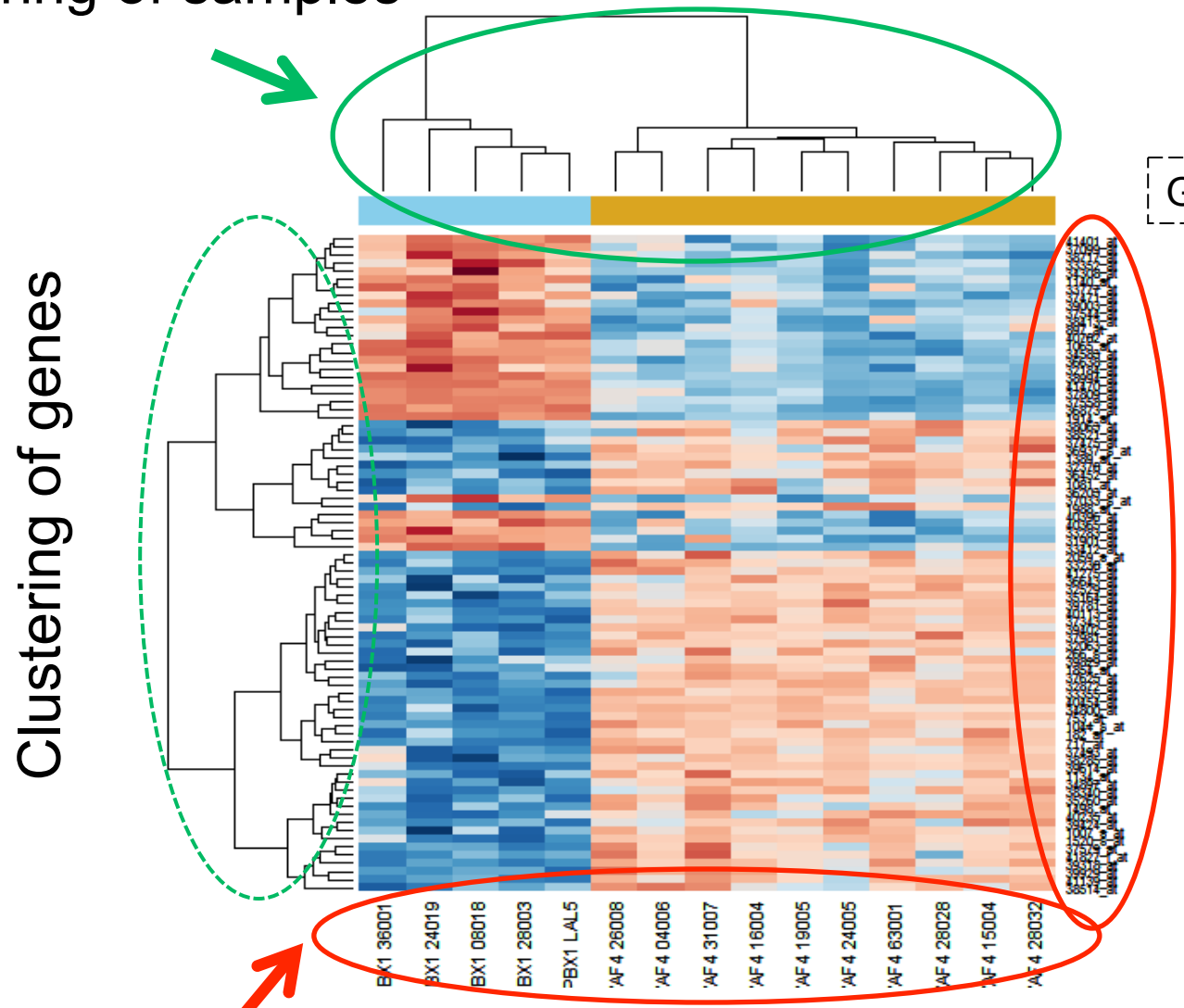


The goal of clustering could be to gather genes or samples into groups. We call those groups as clusters.

A *cluster* is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Clustering

Clustering of samples



Brown = Higher expression
Blue = Lower Expression

Clustering: why cluster genes?

- Identify groups of possibly co-regulated genes (e.g. genes regulated by one transcription factor).
- Identify typical temporal or spatial gene expression patterns (e.g. cell cycle data).
- Arrange a set of genes in a linear order that is at least not totally meaningless (for visualization).

Clustering: why cluster samples?

- Quality control: Detect experimental artifacts/ bad hybridizations
- Check whether samples are grouped according to known categories
- Identify new classes of biological samples (e.g. tumor subtypes)

Clustering: Basic principles

- Issues to be consider before performing a cluster analysis
  - ❑Which genes/arrays to be used?
  - ❑Which distance (similarity) measures?
  - ❑Which method is used to join clusters/ observations?
  - ❑Which clustering algorithm is applied?

Clustering: Basic principles

- Issues to be consider before performing a cluster analysis
  - ❑ Which genes/arrays to be used?
    - It is **advisable to reduce the number of genes** from the full set to some more manageable number, before clustering.
    - A common approach is to perform a cluster analysis based on differentially expressed genes
  - ❑ Which distance (similarity or dissimilarity) measures?
  - ❑ Which method is used to link clusters?
  - ❑ Which clustering algorithm is applied?

Clustering: Basic principles

- Issues to be consider before performing a cluster analysis
  - ❑Which genes/arrays to be used?
  - ❑Which distance (similarity or dissimilarity) measures?
    - Correlation coefficient based distance (scale-independent)
    - Minkowski metric (scale-dependent)
  - ❑Which method is used to link clusters?
  - ❑Which clustering algorithm is applied?

Distance: classes

- Distance between points
  - Minkowski metric
    - Euclidean metric
    - Manhattan metric
  - Correlation distance
    - Pearson sample correlation distance
    - Cosine correlation distance
    - Spearman sample correlation distance

Minkowski distance

For two points $X = (x_1,\ldots,x_m)$ and $Y=(y_1,\ldots,y_m)$

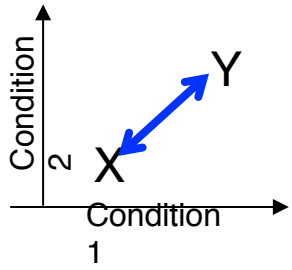
- Minkowski distance

$$F(z_1,\ldots,z_m) = \left(\sum_{k=1}^{m} z_k^{\lambda}\right)^{1/\lambda}$$

$$z_k = d_k(x_k, y_k) = |x_k - y_k|$$

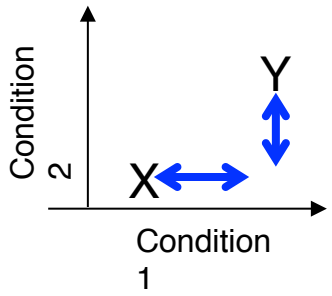
  - EUC Euclidean distance $\lambda = 2$

$$d_{euc}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}.$$



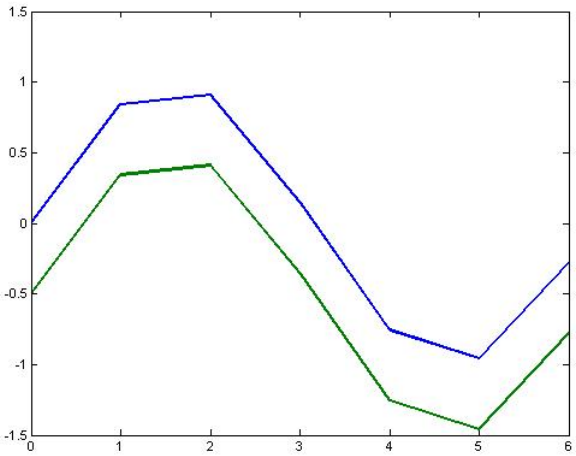
  - Man Manhattan distance $\lambda = 1$

