Intégration Multi-Omiques

Antoine Bodein, Ph.D.
Arnaud Droit, Ph.D.

7 - 12 octobre 2022

Objectifs

- Comprendre les enjeux liés à l'intégration multi-omique.
- Être capable d'expliquer les différents concepts d'intégration multi-omique.
- Savoir appliquer la bonne méthode d'intégration en fonction des données disponibles.
- Savoir exploiter quelques notions de la théorie des graphes pour améliorer l'interprétation de l'intégration multi-omique

Plan

1. Mise en contexte Introduction aux données « omiques »
2. Overview des différentes méthodes d'intégration
3. Les méthodes d'intégration « multivariées »
    1. Concept de l'Analyse en Composante Principale
    2. Présentation de l'outils mixOmics
4. Les méthodes d'intégration à base de réseaux biologiques
5. Exemples d'intégration (ADLab)

Evaluation

- TP:
  - Présentation du travail
  - Même sujet
  - Dataset différent
  - Exposé Oral

Mise en contexte

Différents concepts d'integration

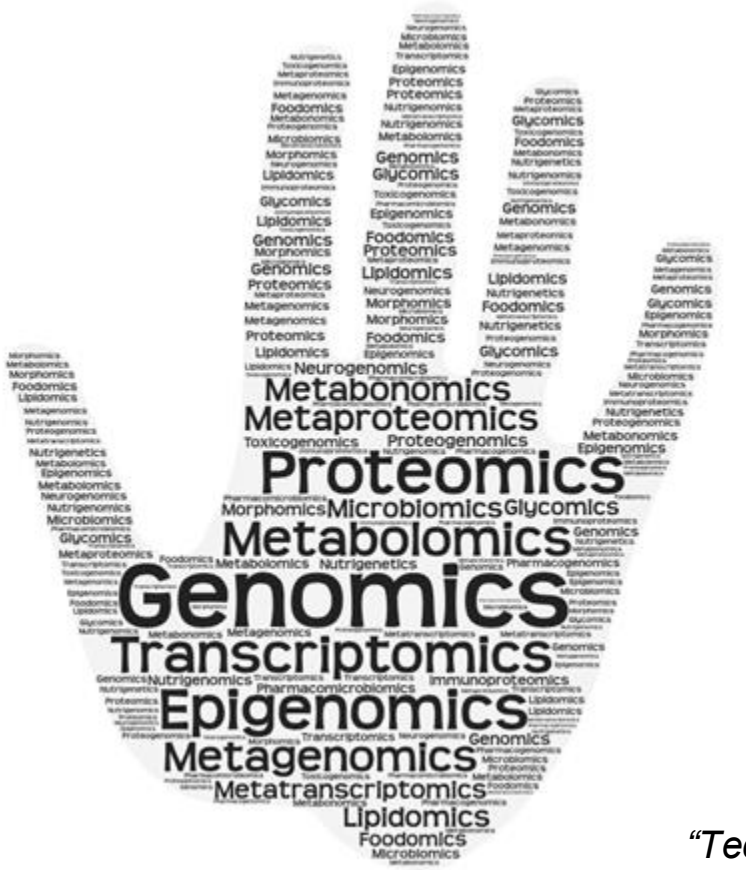
Méthodes multivariées

mixOmics

Réseaux en biologie

Cas d'étude ADLab

Qu’est-ce qu’un omique ?



- “Genomics” inventé dans les 70s
  (Thomas Roderick)
- D’autres termes:
  Protéomique, métabolomique, épigénomique, …
- “Méta”-omique
  Métagénomique, métaprotéomique, …
- Néologisme
  Morphomics, foodomics, ...

“Technologies omiques à haut débit”

Omiques: applications

Génome

- Séquençage *de novo*
- Reséquençage de génomes complets et détection des variations :
- Reséquençage de régions ciblées (exons)

Métagénome

- Diversité des microbes
- Échantillons hétérogènes

Transcriptome (RNA-Seq)

- Profil d'expression des ARNm
- Variants d'épissage
- Expression propre à certaines allèles
- Expression des micro ARN

Épigénome (ChIP-Seq)

- Sites actifs dans la transcription
- Interactions ADN - Protéines
- Modification des histones
- Nucléosomes

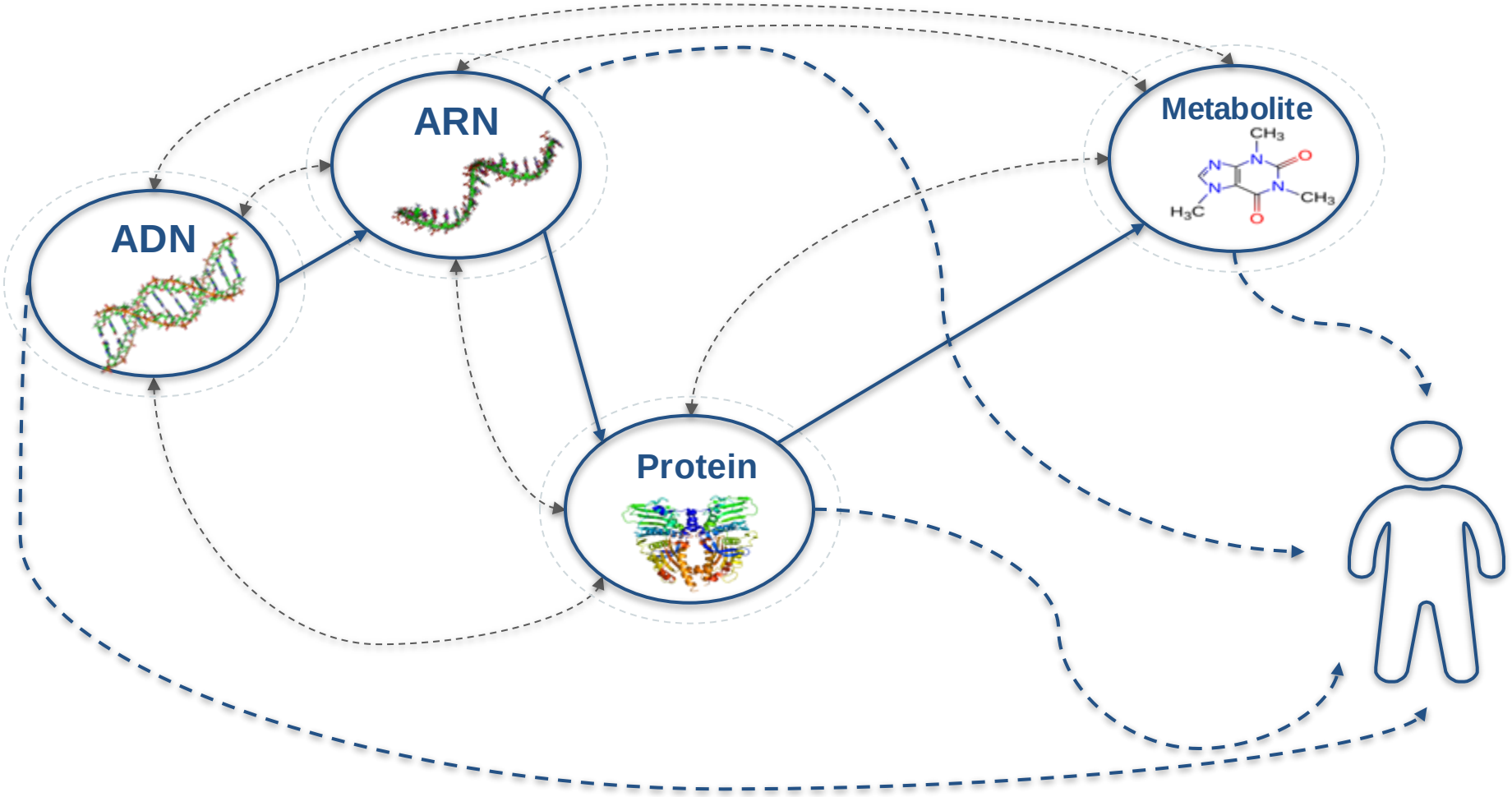
Proteome

- Identification et Quantification
- Profils proteome
- Localisation subcellulaire
- Carte interactome
- …

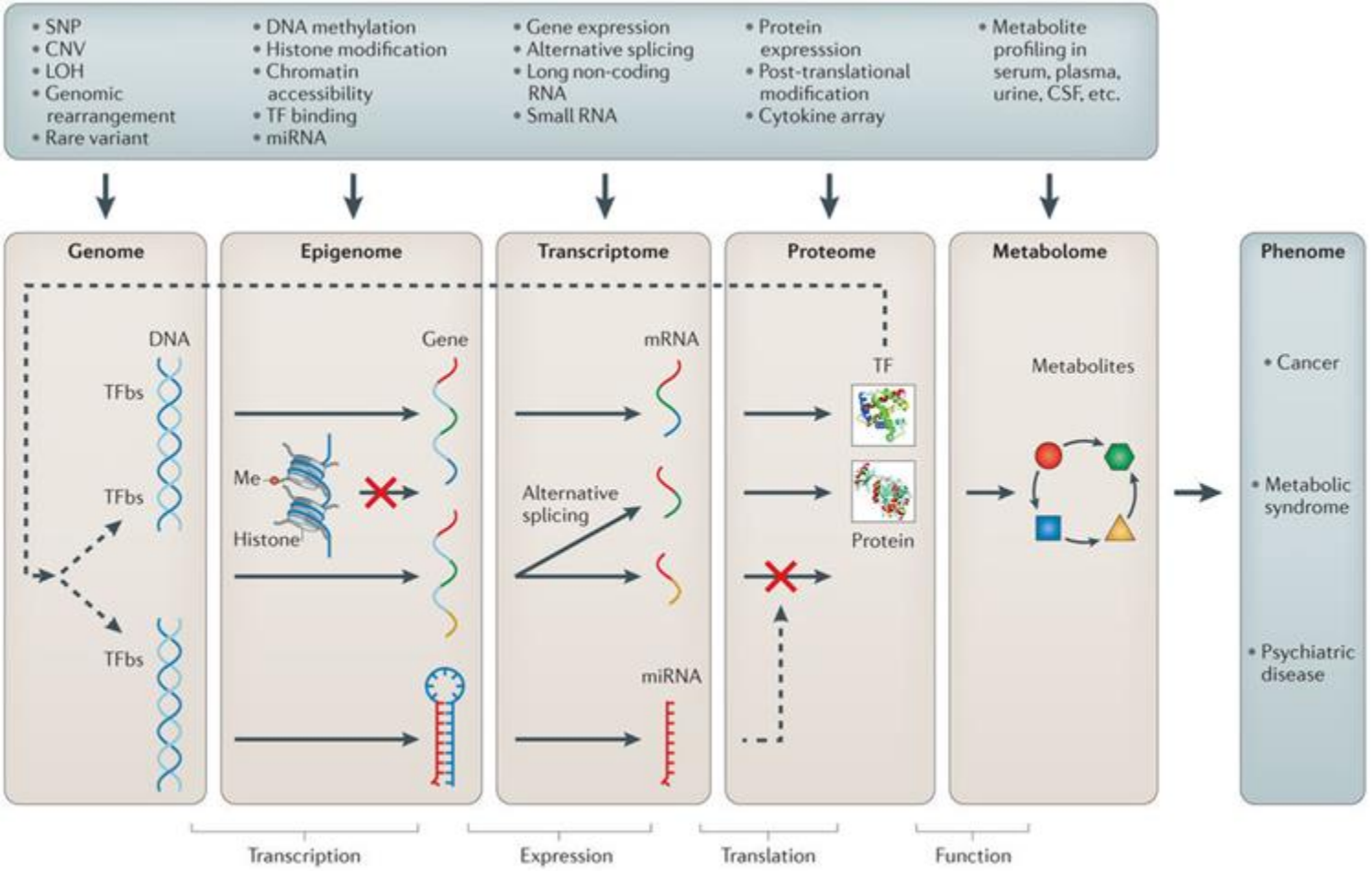
Métabolome

- Identification et Quantification
- Profils metabolome, hormonome, …
- Carte des voies métaboliques
- …

Qu’est-ce que le multi-omique ?

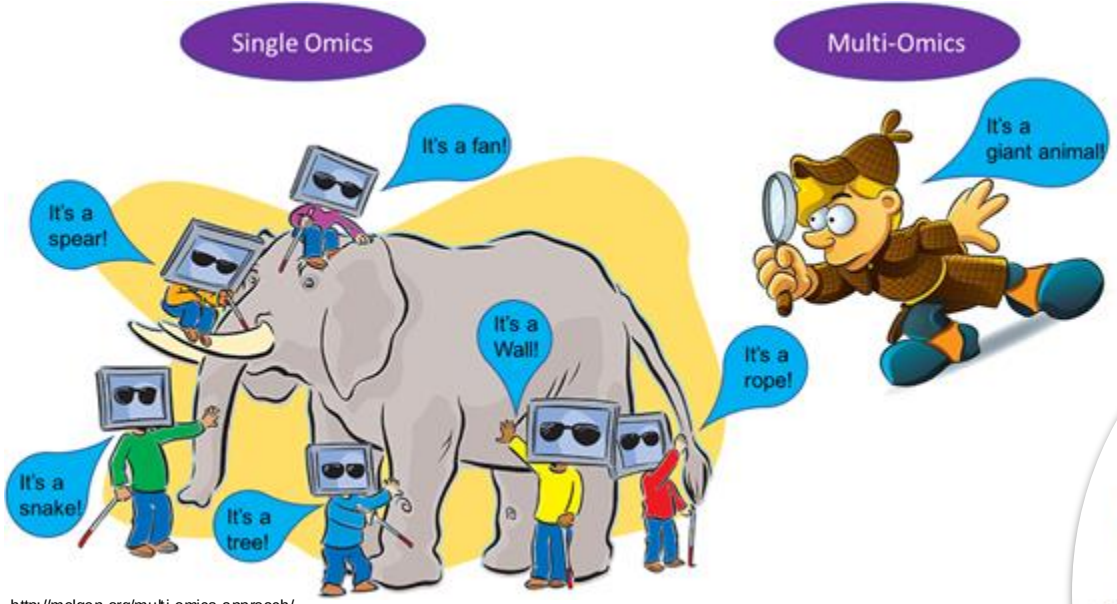


Qu’est-ce que le multi-omique ?



doi:10.1038/nrg3868
Nature Reviews | Genetics

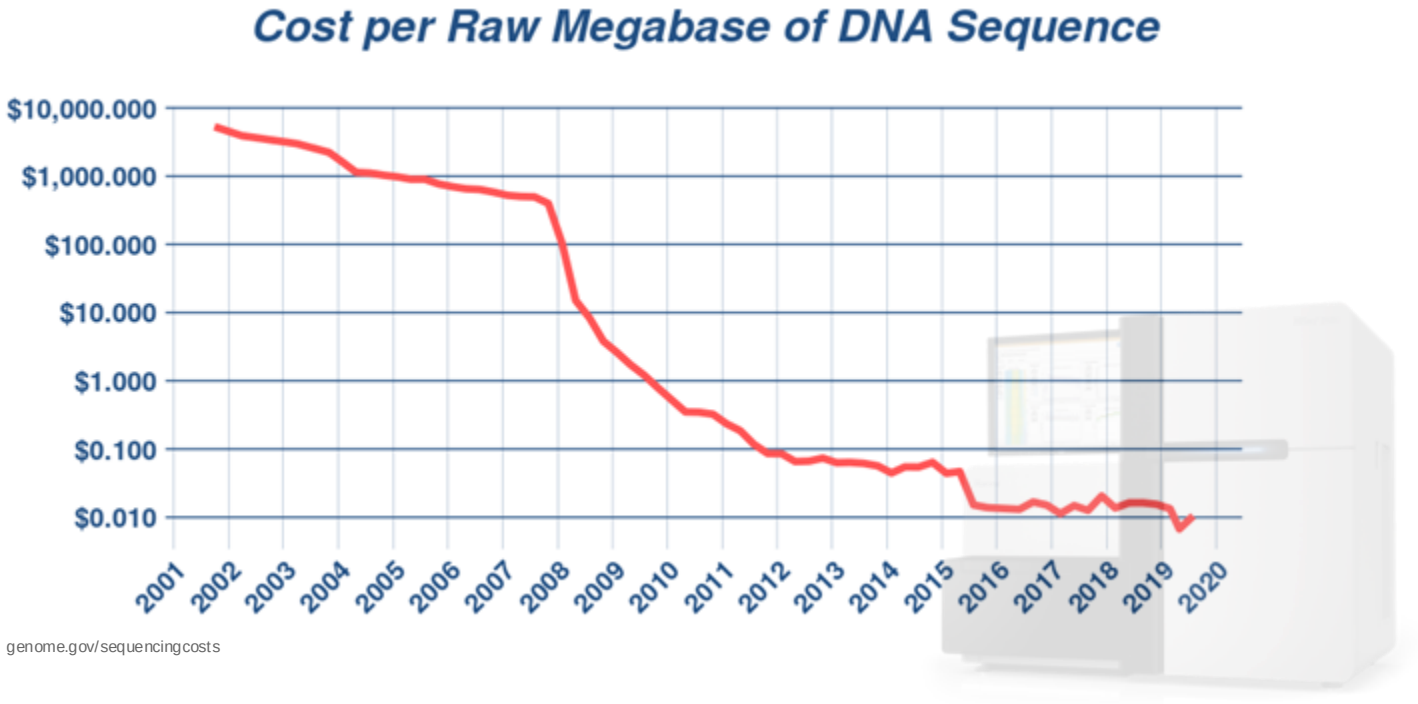
Qu’est-ce que le multi-omique ?



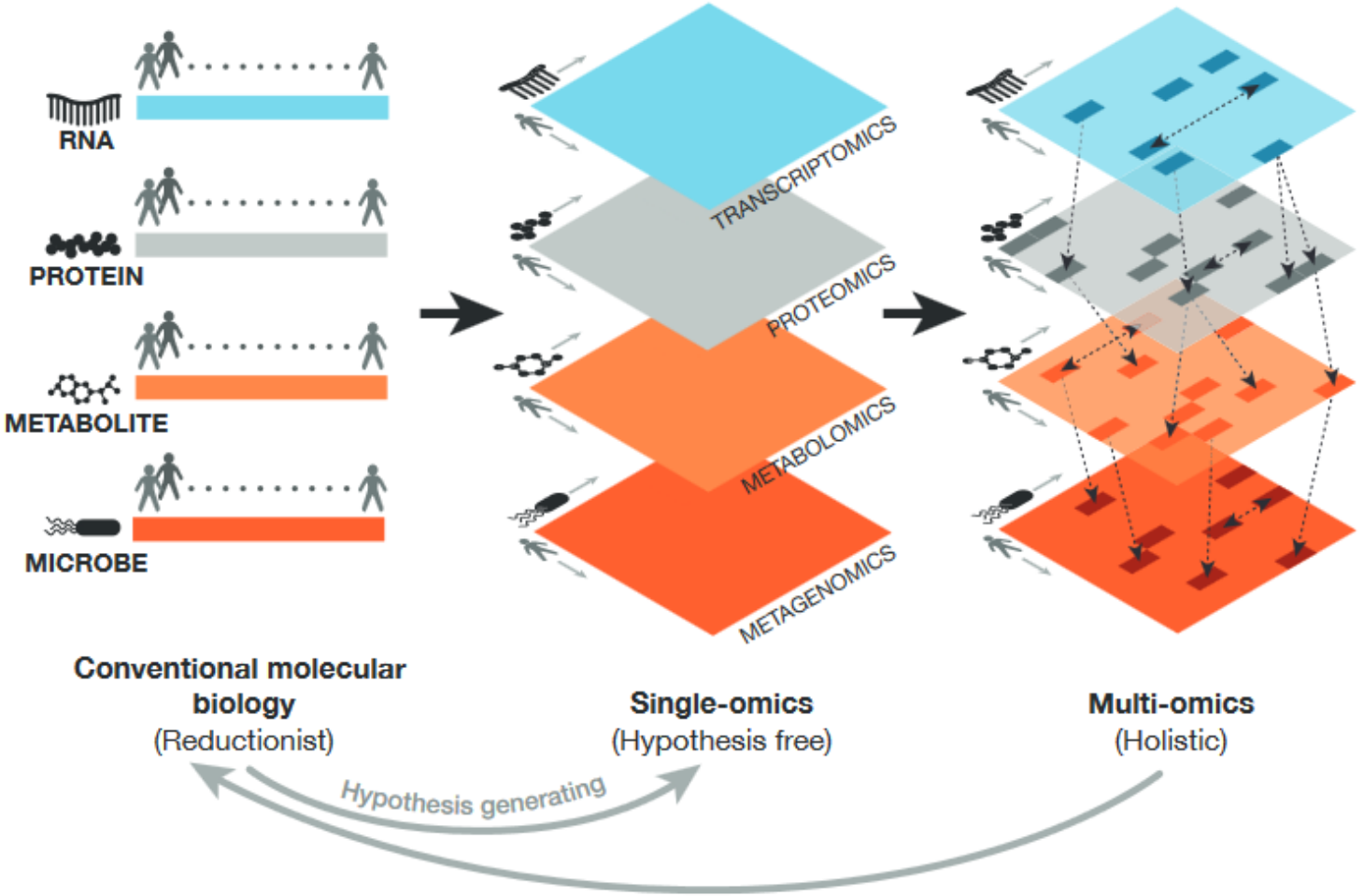
http://melgen.org/multi-omics-approach/

Approche **Réductionniste** (single-omic)
vs
Approche **Holistique** (multi-omic)

Pourquoi ?

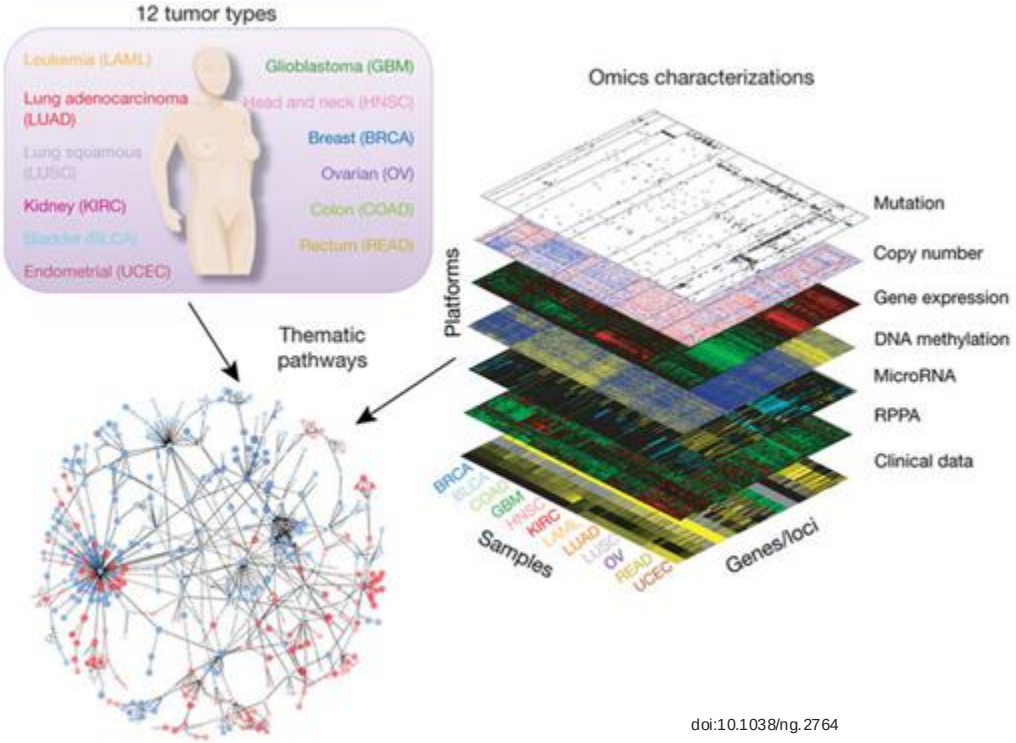

Cost per Raw Megabase of DNA Sequence
$10,000.000
$1,000.000
$100.000
$10.000
$1.000
$0.100
$0.010
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
genome.gov/sequencingcosts

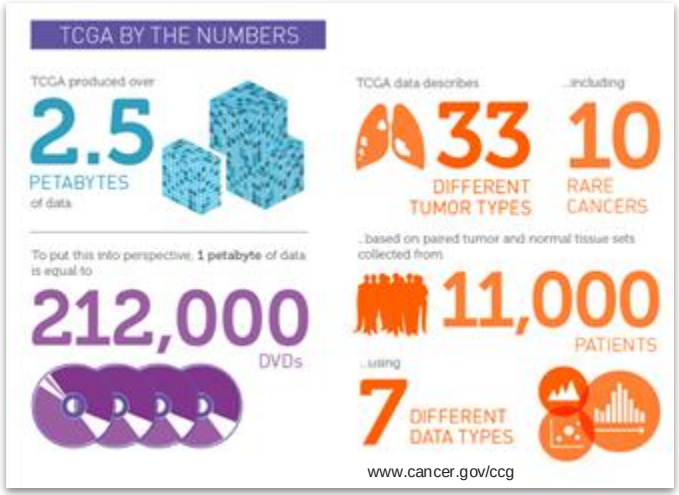
Changement de paradigme



Quelques ressources multi-omique ?

THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

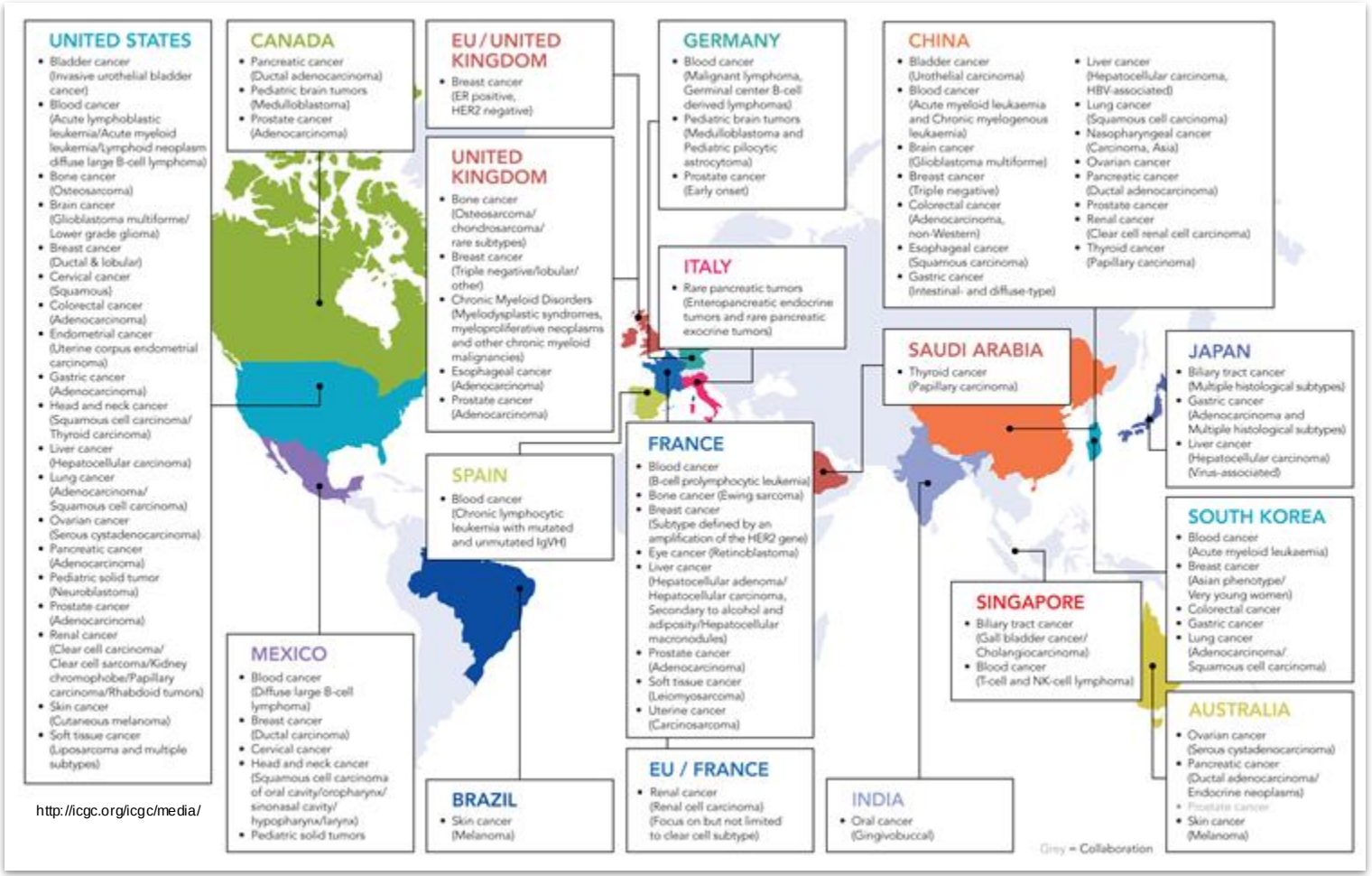




doi:10.1038/ng.2764

Quelques ressources multi-omique ?

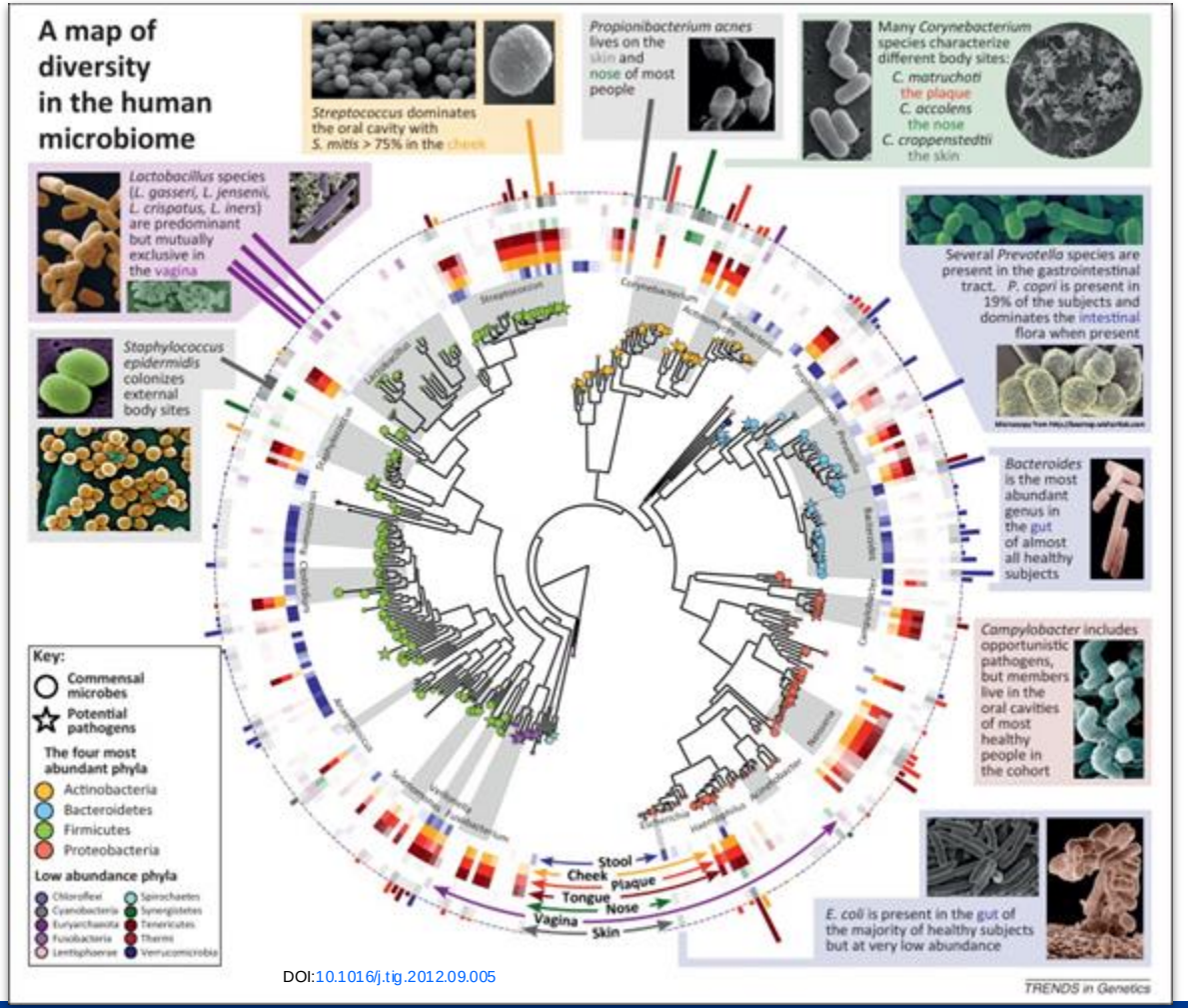
International Cancer Genome Consortium



http://icgc.org/icgc/media/

Quelques ressources multi-omique ?

https://hmpdacc.org



DOI:10.1016/j.tig.2012.09.005

Concrètement …

High Throughput Technology

	Gene-1	Gene-2	Gene-3	Gene-4
A	10	5	6	0
B	42	49	2	9
...				

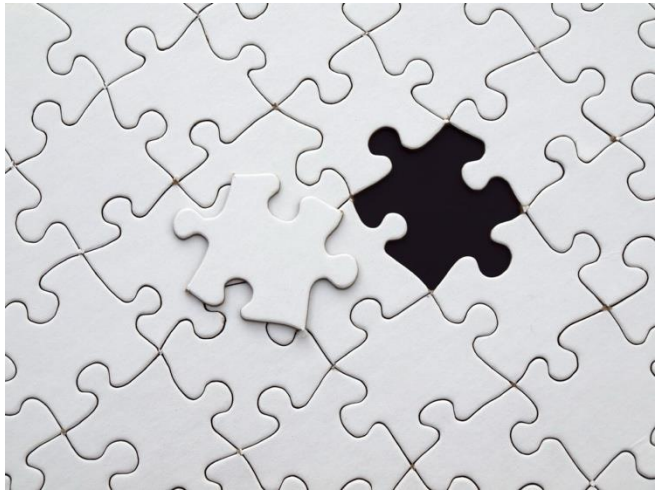
	Prot-1	Prot-2	Prot-3
A	2	9	5
B	8	3	0
...			

	Metab-1	Metab-2	Metab-3	Metab-4	Metab-5
A	58	89	32	73	51
B	72	104	99	43	16
...					

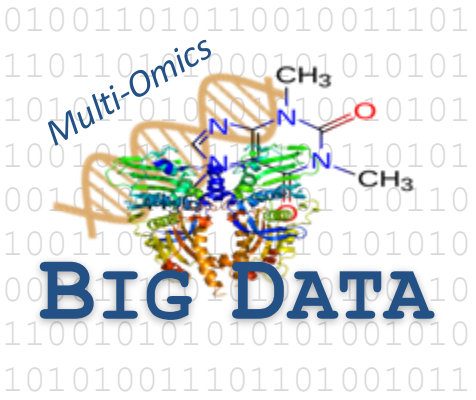
“Bloc”

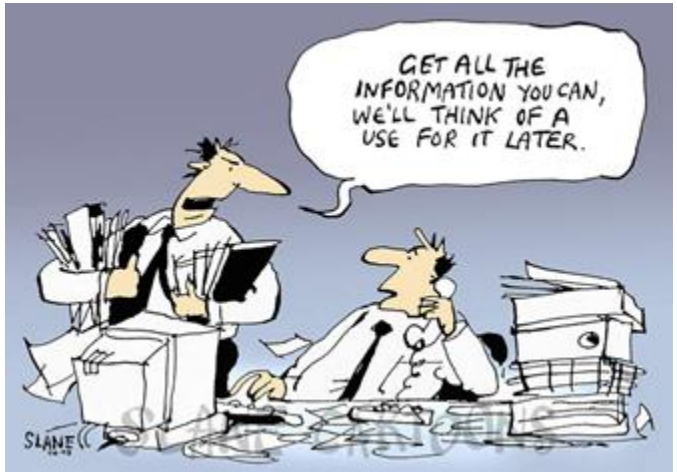
Défis apportés par les données à haut-débits

- Overfitting:
  - Données omiques = Beaucoup de données
  - Modélisation du bruit
- Multi-colinéarité
- Données manquantes et les « 0 »



Défis liés à l'integration multi-omiques





- Sources multiples et hétérogènes
- Puissance de calcul
- Interprétabilité
- Rester informé des avancées dans chaque domaines

Mise en contexte

Différents concepts d'integration

Méthodes multivariées

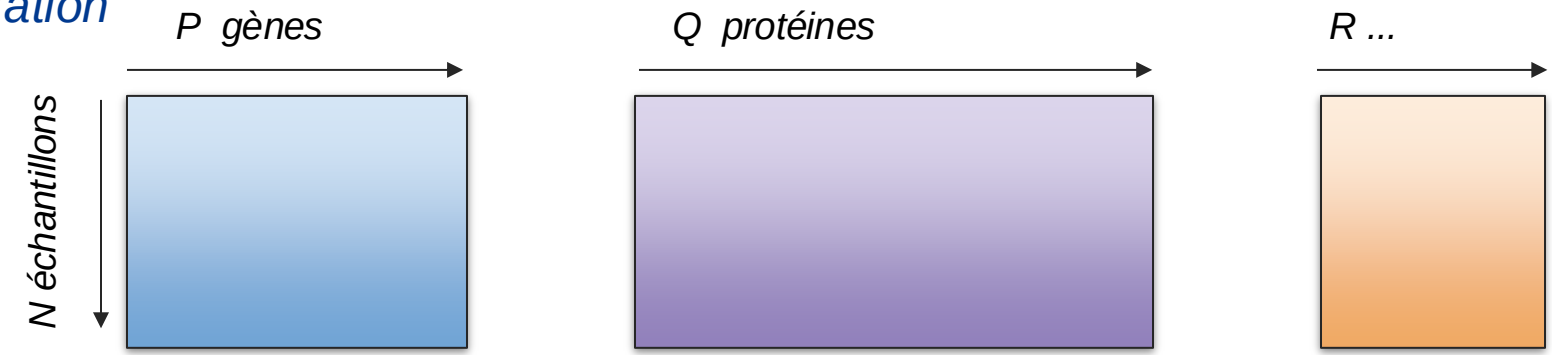
mixOmics

Réseaux en biologie

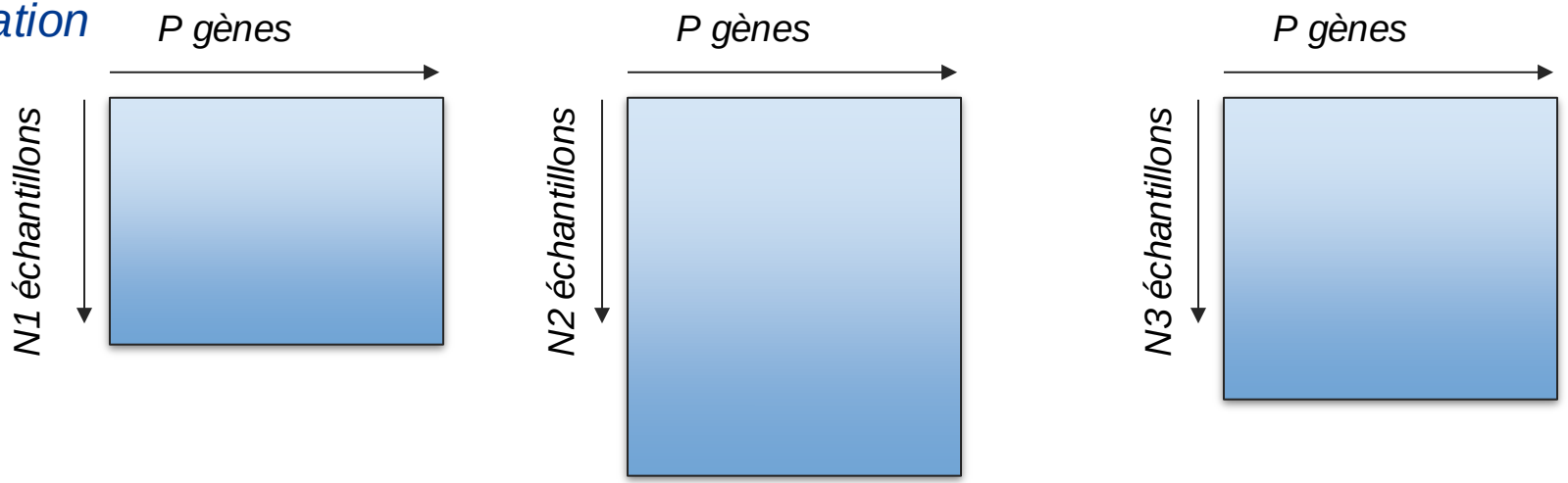
Cas d'étude ADLab

Intégration horizontale et integration verticale

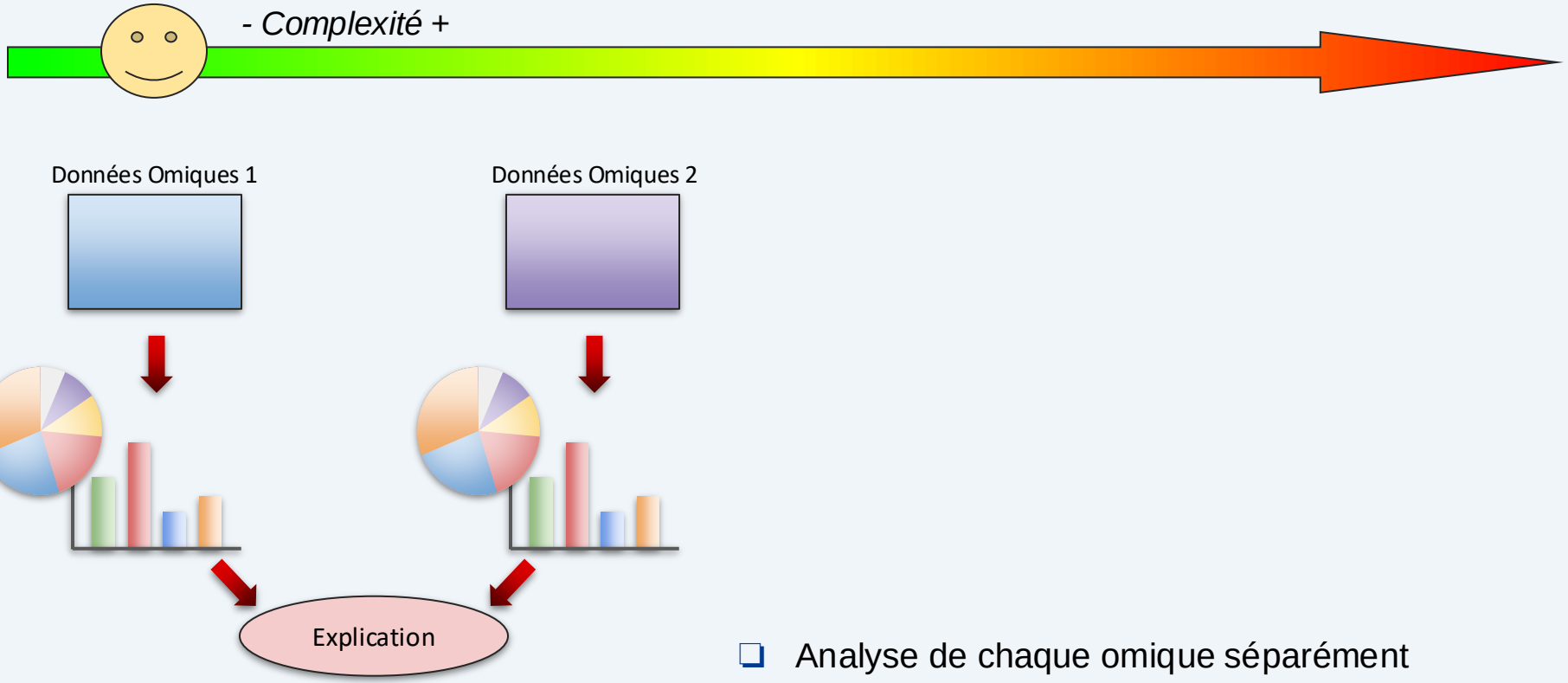
N-Intégration



P-Intégration

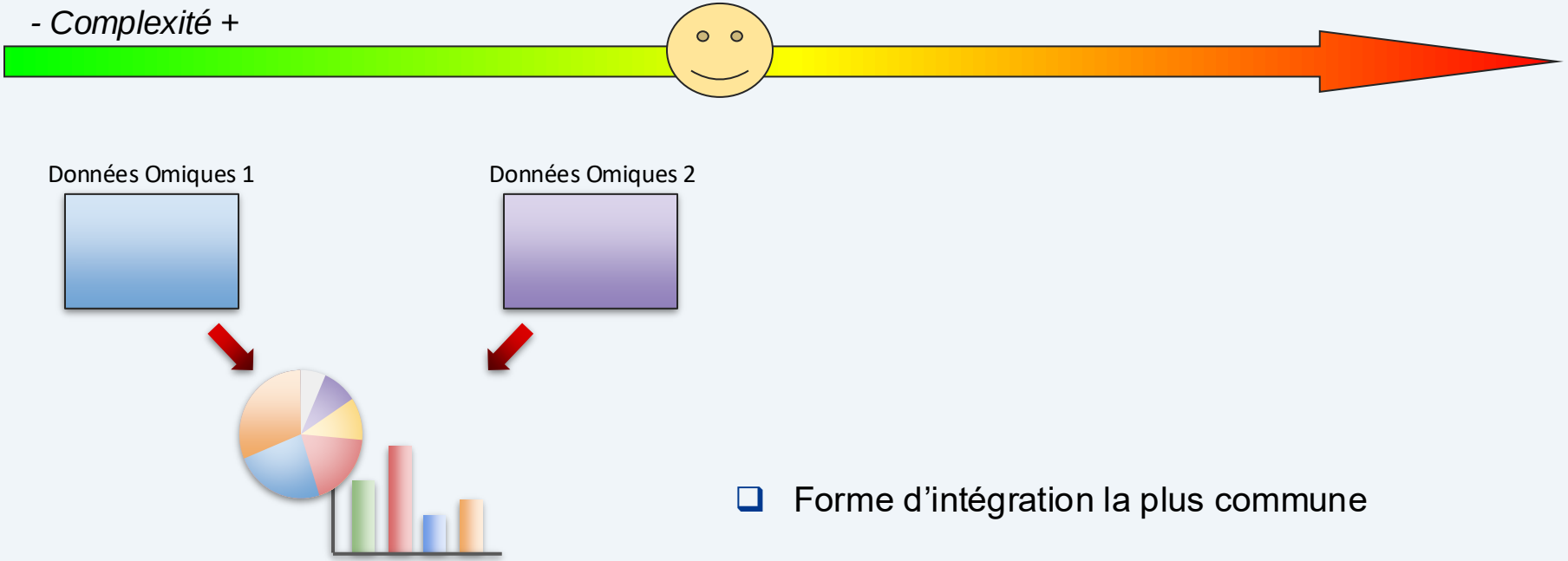


Intégration Conceptuelle



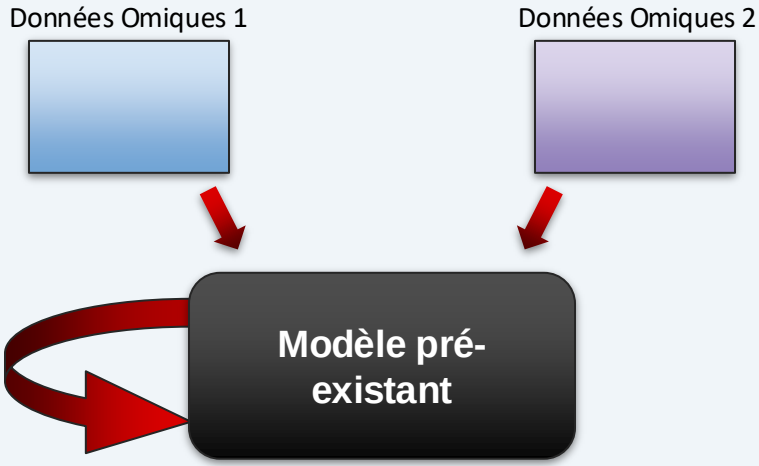
- Analyse de chaque omique séparément
- Manque les relations entre les données

