# Statistical analysis of RNA-Seq data

**G. Marot (Univ. Lille)**
Sources: J. Aubert and C. Hennequet-Antier (Inrae)
M.A. Dillies and H. Varet (Institut Pasteur Paris)
Assistant for practical exercises: Samuel Blanck (Univ. Lille)

14-17 octobre 2022

# Introduction

## Differential analysis

Comparison of treatments, states, conditions, ...
Example : ill vs healthy
$\Rightarrow$ statistical analysis based on tests

Particularities of NGS data :

- Very few individuals
- Many tests (one per variable)
- Count data (statistical distributions different from the ones used for continuous data from microarrays)

# Introduction

Preamble :

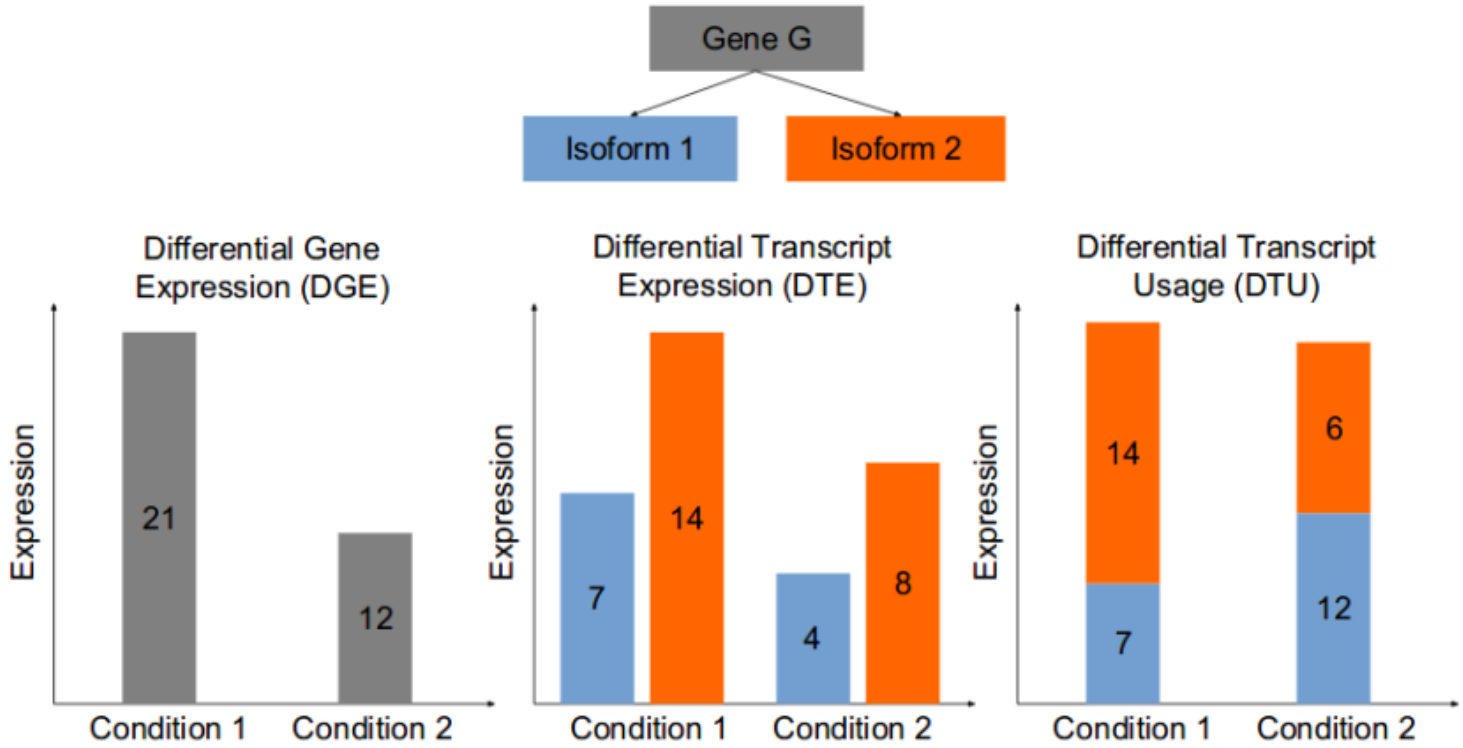
Obtaining a result using a statistical procedure does not mean that this result is reliable. If you do not know the assumptions behind, please be careful with interpretation or ask an expert to help you.

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

DGE : differential gene expression, DTE : differential transcript expression, DTU : differential transcript usage




This course focuses on DGE

# Differential analysis

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.

Key steps for statisticians :

- experimental design
- normalization
- differential analysis
- multiple testing

# Plan

1. Experimental design
2. Exploratory data analysis
3. Normalization
4. Differential analysis
5. Multiple testing
6. Gene Set Enrichment Analysis

# Not a recent idea !

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)

# Make an experimental design

## Context of a RNA-seq experiment

Rule 0 : Share a common language in biology, bioinformatics and statistics.

### Experimental design

All skills are needed to discussions right from project construction.

- Rule 1 : Well define the biological question, get together and collect a priori knowledge (e.g. reference genome, splicing),
- Rule 2 : Anticipate, identify all factors of variation and adapt Fisher's principles (1935), collect metadata from experiment and sequencing,
- Rule 3 : Choose a priori tools/methods for bioinformatics and statistical analyses,
- Rule 4 : Draw conclusions on results.

# Experimental design

A good design is a list of experiments to conduct in order to answer to the asked question which maximize collected information and minimize experiments cost with respect to constraints.

Rule 1 : Well define the biological question : make a choice

- Identify differentially expressed genes,
- Detect and estimate isoforms,
- Construct a de novo transcriptome.

Rule 2 : adapt Fisher's principles : randomization and blocking AVOID CONFUSION between the biological variability of interest and a biological or technical source of variation

# Experimental design

## 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).

Sequencing technology does not eliminate biological variability.
(Nature Biotechnology Correspondence, 2011)

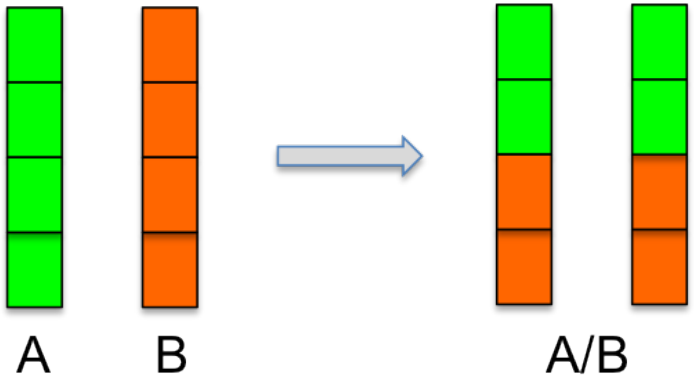
$$\text{lane effect} < \text{run effect} < \text{library prep effect} << \text{biological effect}$$

[Marioni et al., 2008],[Bullard et al., 2010]

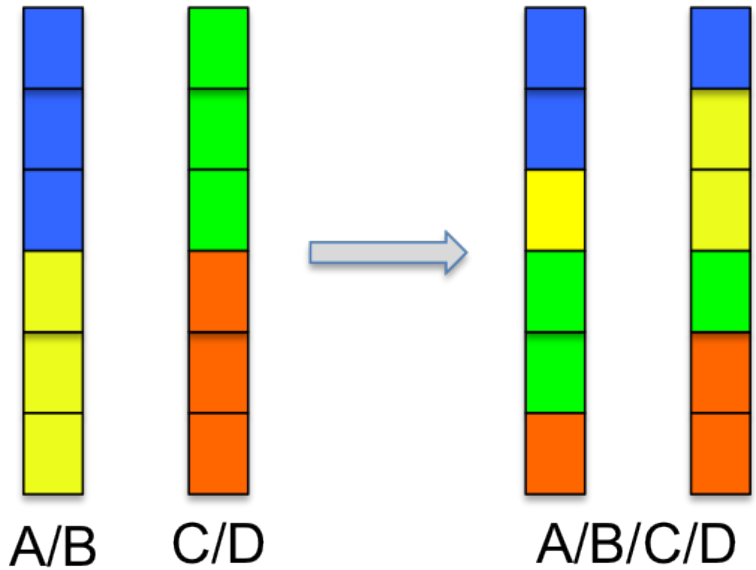
Include at least three biological replicates in your experiments !
Technical replicates are not necessary.

