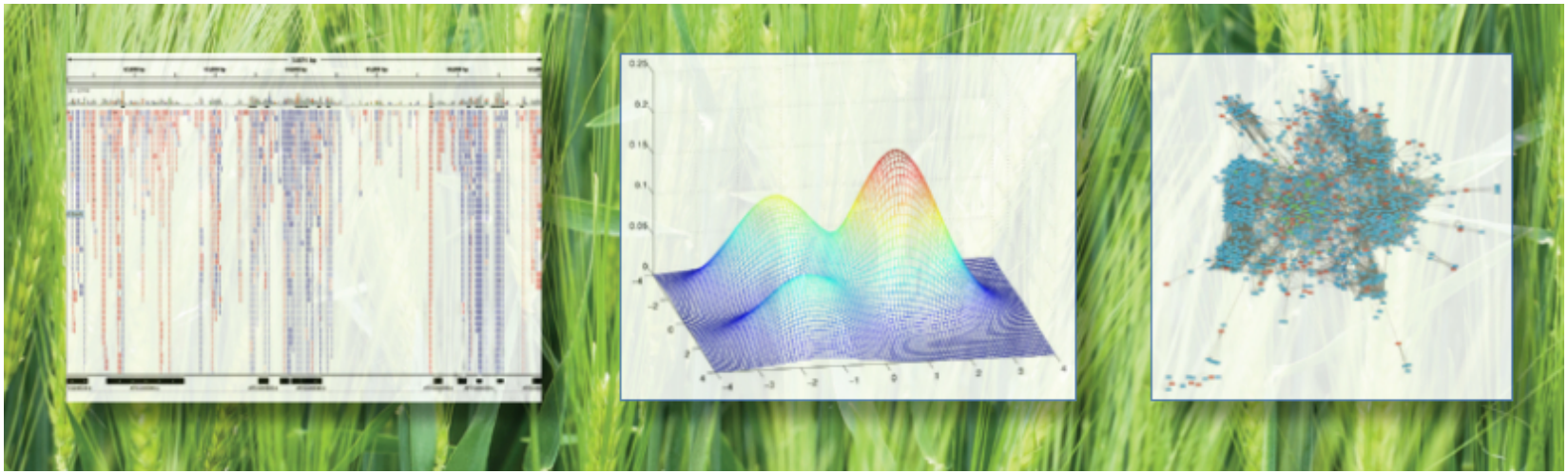
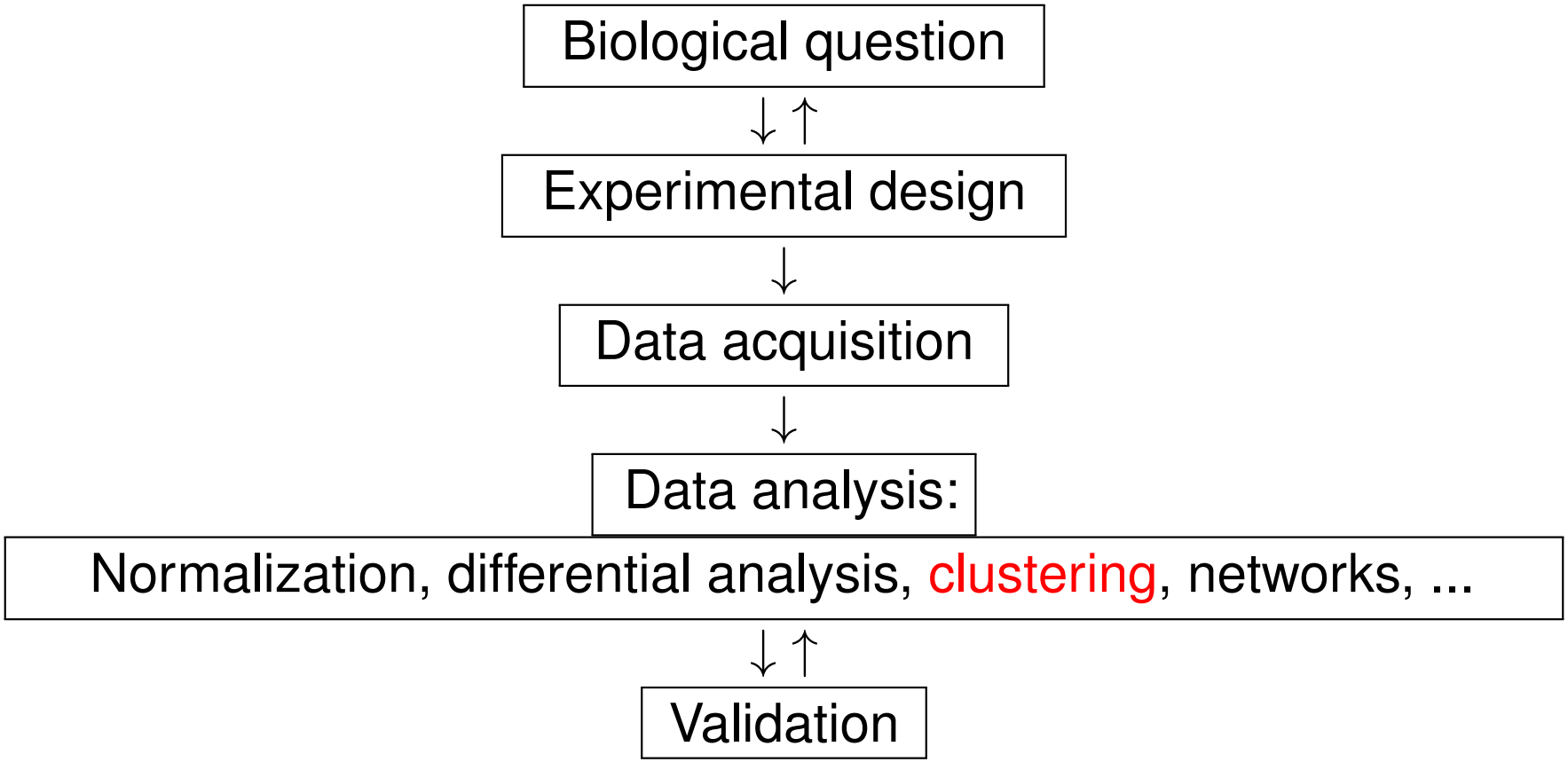
Co-expression analysis

Etienne Delannoy & Marie-Laure Martin-Magniette & Andrea Rau



Outline

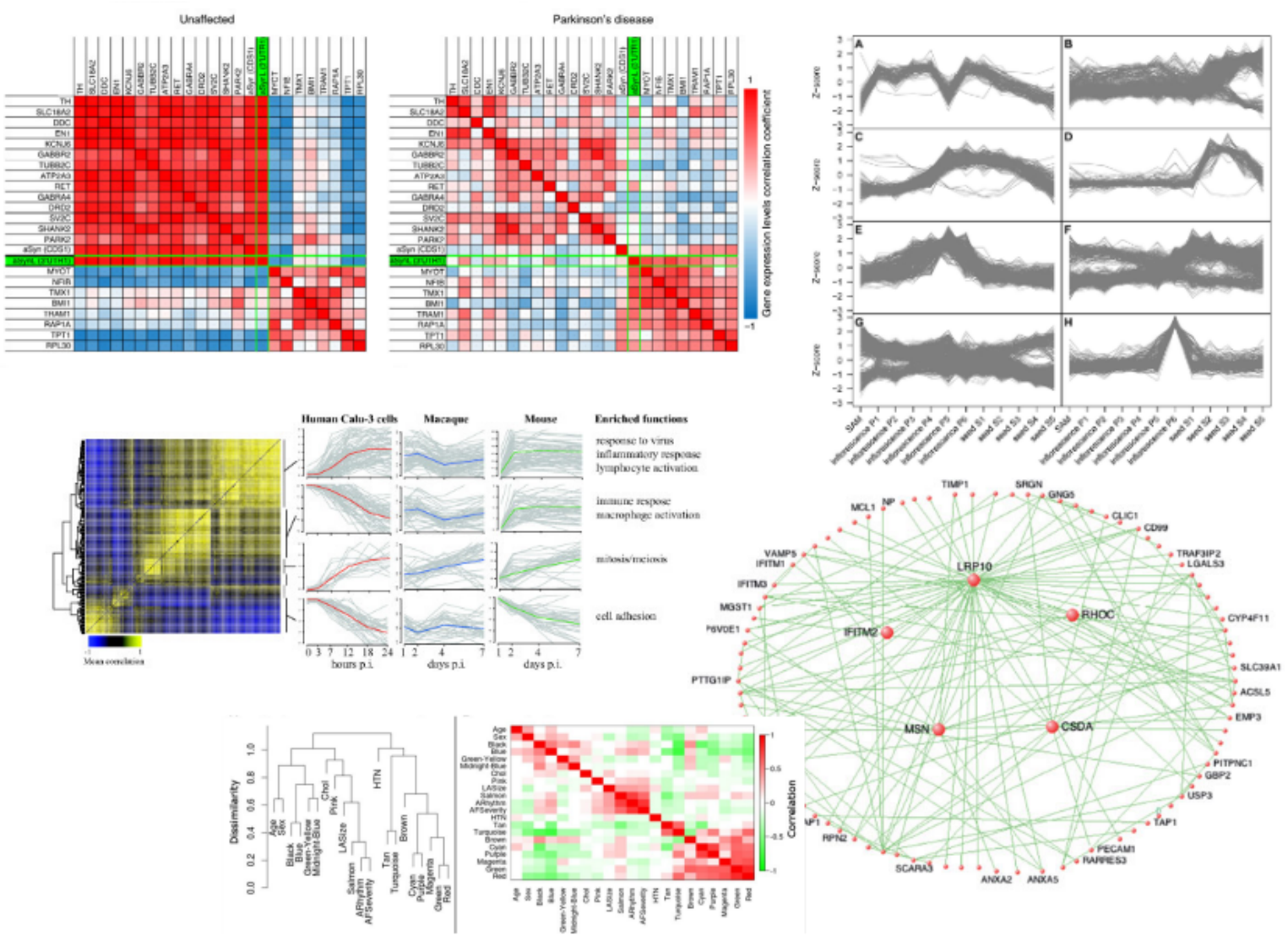
Design of a transcriptomic project



Aims for this talk

- What is the biological/statistical meaning of co-expression analysis?
- What methods exist for performing co-expression analysis?
- How to choose the number of clusters present in data?
- Advantages / disadvantages of different approaches: speed, stability, robustness, interpretability, model selection, ...

Gene co-expression[1]



[1] Google image search: “Coexpression”

Gene co-expression is...

- The simultaneous expression of two or more genes[2]
- Groups of co-transcribed genes[3]
- Similarity of expression[4] (correlation, topological overlap, mutual information, ...)
- Groups of genes that have similar expression patterns[5] over a range of different experiments

[2] https://en.wiktionary.org/wiki/coexpression
[3] http://bioinfow.dep.usal.es/coexpression
[4] http://coxpresdb.jp/overview.shtml
[5] Yeung *et al.* (2001)
[6] Eisen *et al.* (1998)

Gene co-expression is...

- The simultaneous expression of two or more genes[2]
- Groups of co-transcribed genes[3]
- Similarity of expression[4] (correlation, topological overlap, mutual information, ...)
- Groups of genes that have similar expression patterns[5] over a range of different experiments

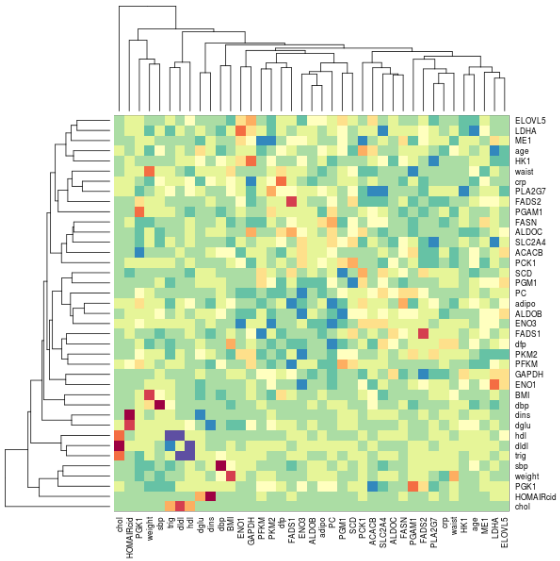
- Related to shared regulatory inputs, functional pathways, and biological process(es)[6]

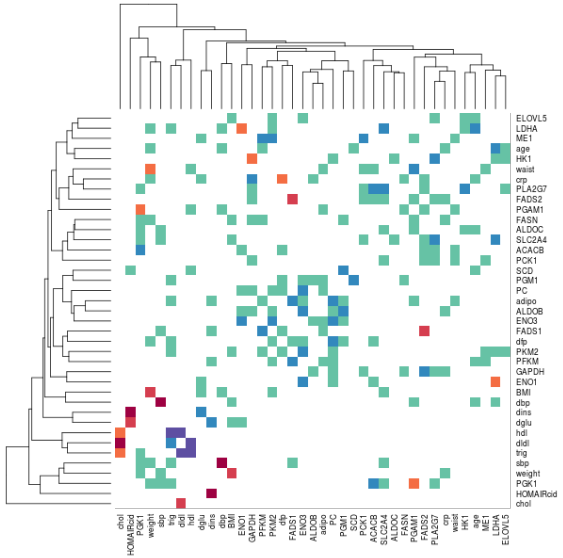
---

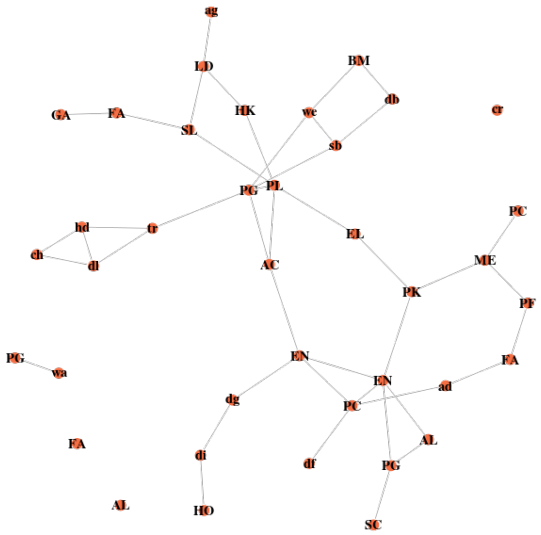
[2] https://en.wiktionary.org/wiki/coexpression
[3] http://bioinfow.dep.usal.es/coexpression
[4] http://coxpresdb.jp/overview.shtml
[5] Yeung *et al.* (2001)
[6] Eisen *et al.* (1998)

Using *correlations*: relevance network [7]

First (naive) approach: calculate correlations between expressions for all pairs of genes, threshold the smallest ones and build the network.