Intégration Statistique



- Forme d'intégration la plus commune
- Différentes méthodes :
  - Corrélation
  - Concaténation
  - Multivariée
  - Pathways Biologiques

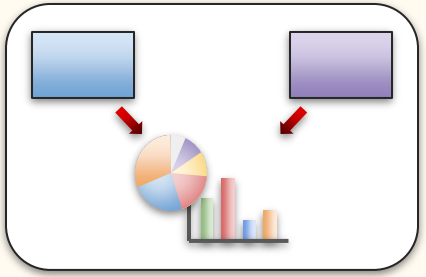
Intégration à base de Modèles



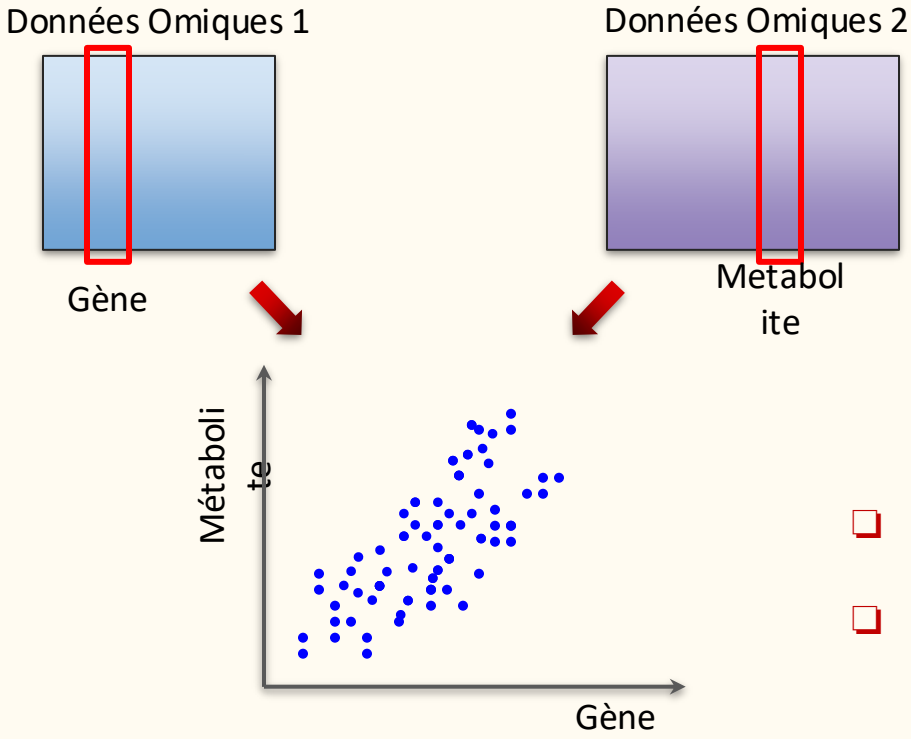
- Idéal à atteindre …

  … mais difficilement atteignable aujourd'hui

Intégration Statistique

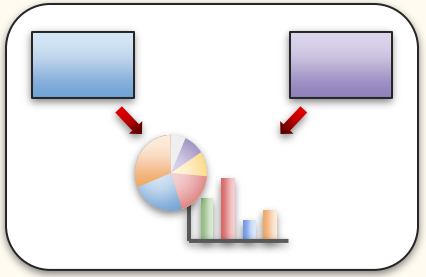


A) Corrélation

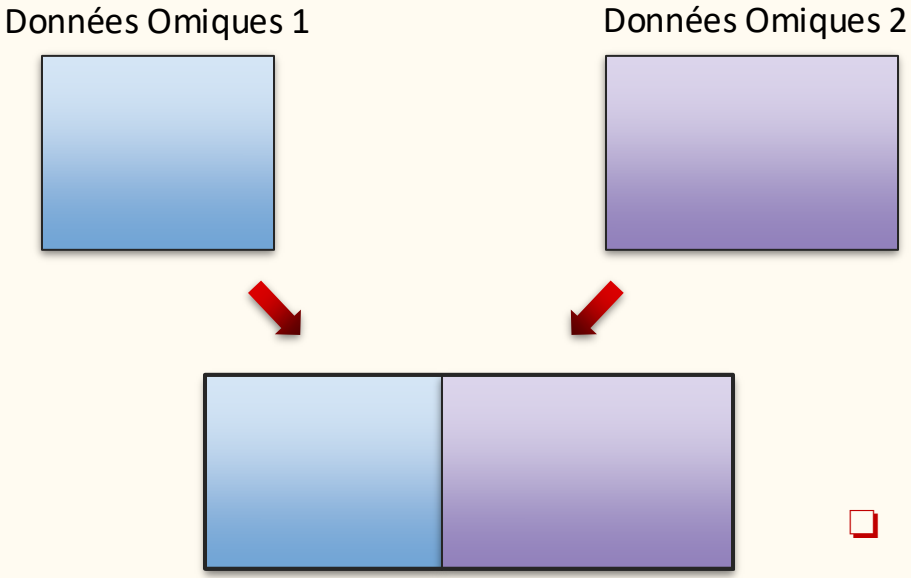


- Méthode la plus simple et plus utilisée
- Corrélation de **Pearson** ou **Spearman**

Intégration Statistique



B) Concaténation



- Concaténation des données en une table
- **Machine Learning**
- Mais : poids des données, sources hétérogènes, schéma de valeurs attendues, ….

Machine learning

- Definition
- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

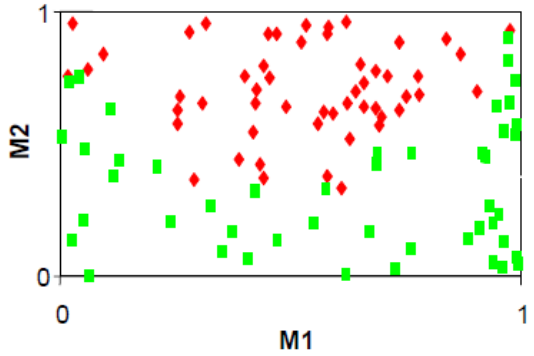
- Learning
  - A computer learns if it improves its predictive performance at some task with experience (i.e. by collecting data)
  - Extracting a model of a system from the sole observation (or the simulation) of this system in some situations

- But: Prédire et mieux caractériser Y ~ variables

Machine learning

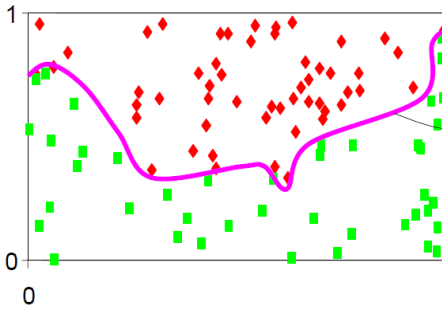
- Ex:
  - Diagnostic medical à partir de 2 (e.g. weights and temperature)
  - Prédire: malade / pas malade

M1	M2	Y
0.52	0.18	Healthy
0.44	0.29	Disease
0.89	0.88	Healthy
0.99	0.37	Disease
...	...	...
0.95	0.47	Disease
0.29	0.09	Healthy



Goal: find a model that classifies at best new cases for which M1 and M2 are known

Learning, model fitting

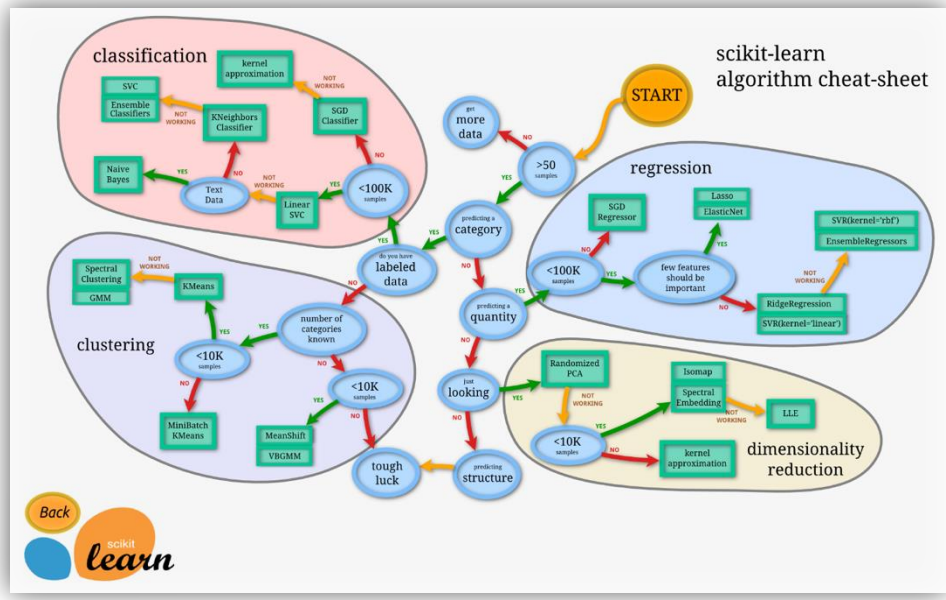


a model obtained by supervised learning

Criteria to evaluate a model:

- Accuracy
- Efficiency (computing time, scalability)
- Interpretability

Machine learning algorithms



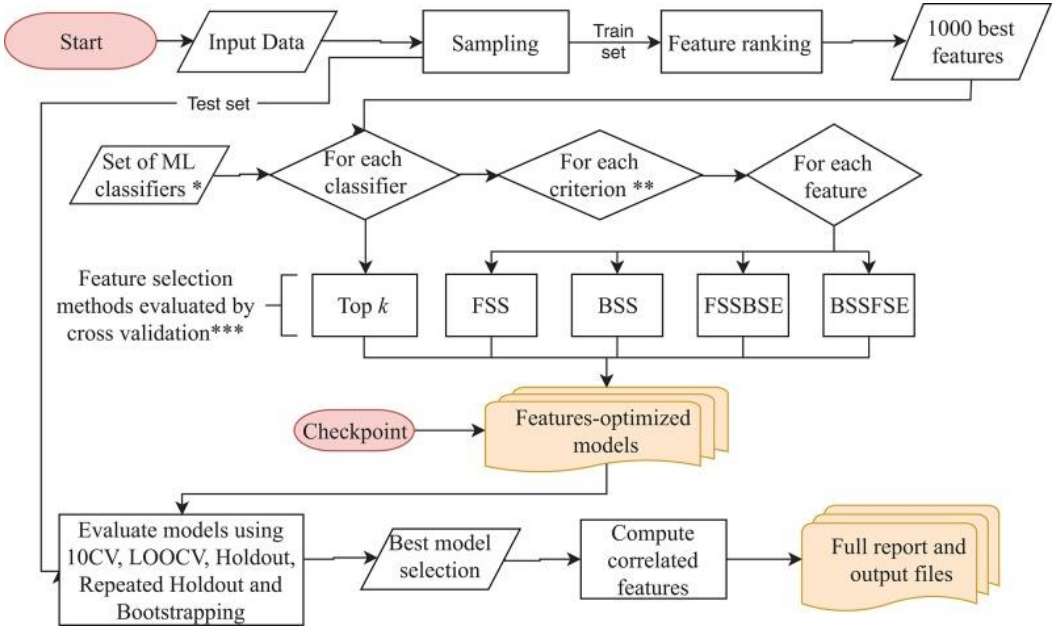
- Biomarker Discovery by Machine Learning
- Detection du meilleur modele et ses hyper-paramètre
- Semi-automatisé
- Parallelisable

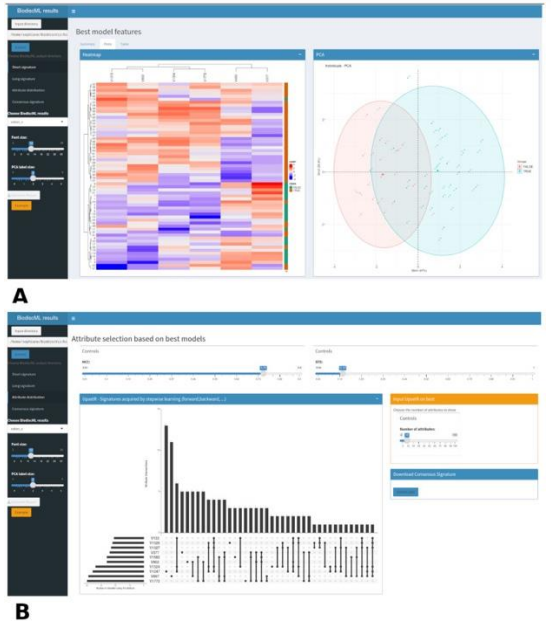
ORIGINAL RESEARCH article
Front. Genet., 16 May 2019 | https://doi.org/10.3389/fgene.2019.00452

BioDiscML

Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data

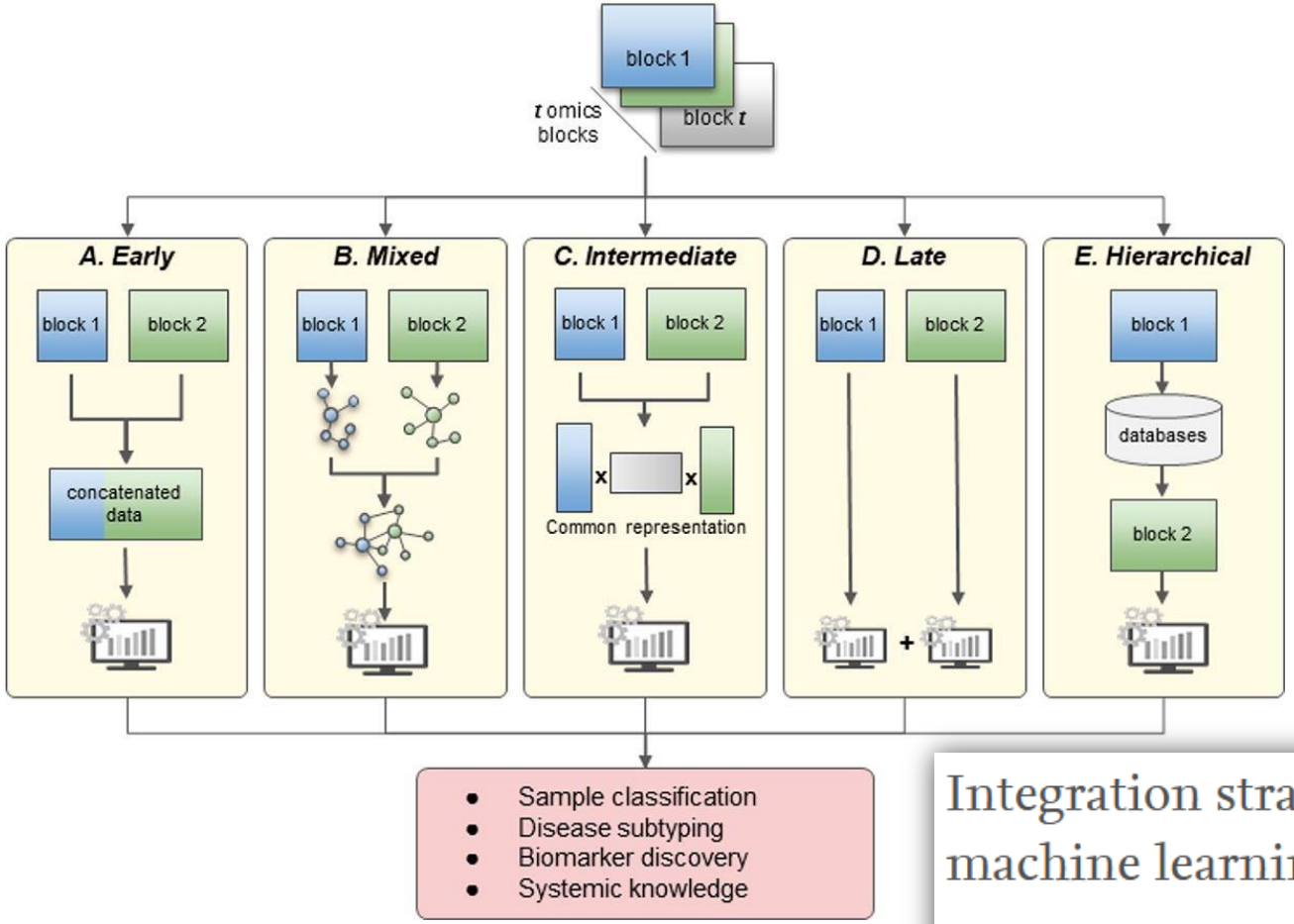
Mickael Leclercq[1,2*], Benjamin Vittrant[1,2], Marie Laure Martin-Magniette[3,4], Marie Pier Scott Boyer[1,2], Olivier Perin[5], Alain Bergeron[1,6], Yves Fradet[1,6] and Arnaud Droit[1,2*†]





BioDiscViz

MO et Machine Learning



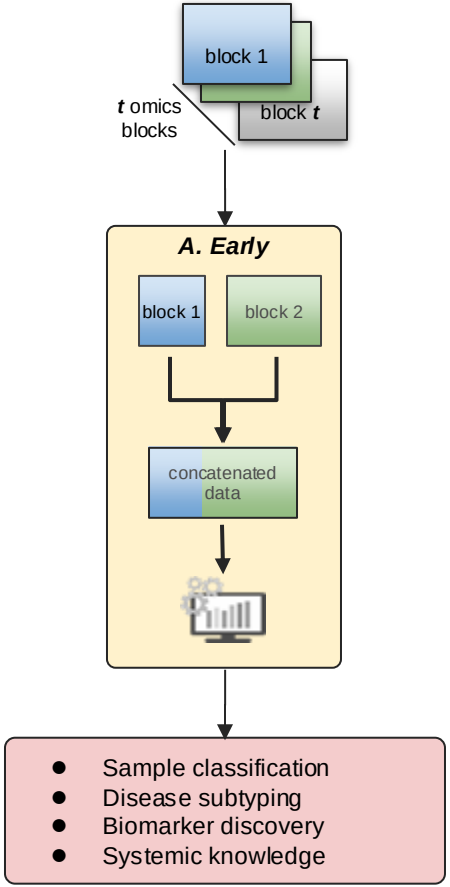
- Données non-supervisées
- **Données supervisées**

Integration strategies of multi-omics data for machine learning analysis

Milan Picard [a], Marie-Pier Scott-Boyer [a], Antoine Bodein [a], Olivier Périn [b], Arnaud Droit [a]

https://doi.org/10.1016/j.csbj.2021.06.030

Early integration



Datasets are concatenated into one matrix

Advantages:

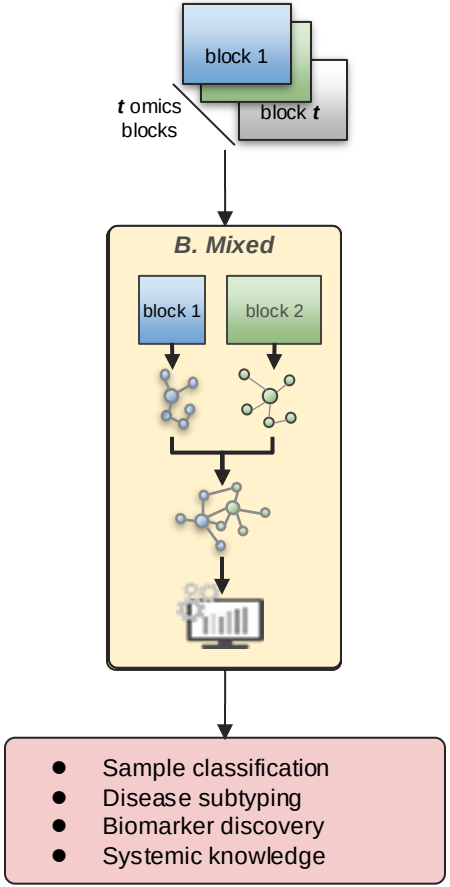
- Easy to implement
- Omics are analysed simultaneously

Drawbacks:

- Increases dimensionality and complexity
- Does not take into account the heterogeneity

Milan Picard, Genomics Festival, Jan 2022

Mixed integration



Datasets are transformed independently, then combined

Advantages:

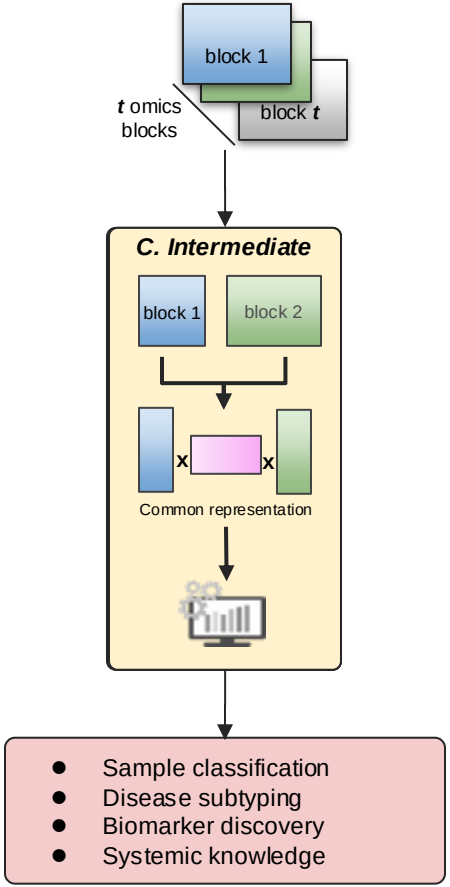
- Deals with dimensionality and complexity
- Can manage multi-omics heterogeneity
- After transformation, omics are analyzed simultaneously

Drawbacks:

- The transformation step is made independently for each omics

Milan Picard, Genomics Festival, Jan 2022

Intermediate integration



Joint integration by finding a common representation

Advantages

- Simultaneous integration
- Deals with dimensionality and complexity

Drawbacks

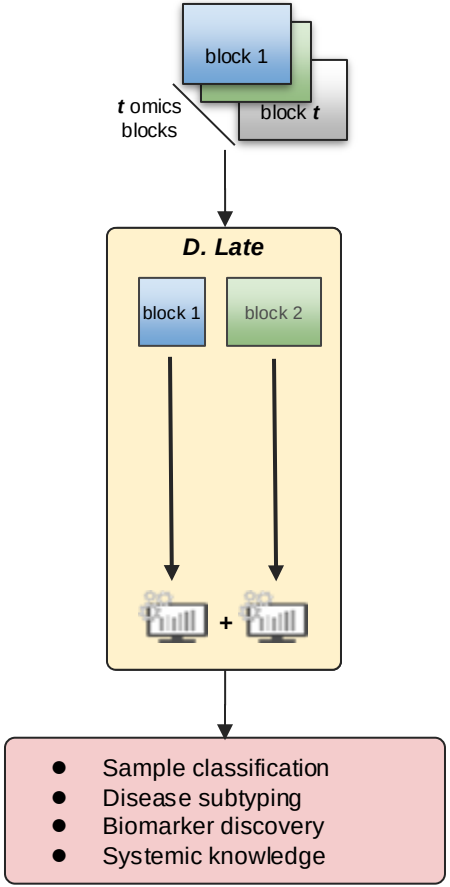
- Mostly unsupervised
- Difficulty integrating prior knowledge and known interactions

Example

- JIVE, MOFA, iClusterBayes, SLIDE, jNMF

Milan Picard, Genomics Festival, Jan 2022

Late integration



Datasets are analyzed separately, final predictions are then combined

Advantages:

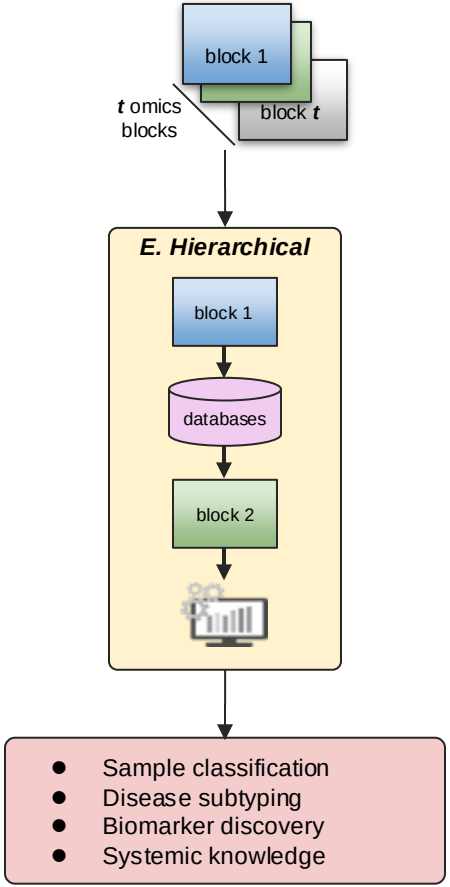
- Easy to implement
- Works with already existing ML models

Drawbacks:

- Datasets are analysed separately

Milan Picard, Genomics Festival, Jan 2022

Hierarchical integration



Datasets are integrated in a way that reflects prior regulatory knowledge in biology

Advantages

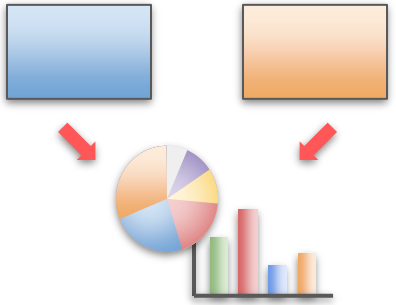
- Make use of known regulatory effects
- More interpretable

Drawbacks

- Prior knowledge isn't always available
- Less generalizable

Milan Picard, Genomics Festival, Jan 2022

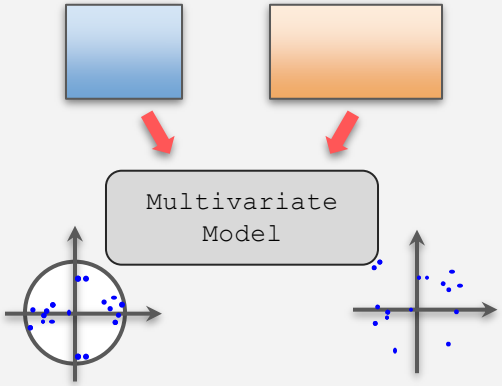
Intégration statistique



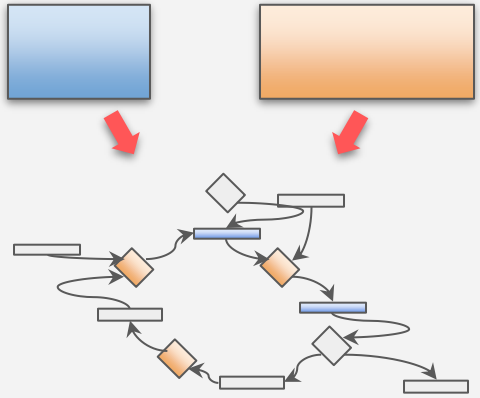
Correlation-based

Concatenation-based

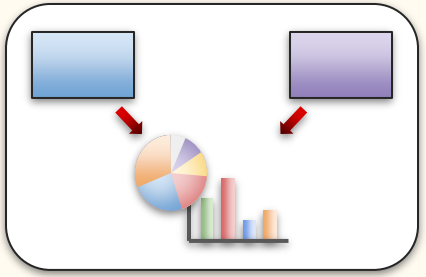
Mulitivariate-based



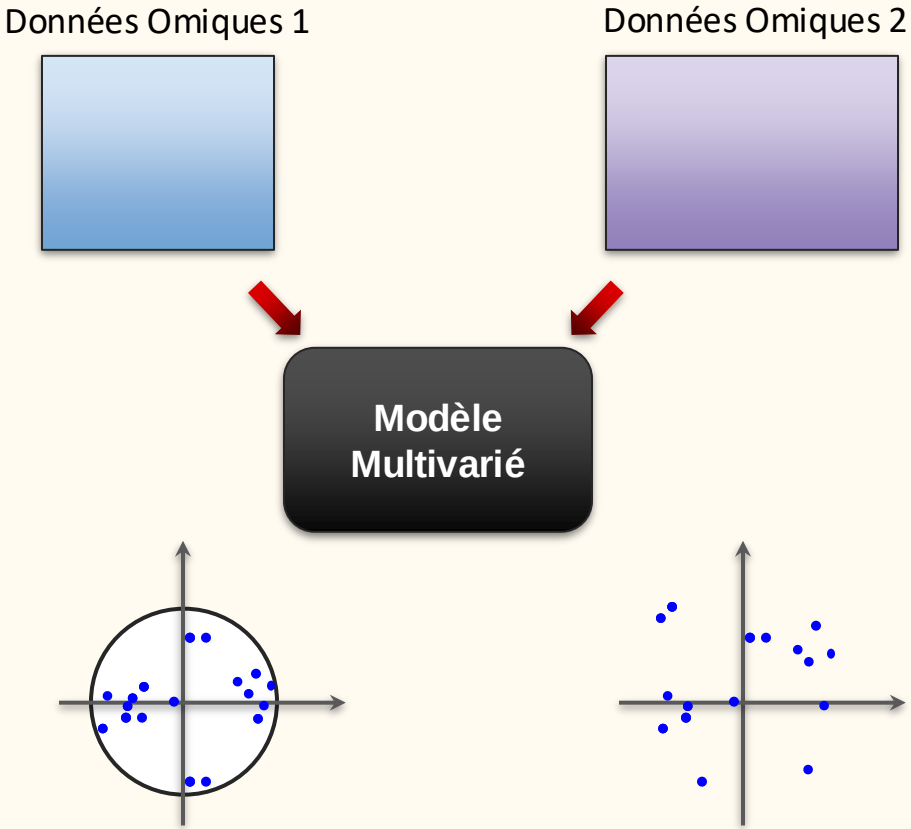
Pathway-based



Intégration Statistique

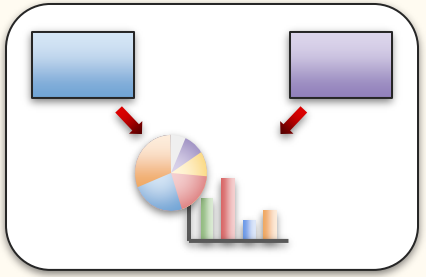


C) Multivarié

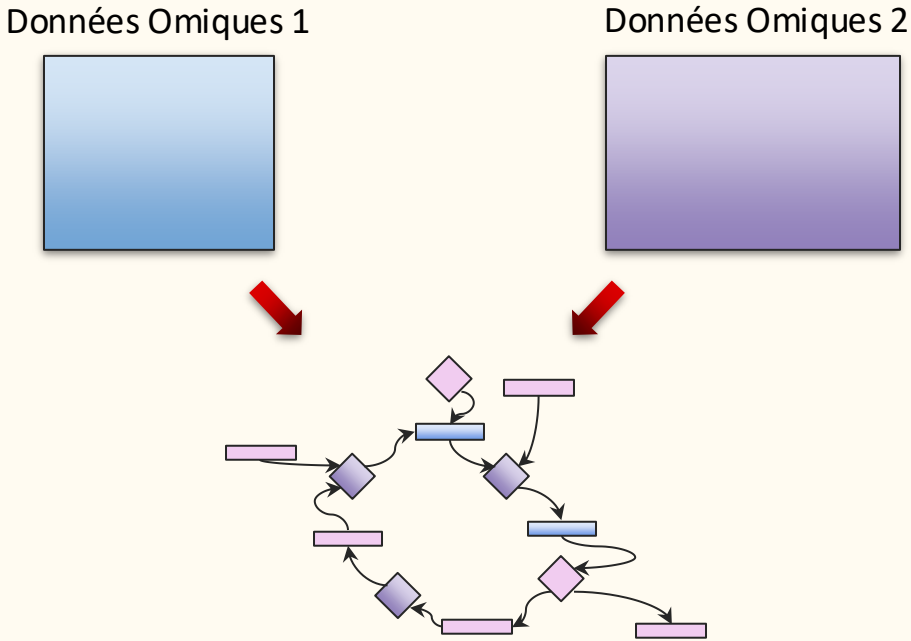


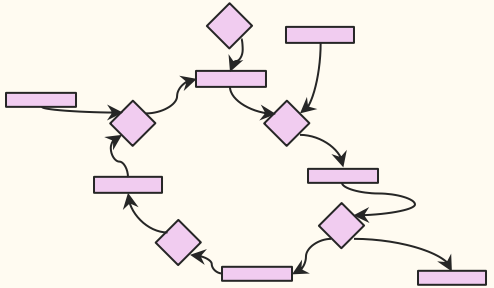
- Méthodes puissantes et à la mode
- Technique de PCA, PLS, …
- Covariance, correlation

Intégration Statistique



D) Pathways Biologiques

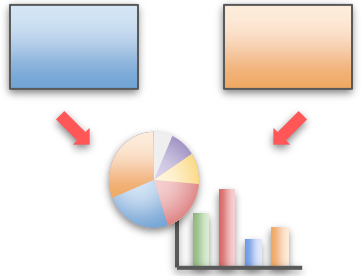




- “*Knowledge driven*”
- Modules fonctionnelles
- Enrichissement

Autres classifications

Intégration statistique



Research | Open Access | Published: 20 January 2016

Methods for the integration of multi-omics data: mathematical aspects

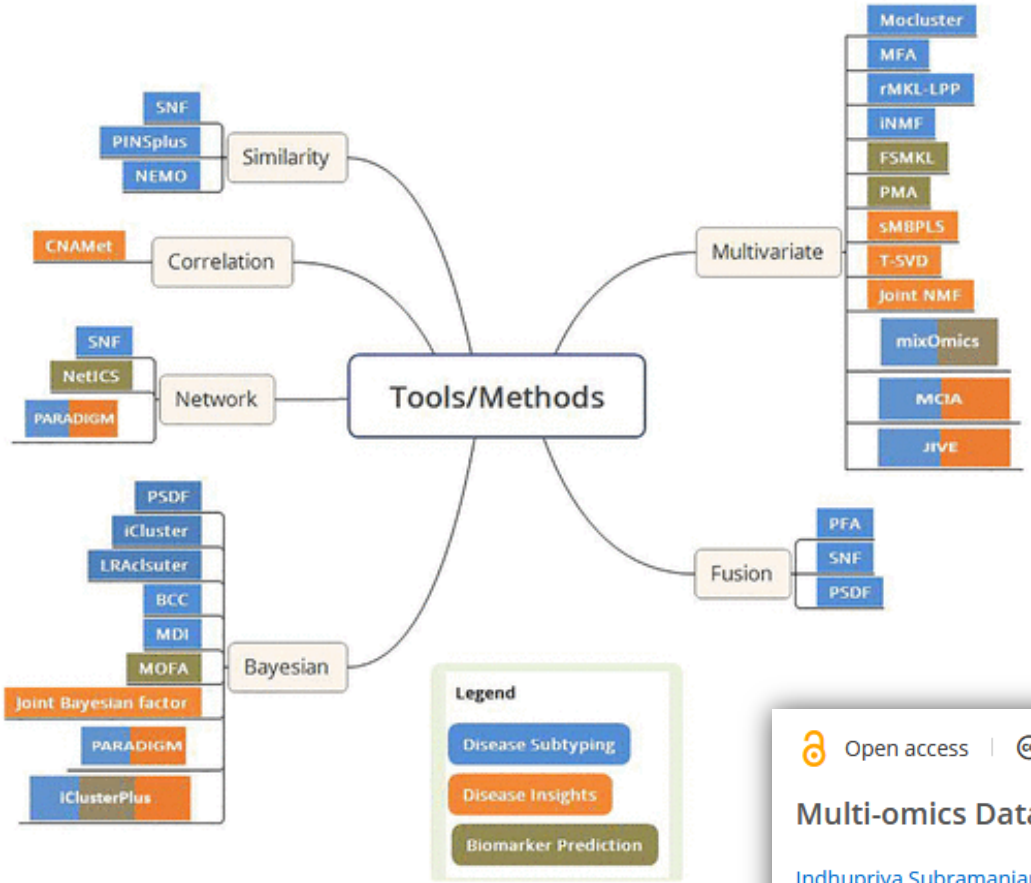
Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani & Luciano Milanesi

BMC Bioinformatics **17**, Article number: S15 (2016) | Cite this article

31k Accesses | **211** Citations | **7** Altmetric | Metrics

	SANS réseaux	AVEC réseaux
Non Bayésien	Corrélation bivariées Approches multivariées (PCA, PLS, RGCCA, ...) ...	SNF Topologie Marche aléatoire ...
Bayésien $P(A\|B) = \frac{P(B\|A)P(A)}{P(B)}$	iCluster MIDI MOFA JIVE ...	PARADIGM CONEXIC ...

Autres classifications



Open access | Review article | First published online January 31, 2020

Multi-omics Data Integration, Interpretation, and Its Application

Indhupriya Subramanian, Srikant Verma, [...], and Krishanpal Anamika, +2 View all authors and affiliations

All Articles | https://doi.org/10.1177/1177932219899051

Type of biological questions

- Sous-typage des maladies et classification
- Prédiction de biomarqueurs
- Compréhension plus générale d'une maladie/phenotype ~ interaction inter-omiques

Mise en contexte

Différents concepts d'integration

Méthodes multivariées

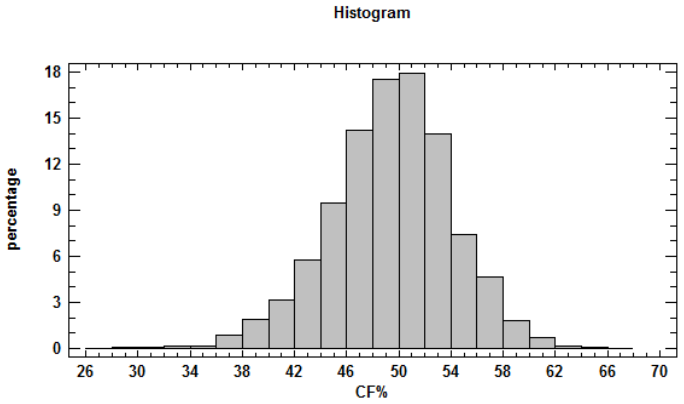
mixOmics

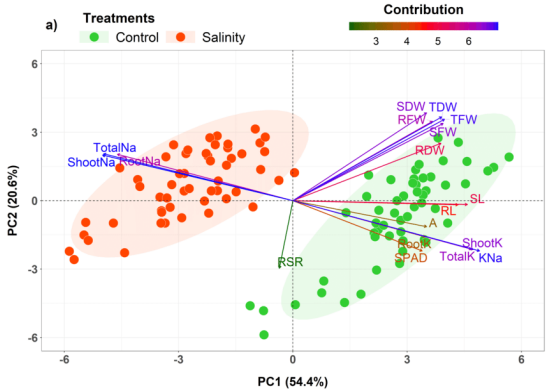
Réseaux en biologie

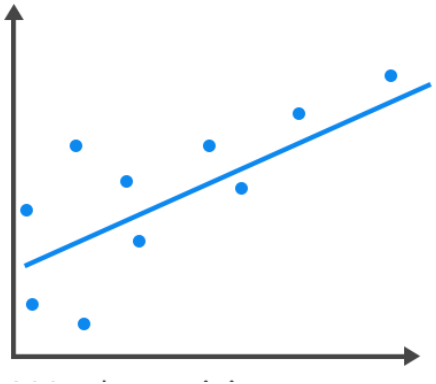
Cas d'étude ADLab

Pourquoi le multivarié ?

