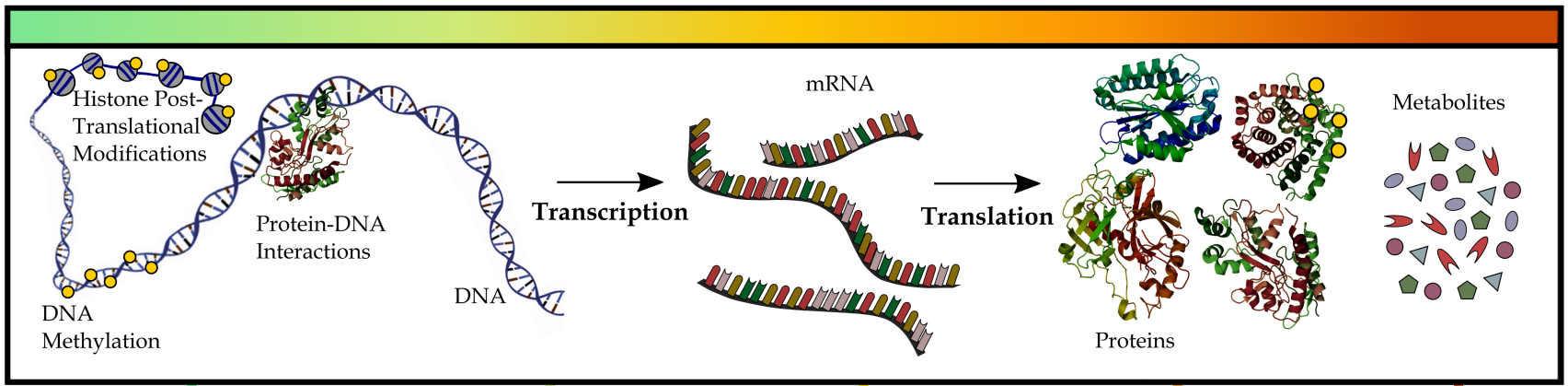
Introduction to statistical analysis of omics data

Pr. Guillemette Marot (UFR3S - Médecine,
Univ. Lille, CHU Lille METRICS & Inria MODAL)

18-19 avril 2024

Omics

Who works on such data ? Please give your name, lab, and kind of omic data you are interested in.



Epigenomics

Measures of DNA modifications including

- Epigenome-wide Association Studies (EWAS)
- Chromatin Immunoprecipitation Sequencing (ChIP-Seq)

Genomics

Measures of DNA sequence variation including

- Genome-wide Association Studies (GWAS)
- Whole-Exome Sequencing (WES)
- Whole-Genome Sequencing (WGS)

Transcriptomics

Measures of transcript expression including

- Gene Expression Microarrays
- RNA-Sequencing (RNA-Seq)

Proteomics

Measures of protein levels and their chemical modifications including

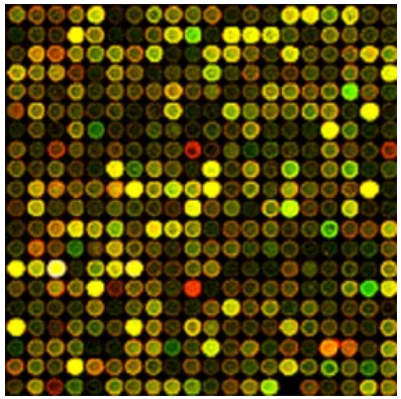
- Quantitative proteomics
- Histone PTMs
- Phosphoproteomics

Metabolomics

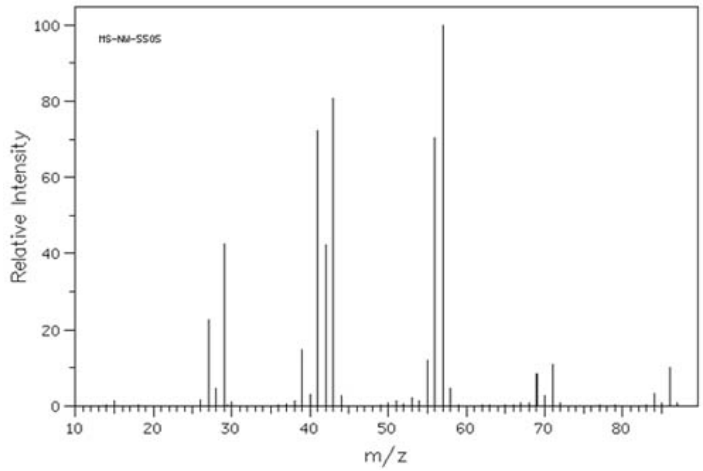
Measures of endogenous and exogenous chemicals in biological systems

Examples of high-throughput experiments

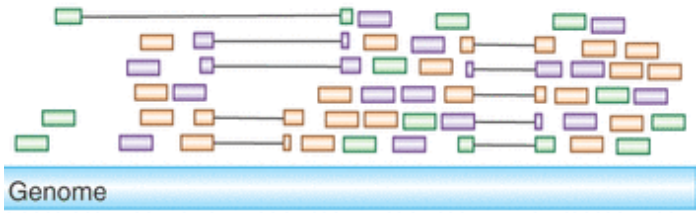
Microarrays



Mass spectrometry

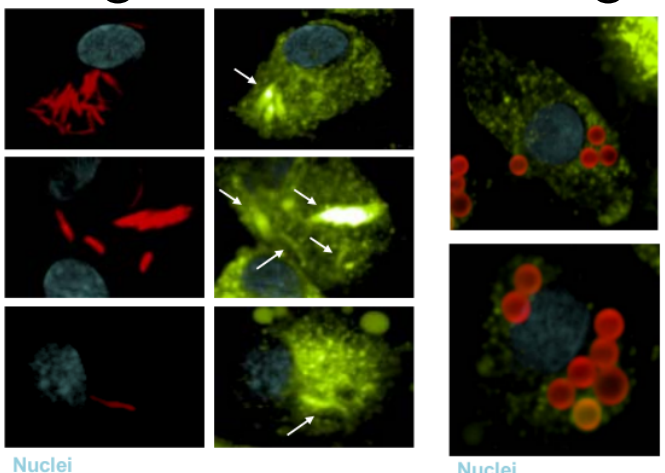


High throughput sequencing



©Haas and Zody, Nature Biotechnology

High content screening



Extract of the program of the lecture

Context : High-throughput experiments using technologies such as microarrays, next-generation sequencing or mass spectrometry usually generate data with much more variables than individuals. Analysis of these data requires either dimensionality reduction methods or special techniques which improve variance estimation in statistical tests.

Requirements : Basic concepts in statistics (e.g. mean, variance, p-value) and basic concepts in molecular biology (e.g. transcription and translation).

Note : it is not necessary to know in detail high-throughput techniques as questions relative to normalisation of each technology will not be detailed in this course. Only statistical techniques common to several omic data analyses will be presented.

Objectives

Objectives :

- Share a common langage with statisticians or bioinformaticiens who are specialists of omics data analysis.
- Understand the main steps of a differential analysis of -omics data and perform a differential analysis on simple use cases.

Comments :

- Statistical analysis of omic data is a field included in the big field of bioinformatics (english comprehension of the word), which covers much more than differential analysis
- The R software enables to study a lot of various applications with omic data, especially using Bioconductor R packages
  $\Rightarrow$ not very user friendly, but essential to master for those who commonly analyse omics data

Introduction

Caution : Simple use cases rely sometimes on assumptions which will not be checked in your own experiment.

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 $\Rightarrow$ 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)

Introduction

All models are wrong (Box, JASA, 1976)

- *Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration.*
- *Since all models are wrong the scientist must be alert to what is importantly wrong.*

GM opinions :

- Be careful with over-fitting, always validate on other patients or samples a model trained from a training set.
- Do not hesitate to call an expert to help with interpretation and evaluate how wrong the model is
- While statistics is very helpful to take decision, especially statistical learning in artificial intelligence, human people must remain the final decision-makers.