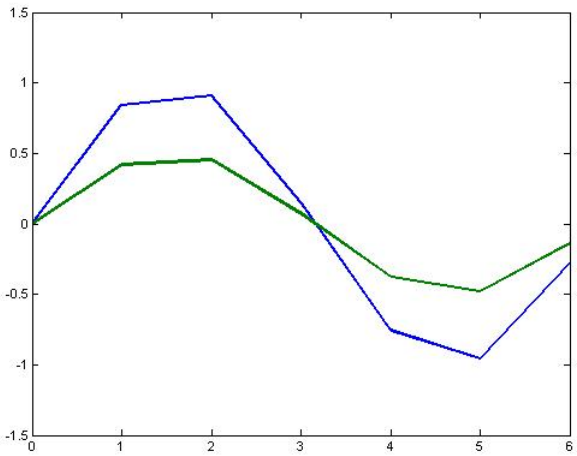
$$d_{man}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m}|x_i - y_i|.$$



Distance

- In most case, we care more about the overall shape of expression profiles rather than the actual magnitudes
- That is, we might want to consider genes similar when they are “up” and “down” together





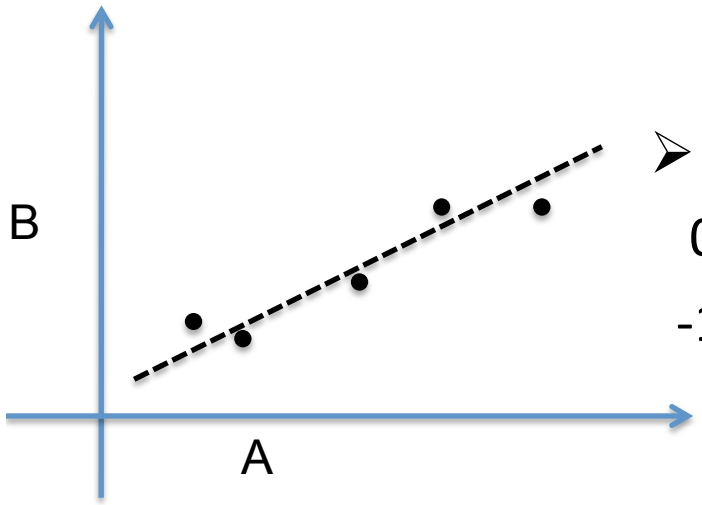
Distance: correlations

- Pearson correlation coefficient

$$r(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{m}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\overline{x})^2\sum_{i=1}^{m}(y_i-\overline{y})^2}}.$$

$$r(x,y) \in [-1,1]$$

- Pearson's correlation reflects the degree of linear relationship between X and Y.



- 1 – perfect similarity(positive linear)
  0 – no similarity
  -1 – perfect dissimilarity(negative linear)

Distance: correlations

- Pearson sample correlation distance

$$d_{cor}(\mathbf{x},\mathbf{y}) = 1 - r(\mathbf{x},\mathbf{y}) = 1 - \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2 \sum_{i=1}^{m}(y_i - \overline{y})^2}}.$$

- Cosine correlation distance

$$d_{eisen}(\mathbf{x},\mathbf{y}) = 1 - \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = 1 - \frac{|\sum_{i=1}^{m} x_i y_i|}{\sqrt{\sum_{i=1}^{m} x_i^2 \sum_{i=}^{m} y_i^2}}$$

- Spearman's rank correlation distance

Correlation Coefficient in R

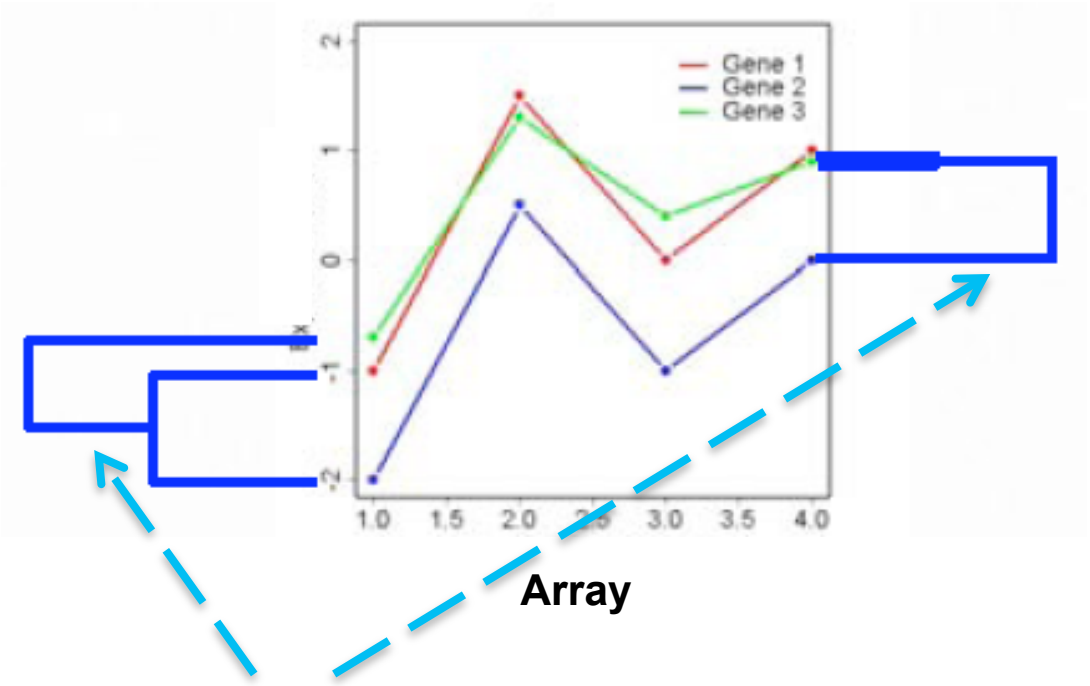
- cor(x, method=“pearson")
- "pearson", "kendall", "spearman”

```
Pairwise calculation to any two collums
> cor(d_matrix, methods=”pearson”)
> cor(t(d_matrix), methods=”pearson")
```

```
For distance

> d=1- cor(d_matrix, methods=”pearson”)
```

	S1	S2	S3	S4	S5	S6
g1						
g2						
g3						
g4						
g5						
g6						

Euclidean vs. Correlation



Are these clustering based on euclidean distance or correlation coeffecient?

Clustering: Basic principles

- Issues to be consider before performing a cluster analysis
  - ❑Which genes/arrays to be use?
  - ❑Which distance (similarity or dissimilarity) measures?
  - ❑Which method to use to link clusters?
    - ❑How to compute the cluster similarity in order to link them?
  - ❑Which clustering algorithm?