"Correlations" Thresholding Graph

[7] Butte and Kohane (1999,2000)

1 Introduction

2 **Unsupervised clustering**
- Distance-based clustering
- Model-based clustering
- Conclusion / discussion

3 Mixture models for transcriptomic data

Unsupervised clustering

Objective

Define homogeneous and well-separated groups of genes from transcriptomic data

What does it mean for a pair of genes to be close?
Given this, how do we define groups?

Unsupervised clustering

Objective

Define homogeneous and well-separated groups of genes from transcriptomic data

What does it mean for a pair of genes to be close?
Given this, how do we define groups?

Two broad classes of methods typically used:

1. Distance-based clustering (hierarchical clustering and K-means)
2. Model-based clustering (mixture models)

Hierarchical clustering analysis (HCA)

Objective Construct embedded partitions of $(n, n-1, \ldots, 1)$ groups, forming a tree-shaped data structure (dendrogram)

Algorithm

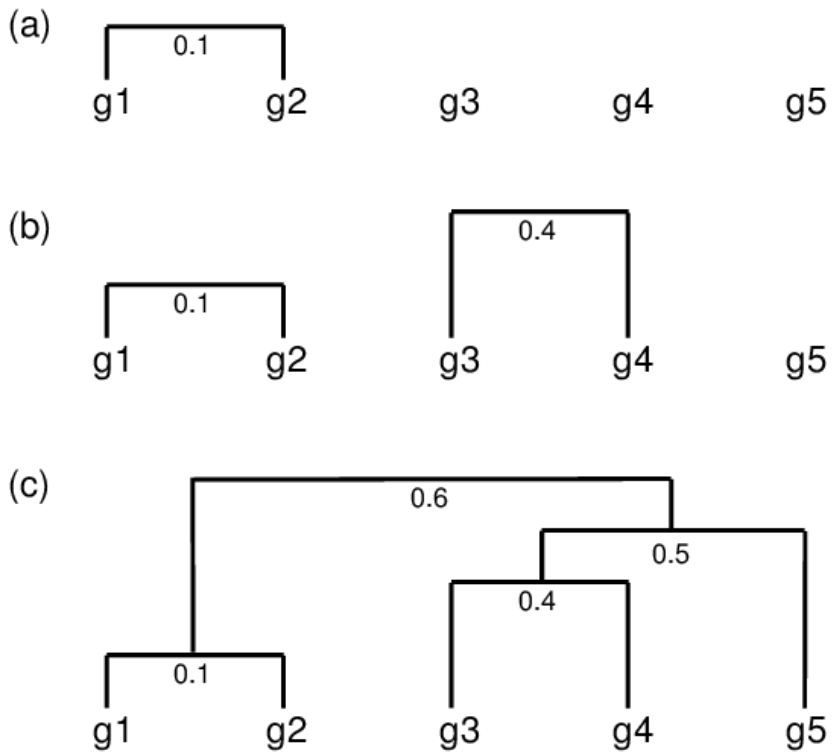
- **Initialization** n groups for n genes
- **At each step:**
  - Closest genes are clustered
  - Calculate distance between this new group and the remaining genes

Hierarchical clustering analysis (HCA)

Objective Construct embedded partitions of $(n, n-1, \ldots, 1)$ groups, forming a tree-shaped data structure (dendrogram)

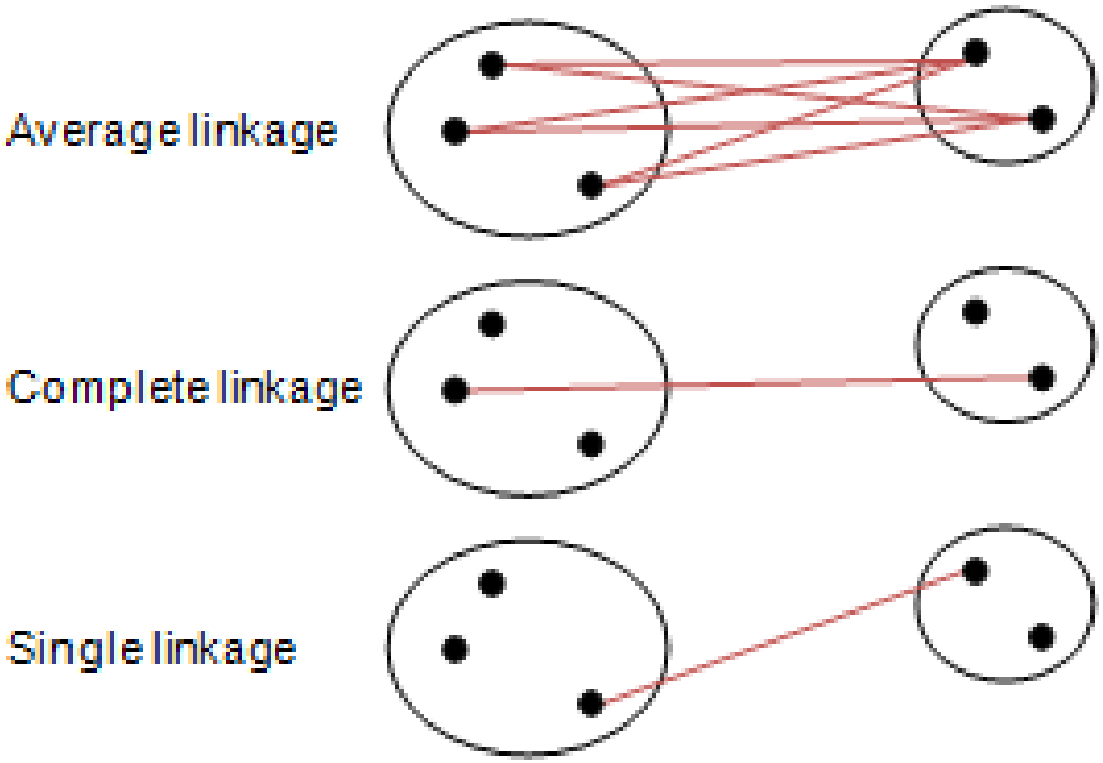
Algorithm

- **Initialization** n groups for n genes
- **At each step:**
  - Closest genes are clustered
  - Calculate distance between this new group and the remaining genes



Source: http://girke.bioinformatics.ucr.edu/GEN242/mydoc_Rclustering_3.html

Distances between groups for HCA



Source: https://www.multid.se/genex/onlinehelp/hs515.htm

Distances between groups for HCA

Average-linkage clustering	$\frac{\sum_{y \in C_k} \sum_{y' \in C_{k'}} d(y,y')}{\lvert C_k \rvert \lvert C_{k'} \rvert}$
Complete-linkage clustering	$\max_{y \in C_k} \max_{y' \in C_{k'}} d(y, y')$
Single-linkage clustering	$\min_{y \in C_k} \min_{y' \in C_{k'}} d(y, y')$
Ward distance d= Euclidian distance	$\sum_{y \in C_k \cup C_{k'}} d(y, y_{C_k \cup C_{k'}})$ $-\{\sum_{y \in C_k} d(y, y_{C_k}) + \sum_{y \in C_{k'}} d(y, y_{C_{k'}})\}$

Distance between genes: similarity measures

- **Manhattan distance**

$$\sum_{\ell=1}^{p} |y_{i\ell} - y_{i'\ell}|$$

Distance between genes: similarity measures

- **Manhattan distance**

$$\sum_{\ell=1}^{p} |y_{i\ell} - y_{i'\ell}|$$

- **Euclidian distance**

$$d^2(\mathbf{y}_i, \mathbf{y}_{i'}) = \sum_{\ell=1}^{p} (y_{i\ell} - y_{i'\ell})^2$$

$\Rightarrow$ Sensitive to scaling and differences in average expression level

Distance between genes: similarity measures

- **Manhattan distance**

$$\sum_{\ell=1}^{p} |y_{i\ell} - y_{i'\ell}|$$

- **Euclidian distance**

$$d^2(\mathbf{y}_i, \mathbf{y}_{i'}) = \sum_{\ell=1}^{p} (y_{i\ell} - y_{i'\ell})^2$$

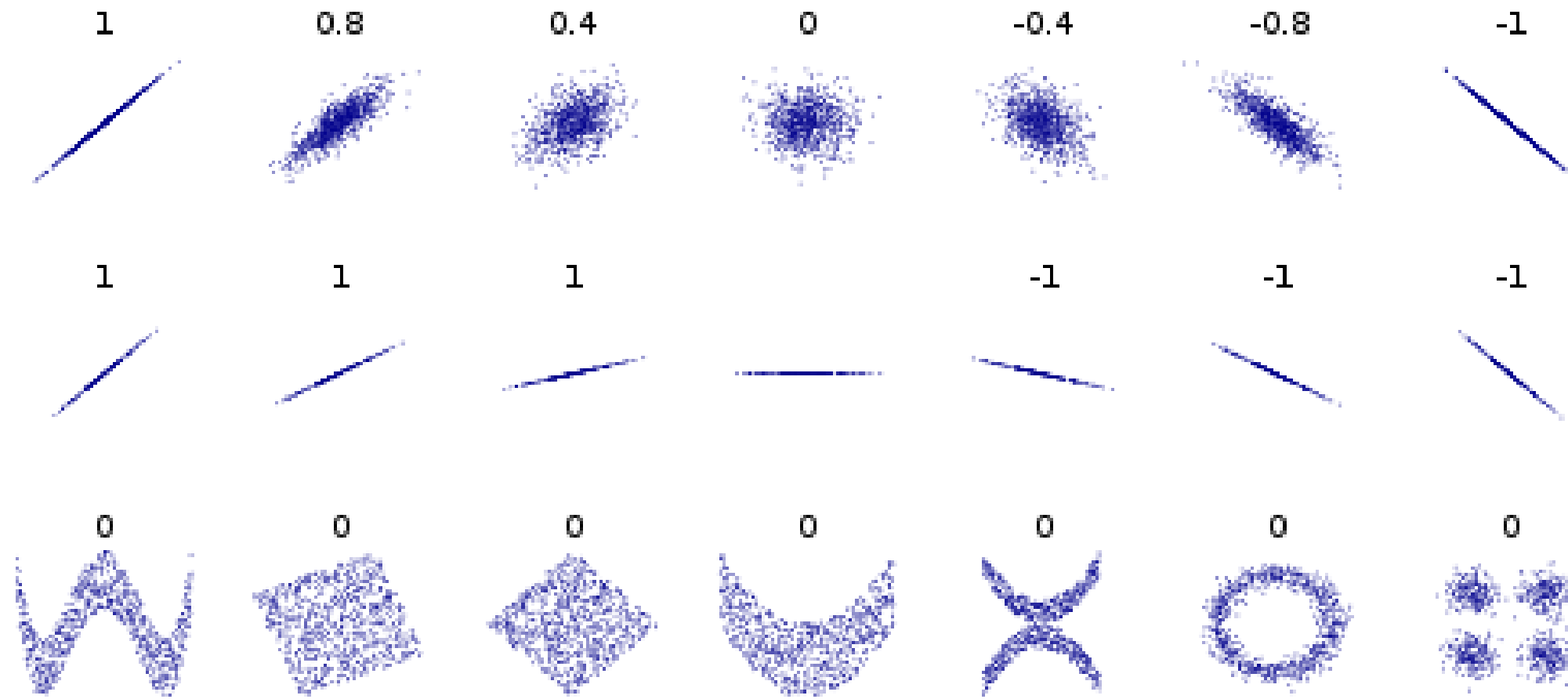
$\Rightarrow$ Sensitive to scaling and differences in average expression level

- **Pearson correlation distance**:

