# Introduction to statistics for omics data

**Guillemette Marot**
Univ. Lille, CHU Lille METRICS & Inria MODAL

17 November 2022

# Introduction

In mathematics, statistics refers to the discipline which concerns the collection, the analysis and interpretation of data. It can be applied in many fields : industry, society, economics, social, biology ....

BIOSTATISTICS : LARGE FIELD $\Rightarrow$ As an illustration, there exists more than 2000 R packages available in Bioconductor !

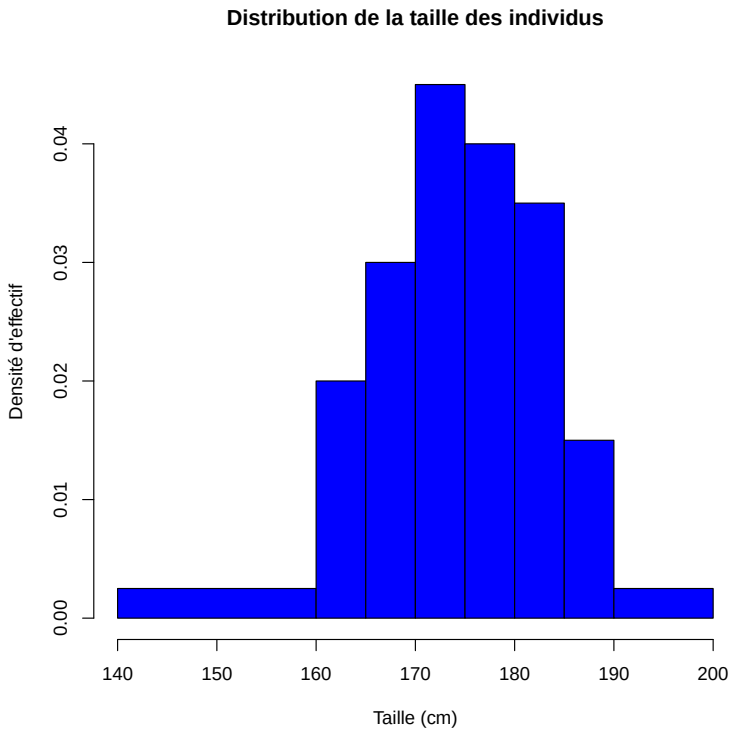
Main characteristic of omics data for statisticians : more variables (metabolites, peptides, proteins, ...) than individuals.

Need for appropriate statistics in addition or in extension of standard tools.

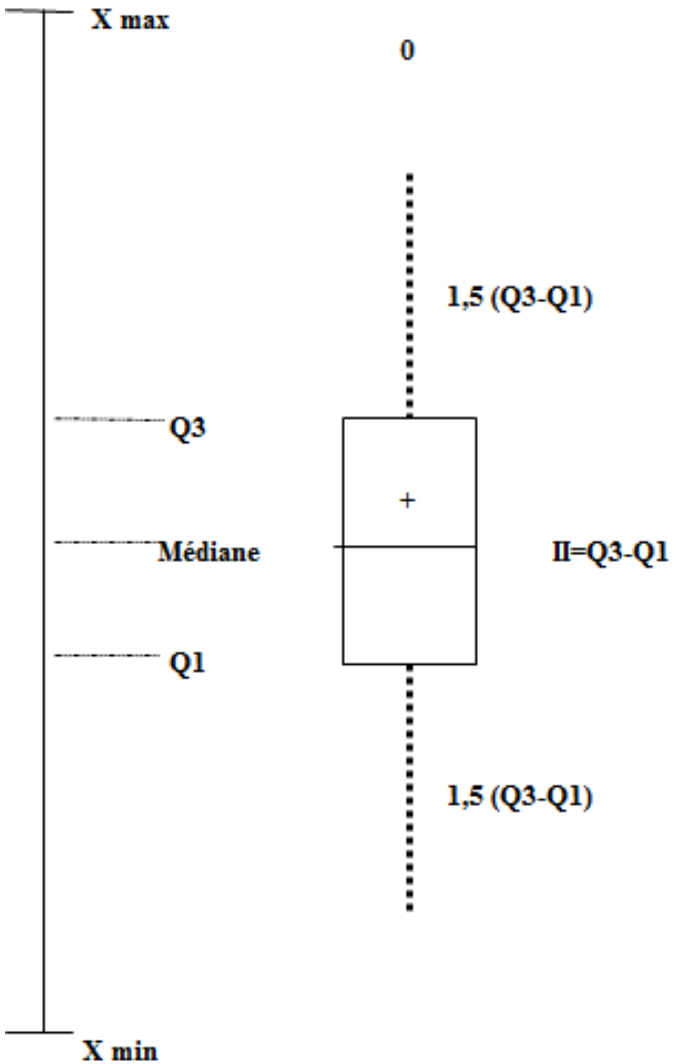
# Overview

# Statistics

- Descriptive statistics : mean, mode, variance, . . .