# Experimental design

AVOID CONFUSION between the biological variability of interest and a biological or technical source of variation




*Problem :* Confusion between lane and condition
*Solution :* Distribute the conditions evenly on both lanes




*Problem :* Partial confusion between lane and condition
*Solution :* Distribute the conditions "evenly" on both lanes

# 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, 2021 |
| S2 | control | July 12th, 2021 |
| S3 | control | July 12th, 2021 |
| S4 | wound | July 20th, 2021 |
| S5 | wound | July 20th, 2021 |
| S6 | wound | July 20th, 2021 |

**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, 2021 |
| S2 | control | July 20th, 2021 |
| S3 | control | July 25th, 2021 |
| S4 | wound | July 12th, 2021 |
| S5 | wound | July 20th, 2021 |
| S6 | wound | July 25th, 2021 |

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, 2021 | m1 |
| S2 | control | July 20th, 2021 | m2 |
| S3 | control | July 25th, 2021 | m3 |
| S4 | wound | July 12th, 2021 | m1 |
| S5 | wound | July 20th, 2021 | m2 |
| S6 | wound | July 25th, 2021 | 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.

# Experimental design

## Why increasing the number of biological replicates ?

- To generalize to the population level
- To estimate with a higher degree of accuracy variation in individual transcript [Hart et al., 2013]
- To improve detection of DE transcripts and control of false positive rate [Soneson and Delorenzi, 2013]
- To focus on detection of low mRNAs, inconsistent detection of exons at low levels ($\leq$ 5 reads) of coverage [McIntyre et al., 2011]

## More biological replicates or increasing sequencing depth ?

It depends ! [Haas et al., 2012], [Liu et al., 2014]

- DE transcript detection : (+) biological replicates
- Construction and annotation of transcriptome : (+) depth and (+) sampling conditions
- Transcriptomic variants search : (+) biological replicates and (+) depth

## Support

- An experimental design using multiplexing,
- Tools for experimental design decisions : Scotty [Busby et al., 2013], RNAseqPower [Hart et al., 2013], PROPER [Wu et al., 2015]

And do not forget : budget also includes cost of biological data acquisition, sequencing data backup, bioinformatics and statistical analysis.

# For a good (nice) experiment design ...

## Before the experiment

- Ask a precise and well defined biological question
- List all possible biological confounding effects (sex, age, ...)
- Collect samples while taking care of the distribution of unwanted sources of variation across samples
- Include at least three biological replicates per condition. Technical replicates are not necessary
- Distribute samples on lanes and flow cells ...
  - according to the comparisons to be made
  - without introducing a confusion between technical effects and the biological effects of interest
  - applying the same multiplexing rate on all samples

# Plan

1 Experimental design

2 Exploratory data analysis

3 Normalization

4 Differential analysis

5 Multiple testing

6 Gene Set Enrichment Analysis

# 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

# Exploratory data analysis

Sample comparison for RNA-Seq [Schulze et al., 2012]

## Pearson's correlation coefficient

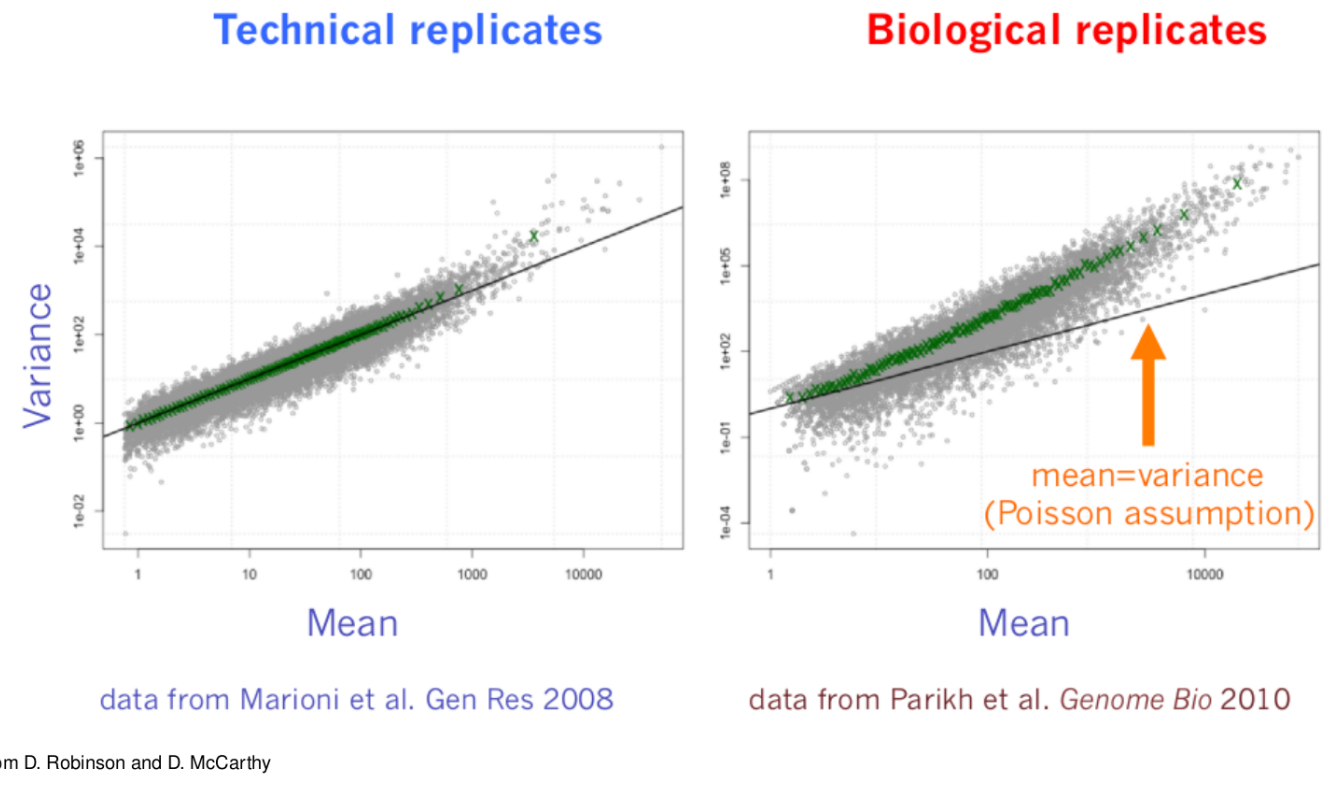
- widely used . . .
- . . .but highly dependent on sequencing depth and the range of expression samples inherent to the sample.

## 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

# Exploratory data analysis

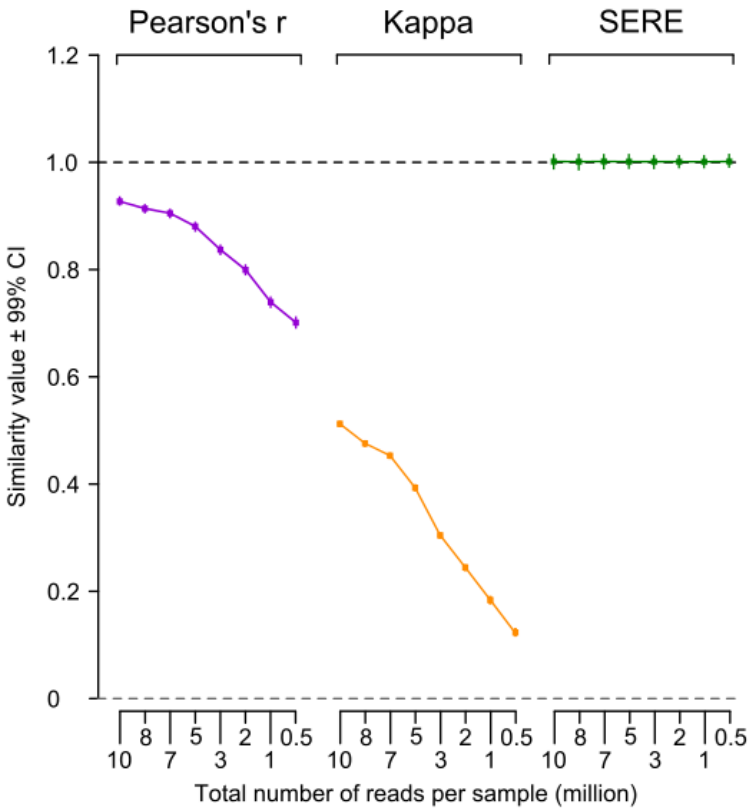
scores between 0 and 1 $\Rightarrow$ underdispersion (variance smaller than mean)