Introduction

Some are useful *(Box, 1978)*

- *cunningly chosen parsimonious models often do provide remarkably useful approximations.*
- *any model is at best a useful fiction, there never was, or ever will be, an exactly normal distribution or an exact linear relationship. Nevertheless, enormous progress has been made by entertaining such fictions and using them as approximations.*
- *the question you need to ask is not "Is the model true ?" (it never is) but "Is the model good enough for this particular application ?"*

GM opinion :
Statistics offers great possibilities, the field is not limited to calculating p-values. Do not hesitate to learn more and more.

Content

Content :

- Presentation of main steps of a statistical analysis involving data from high-throughput experiments
- Dimensionality reduction
- Differential analysis of omics data, including multiple testing
- Practice

A gene is declared differentially expressed if the observed difference between two conditions is statistically significant, that is to say higher than some natural random variation.

Plan

1. Main steps
2. Dimensionality reduction
3. Differential analysis of omics data
4. Multiple testing
5. Gene Set Enrichment Analysis
6. Conclusions

Main steps

Key steps for statisticians :

- experimental design
- normalization
- choice of the type of analysis : differential analysis, score building, network inference, ...

In the case of differential analysis :

- choice of the appropriate test statistic
- multiple testing

Experimental design

To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of (Ronald A. Fisher, Indian statistical congress, 1938, vol. 4, p 17).

While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment (Kathleen Kerr, Inserm workshop 145, 2003)

Experimental design

- Anticipate, identify all factors of variation and collect metadata from experiment
- Adapt Fisher's principles (1935) : randomization and blocking
  AVOID CONFUSION between the biological variability of interest and a biological or technical source of variation

Biological vs technical replicate

Biological replicate : Repetition of the same experimental protocol but independent data acquisition (several samples).
Technical replicate : Same biological material but independent replications of the technical steps (several extracts from the same sample).

Experimental design

Find genes that are differentially expressed between a normal skin and a damaged skin on mouse

Sample	Condition	RNA extraction date
S1	control	July 12th, 2016
S2	control	July 12th, 2016
S3	control	July 12th, 2016
S4	wound	July 20th, 2016
S5	wound	July 20th, 2016
S6	wound	July 20th, 2016

Confusion between skin status and RNA extraction date : comparing healthy and damaged skin is comparing RNAs extracted July 12th and 20th

Experimental design

Find genes that are differentially expressed between a normal skin and a damaged skin on mouse

Sample	Condition	RNA extraction date
S1	control	July 12th, 2016
S2	control	July 20th, 2016
S3	control	July 25th, 2016
S4	wound	July 12th, 2016
S5	wound	July 20th, 2016
S6	wound	July 25th, 2016

One solution : the day effect is evenly distributed across conditions.

Experimental design

Find genes that are differentially expressed between a normal skin and a damaged skin on mouse

Sample	Condition	RNA extraction date	mouse
S1	control	July 12th, 2016	m1
S2	control	July 20th, 2016	m2
S3	control	July 25th, 2016	m3
S4	wound	July 12th, 2016	m1
S5	wound	July 20th, 2016	m2
S6	wound	July 25th, 2016	m3

One solution : the day effect is evenly distributed across conditions.
In case of paired data the pairing may be confounded with the batch effect.
These effects are NOT confounded with the biological effect of interest.

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

Examples of sources of within-sample biases

- fluorochrome in two-color microarrays
- GC content
- gene length in high throughput sequencing

Examples of sources of between-sample biases

- Depth in high throughput sequencing(total number of sequenced and mapped reads)
- Sampling bias in library construction
- Protein degradation
- Presence of majority fragments
- ...

Normalization

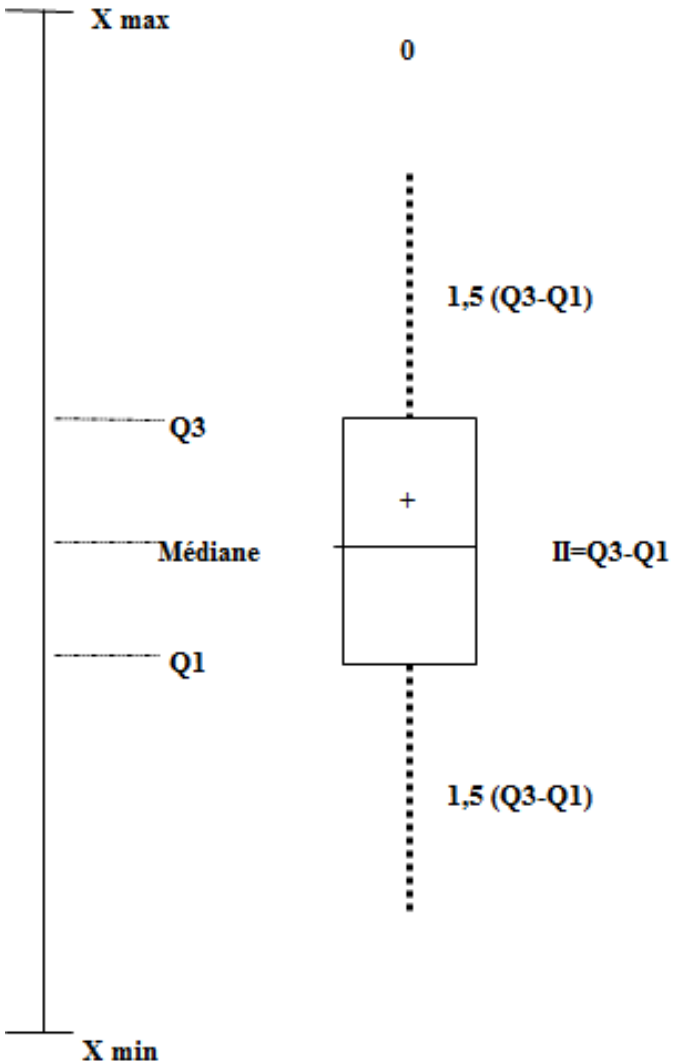
Normalization does not only depend on the technology but also on the statistical question raised.

Example 1 : in a differential analysis context where the analysis is performed gene-by-gene, it is not necessary to normalize for gene length.

Example 2 : while dimension reduction techniques based on distances need to standardise the data (center and reduce), reduction is clearly unappropriate when performing differential analyses based on special modeling of variances.

Boxplots are very useful to check the quality of a between-sample normalization.

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.

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.

Normalization

Caution : Log-scaling might not be sufficient to obtain normal distributions.

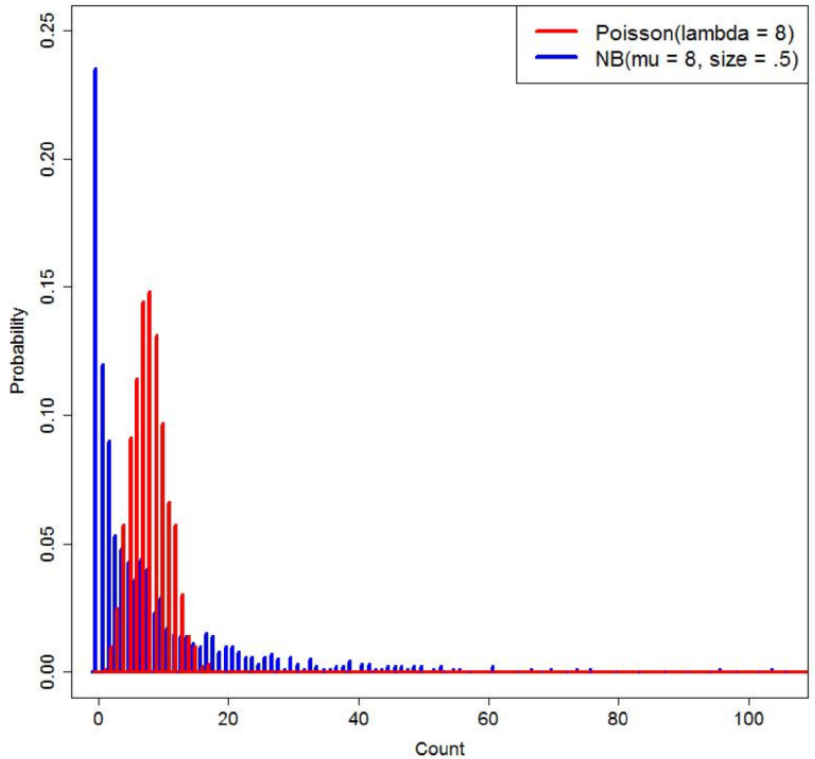
Example : In high throughput sequencing, techniques developed for microarrays can not be directly applied as counts and intensities do not have the same natural distribution (poisson or negative binomial vs normal).