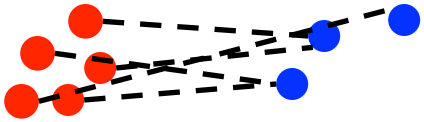
Clustering: Cluster similarity

- Four major methods to compute group similarity:
  - Given two clusters c1 and c2
    - Single-link: s(g1,g2)= similarity of the **closest pair** of points between the two clusters
    - Complete-link: s(g1,g2)= similarity of the **furtherest pair** of points between the two clusters
    - Average-link: s(g1,g2)= **average** of similarity of all pairs of points between the two clusters
    - Centroid-link: s(g1,g2)= distance between **centroids** of the two clusters

Clustering: cluster similarity

Single-link: similarity of the closest pair of points between the two clusters

Complete-link: similarity of the furtherest pair of points between the two clusters



Average-link: average of similarity of all pairs of points between the two clusters

Centroid-link: distance between centroids of the two clusters

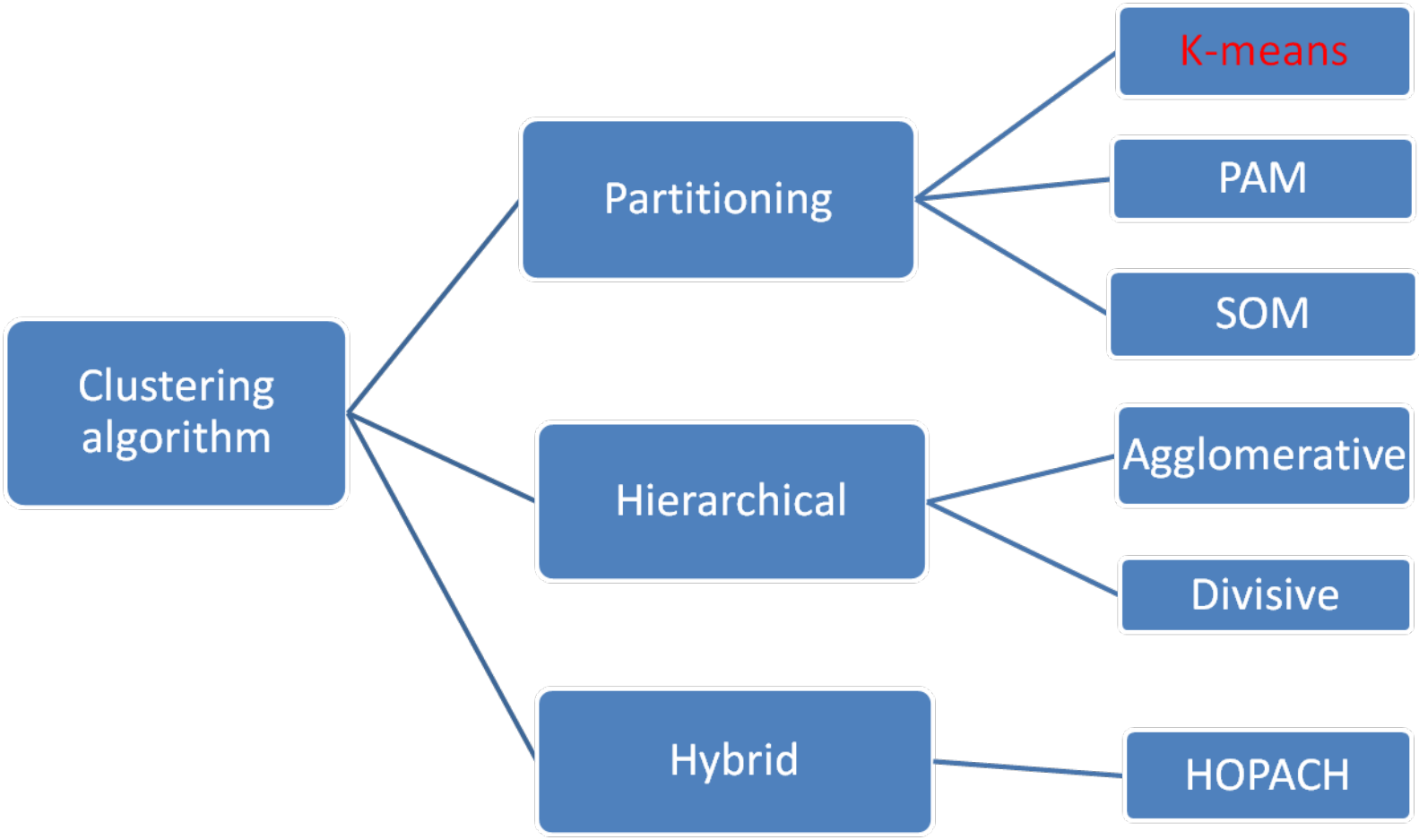
Clustering: Cluster similarity

- A comparison of cluster linkage methods:
  - Single-link and complete link: individual decision, more sensitive to outliers
  - Average-link and centroid-link: group decision, less sensitive to outliers

Clustering: Basic principles

- Issues to be consider before performing a cluster analysis
  - ❑Which genes/arrays to be use?
  - ❑Which distance (similarity or dissimilarity) measures?
  - ❑Which method to use to link clusters?
  - ❑Which clustering algorithm?

Type of Clustering algorithm



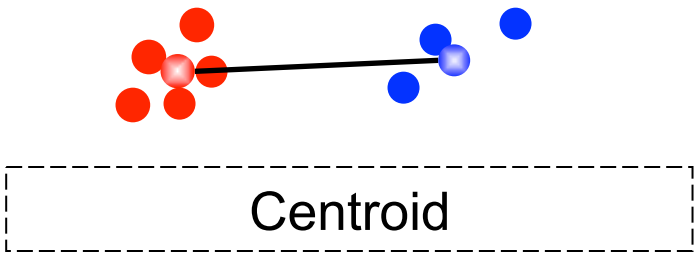
Partitioning: K-means

- A partitioning algorithm with a prefixed number k of clusters, that tries to minimize the sum of within-cluster-variances

$$\min\left(\sum_{i=1}^{k}\sum_{x_j\in S_i}\left\|x_j-\mu_i\right\|^2\right)$$

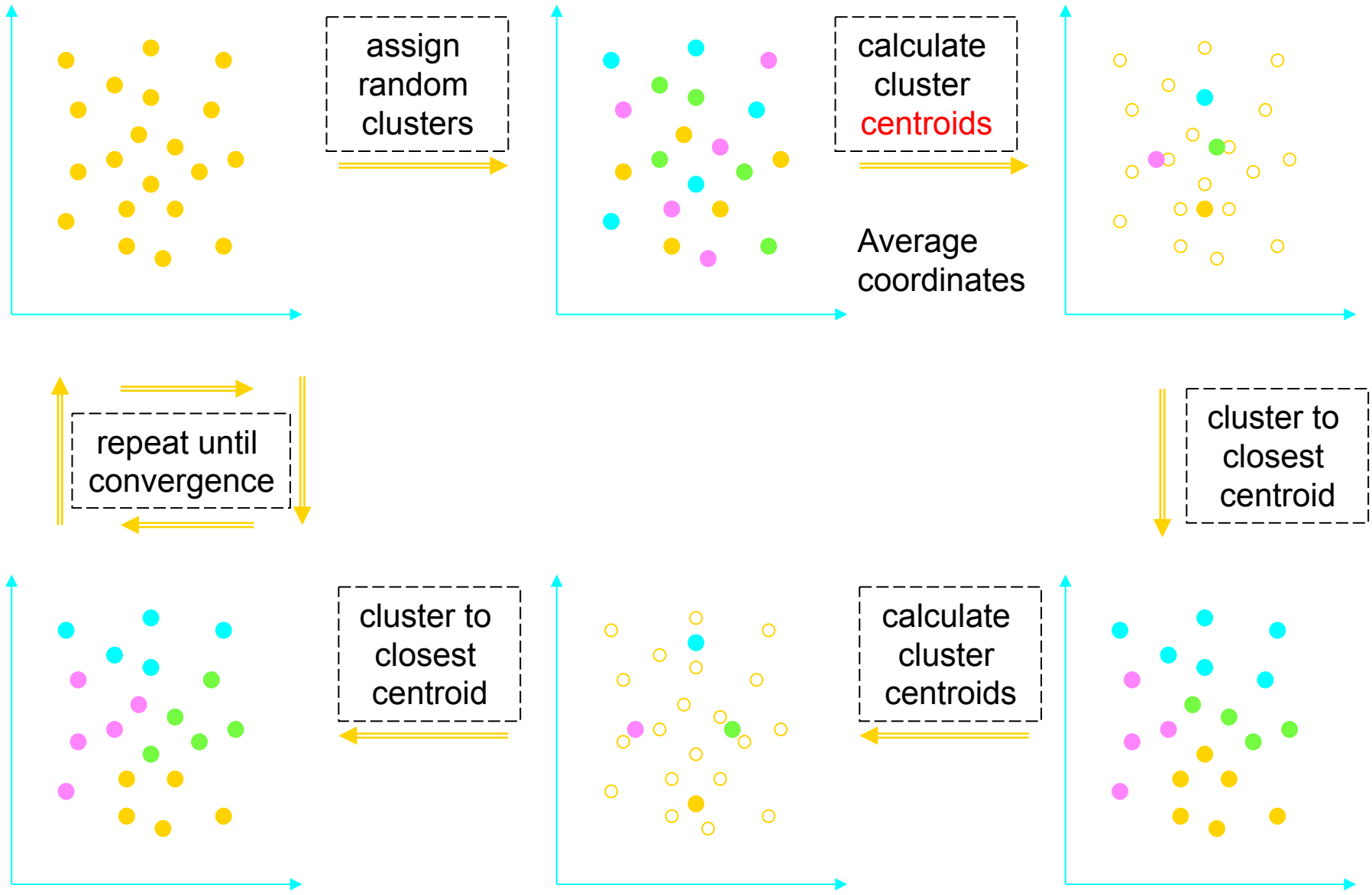
- Algorithm
  1. Randomly choose K points as the center of the K clusters
  2. Visit each point to its closest cluster
  3. Update the center of each newly formed cluster
  4. Repeat steps 2-4 until there is no change to the centers (centroids)(or reach the maximum cycles)