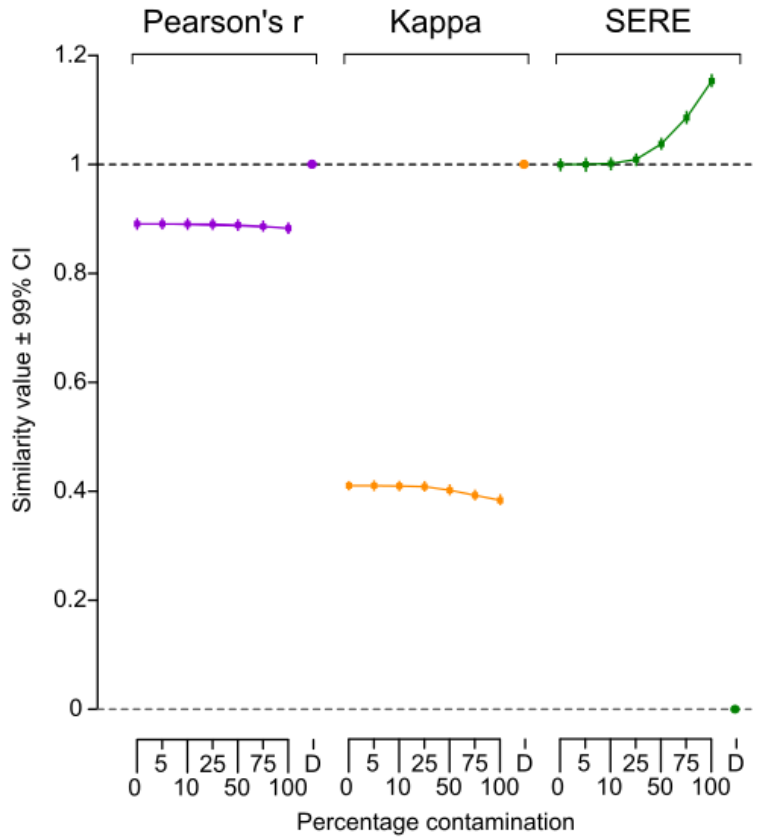

From D. Robinson and D. McCarthy

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

# Sample comparison for RNA-Seq




total read count dependence




sensitivity to contamination

source : [Schulze et al., 2012]

# Exploratory data analysis

## Multivariate exploratory data analysis

Main goal : explore the structure of the dataset to better understand the proximity between samples and detect possible problems. **This is a quality control step**

### Two main tools

- Principal Component Analysis (PCA) or MultiDimensional Scaling (MDS)
- Clustering

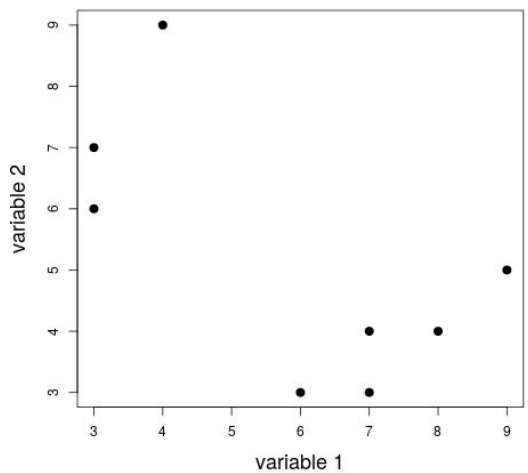
# Dimensionality reduction

Problem : $n$ individuals, $p$ genes

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

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

Possibility to visualize pair-wise relations by scatter plots :




When p is large, this is not efficient !

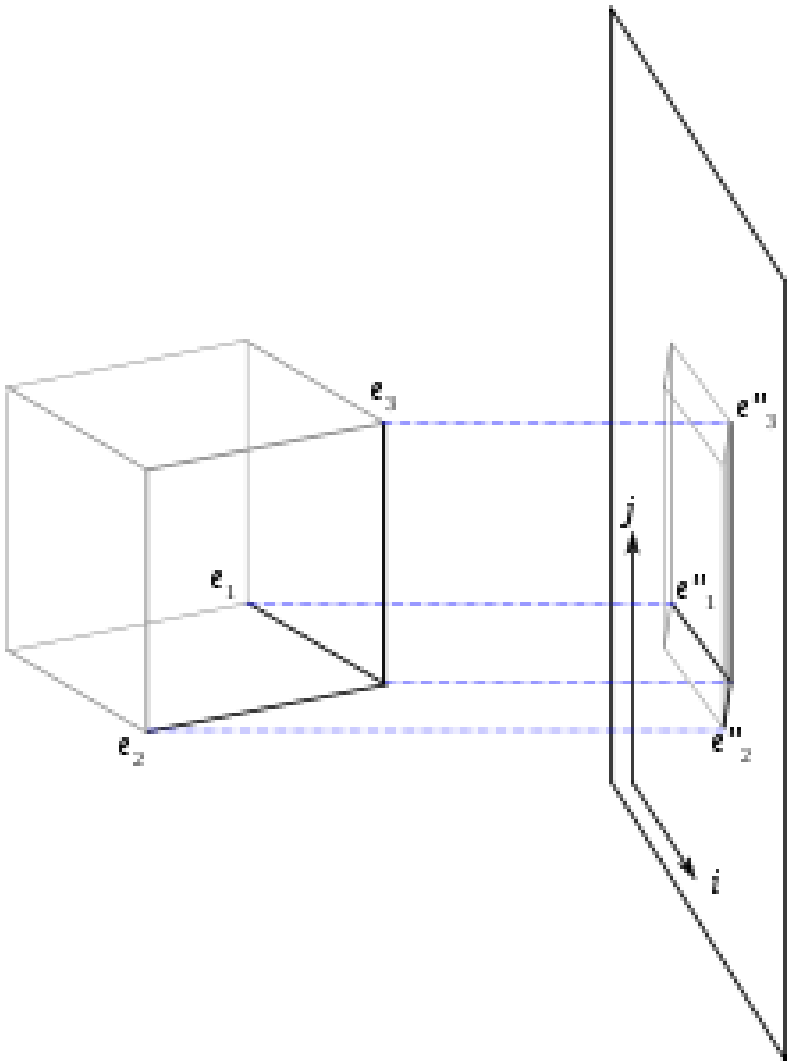
# Principal components analysis

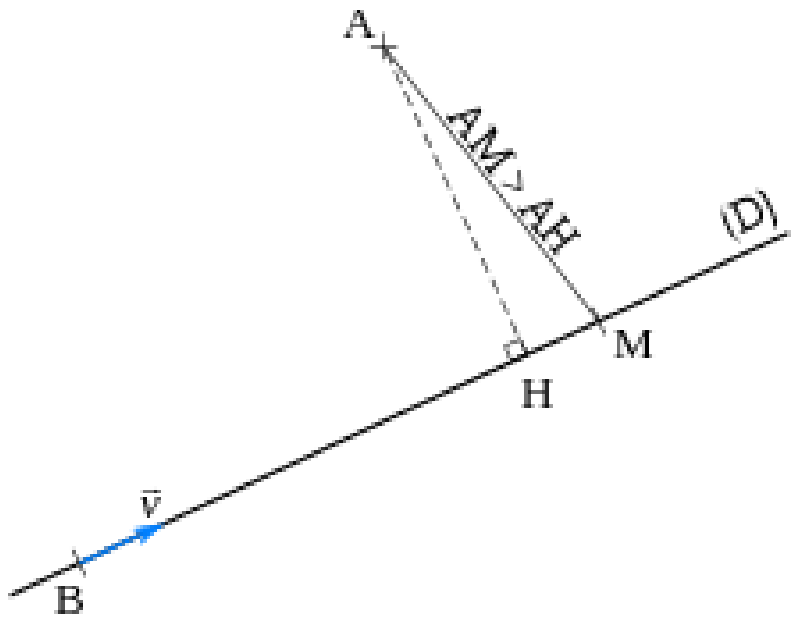
## Principal component analysis (PCA) :

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

# Analyse en composantes principales



# 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.

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

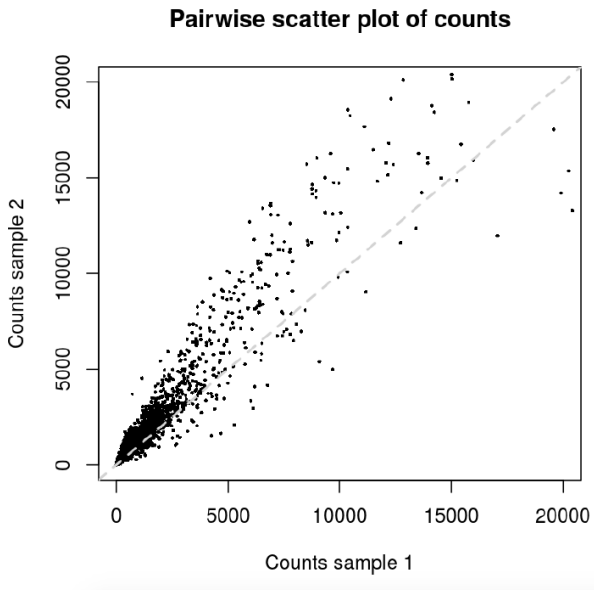
A pre-requisite to apply PCA is to make the data homoscedastic : **the variance must be independent of the intensity**.

