



HAL
open science

Do my clusters make sense? Some statistical and biological clues to help...

Andrea Rau

► **To cite this version:**

Andrea Rau. Do my clusters make sense? Some statistical and biological clues to help... École thématique. DIGIT-BIO tutorial, Ecully, France. 2022. hal-04482513

HAL Id: hal-04482513

<https://hal.inrae.fr/hal-04482513v1>

Submitted on 28 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Do my clusters make sense? Some statistical and biological clues to help..

Andrea Rau

December 8, 2022
DIGIT-BIO seminar @ Ecully



Outline

- 1 Introduction: Clustering and RNA-seq co-expression
- 2 Validating co-expression clusters: why and how?
 - Internal metrics
 - Stability metrics
 - External metrics
- 3 Let's try it out!
- 4 Wrapping up

Gene co-expression is...

- **Simultaneous expression**¹ or **co-transcription**² of several genes
- **Similarity**³ (correlation, mutual information, ...) **of expression patterns** over a range of different experiments⁴

Related to shared regulatory inputs,
functional pathways, and biological process(es)⁵
+ a tool to study genes without known or predicted function

¹<https://en.wiktionary.org/wiki/coexpression>

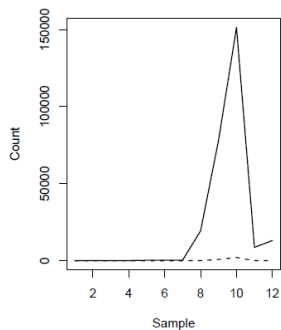
²<http://bioinfow.dep.usal.es/coexpression>

³<http://coexpresdb.jp/overview.shtml>

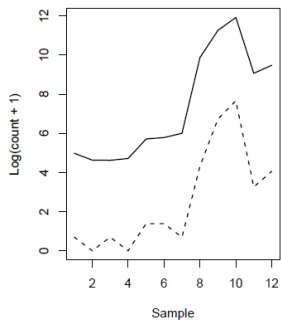
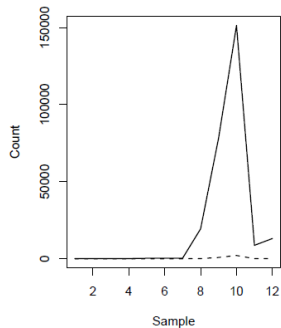
⁴Yeung *et al.* (2001)

⁵Eisen *et al.* (1998)

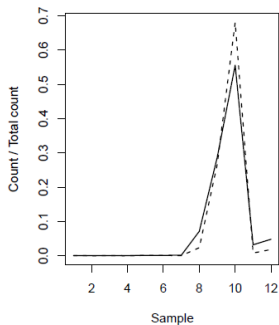
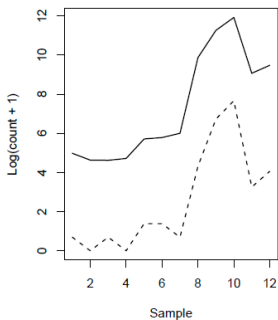
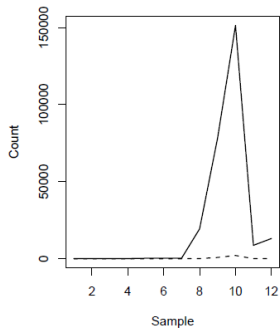
RNA-seq co-expression: counts, transformed counts, or profiles?



RNA-seq co-expression: counts, transformed counts, or profiles?



RNA-seq co-expression: counts, transformed counts, or profiles?



- y_{ij} = raw count for gene i in sample j , with library size normalization factor s_j
- Normalized profile for gene i : $p_{ij} = \frac{y_{ij}/s_j}{\sum_{\ell} y_{i\ell}/s_j}$

Unsupervised classification (aka 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 Centroid-based clustering (**K-means** and hierarchical clustering)
- 2 **Model-based clustering** (mixture models)

Model-based clustering for co-expression

Assume data \mathbf{y} come from K subpopulations, each modeled separately:

$$f(\mathbf{y}|K, \boldsymbol{\Psi}_K) = \prod_{i=1}^n \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k)$$

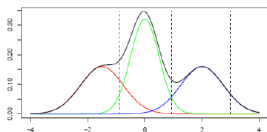
- $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)'$ are the mixing proportions, where $\sum_{k=1}^K \pi_k = 1$
- f_k are the densities of each of the components
- Microarrays: typically assume $\mathbf{y}_i|k \sim \text{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- RNA-seq: Poisson distribution for counts (HTSCluster), Gaussian distribution for **arcsine or CLR-transformed normalized profiles** (coseq)

→ Estimation (EM algorithm), model selection (BIC/ICL/slope heuristics), ...
 See DIGIT-BIO *Concepts en IA* talk by Cathy Maugis-Rabusseau for more!