Clustering: cluster similarity



$$1/n \sum_{x_j \in S_i} x_j$$

Partitioning: K-means



Partitioning: K-means

- A partitioning algorithm with a prefixed number k of clusters, that tries to minimize the sum of within-cluster-variances
- MUST choose number of clusters K as a priori
  - If K =2, the data will be clustered (partitioned) into two clusters…
  - If K =4, the data will be clustered (partitioned) into two clusters..
  - …

Partitioning: K-means

- Use “cclust” package in R
- http://cran.r-project.org/web/packages/cclust/index.html

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("cclust")

> library(cclust)
> ALL_exp=exprs(ALL)
> kc=cclust(ALL_exp,10,200,verbose=TRUE,
   dist="euclidean", method="kmean")
```

Distance: euclidean or manhattan

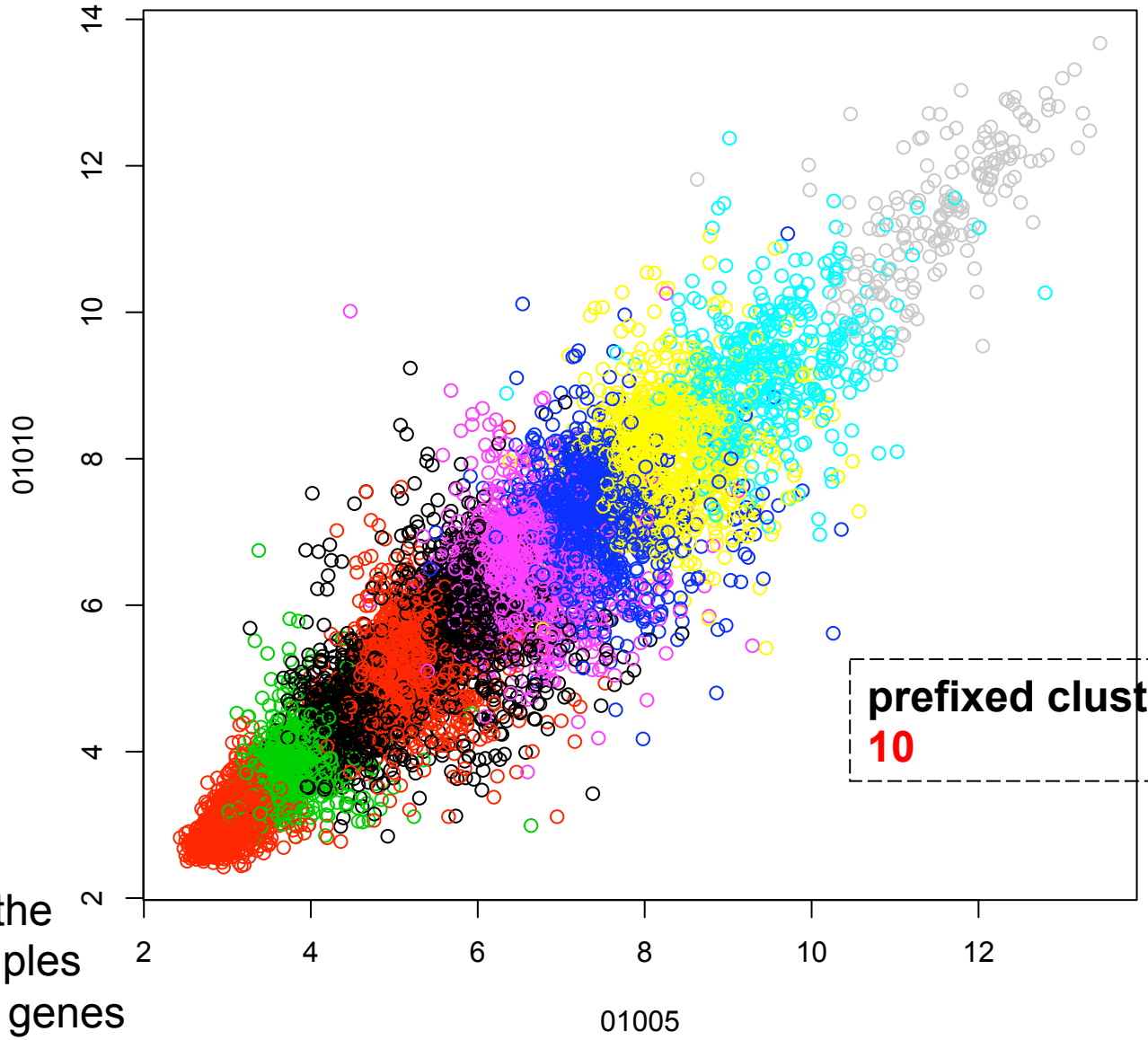
Partitioning: K-means

- Use "cclust" package in R