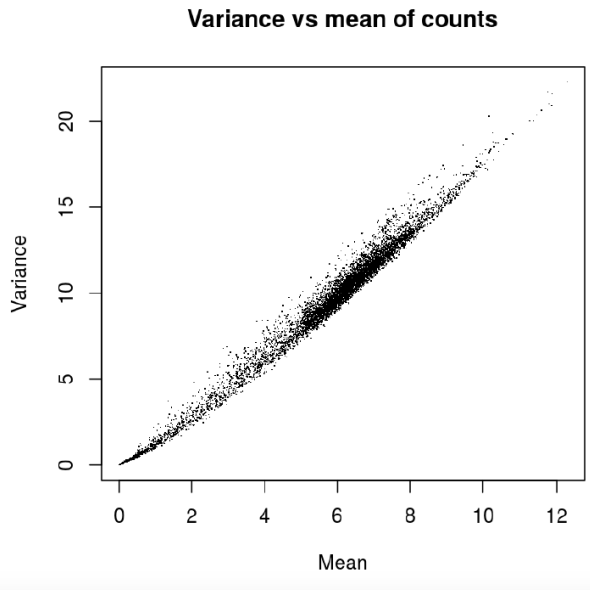
# Exploratory data analysis

Transformations proposed :

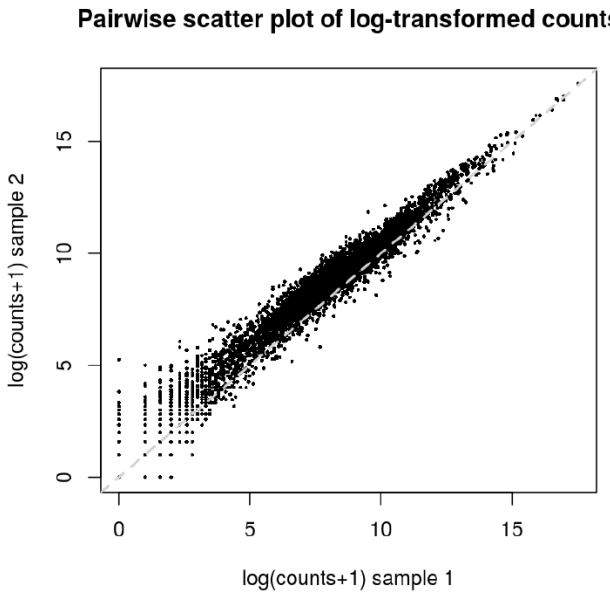
- DESeq2 : VST (Variance Stabilizing Transformation) or rlog (Regularized Log Transformation)
- edgeR : transformation of the count data as moderated log-counts-per-million

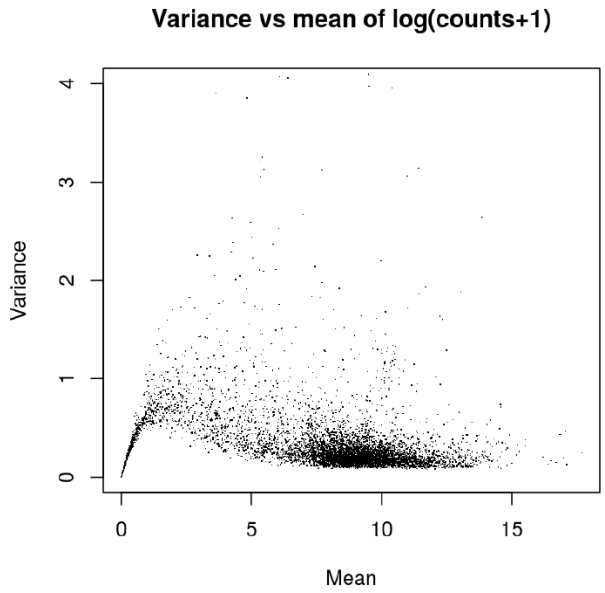
Illustration : Without transformation : variance increases with mean

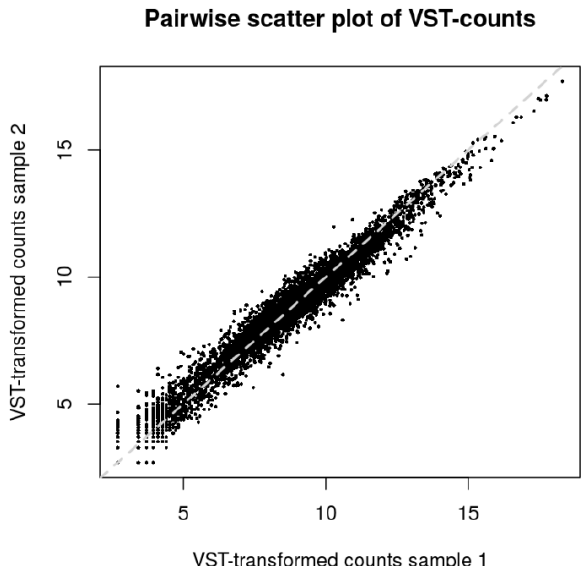


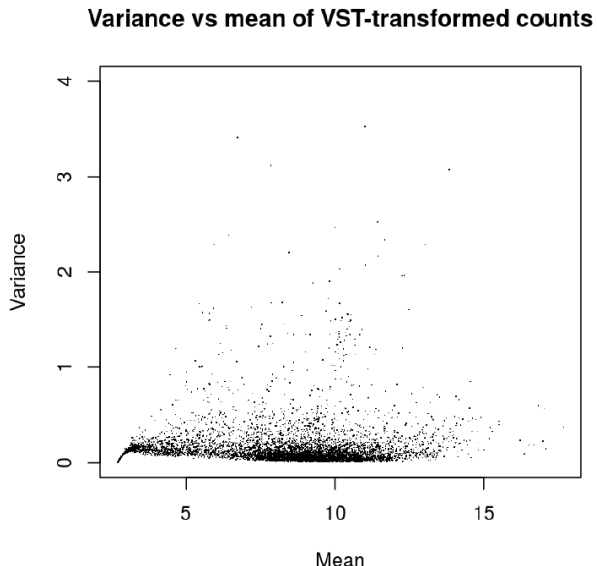



# Exploratory data analysis - VST transformation



# SARtools

PRACTICE

# Plan

1 Experimental design

2 Exploratory data analysis

3 Normalization

4 Differential analysis

5 Multiple testing

6 Gene Set Enrichment Analysis

# Normalization

## 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 fragments (genes) from a same sample.
No need in a differential analysis context.

## Between-sample normalization

Normalization enabling comparisons of fragments (genes) from different samples.

# Sources of variability

Read counts are proportional to expression level, gene length and sequencing depth (same RNAs in equal proportions).

## Within-sample

- Gene length
- Sequence composition (GC content)

## Between-sample

- Depth (total number of sequenced and mapped reads)
- Sampling bias in library construction ?
- Presence of majority fragments
- Sequence composition due to PCR-amplification step in library preparation [Pickrell et al., 2010], [Risso et al., 2011]

# Comparison of normalization methods

## A lot of different normalization methods...

- Some are part of models for DE, others are 'stand-alone'
- They do not rely on similar hypotheses
- But all of them claim to remove technical bias associated with RNA-seq data

Which one is the best ?

[Dillies et al., 2013], on behalf of StatOmique Group
Evaluation of normalization methods for RNA-Seq differential analysis at the gene level

# Comparison of normalization methods

**Focus on methods which aim at making read counts comparable across samples**

## Two main types

1. Methods that make read count distributions similar (if not equal)
2. Methods assuming that most genes are not differentially expressed

## Note that :

- These methods apply on raw (integer) count data, to RNA-seq data (metagenomics), for differential expression analysis
- Other more complex methods have been proposed after the comparison [Risso et al., 2014]
- Library size : Number of reads that have been sequenced, mapped and counted for a given sample (sum on columns on the count table)