When many replicates are available, the distribution of the mean tends to follow a normal distribution. However, it is frequent in studies with omics data not to have enough replicates to use this approximation ...

Negative Binomial Models

A supplementary dispersion parameter ϕ to model the variance

Poisson vs Negative Binomial models



Technical variability is the main source of variability in low counts, whereas biological variability is dominant in high counts

Normalization

Another common normalization (but not necessary always the best)

Quantile normalization :

Hypothesis : the distribution of peptide abundances or gene expression in different samples is similar. This expectation can be accounted for by adjusting observed distributions.

Exercise : Calculate the means row by row and for each column, replace the values by the rank in the column.

3	9	3	8	7	7	6	4
6	5	7	4	3	4	3	9
9	4	8	3	8	8	9	6
4	8	4	5	4	9	5	7
7	6	6	7	9	3	4	8

Normalization

Means : 5.9 ; 5.1 ; 6.9 ; 5.8 ; 6.2

Ranks :

1	5	1	5	3	3	4	1
3	2	4	2	1	2	1	5
5	1	5	1	4	4	5	2
2	4	2	3	2	5	3	3
4	3	3	4	5	1	2	4

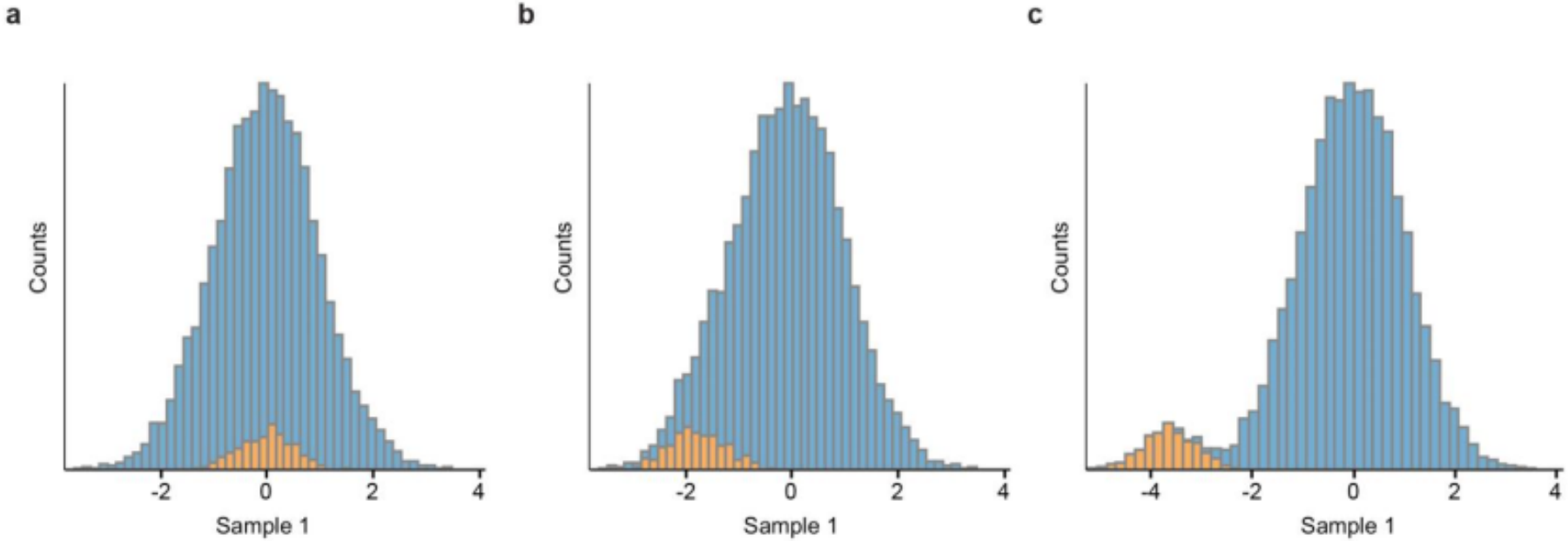
Replace, in each column, the rank k by the kth sorted mean value. You will thus obtain the quantile-normalized table.

Normalization

5.1	6.9	5.1	6.9	5.9	5.9	6.2	5.1
5.9	5.8	6.2	5.8	5.1	5.8	5.1	6.9
6.9	5.1	6.9	5.1	6.2	6.2	6.9	5.8
5.8	6.2	5.8	5.9	5.8	6.9	5.9	5.9
6.2	5.9	5.9	6.2	6.9	5.1	5.8	6.2

Caution : this normalization provides very good boxplots but can heavily change the measures. It can also favor null variances on rows. Be careful when using it, if not recommanded by the platform which generated the data.

Imputation of missing values in Perseus



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.

Plan

1 Main steps

2 Dimensionality reduction

3 Differential analysis of omics data

4 Multiple testing

5 Gene Set Enrichment Analysis

6 Conclusions

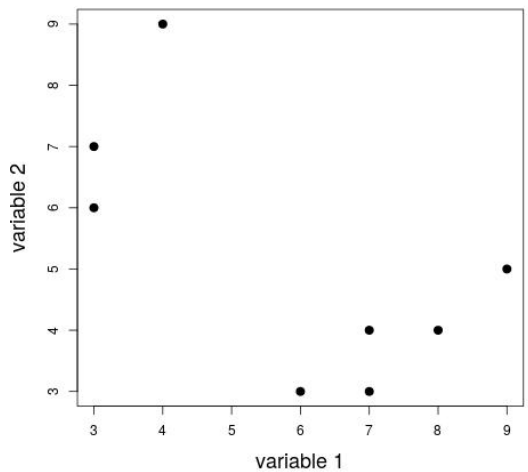
Dimensionality reduction

Problem : n individuals, p quantitative variables (e.g. genes, peptides, proteins, siRNA, ...)

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

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 !