```
> kc=cclust(d,10,200,verbose=TRUE, dist="euclidean", method="kmean")
Iteration:   1    Changes:        10559
Iteration:   2    Changes:         2293
Iteration:   3    Changes:          929
Iteration:   4    Changes:          795
Iteration:   5    Changes:          789
Iteration:   6    Changes:          890
Iteration:   7    Changes:          871
Iteration:   8    Changes:          747
Iteration:   9    Changes:          645
Iteration:  10    Changes:          564
……..
Iteration: 117    Changes:           14
Iteration: 118    Changes:           11
Iteration: 119    Changes:            5
Iteration: 120    Changes:            6
Iteration: 121    Changes:            8
Iteration: 122    Changes:            5
Iteration: 123    Changes:            5
Iteration: 124    Changes:            6
Iteration: 125    Changes:            3
Iteration: 126    Changes:            3
Iteration: 127    Changes:            5
Iteration: 128    Changes:            3
Iteration: 129    Changes:            0
```

Partitioning: K-means

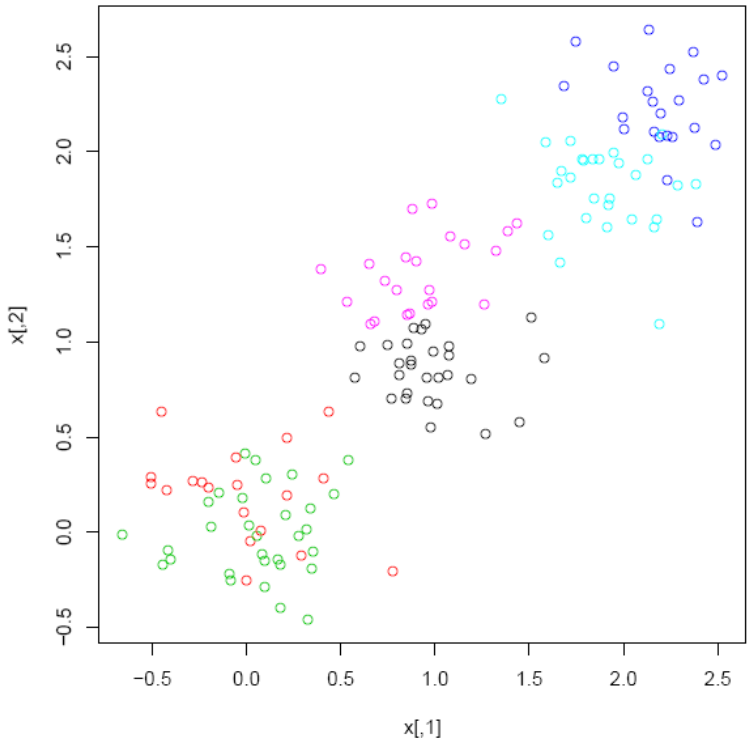