# Which method should I use ? [Dillies et al., 2013]

## In most cases

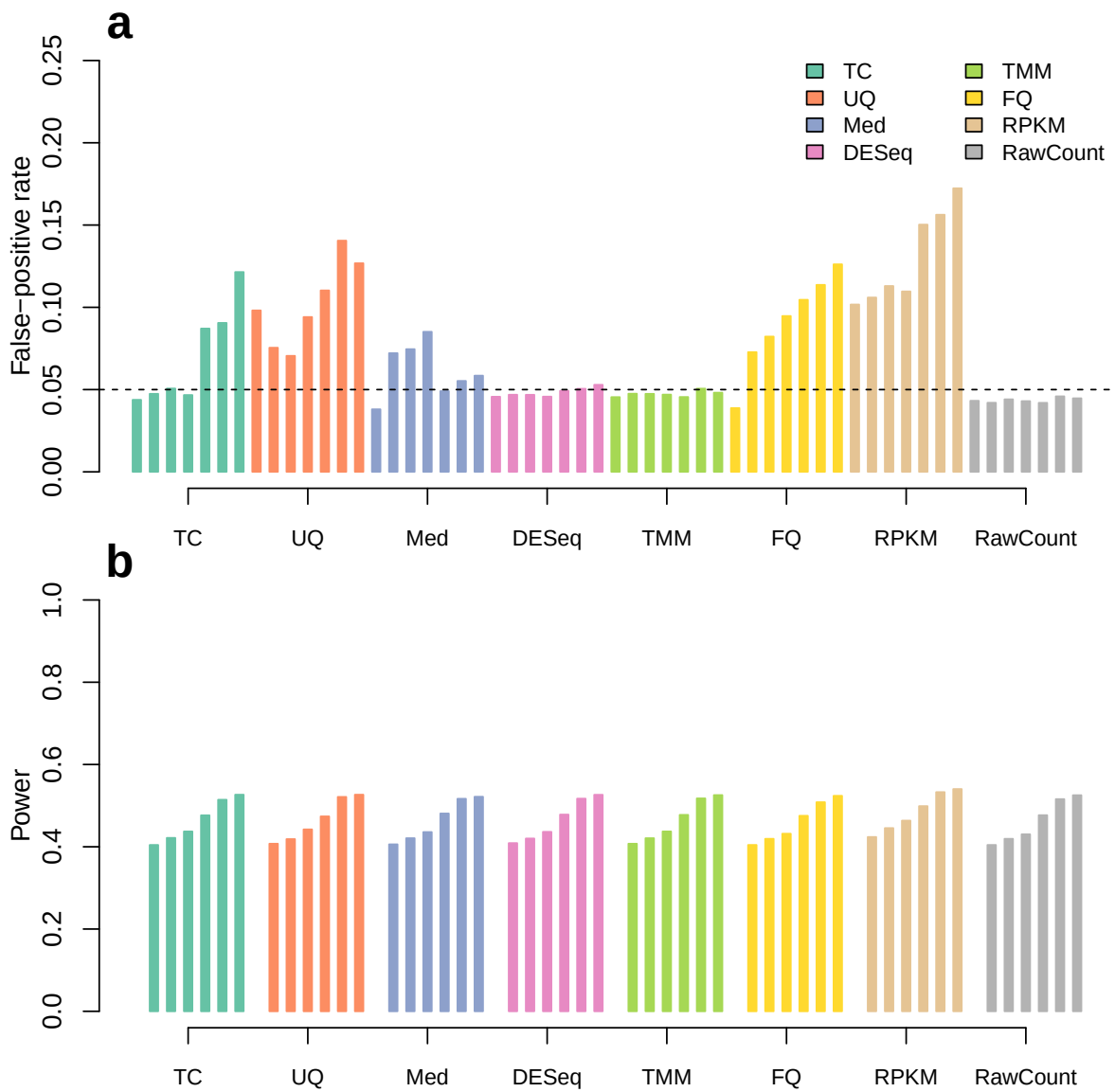
All methods provide comparable results

## Anyway ...

Clear differences appear in the presence of high count genes or when the expressed RNA repertoire varies notably across samples

| Method | Distribution | Intra-Variance | Housekeeping | Clustering | False-positive rate |
|---|---|---|---|---|---|
| TC | – | + | + | – | – |
| UQ | ++ | ++ | + | ++ | – |
| Med | ++ | ++ | – | ++ | – |
| **DESeq** | ++ | ++ | ++ | ++ | ++ |
| **TMM** | ++ | ++ | ++ | ++ | ++ |
| FQ | ++ | – | + | ++ | – |
| RPKM | – | + | + | – | – |

# Which method should I use ? [Dillies et al., 2013]



# Conclusions

- Hypothesis : the majority of genes is invariant between two samples.
- Differences between methods when presence of majority sequences, very different library depths.
- TMM and DESeq : performant and robust methods in a DE analysis context on the gene scale.
- Normalization is necessary and not trivial.
- Detection of differential expression in RNA-seq data is inherently biased (more power to detect DE of longer genes)
- Do not normalise by gene length in a context of differential analysis.

# Plan

1 Experimental design

2 Exploratory data analysis

3 Normalization

4 Differential analysis

5 Multiple testing

6 Gene Set Enrichment Analysis

# Statistical significance and practical importance

**Differential analysis :**

Detect differentially expressed genes between two conditions

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 ?

# 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.

# 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.

Practical importance and statistical significance 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.

# Differential analysis

Aim : Detect differentially expressed genes between two conditions

- Discrete quantitative data
- Few replicates
- Overdispersion problem

Challenge : method which takes into account overdispersion and a small number of replicates

- Proposed methods : edgeR, DESeq for the most used and known [Anders et al., 2013]
- An abundant litterature
- Comparison of methods : [Pachter, 2011], [Kvam and Liu, 2012], [Soneson and Delorenzi, 2013], [Rapaport et al., 2013]

# Statistical test

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

  $H_1$= {the mean expression ((or proportion) of the gene 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 ($\alpha$), 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.

# Critical region and p-value

**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.

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

# Critical region and p-value

## 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.

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

## For each gene : is it differentially expressed between A and B ?

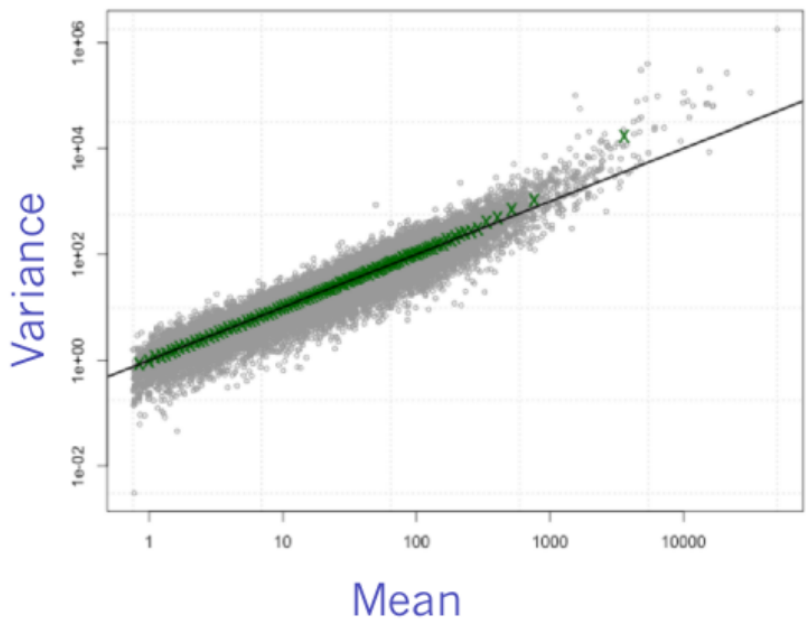
- Generalized linear framework
- Hypothesis to test : $H_{0i}$ Equality of relative abundance of gene i in condition A and B vs $H_{1i}$ non-equality
- Critical region - Wald Test or Likelihood Ratio Test

# Mean-Variance Relationship

## The Poisson distribution to model counts

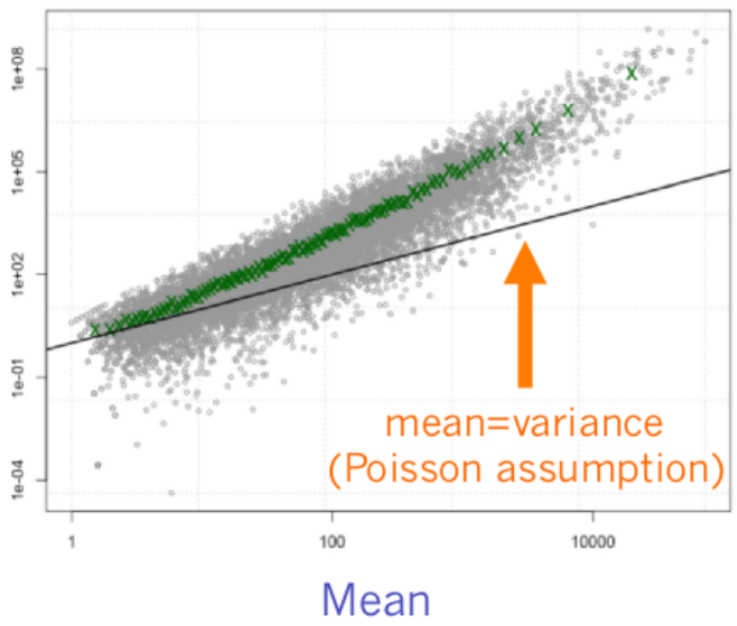
- Describes the number of occurences of rare events during a given time interval
- Property : Mean = Variance

**Technical replicates**




data from Marioni et al. Gen Res 2008

**Biological replicates**




data from Parikh et al. *Genome Bio* 2010

# Overdispersion in RNA-seq data

Counts from biological replicates tend to have variance exceeding the mean (= overdispersion). Poisson describes only technical variation.