- Univarié : hypothesis driven, test statistique et pvalue (t-test, f-test, …)
- Bivarié : x ~ y, cor(x, y), pairwise, combursum for 1000+ genes
- Multivarié: tout à la fois, avec les problems que ça engender
  - Mais visualiation intuitives, et interprétabilité des modèles proposes
  - Importance du plan d'experience
  - Pré-processing des données avant l'intégration



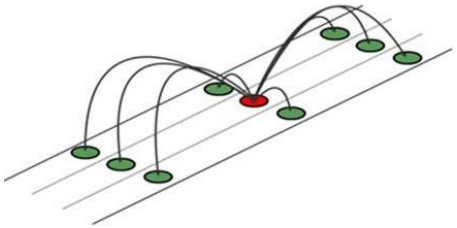




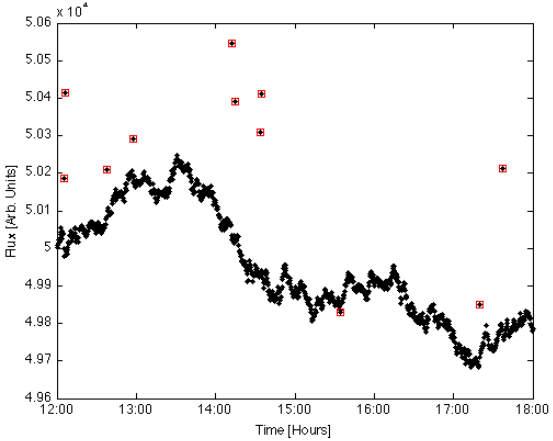
Pré-processing

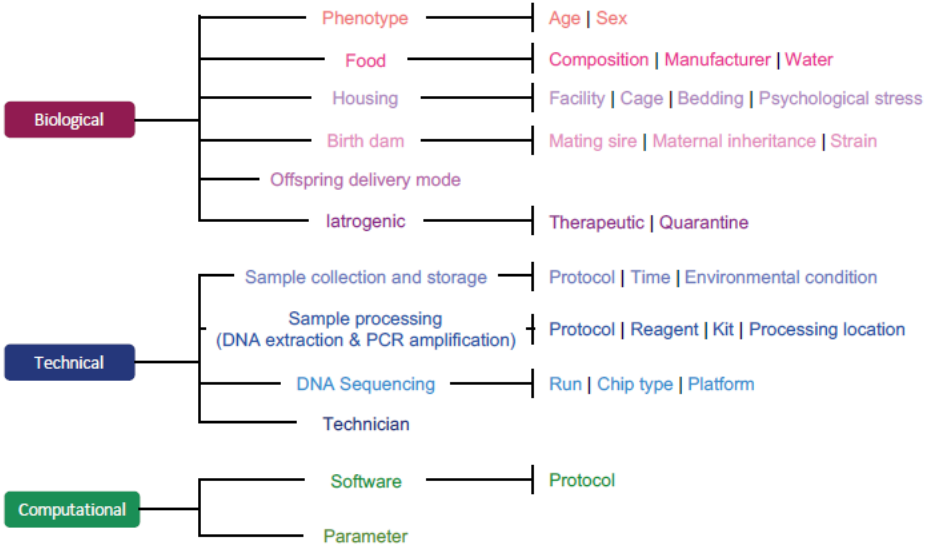
- Bad input = bad model
- Best solution: delete sample having missing data
- Impute missing data
  - mean of the numerical distribution
- Detect erroneous data
  - Outlier detection
- Batch Effect
- Normalisation : échantillons comparables
- Filtering
- Missing value

Imputation



Outlier detection



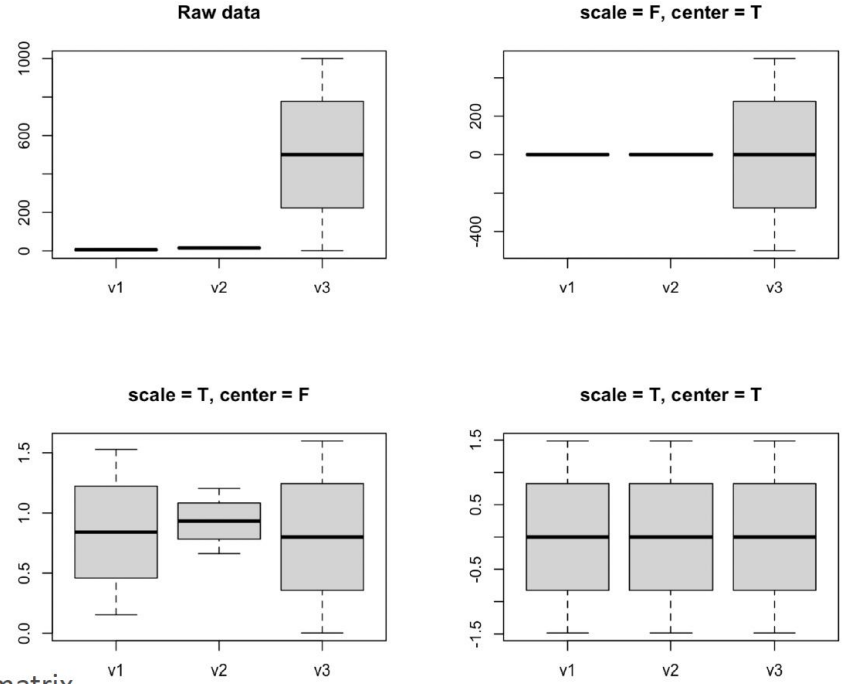


Normalisation

- LOG,
- STANDARDISATION

- scale_value <- data.frame(v1 = 1:10, v2 = 11:20, v3 = seq(from = 1, to = 1000, len = 10))

v1	v2	v3
1	11	1
2	12	112
3	13	223
4	14	334
5	15	445
6	16	556
7	17	667
8	18	778
9	19	889
10	20	1000



scale s'applique sur les colonnes d'un data.frame / matrix

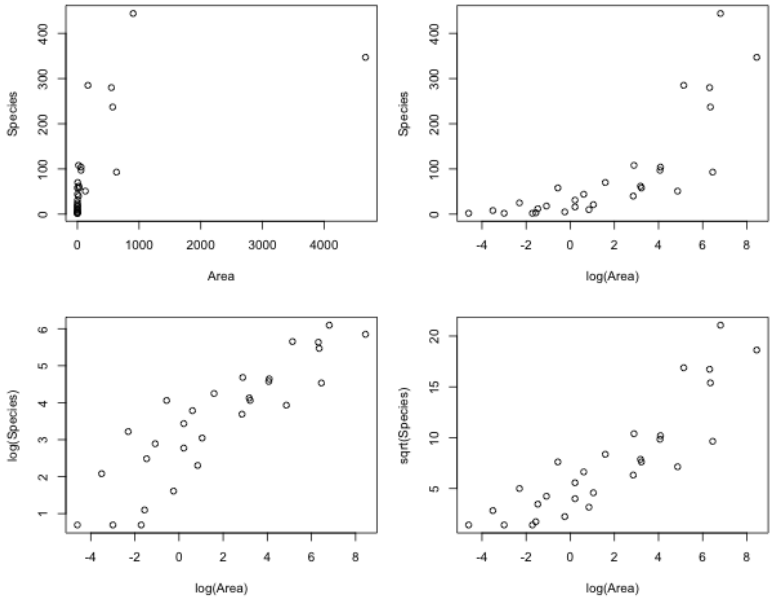
Feature scaling

- Standardize the range of independent variables or features of data
- Methods:
  - Rescaling the range of features to scale the range in [0, 1] or [−1, 1]
  - Where x is an original value, x' is the normalized value $$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$
  - Mean normalization
  - Where x is an original value, x' is the normaliz $$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$
  - Standardization
  - Where x is the original feature vector, is the mean $$x' = \frac{x - \bar{x}}{\sigma}$$ ıre vector, and $\bar{x}$ is its standard deviation σ

50

Feature scaling

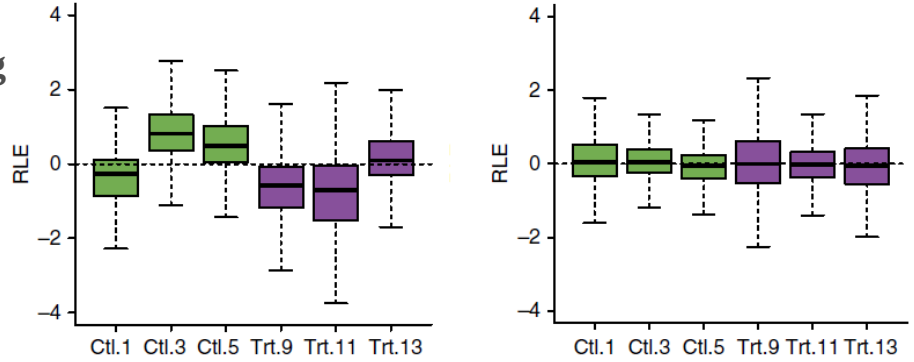
- Transformations:
- Log, square root



- **Normalization by Housekeeping genes (RUV method)**

  www.nature.com/articles/nbt.2931

RLE = relative log expression



L'Analyse en Composante Principale

Les données

Men					
ID	**Shoulders**	**Chest**	**Waist**	**Mass**	**Height**
M1	106.2	89.5	71.5	65.6	174.0
M2	110.5	97.0	79.0	71.8	175.3
M3	115.1	97.5	83.2	80.7	193.5
M4	104.5	97.0	77.8	72.6	186.5
M5	107.5	97.5	80.0	78.8	187.2
M6	119.8	99.9	82.5	74.8	181.5
M7	123.5	106.9	82.0	86.4	184.0
M8	120.4	102.5	76.8	78.4	184.5
M9	111.0	91.0	68.5	62.0	175.0
M10	119.5	93.5	77.5	81.6	184.0

Women					
ID	**Shoulders**	**Chest**	**Waist**	**Mass**	**Height**
W1	105.0	89.0	71.2	67.3	169.5
W2	100.2	94.1	79.6	75.5	160.0
W3	99.1	90.8	77.9	68.2	172.7
W4	107.6	97.0	69.6	61.4	162.6
W5	104.0	95.4	86.0	76.8	157.5
W6	108.4	91.8	69.9	71.8	176.5
W7	99.3	87.3	63.5	55.5	164.4
W8	91.9	78.1	57.9	48.6	160.7
W9	107.1	90.9	72.2	66.4	174.0
W10	100.5	97.1	80.4	67.3	163.8

L’Analyse en Composante Principale

Comment représenter les données ?

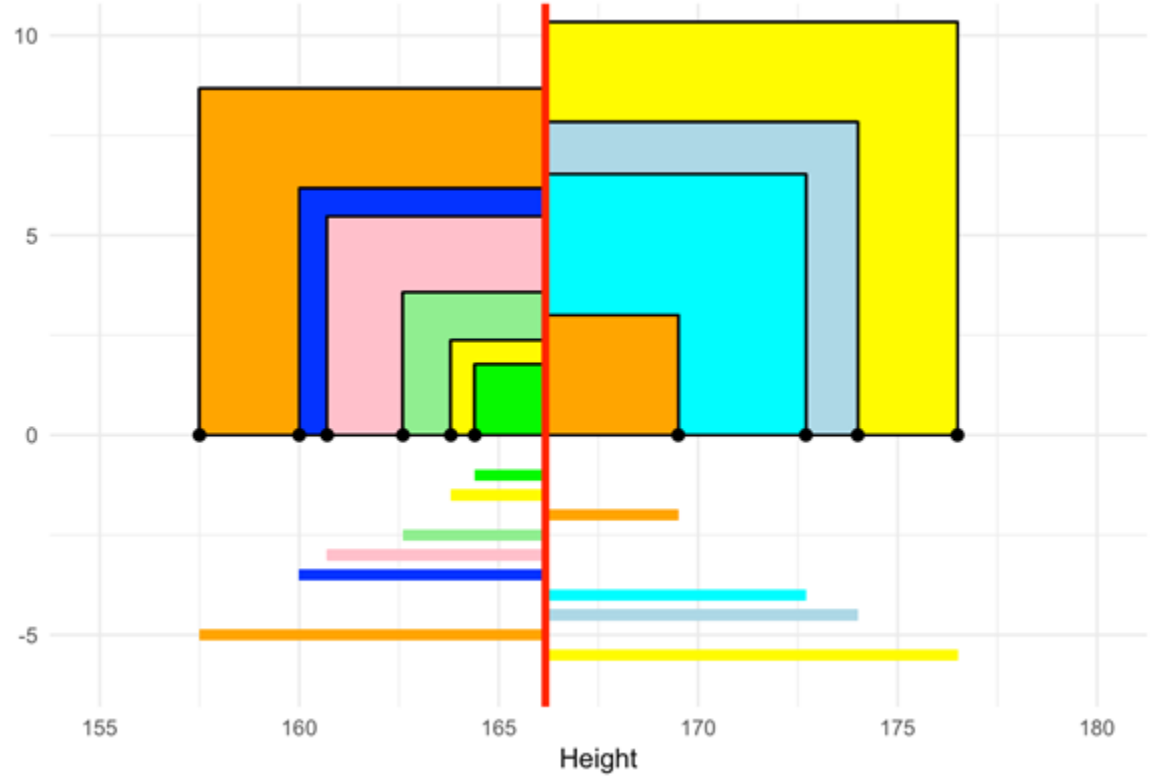
L'Analyse en Composante Principale

Indicateur de dispersion

- La **moyenne** $$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
- L'**étendue** $$E = x_{\max} - x_{\min}$$
- La **variance** $$V = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$
- L'**écart-type** $$\sigma = \sqrt{V} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

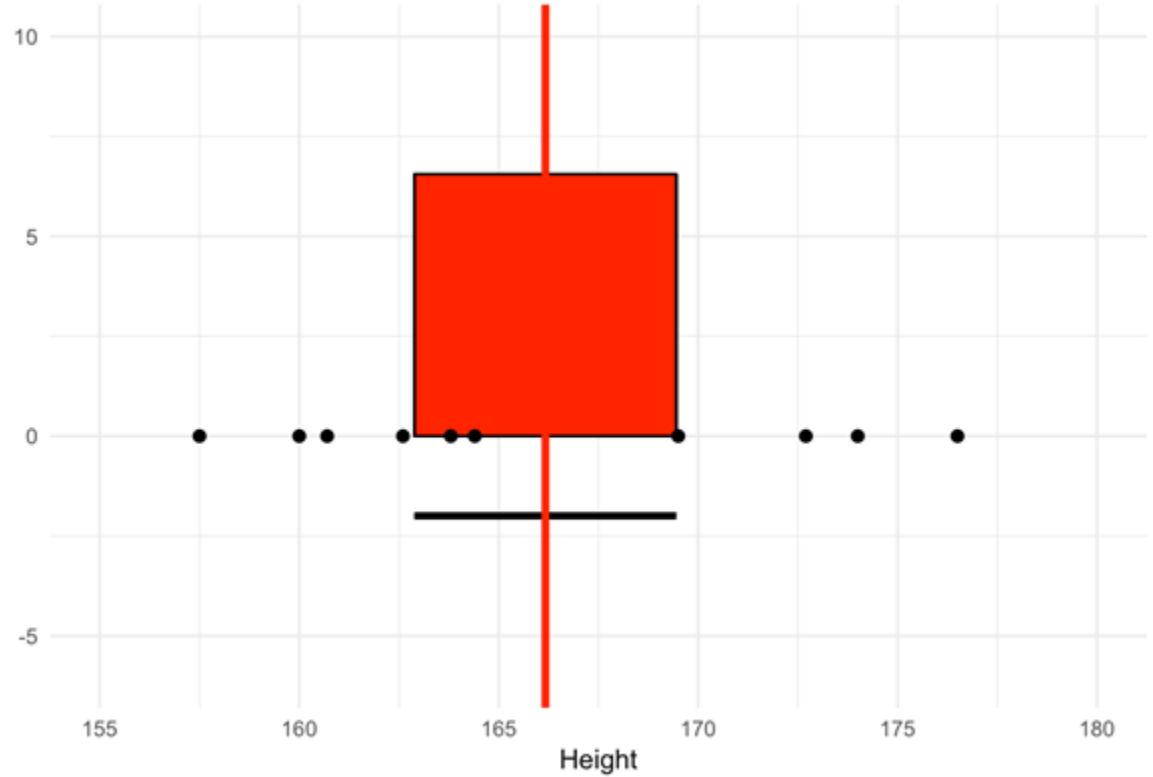
L’Analyse en Composante Principale

Variance et Écart-type



L’Analyse en Composante Principale

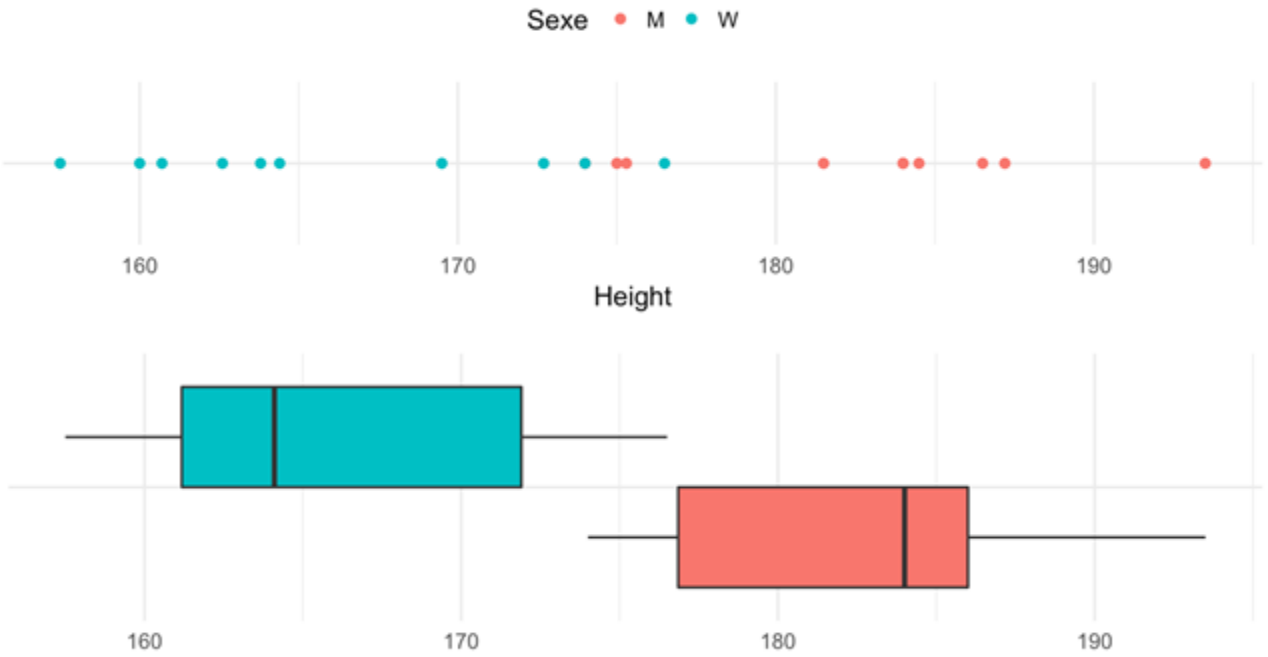
Variance et Écart-type



L'Analyse en Composante Principale

Comment représenter les données ?

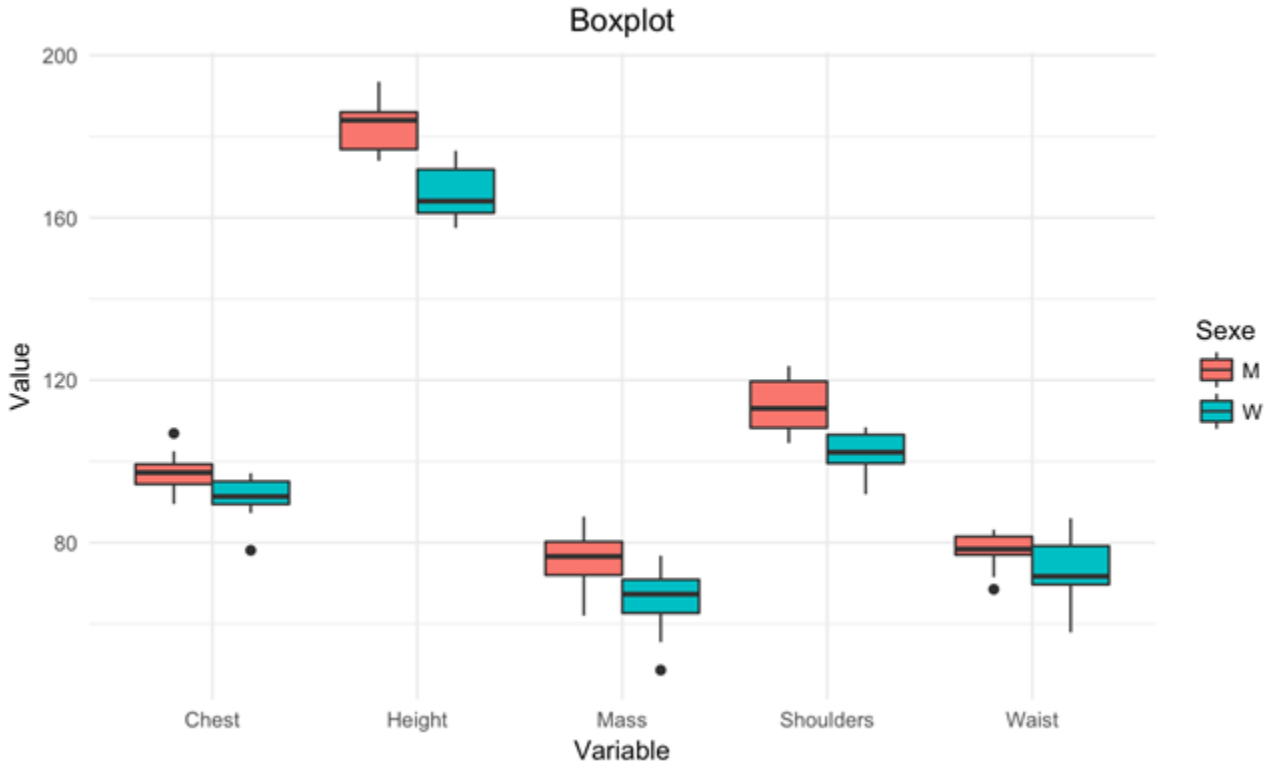
1D



L'Analyse en Composante Principale

Comment représenter les données ?

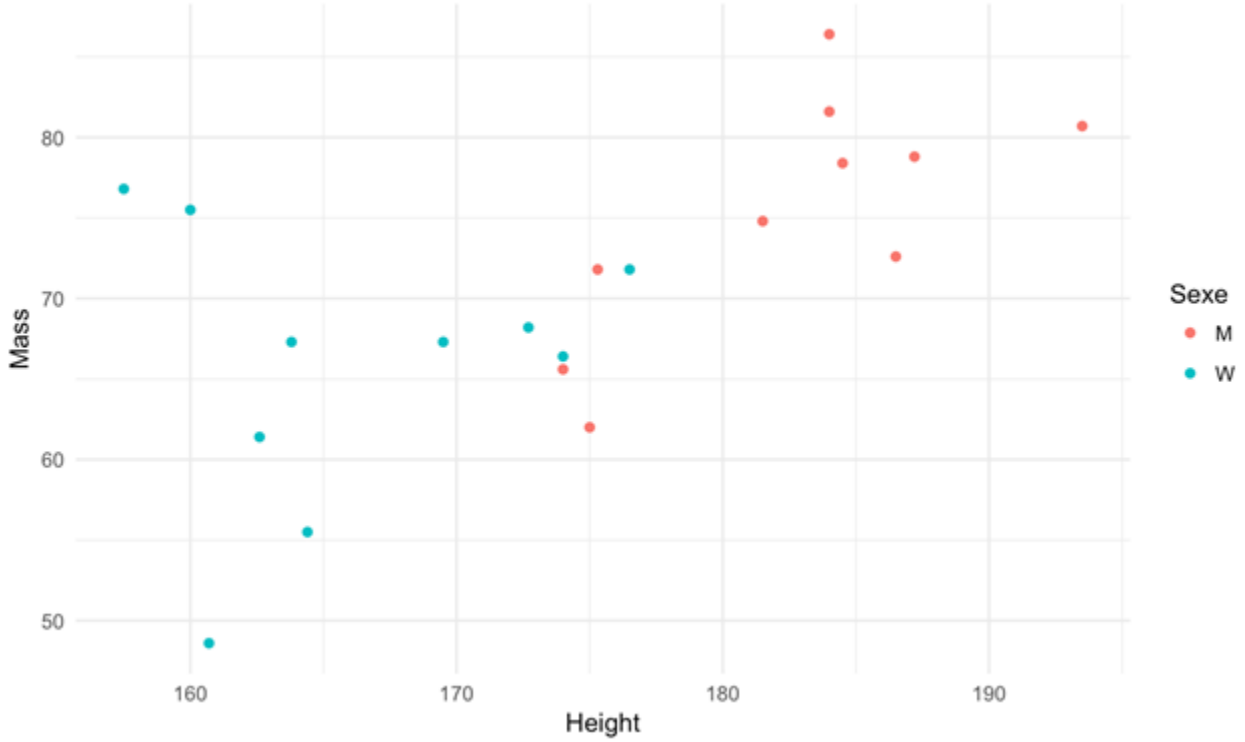
1D



L'Analyse en Composante Principale

Comment représenter les données ?

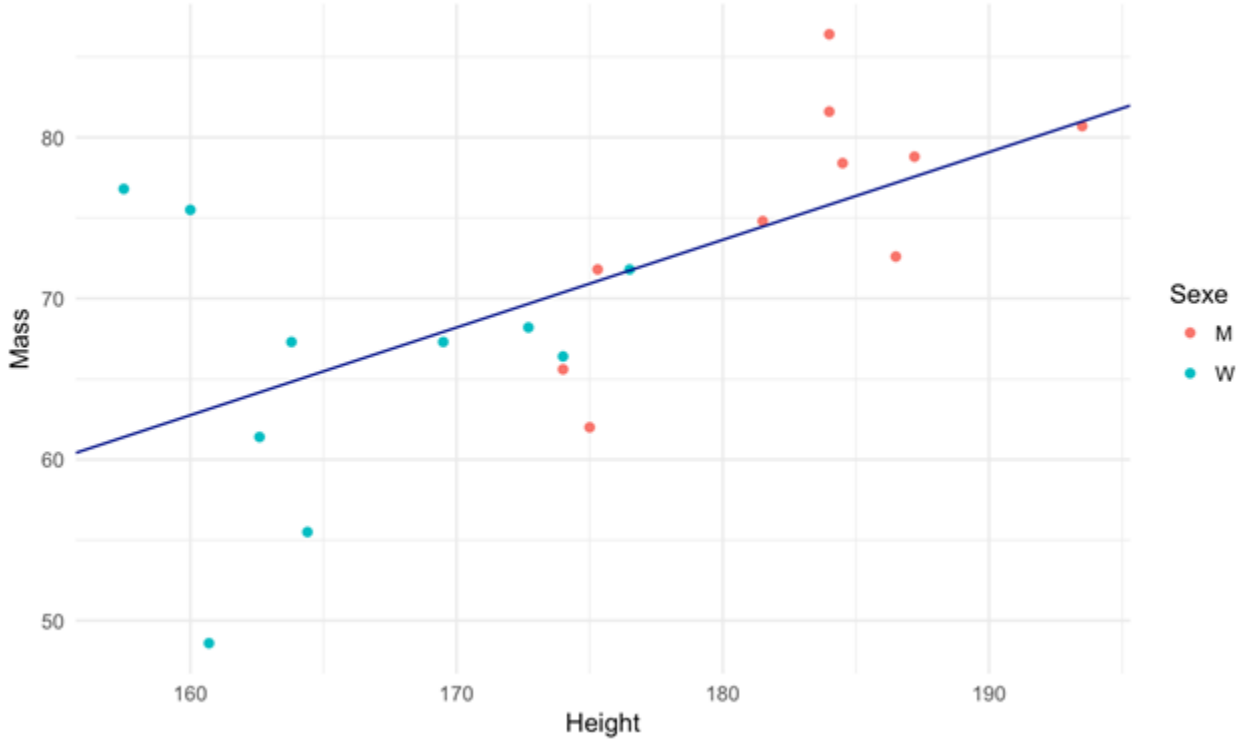
2D



L'Analyse en Composante Principale

Comment représenter les données ?

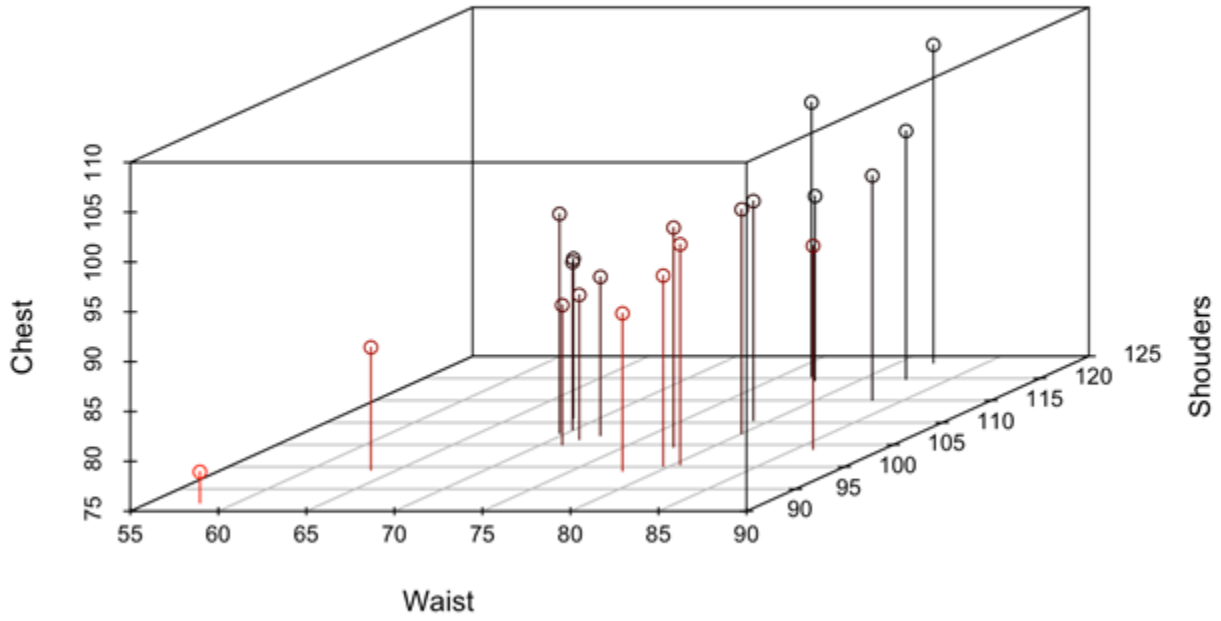
2D



L'Analyse en Composante Principale

Comment représenter les données ?

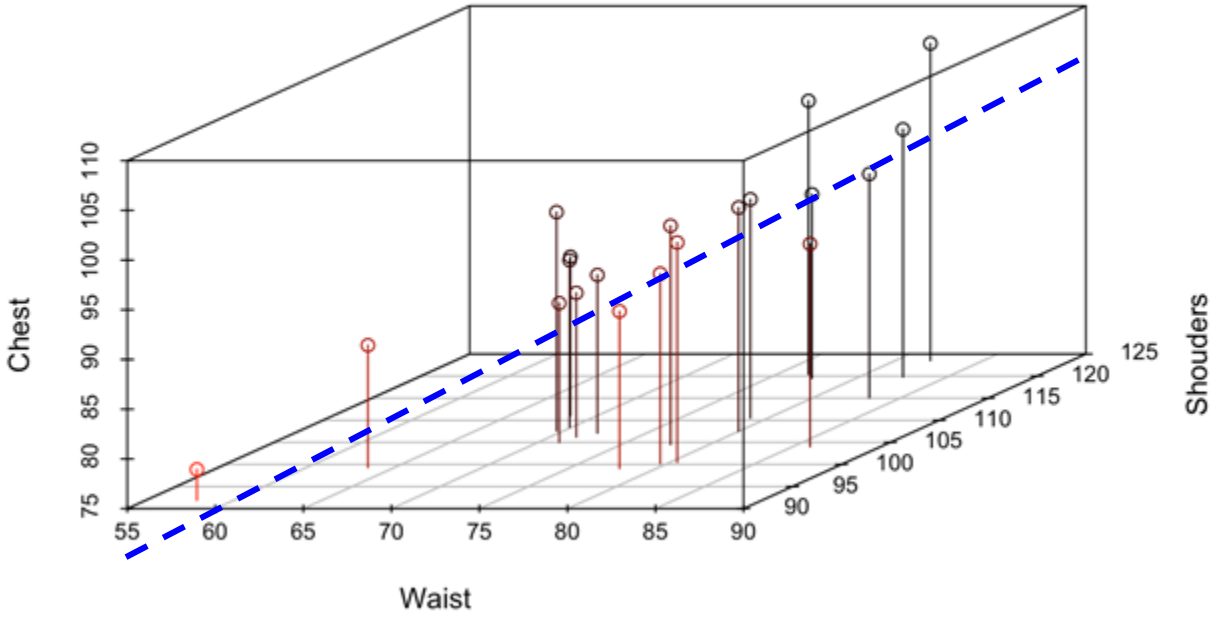
3D



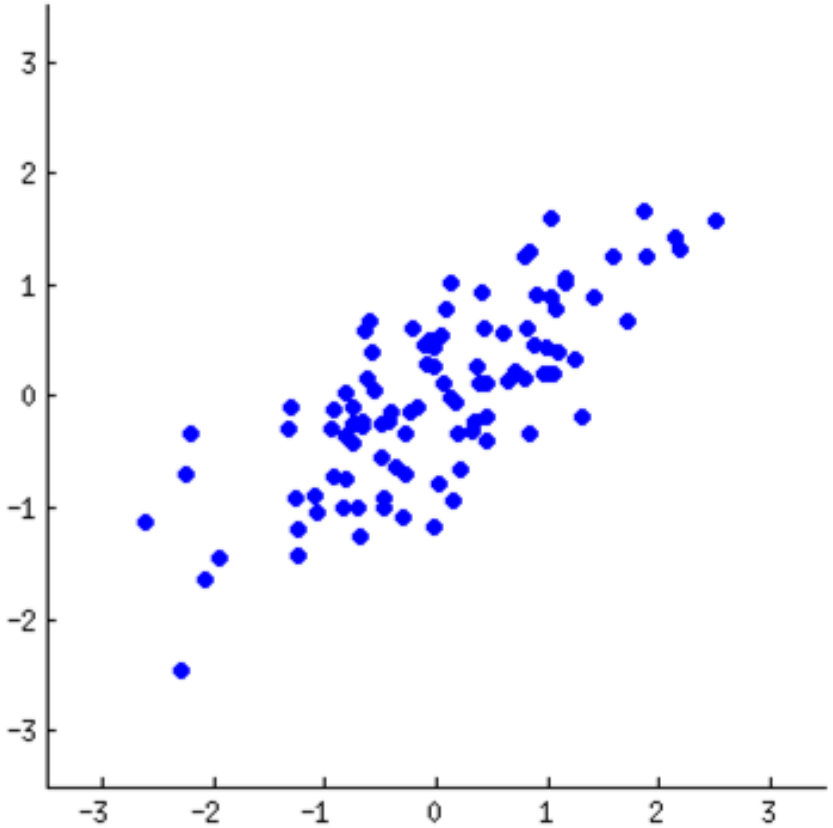
L'Analyse en Composante Principale

Comment représenter les données ?

3D

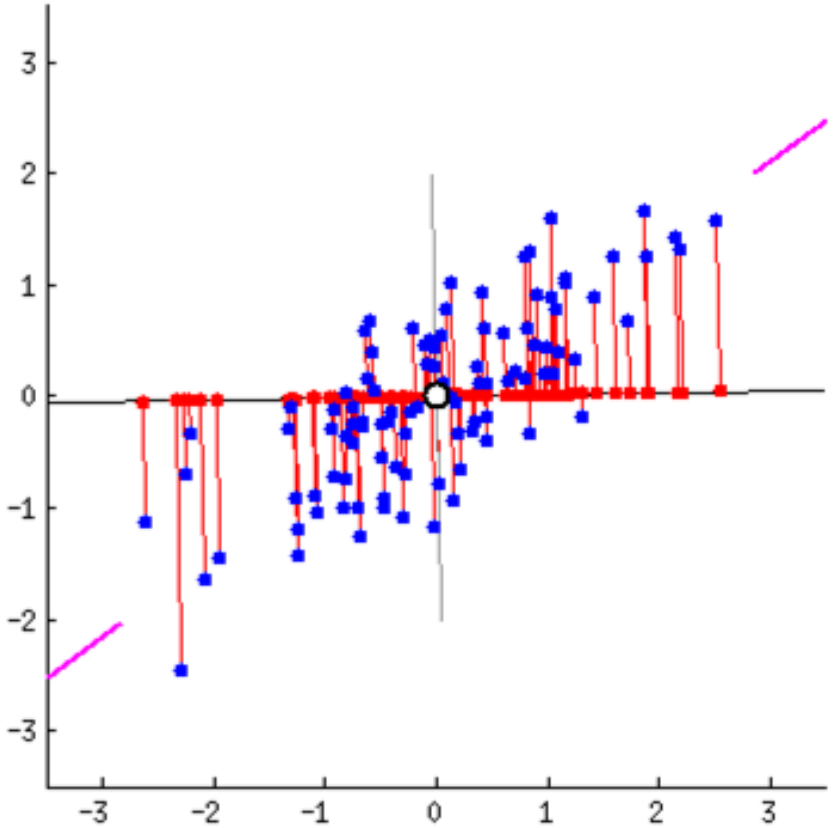


L’Analyse en Composante Principale



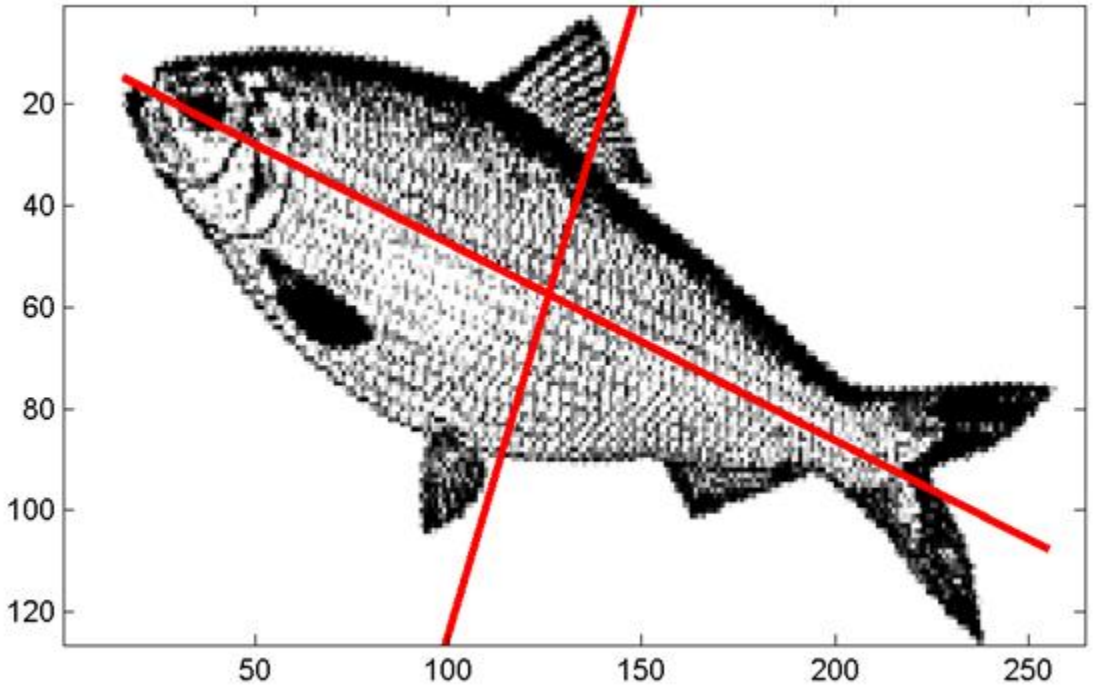
https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

L’Analyse en Composante Principale



https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

L’Analyse en Composante Principale



https://upload.wikimedia.org/wikipedia/commons/9/90/PCA_fish.png

L'Analyse en Composante Principale

X =

	Gene-A	Gene-B	Gene-C
Ech1	10	5	6
ECh2	42	49	2
...			

components =

	PC1	PC2	PC3
Ech1	6.1	1.2	1.5
ECh2	42.9	8.6	2.5
...			

0.2 x *(a1)*

Gene-A
10
42

\+ 0.7 x *(a2)*

Gene-B
5
49

\+ 0.1 x *(a3)*

Gene-C
6
2

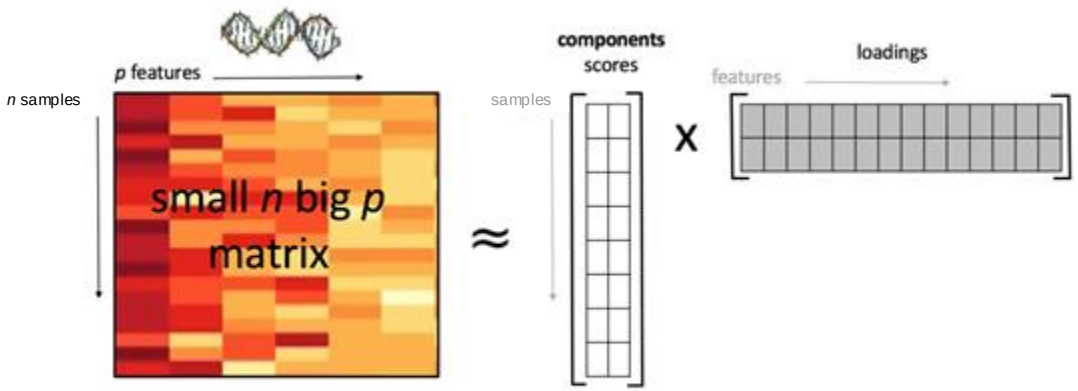
=

Linear Comb.
6.1
42.9

a = loading vectors

= PC

L'Analyse en Composante Principale



$$\arg \max_{||a^h||=1} \text{var}(Xa^h)$$

Solved with SVD:

$$X = U\Delta A^T$$

Singular vectors:

- $T = U\Delta$, T contains the PCs $\boldsymbol{t}^h$
- A contains the loading vectors $\boldsymbol{a}^h$

Singular values:

- Δ diagonal matrix with $\sqrt{\delta_h}$

PCA dimension: $h = 1..H$

X (n x p)	is decomposed into:	• principal components PC • loading vectors associated to each PC => such that var(PC) is max.

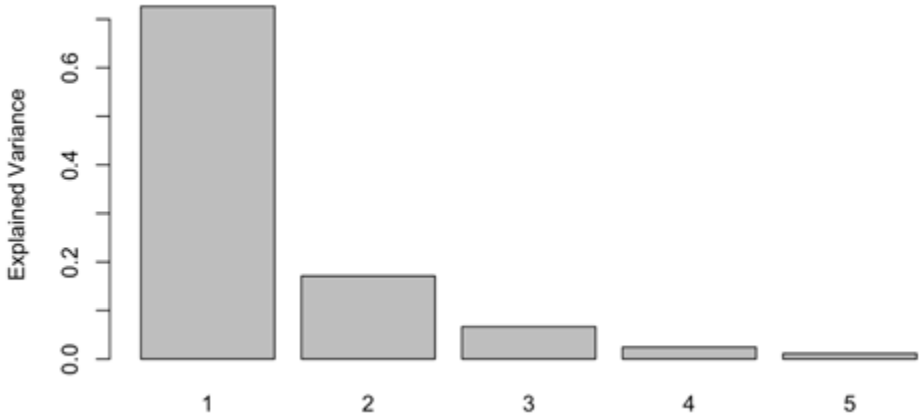
The variance of the first principal component $\boldsymbol{t}^1$ is the largest ($= \delta_1$).

The eigenvalues δ_h decrease and correspond to the explained variance per component.