> plot(ALL_exp, col=kc$cluster)



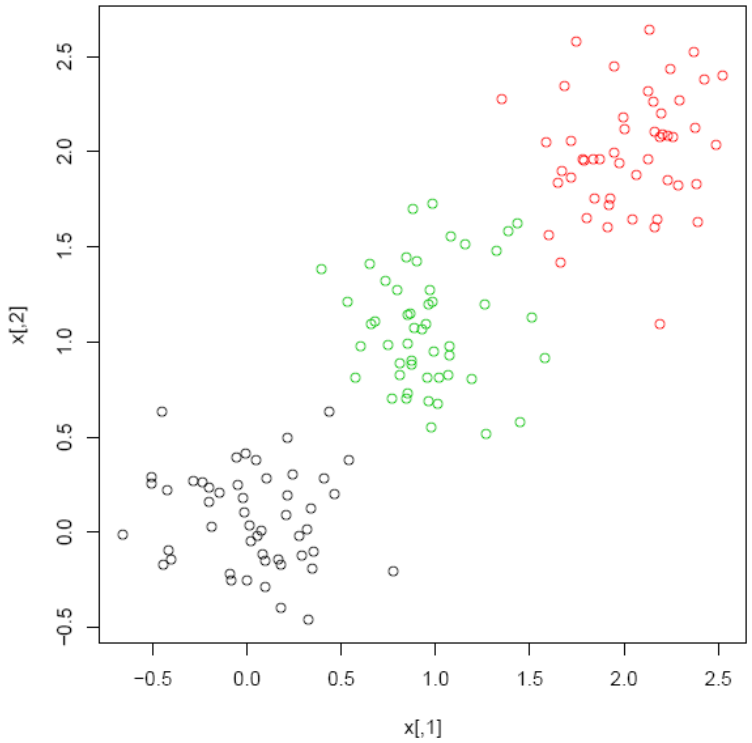
prefixed cluster number are 10

x,y-axis are the first two samples in ALL for all genes

Partitioning: K-means

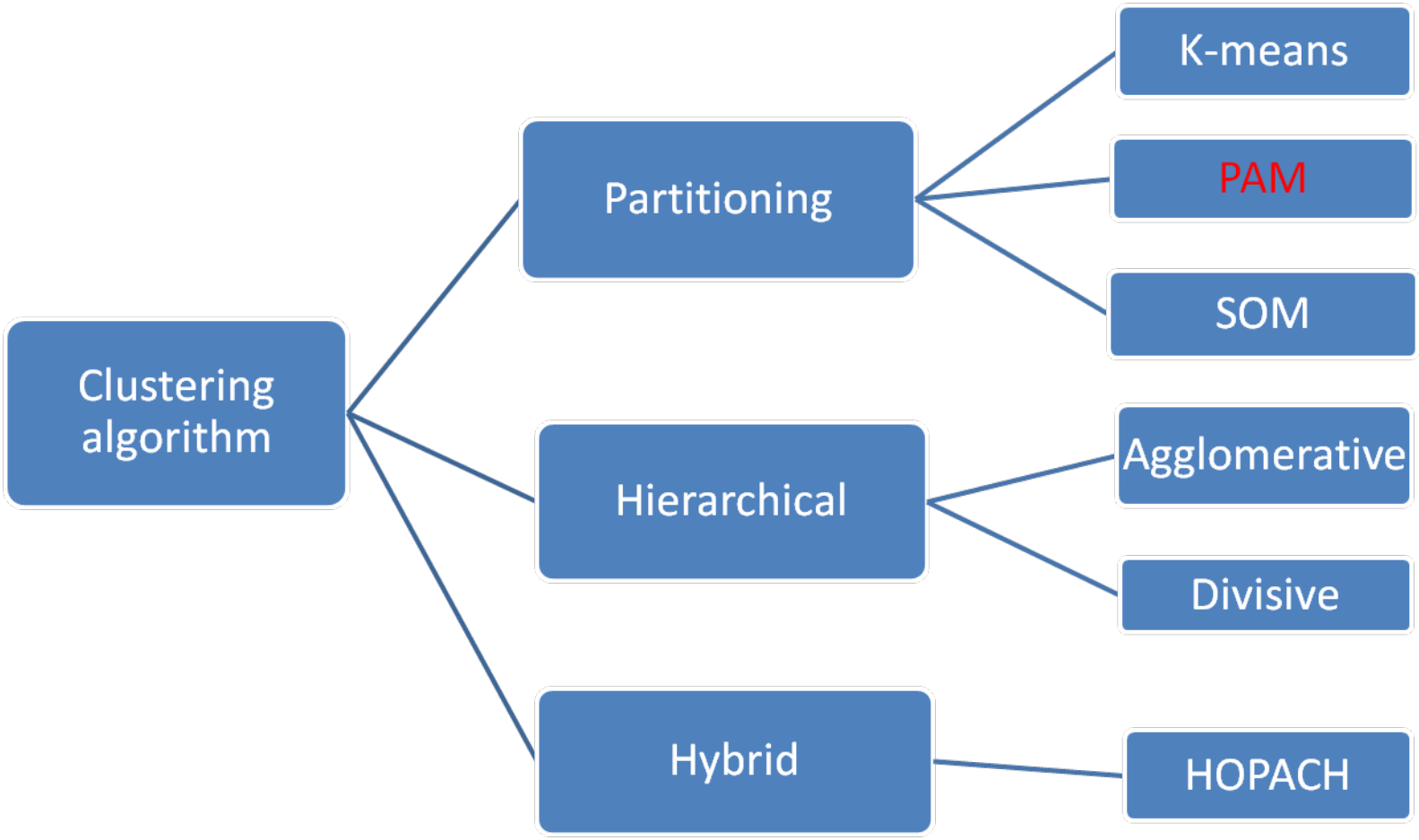


prefixed cluster number are 6



prefixed cluster number are 3

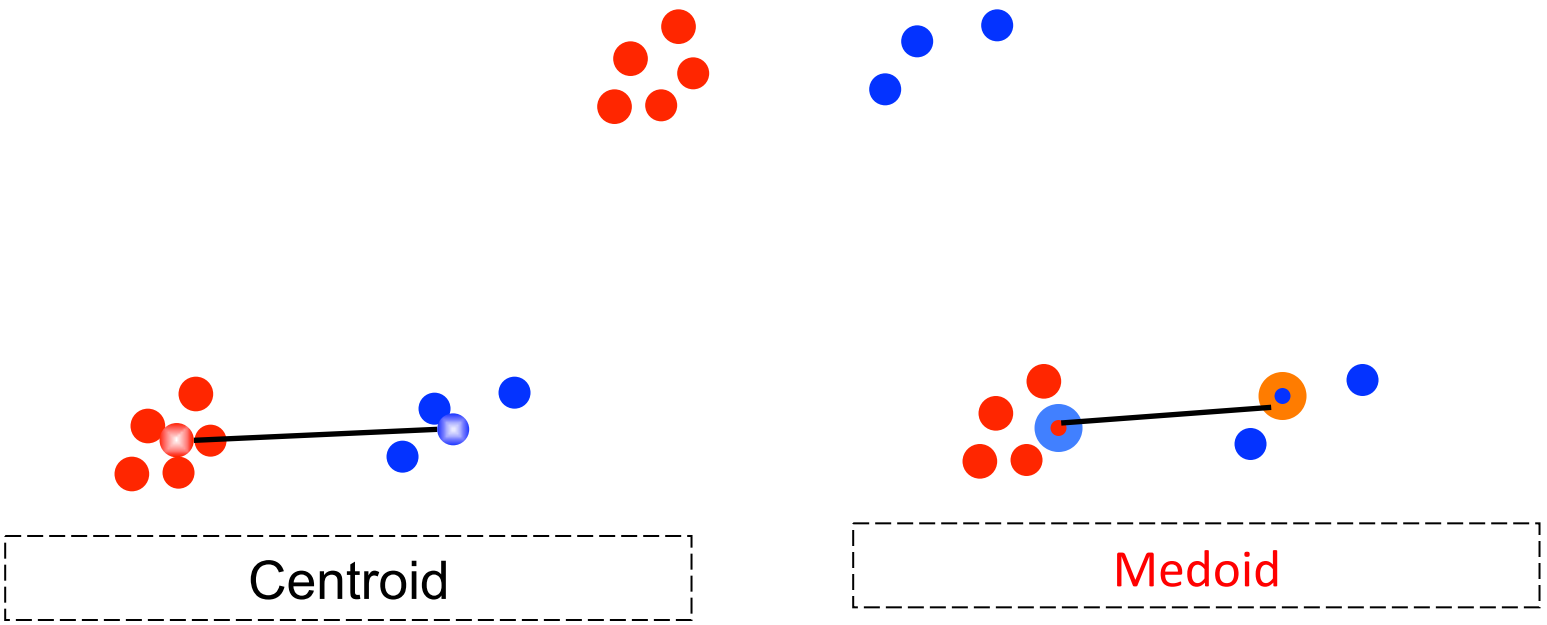
Type of Clustering algorithm



Partitioning: PAM

- PAM: Partitioning around medoids or k-medoids
- Mediod is the "representative point" within a cluster
  - It is different from "centroid" used by k-means, which is the average of the samples within a cluster
  - For example, medoid can be a point which has the smallest sum distance to all other points within the cluster
- The iterative procedure is analogous the one in *K*-means clustering

Clustering: cluster similarity



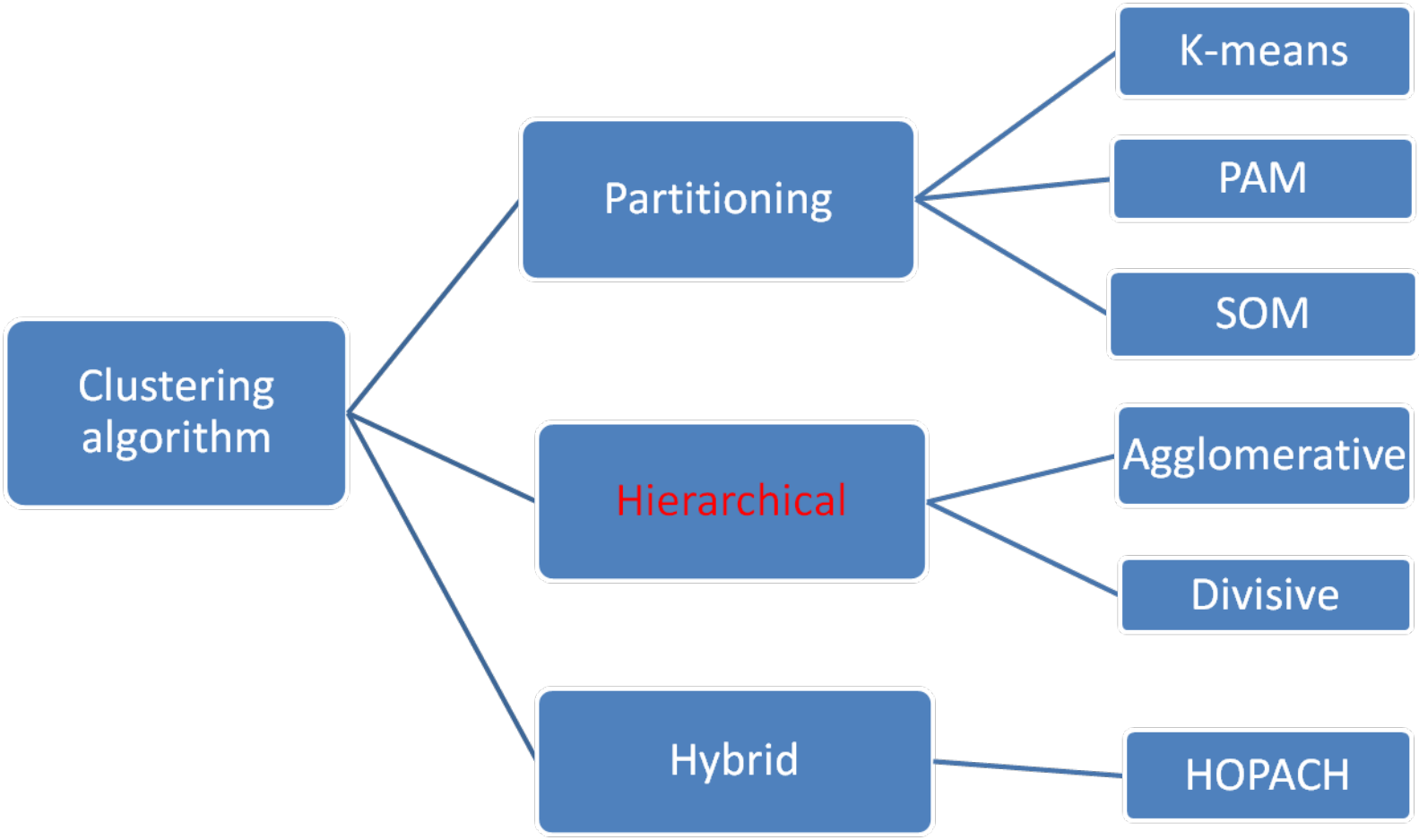
Partitioning: PAM

- Use “cluster” package in R
- http://cran.r-project.org/web/packages/cluster/index.html

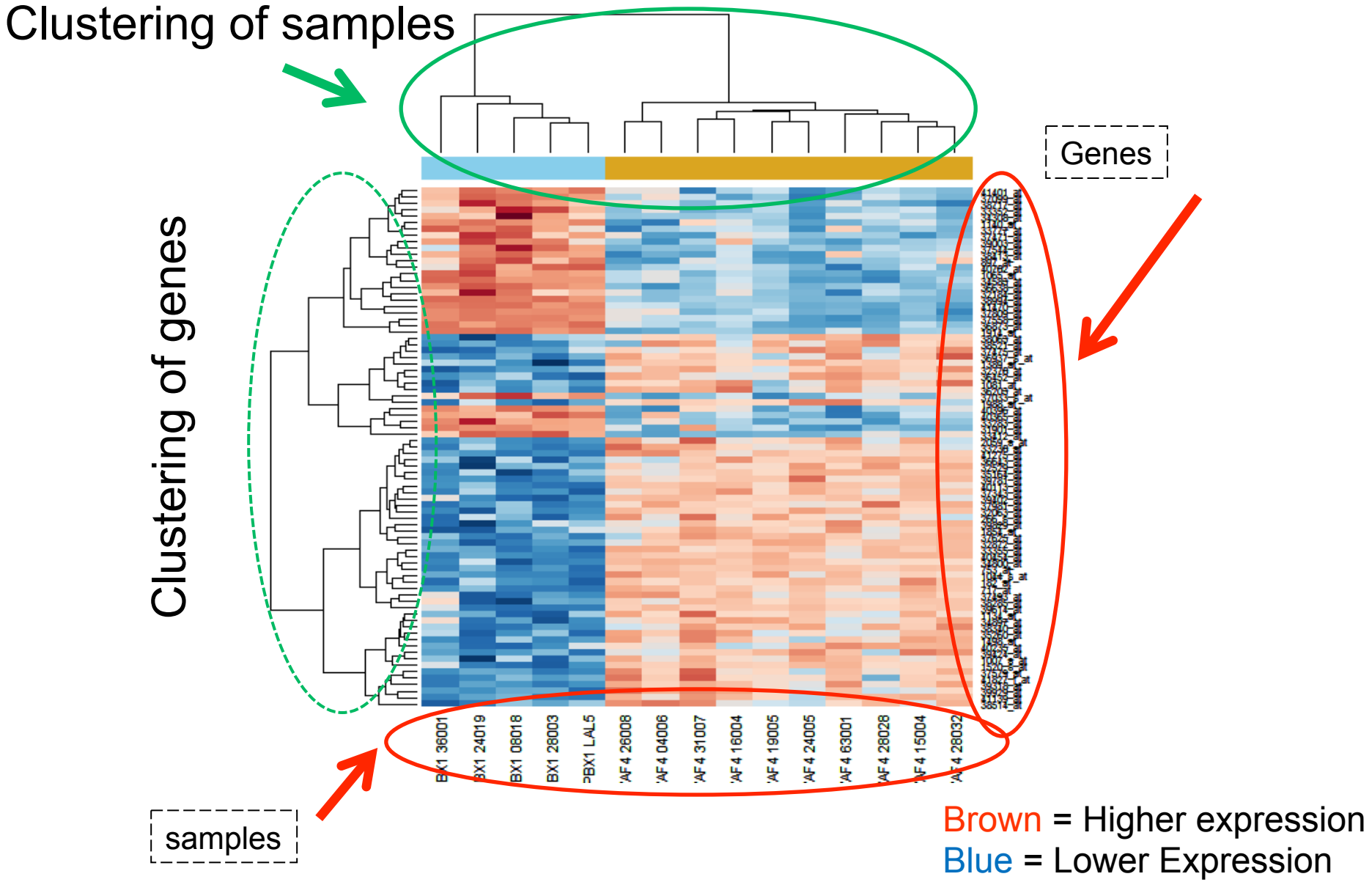
```
> source("http://bioconductor.org/biocLite.R")
> biocLite(“cluster")