**What causes this overdispersion ?**

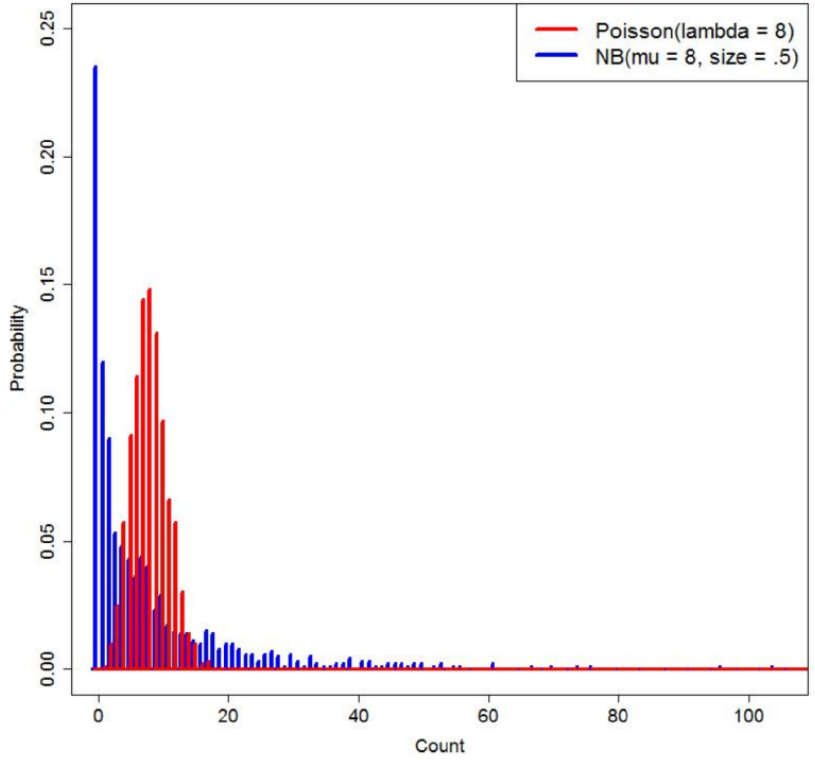
- Correlated gene counts
- Clustering of subjects
- Within-group heterogeneity
- Within-group variation in transcription levels
- Different types of noise present...

In case of overdispersion, increase of the type I error rate (probability to declare incorrectly a gene DE).

# Negative Binomial Models

A supplementary dispersion parameter $\phi$ 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

# Available tests

## Models of count data

- Data transformation and gaussian-based model : limma - voom
- Poisson : TSPM
- Negative Binomial : edgeR, DESeq(2), NBPSeq, baySeq, ShrinkSeq, ...

## Statistical approaches

- Frequentist Approach : edgeR, DESeq(2), NBPSeq, TSPM, ...
- Bayesian Approach : baySeq, ShrinkSeq, EBSeq, ...
- Non-parametric approach : SAMSeq, NOISeq, ...

# Comparison of two conditions [Soneson and Delorenzi, 2013]

A comparison of methods for differential analysis of RNA-Seq data [Soneson and Delorenzi, 2013]

- 11 statistical tests included in the study
- R packages
- input data are raw counts (gene-level analysis)
- TMM or DESeq normalization

# Main results

- **With only two biological replicates, all the methods show low performances.** They either lack power or poorly control the false positive rate.
- No method outperforms the others in all circumstances : **the method should be chosen according to the dataset**

## How to choose ?

- Number of replicates of the experiment
- Presence / absence of outliers
- Constant / variable within-group dispersion
- Balanced / unbalanced differential expression
  (results are more accurate and less variable between methods if DE genes are regulated in both directions)
- Simple / complex experiment design

# edgeR and DESeq(2)

## DESeq2 et edgeR : similarities . . .

- Easy to use and well documented R packages
- A 3-step analysis process : normalization, dispersion estimation, statistical test
- Negative Binomial distribution of counts and Generalized Linear Models (GLM) : allows analysis of simple and complex designs

## . . . and differences

- outlier detection and processing
- low counts filtering
- **dispersion estimation**

In both cases, the version matters

# Estimating the dispersion : the key question

## Problem

Estimate a reliable dispersion from a very small number of replicates (sometimes less than 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

# Empirical bayesian approaches

## Principles

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

$$\tilde{\boldsymbol{\theta}_g} = \widehat{\boldsymbol{\theta}_c} + b(\widehat{\boldsymbol{\theta}_g} - \widehat{\boldsymbol{\theta}_c})$$

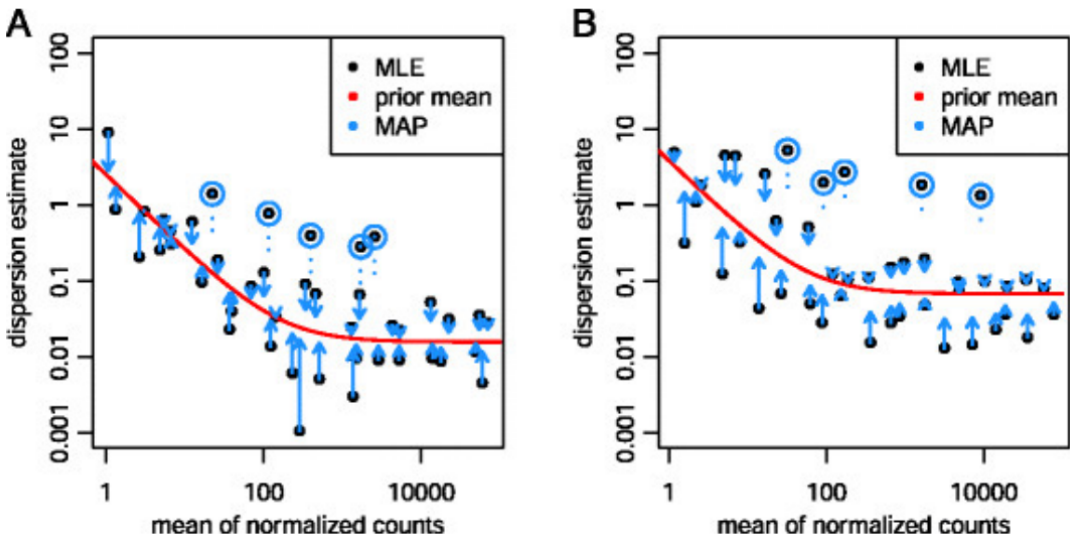
with $\tilde{\boldsymbol{\theta}_g}$ = shrinkage estimator
$\widehat{\boldsymbol{\theta}_c}$ = estimator of the mean population
$\widehat{\boldsymbol{\theta}_g}$ = usual empirical estimator gene by gene
$b$ =shrinkage factor

$b = 1 \Rightarrow \tilde{\boldsymbol{\theta}_g} = \widehat{\boldsymbol{\theta}_g}$
$b = 0 \Rightarrow \tilde{\boldsymbol{\theta}_g} = \widehat{\boldsymbol{\theta}_c}$

# Dispersion estimation with 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).



# Dispersion estimation with edgeR

1. Estimate gene-wise dispersion estimates using ML
2. Estimate a common dispersion parameter by ML
3. Moderate gene-wise dispersion estimates toward a common estimate or toward a local estimate from genes with similar expression strength using a weighted conditional likelihood.

Differences :

- DESeq2 estimates the width of the prior distribution from the data and therefore automatically controls the amount of shrinkage based on the observed properties of the data.
- edgeR requires a user-adjustable parameter, the prior degrees of freedom, which weights the contribution of the individual gene estimate and edgeR's dispersion fit.

# Differences between edgeR and DESeq(2)

- **edgeR** : borrow information across genes for stable estimates of $\phi$ ; 3 ways to estimate $\phi$ (common, trend, moderated)
- **DESeq2** : relationship of variance and mean + dispersion and fold change shrinkage (for PCA and Gene Set Enrichissment Analysis) + detection of outliers

Robustness

- **edgeR :** one option : moderate dispersion less towards trend
  Allows dispersions to be driven more by the data
- **DESeq2 :** calculate Cook's distance and filter genes with outliers
  Can inadvertently filter interesting genes

# Robustness - edgeR and DESeq(2)

- Robust edgeR (not by default in R) suffers a tiny bit in power with no outliers, but has good capacity to dampen their effect if present (be careful with reviews which take the value by default of edgeR) resulting in (sometimes drastic) drop in power
- DESeq2 is very powerful in the absence of outliers, but policy to filter outliers results in loss of power
- edgeR and edgeR robust are a bit liberal (5% FDR might mean 6% or 7%)

# Comparaison of differential analysis methods

[Soneson and Delorenzi, 2013]

- Small number of replicates (2-3) or low expression $\rightarrow$ be careful ! !
- Large number of replicates (10 or so) or very high expression $\rightarrow$ method choice does not matter much.
- Outlier counts affect different methods in different ways. Removing genes with outlier counts or using non-parametric methods reduce the sensitivity to outliers
- Allow tagwise dispersion values
- Normalization methods have problems when all DE genes are regulated in one direction. Results are more accurate and less variable between methods if DE genes are regulated in both directions.

# Comparaison of differential analysis methods

[Rapaport et al., 2013]

Evaluation on methods using SEQC benchmark dataset and ENCODE data.