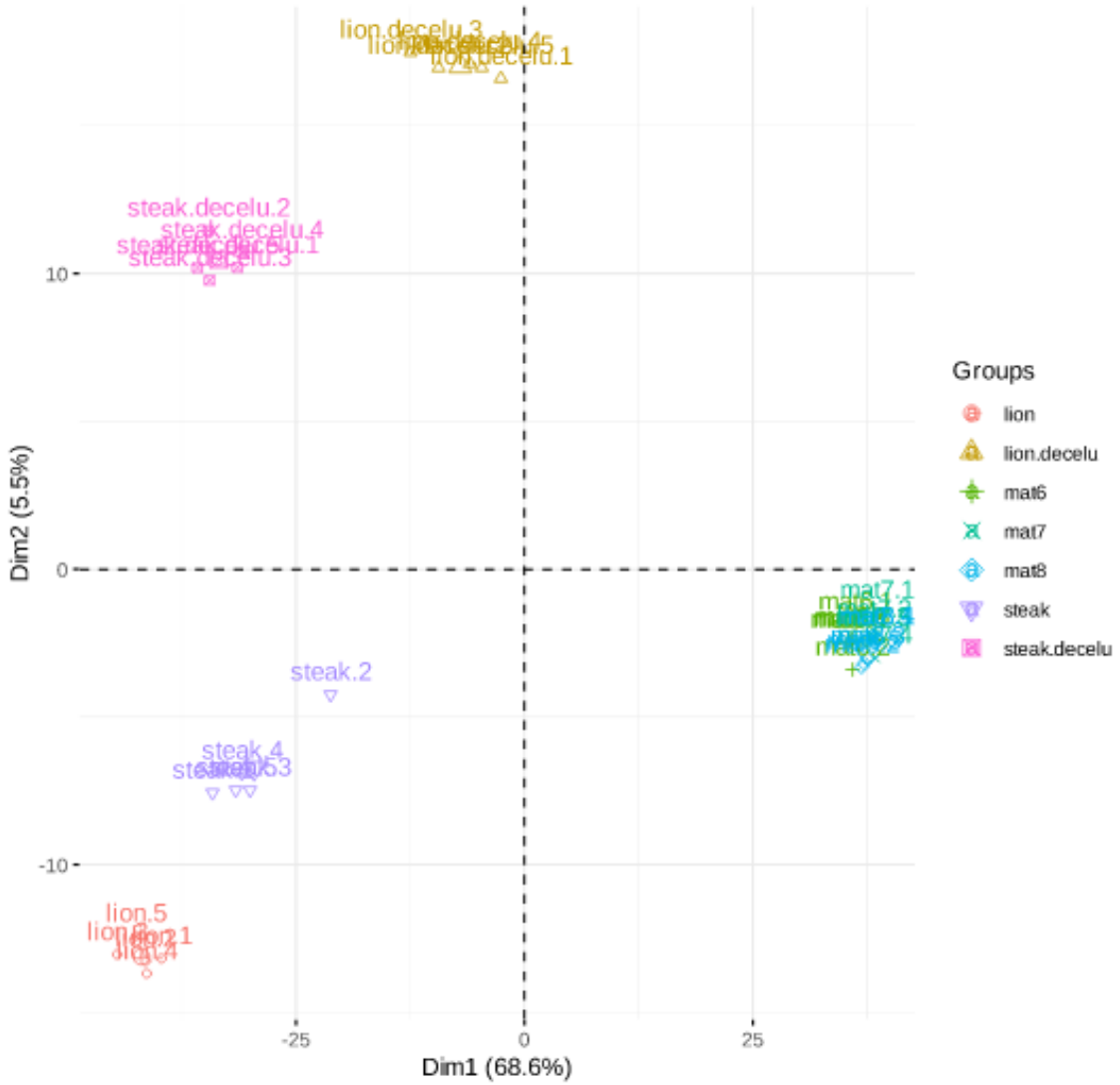
Principal components analysis

Principal components 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

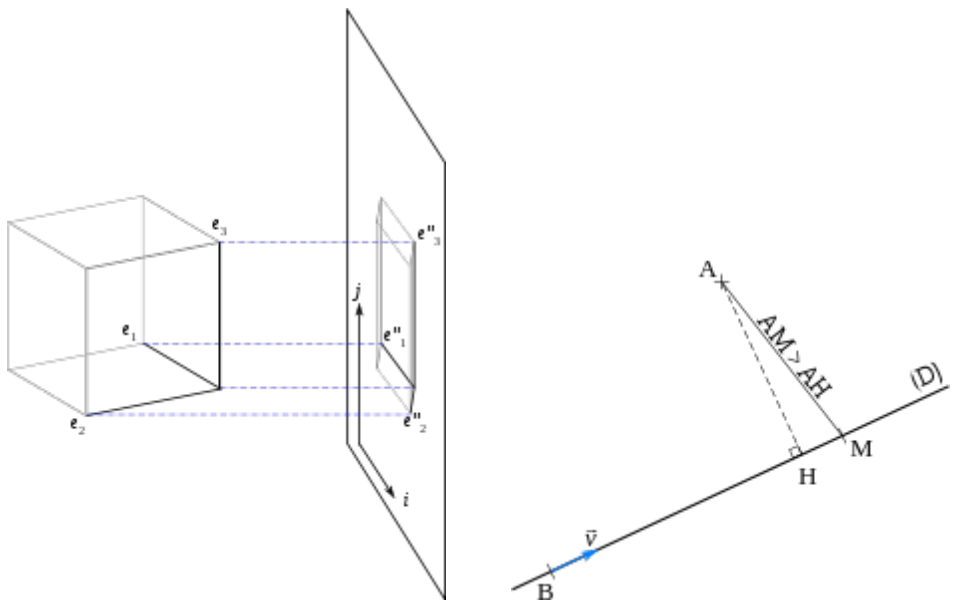


Principal components analysis

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 variance 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).

Plan

1 Main steps

2 Dimensionality reduction

3 Differential analysis of omics data

4 Multiple testing

5 Gene Set Enrichment Analysis

6 Conclusions

Preamble : Interpretation - Statistical significance and practical importance

A gene is declared differentially expressed (DE) if the observed difference between two conditions is statistically significant, that is to say higher than some natural random variation.

Fold change : measure describing how much a quantity changes. Various definitions (see Wikipedia, ipfs.io). In this course : ratio between measurements. If condition A measures 50 and condition B measures 100, fold change = 100/50 =2 and measure B is twice higher than measure A.

Log fold change : 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 genes ?

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 :

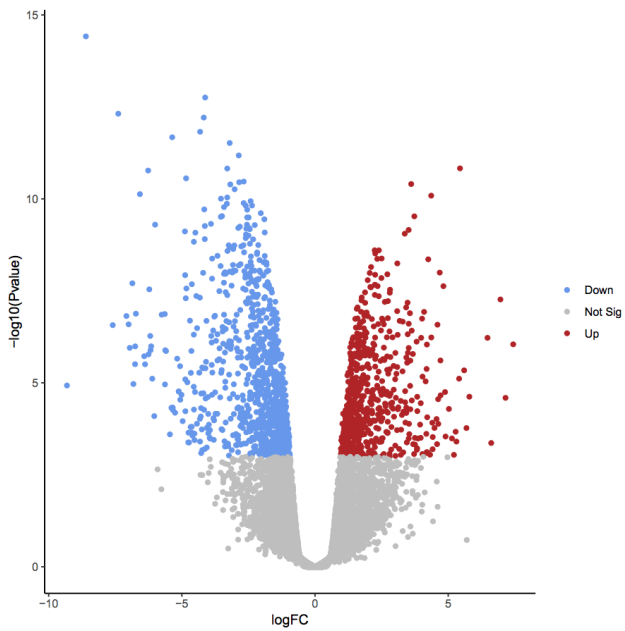
$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma}$$

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.

Volcano plot

Compromise between statistical significance and importance.
One can adapt the definition of differentially expressed by saying for exemple "A gene 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"



Statistical test

- State the null and the alternative hypotheses
  H_0= {the mean expression of the gene (or protein) is identical between the two conditions}

  H_1= {the mean expression of the gene (or protein) is different between the two conditions}
- Consider the statistical assumptions (e.g. independence) and distributions (e.g. normal, negative binomial, ...)
- Calculate the appropriate test statistic T
- Derive the distribution of the test statistic *under the null hypothesis* from the assumptions.
- Select a significance level (α), a probability threshold below which the null hypothesis will be rejected.

Remark : H_0 is always preferred. No sufficient proof $\rightarrow$ no rejection. When we can not reject $H0$, this does not mean that $H0$ is true.