Adapted from from K-A Lê Cao

Exemple d'ACP - Nombre de composantes

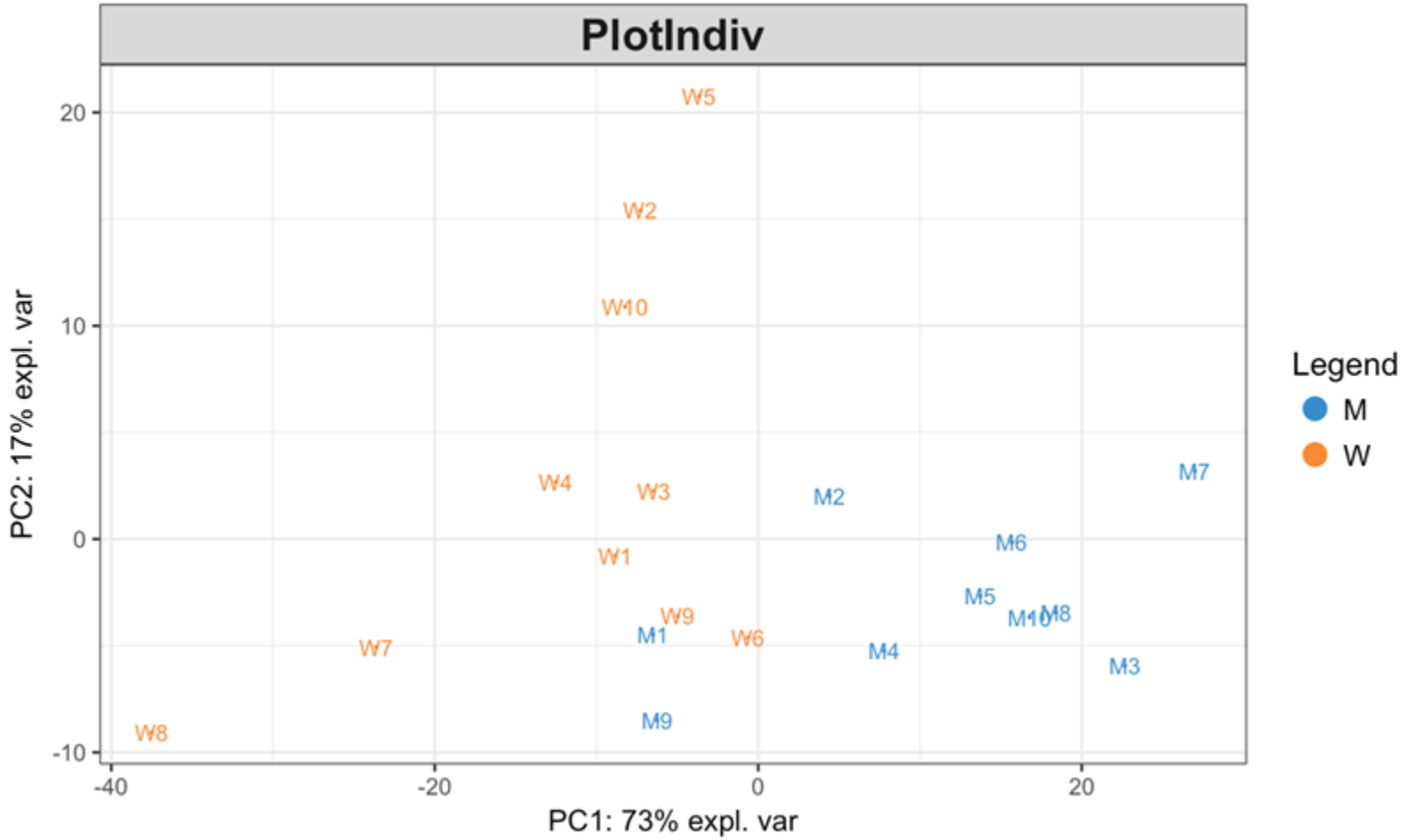
```
## Eigenvalues for the first 5 principal components, see object$sdev^2:
##        PC1        PC2        PC3        PC4        PC5
## 255.655833  60.183388  23.484105   8.607990   4.010947
##
## Proportion of explained variance for the first 5 principal components, see object$explained_variance:
##        PC1        PC2        PC3        PC4        PC5
## 0.72641413 0.17100358 0.06672715 0.02445853 0.01139661
##
## Cumulative proportion explained variance for the first 5 principal components, see object$cum.var:
##       PC1       PC2       PC3       PC4       PC5
## 0.7264141 0.8974177 0.9641449 0.9886034 1.0000000
```



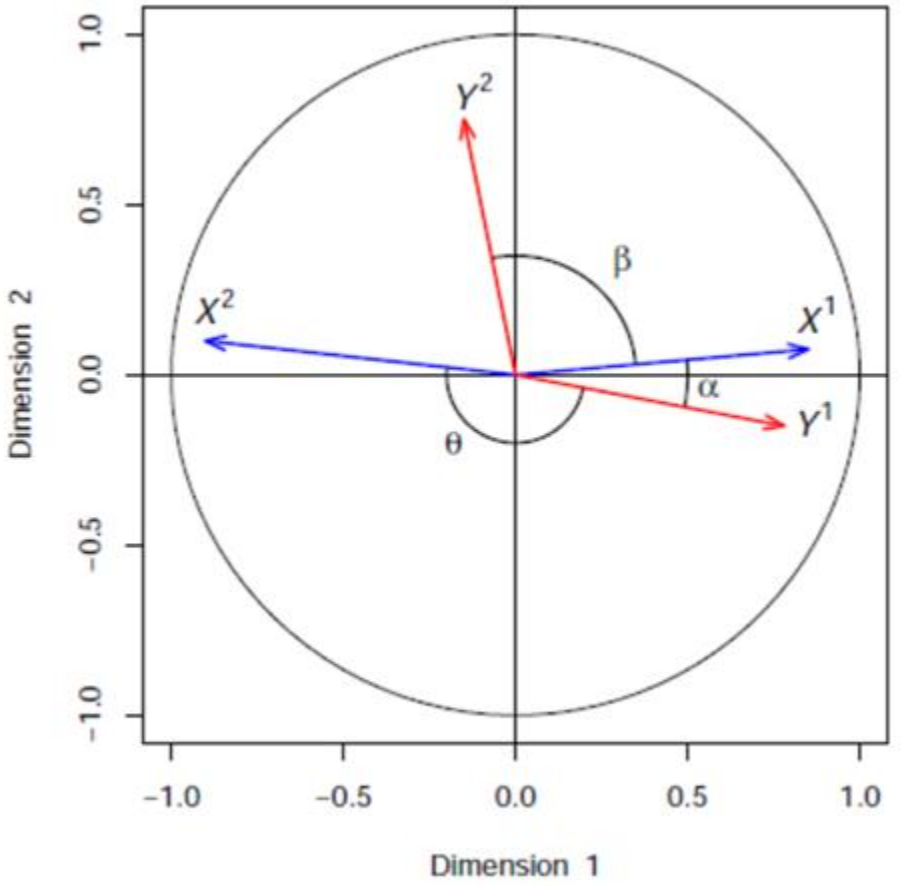
- Critère de Kaiser
- Seuil de variance expliquée
- Critère du coude

```
> plot(pca.res)
```

Exemple d'ACP - graphe des individus

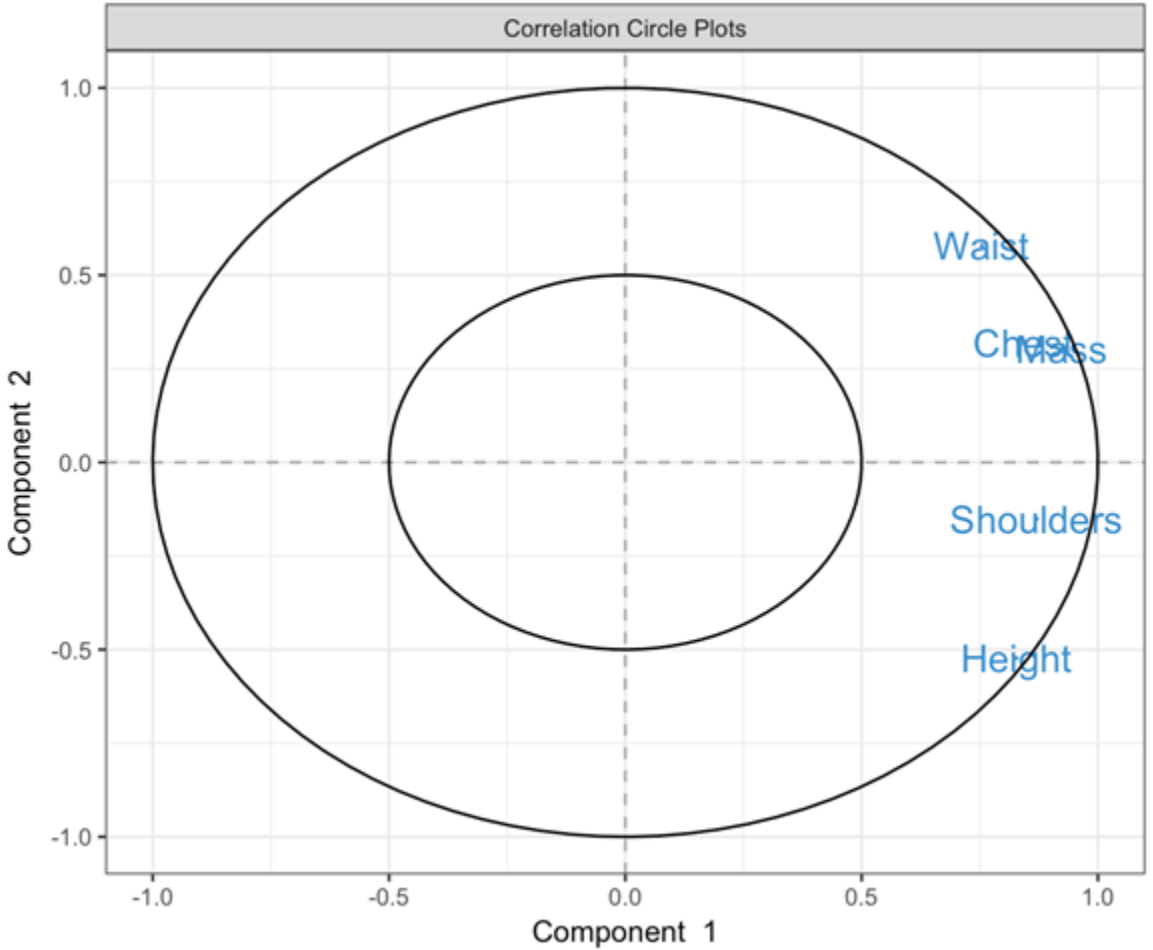


Exemple d'ACP - graphe des variables

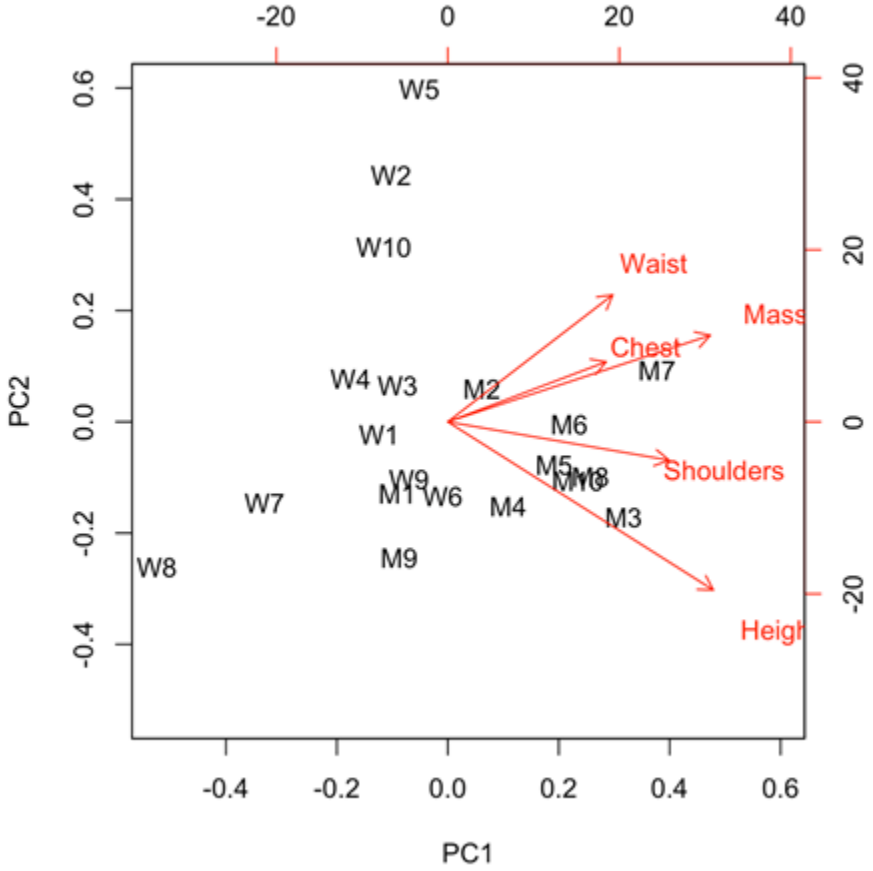


- Hypersphères des corrélations
- Corrélation entre 1 variables :
  - positive
  - négative
  - nulle
- Contribution aux axes

Exemple d'ACP - graphe des variables

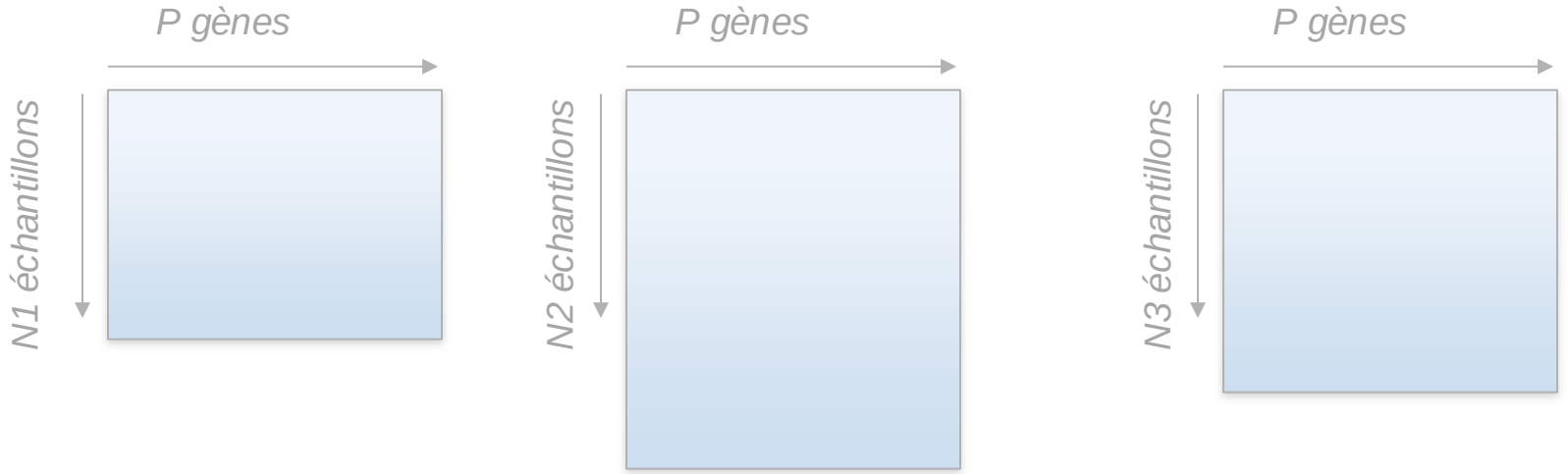


Exemple d'ACP - biplot



N-Intégration

P-Intégration

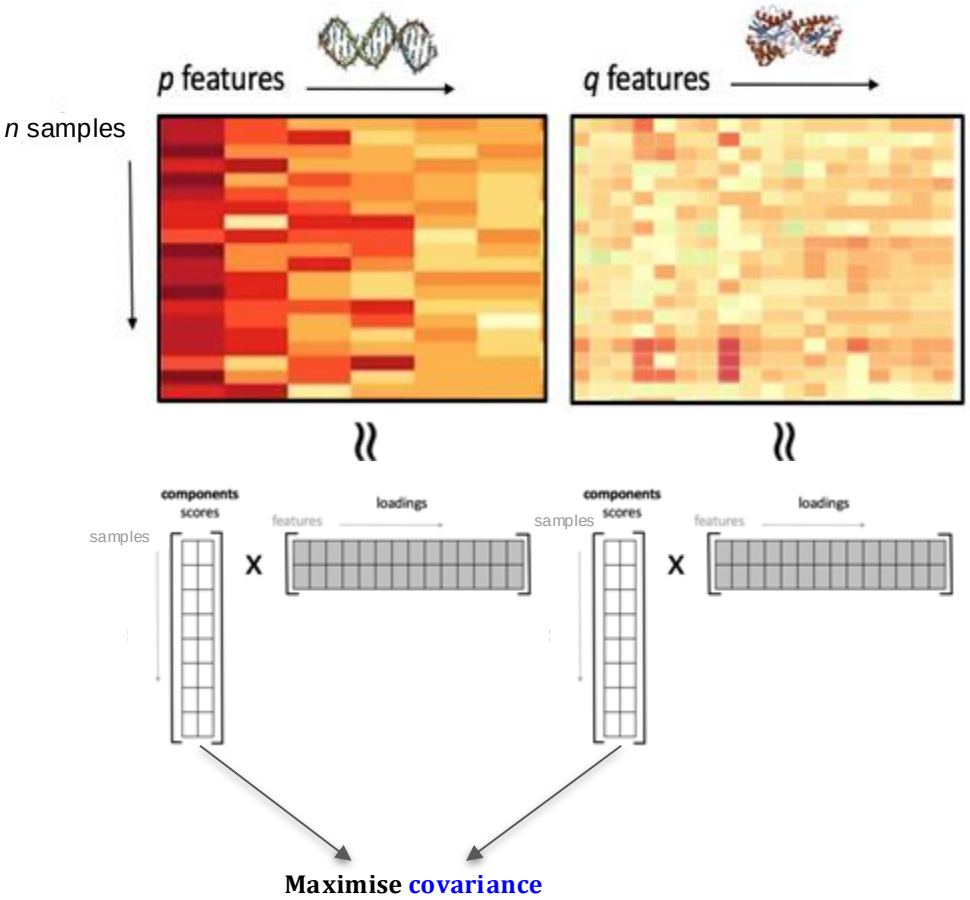


PLS: Projection on Latent Structures

PLS maximises the **covariance** between 2 sets of data

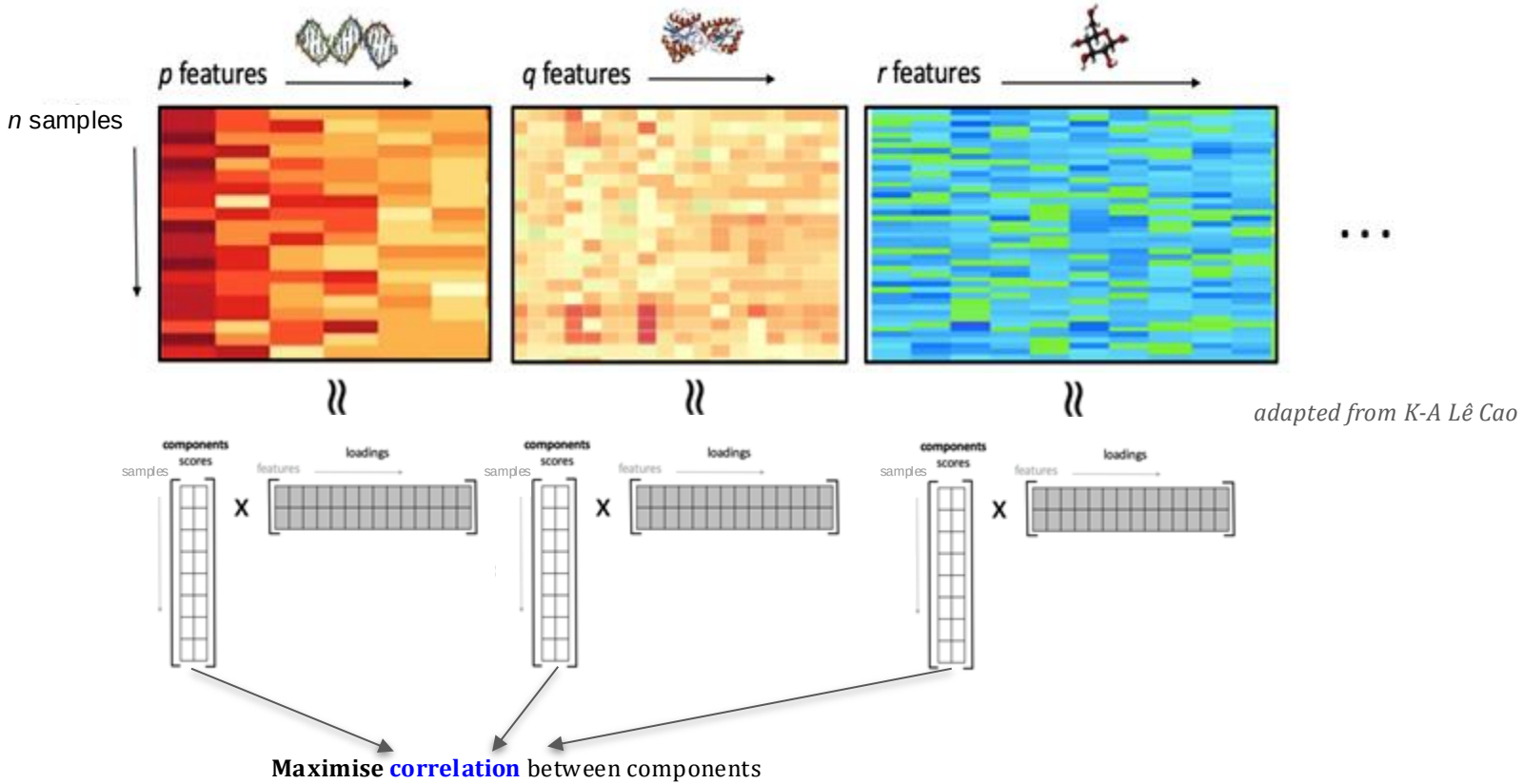
$$\arg \max_{||a^h||=1, ||b^h||=1} \text{cov}(Xa^h, Yb^h)$$

Loading vectors obtained from $\texttt{svd}(X_1^T X_2)$



Adapted from from K-A Lê Cao

Multi-block PLS



adapted from K-A Lê Cao

Adapted from from K-A Lê Cao

Mise en contexte

Différents concepts d'integration

Méthodes multivariées

mixOmics

Réseaux en biologie

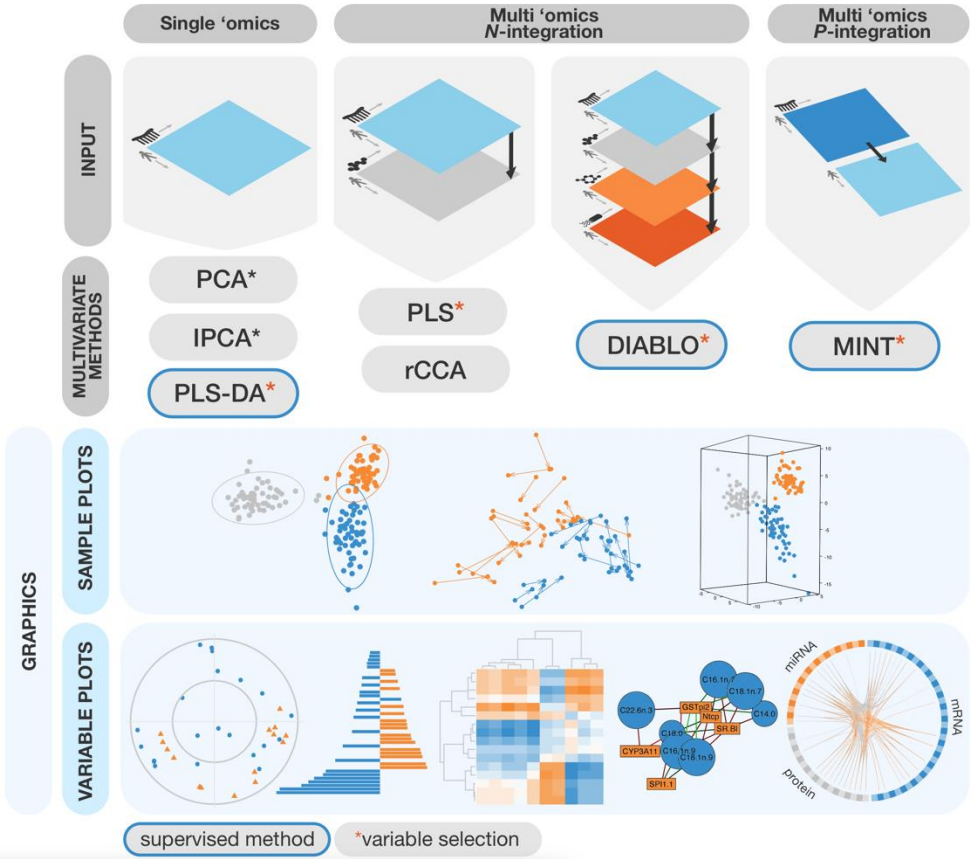
Cas d'étude ADLab

Présentation de l'outil

mixOmics

platforms all | rank 109 / 2140 | support 0 / 0 | in Bioc 4 years
build ok | updated before release | dependencies 67

DOI: 10.18129/B9.bioc.mixOmics



[HTML] **mixOmics**: An R package for 'omics feature selection and multiple data integration [HTML] plos.org

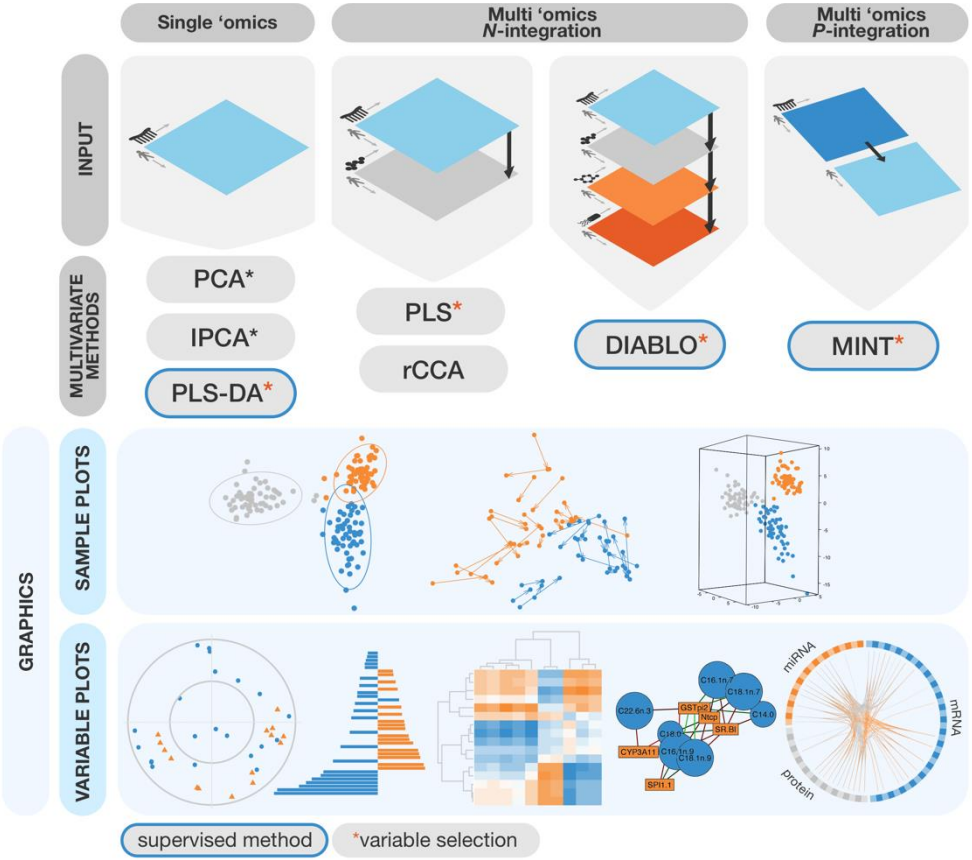
F Rohart, B Gautier, A Singh... - PLoS computational ..., 2017 - journals.plos.org

... We introduce **mixOmics**, an R package dedicated to the ... We illustrate our latest **mixOmics** integrative frameworks for ... We introduce **mixOmics** in the context of supervised analysis, where ...

☆ Enregistrer 99 Citer Cité 1570 fois Autres articles Les 20 versions »

http://mixomics.org/

Paysage des méthodes disponibles

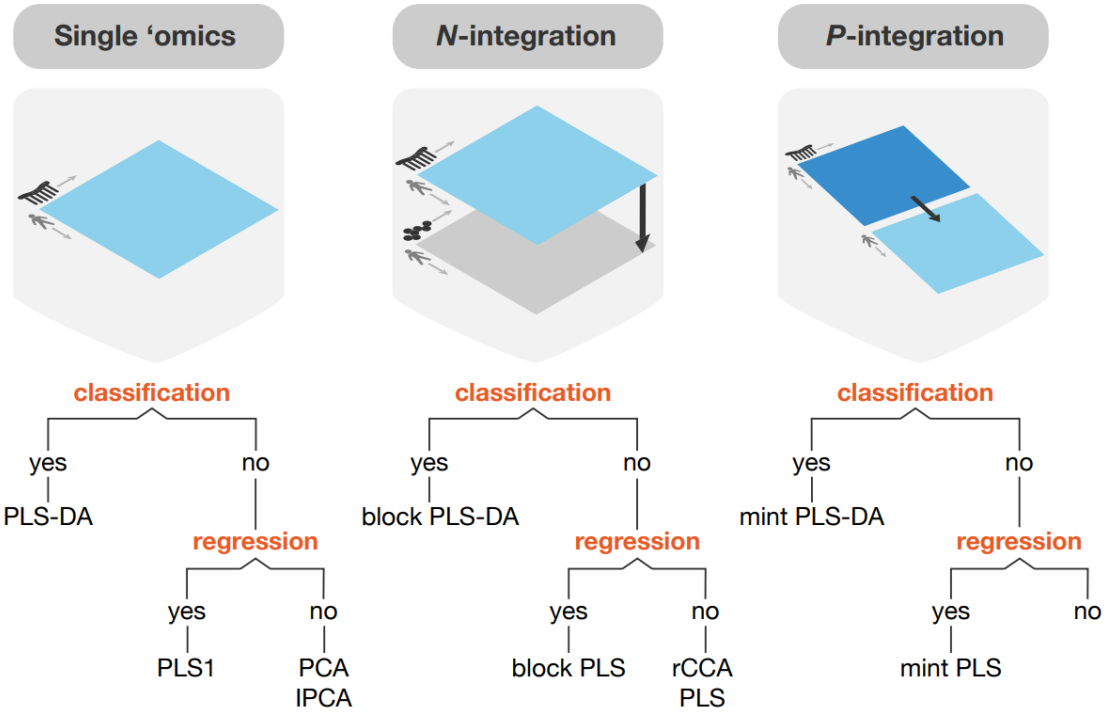


Framework:

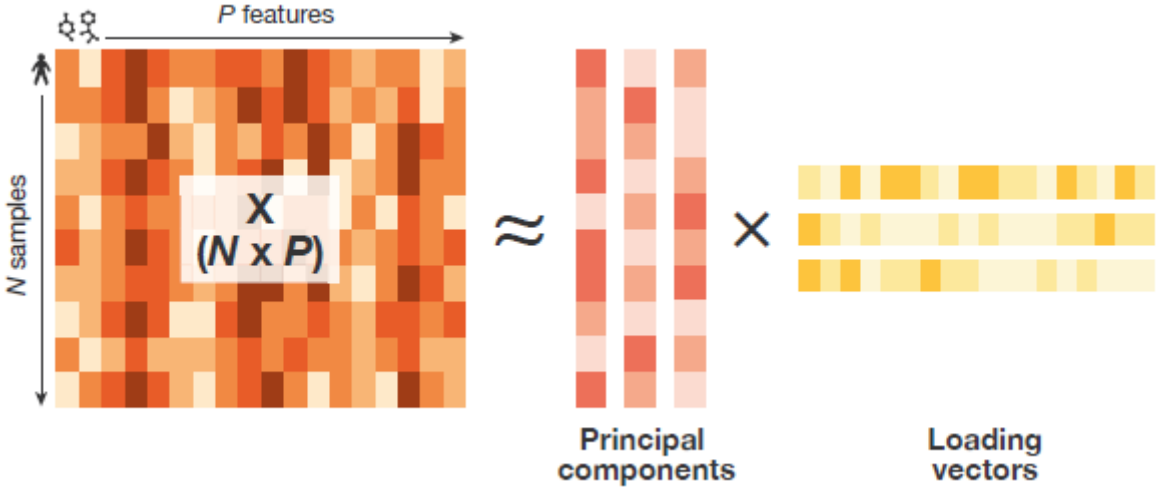
- Méthodes multivariées (1 bloc, 2 blocs, n blocs)
- Supervisée / non supervisée
- ~ Feature selection
- Tuning du modèle (nombre de composante, nombre de feature)
- Evaluation des performances
- Visualisation

Choix de la méthode

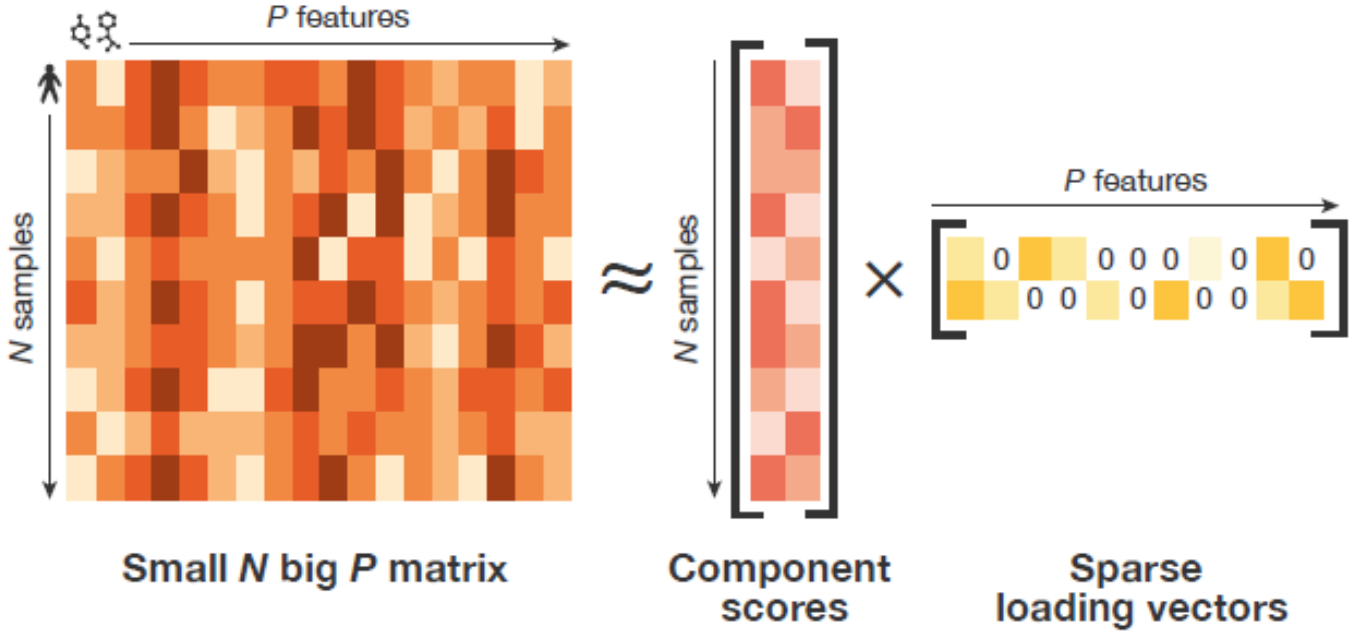
- 19 méthodes disponibles (2022)
- Type de data ?
  - Omiques “classique”:
    - RNAseq
    - Protéomique
    - Métabolomique
  - Microbiome
  - Genotypage (SNP categorique)
  - Type de question



Principal Componant Analysis



Sparse PCA

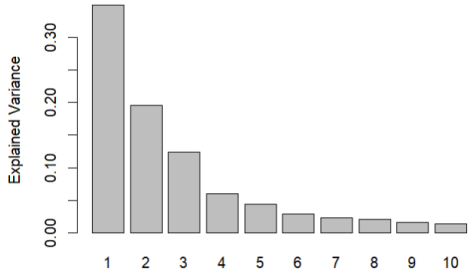


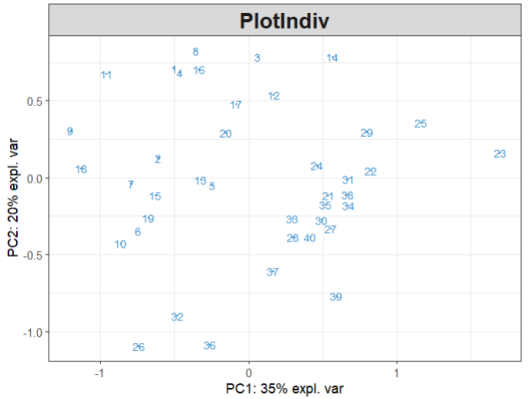
* Feature selection

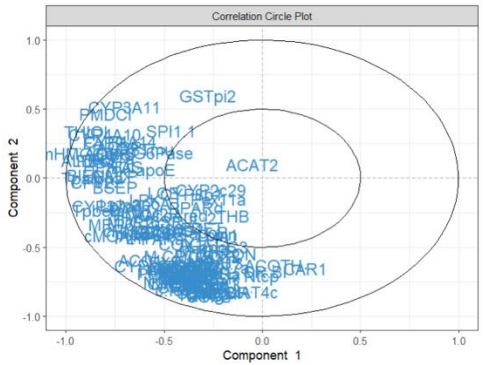
ACP

```r
library(mixOmics)
data("nutrimouse")

pca.res <- pca(X = nutrimouse$gene, ncomp = 10)
pca.res

# scree plot
plot(pca.res)

# sample plot
plotIndiv(pca.res)

# variable plot
plotVar(pca.res)
```

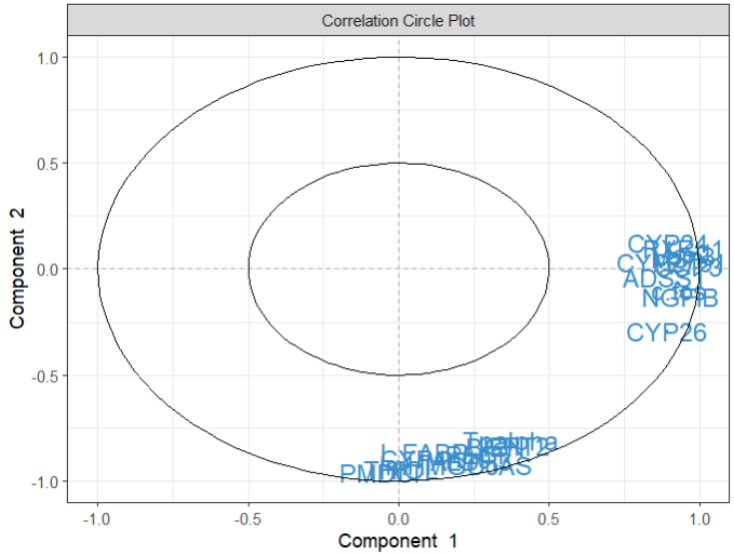
```
> pca.res
  Eigenvalues for the first 10 principal components, see object$sdev^2:
       PC1        PC2        PC3        PC4        PC5        PC6        PC7        PC8
0.45737417 0.25644100 0.16268043 0.07955690 0.05838751 0.03780991 0.03066913 0.02721979
       PC9       PC10
0.02189256 0.01855921

  Proportion of  explained variance for the first 10 principal components, see object$prop_expl_var:
       PC1             PC2             PC3             PC4             PC5             PC6
0.34974173      0.19609354      0.12439735      0.06083503      0.04464736      0.02891222
       PC7             PC8             PC9            PC10
0.02345186      0.02081424      0.01674065      0.01419173

  Cumulative proportion of  explained variance for the first 10 principal components, see object$cum.var:
      PC1            PC2            PC3            PC4            PC5            PC6
0.3497417      0.5458353      0.6702326      0.7310676      0.7757150      0.8046272
      PC7            PC8            PC9           PC10
0.8280791      0.8488933      0.8656340      0.8798257

  Other available components:
 --------------------
  loading vectors: see object$rotation
  Other functions:
 --------------------
  plotIndiv, plot, plotVar, selectVar, biplot
```







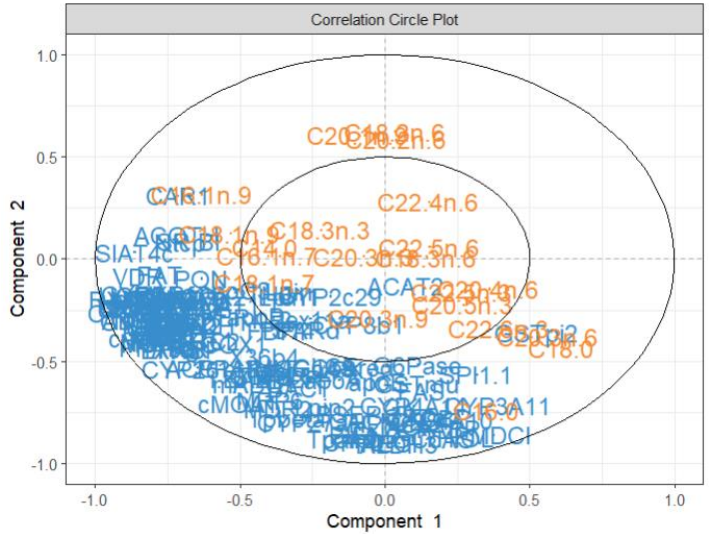
```r
# spca
spca.res <- spca(X = nutrimouse$gene, ncomp = 2,  keepX = c(10, 10))
plotVar(spca.res)
```

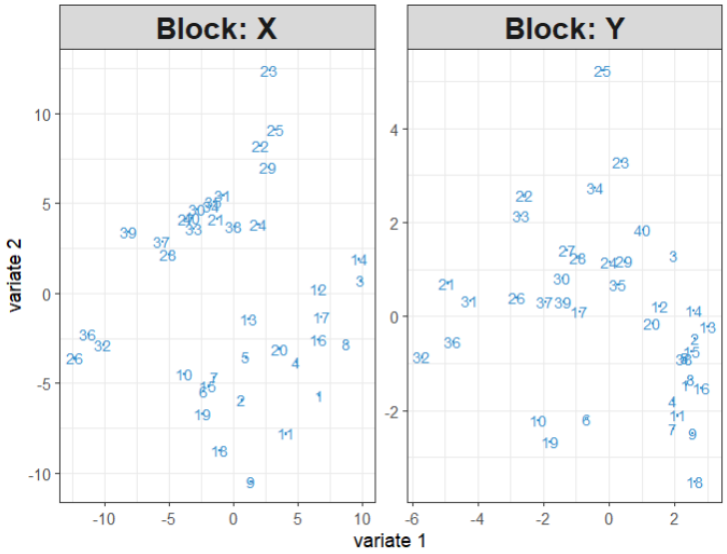


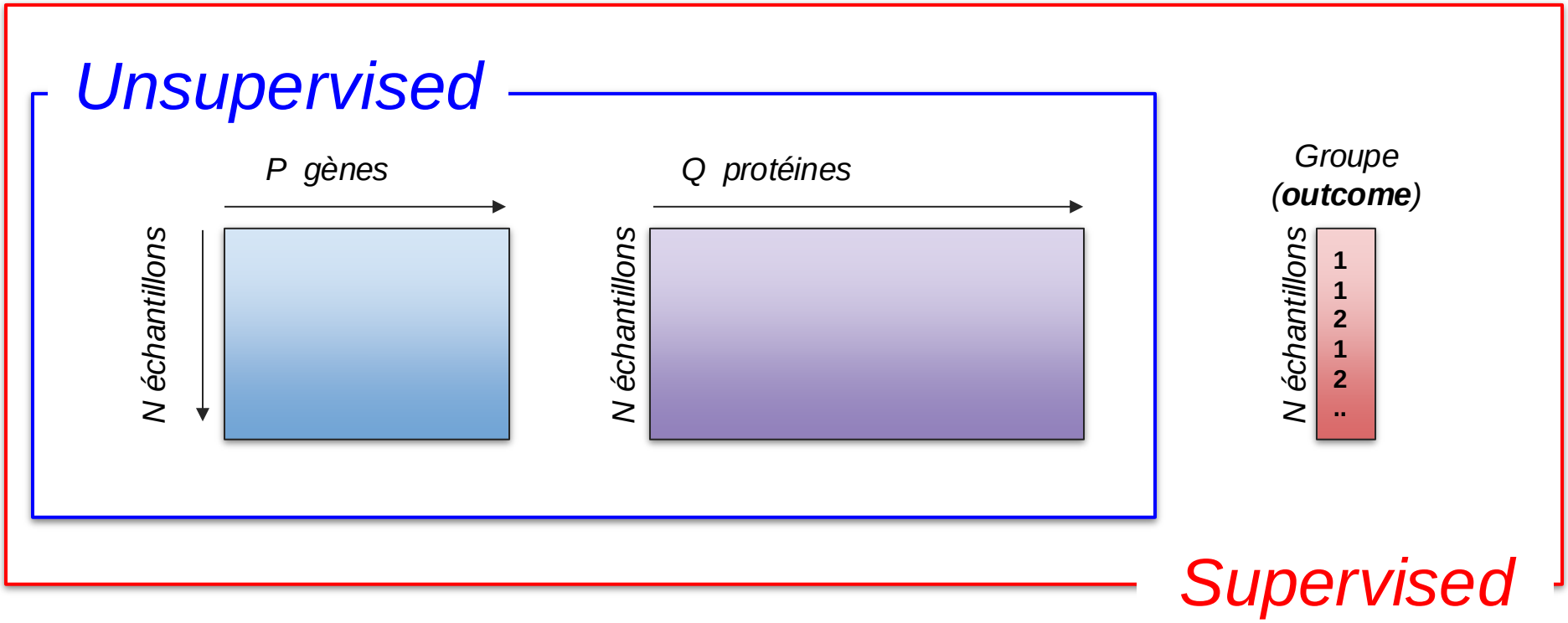
PLS

```
# pls
pls.res <- pls(X = nutrimouse$gene, Y = nutrimouse$lipid, ncomp = 2)

# sample plot
plotIndiv(pls.res)

# variable plot
plotVar(pls.res)


# spls
spls.res <- spls(X = nutrimouse$gene, Y = nutrimouse$lipid, ncomp = 2,
                 keepX = c(5,3),
                 keepY = c(3,2))
plotVar(spls.res)
```