# Boxplot

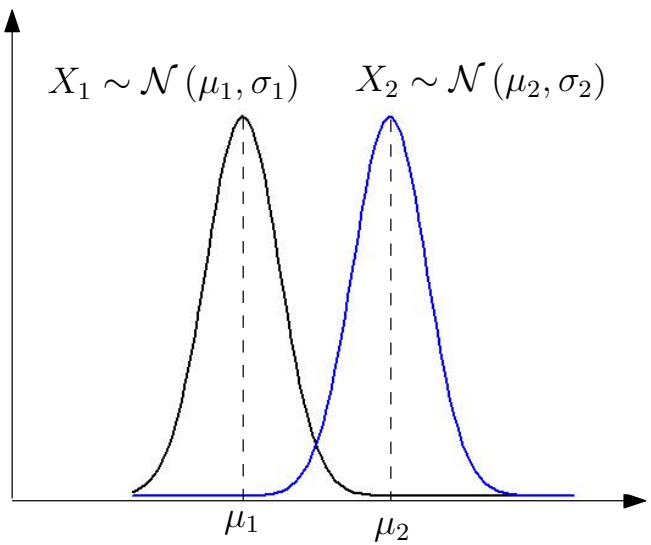



Key elements of a boxplot

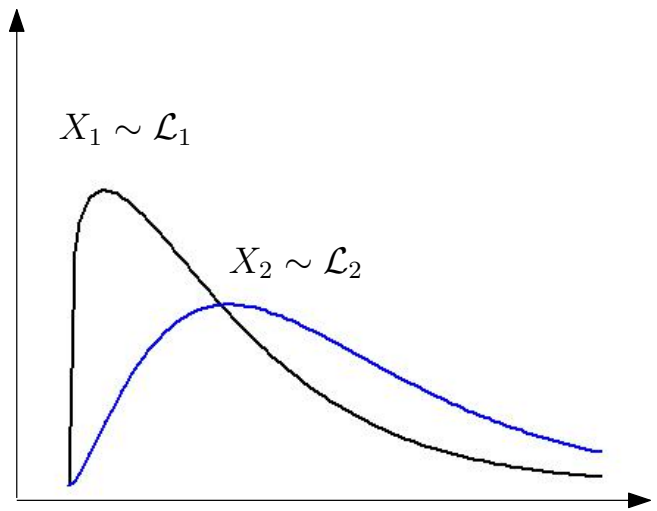
- box : rectangle whose length is between the $1^{st}$ (lower) and the $3^{rd}$ (upper) quartiles
  $\Rightarrow$ 50% of values belong to the interval.
- whiskers : vertical lines of length $1,5 * (Q3 - Q1)$, shortened to minimum and maximum of observations if there are no values outside the whiskers.
- a line within the rectangle : the median.

# Statistics

- Inferential statistics : draw meaningful conclusions about the entire population (confidence intervals, hypothesis testing)



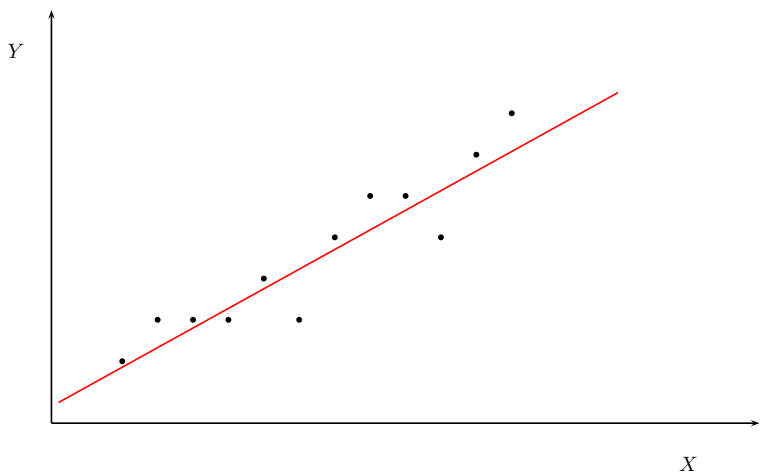

$\mathrm{H}_0 \ : \ \mu_1 = \mu_2$
$\mathrm{H}_1 \ : \ \mu_1 \neq \mu_2$

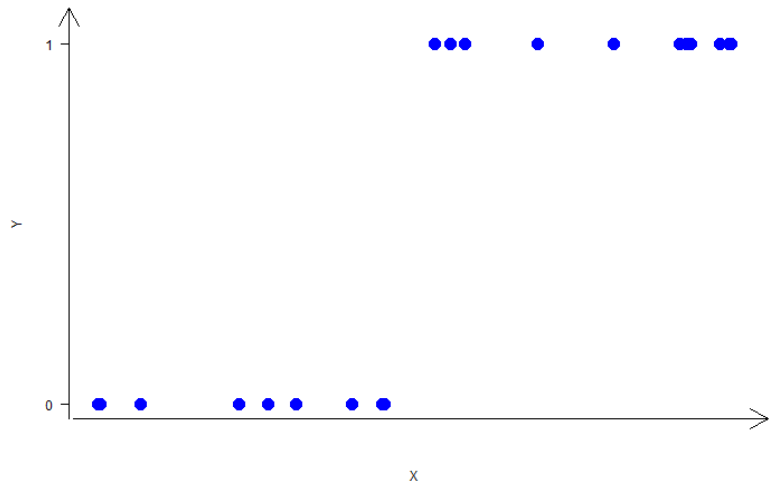



$\mathrm{H}_0 \ : \ \mathcal{L}_1 = \mathcal{L}_2$
$\mathrm{H}_1 \ : \ \mathcal{L}_1 \neq \mathcal{L}_2$

# Statistics

- Modeling relationships (linear regression, logistic regression) or survival (cox regression) and predict a response.