Statistical test

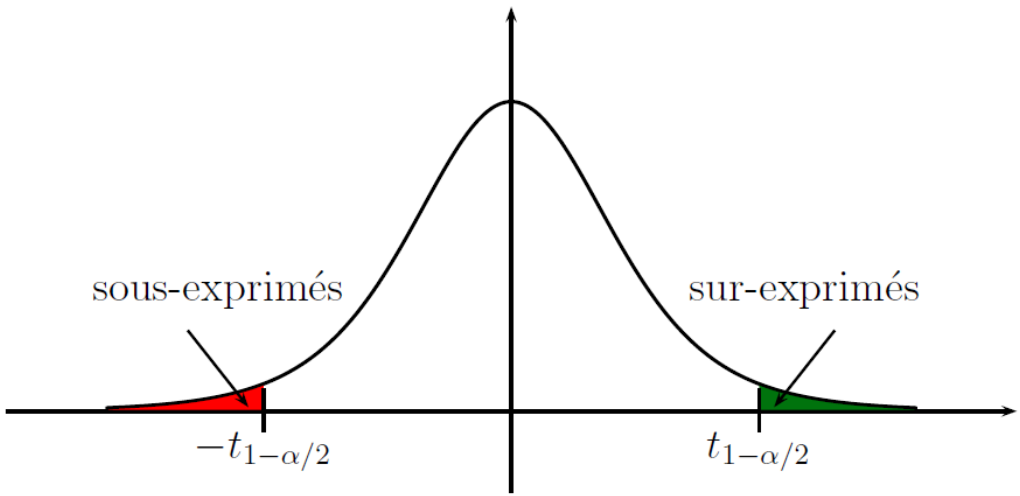
p-value $p(t)$

For a realisation t of the T test statistic $p(t)$ is the probability (calculating under H_0) of obtaining a test statistic at least as extreme as the one that was actually observed.

In bilateral case :

$$p(t) = \mathbb{P}_{H_0}\{|T| \geq |t|\}$$

The p-value measures the agreement between H_0 and the obtained result.



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 ?

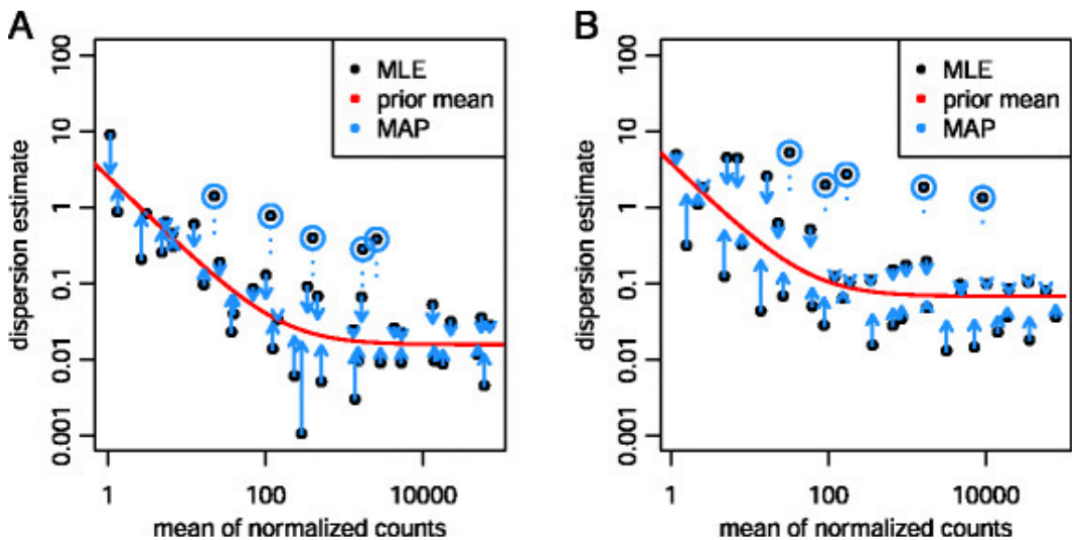
- gene-specific tests $\Rightarrow$ lack of sensitivity (proportion of true positives among positives) due to the lack of information
- common dispersion parameter for all tests $\Rightarrow$ many false positives

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

Example of empirical bayesian approach in DESeq2

Hypothesis : genes of similar average expression strength have similar dispersion

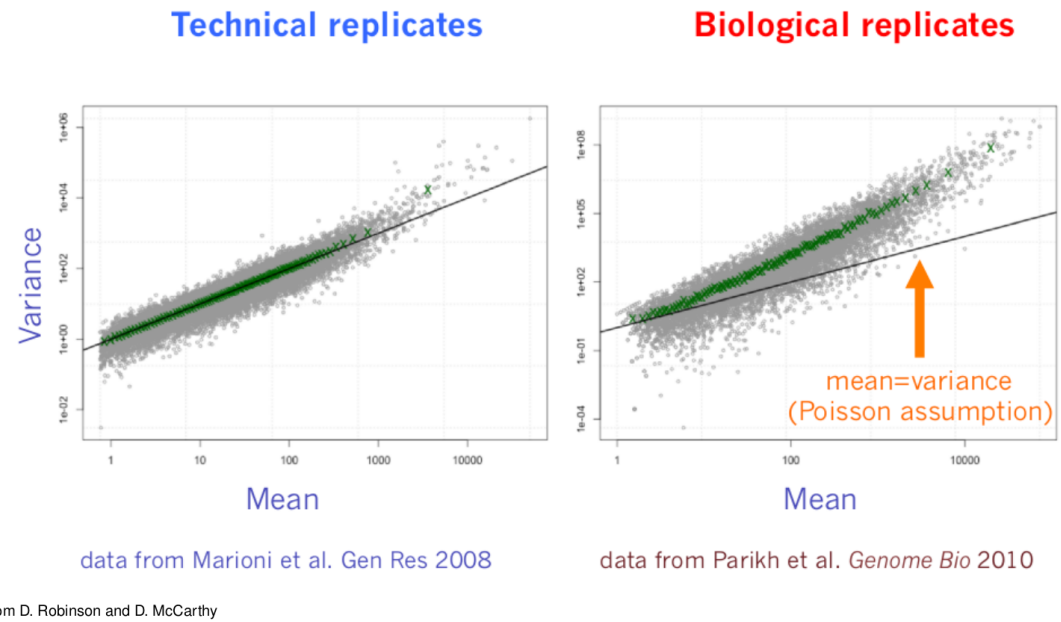
1. Estimate gene-wise dispersion estimates using maximum likelihood (ML) (black dots)
2. Fit a smooth curve (red line)
3. Shrink the gene-wise dispersion estimates (empirical Bayes approach) toward the values predicted by the curve to obtain final dispersion values (blue arrow heads).



Empirical bayesian approaches

For microarrays : package limma

For bulk RNA-Seq : edgeR, DESeq2 (negative binomial distributions) or limma-voom



From D. Robinson and D. McCarthy

For proteomic and metabolomic data : limma can be used after appropriate normalization

Plan

1 Main steps

2 Dimensionality reduction

3 Differential analysis of omics data

4 Multiple testing

5 Gene Set Enrichment Analysis

6 Conclusions

Multiple Testing

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

For all 'genes', we test H_0 (gene i is not DE) vs H_1 (the gene is DE) using a statistical test