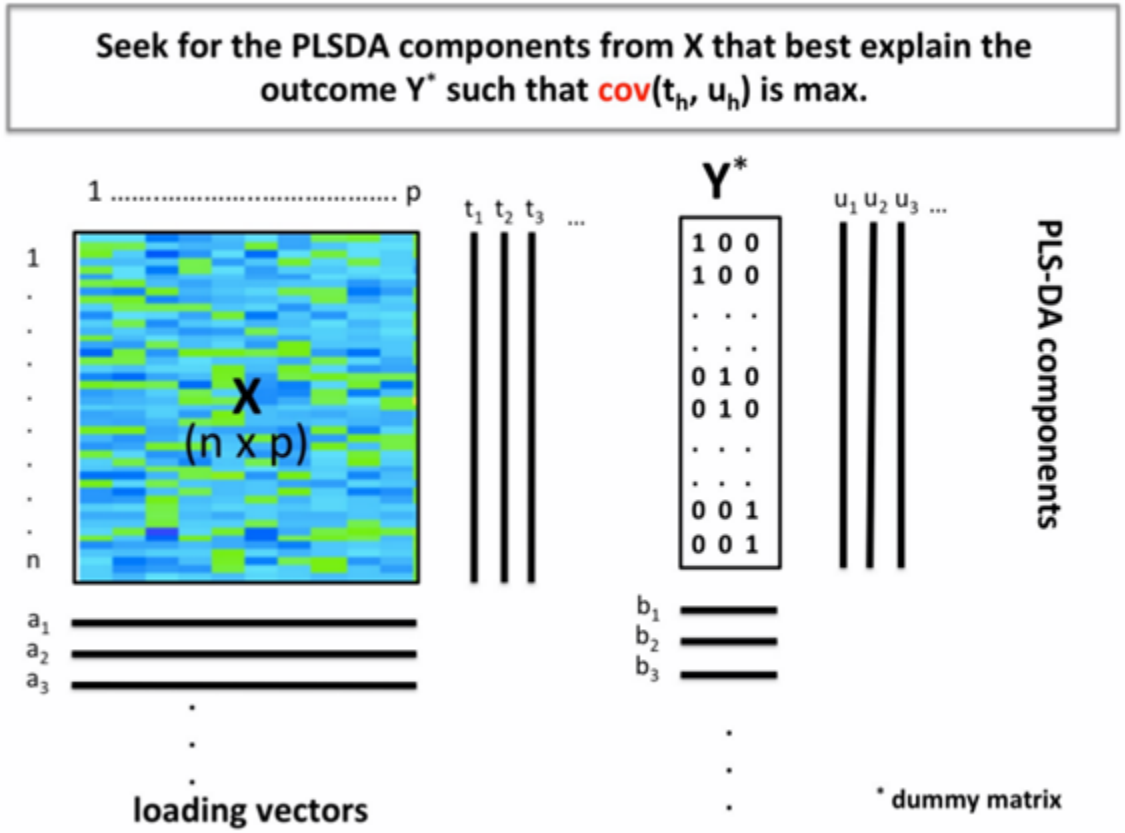
```
> selectVar(spls.res, comp = 1)
$X
$X$name
[1] "PMDCI"   "SPI1.1"  "SR.BI"   "CYP3A11" "Ntcp"

$X$value
         value.var
PMDCI    0.6625305
SPI1.1   0.5321982
SR.BI   -0.3955914
CYP3A11  0.3070180
Ntcp    -0.1645169


$Y
$Y$name
[1] "C18.0"    "C16.1n.9" "C16.0"

$Y$value
          value.var
C18.0     0.7726563
C16.1n.9 -0.6035211
C16.0     0.1968871


$comp
[1] 1
```






Unsupervised
P gènes
N échantillons
Q protéines
N échantillons
Groupe
(outcome)
N échantillons
1
1
2
1
2
..
Supervised

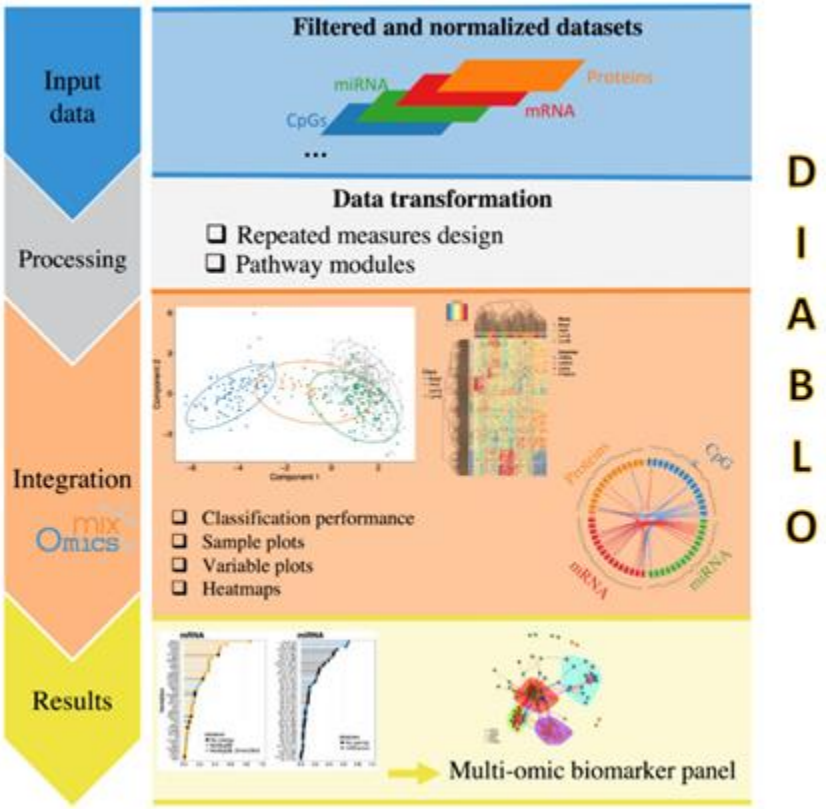
PLS - Discriminant Analysis (PLS-DA)

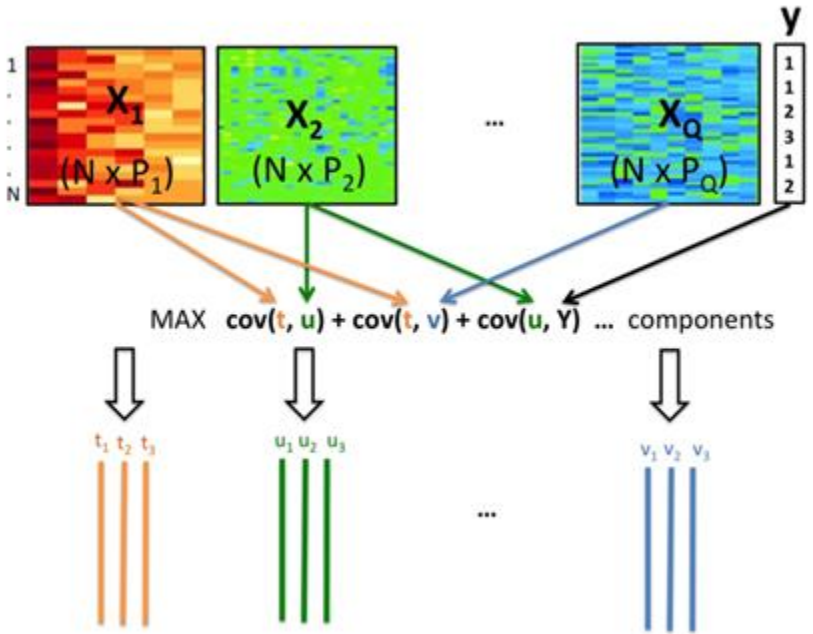


Adapted from from K-A Lê Cao

DIABLO

Data **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent Variables approaches for **O**mics studies



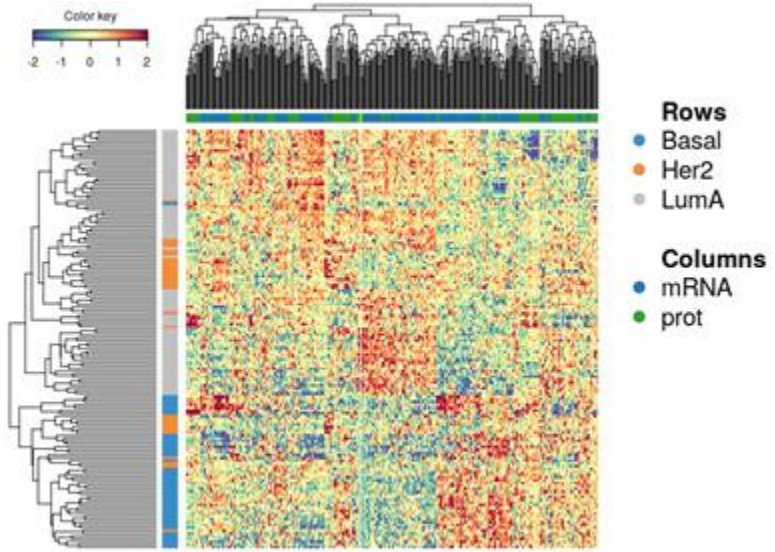


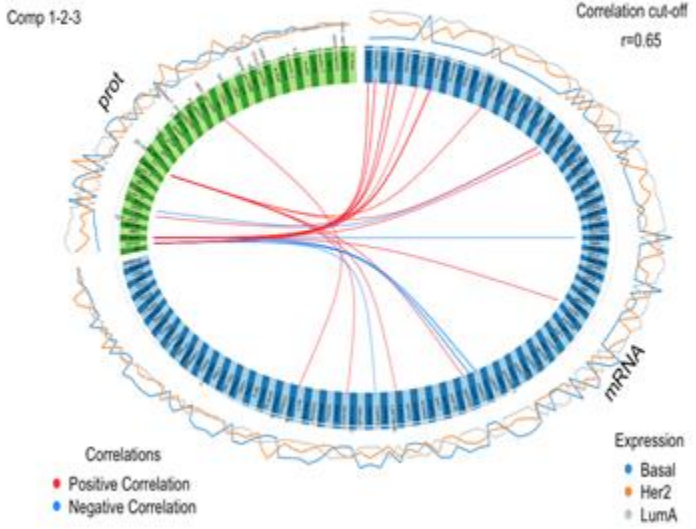
From K-A Lê Cao

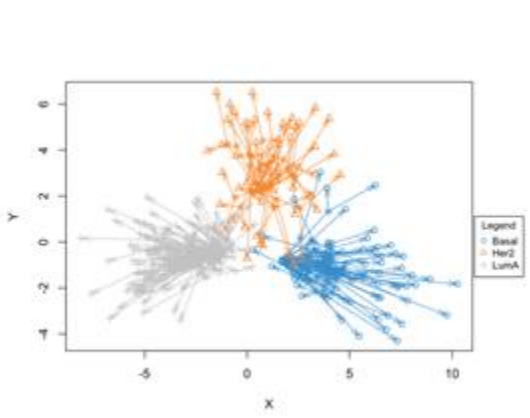
Adapted from from K-A Lê Cao

DIABLO


mRNA
0.93
prot
Basal
Her2
LumA


Color key
Rows
Basal
Her2
LumA
Columns
mRNA
prot

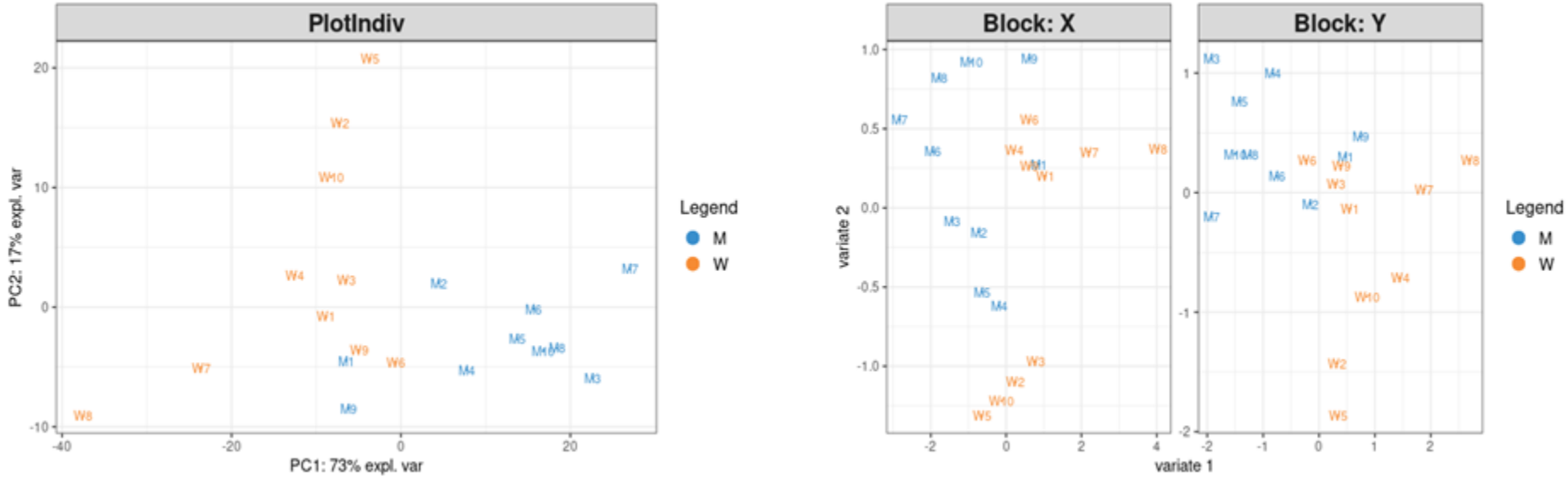

Comp 1-2-3
Correlation cut-off
r=0.65
prot
mRNA
Correlations
Positive Correlation
Negative Correlation
Expression
Basal
Her2
LumA


Legend
Basal
Her2
LumA
X
Y


SLC43A3
ZEB1
Caveolin.1
GATA3
KDM4B
ZNF552
ER.alpha

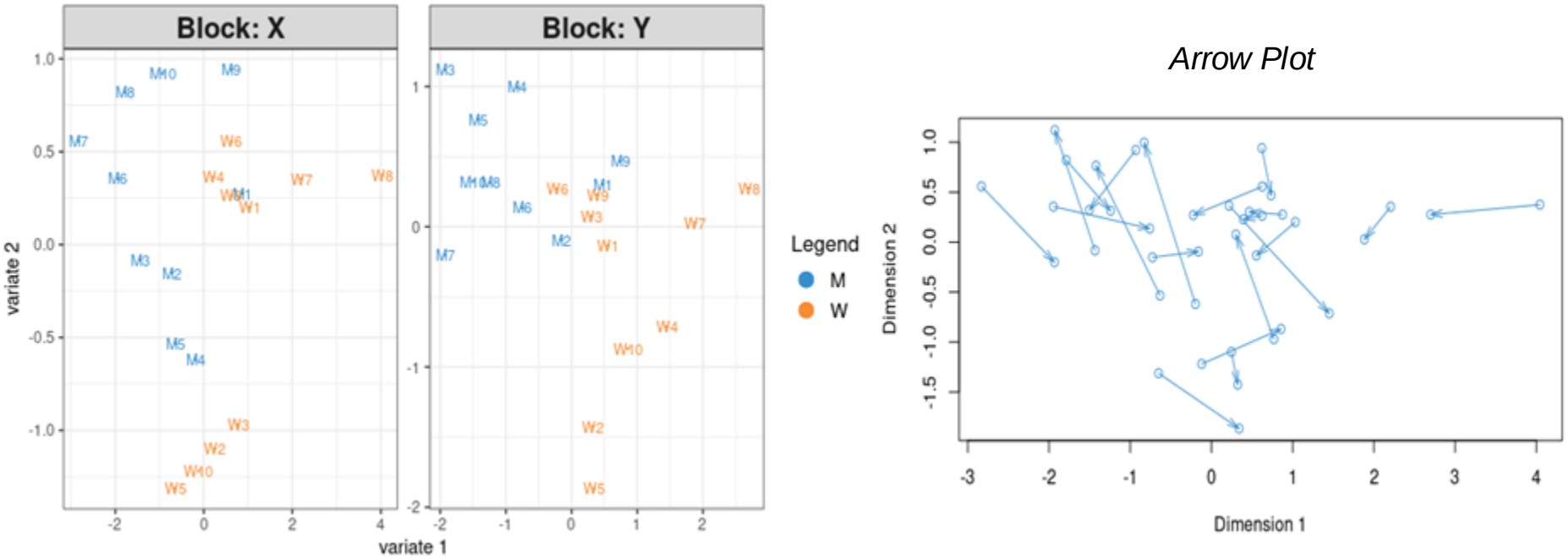
mixOmics Graphical Output

1. Samples plots



- Projection des échantillons (samples) dans l'espace produit par les nouvelles variables instrumentales (components, latent structures, …)
- similarité entre les échantillons (clustering)
- autres méthodes pour représenter les échantillons dans mixOmics: *plotArrow* (paired: X -> Y)

```
> plotIndiv(...)
```

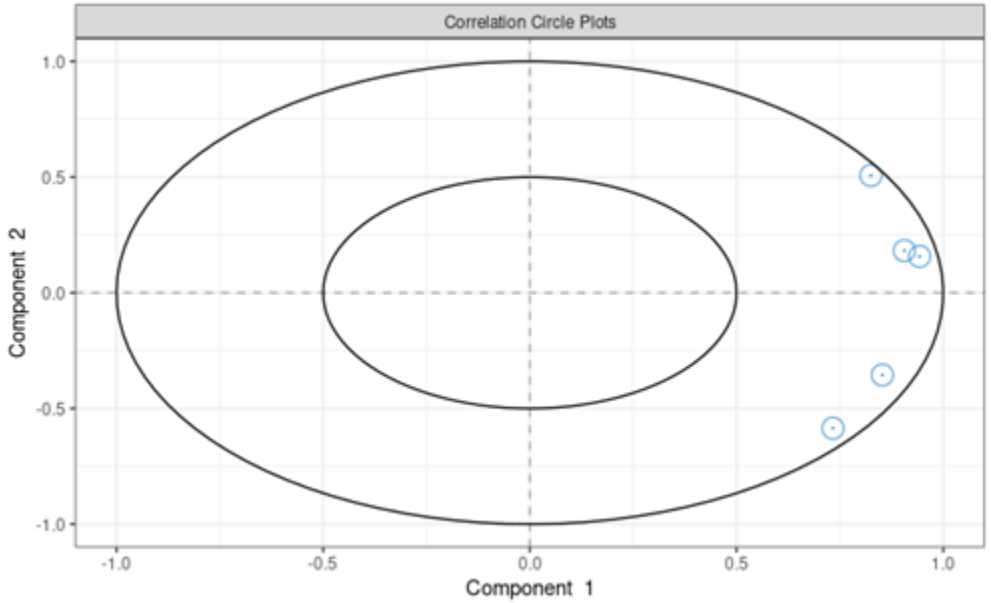


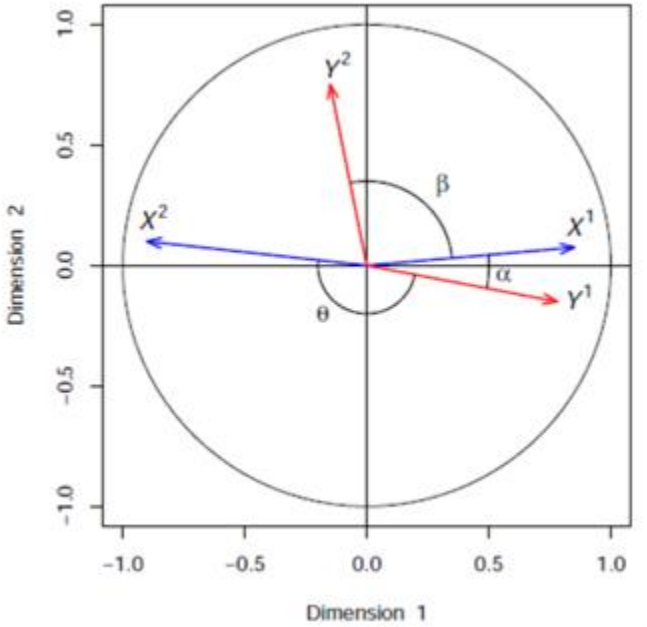
- 2 *blocs*
- 1 flèche = 1 échantillon.
- Origine = l'emplacement de l'échantillon en X, la pointe = l'emplacement dans Y
- Courtes flèches = X et Y sont en accord pour cet échantillon et inversement.

```
> plotArrow(...)
```

2. Variables plots

Circle Correlation plot

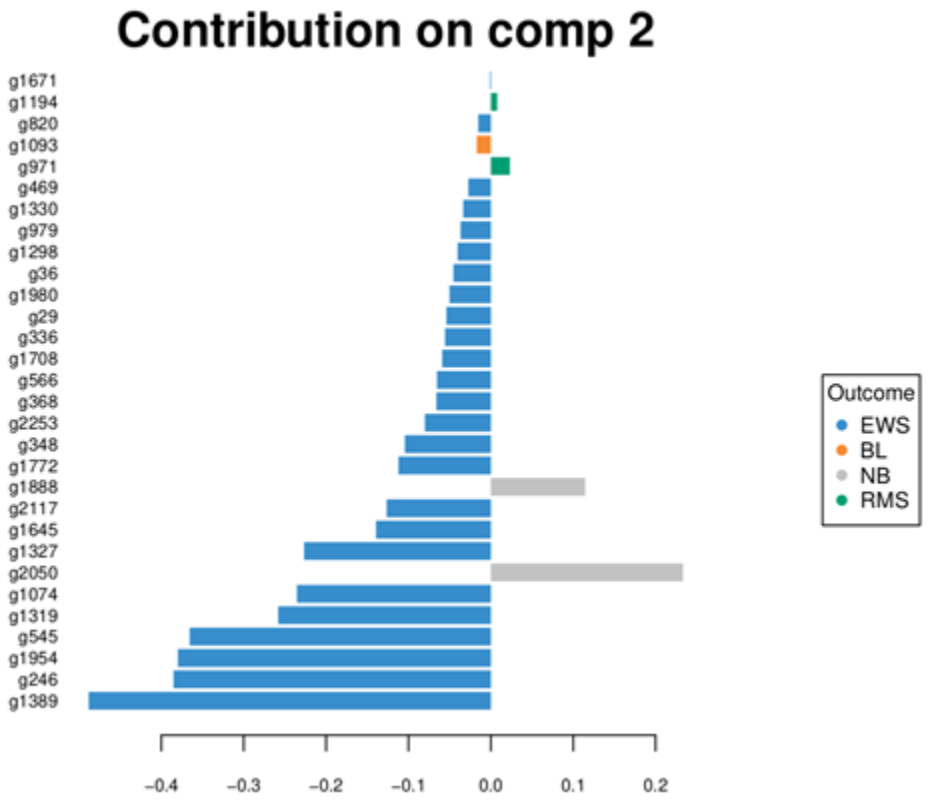




- Visualise les relations entre les variables et les composantes (*structure latentes*)
- Importance de la longueur du vecteur et du cosinus de l'angle avec l'axe
- Montre l'importance des variables / contribution par rapport aux composantes

```
> plotVar(...)
```

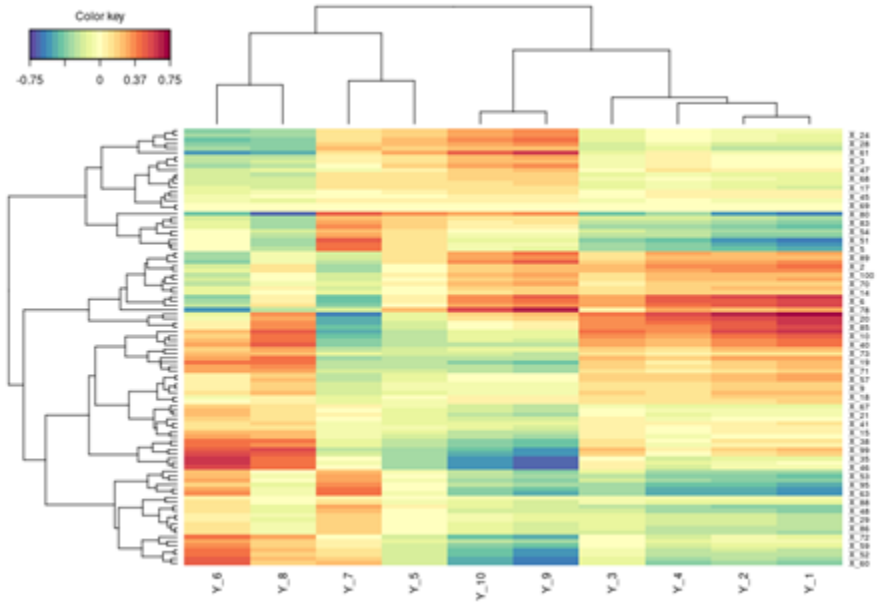
Loading plot

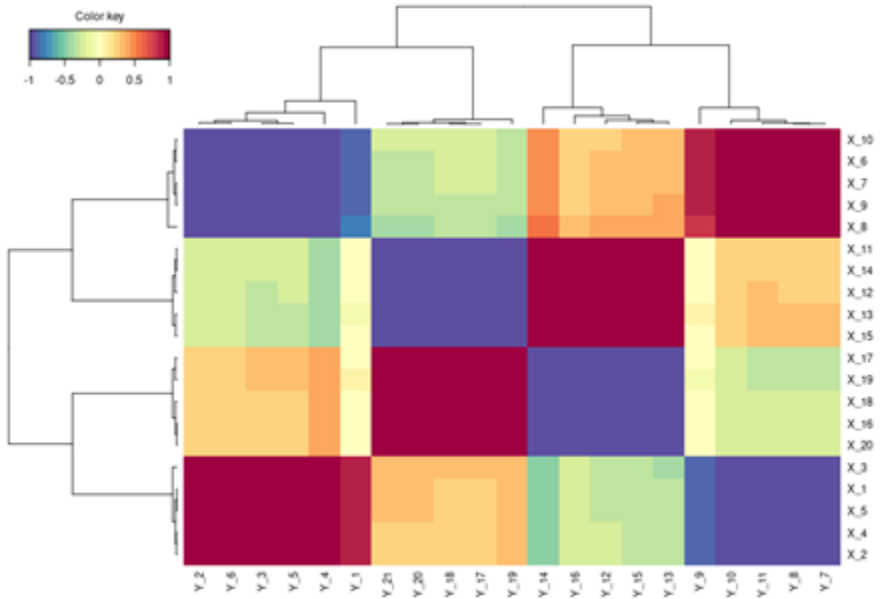


- Très similaire au *CCplot*
- Importance des variables chaque composantes (taille des barres)
- Ordonnée par ordre d'importance
- en mode *supervisé*, la couleur correspond au groupe le plus proche

```
> plotLoadings(...)
```

Clustered Image Map

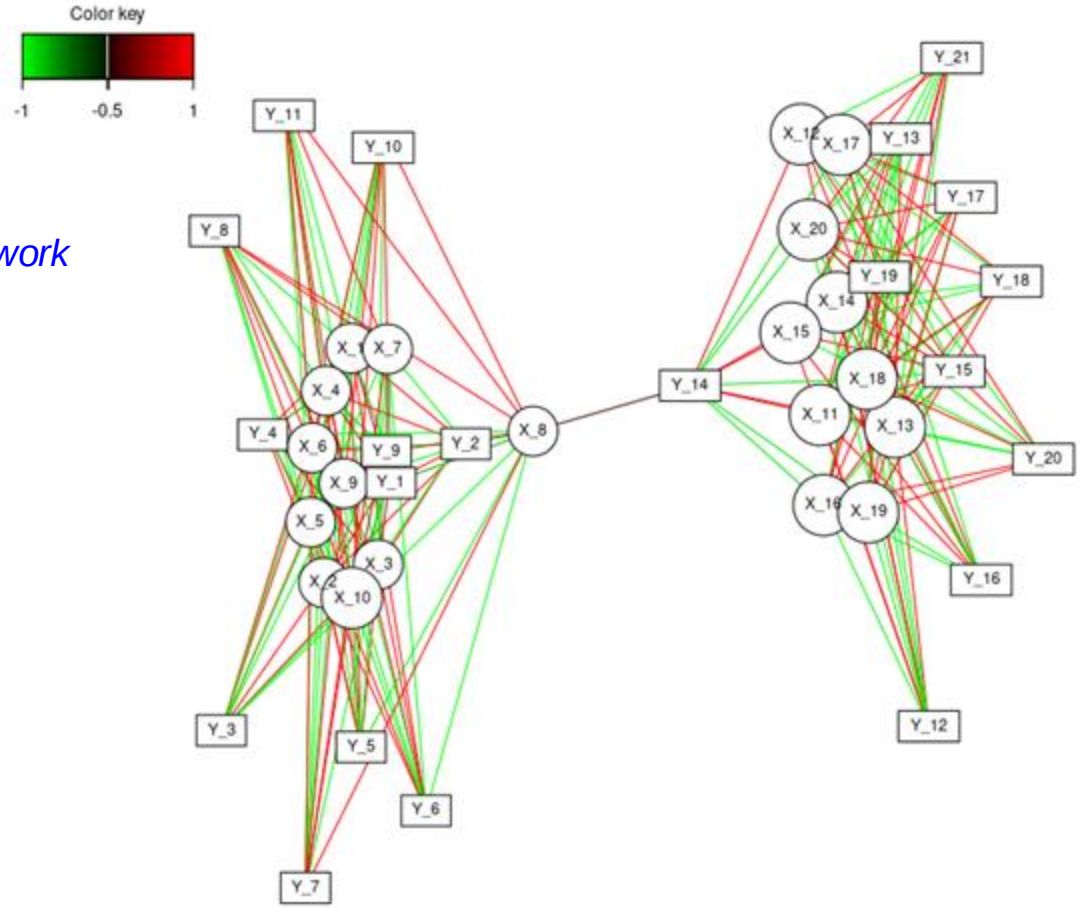




- Correlation = couleur
- Montre la corrélation entre des ensembles de variables de type différents (2 blocs)
- Clusters
- Dendrogramme

```
> cim(...)
```

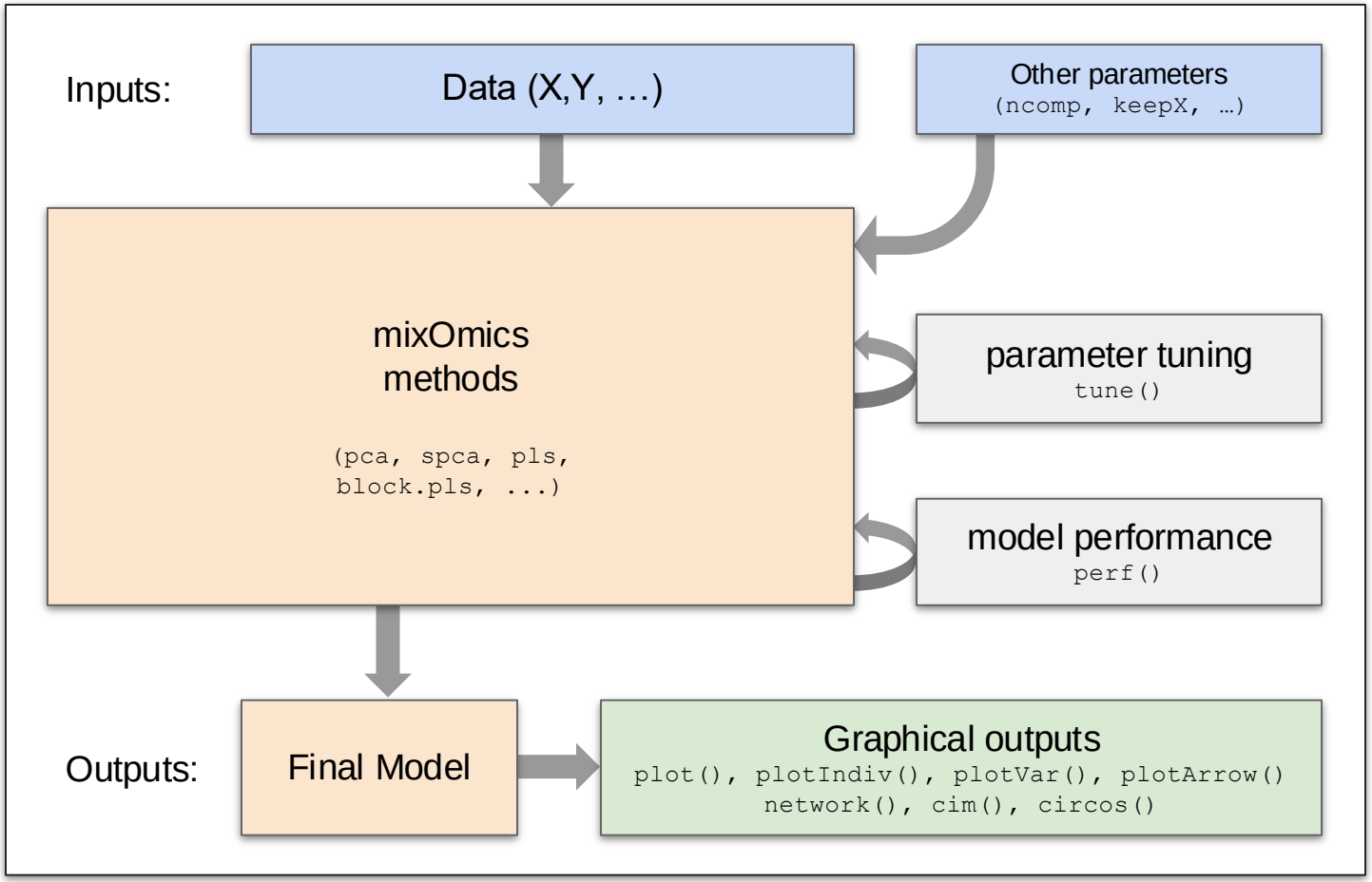
Relevance Network



- Montre la corrélation entre 2 variables (2 à 2) par un lien (tous les blocs).
- Corrélation = couleur du lien *(positif/négatif, fort/faible)*
- choix du *cutoff* *(affichage des liens)*
- Recherche de clique / sous-réseau

```
> network(...)
```

Cycle d'une analyse

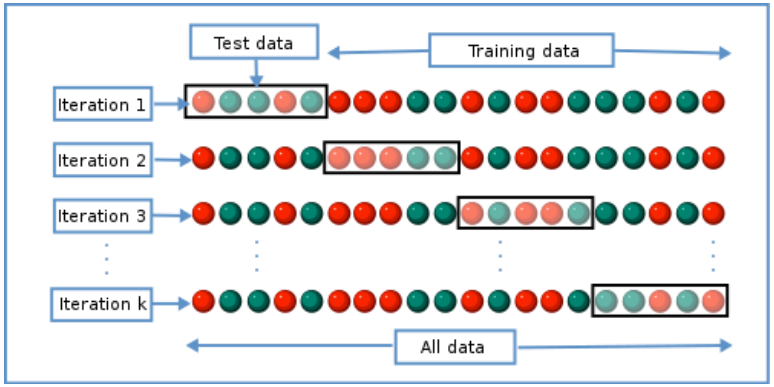


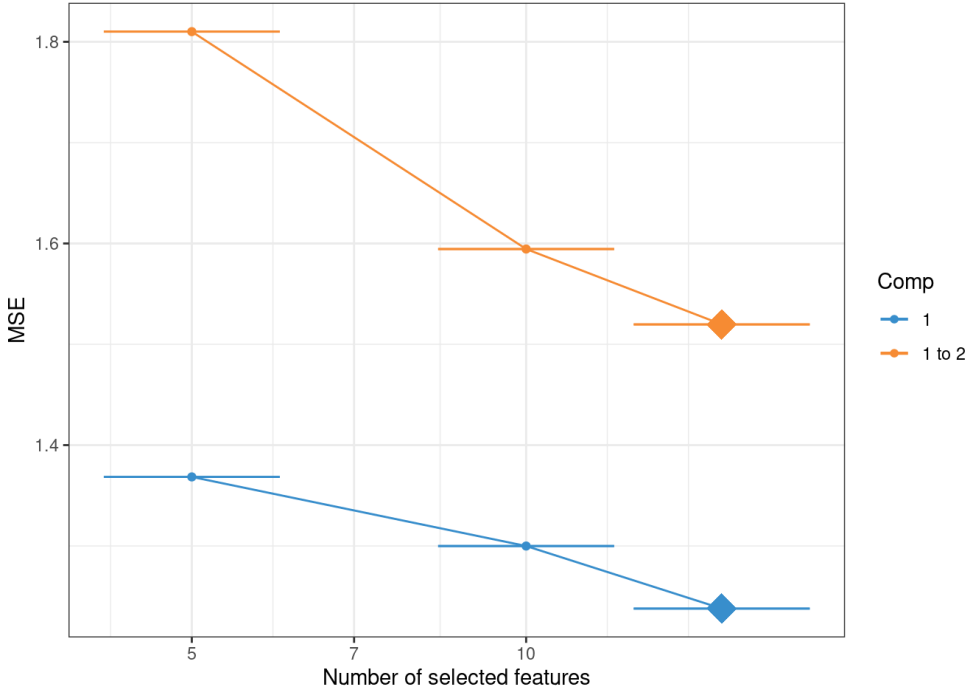
Tuning d’un modèle

- Choix du nombre de composantes (ou variable latente)
- Choix du nombre de feature à retenir (sparse) sur X (keepX), sur Y (keepY)

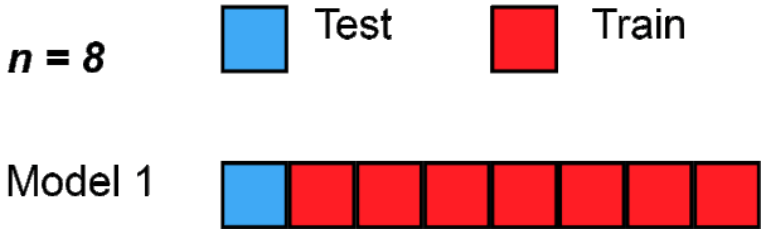
Evaluer les performances d'un modèle

M-fold cross validation





Leave One Out



Mise en contexte

Différents concepts d'integration

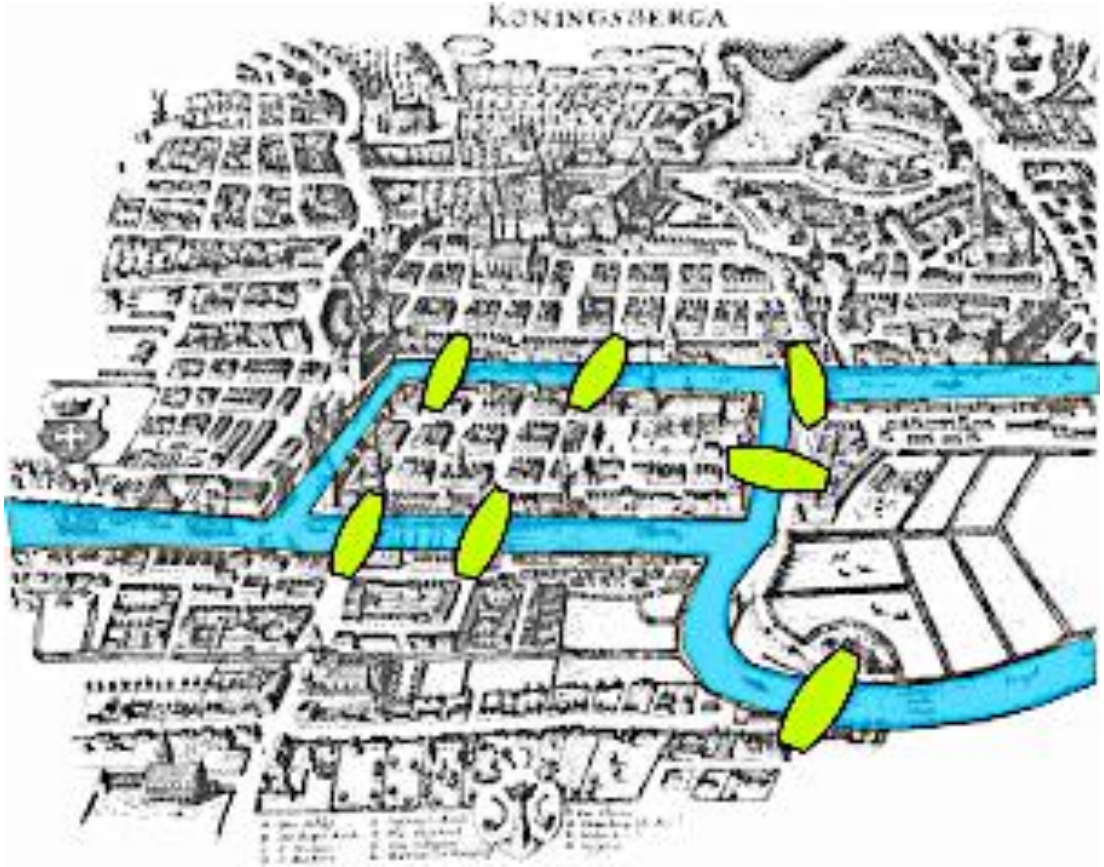
Méthodes multivariées

mixOmics

Réseaux en biologie

Cas d'étude ADLab

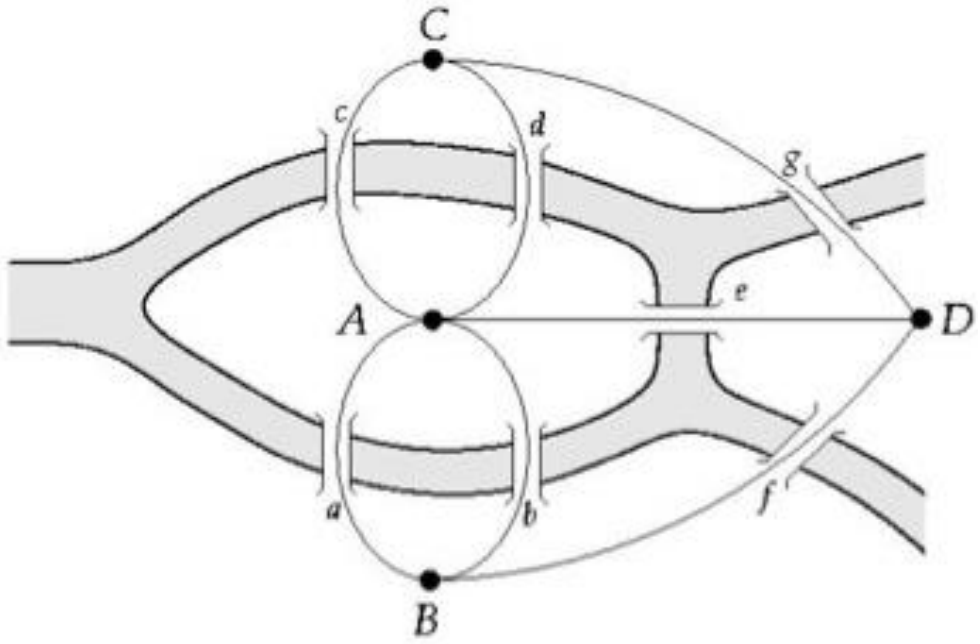
Les 7 ponts de Koningsberg



Peut-on marcher sur les sept ponts et
ne jamais traverser deux fois le même pont ?

http://networksciencebook.com/chapter/2#bridges

Les 7 ponts de Koningsberg



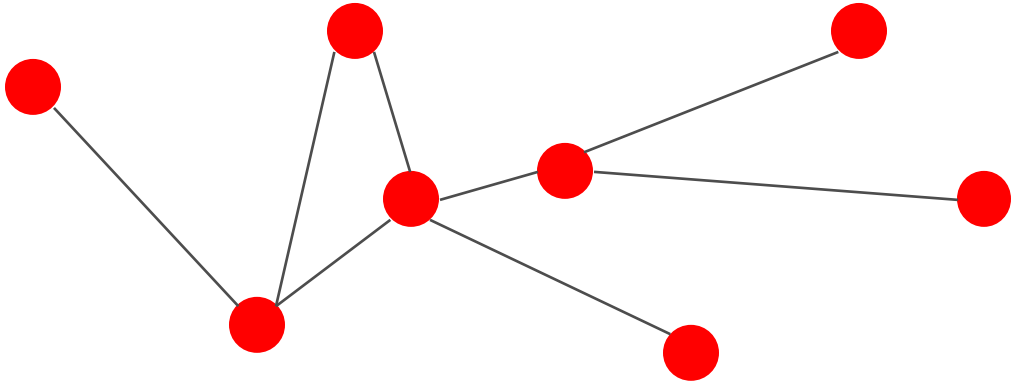
Un nœud avec nombre de connexion impair peut être soit un l'entré du chemin, soit la sortie

Théorème d'Euler (1735):

- Si un graphe a plus de deux noeuds de degré impair, il n'y a pas de chemin.
- Si un graphe est connecté et n'a pas de noeuds de degré impair, il a au moins un chemin.