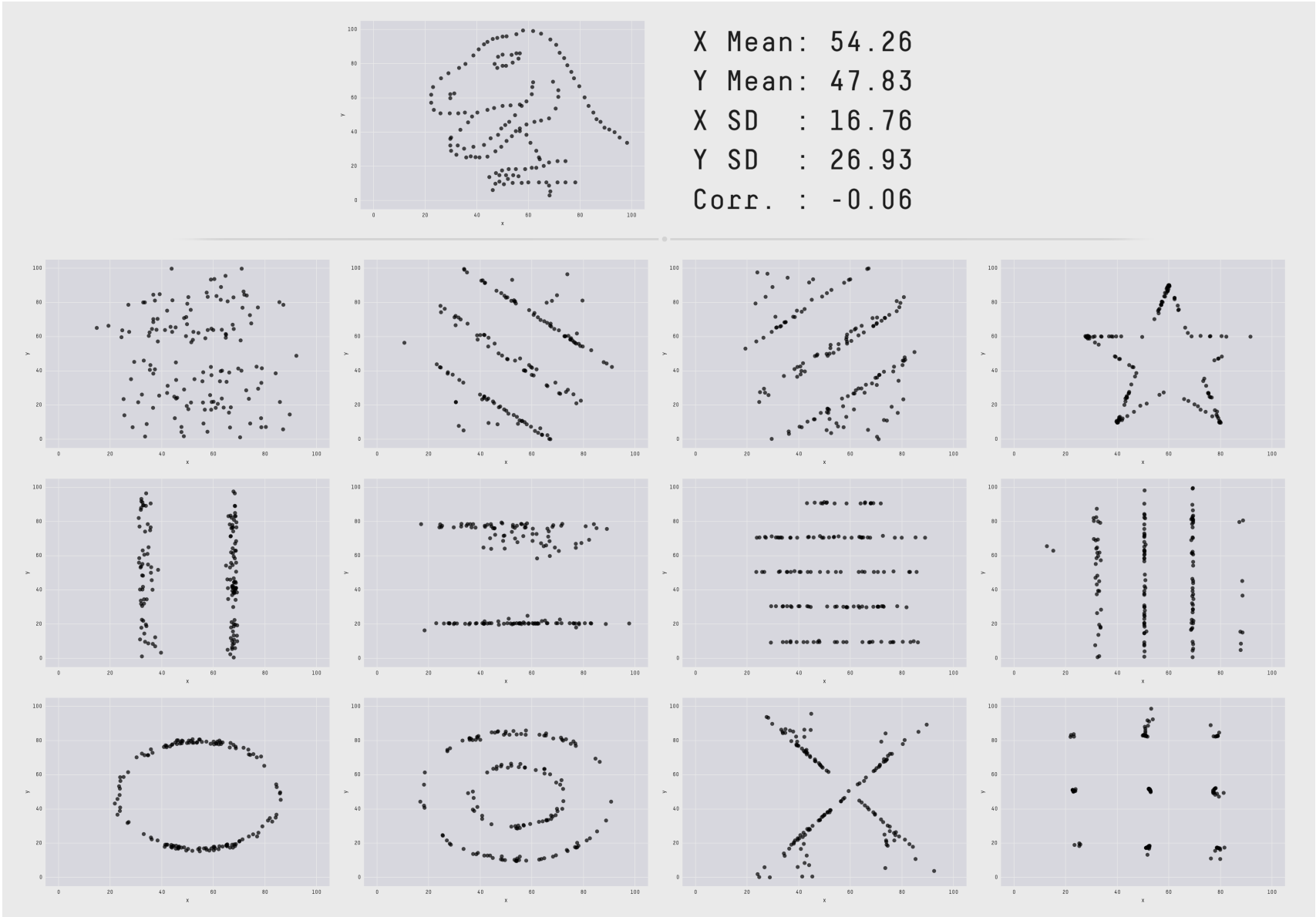
$$1 - \frac{Cov(\mathbf{y}_i, \mathbf{y}_{i'})}{\sigma(\mathbf{y}_i)\sigma(\mathbf{y}_{i'})}$$

$\Rightarrow$ Assessment of linear relationships

Examples of Pearson correlation values



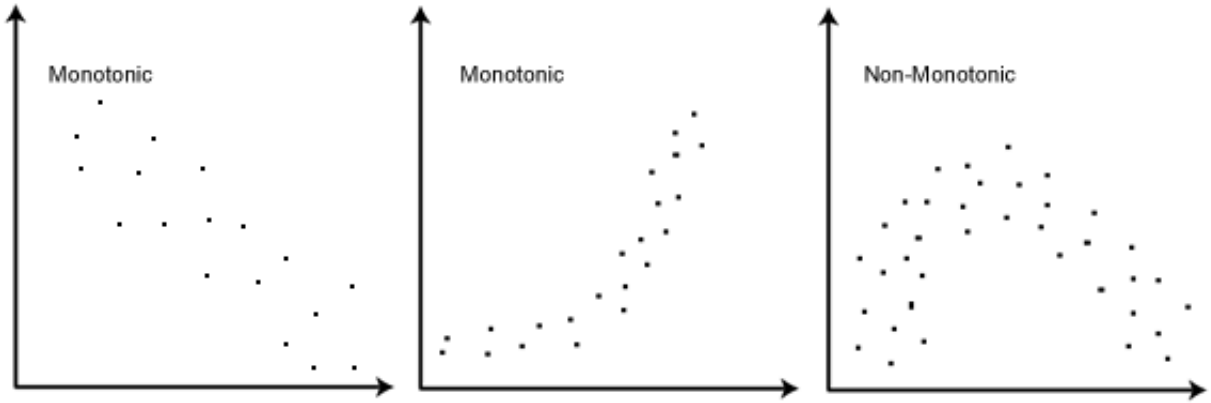
Examples of Pearson correlation values



Source: https://www.autodeskresearch.com/publications/samestats

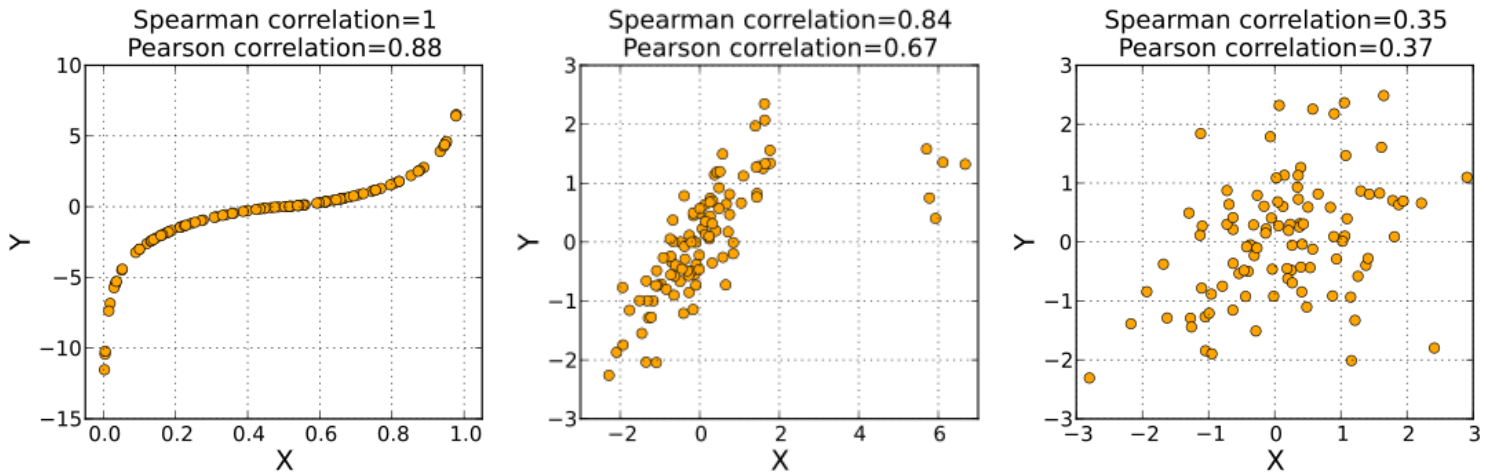
Similarity measures

- **Spearman correlation distance**: Pearson correlation distance between the rank values: y_{ij} replaced with rank of sample j across all samples
  $\Rightarrow$ Assessment of monotonic relationships (whether linear or not)



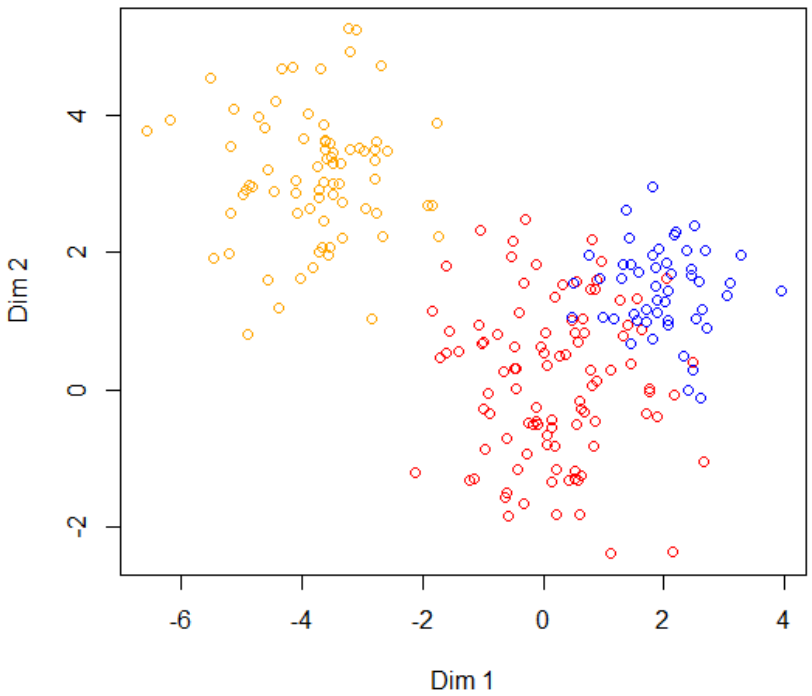
Similarity measures

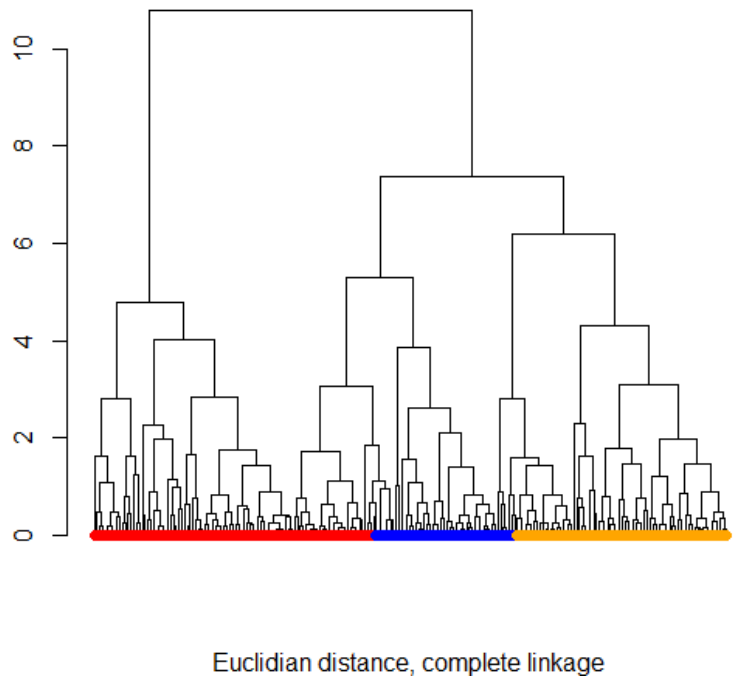
- **Spearman correlation distance**: Pearson correlation distance between the rank values: y_{ij} replaced with rank of sample j across all samples
  $\Rightarrow$ Assessment of monotonic relationships (whether linear or not)



HCA properties

- HCA is stable
- Results strongly depend on the chosen distances
- The number of clusters is chosen according to the tree
- Branch lengths are proportional to the percentage of inertia loss
  $\Rightarrow$ a long branch indicates that the 2 groups are not homogeneous





K-means algorithm

Initialization K centroids are chosen ramdomly or by the user

Iterative algorithm

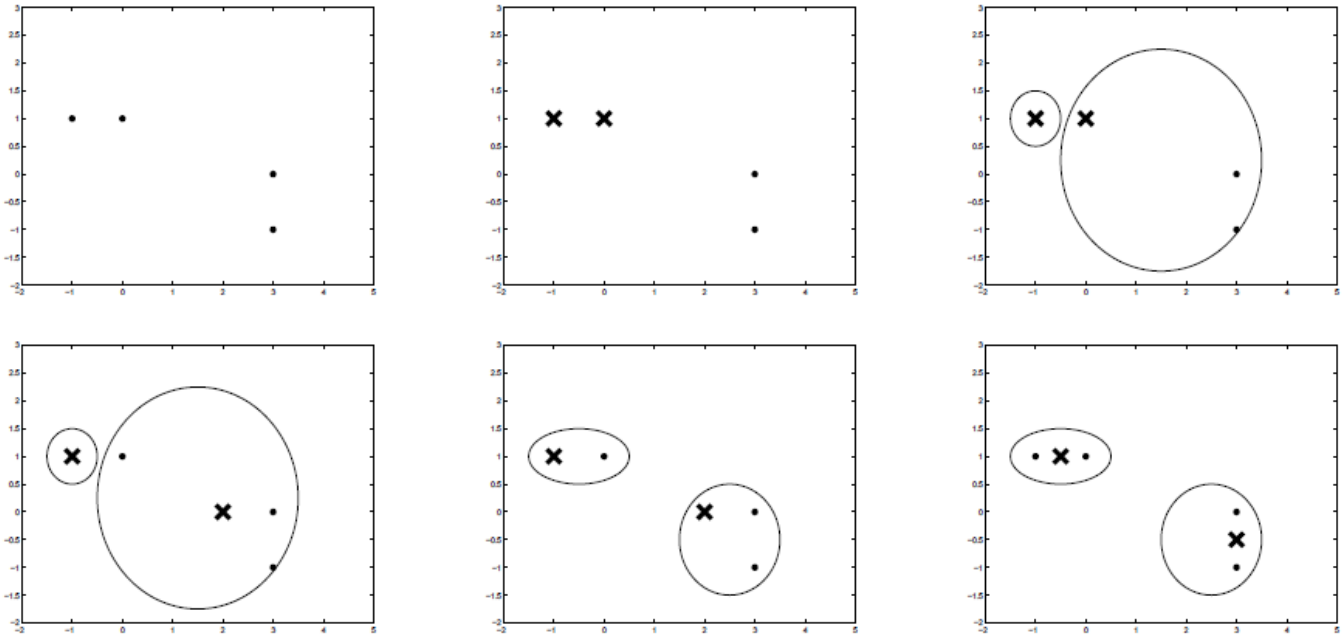
1. **Assignment** Each gene is assigned to a group according to its euclidian distance to the centroids.
2. **Calculation of the new centroids**

Stopping criterion: when the maximal number of iterations is achived OR when groups are stable