Problem

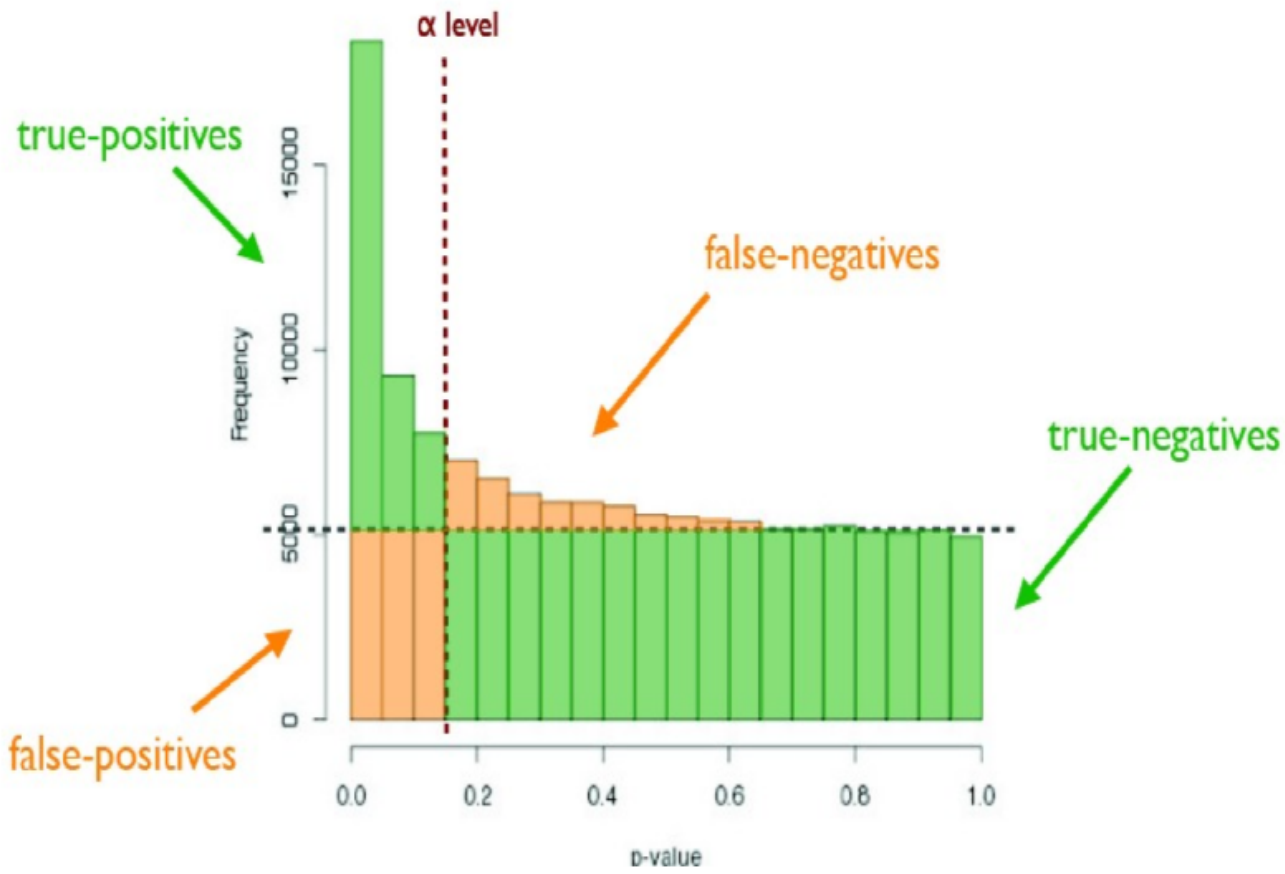
Let assume all the G genes are not DE. Each test is performed at α level
Ex : $G = 10000$ genes and $\alpha = 0.05 \rightarrow E(FP) = 500$ genes.

Simultaneous tests of G null hypotheses

Reality	Declared non diff. exp.	Declared diff. exp.
G_0 non DE genes	**True Negatives** (*TN*)	**False Positives** (*FP*)
G_1 DE genes	**False Negatives** (*FN*)	**True Positives** (*TP*)
G Genes	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 gene.

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

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 Hochberg, 1995) is the expected proportion of Type I errors among the rejected hypotheses

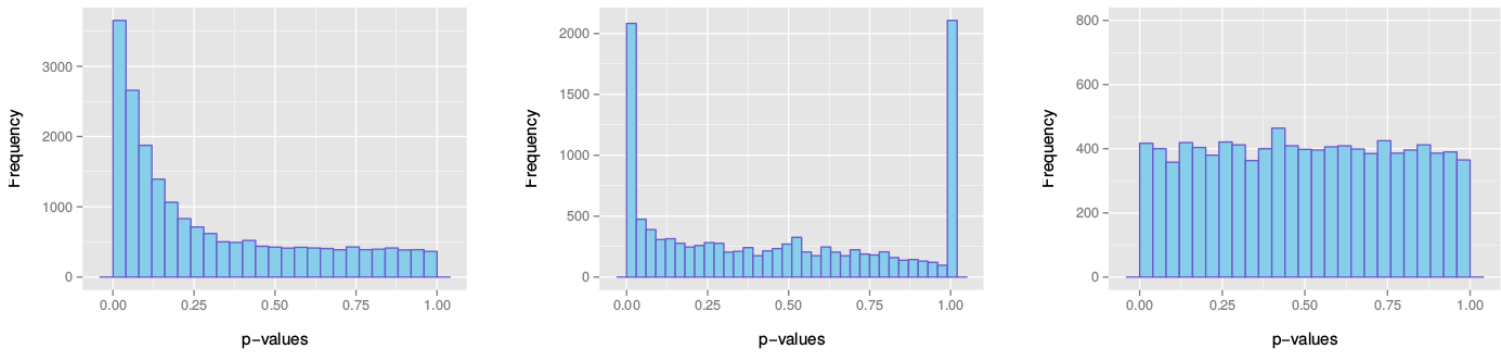
$$\text{FDR} = \mathbb{E}(FP/P) \text{ if } P > 0 \text{ and } 0 \text{ if } P = 0$$

Prop

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

p-values histograms for diagnosis

Examples of expected overall distribution



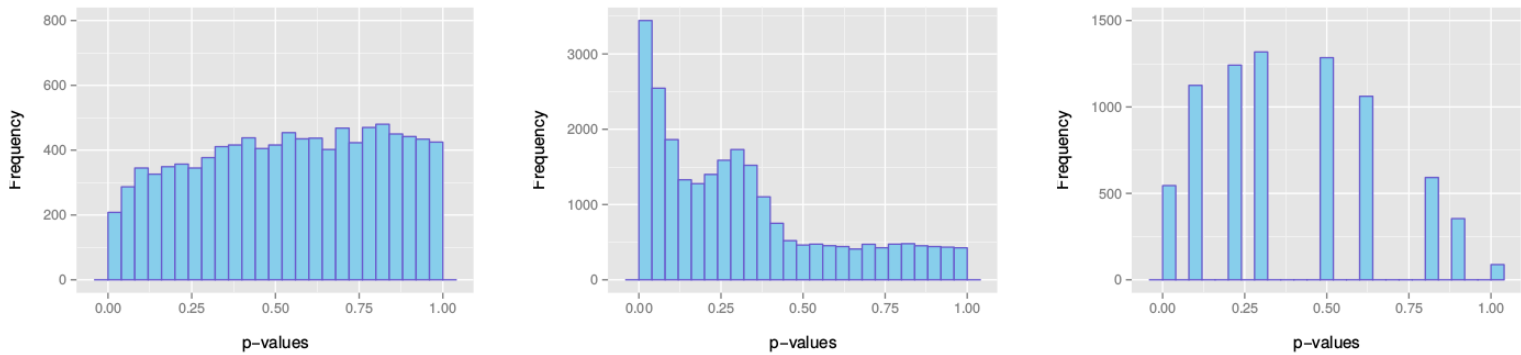
(a) : the most desirable shape

(b) : very low counts genes 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

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.

Plan

1 Main steps

2 Dimensionality reduction

3 Differential analysis of omics data

4 Multiple testing

5 Gene Set Enrichment Analysis

6 Conclusions

Gene Set Enrichment Analysis (GSEA)

Gene sets (Subramanian et al., 2005) : groups of genes that share common biological function, chromosomal location, or regulation.