# Overview

# Normalisation

## Definition

Normalization is a process designed to identify and correct **technical biases** removing the least possible biological signal. This step is technology and platform-dependant.

## Within-sample normalization

Normalization enabling comparisons of measures from a same sample.

## Between-sample normalization

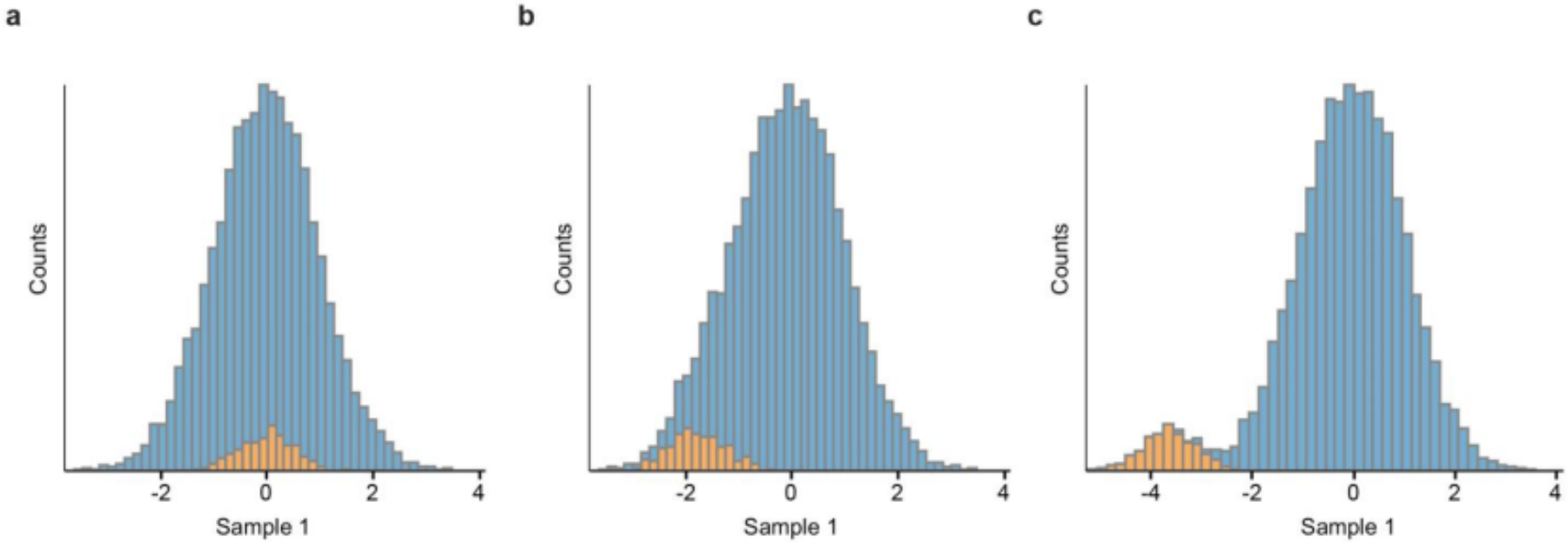
Normalization enabling comparisons of measures from different samples.

# Normalization

Common normalization : Log transform the data

(Callister et al., 2007)
*Even though the relationship between peptide abundance and detector measurement is expected to be linear, log transformation has several advantages similar to those highlighted for microarray data. Using such a transform converts the distribution of ratios of abundance values of peptides into a more symmetric, almost normal distribution. This allows the use of several robust normalization techniques that have been developed for such data. Also, a log transform reduces the leverage of a low number of highly abundant species on the regression analysis used by these robust techniques.*

# Imputation of missing values




Source : Tyanova et al., Nature Methods, 2016

(a) No down-shift a do not simulate low abundant missing values.
(b) Down-shift of 1.8 and distribution width of 0.5 simulate the assumption of low abundant proteins giving rise to missing values.
(c) Down-shift of 3.6 results in an undesirable bi-modal distribution.

# Overview

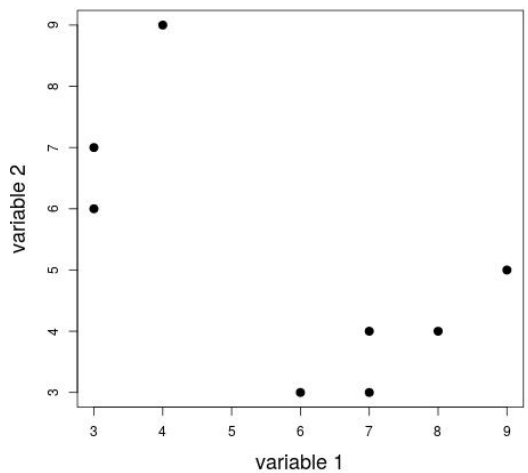
# Dimensionality reduction

Problem : $n$ individuals, $p$ quantitative variables (e.g. peptides, proteins, metabolites, mRNA, ...)

$$X = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \\ \dots & \dots & \dots \\ x_{p1} & \dots & x_{pn} \end{bmatrix}$$

$x_{ij}$ : value of variable $j$ for individual $i$.