Adapted from Barabási, Network Science

Eléments constitutifs



- **components**: nodes (noeud), vertices (sommet) N
- **interactions**: links (liens), edges (arêtes) L
- **system**: network, graph (N,L)

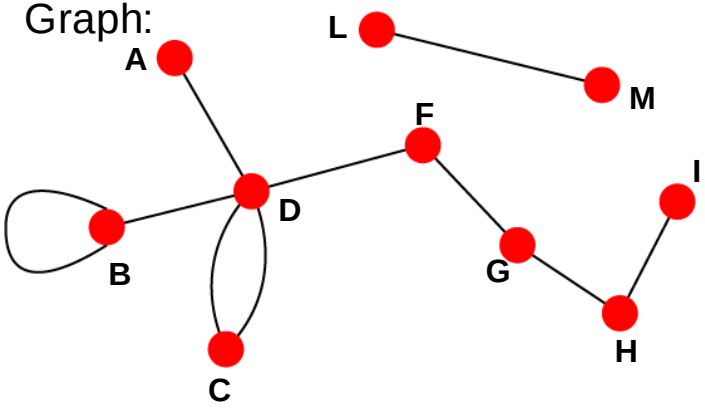
Adapted from Barabási, Network Science

Eléments constitutifs

Undirected

Links: undirected (*symmetrical*)

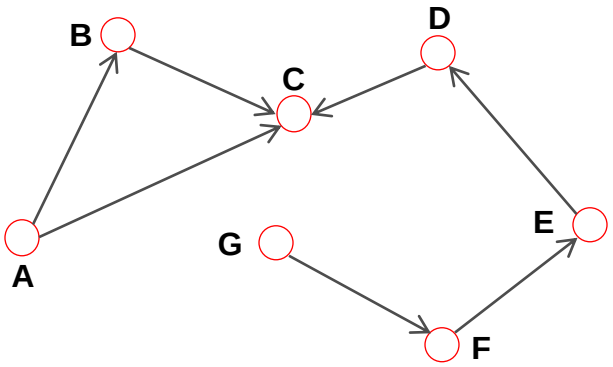
Graph:



Directed

Links: directed (*arcs*).

Digraph = directed graph:

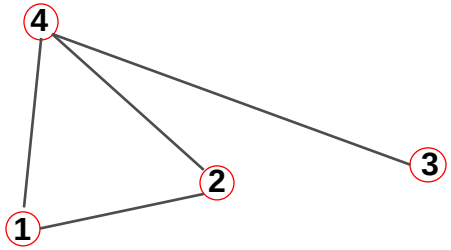


An undirected link is the superposition of two opposite directed links.

Adapted from Barabási, Network Science

Matrice d'adjacence

Undirected

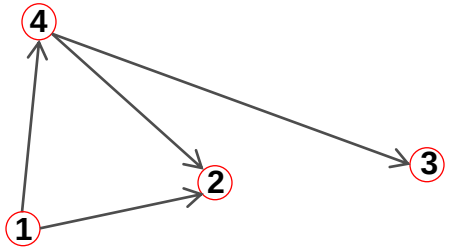


$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

symétrique

A_{ij}=1 si il existe un lien entre i et j
A_{ij}=0 sinon

Directed



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

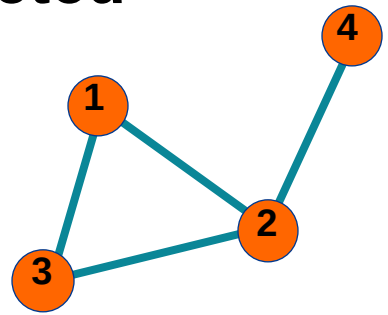
non-symétrique

A_{ij}=1 si il existe un lien **de** i **vers** j
A_{ij}=0 si auncun lien ne pointe de i vers j

Adapted from Barabási, Network Science

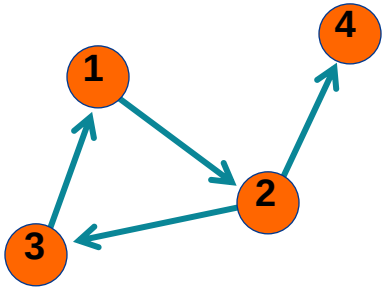
Matrice d'adjacence

Undirected



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Directed

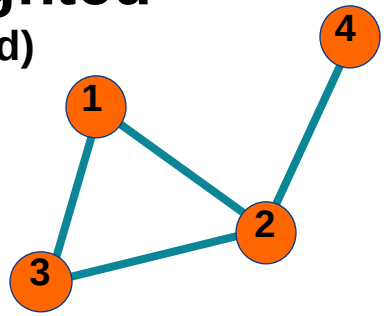


$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Adapted from Barabási, Network Science

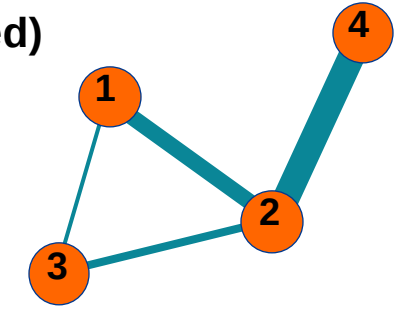
Matrice d'adjacence

Unweighted
(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Weighted
(undirected)

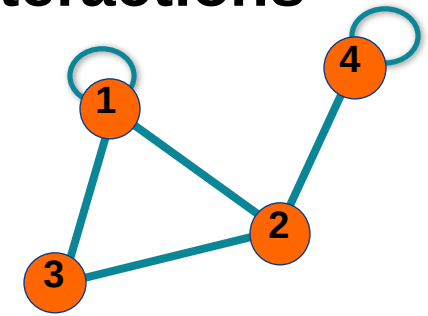


$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

Adapted from Barabási, Network Science

Matrice d'adjacence

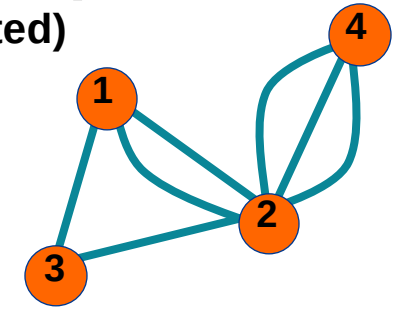
Self-interactions



$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

Multigraph

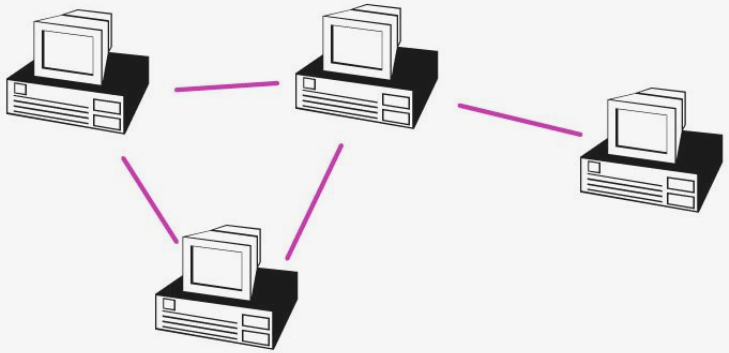
(undirected)

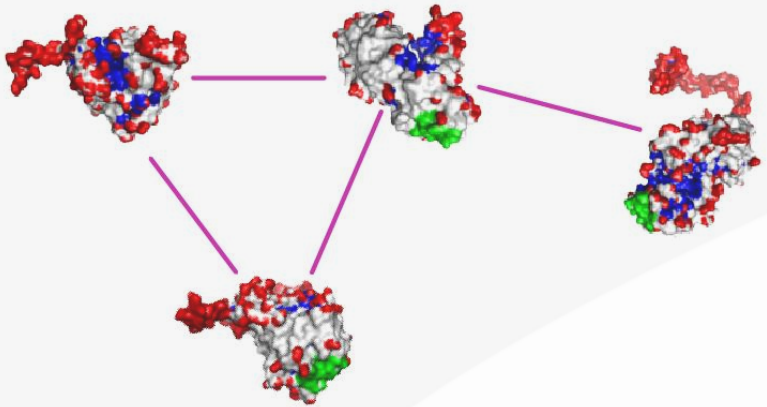


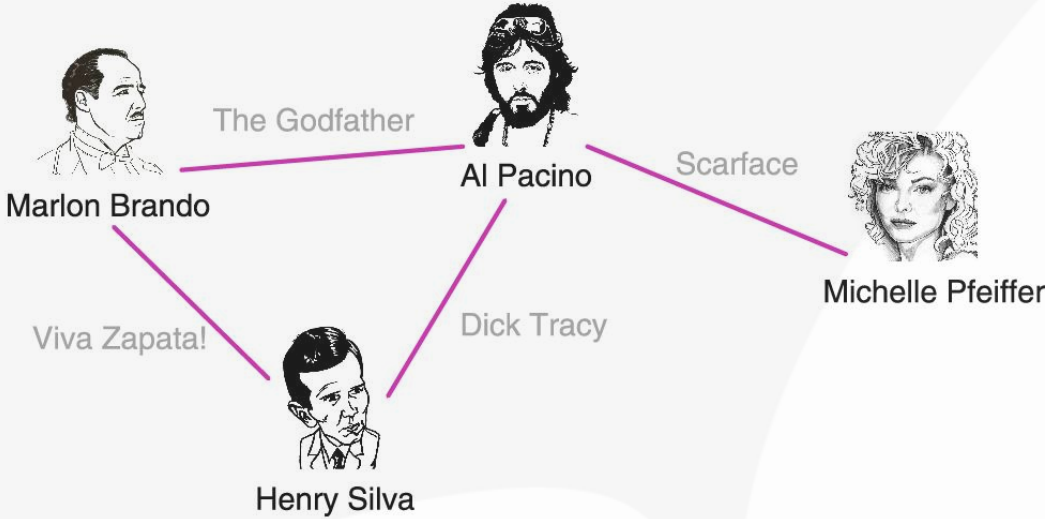
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

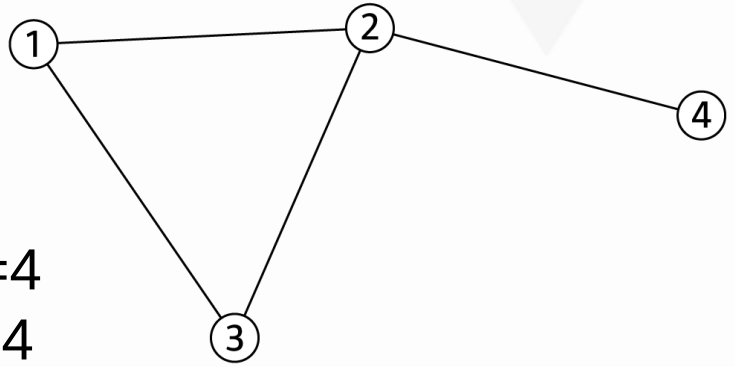
Adapted from Barabási, Network Science

Un même graphe … plusieurs réseaux







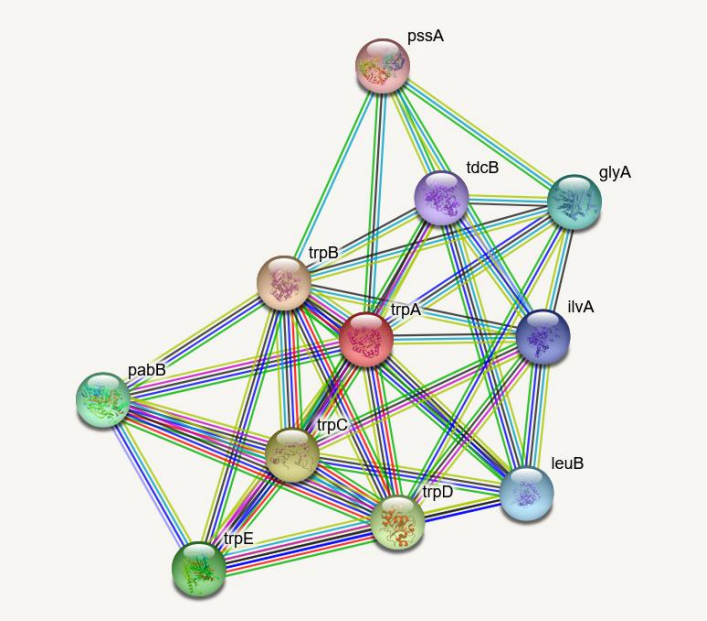


N=4
L=4

Adapted from Barabási, Network Science

Choisir la bonne répresentation

- Parfois representation unique et non ambiguë
- Parfois plusieurs graphes sont possible
- Trouver la meilleure représentation pour solutionner le problème.



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes: query proteins and first shell of interactors

white nodes: second shell of interactors

Node Content

empty nodes: proteins of unknown 3D structure

filled nodes: some 3D structure is known or predicted

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.

Known Interactions

from curated databases

experimentally determined

Predicted Interactions

gene neighborhood

gene fusions

gene co-occurrence

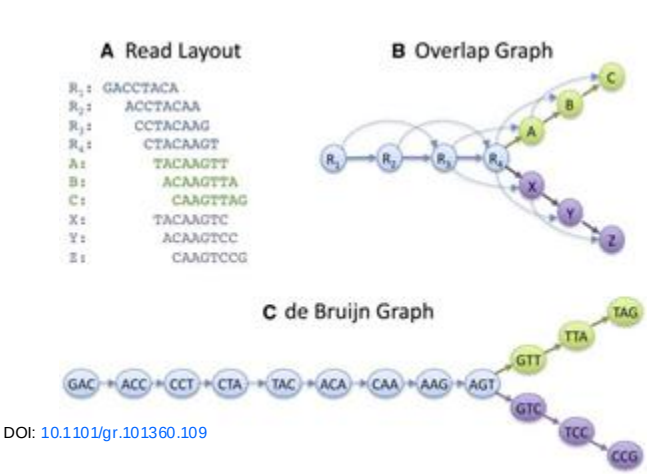
Others

textmining

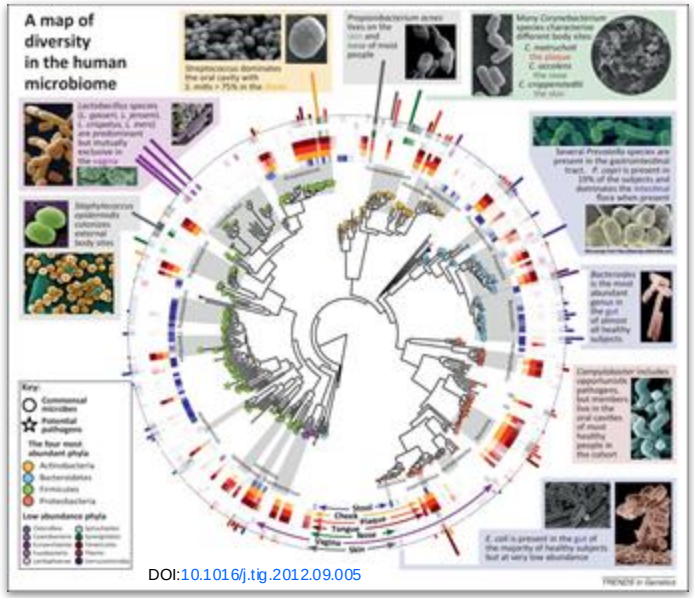
co-expression

protein homology

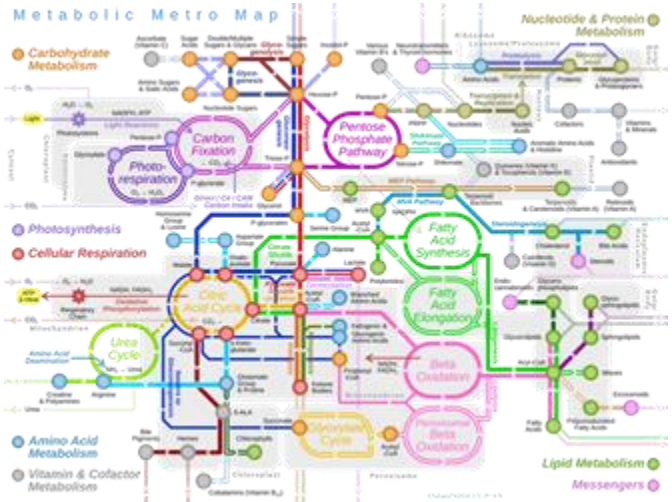
Les Réseaux en Biologie



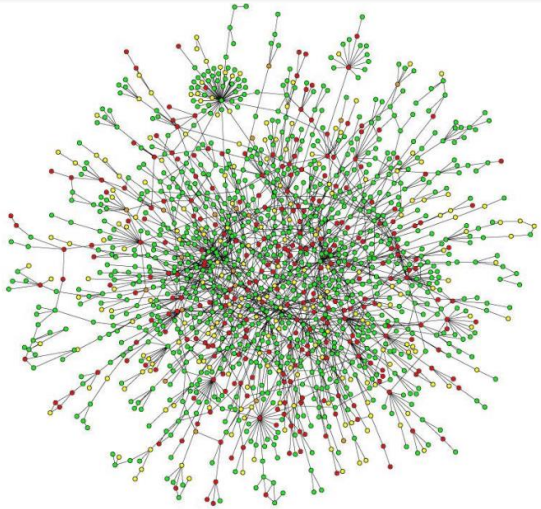
DOI: 10.1101/gr.101360.109



DOI:10.1016/j.tig.2012.09.005



https://upload.wikimedia.org/wikipedia/commons/thumb/6/6e/Metabolic_Metro_Map.svg/1280px-Metabolic_Metro_Map.svg.png



PPI of Saccharomyces cerevisiae

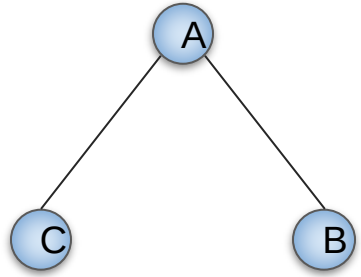
Gene Co-Expression / Regulation Network

	Gene-A	Gene-B	Gene-C
Ech1	10	5	6
ECh2	42	49	2
...			

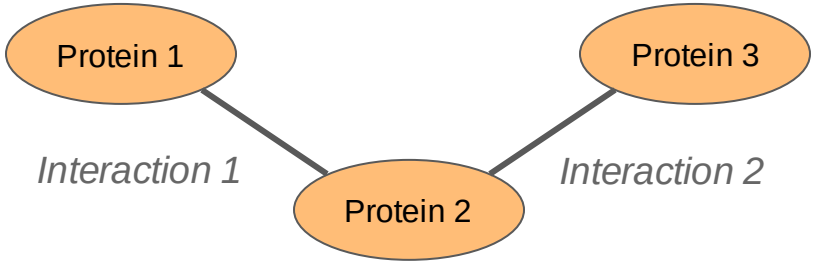
Matrice d'adjacence

	Gene-A	Gene-B	Gene-C
Gene-A	1	1	1
Gene-B	1	1	0
Gene-C	1	0	1

1: connecté 0: pas connecté

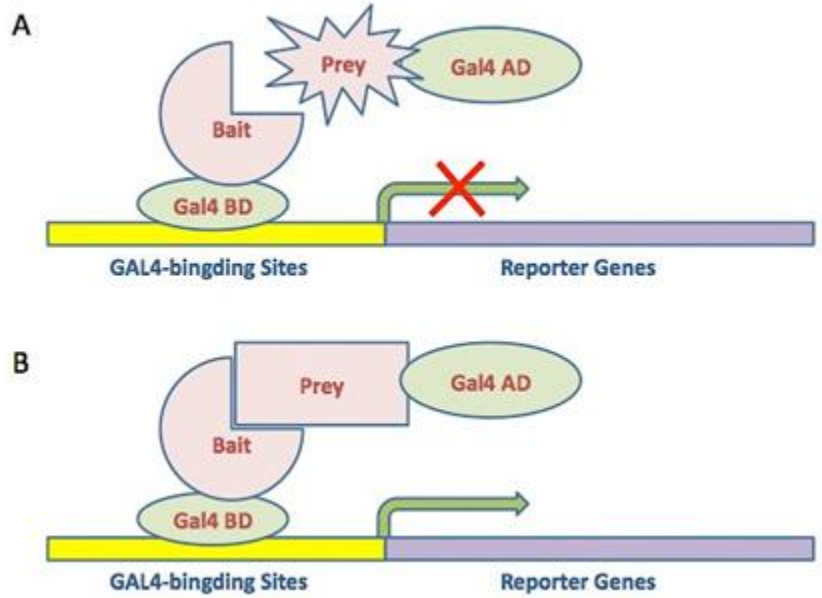


Réseaux d'interaction Protéine-Protéine (PPi)



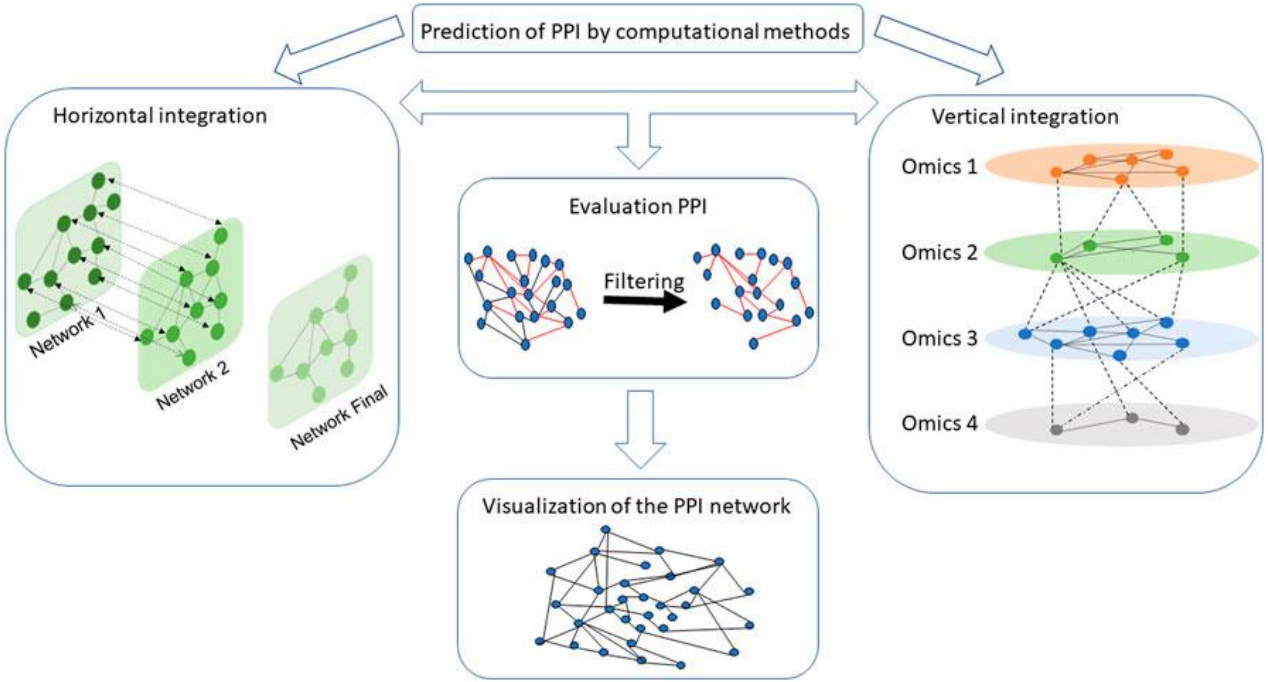
Interaction = (activation, binding to, phosphorylation…)

Two-hybrid screening



http://www.myfolio.com/art/fez57ey2az

Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature, 411(6833), 41.



[CITATION] **Overview** of **methods** for characterization and visualization of a protein-protein interaction network in a multi-omics integration context

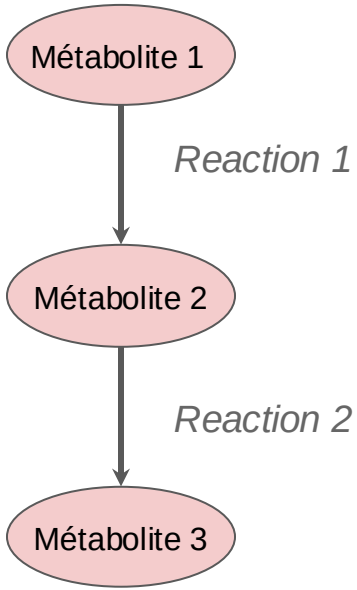
V **Robin**, A Bodein, MP Scott-Boyer, M Leclercq… - Frontiers in Molecular … - Frontiers

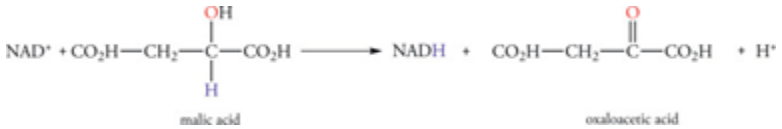
☆ Enregistrer 99 Citer

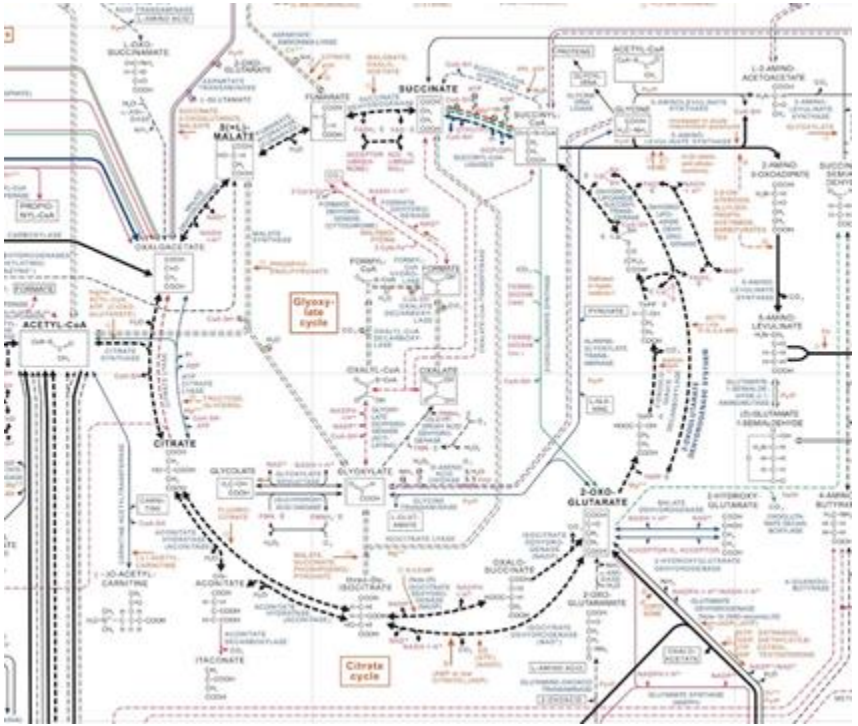
- Genomic context
- **Domain fusion**
- **Conserved gene neighborhood**
- **Phylogenetic profiles**
- **Coevolution**

- Machine learning
- Text mining (natural language processes)

Metabolic Pathways



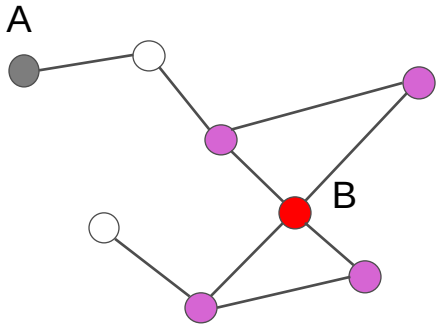




Interaction = (synthesis, oxydation, hydratation, decarboxylation, ...)

Propriétés topologiques: Degré

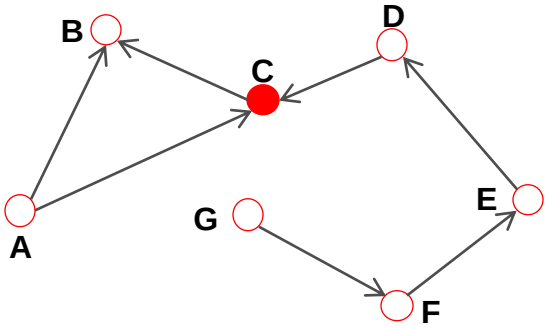
Undirected



Degré d'un noeud = nombre de connection avec ses voisins.

$$k_A = 1 \qquad k_B = 4$$

Directed



On fait la distinction entre degré entrant, degré sortant.

Degré total = degré entrant + degré sortant

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

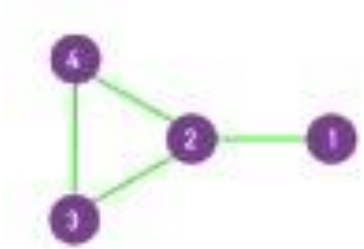
Adapted from Barabási, Network Science

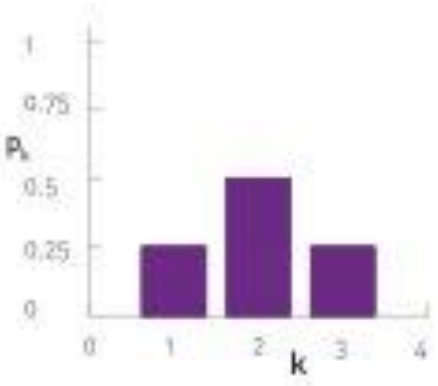
Distribution des degrés

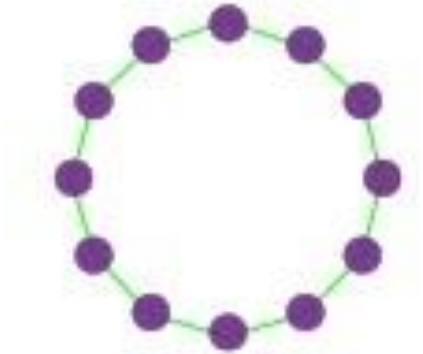
On note P(k) la probabilité pour un nœud au hasard d'avoir un degré k

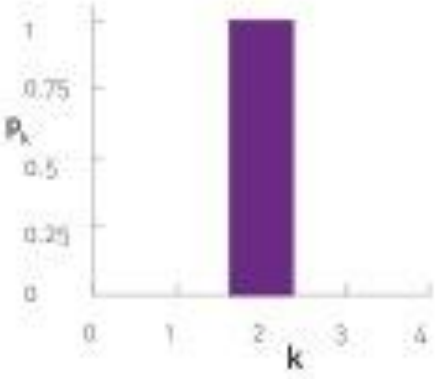
P(k) = Nk /N

Avec Nk = # noeuds avec degré k



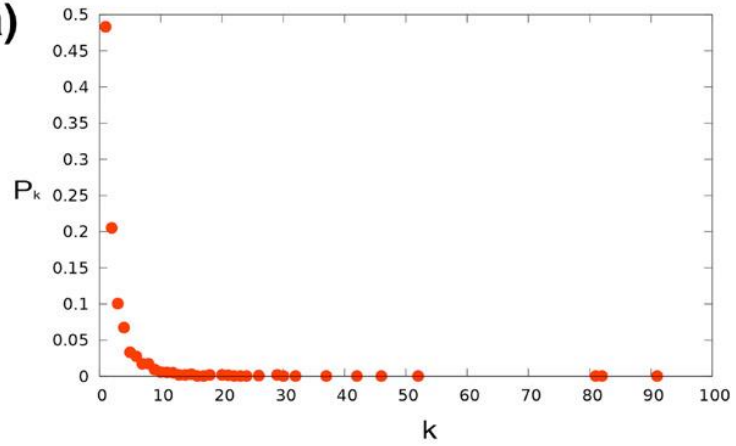




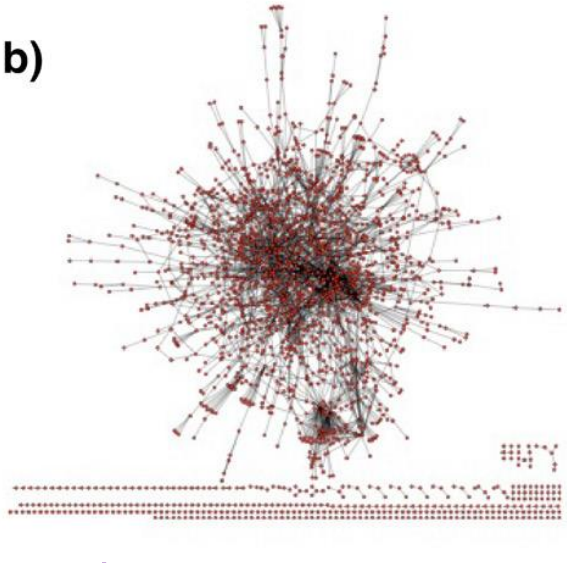


Adapted from Barabási, Network Science

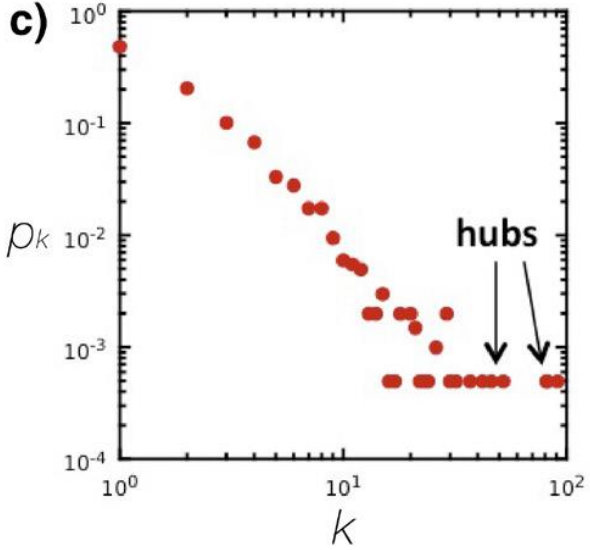
a)



b)

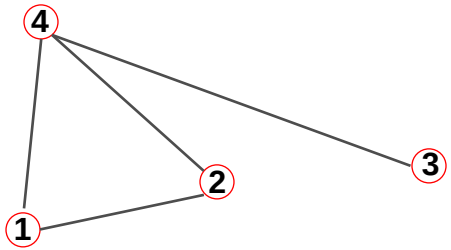


c)



Adapted from Barabási, Network Science

Undirected



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$k_i = \sum_{j=1}^{N} A_{ij}$$

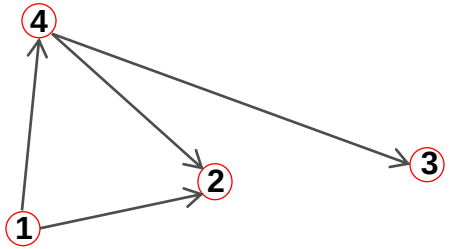
$$k_j = \sum_{i=1}^{N} A_{ij}$$

$$A_{ij} = A_{ji}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij} A_{ij}$$

Directed



$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$k_i^{in} = \sum_{j=1}^{N} A_{ij}$$

$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

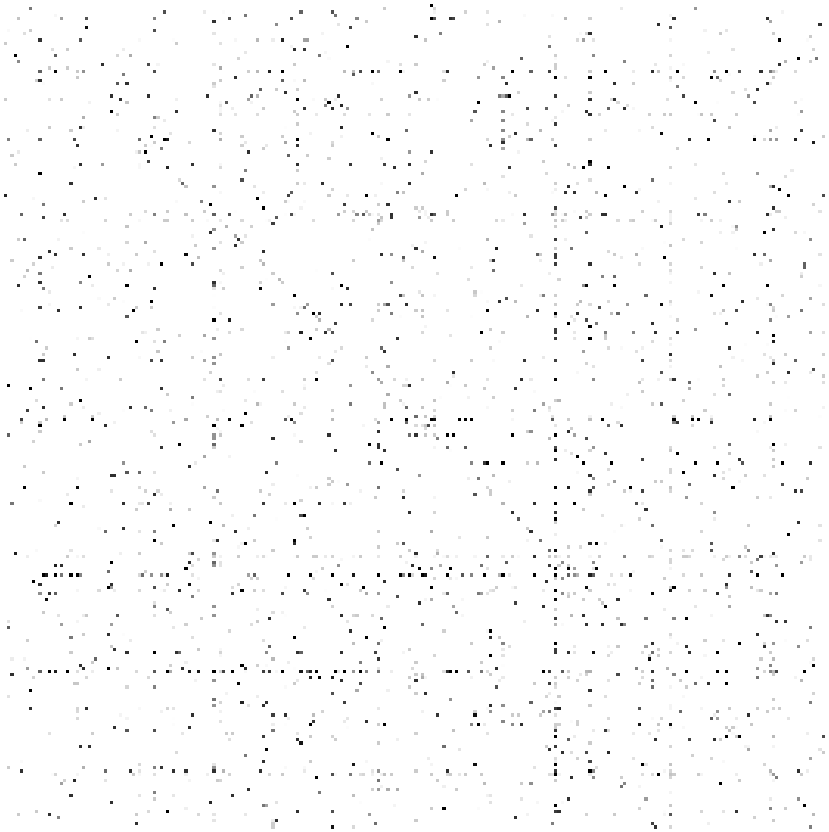
$$A_{ij} \neq A_{ji}$$

$$A_{ii} = 0$$

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j} A_{ij}$$

Adapted from Barabási, Network Science

Les reseaux biologiques sont *sparses*



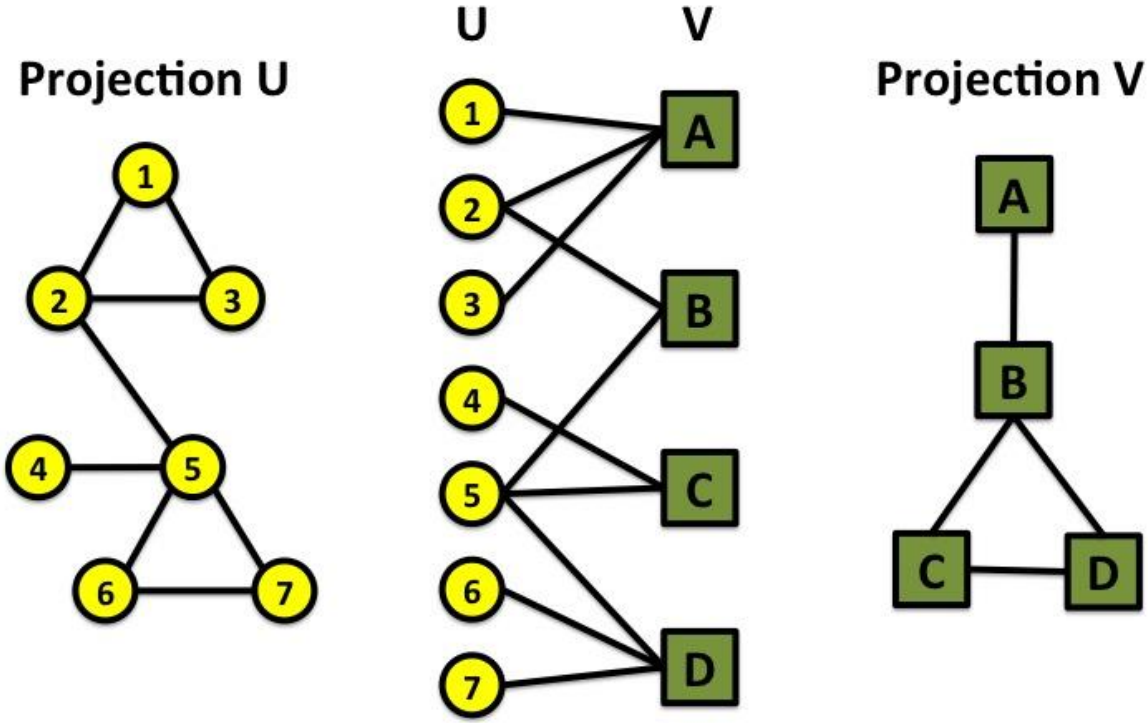
- Protein-protein interaction network
- 2018 noeuds
- 1 point: Aij = 1

Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature, 411(6833), 41.

Adapted from Barabási, Network Science

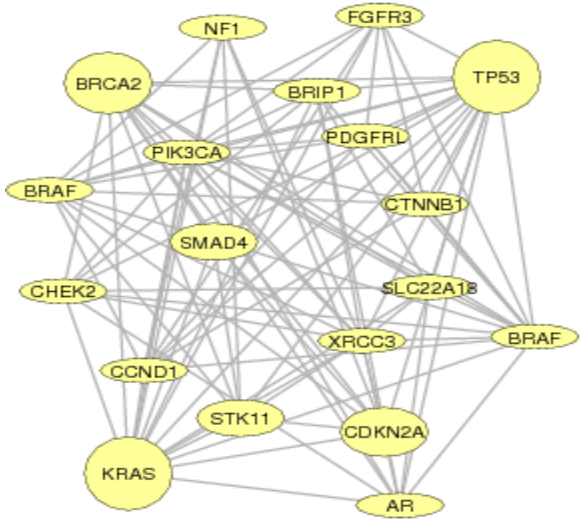
Réseau Bipartite (et multi-partite)

Un graphe **bipartite** (ou bigraphe) est un graphe dont les nœuds peuvent être **divisés en deux** ensembles disjoints U et V tels que chaque lien relie un nœud de U à un nœud de V ; autrement dit, U et V sont **des ensembles indépendants**.