Motivation :

GSEA can reveal many biological pathways in common where single-gene analyses find little similarities between independent studies (Subramanian et al., 2005)

Moelcular Signatures Database available at :
`http://software.broadinstitute.org/gsea/msigdb/index.jsp`

Over-Representation Analysis (ORA)

Compute overlaps with other gene sets in MSigDB

Use of the hypergeometric distribution which describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, *without replacement*, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure.

The hypergeometric uses the hypergeometric distribution to identify which gene-sets are over-represented in the list of differentially expressed genes. This test is identical to the corresponding one-tailed version of Fisher's exact test.

GSEA history

History of a very cited procedure implemented in the software available on the Broad Institute website :

- first paper : Mootha et al., Nature Genetics, 2004
- Damian and Gorfine published Statistical concerns about the GSEA procedure, Nature Genetics, 2004
- Subramanian et al., PNAS, 2005 : definition of a normalized enrichment score (NES)

GSEA

To compute the enrichment score (ES), no need to pre-specify cut-offs on p-values and log fold changes, the method asks for a ranked list L.

The user can load raw or normalised data and ask the software to rank the data according to a criterion. Otherwise, it is possible to give a pre-ranked list calculated outside the software, e.g. by limma.

Various criteria provided in the guide : `http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html`

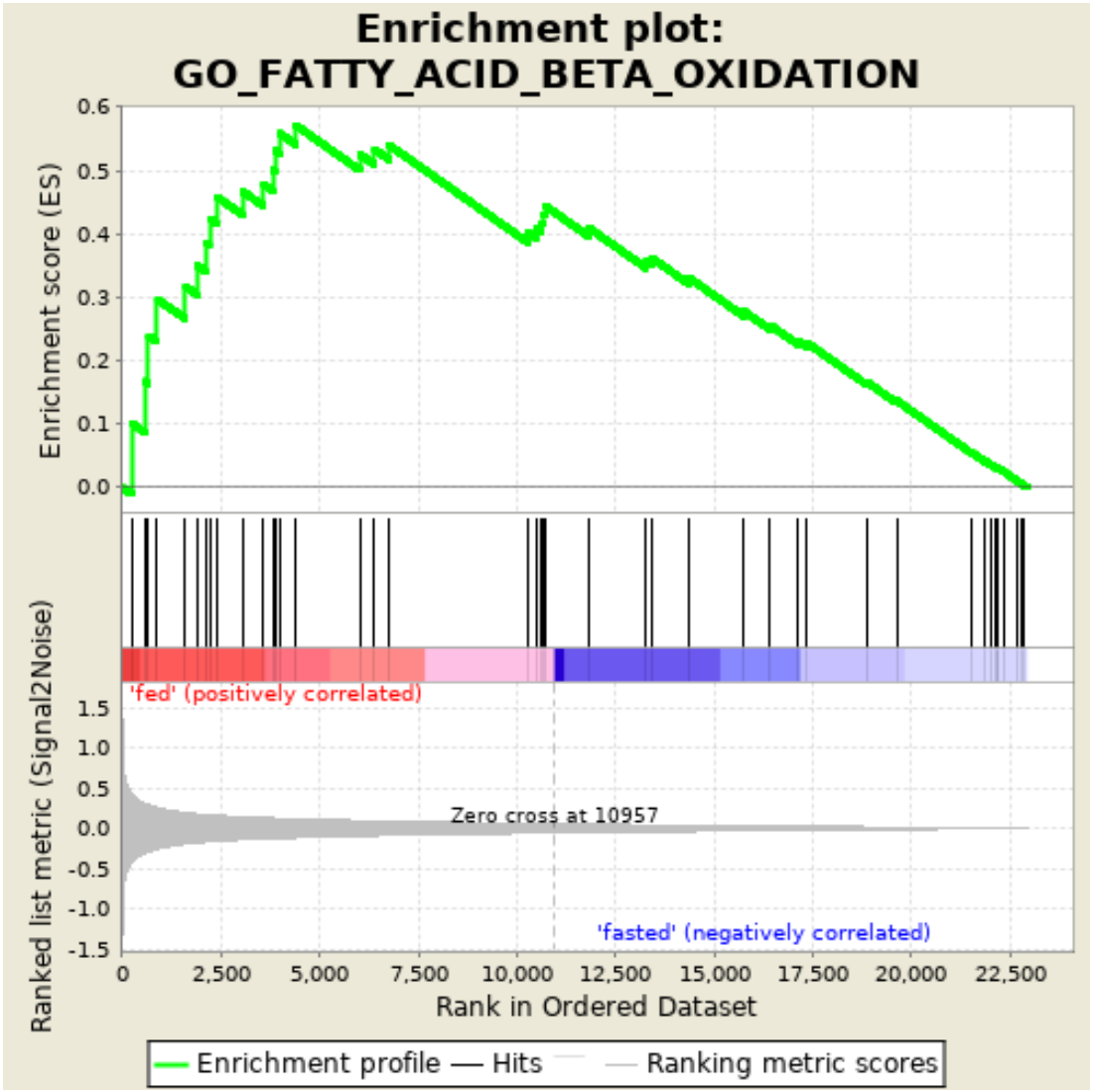
GSEA

The ES reflects the degree to which a set S is over-represented at the extremes (top or bottom) of the entire ranked list L.

The score is calculated by walking down the list L, increasing a running-sum statistic when we encounter a gene in S and decreasing it when we encounter genes not in S.

The magnitude of the increment depends on the ranking metric of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk.

GSEA


Enrichment plot:
GO_FATTY_ACID_BETA_OXIDATION
Enrichment score (ES)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
'fed' (positively correlated)
Ranked list metric (Signal2Noise)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
Zero cross at 10957
'fasted' (negatively correlated)
0
2,500
5,000
7,500
10,000
12,500
15,000
17,500
20,000
22,500
Rank in Ordered Dataset
Enrichment profile — Hits
Ranking metric scores

GSEA

Estimation of the statistical significance (raw p-value) using phenotype permutations.

- advantage of phenotype permutations : preserving the correlation structure between genes
- not advised to use phenotype permutations when less than 7 samples per condition. In that case, use gene permutations
- in the case of a pre-ranked list, the only possibility is to perform gene permutations

Normalization of the ES for each gene set to account for the size of the set

Adjustment for multiple testing with False Discovery Rate (q-value)

Plan

1 Main steps

2 Dimensionality reduction

3 Differential analysis of omics data

4 Multiple testing

5 Gene Set Enrichment Analysis

6 Conclusions

Conclusions

- Include replicates when you want to generalize !
- Normalisation depends both on the type of omics and on the statistical question.
- The statistical procedure also depends on the biological question.
- Differential analysis is exploratory. Do not forget to confirm by biological experiments.
- Do not forget to correct for multiple testing if you do not want to waste money in validation studies !
- The cost of statistical analysis can be inversely related to the number of replicates : experiments with few replicates necessit specific methods. The need for 'sophisticated' methods decreases when the number of replicates increases.

Want to go further ?

To learn more :

- training in R : `https://doctorat.univ-lille.fr/college-doctoral/formations/`
- training in high-throughput sequencing analysis : `https://bilille.univ-lille.fr/training/training-offer`

To obtain help in statistical analysis of omics data :
write an e-mail to bilille@univ-lille.fr

Annex : SARtools

SARTools : Statistical Analysis of RNA-Seq Tools (Varet et al., 2016)

- exports the results into easily readable **tab-delimited files**
- generates **a HTML report** which displays all the figures produced, explains the statistical methods and gives the results of the differential analysis.