From finite mixtures to clusters

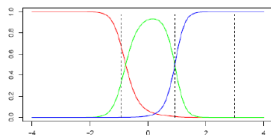
Distributions:

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



Conditional probabilities:

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



Maximum a posteriori (MAP) rule: Assign genes to the component with highest conditional probability τ_{ik} :

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

Outline

- 1 Introduction: Clustering and RNA-seq co-expression
- 2 Validating co-expression clusters: why and how?
 - Internal metrics
 - Stability metrics
 - External metrics
- 3 Let's try it out!
- 4 Wrapping up

Why validate co-expression clusters?

- Avoid finding patterns in noise: does non-random structure actually exist in the data?
- Evaluate fit of clustering on data, compare different algorithms or results ⇒ Internal
- Determine “correct” number of clusters
- Determine robustness of clustering results ⇒ Stability
- Compare clusters to externally known labels
- Characterize clusters with respect to externally known information ⇒ External

Internal cluster validation

Evaluate goodness of clustering structure using data alone \Rightarrow similarity of genes within cluster vs distinctness of genes between clusters

- **Compactness**, or within-cluster variation
- **Connectivity**, or extent to which genes clustered together are also neighbors in the data space
- **Separation** between clusters



Image: 10.1093/bioinformatics/bti517

Examples of internal validation metrics

- **Silhouette statistic**: measure of how closely data within a cluster is matched (compactness) and how loosely it is matched to neighboring clusters (separation)

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \text{ and } b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j), \text{ for } i \in C_I$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad -1 \leq s(i) \leq 1$$

Examples of internal validation metrics

- **Silhouette statistic**: measure of how closely data within a cluster is matched (compactness) and how loosely it is matched to neighboring clusters (separation)

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \text{ and } b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j), \text{ for } i \in C_I$$

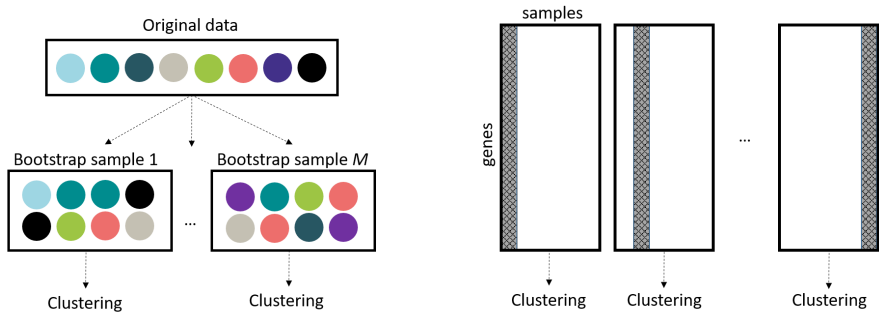
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad -1 \leq s(i) \leq 1$$

- **Dunn index**: ratio of smallest distance between clusters and their diameter (largest intra-cluster distance)

$$DI = \frac{\min_{C_I, C_J, I \neq J} \left(\min_{i \in C_I, j \in C_J} d(i, j) \right)}{\max_{C_M} \text{diam}(C_M)}$$

where $\text{diam}(C_M)$ is maximum distance among $i \in C_M$, $0 < DI$ (**maximize**)

Validation of cluster stability



- `fpc`: resampling of data using bootstrap → calculate Jaccard similarities to original clusters
- `c1Valid`: remove columns one by one, compare to clustering from full data (several criteria proposed)

External cluster validation

Evaluate ability of clustering algorithm to produce biologically meaningful clusters with respect to:

- Gene ontology (GO) term annotations
- A priori functional categorizations of genes
- Pathway membership
- ...

⇒ Assume set of F known biological classes (not necessarily disjoint)

Examples of external validation metrics: *comparison*

- **Adjusted Rand Index** (ARI): corrected-for-chance Rand index

$$R = \frac{TP + TN}{TP + FP + FN + TN} \in [0, 1]$$

Examples of external validation metrics: *comparison*

- **Adjusted Rand Index** (ARI): corrected-for-chance Rand index

$$R = \frac{TP + TN}{TP + FP + FN + TN} \in [0, 1]$$

- **Biological Homogeneity Index** (BHI): if $B(i)$ is the known biological class of gene i ,

$$BHI = \frac{1}{K} \sum_{k=1}^K \frac{1}{n_k(n_k - 1)} \sum_{i \neq j \in C_k} I(B(i) = B(j)) \in [0, 1]$$

Examples of external validation metrics: *comparison*

- **Adjusted Rand Index** (ARI): corrected-for-chance Rand index

$$R = \frac{TP + TN}{TP + FP + FN + TN} \in [0, 1]$$

- **Biological Homogeneity Index** (BHI): if $B(i)$ is the known biological class of gene i ,

$$BHI = \frac{1}{K} \sum_{k=1}^K \frac{1}{n_k(n_k - 1)} \sum_{i \neq j \in C_k} I(B(i) = B(j)) \in [0, 1]$$