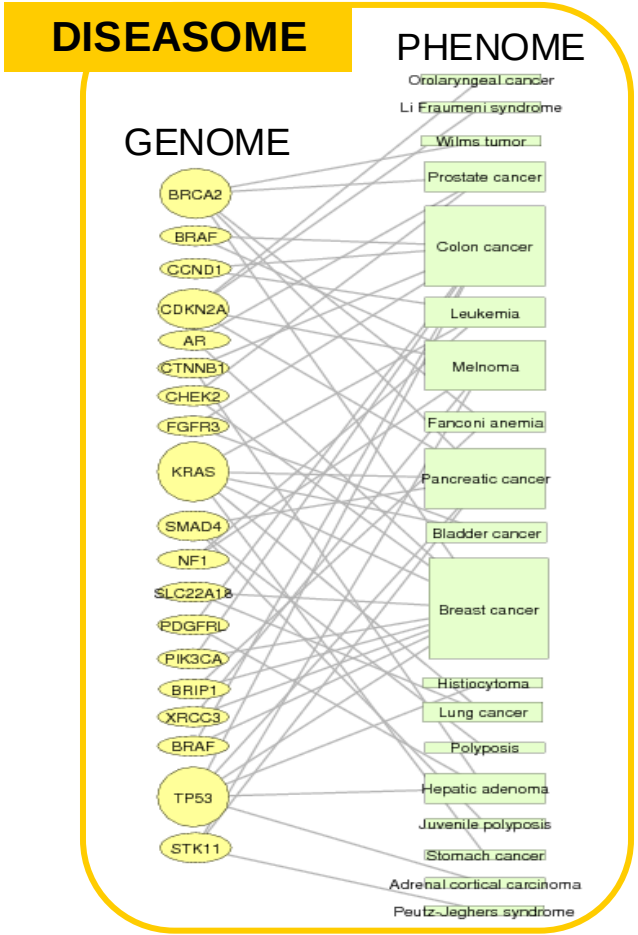


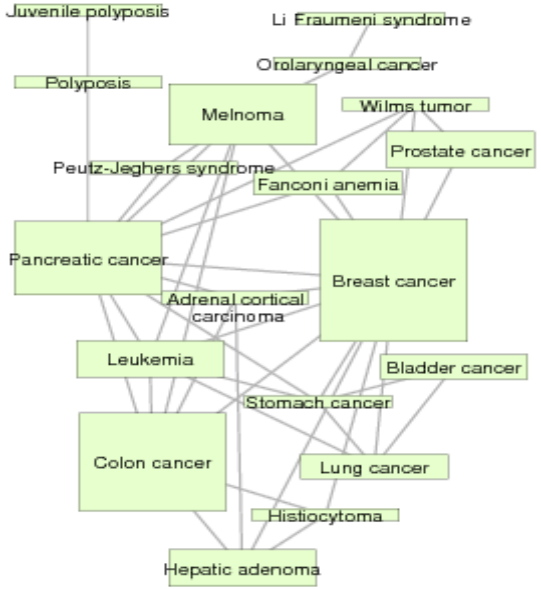
Adapted from Barabási, Network Science

Exemple



Gene network



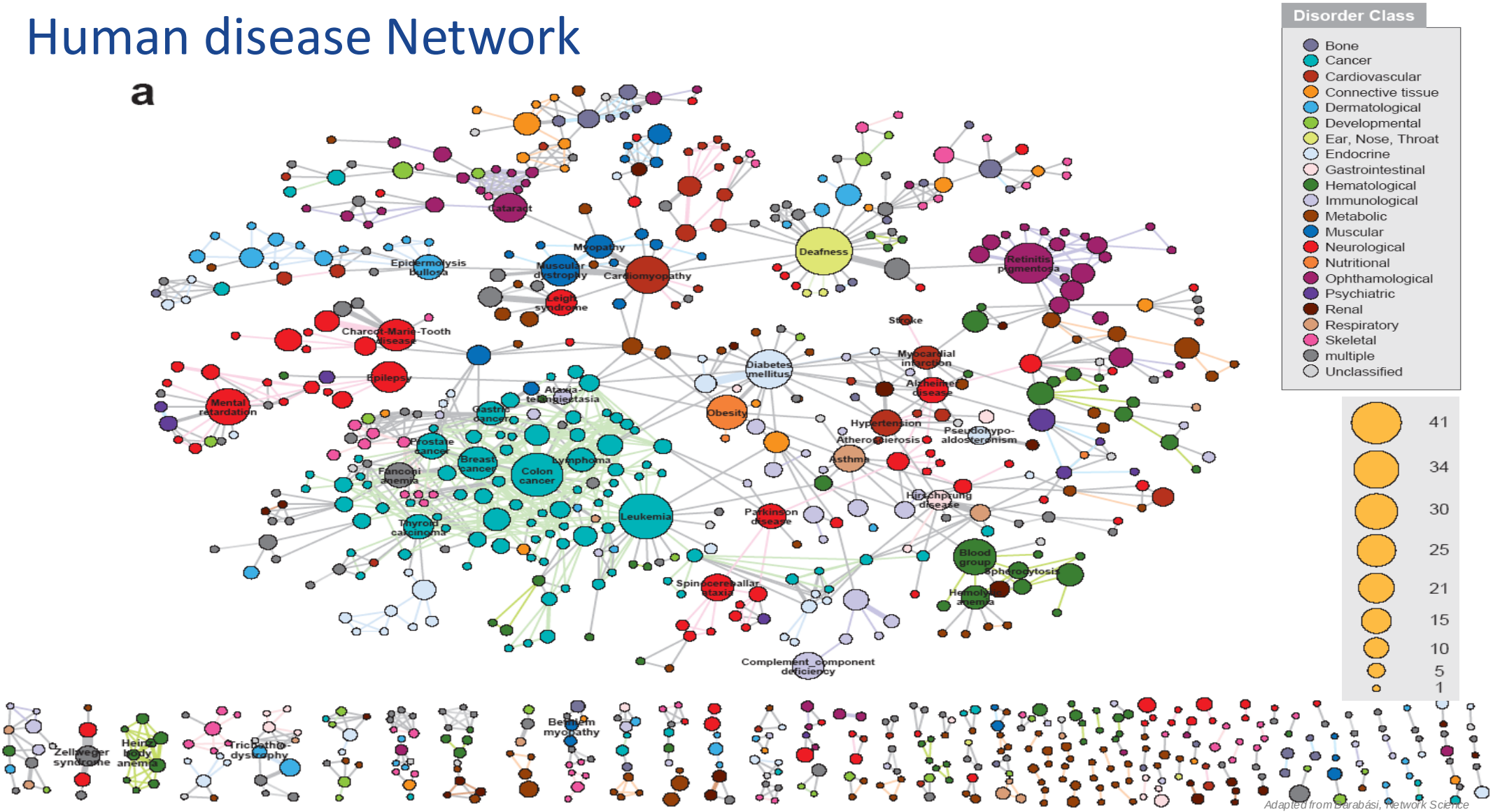


Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Adapted from Barabási, Network Science

Human disease Network

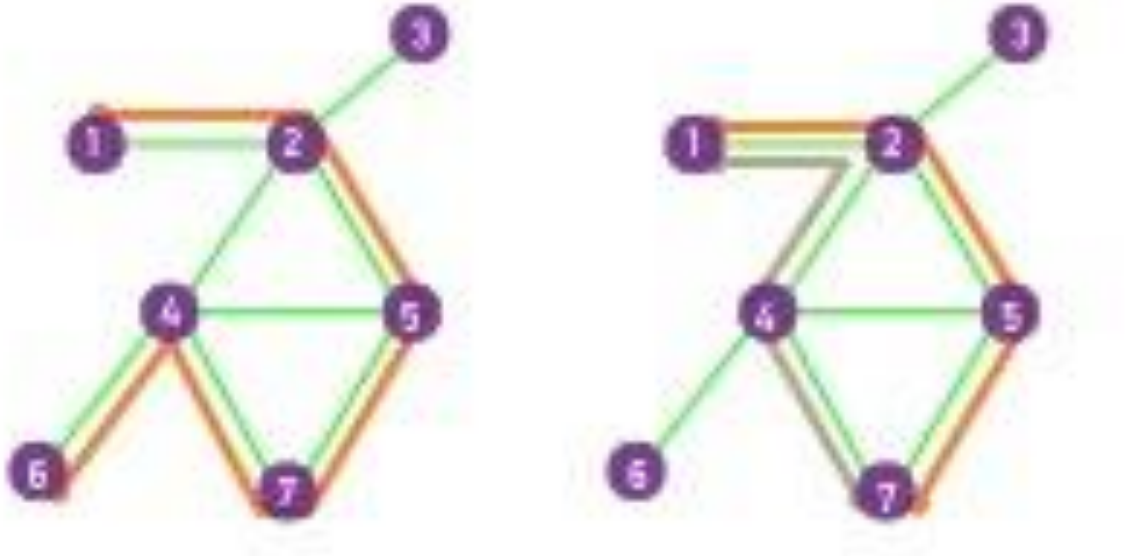


Adapted from Barabási, Network Science

Path

Un chemin est une séquence de nœuds dans laquelle chaque nœud est adjacent au suivant.
$P_{i0,in}$ de longueur n entre les nœuds i_0 et i_n est une collection ordonnée de n+1 nœuds et n liens.

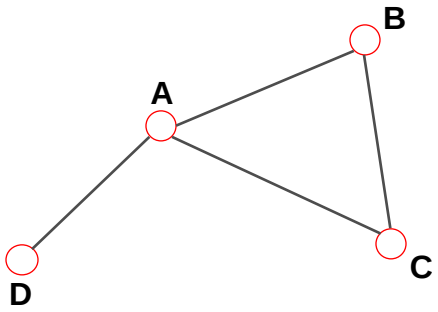
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$



Graphe dirigé = sens des flèches

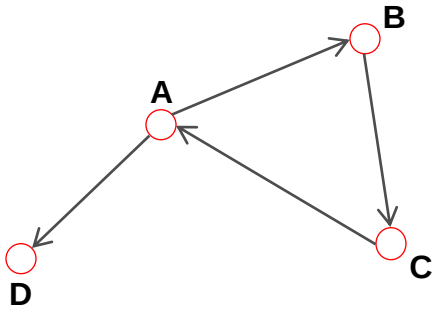
Adapted from Barabási, Network Science

Distance



The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them.

*If the two nodes are disconnected, the distance is infinity.

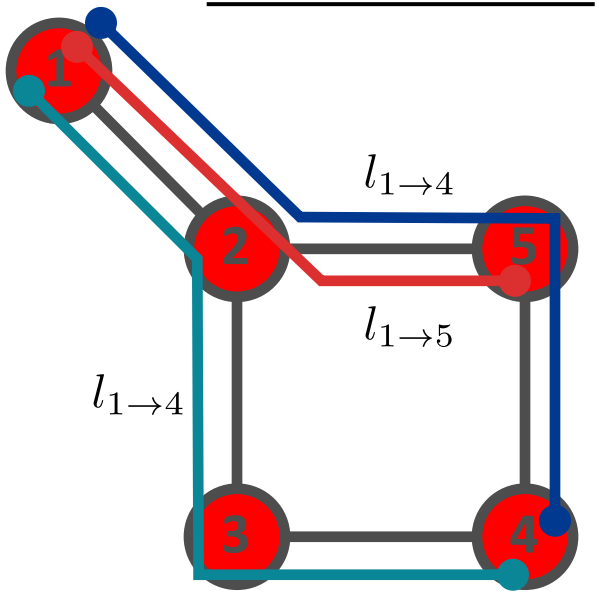


In directed graphs each path needs to follow the direction of the arrows.
Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

Adapted from Barabási, Network Science

Distance

Shortest Path



$$l_{1\to4} = 3$$

$$l_{1\to5} = 2$$

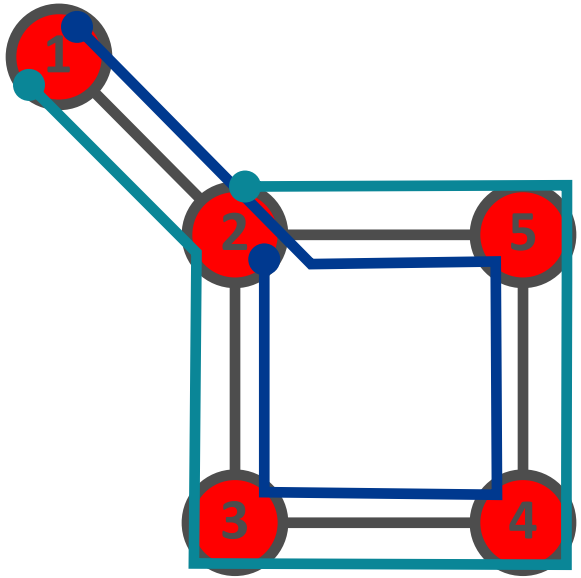
Chemin avec la plus courte distance entre 2 noeuds.

- Diametre: Longueur du plus long shortest path
- Longueur moyenne de tous les shortest paths
- Cycle
- Graphe Complet

Adapted from Barabási, Network Science

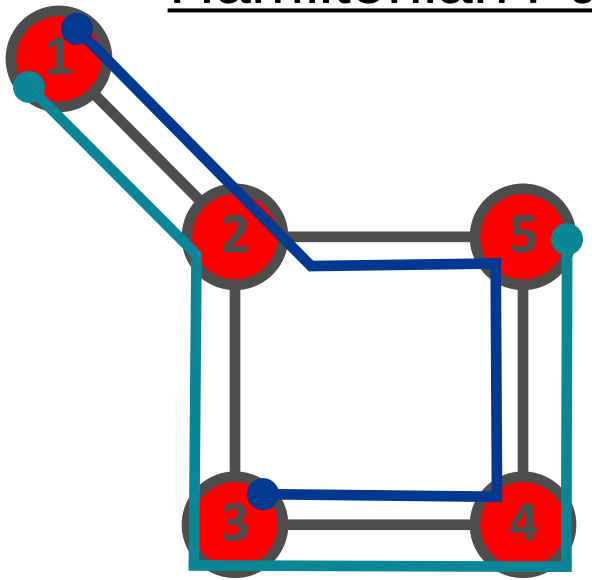
Distance

Eulerian Path



Un chemin qui traverse chaque **lien** exactement une fois.

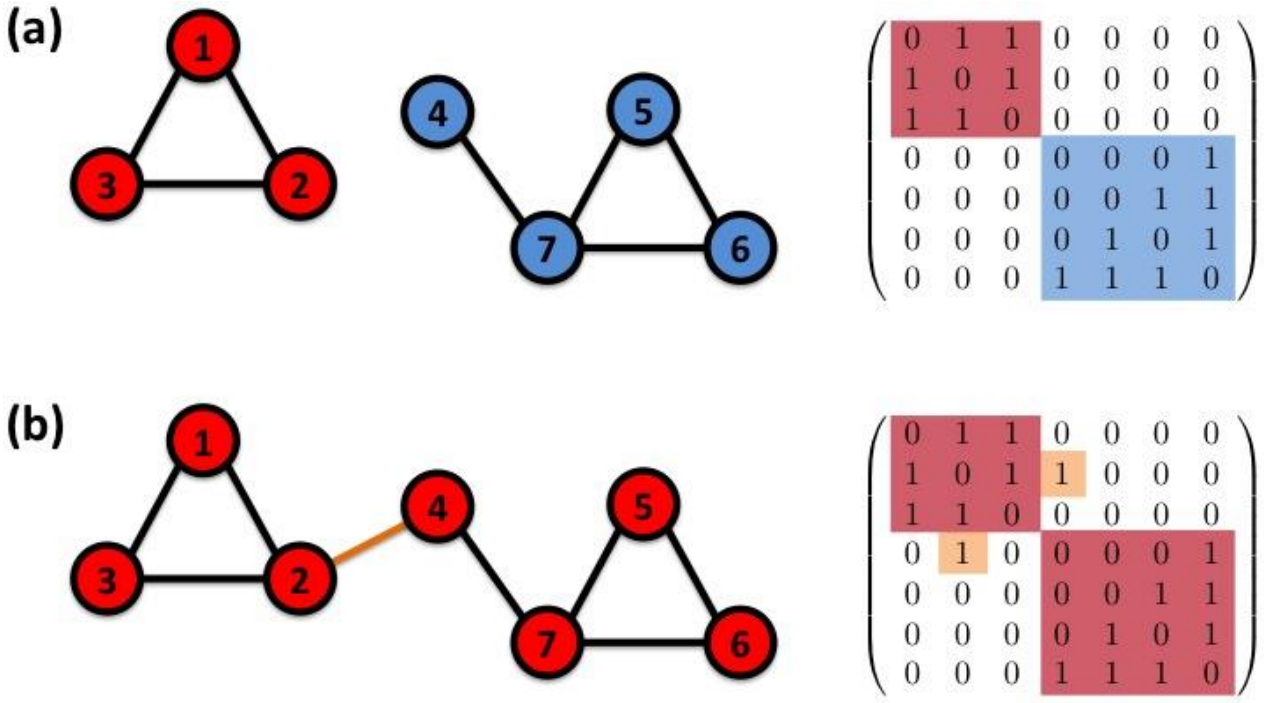
Hamiltonian Path



Un chemin qui visite chaque **nœud** exactement une fois.

Adapted from Barabási, Network Science

Connectivité



$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Adapted from Barabási, Network Science

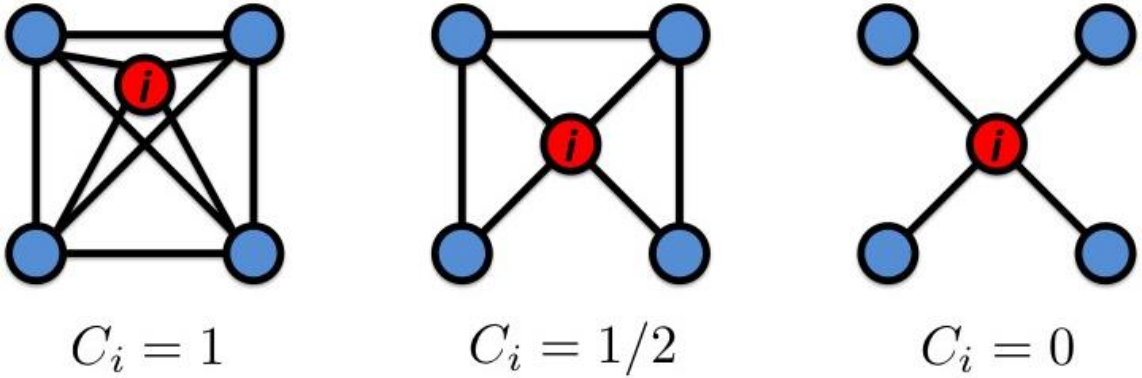
Coefficient de clustering

Pour un nœud i avec un degré k_i, quelle proportion des voisins sont connectés entre eux ?

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

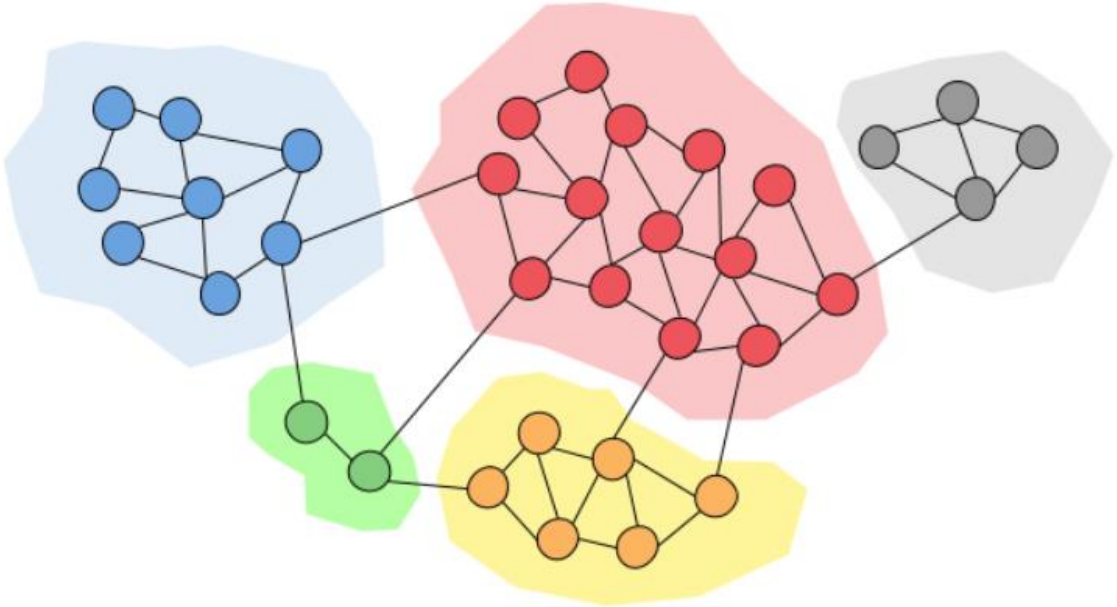
E_i = nombre de lien entre les voisins de k
C_i in [0,1]

Watts & Strogatz, Nature 1998.



Adapted from Barabási, Network Science

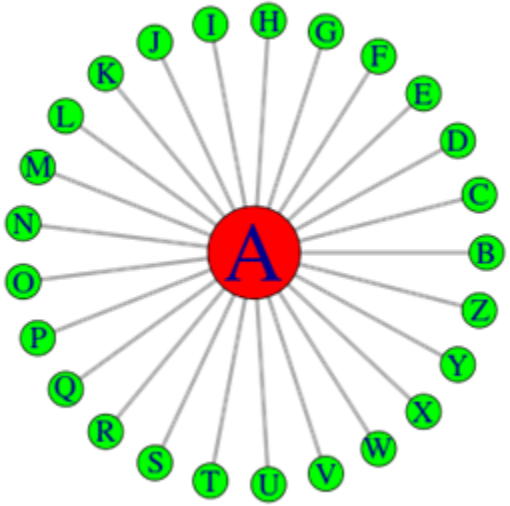
Modularité



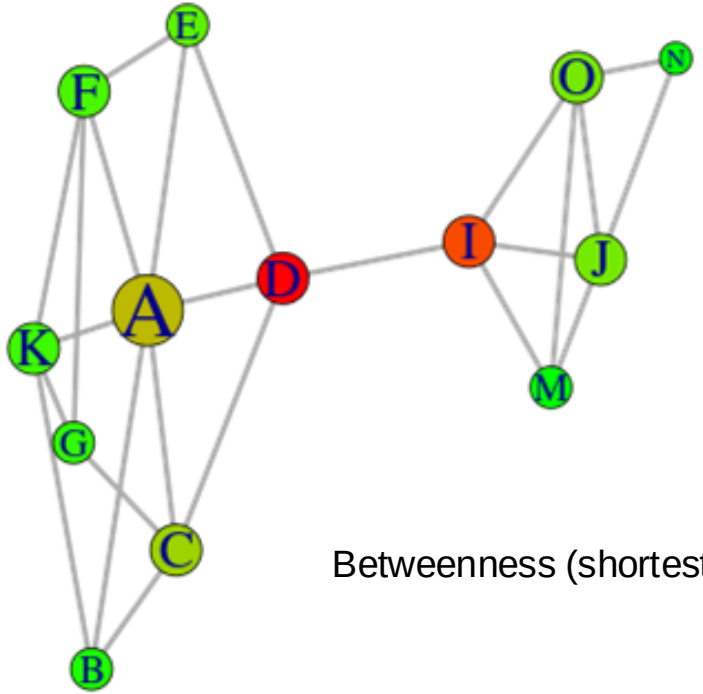
- Méthodes *agglomérantes (Louvain)*
- Méthode *divisantes*

- *Guilt by associations*

Autres mesures de centralités



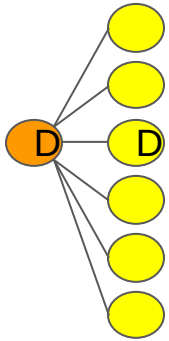
Degrés et hubs



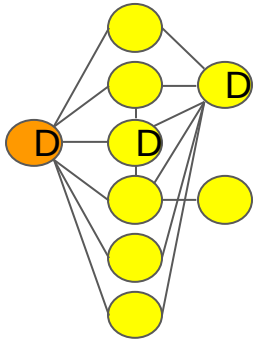
Betweenness (shortest path)

Propagation de signal: Random walk

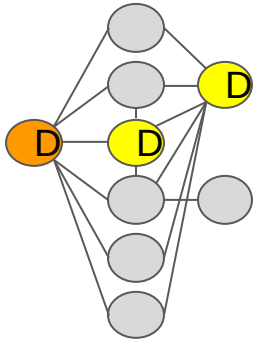
- Random Walk



Direct neighbors
- False positives
- False negatives

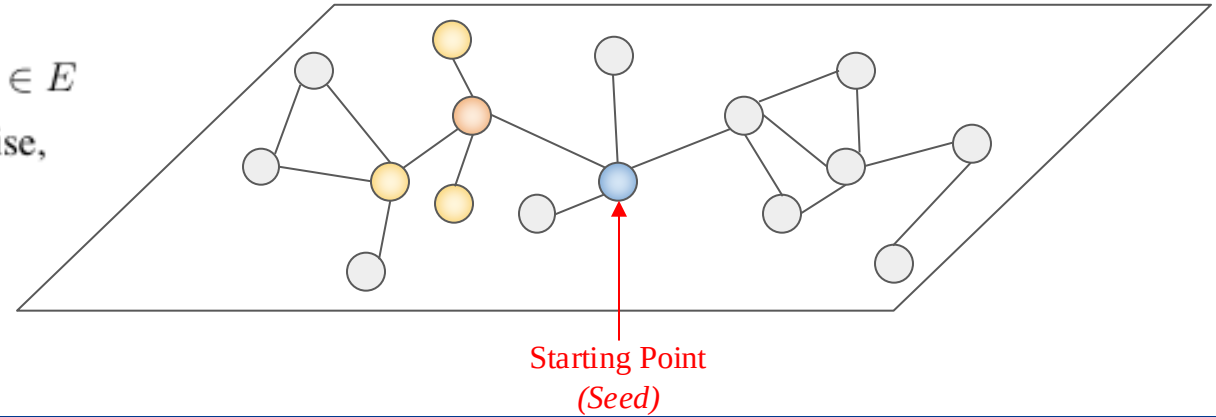


Shortest Path
- *Small world*
- False positives



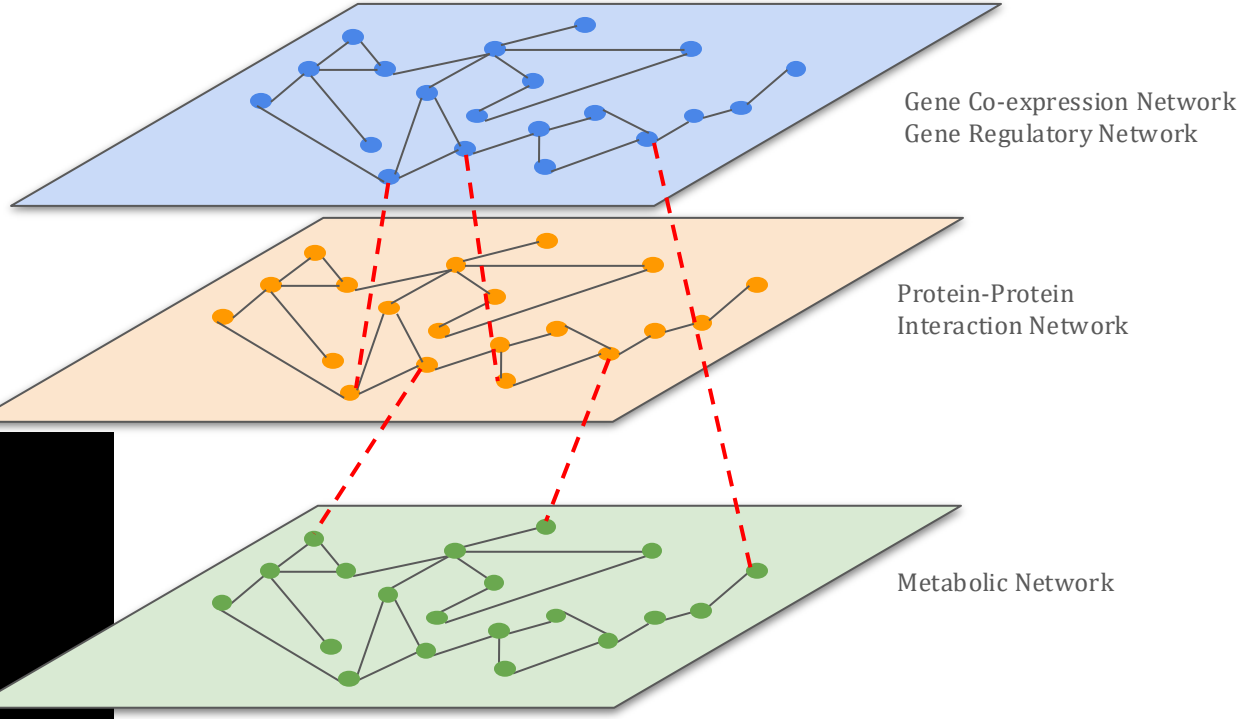
Network Propagation
- Random Walk
- Node Ranking

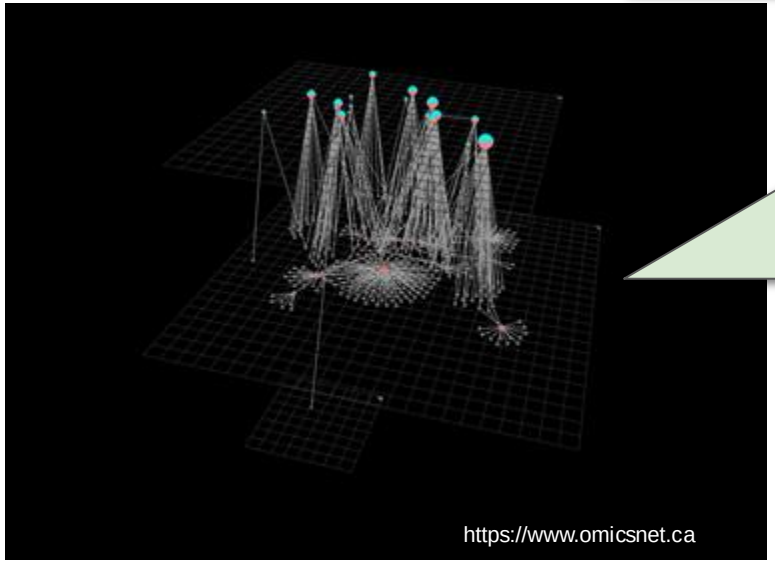
$$\mathbb{P}(v_{t+1} = y | v_t = x) = \begin{cases} \frac{1}{d(x)} & \text{if}(x, y) \in E \\ 0 & \text{otherwise,} \end{cases}$$



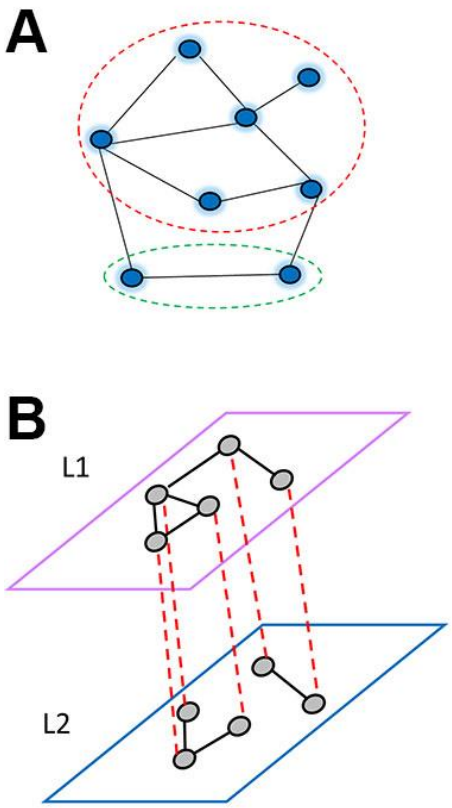
Multi-Omics Networks

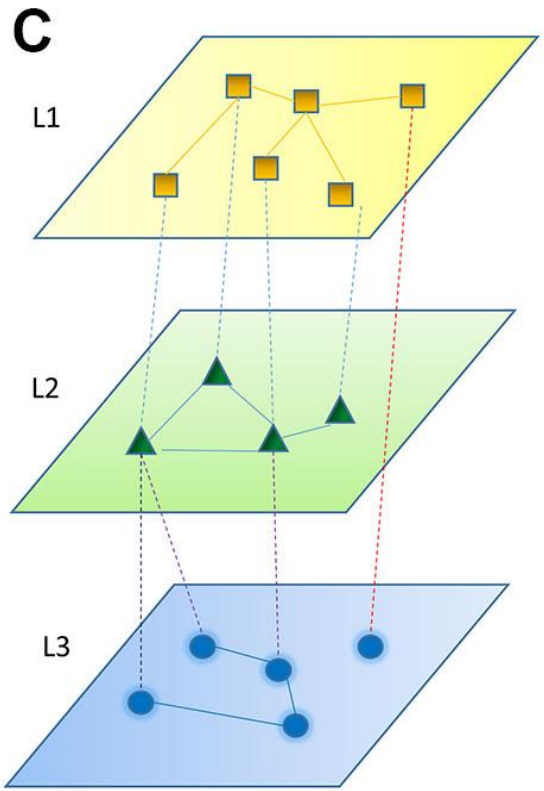
Multi-Layered

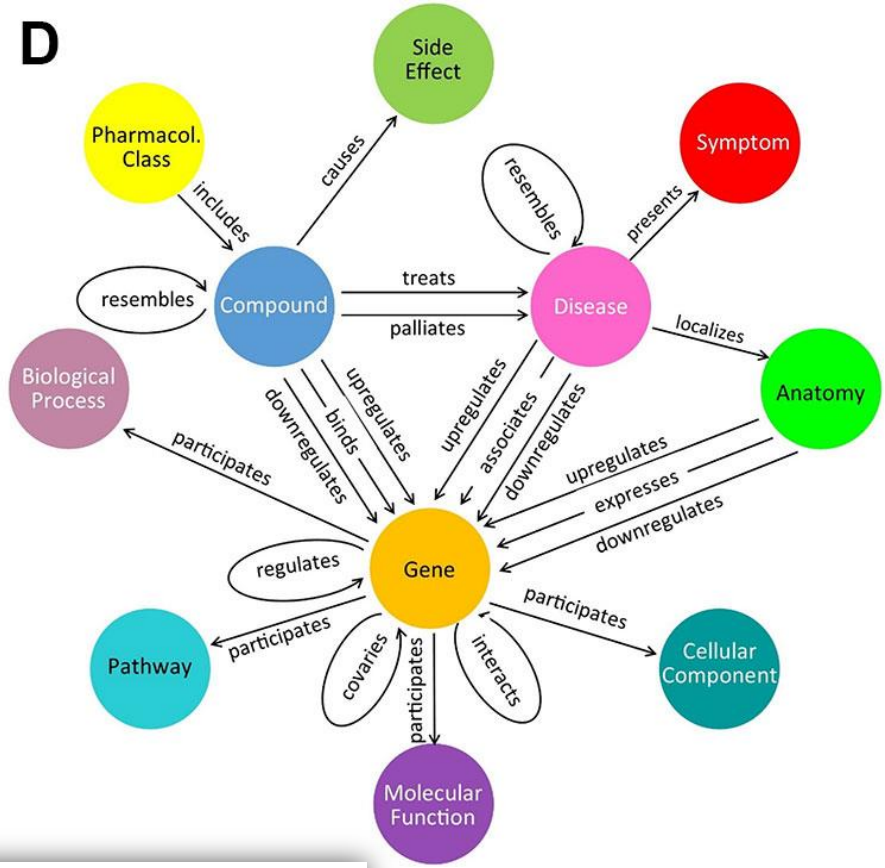




Multiplex Heterogeneous Network







[HTML] Heterogeneous multi-layered network model for omics data integration and analysis

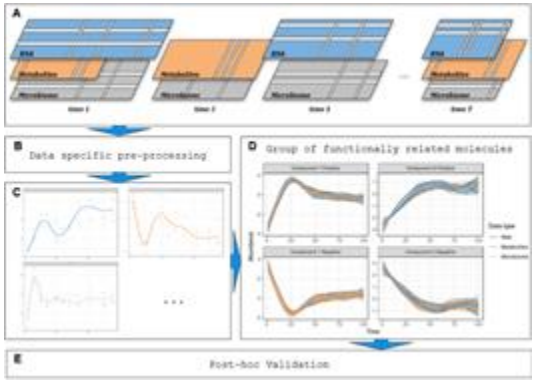
B Lee, S Zhang, A Poleksic, L Xie - Frontiers in genetics, 2020 - frontiersin.org

Dernières tendances

- Omiques inhabituels
  - Microbiome: iHMP (https://www.hmpdacc.org/ihmp/)
  - Single-cell Multi-omics
  - Wearable (montres connectés, …)

- Aspect Longitudinale
  Suivre la dynamique des systèmes (complexe) au niveau moléculaire

- Méthode d'exploration des graphes
  (Random Walk, Steiner trees, …)



timeOmics DOI:10.3389/fgene.2019.00963

Mise en contexte

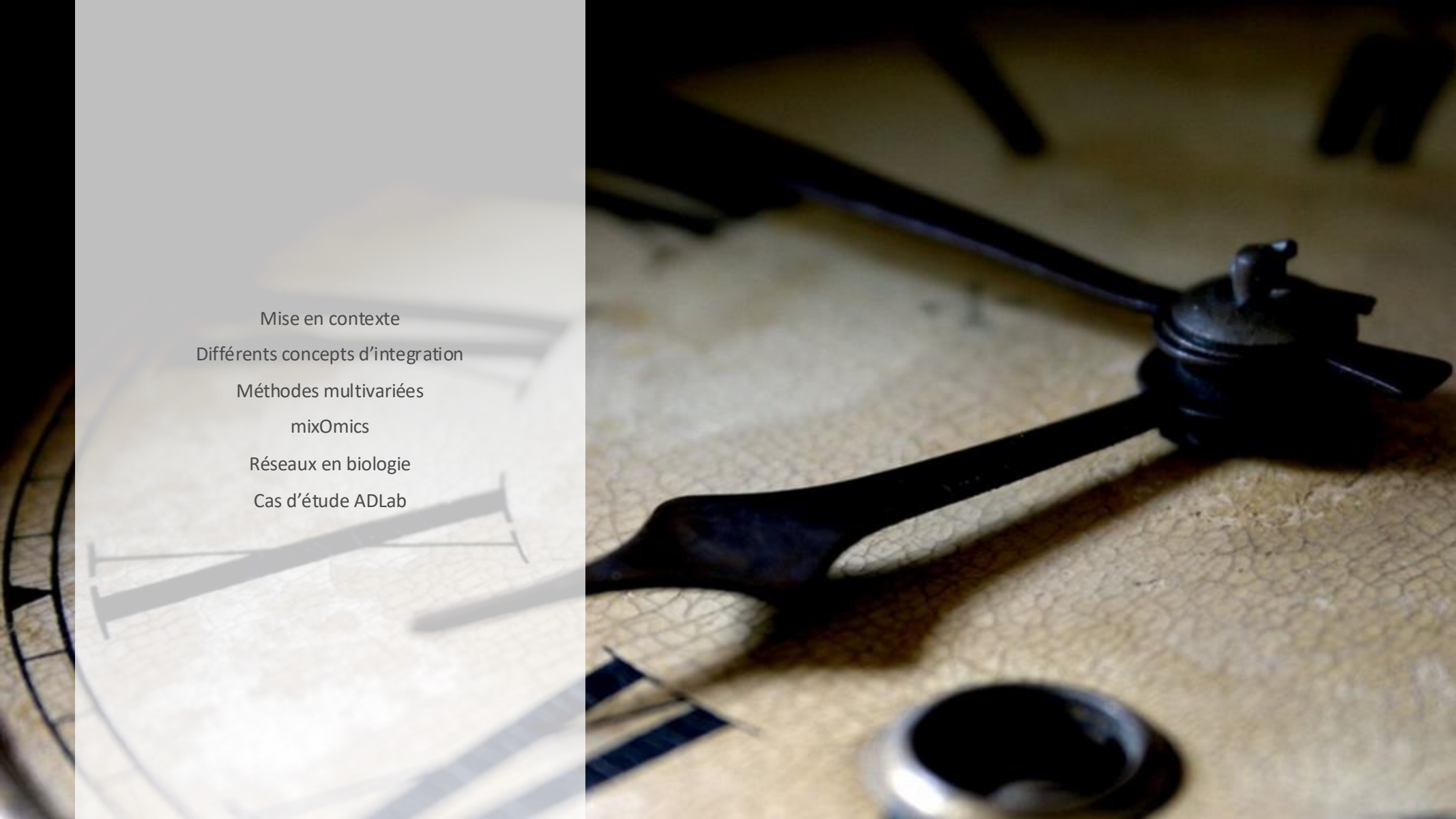
Différents concepts d'integration

Méthodes multivariées

mixOmics

Réseaux en biologie

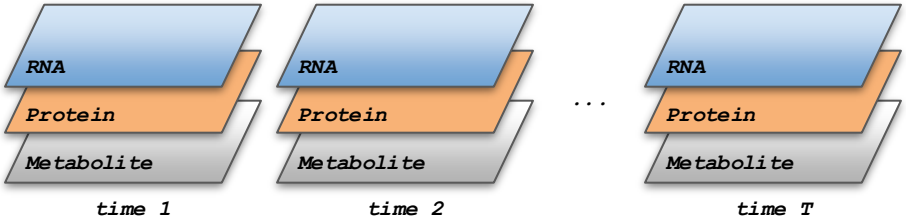
Cas d'étude ADLab

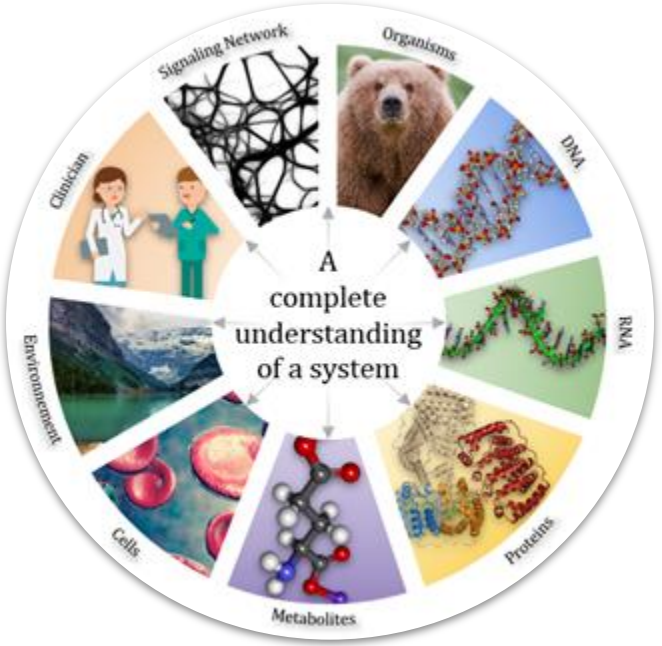


Intégration longitudinale

New experimental **design**

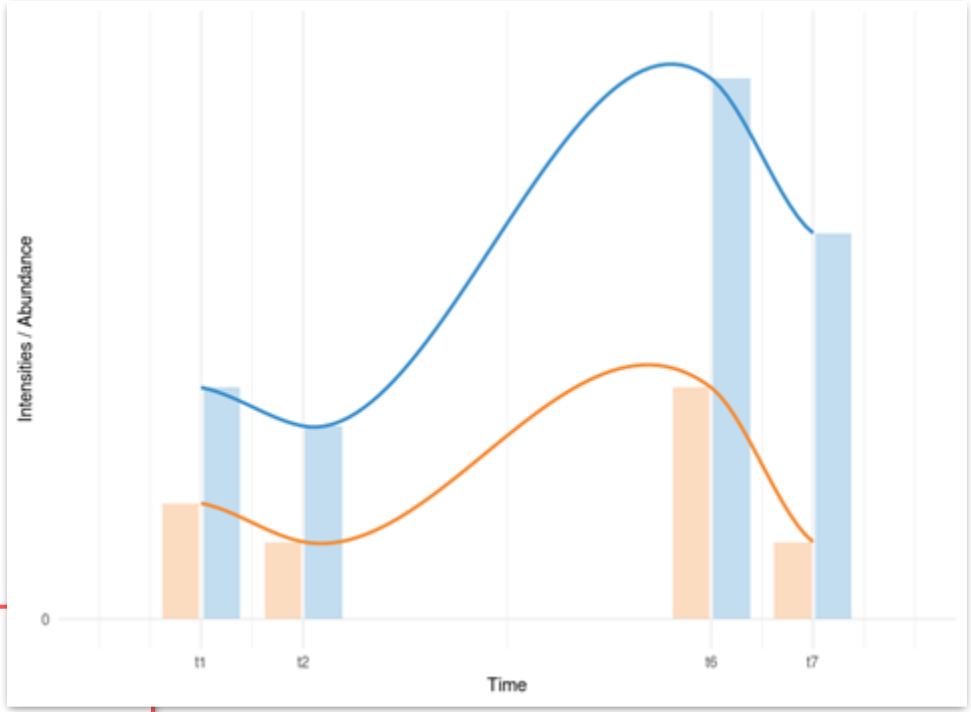
Multi-OMICS **time** series data





Intégration longitudinale

- ➔ Time effect
  - ◆ between two time points
  - ◆ between intervals of time points
- ➔ Group effect
- ➔ **Detect similar pattern**

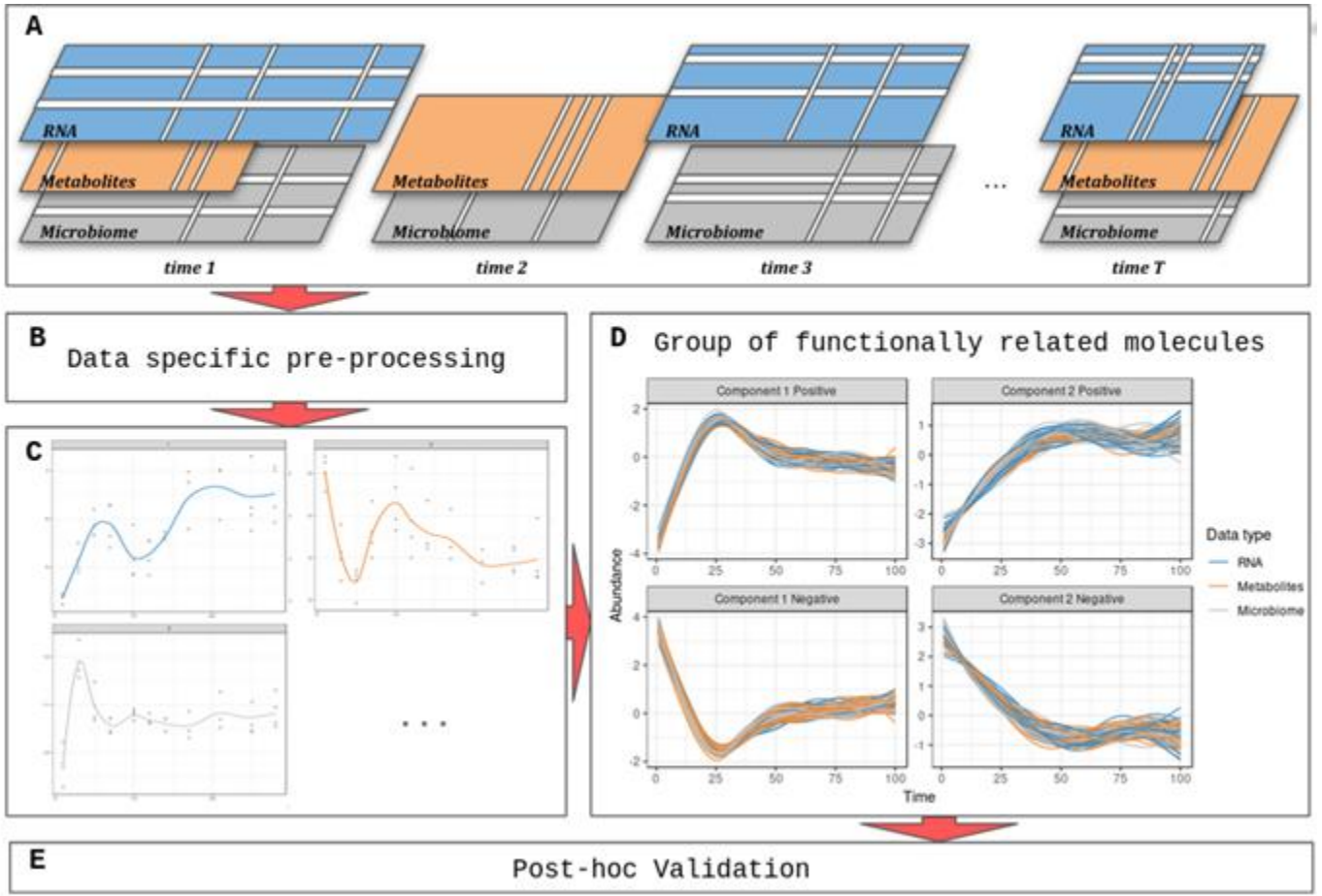


Why cluster data ?