Possibility to visualize pair-wise relations by scatter plots :




When p is large, this is not efficient !

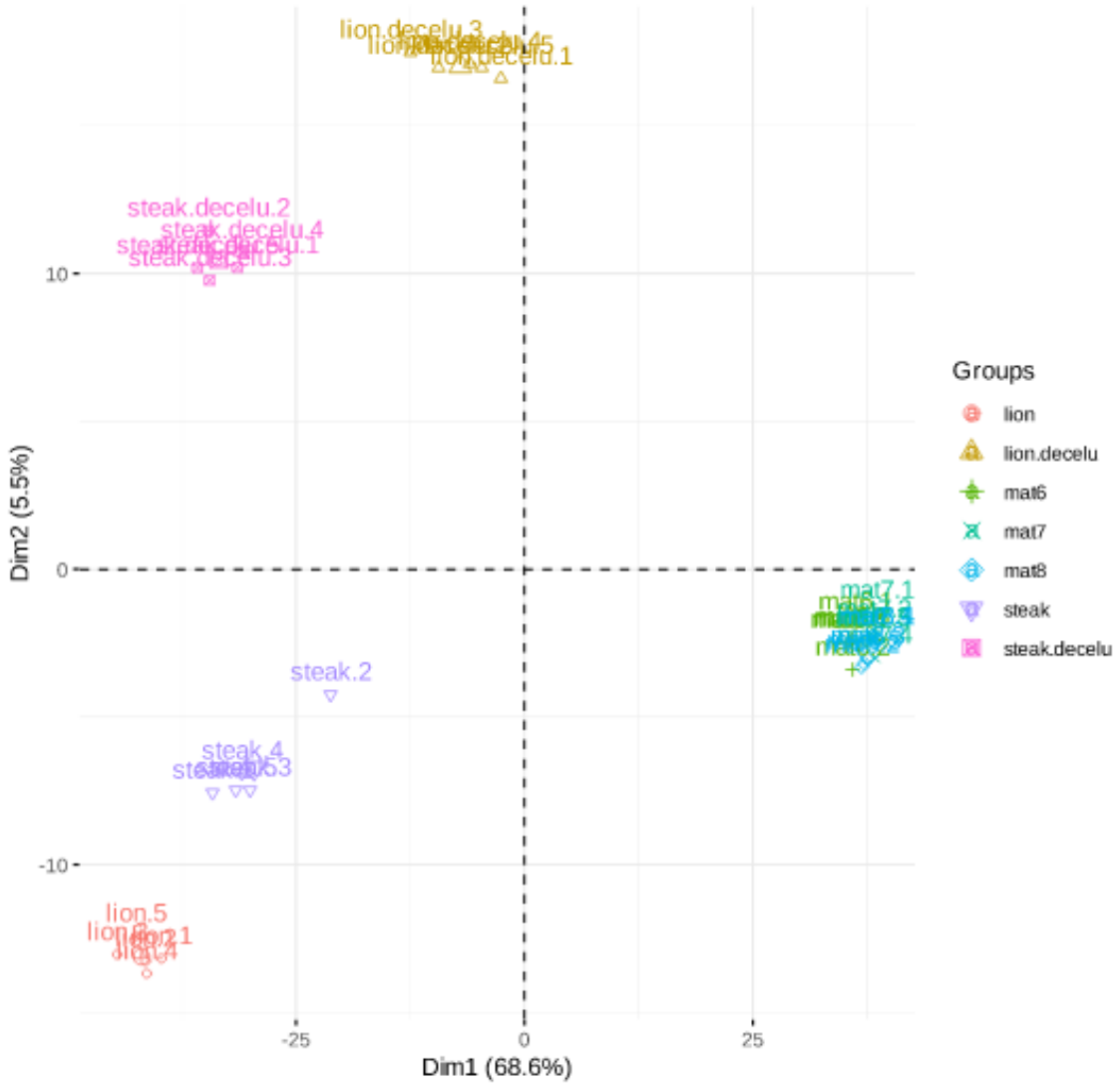
# Principal components analysis

## Principal component analysis (PCA) :

Main goal : explore the structure of the dataset to better understand the proximity between samples and detect possible problems $\rightarrow$ often used as a quality control step

- synthetize information and visualize points in a space of reduced dimension
- describe links between variables and which ones explain most variability
- highlight homogeneous subgroups
- detect aberrant individuals