Properties

- Rapid and easy
- Results depend strongly on initialization
- Number of groups K is fixed a priori

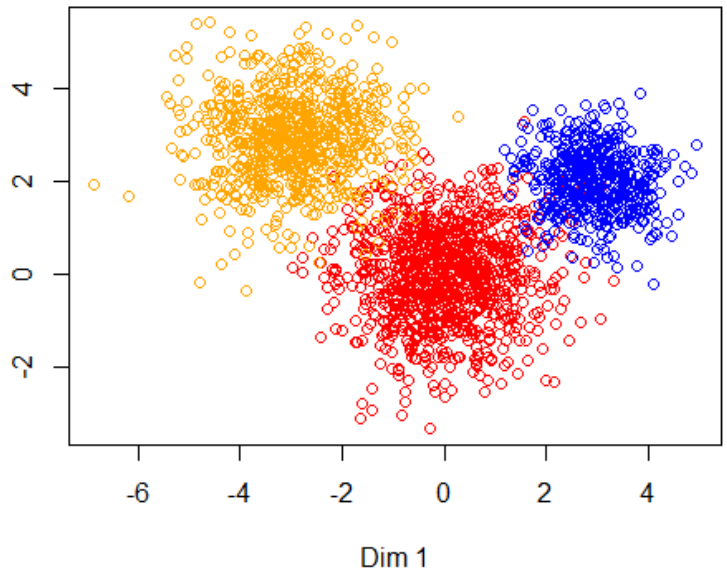
K-means illustration

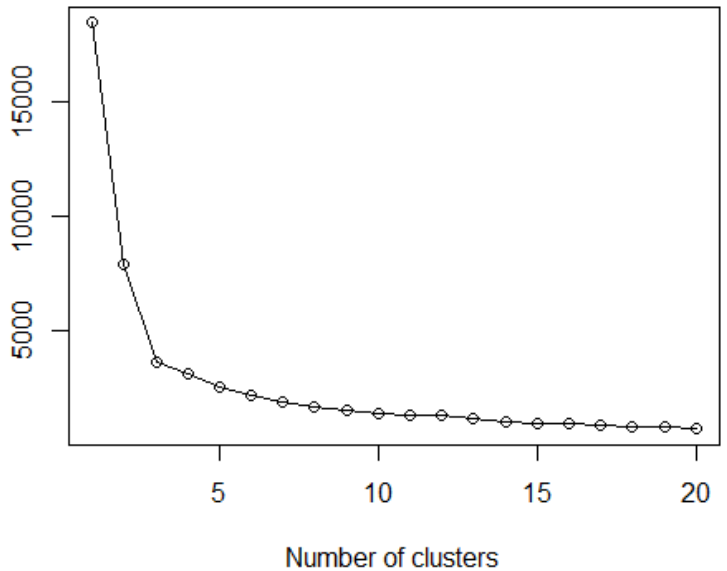


Animation: `http://shabal.in/visuals/kmeans/1.html`

K-means algorithm: Choice of K?

- **Elbow plot of within-sum of squares**: examine the percentage of variance explained as a function of the number of clusters

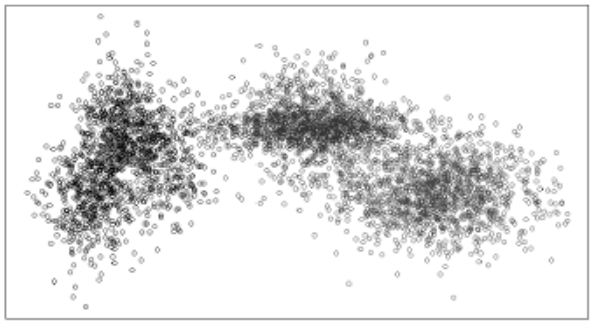




Model-based clustering

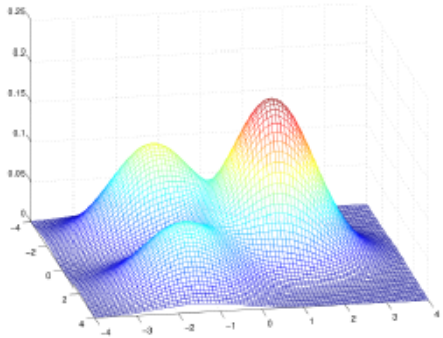
- Probabilistic clustering models : data are assumed to come from distinct subpopulations, each modeled separately
- Rigourous framework for parameter estimation and choice of the number of groups
- It assigns a probability of cluster membership for each observation

what we observe

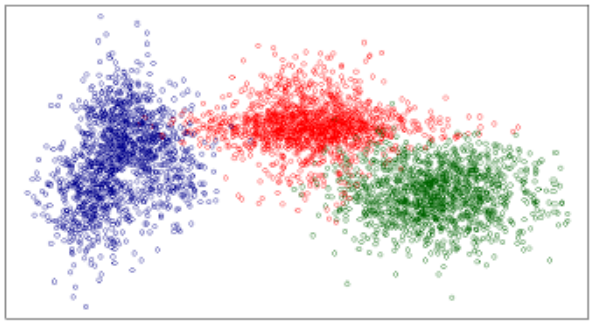


$Z = ?$

the model



the expected results



$Z : 1 = \bullet, 2 = \bullet, 3 = \bullet$

How to construct a mixture model?

- Distribution: what distribution to use for each group?
  $\rightsquigarrow$ depends on the observed data.
- Inference: how to estimate the parameters?
  $\rightsquigarrow$ usually done with an EM-like algorithm (Dempster *et al.*, 1977)
- Model selection: how to choose the number of groups?
  - A collection of mixtures with **a varying number of groups** is usually considered
  - A penalized criterion is used to select the best model from the collection

Key ingredients of a mixture model

- Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote the observations with $\mathbf{y}_i \in \mathbb{R}^p$ and $n >> p$
- We introduce a latent variable to indicate the group from which each observation arises:

$$\mathbf{Z} \sim \mathcal{M}(n; \pi_1, \ldots, \pi_K), \quad \sum_{\ell=1}^{K} \pi_\ell = 1$$

$$P(Z_i = \ell) = \pi_\ell$$

- Assume that $\mathbf{y}_i$ are conditionally independent given $\mathbf{Z}$
- Model the distribution of $\mathbf{y}_i | Z_i$ using a parametric distribution:

$$(\mathbf{y}_i | Z_i = \ell) \sim f(\cdot; \theta_\ell)$$

Key ingredients of a mixture model

- Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote the observations with $\mathbf{y}_i \in \mathbb{R}^p$ and $n >> p$
- We introduce a latent variable to indicate the group from which each observation arises:

$$\mathbf{Z} \sim \mathcal{M}(n; \pi_1, \ldots, \pi_K), \quad \sum_{\ell=1}^{K} \pi_\ell = 1$$

$$P(Z_i = \ell) = \pi_\ell$$

- Assume that $\mathbf{y}_i$ are conditionally independent given $\mathbf{Z}$
- Model the distribution of $\mathbf{y}_i | Z_i$ using a parametric distribution:

$$(\mathbf{y}_i | Z_i = \ell) \sim f(\cdot; \theta_\ell)$$

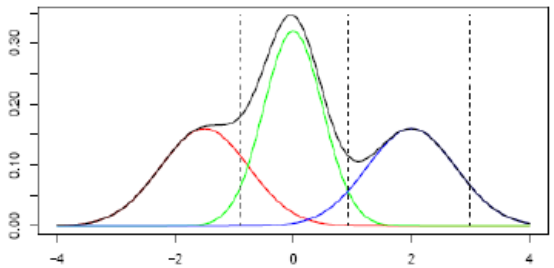
- After parameter estimation, calculate the conditional probabilities

$$\tau_{i\ell} = P(Z_i = \ell | \mathbf{y}_i)$$

Clustering data into groups

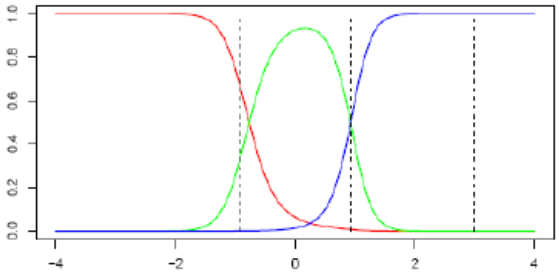
Distributions:

$$g(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \pi_3 f_3(x)$$



Conditional probabilities:

$$\tau_{ik} = \frac{\pi_k f_k(x_i)}{g(x_i)}$$



Maximum a posteriori (MAP) rule: Assign genes to the group with highest conditional probability:

τ_{ik} (%)	$k = 1$	$k = 2$	$k = 3$
$i = 1$	65.8	34.2	0.0
$i = 2$	0.7	47.8	51.5
$i = 3$	0.0	0.0	100
...	...	...	...

Model selection for mixture models

Asymptotic penalized criteria[8]

- BIC aims to identify the best model wrt the global fit of the data distribution:

$$BIC(k) = \log P(\mathbf{y}|k, \hat{\theta}_k) - \frac{D_k}{2} \log(n)$$

where D_k is the # of free parameters and $\hat{\theta}_k$ is the MLE of the model with k clusters

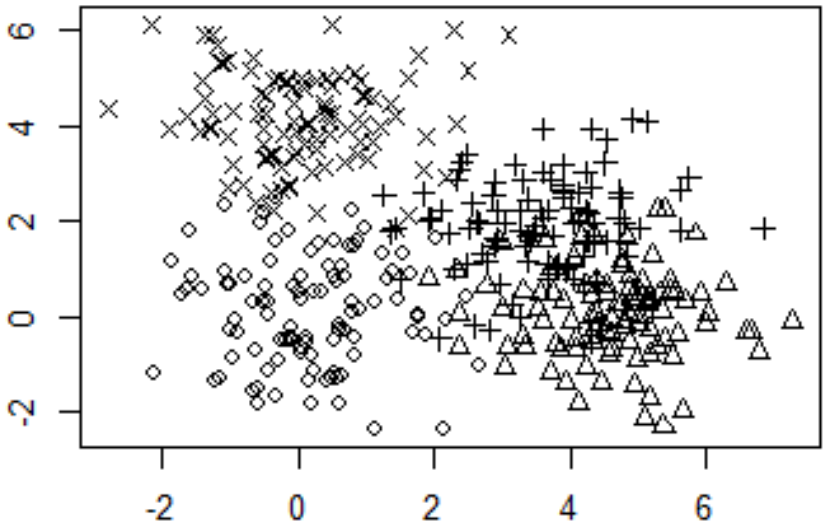
- ICL aims to identify the best model wrt cluster separation:

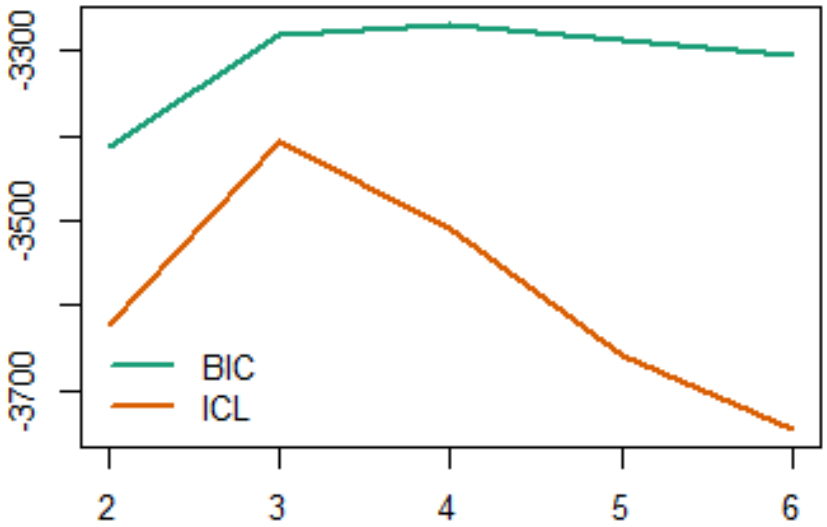
$$ICL(k) = BIC(k) - \left(- \sum_{i=1}^{n} \sum_{\ell=1}^{k} \tau_{i\ell} \log \tau_{i\ell} \right)$$

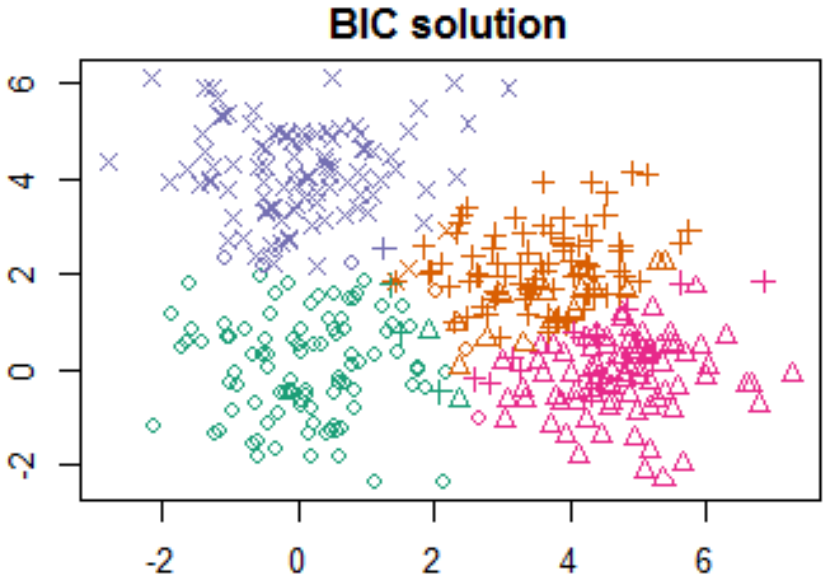
$\rightsquigarrow$ Select K that maximizes BIC or ICL (but be careful about their sign!)