- Significant differences between methods.
- Array-based methods adapted perform comparably to specific methods.
- Increasing the number of replicates samples significantly improves sensitivity over increased sequencing depth.

# Plan

1 Experimental design

2 Exploratory data analysis

3 Normalization

4 Differential analysis

5 Multiple testing

6 Gene Set Enrichment Analysis

# 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**

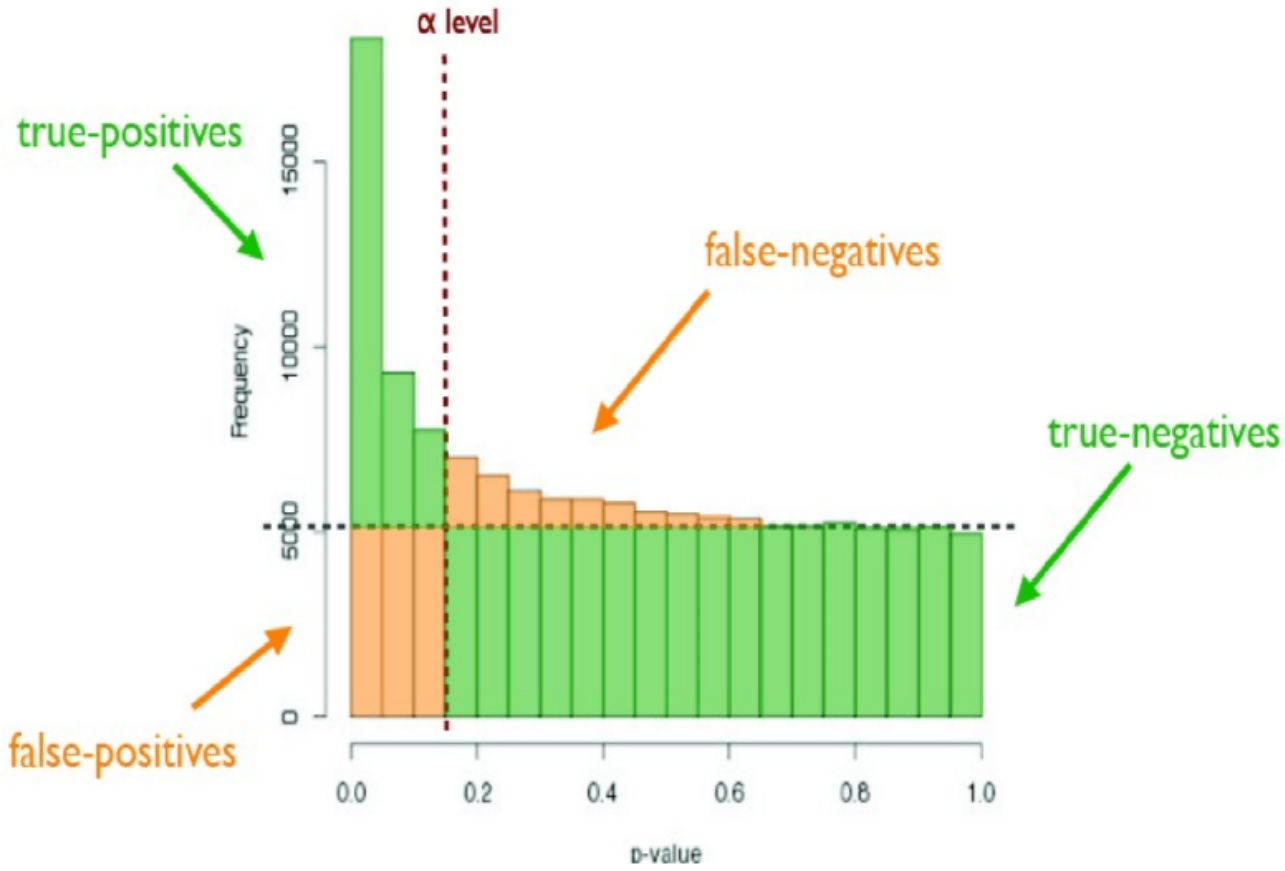
Let assume all the G genes are not DE.
Each test is realized at $\alpha$ level
Ex : $G = 10000$ genes and $\alpha = 0.05 \rightarrow E(FP) = 500$ genes.

# 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 and Hochberg, 1995] is the expected proportion of Type I errors among the rejected hypotheses

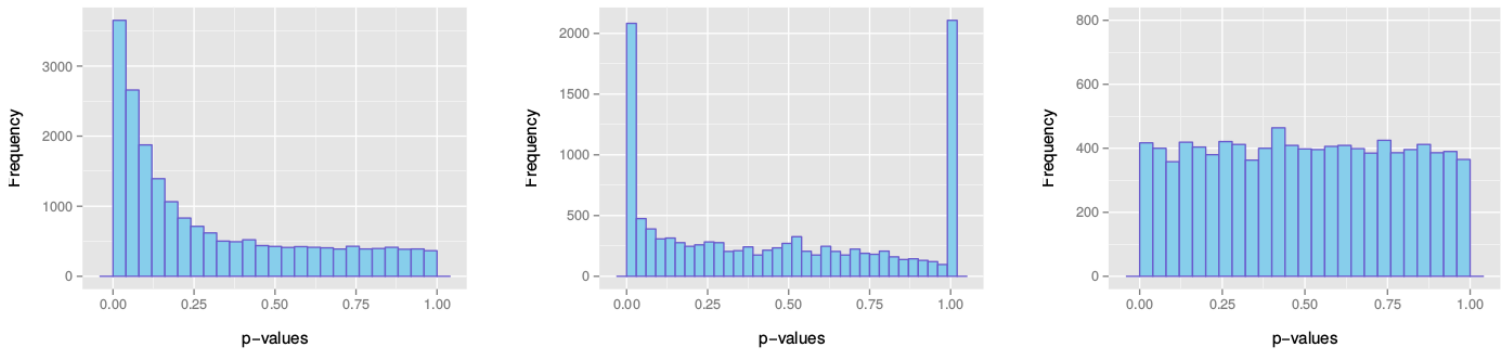
$$\text{FDR} = \mathbb{E}(FP/P) \text{ if } P > 0 \text{ and } 0 \text{ if } P = 0$$

**Prop**

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

# p-values histograms for diagnosis

Examples of expected overall distribution



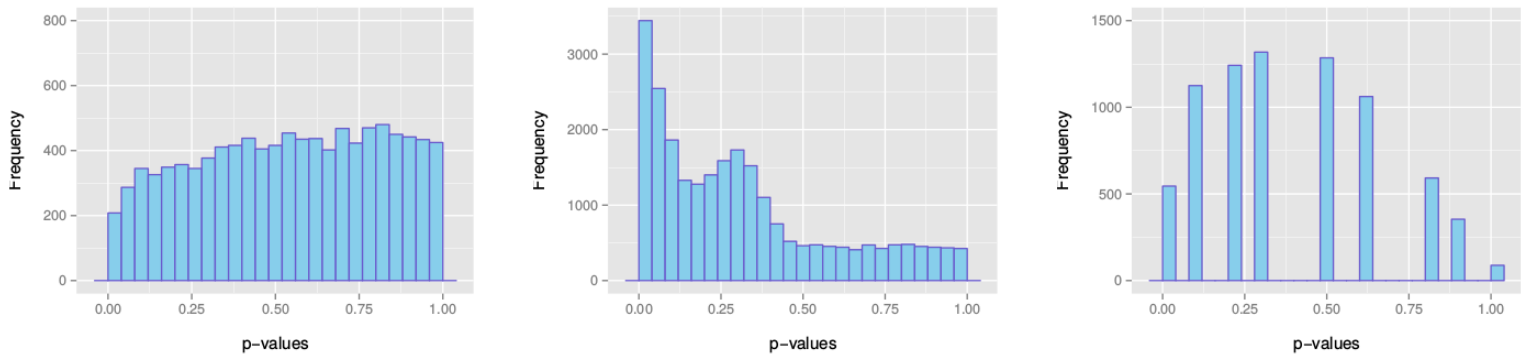

(a) : the most desirable shape

(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 :**

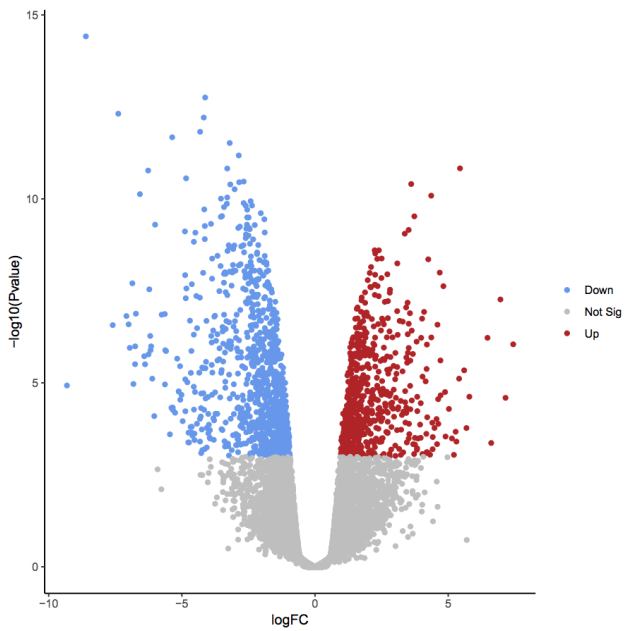
Having a great confidence on the DE elements (strong control). Accepting to not detect some elements (lack of sensitivity $\Leftrightarrow$ a few DE elements)