# Principal components analysis



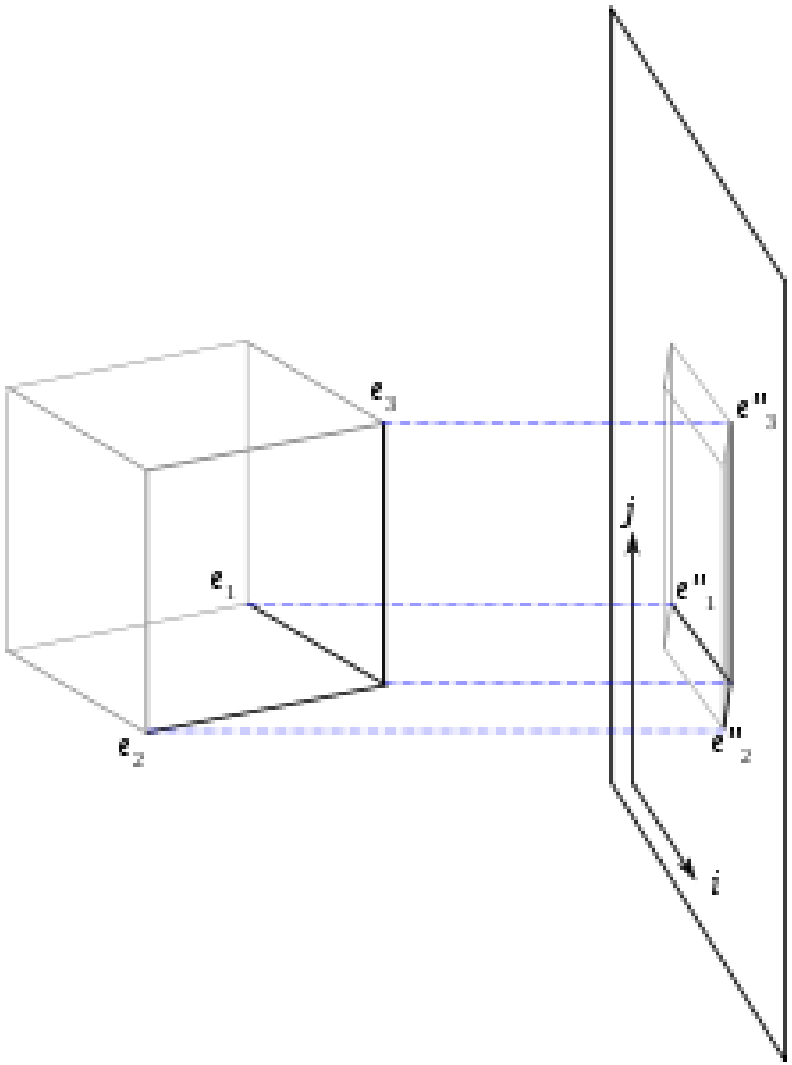
# Principal components analysis

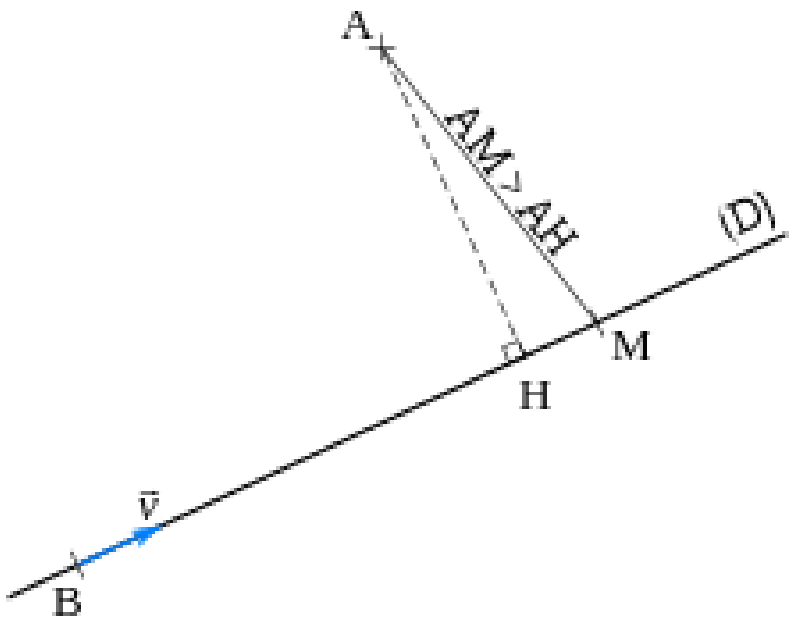
**Principle :**

Find axes on which one can project points to obtain a space of reduced dimension comprehensible by the eye.

Projection is a distorting operation $\Rightarrow$ we begin by looking for an axis on which the cloud of points is distorting the less possible during the projection.

# Principal components analysis



PCA uses a criterion based on variance to build new axes, also called components, in order to preserve variability.

A pre-requisite to apply PCA is to make the data homoscedastic : the variance must be independent of the mean.

PCA uses a criterion based on variance to build new axes, also called components, in order to preserve variability.

A pre-requisite to apply PCA is to make the data homoscedastic : the variance must be independent of the mean.

New components are linear combinations of the initial variables. If you want to force the method to select only a few variables, you need to use variations of PCA, such as sparse PCA, which often includes a Lasso penalty (Tibshirani, 1996).