- ❑ Group molecules with similar expression profile
- ❑ Functionally related
- ❑ Can be used to infer regulatory relationships

timeOmics

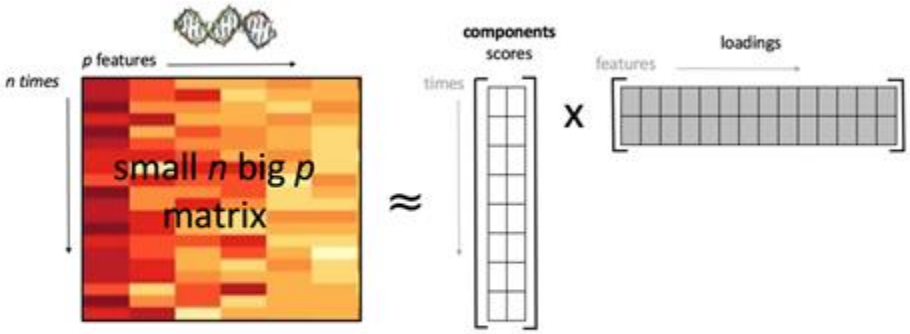
A. Data Acquisition
B. Pre-processing
C. Modelling and Filtering
D. Clustering / Integration
E. Validation



timeOmics DOI:10.3389/fgene.2019.00963

timeOmics

Principal Component Analysis - PCA

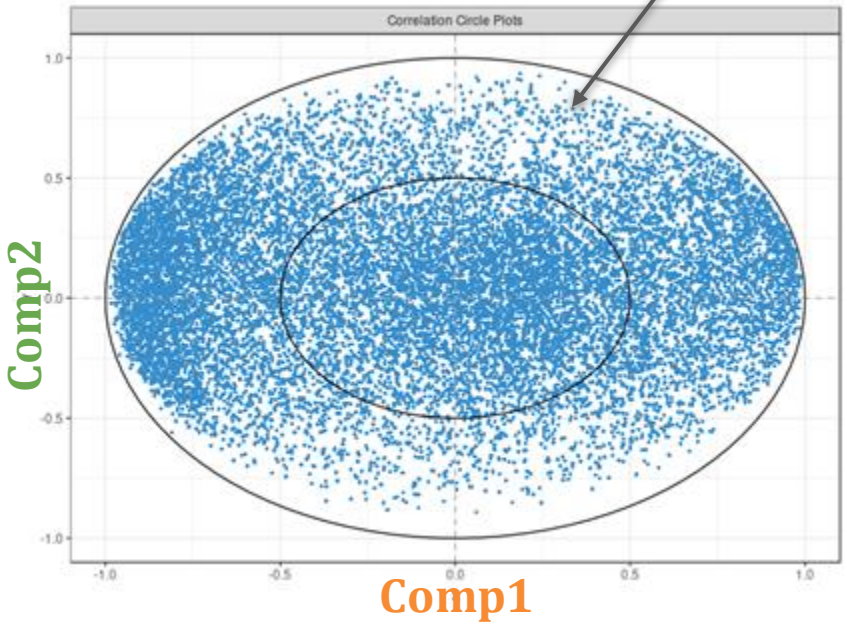


- Block = OTU table
- Unsupervised
- Dimension reduction
- Maximize **variance**

(`mixOmics` R package)

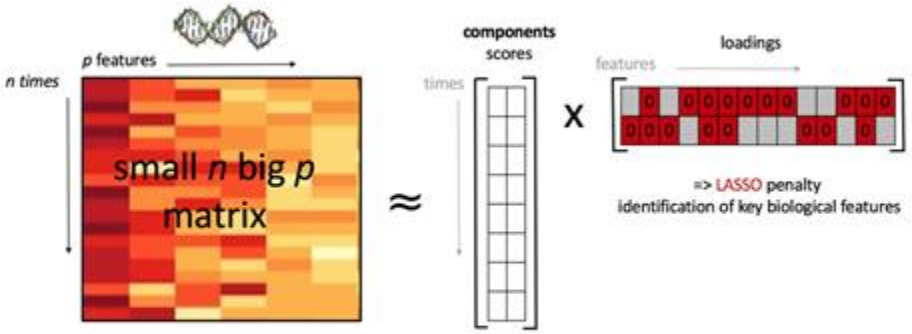
1 point = 1 time profile (genes, proteins, ...)

Circle Correlation Plot



timeOmics

sparse Principal Component Analysis - sPCA

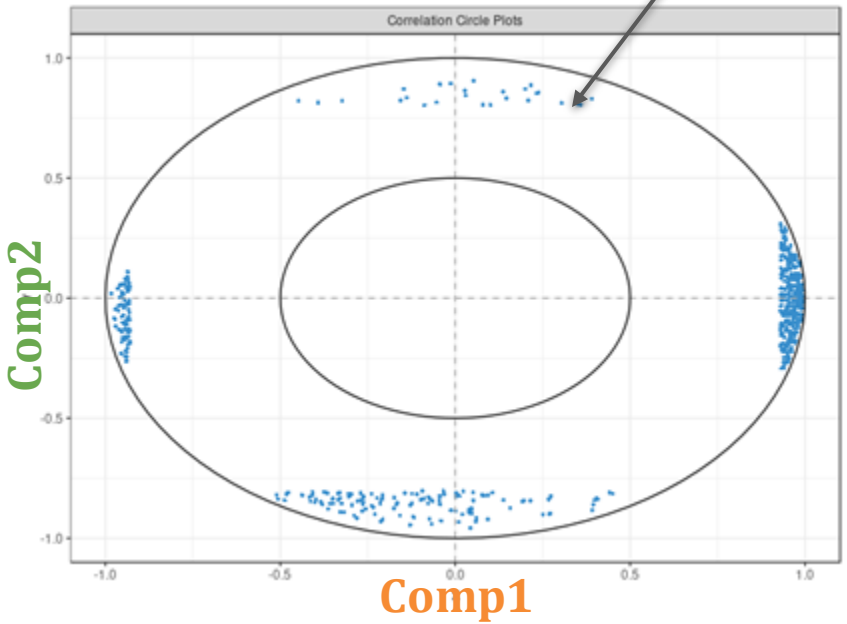


- Block = OTU table
- Unsupervised
- Dimension reduction
- Maximize **variance**
- Sparse version for **feature selection**

(`mixOmics` R package)

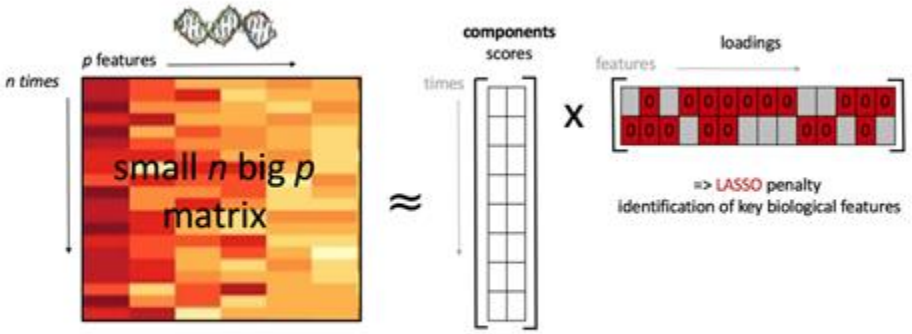
1 point = 1 time profile (genes, proteins, ...)

Circle Correlation Plot

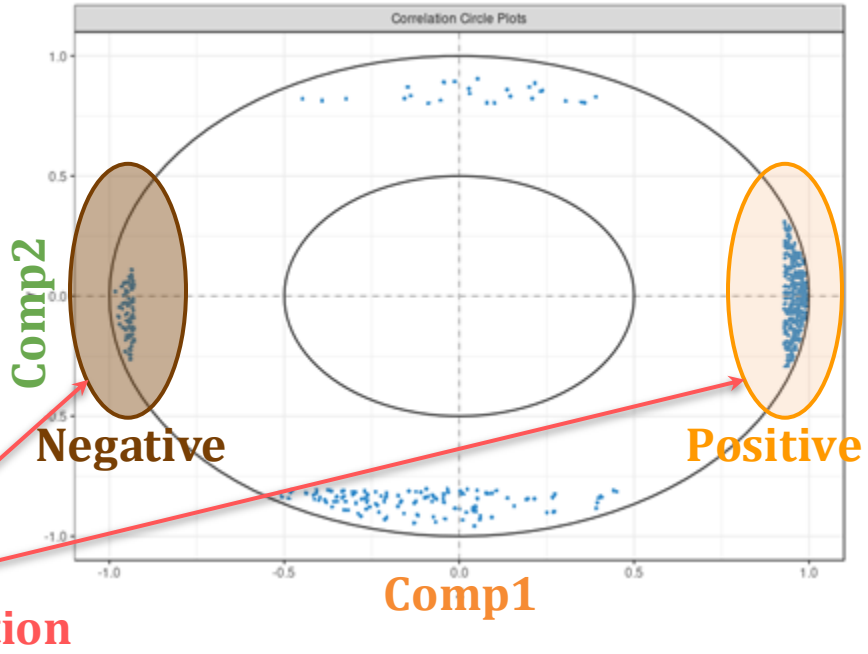


timeOmics

sparse Principal Component Analysis - sPCA



Circle Correlation Plot



Each component selects **2 clusters of OTUs**

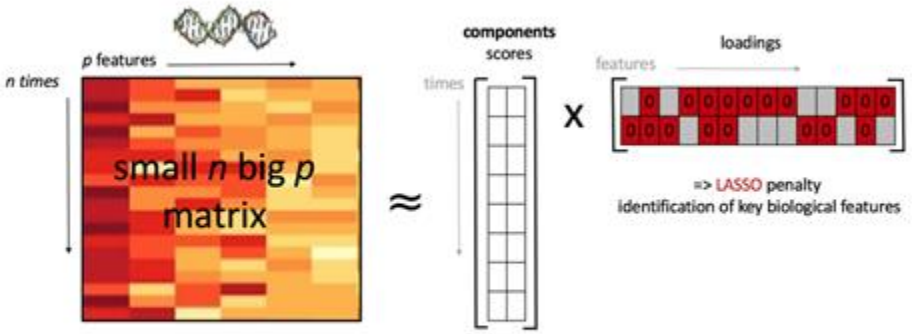
- Positively correlated
- Negatively correlated

Negative Correlation

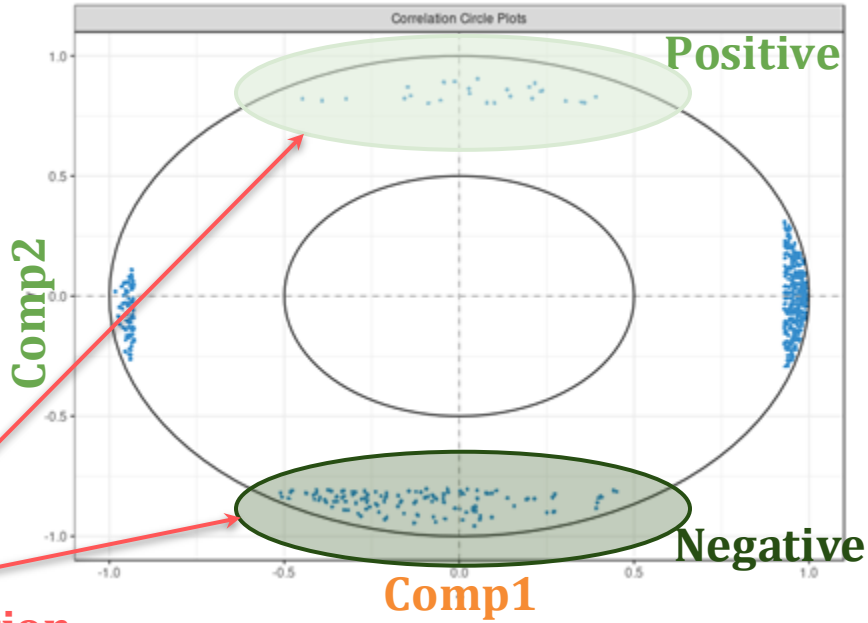
Two clusters identified on **comp 1**

timeOmics

sparse Principal Component Analysis - sPCA



Circle Correlation Plot



Each component selects **2 clusters of OTUs**

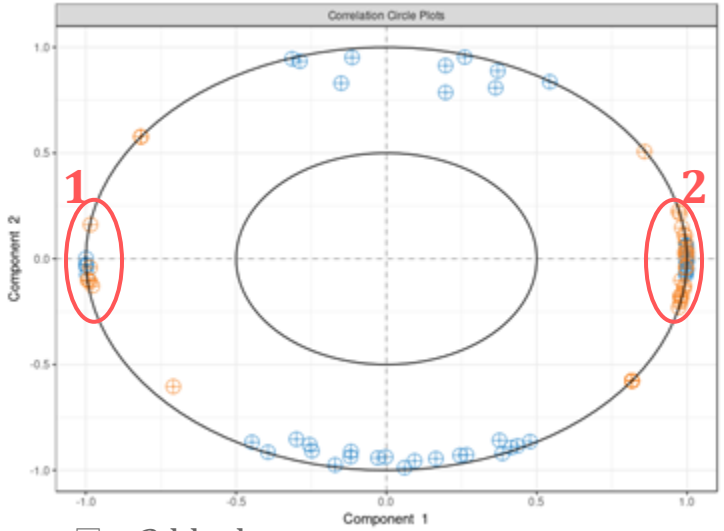
- Positively correlated
- Negatively correlated

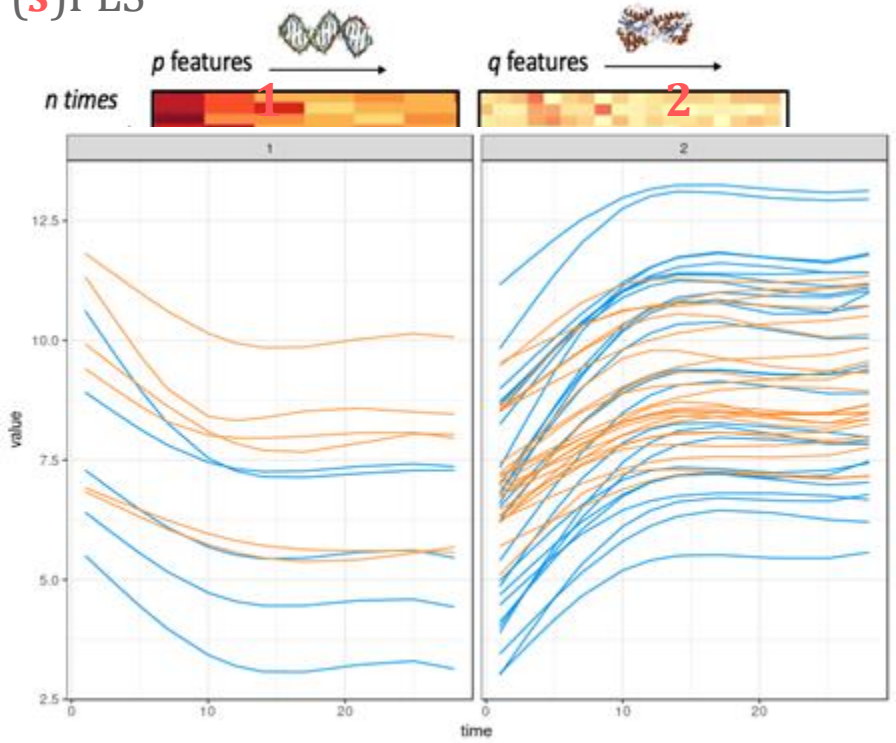
Negative Correlation

Two clusters identified on **comp 2**

timeOmics

(***sparse***) Projection on Latent Structures - (***s***)PLS





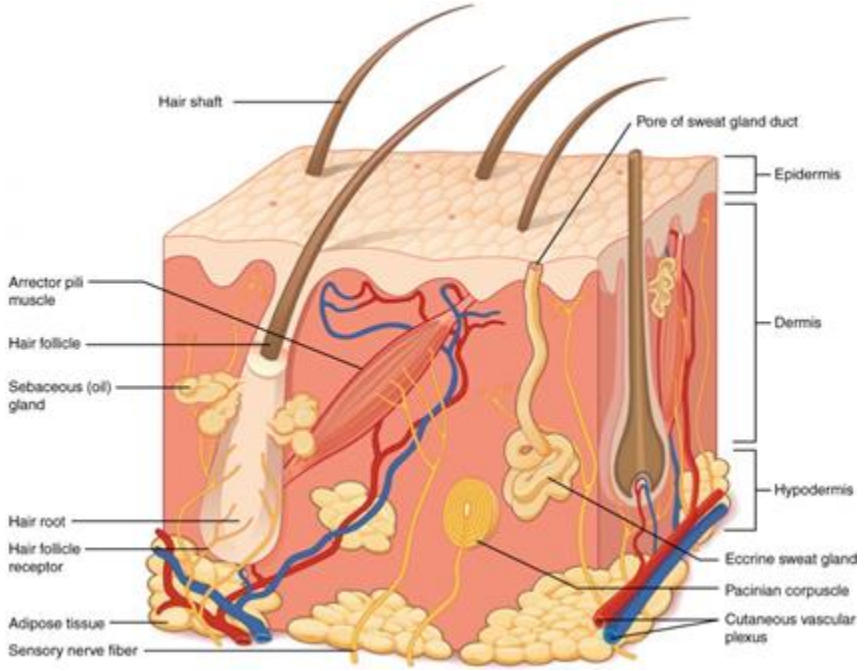
- ❑ 2 blocks
- ❑ sparse PLS = feature extraction

(`mixOmics` R package)

Example: epidermis reconstruction

Main function of the human skin:

- ❑ Organ of Sensation
- ❑ Thermo-regulation
- ❑ Secretion
- ❑ Immunité
- ❑ Psycological
- ❑ **Protection**

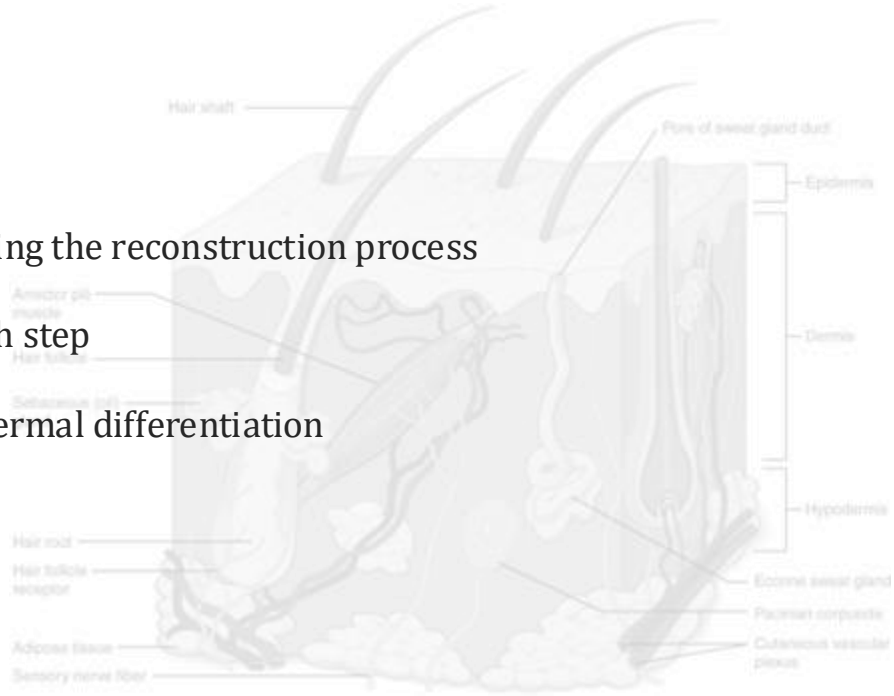


High capacity for regeneration

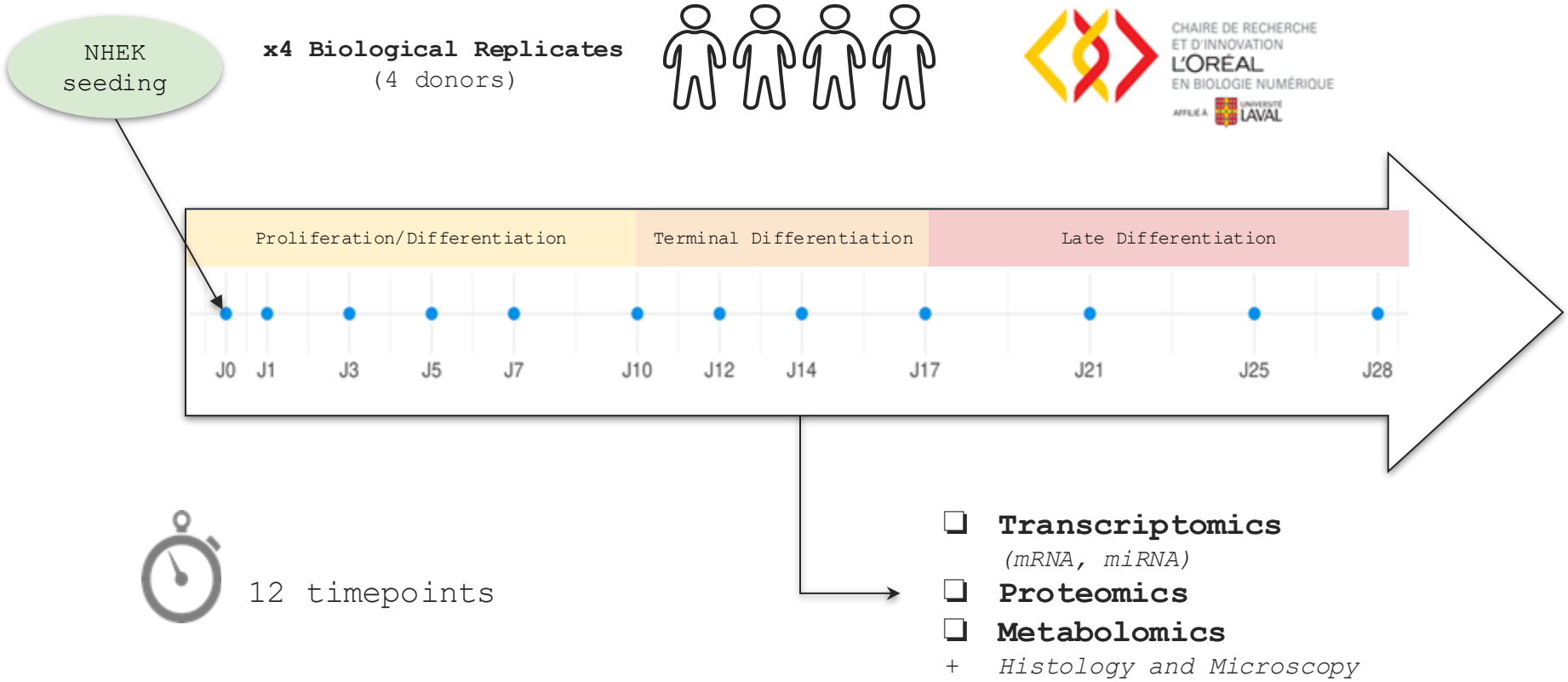
We aim to study the kinetics of epidermal differentiation during the reconstruction process.

Example: epidermis reconstruction

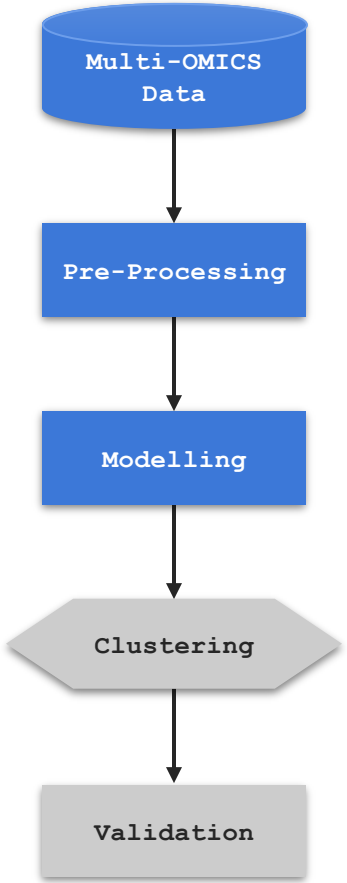
- ❏ Understand the epidermal differentiation during the reconstruction process
- ❏ Identify key biological players involved in each step
- ❏ Identify key and transitional steps of the epidermal differentiation
- ❏ Identify key missing players



Example: epidermis reconstruction



Example: epidermis reconstruction



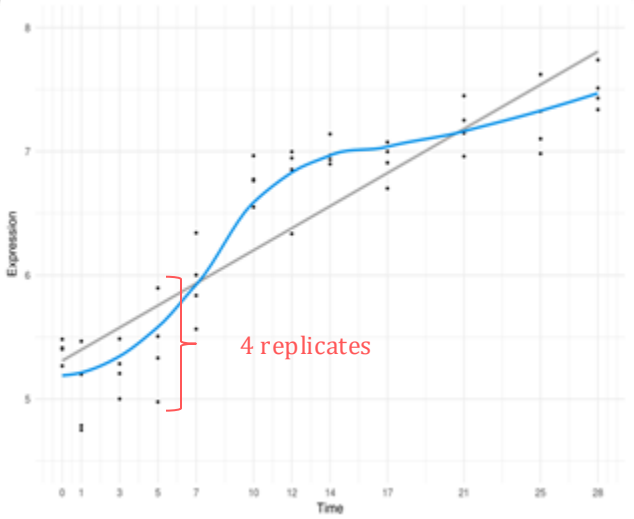
OMICS	# Molecules	# Molecule after filtering
mRNA	15,348	1,979 *(13%)*
miRNA	896	171 *(19%)*
Proteomics	1,820	987 *(54%)*
Metabolomics	428	338 *(79%)*

Filtering

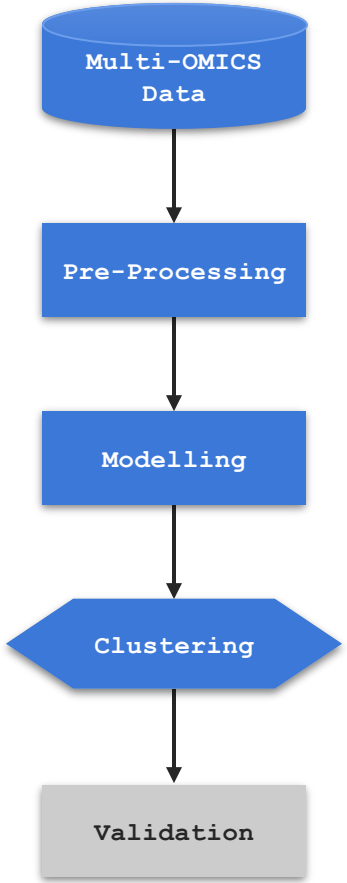
- ❑ "*flat*" profiles
- ❑ Molecules with no D.E. in time

Modelling

- ❑ Linear Mixed Model Spline
- ❑ Filter noisy profile

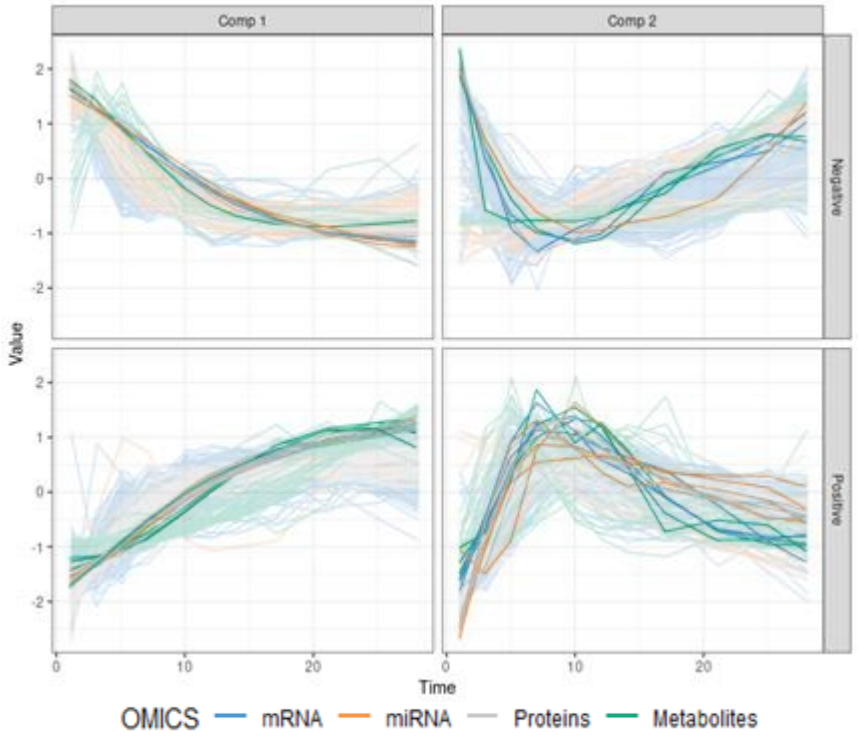


Example: epidermis reconstruction



Cluster -1 *(Comp1 Neg)*
mRNA: 418 *(5)*
miRNA: 87 *(18)*
Proteins: 179 *(5)*
Metabolites: 62 *(16)*

Multi-OMICS (sparse) clusters

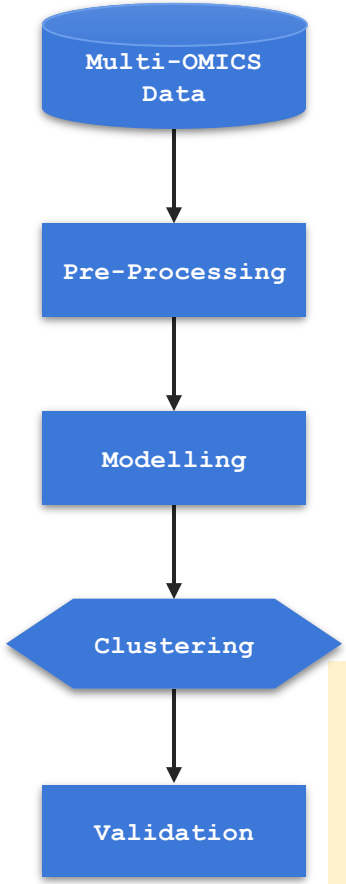


Cluster -2 *(Comp2 Neg)*
mRNA: 228 *(7)*
miRNA: 10 *(7)*
Proteins: 197 *(3)*
Metabolites: 58 *(6)*

Cluster 1 *(Comp1 Pos)*
mRNA: 898 *(10)*
miRNA: 21 *(17)*
Proteins: 317 *(20)*
Metabolites: 127 *(9)*

Cluster 2 *(Comp2 Pos)*
mRNA: 435 *(8)*
miRNA: 53 *(8)*
Proteins: 294 *(12)*
Metabolites: 91 *(9)*

Example: epidermis reconstruction



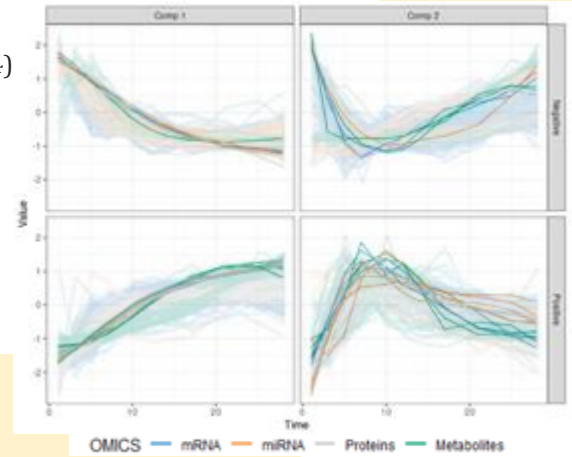
❑ **Single OMIC enrichment**
Hypergeometric test against OMIC specific background set
(Gene Ontology BP, Kegg, Reactome, WikiPathways)

❑ **Multi-OMICS enrichment**
significant function represented by different omics
(Fisher's combined probability test)

(`grofiler2` `FELLA` R packages)

Cluster -1
(mRNA-miRNA-proteins): regulation of cell migration (10e-16), anatomical structure morphogenesis (10e-21), ..
(all): Focal adhesion (10e-9)
(mRNA): angiogenesis (10e-4), nervous system development (10e-4)

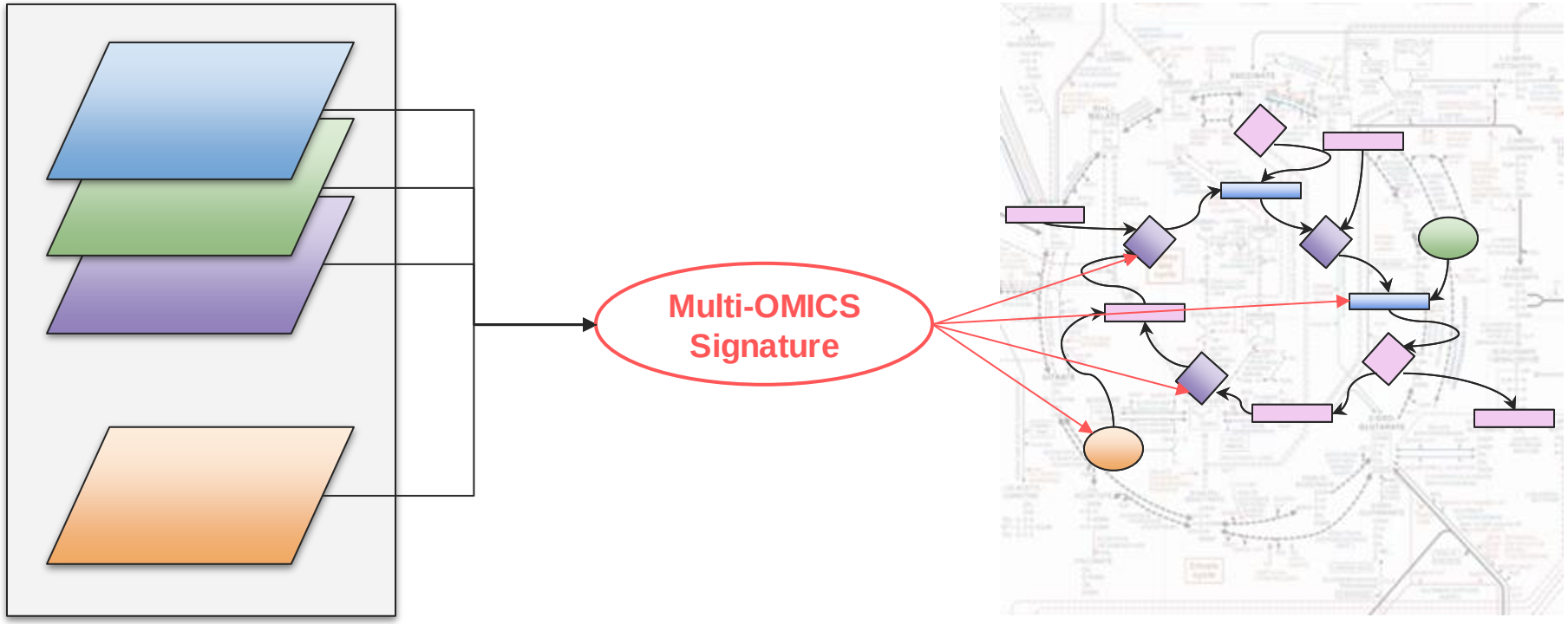
Cluster -2
(mRNA-miRNA-proteins): Extracellular matrix organization (10e-7)
(metabolite): Glycosphingolipid biosynthesis (10e-4)
(proteins): cornification (10e-5), keratinization (10e-4)



Cluster 1
(mRNA-miRNA-proteins):
Keratinisation (10e-36),
Cornification (10e-24), **epidermis** development (10e-22),
..
(all): Sphyngolipid metabolism (10e-4), ...

Cluster 2
(mRNA-miRNA-proteins): **cornification** (10e-26),
epithelial cell differentiation (10e-17),
tissue development (10e-13)

Example: epidermis reconstruction

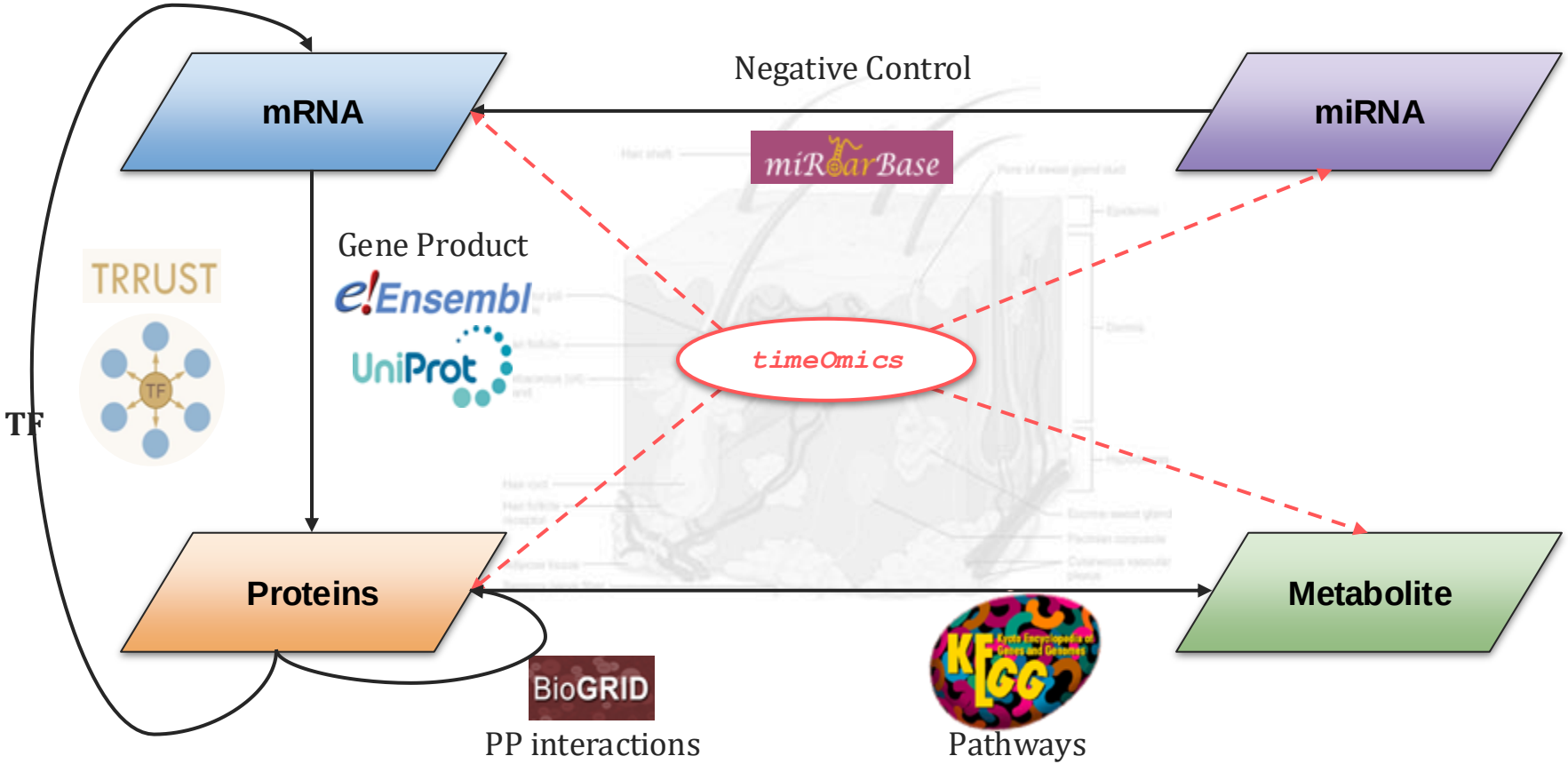


Multi-OMICS Data

timeOmics

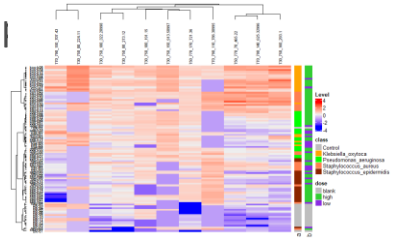
Knowledge

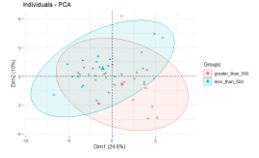
Example: epidermis reconstruction

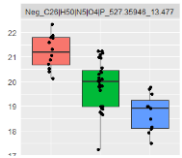


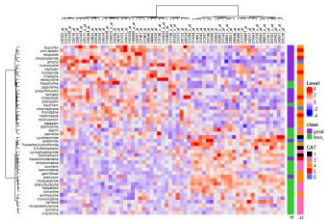
6. Signature visualization and characterization

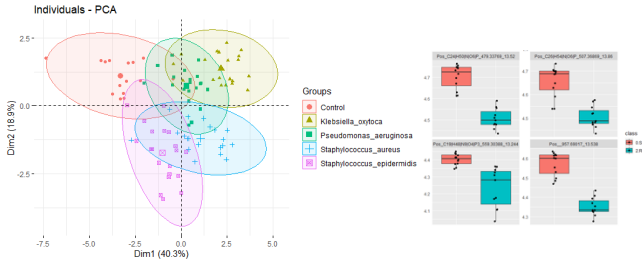
- R scripts to generate heatmaps, PCA and boxplots
- Characterization using enrichment analysis tools when transcri are involved
  - EnrichR
    - https://maayanlab.cloud/Enrichr/
  - ToppGene Suite
    - https://toppgene.cchmc.org/
  - g:Profiler
    - https://biit.cs.ut.ee/gprofiler/gost









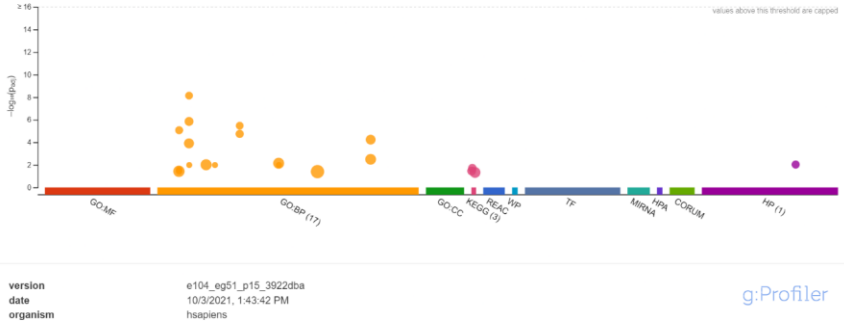


ToppGene

pValue Method: Probability density function

Feature	Correction	p-Value cutoff		Gene Limits	
All	FDR	0.05	1	≤ n ≤	2000
GO: Molecular Function	FDR	0.05	1	≤ n ≤	2000
GO: Biological Process	FDR	0.05	1	≤ n ≤	2000
GO: Cellular Component	FDR	0.05	1	≤ n ≤	2000
Human Phenotype	FDR	0.05	1	≤ n ≤	2000
Mouse Phenotype	FDR	0.05	1	≤ n ≤	2000
Domain	FDR	0.05	1	≤ n ≤	2000
Pathway	FDR	0.05	1	≤ n ≤	2000
Pubmed	FDR	0.05	1	≤ n ≤	2000
Interaction	FDR	0.05	1	≤ n ≤	2000
Cytoband	FDR	0.05	1	≤ n ≤	2000
Transcription Factor Binding Site	FDR	0.05	1	≤ n ≤	2000
Gene Family	FDR	0.05	1	≤ n ≤	2000
Coexpression	FDR	0.05	1	≤ n ≤	2000
Coexpression Atlas	FDR	0.05	1	≤ n ≤	2000
ToppCell Atlas	FDR	0.05	1	≤ n ≤	2000
Computational	FDR	0.05	1	≤ n ≤	2000
MicroRNA	FDR	0.05	1	≤ n ≤	2000
Drug	FDR	0.05	1	≤ n ≤	2000
Disease	FDR	0.05	1	≤ n ≤	2000

G:Profiler



- http://melgen.org/multi-omics-approach/

- Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in Genetics. 2017;8:84. doi:10.3389/fgene.2017.00084.

- Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome biology, 18(1), 83.

- Cavill, R., Jennen, D., Kleinjans, J., & Briedé, J. J. (2015). Transcriptomic and metabolomic data integration. Briefings in bioinformatics, 17(5), 891-901.

- Wanichthanarak, K., Fahrmann, J. F., & Grapov, D. (2015). Genomic, proteomic, and metabolomic data integration strategies. Biomarker insights, 10(Suppl 4), 1.

- https://www.frontiersin.org/research-topics/2280/multi-omic-data-integration

- http://mixomics.org

- L'analyse en Composante Principale, Sébastien Déjean - math.univ-toulouse.fr

Lê Cao, K. A., & Welham, Z. (2021). *Multivariate Data Integration Using R: Methods and Applications with the mixOmics Package*. Chapman and Hall/CRC.

Barabási, A. L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1987), 20120375.