> library(cluster)
> ALL_exp=exprs(ALL)
> kc=pam(ALL_exp,10,metric = "euclidean”)
```

Distance (metric) : euclidean or manhattan

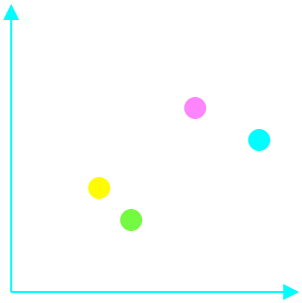
Type of Clustering algorithm



Dendrogram: Hierarchical Clustering



Agglomerative Hierarchical clustering



calculate distance matrix

	gene 1	gene 2	gene 3	gene 4
gene 1	0			
gene 2	2	0		
gene 3	8	7	0	
gene 4	10	12	4	0

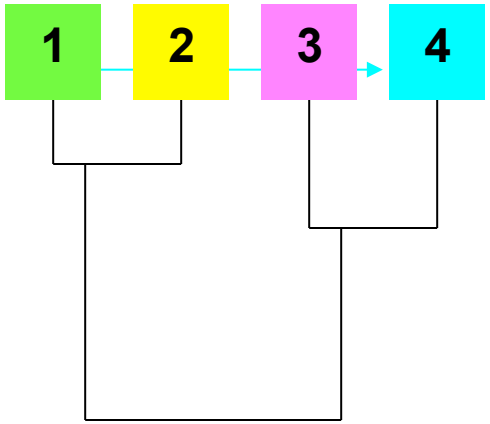
calculate averages of most similar

	gene 1,2	gene 3	gene 4
gene 1,2	0		
gene 3	7.5	0	
gene 4	11	4	0

calculate averages of most similar

	gene 1,2	gene 3,4
gene 1,2	0	
gene 3,4	9.25	0

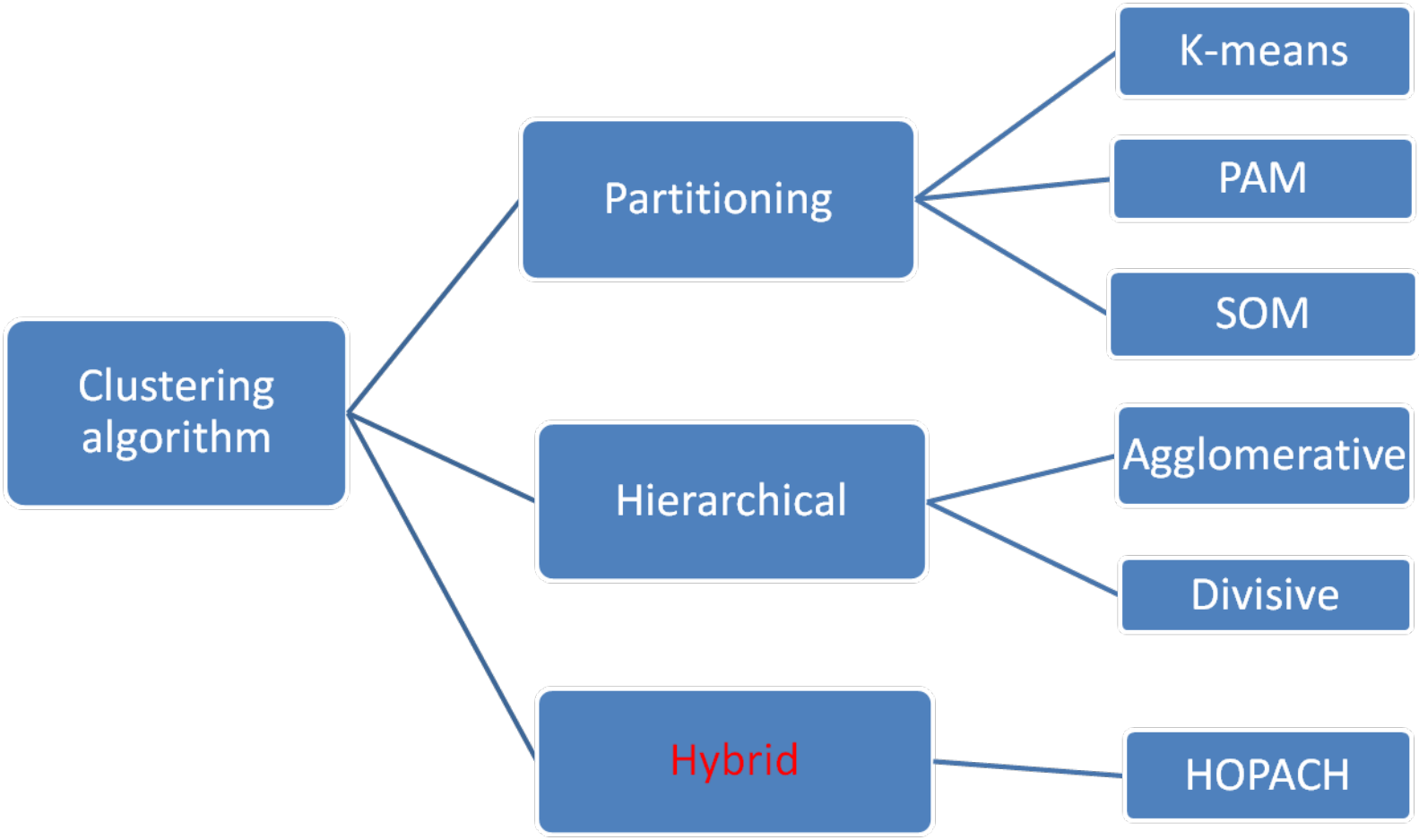
Dendrogram



Hierarchical clustering