# Overview

# Studies with small sample sizes . . .

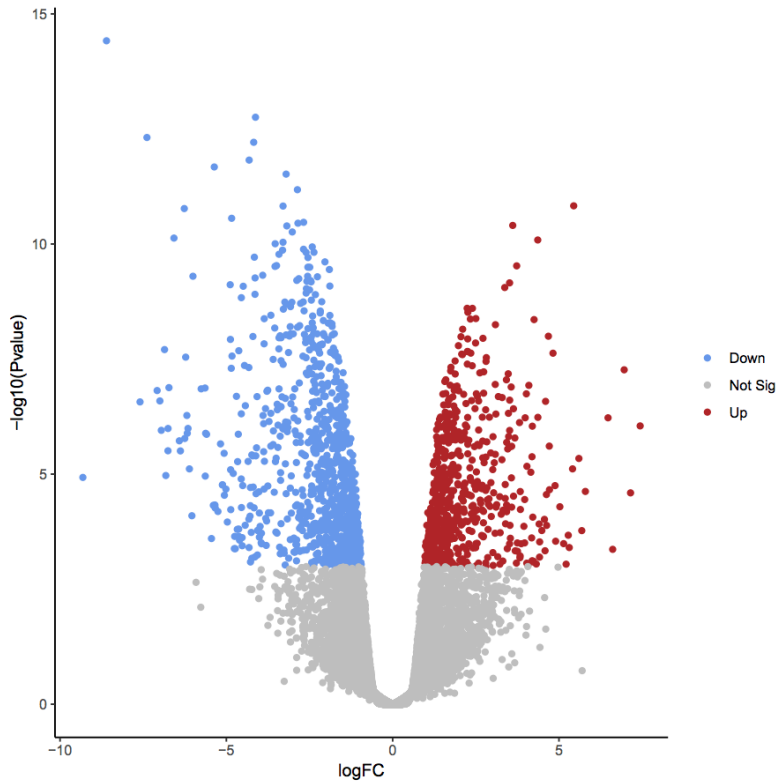
Historically, very few individuals in omics studies, due to the cost of experiments.

Exploratory studies are performed to detect differentially expressed proteins (proteins for which the observed difference between two conditions is statistically significant, that is to say higher than some natural random variation).

Volcano plots are used to find a compromise between statistical significance and importance and adapt the definition of differentially expressed proteins.

# Volcano plot

"A protein is declared differentially expressed (DE) if the observed difference between two conditions is statistically significant at 5% and the fold change is higher than 2"



# Volcano plot

The p-value measures the agreement between a null hypothesis (e.g. the protein is not differentially expressed) and the obtained result.

The fold change is a ratio describing how much a quantity changes : if condition A measures 50 and condition B measures 100, fold change = 100/50 =2 and measure B is twice higher than measure A.

The log fold change is equal to the mean of normalised values in condition 1 - mean of normalised values in condition 2 (log B/A= log B - log A)

Question : Why not only using the fold change or log fold change to find differentially expressed proteins ?

# Preamble : Interpretation - Statistical significance and practical importance

- Fold change does not take the variance of the samples into account. Problematic since variability in omic data is partially marker-specific.
- The difference between 102 and 100 is the same as between 4 and 2 but does not seem to have the same importance, regarding the baseline value.

Example of test statistic, which takes into account the variance of the samples :

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma}$$

# Preamble : Interpretation - Statistical significance and practical importance

- Practical importance and statistical significance (detectability) have little to do with each other.
- An effect can be important, but undetectable (statistically insignificant) because the data are few, irrelevant, or of poor quality.
- An effect can be statistically significant (detectable) even if it is small and unimportant, if the data are many and of high quality.

# Estimating the variance : the key question

## Problem

Estimate a reliable variance from a very small number of replicates (sometimes 3 or 5)

## Why using sophisticated approaches ?

- protein-specific tests $\Rightarrow$ lack of sensitivity (proportion of true positives among positives) due to the lack of information
- common dispersion parameter for all tests $\Rightarrow$ many false positives

Example : empirical bayesian approaches = compromise between protein-specific and common dispersion parameter estimation

# Empirical bayesian approaches

## Principles

- Bayes theorem : $P(A/B) = P(B/A)P(A)$
- "empirical" $\Rightarrow$ priors from the observed data

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}} + b(\widehat{\boldsymbol{\theta}_{\boldsymbol{g}}} - \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}})$$

with $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$ = shrinkage estimator
$\widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}$ = estimator of the mean population
$\widehat{\boldsymbol{\theta}_{\boldsymbol{g}}}$ = usual empirical estimator protein by protein
$b$ =shrinkage factor

$b = 1 \Rightarrow \tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{g}}}$
$b = 0 \Rightarrow \tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} = \widehat{\boldsymbol{\theta}_{\boldsymbol{c}}}$

# Multiple Testing

Empirical bayesian approaches enable to obtain reliable results from small sample size, due to a better estimation of parameters. This does not solve the problem of multiple testing.

## Problem

Let assume all the G proteins are not DE. Each test is performed at $\alpha$ level
Ex : $G = 10000$ proteins and $\alpha = 0.05 \rightarrow E(FP) = 500$ proteins.