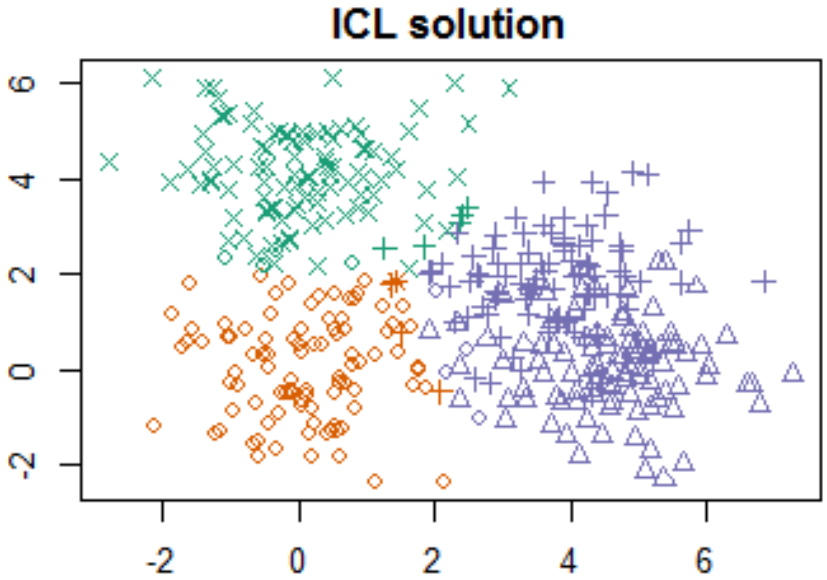
[8] Asymptotic: approaching a given value as the number of observations $n \to \infty$

Model selection for mixture models: BIC vs ICL







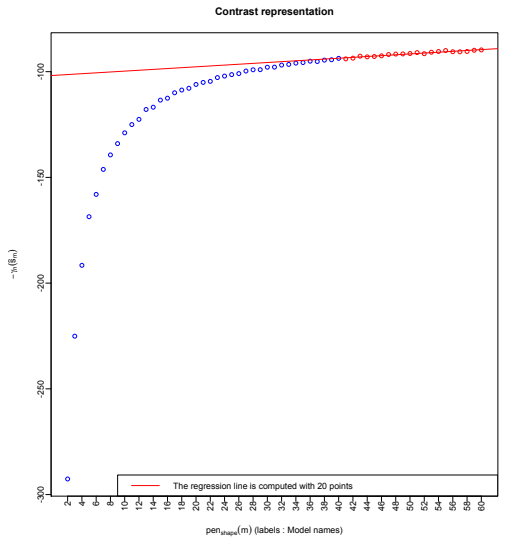


Model selection for mixture models

Non-asymptotic penalized criterion

The slope heuristics[9] is defined by $SH(k) = \log P(\mathbf{y}|k, \hat{\theta}_k) - \kappa \frac{D_k}{n}$

- In large dimensions, $\log P(\mathbf{y}|k, \hat{\theta}_k)$ must be linear in $\frac{D_k}{n}$
- Estimation of slope to calibrate κ in a data-driven manner [10]



$\rightsquigarrow$ Select K that maximizes SH(k)

[9] Birgé & Massart (2007)

[10] Data-Driven Slope Estimation (DDSE) available in `capushe` R package

A note about evaluating clustering approaches[11]

- Clustering results can be evaluated based on internal criteria (e.g., statistical properties of clusters) or external criteria (e.g., functional annotations)
  - **Adjusted Rand index**: measure of similarity between two data clusterings, adjusted for the chance grouping of elements
    $\rightsquigarrow$ ARI has expected value of 0 in the case of a random partition, and is bounded above by 1 in the case of perfect agreement
- Methods that give different results depending on the initialization should be rerun multiple times to check for stability
- Most clustering methods will find clusters even when no actual structure is present $\Rightarrow$ good idea to compare to results with randomized data!

[11] D'haeseller, 2005

1 Introduction

2 Unsupervised clustering

3 **Mixture models for transcriptomic data**

- For microarray data
- For RNA-seq data
- Conclusion / discussion

From mixture models to co-expression analysis

- Transcriptomic data: main source of 'omic information available for living organisms
  - Microarrays ($\sim$1995 -)
  - High-throughput sequencing: RNA-seq ($\sim$2008 -)

Co-expression (clustering) analysis

- Study patterns of relative gene expression (*profiles*) across several conditions
- $\Rightarrow$ Co-expression is a tool to study genes without known or predicted function (orphan genes)
- Exploratory tool to identify expression trends from the data ($\neq$ sample classification, identification of differential expression)

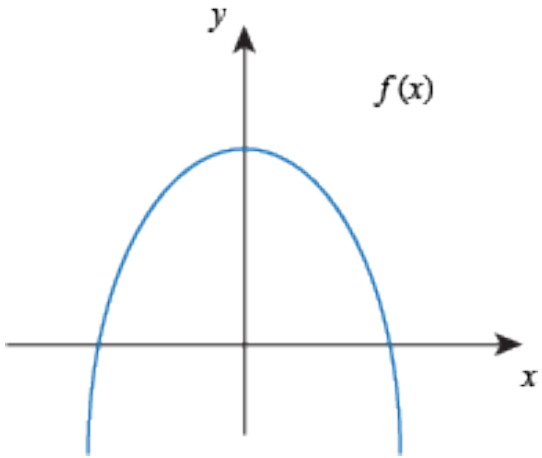
For microarray data

Development of model-based clustering **with variable selection** for Gaussian mixture models (Maugis et al., 2009a, 2009b, 2009c)

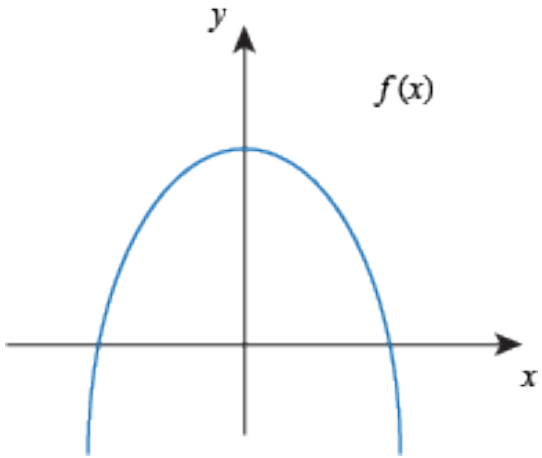
From Gene Expression Modeling to genomic Networks

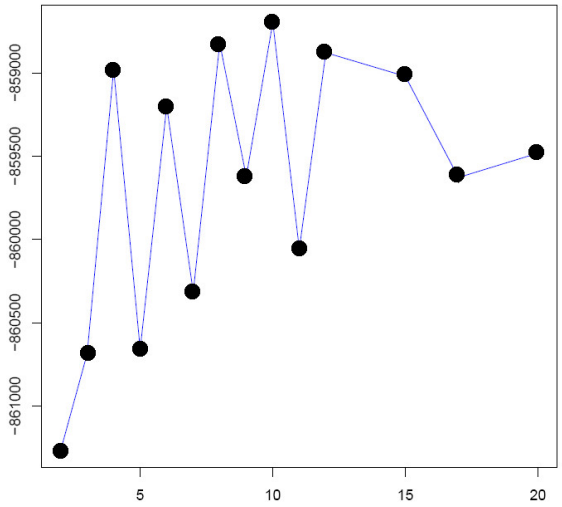
- Mixtures for 18,110 genes described by 387 expression differences between two conditions (stress/no stress), categorized in 18 types of stress.
- A new module of CATdb, for the integration of other sources of data and the visualization of all the results (Zaag *et al.*, 2015, NAR)
- Methodology also available for small datasets (Frei-dit Frey *et al.*, 2014, Genome Biology)

BIC used to create stress categories



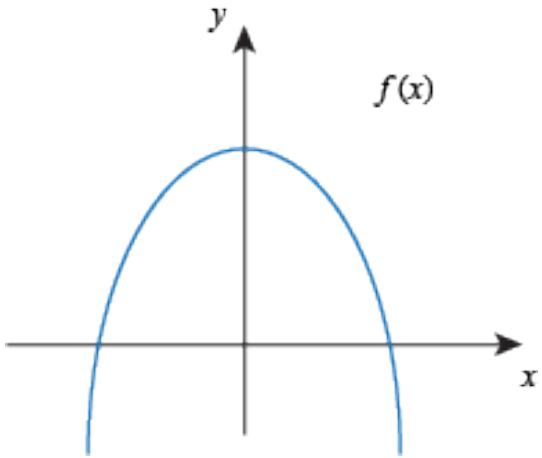
BIC used to create stress categories

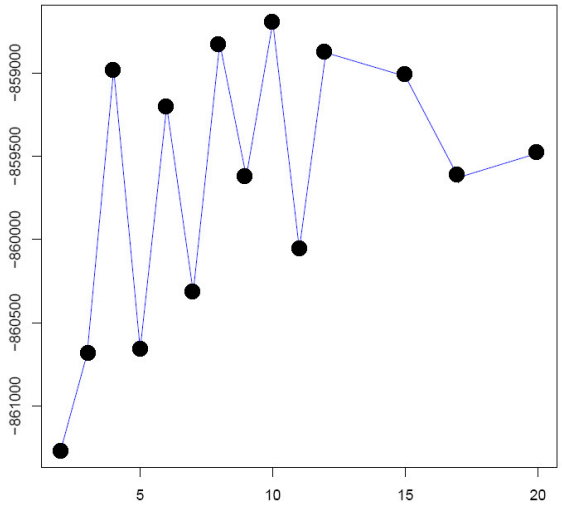




⇝ It suggests that several latent structures may exist

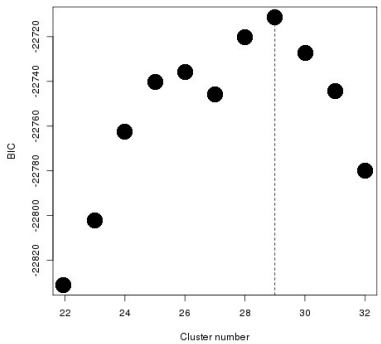
BIC used to create stress categories



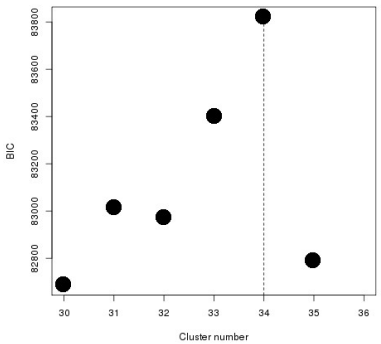


⇝ It suggests that several latent structures may exist

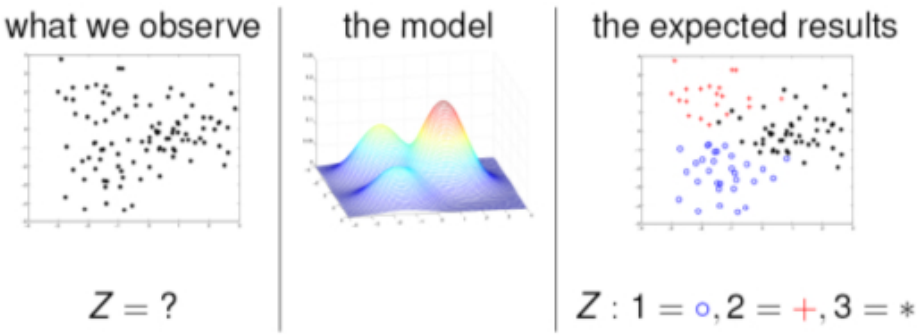
Nematode



Drought



Large scale co-expression study of Arabidopsis

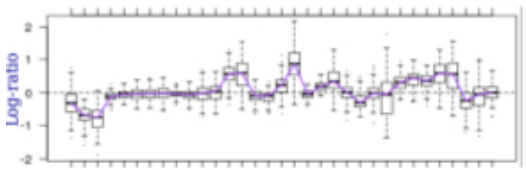


Matrix by stress
{ genes x log-ratios}

Gaussian mixture

Data-driven method

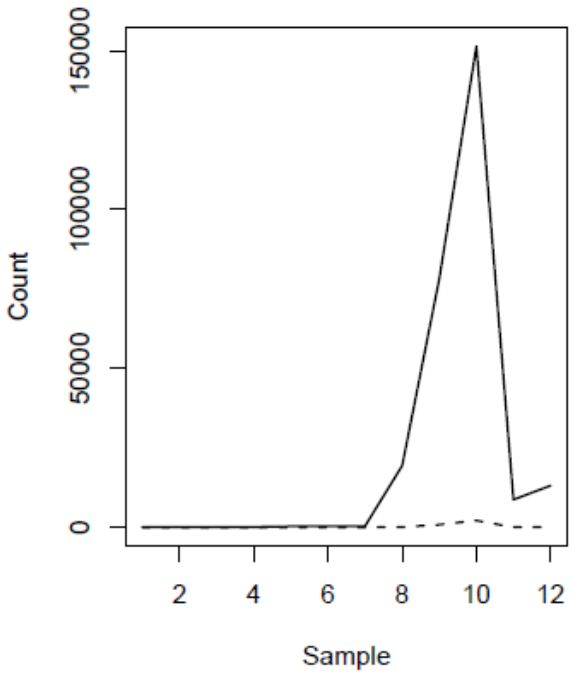
- number of cluster chosen by BIC
- gene classification based on the conditional probabilities