- **Biological Stability Index** (BSI):

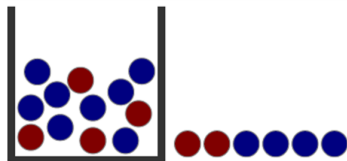
$$BSI = \frac{1}{F} \sum_{k=1}^F \frac{1}{|B_k|(|B_k| - 1)M} \sum_{\ell=1}^M \sum_{i \neq j \in B_k} \frac{|C^{i,0} \cap C^{j,\ell}|}{|C^{i,0}|} \in [0, 1]$$

based on removing each of the M data columns one at a time

Examples of external validation metrics: *characterization*

What biological processes are over-represented in each cluster?

- Gene Ontology (GO) terms = group genes into categories by a common biological property
- Assumptions: under the null hypothesis, genes are independent and equally likely to be grouped together in the list of interest (= cluster)
- Test for over-representation using a hypergeometric distribution (Fisher's exact test)



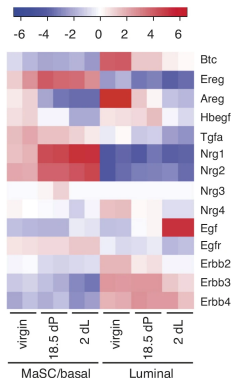
	Red	Blue	Total
Chosen	2	4	6
Remaining	4	8	12
Total	6	12	18

Outline

- 1 Introduction: Clustering and RNA-seq co-expression
- 2 Validating co-expression clusters: why and how?
 - Internal metrics
 - Stability metrics
 - External metrics
- 3 Let's try it out!
- 4 Wrapping up

Data description: Fu *et al.* (2015)

- RNA-seq experiment to study lineage of luminal cells in mouse mammary gland and changes in expression upon pregnancy & lactation
- **2 cell types** {basal stem-cell enriched cells, committed luminal cells} × **3 statuses** {virgin, pregnant, lactating} × **2 biological replicates**
 - Illumina HiSeq → 30 million 100bp SE reads
 - *Pre-processing*: Subread to align reads to mm10 genome + featureCounts for quantification of Entrez genes (RefSeq)
 - *Initial analysis*: filter genes with weak expression or unambiguous/missing IDs + DESeq2 normalization/differential analysis



Supp Fig 4: 10.1038/ncb3117

Outline

- 1 Introduction: Clustering and RNA-seq co-expression
- 2 Validating co-expression clusters: why and how?
 - Internal metrics
 - Stability metrics
 - External metrics
- 3 Let's try it out!
- 4 **Wrapping up**

Validating clustering approaches in practice

Clustering results can be evaluated based on **internal and stability** criteria (e.g., statistical properties of clusters) or **external** criteria (e.g., functional annotations)

- Preprocessing steps will affect clustering outcome
- Methods that give different results depending on the initialization should be rerun multiple times to check for stability
- Repeated subsampling to identify consensus clusters (ConsensusClusterPlus)
- Most methods will find clusters even when no structure is present \Rightarrow good idea to compare to results with randomized data

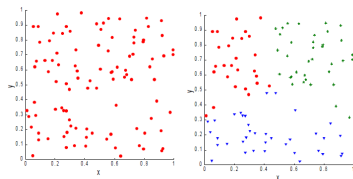


Image: Ruming Jin, *Cluster validation*

Final thoughts⁶

“

There is no single best criterion for obtaining a partition because no precise and workable definition of *cluster* exists. Clusters can be of any arbitrary shapes and sizes in a multidimensional pattern space. Each clustering criterion imposes a certain structure on the data, and if the data happen to conform to the requirements of a particular criterion, the true clusters are recovered.

”

⁶Jain & Dubes, 1988



MixStatSeq ANR-JCJC grant (2014-2018, PI: C. Maugis-Rabuseau)

Thanks to Gilles Celeux (Inria Saclay - Île-de-France), Cathy Maugis-Rabuseau (INSA / IMT Toulouse), Etienne Delannoy, Marie-Laure Martin (IPS2), and Panos Papastamoulis (Athens University of Economics and Business)

Some useful references

- Jain & Dubes (1988) *Algorithms for Clustering Data*. Prentice-Hall, Upper Saddle River, NJ.
- Rau *et al.* (2015) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics* 31(9):1420-7.
- Rau & Maugis-Rabusseau (2018) Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics* bbw128.
- Godichon-Baggioni *et al.* (2018) Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics* 46(1):47-65.
- Brock *et al.* (2008) cValid: an R package for cluster validation. *Journal of Statistical Software* 25:4.
- Handl *et al.* (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15):3201-3212.
- Ullman *et al.* (2021) Validation of cluster analysis results on validation data: a systematic framework. *WIREs: Data Mining and Knowledge Discovery* 12:e14444.