- Exploratory data analysis
- Differential analysis including normalization and multiple testing

Available on R and Galaxy

Annex : Exploratory data analysis

Sample comparison for RNA-Seq (Schulze et al., 2012)

Pearson's correlation coefficient

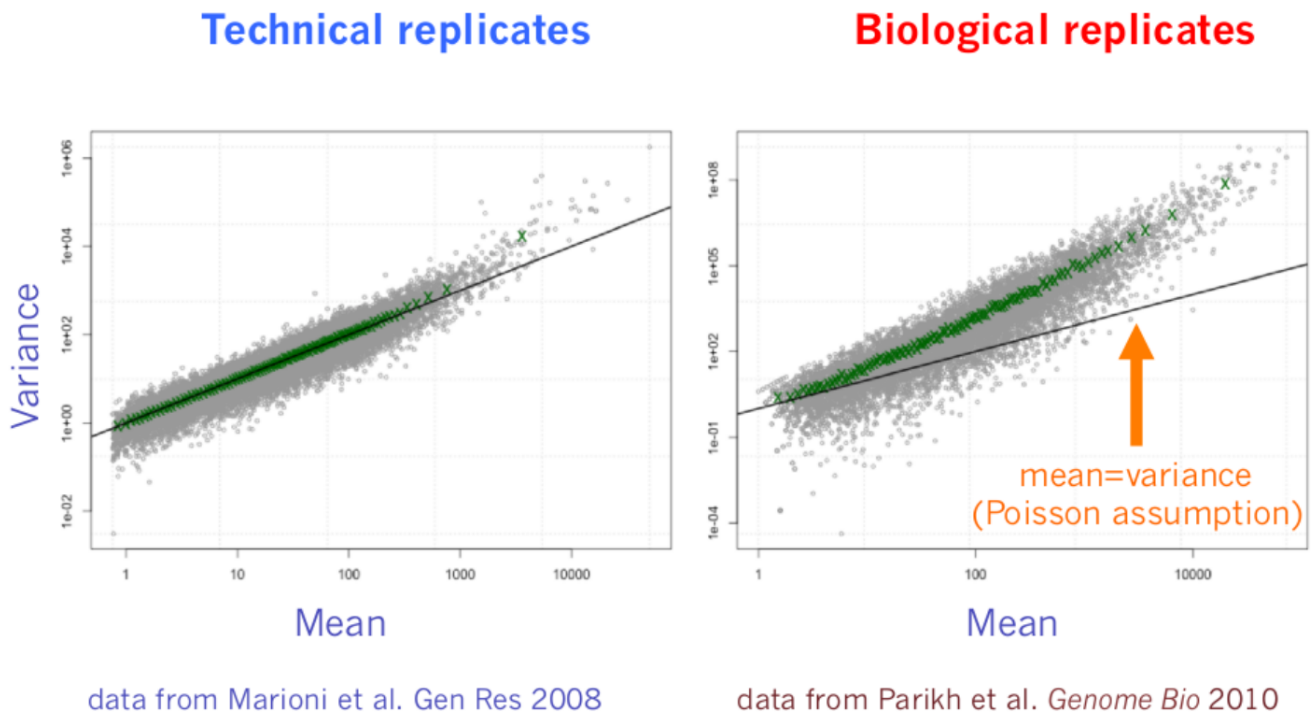
- widely used . . .
- . . .but highly dependent on sequencing depth and the range of expression samples inherent to the sample.

SERE : Simple Error Ratio Estimate

- ratio of observed variation to what would be expected from an ideal Poisson experiment
- interpretation unambiguous regardless of the total read count or the range of expression
- score of 1 : faithful replication
- score of 0 : data duplication
- scores > 1 true global differences between RNA-Seq libraries

Annex : SERE

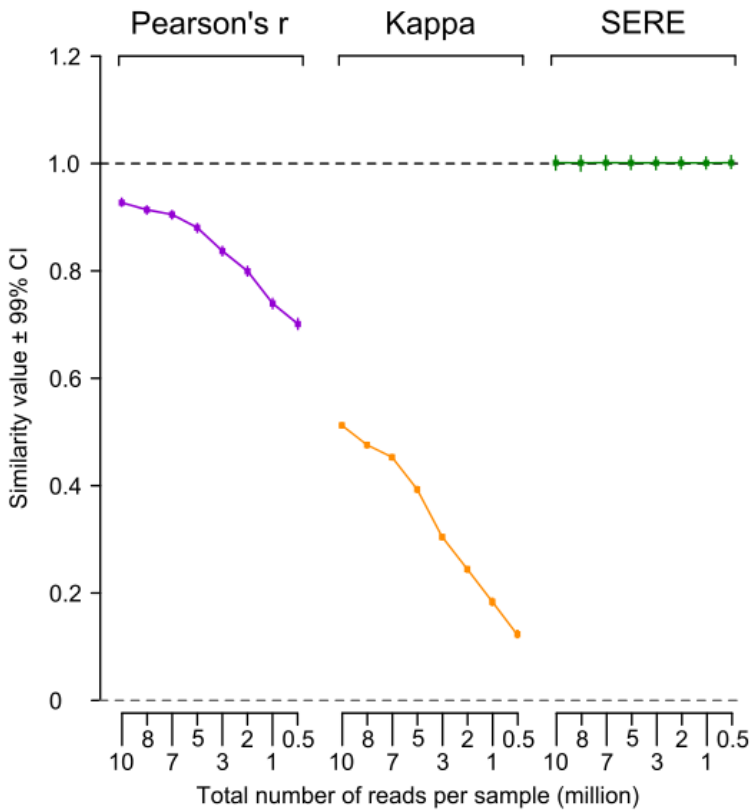
scores between 0 and 1 $\Rightarrow$ underdispersion (variance smaller than mean)



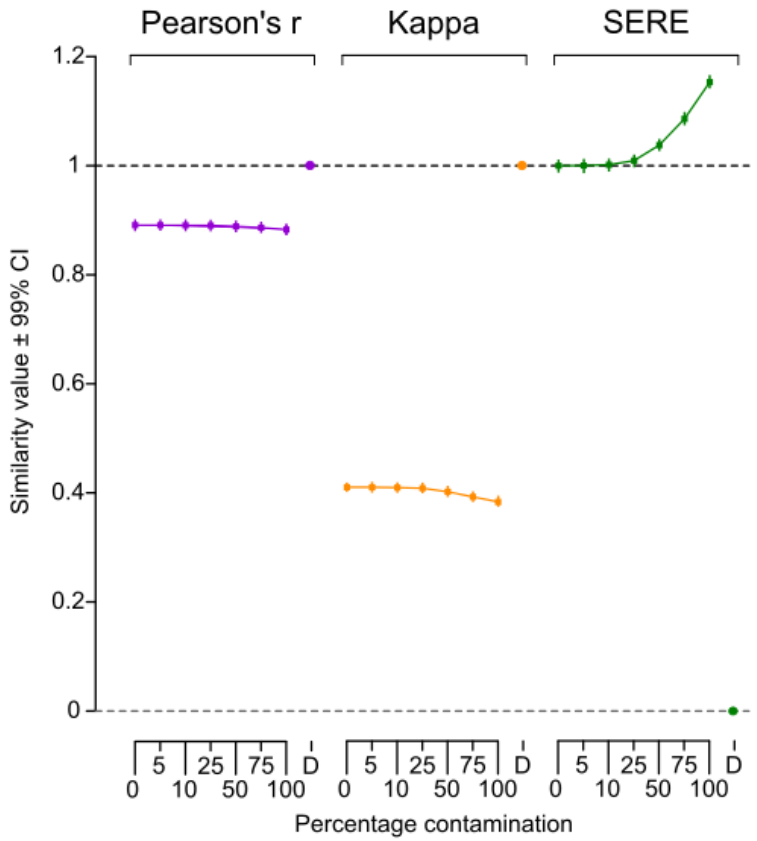
From D. Robinson and D. McCarthy

scores greater than 1 : overdispersion $\Rightarrow$ adapted to biological replicates

Annex : Sample comparison for RNA-Seq



total read count dependence



sensitivity to contamination

source :

(Schulze et al., 2012)