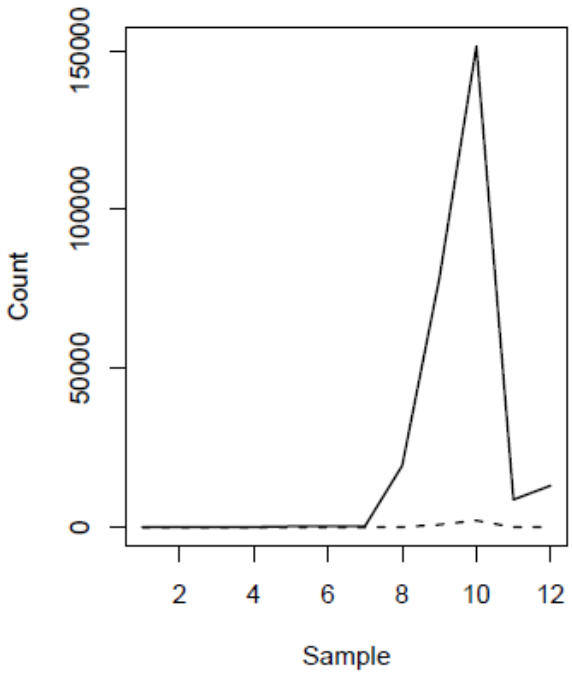
Stress category	Gene_nb	Cluster_nb
Nitrogen	13 495	59
Temperature	11 365	34
Drought	8 143	34
Salt	5 729	30
Heavy metal	10 617	57
UV	7 894	37
Gamma	5 350	32
Oxydative stress	10 127	52
Nectrophic bacteria	11 220	50
Biotrophic bacteria	12 023	56
Fungi	9 773	51
Rhodococcus	1 900	13
Oomycete	5 508	31
Nematode	7 413	27
Stifenia	1 525	17
Virus	11 832	54

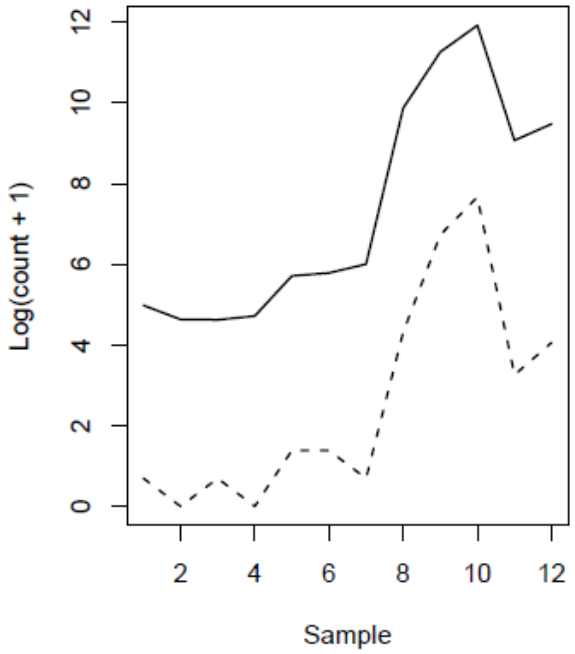
~ 700 clusters of co-expression

For RNA-seq data

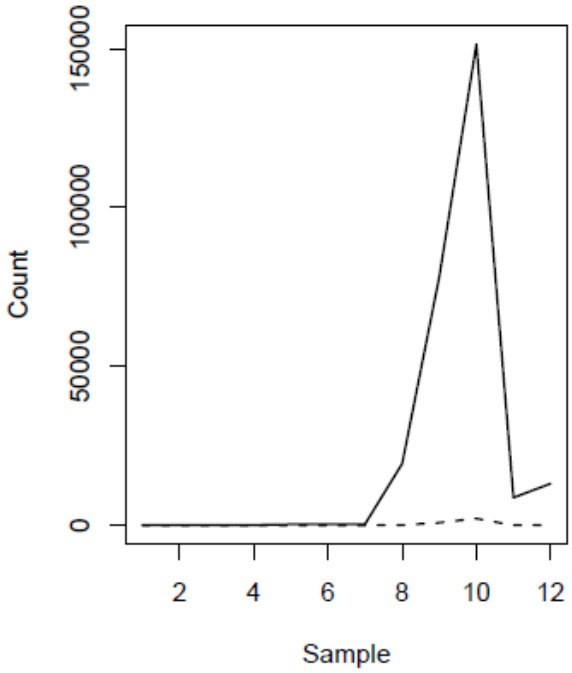


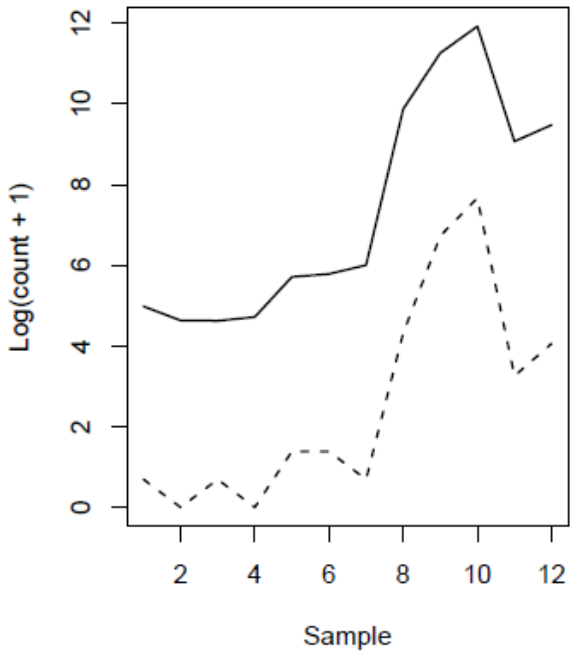
For RNA-seq data

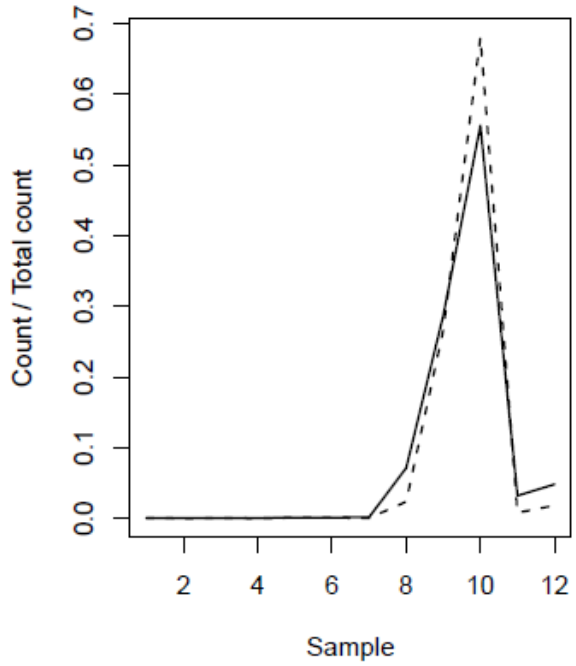




For RNA-seq data

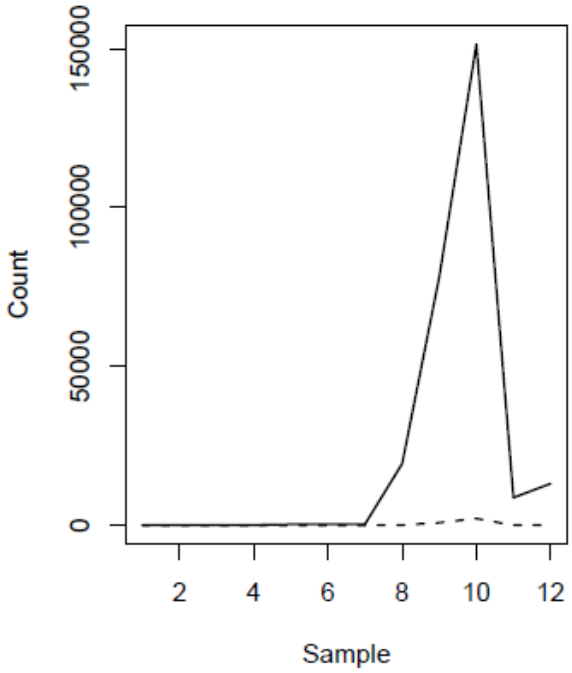


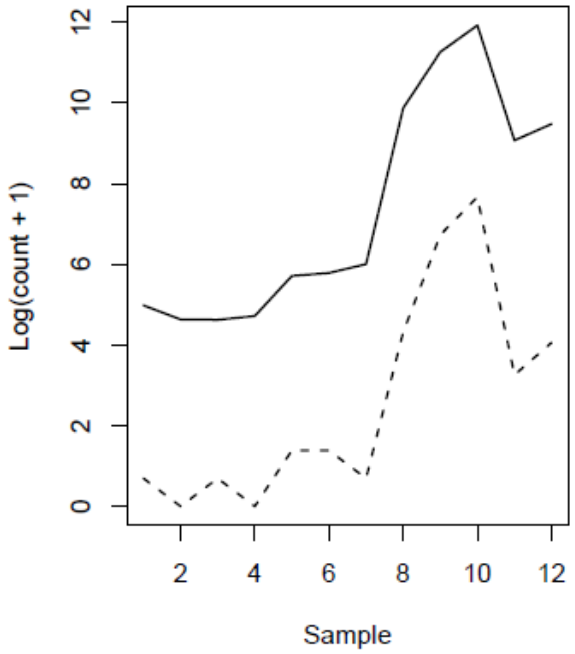


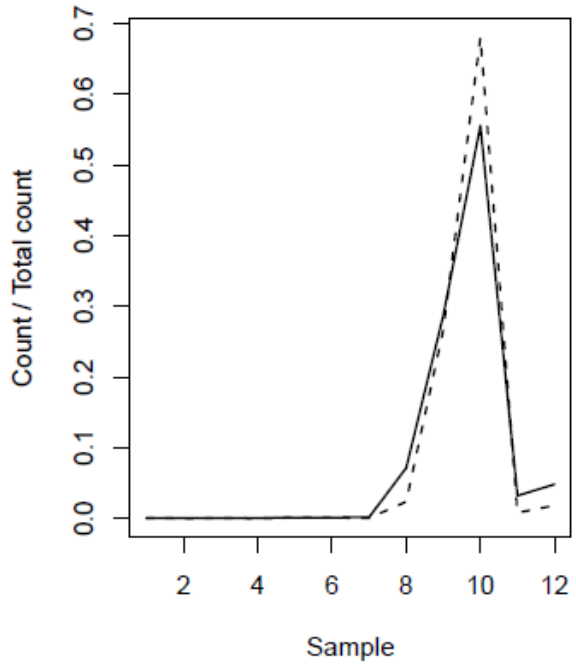


- Let y_{ij} be the raw count for gene i in sample j, with library size s_j
- Profile for gene i: $p_{ij} = \frac{y_{ij}}{\sum_\ell y_{i\ell}}$

For RNA-seq data







Normalized profile for gene i: $p_{ij} = \frac{y_{ij}/s_j}{\sum_\ell y_{i\ell}/s_\ell}$

Finite mixture models for RNA-seq

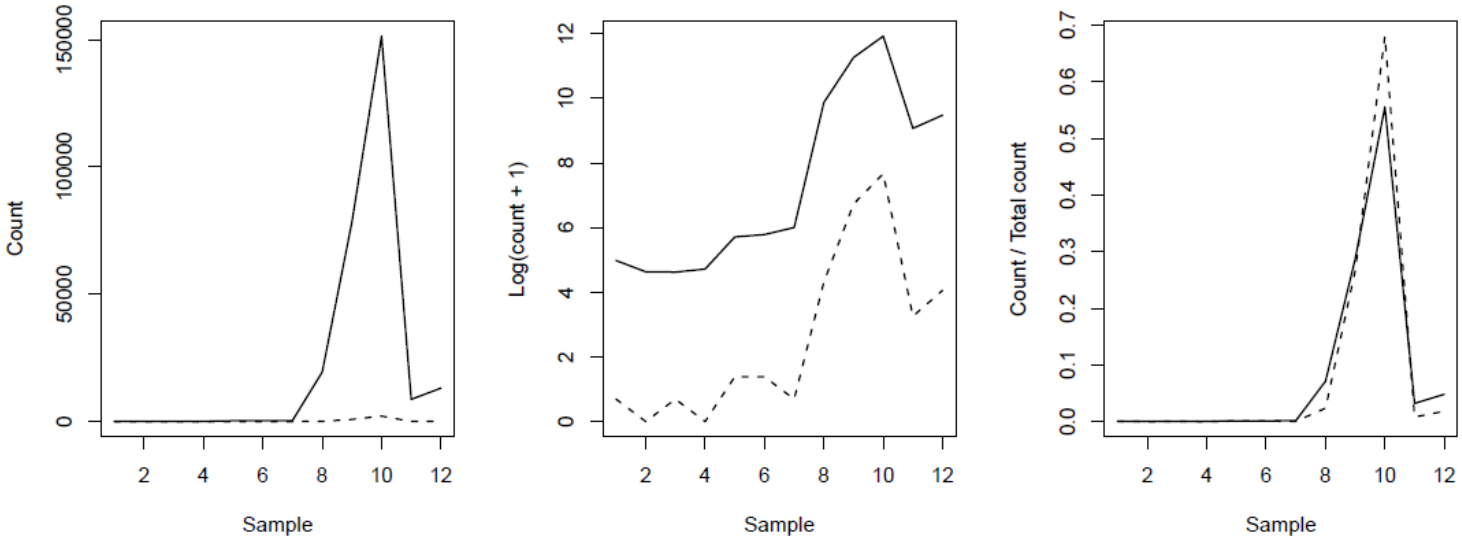
- Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote the observations with $\mathbf{y}_i \in \mathbb{R}^p$ and $n >> p$
- We introduce a latent variable to indicate the group from which each observation arises:

$$\mathbf{Z} \sim \mathcal{M}(n; \pi_1, \ldots, \pi_K), \quad \sum_{k=1}^{K} \pi_k = 1$$

$$P(Z_i = k) = \pi_k$$

- Assume that $\mathbf{y}_i$ are conditionally independent given $\mathbf{Z}$
- Model the distribution of $\mathbf{y}_i | Z_i$ using a parametric distribution:
  - For microarray data, we often assume $\mathbf{y}_i | Z_i = k \sim \mathcal{N}_p(\mu_k, \Sigma_k)$
  - What about RNA-seq data?