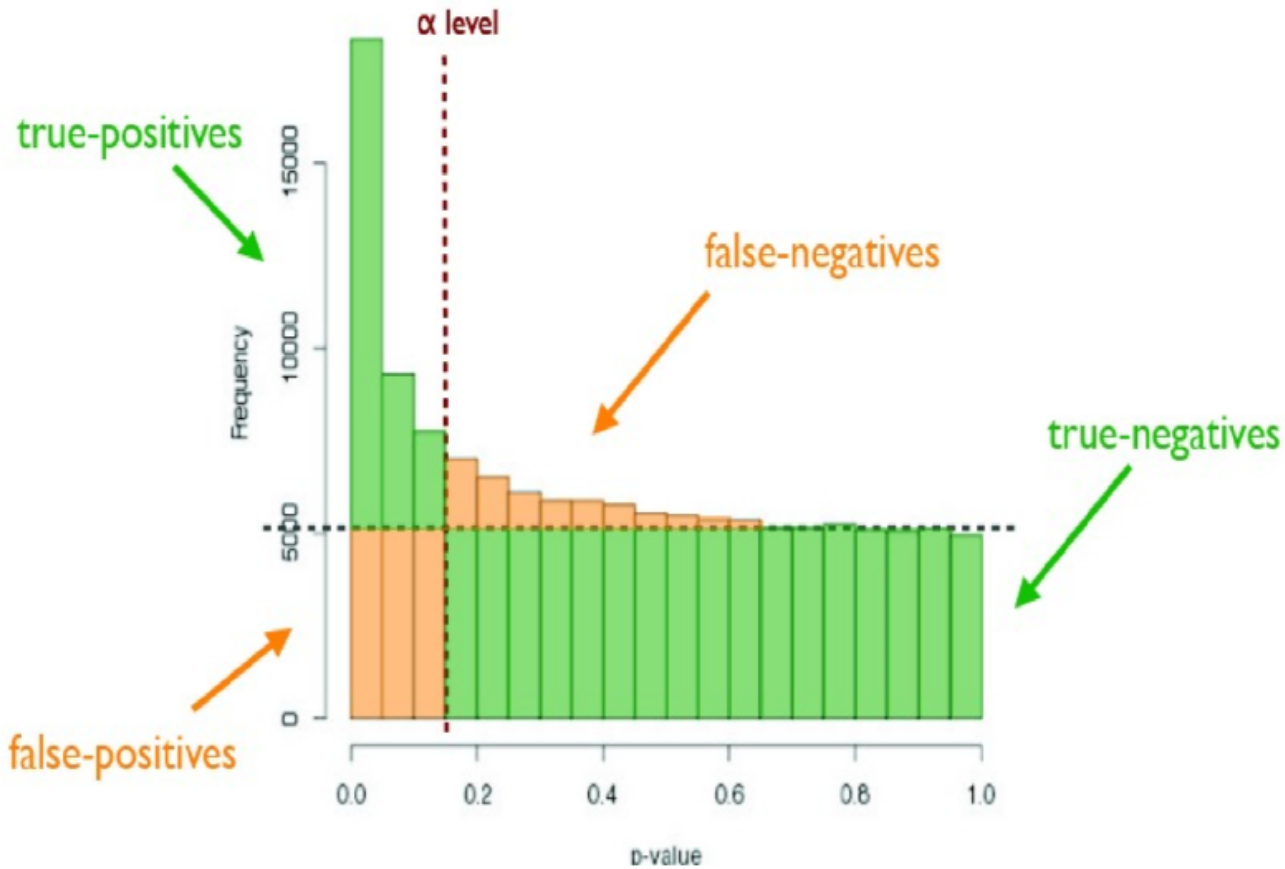
(False positive (FP) : A non differentially expressed (DE) protein which is declared DE)

# Simultaneous tests of $G$ null hypotheses

| Reality | Declared non diff. exp. | Declared diff. exp. |
|---|---|---|
| $G_0$ non DE proteins | **True Negatives** (*TN*) | **False Positives** (*FP*) |
| $G_1$ DE proteins | **False Negatives** (*FN*) | **True Positives** (*TP*) |
| $G$ Proteins | $N$ Negatives | $P$ Positives |

**Aim :** minimize *FP* and *FN*.

# Standard assumption for p-value distribution




Source : M. Guedj, Pharnext

# The Family Wise Error Rate (FWER)

## Definition

Probability of having at least one Type I error (false positive), of declaring DE at least one non DE protein.

$$FWER = \mathbb{P}(FP \geq 1)$$

## The Bonferroni procedure

Either each test is realized at $\alpha = \alpha^* / G$ level
or use of adjusted pvalue $pBonf_i = min(1, p_i * G)$ and FWER $\leq \alpha^*$.
*For* $G = 2000$ *and* $\alpha^* = 0.05$ *;* $\alpha = 2.5.10^{-5}$.

**Easy but conservative and not powerful.**

# The False Discovery Rate (FDR)

Idea : Do not control the error rate but the proportion of error
$\Rightarrow$ less conservative than control of the FWER.