```
> ALL_exp=exprs(ALL)

> dd <- dist(t(ALL_exp), method="euclidean")
> hc <- hclust(dd, method="centroid")
 or
> dd <- as.dist((1 – cor(ALL_exp))/2)
> hc <- hclust(dd, method="centroid")
> plot(hc)
```

Type of Clustering algorithm



HOPACH in R

http://cran.r-project.org/web/packages/hopach/index.html

> source("http://bioconductor.org/biocLite.R")

> biocLite(”hopach”)

> library(hopach)

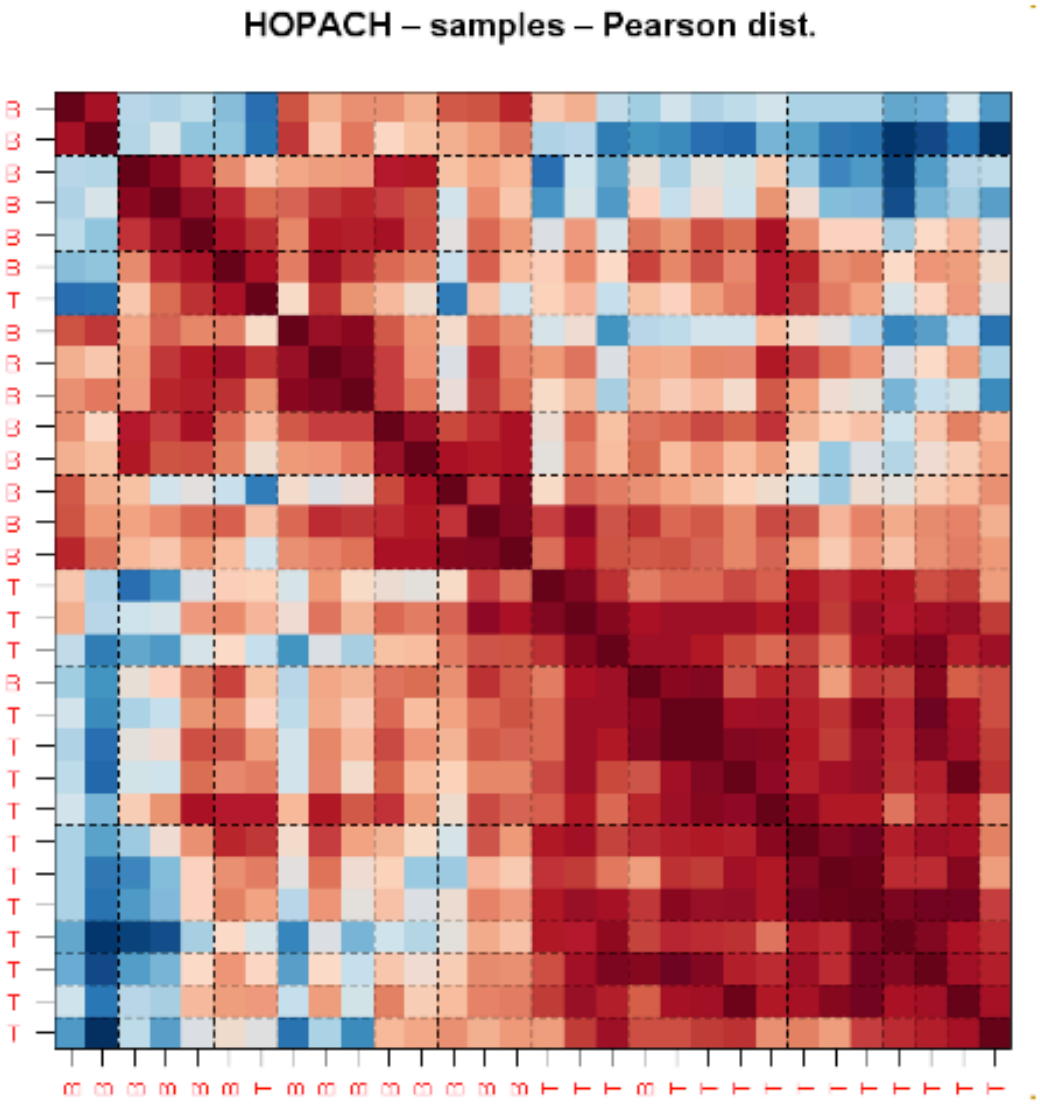
> ALL_exp=exprs(ALL)

> hc=hopach(ALL_exp, d = "cor”)

d= "cosangle”, "abscosangle”, "euclid”, "abseuclid”, "cor”, and "abscor”.

HOPACH in R

>dplot(ALL_exp[1:100,], kc)



- Heatmap shows the distance between genes
- Dotted lines: “breaks” between clusters

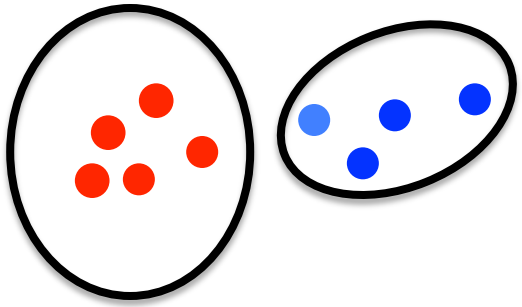
Fuzzy clustering

- In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster.

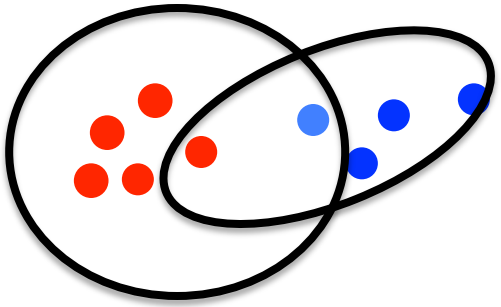
- In **fuzzy clustering** (also referred to as **soft clustering)**, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster.

- Fuzzy clustering can assign data elements to one or more clusters.

Fuzzy clustering



Hard clustering



Fuzzy Clustering

Fuzzy C-Means Algorithm

fanny() computes a fuzzy clustering of the data into k clusters.

```
> library(cluster)
> ALL_exp=exprs(ALL)
> fannyx <- fanny(ALL_exp, 3, memb.exp =2)
```

Midterm

- Will be posted on Friday 10/30.
- Midterm Exam is due by 11/11, Sunday, 11:59PM. Late submission is not accepted.
- Open book
- You can ask me, but cannot discuss with any other people.
- Including some topics, such as DESeq/edgeR, multiple test, enrichment test etc.