What family & parameterization for RNA-seq data?



1. Directly model read counts (`HTSCluster`):

$$(\mathbf{y}_i|Z_i = k) \sim \prod_{j=1}^{p} \text{Poisson}(y_{ij}|\mu_{ijk})$$

2. Apply appropriately chosen data transformation (`coseq`):

$$(\tilde{\mathbf{y}}_i|Z_i = k) \sim \mathcal{N}_p(\mu_k, \Sigma_k)$$

HTSCluster: poisson mixture models [12]

$$\mathbf{y}_i | Z_i = k \sim \prod_{j=1}^{J} \text{Poisson}(y_{ij} | \mu_{ijk})$$

Question: How to parameterize the mean μ_{ijk} to obtain meaningful clusters of co-expressed genes?

[12] Rau *et al.* (2015)

HTSCluster: poisson mixture models [12]

$$\mathbf{y}_i | Z_i = k \sim \prod_{j=1}^{J} \text{Poisson}(y_{ij} | \mu_{ijk})$$

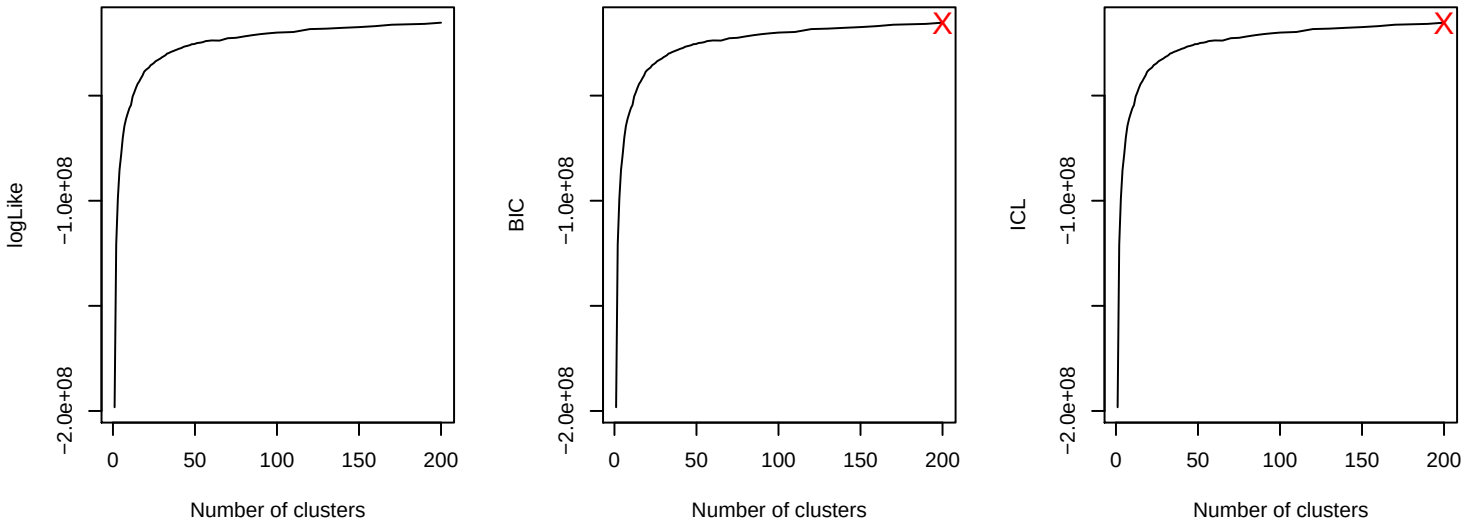
Question: How to parameterize the mean μ_{ijk} to obtain meaningful clusters of co-expressed genes?

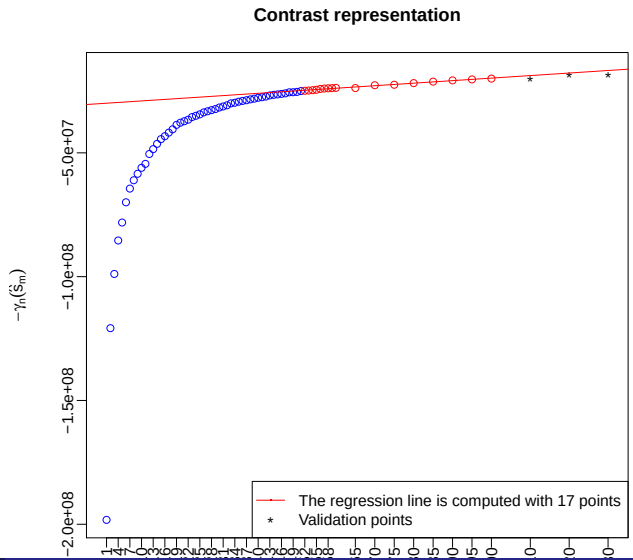
$$\mu_{ijk} = w_i \lambda_{jk} s_j$$

- w_i : overall expression level of observation i ($y_{i.}$)
- $\lambda_k = (\lambda_{jk})$: clustering parameters that define the profiles of genes in cluster k (variation around w_i)
- s_j : normalized library size for sample j, where $\sum_j s_j = 1$

[12] Rau *et al.* (2015)

Behavior of model selection in practice for RNA-seq





Poisson mixture models for RNA-seq data

Advantages:

1. Directly models counts (no data transformation necessary)
2. Clusters interpreted in terms of profiles around mean expression
3. Implemented in `HTSCluster` package on CRAN (v1.0.8)
4. Promising results on real data...

Poisson mixture models for RNA-seq data

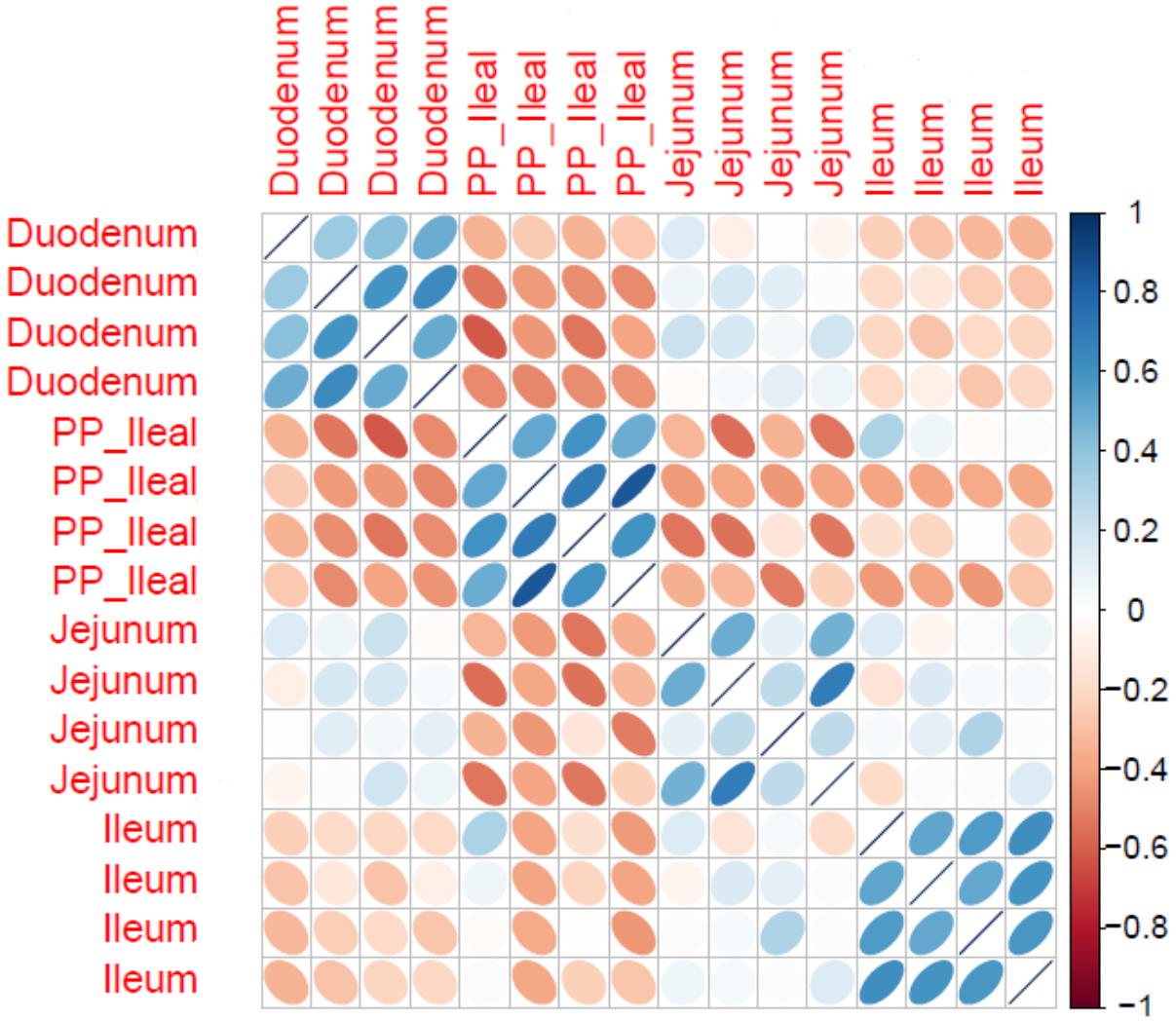
Advantages:

1. Directly models counts (no data transformation necessary)
2. Clusters interpreted in terms of profiles around mean expression
3. Implemented in `HTSCluster` package on CRAN (v1.0.8)
4. Promising results on real data...

Limitations:

1. Slope heuristics requires a very large collection of models to be fit
2. Restrictive assumption of conditional independence among samples
3. Cannot model per-cluster correlation structures
4. Poisson distribution requires assuming that mean = variance

Correlation structures in RNA-seq data



Example: data from Mach *et al.* (2014) on site-specific gene expression along the gastrointestinal tract of 4 healthy piglets

Coseq: Gaussian mixture models[13]

Idea: Transform RNA-seq data, then apply Gaussian mixture models

Several data transformations have been proposed

- $\log_2(y_{ij} + c)$
- Variance stabilizing transformation (`DESeq`)
- Moderated log counts per million (`edgeR`)
- Regularized log-transformation (`DESeq2`)

... but recall that we wish to cluster the normalized profiles

$$p_{ij} = \frac{y_{ij}/s_j}{\sum_\ell y_{i\ell}/s_\ell}$$

[13] Rau & Maugis-Rabusseau (2017)

Remark: transformation needed for normalized profiles

- The normalized profiles are *compositional data*, i.e. the sum for each gene $p_{i.} = 1$
- This implies that the vector $\mathbf{p}_i$ is linearly dependent $\Rightarrow$ imposes constraints on the covariance matrices Σ_k that are problematic for the general Gaussian mixture models

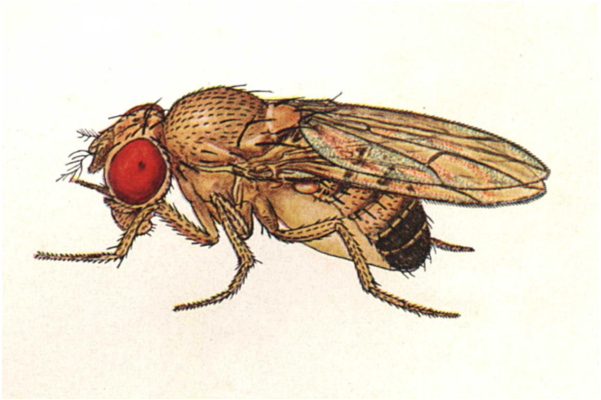
- As such, we consider a transformation on the normalized profiles to break the sum constraint:

$$\arcsin\left(\sqrt{p_{ij}}\right)$$

- And fit a Gaussian mixture model to the transformed normalized profiles:

Real data analysis: Embryonic fly development

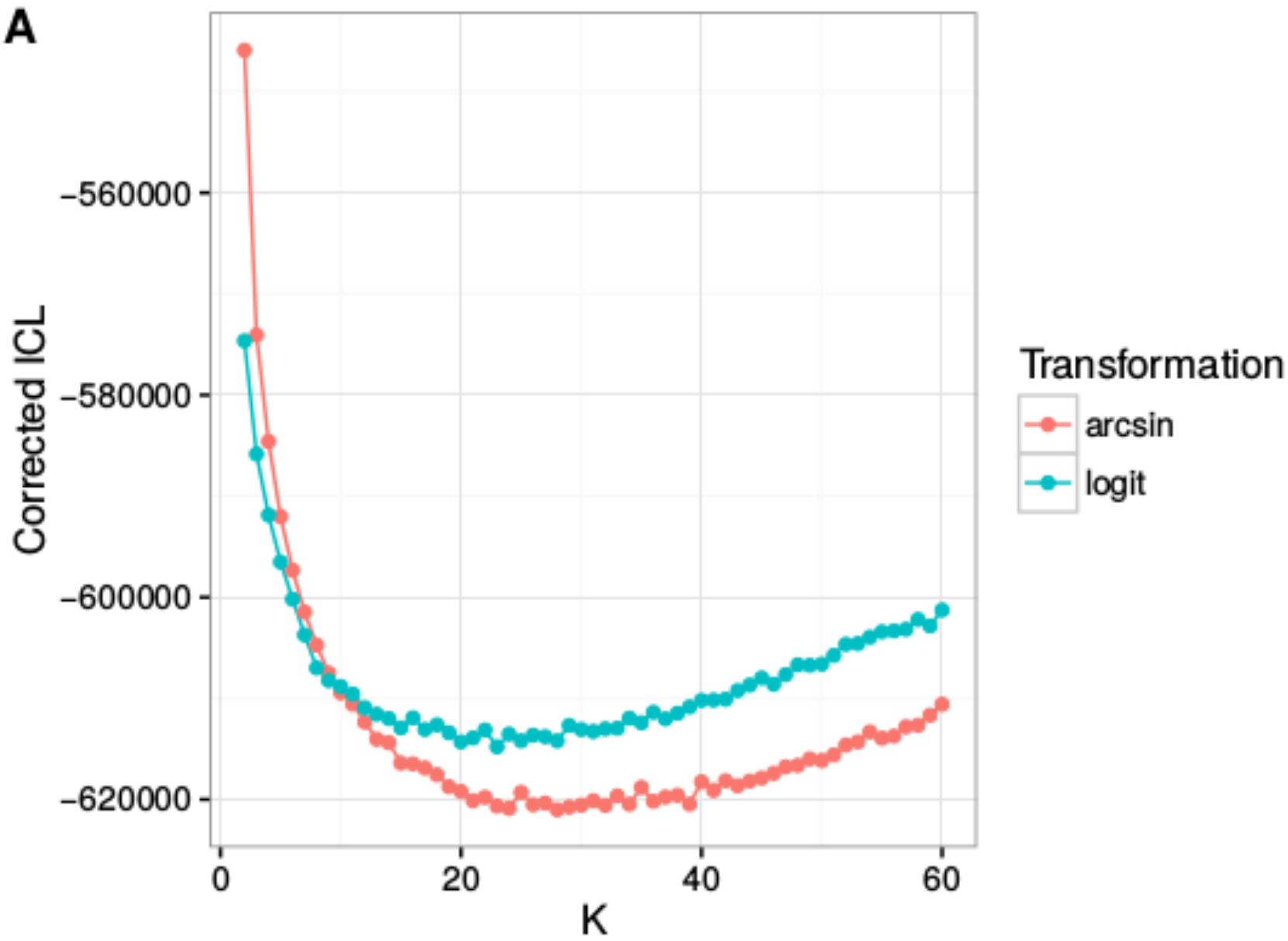
- modENCODE project to provide functional annotation of Drosophila (Graveley et al., 2011)
- Expression dynamics over 27 distinct stages of development during life cycle studied with RNA-seq
- 12 embryonic samples (collected at 2-hr intervals over 24 hrs) for 13,164 genes downloaded from ReCount database (Frazee et al., 2011)



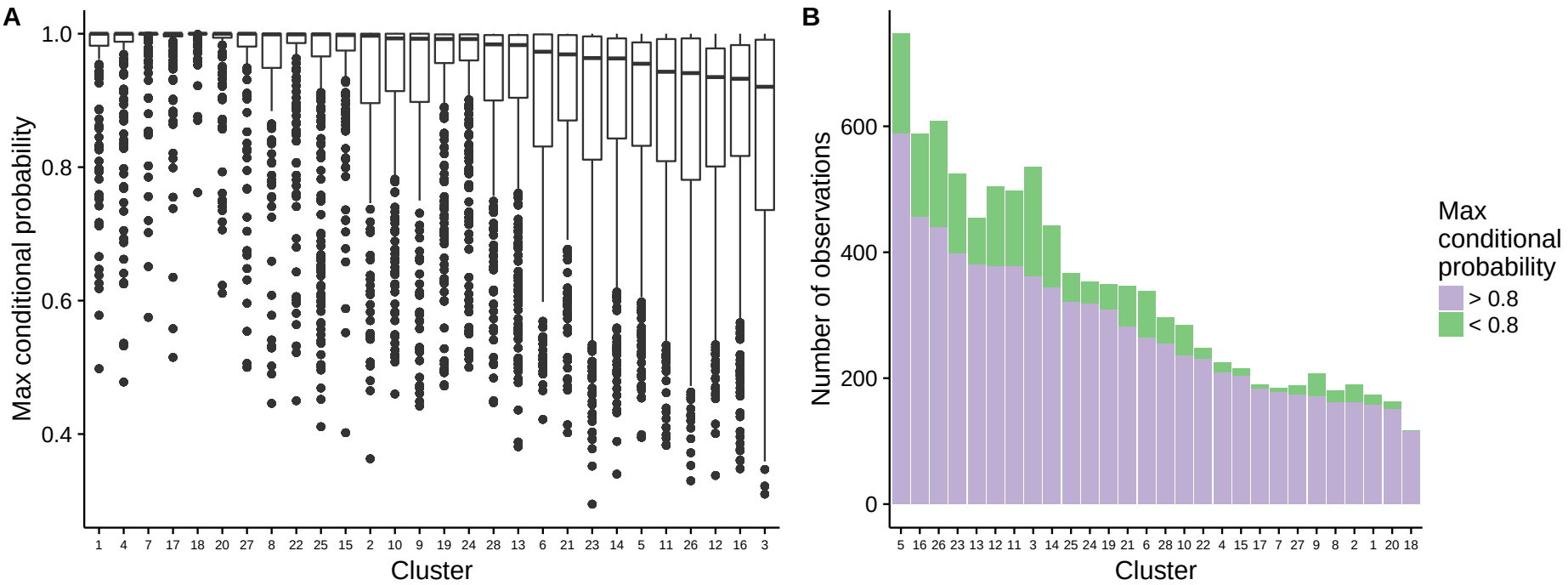
Running the PMM or GMM for RNA-seq data with coseq

```
> library(coseq)
>
> GMM <- coseq(counts, K=2:10, model="Normal",
>              transformation="arcsin")
> summary(GMM)
> plot(GMM)
>
> ## Note: indirectly calls HTSCluster for PMM
> PMM <- coseq(counts, K=2:10, model="Poisson",
>              transformation="none")
> summary(PMM)
> plot(PMM)
```

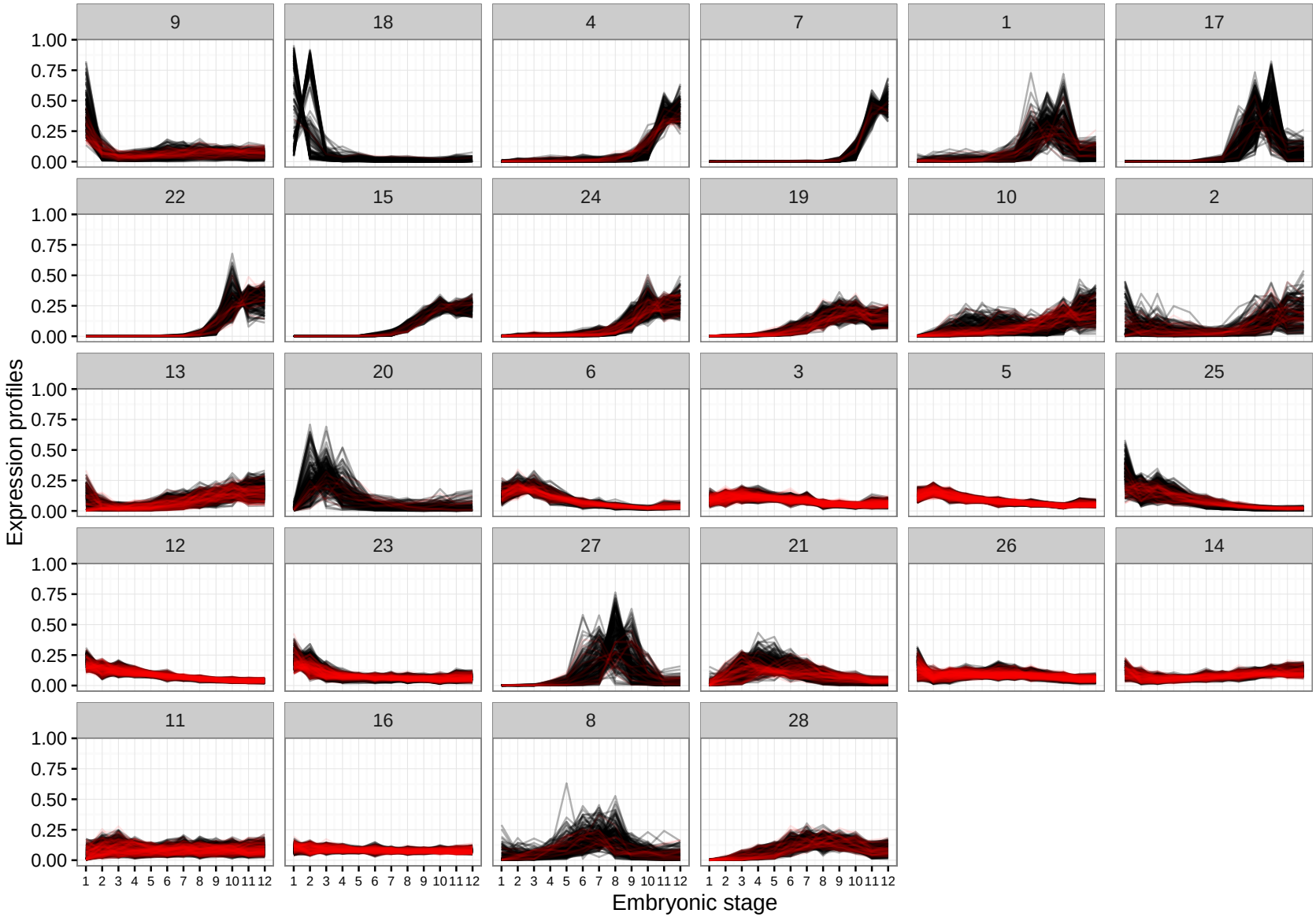
Evaluation of clustering quality



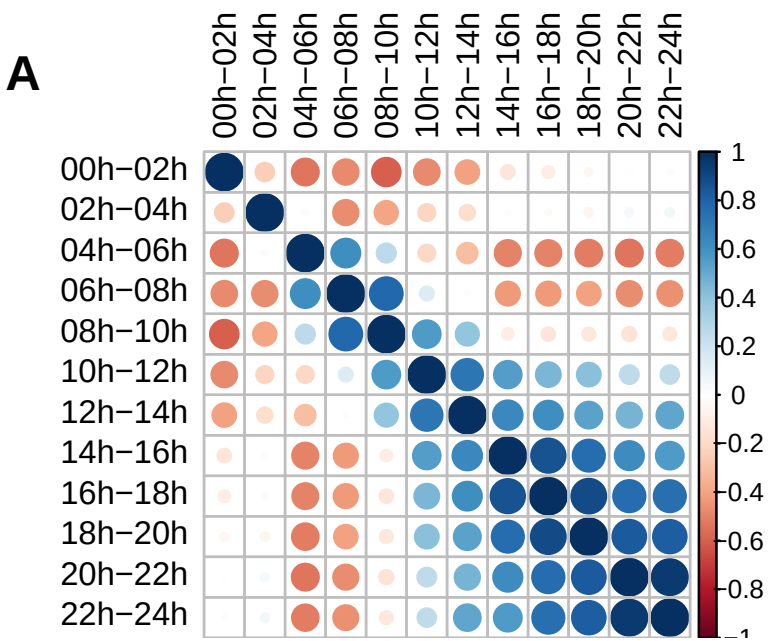
Evaluation of clustering quality

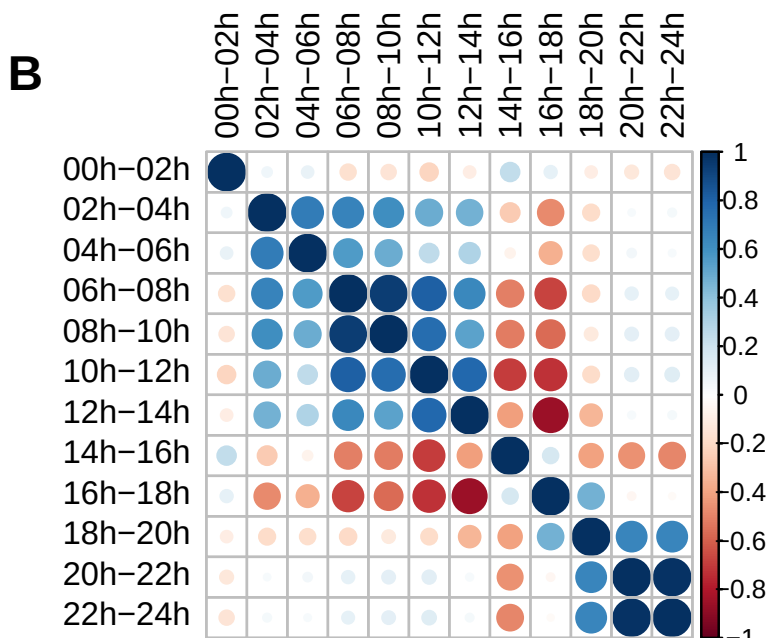


Examining clustering results



Examining clustering results





Some practical remarks for co-expression analysis

Preprocessing details (normalization, filtering, dealing with missing values) can affect clustering outcome

- Should all genes be included?
  Screening via differential analysis or a filtering step (based on mean expression or coefficient of variation)...
  $\rightsquigarrow$ Usually a good idea, genes that contribute noise will affect results!

- What to do about replicates?
  Average, or model each one independently.

Acknowledgements & References

Jain & Dubes (1988) *Algorithms for Clustering Data*. Prentice-Hall, Upper Saddle River, NJ.
D'haeseller (2005) How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499-501.
Yeung *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977-87.
Eisen *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863-8.
Dempster *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *JRSS B*, 39(1):1-38.
Birgé & Massart (2007) Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields* 138(1):33-73.

Rau *al.* (2015) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics* 31(9):1420-7.
Rau & Maugis-Rabusseau (2017). Transformation and model choice for co-expression analysis of RNA-seq data. *Briefings in Bioinformatics.*
Godichon-Baggioni A, Maugis-Rabusseau C and Rau A (2017). Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. arXiv.

MixStatSeq ANR-JCJC grant coordinated by
Cathy Maugis-Rabusseau (INSA / IMT Toulouse)