## Definition

The false discovery rate of [Benjamini and Hochberg, 1995] is the expected proportion of Type I errors among the rejected hypotheses

$$\text{FDR} = \mathbb{E}(FP/P) \text{ if } P > 0 \text{ and } 0 \text{ if } P = 0$$

## Prop

$$\text{FDR} \leq \text{FWER}$$

# Multiple testing : key points

- Important to control for multiple tests
- FDR or FWER depends on the cost associated to FN and FP

**Controlling the FWER :**

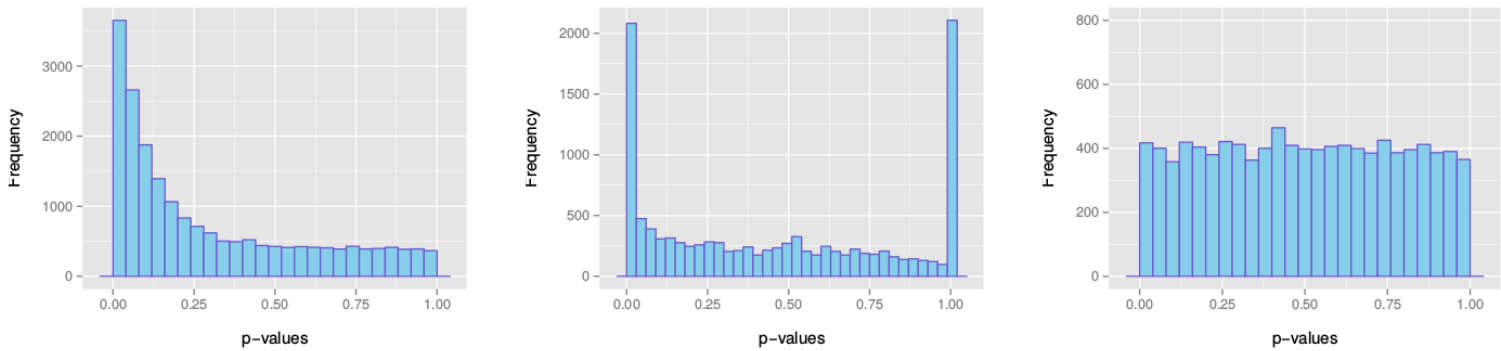
Having a great confidence on the DE elements (strong control). Accepting to not detect some elements (lack of sensitivity $\Leftrightarrow$ a few DE elements)

**Controlling the FDR :**

Accepting a proportion of FP among DE elements. Very interesting in exploratory study.

# p-values histograms for diagnosis

Examples of expected overall distribution



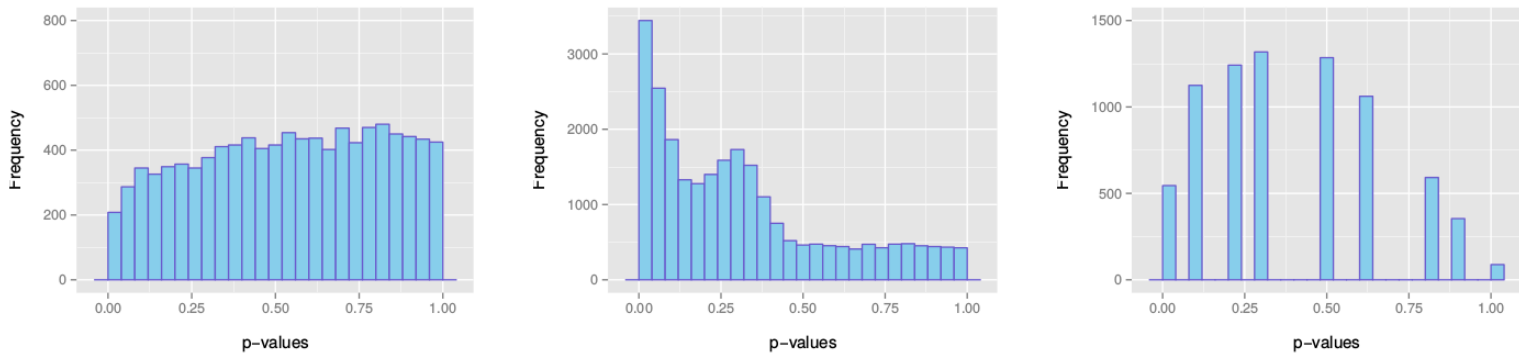

(a) : the most desirable shape

(b) : very low counts usually have large p-values

(c) : do not expect positive tests after correction

# p-values histograms for diagnosis

Examples of not expected overall distribution




(a) : indicates a batch effect (confounding hidden variables)

(b) : the test statistics may be inappropriate (due to strong correlation structure for instance)

(c) : discrete distribution of p-values : unexpected

# Overview

# Conclusions

In omics data analysis,

- Empirical bayesian approaches are used when very few replicates are available $\Rightarrow$ Exploratory analysis to be confirmed by biological experiments
- The need for 'sophisticated' methods decreases when the number of replicates increases.
- Score building can be performed by penalized regression frameworks, it necessits more replicates even if statistical learning algorithms are used.
- A very small methodology paragraph described in a biology paper sometimes hides several months of work.
- Interdisciplinarity is necessary to reach high level journals.

# Conclusions

Do not forget :

- Obtaining a result using a statistical procedure does not mean that this result is reliable. If you do not know the assumptions behind, please be careful with interpretation or ask an expert to help you.
- Most of the time, not a unique solution ; statisticians do not know all statistical procedures developped (example of the Bioconductor project : more than 2000 R packages) but have competences to understand them.
- "All models are wrong but some are useful" (G. Box, 1978)

## Want to go further in the analysis of your data ?

- Contact : bilille@univ-lille.fr (Guillemette Marot, Pierre Pericard, Jimmy Vandel)
- Site web : `https://wikis.univ-lille.fr/bilille/`