**Controlling the FDR :**

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

# 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"



# Plan

1 Experimental design

2 Exploratory data analysis

3 Normalization

4 Differential analysis

5 Multiple testing

6 Gene Set Enrichment Analysis

# Gene Set Enrichment Analysis

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`

# Gene Set Enrichment Analysis

## 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 test 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 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`

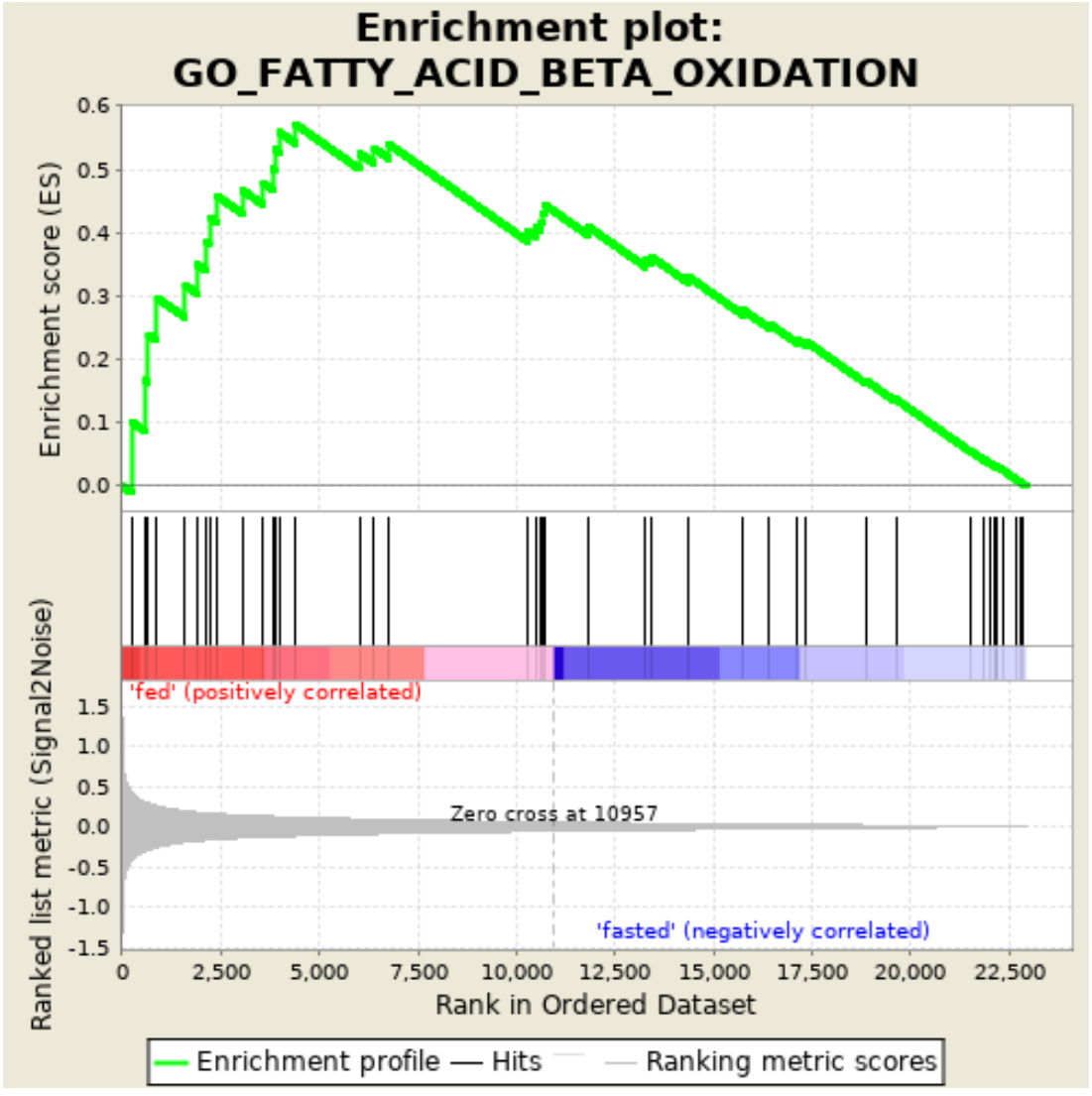
# GSEA

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)

# Conclusions

- Methods dedicated to microarrays are not directly applicable to RNA-seq
- Normalization depends on the statistical question
- Include at least 3 replicates per condition for differential analysis
- Large number of replicates (10 or so) or very high expression $\rightarrow$ method choice of differential analysis does not matter much.
- Removing genes with outlier counts or using non-parametric methods reduce the sensitivity to outliers
- Don't forget to correct for multiple testing !

# Conclusions

## Adapt the method to your data

Specific methods have been developped for few replicates.
The need for 'sophisticated' methods decreases when the number of replicates increases.

GSEA helps finding differentially expressed genes when not enough replicates were present in the initial study. Avoid merging the data when a high study effect is expected, prefer an appropriate statistical analysis !

# Want to go further ?

To practice more : Galaxy permanences
`https://wikis.univ-lille.fr/bilille/permanences`

To obtain help in statistical analysis of omic data :
bilille call for projects (around december each year, to plan the calendar of engineers)


Anders S and Huber W.
Differential expression analysis for sequence count data.
Genome Biology 2010, 11 :R106.


Anders S, Reyes A and Huber W
Detecting differential usage of exons from RNA-seq data
Genome Research 2012 :22


Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W and Robinson MD
Count-based differential expression analysis of RNA sequencing data using R and Bioconductor
Nature Protocols 2013, 8, 1765-1786


Anders A
Comparative analysis of RNA-seq data with DESeq and DEXseq
http ://www.bioconductor.org/help/course-materials/2013/CSAMA2013/tuesday/morning/Anders_DESeq_DEXSeq.pdf


Benjamini Y and Hochberg Y
Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing
Journal of the Royal Statistical Society, 1995, 57 :1, 289–300


Benjamini Y and Speed TP
Summurizing and correcting the GC content bias in high throughput sequencing
Nucleic Acids Research, 2012, 1-14.


Bolstad BM, Irizarry RA, Astrand M, and Speed TP
A comparison of normalization methods for high density oligonucleotide array data based on bias and variance.
Bioinformatics 19, 185-193, 2003.


Bullard JH, Purdom E, Hansen KD, Dudoit S.
Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments.
BMC Bioinformatics 2010, 11 :94


Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT

Scotty : a web tool for designing RNA-Seq experiments to measure differential gene expression.
Bioinformatics 2013, 29(5),656 :657.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C et al

A comprehensive comparison of normalization methods for Illumina high-throughput RNA-sequencing data analysis
Briefings in Bioinformatics 2013, 14 :6, 671-683.

Dudoit S, Maya O and Jacob L.

Short course on RNA seq and CHiP seq data analysis.
Valencia, Nov. 2010.

Eisenberg EE and Levanon EY.

Human housekeeping genes are compact.
Trends Genet, 19(7) :362-365.

Fisher RA

The Design of experiments
Oliver and Boyd 1935, 1-252

Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J

How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes ?
BMC genomics 2012, 1 (13),734.

Hansen KD, Brenner SE, Dudoit S.

Biases in Illumina transcriptome sequencing caused by random hexamer priming.
Nucleic Acids Research, 2010, 1-7.

Hansen KD, Irizarry RA and Wu Z

Removing technical variability in RNA-seq data using Conditional Quantile Normalization
Biostatistics 2011, 13 :2, pp204-216

Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P

Calculating Sample Size Estimates for RNA Sequencing Data.
Journal of Computational Biology 2013, 12(20), 970 :978

Kvam V, Liu P

A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data
American Journal of Botany 2012 99(2), 248-256

Liu Y, Zhou J, White K

RNA-seq differential expression studies : more sequence or more replication ?
Bioinformatics 2014, 30(3),301 :304.

Love MI, Huber W and Anders S

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.
Genome Biology 2014, 15 :550.

Marioni JC, Mason CE et al.

RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays.
Genome Research 2008, 18 : 1509-1517

Marot G, Foulley JL, Mayer CD, Jaffrézic F.

Moderated effect size and P-value combinations for microarray meta-analyses.
Bioinformatics 2009, 25(20) :2692-9.

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV

RNA-seq : technical variability and sampling
BMC Genomics 2011, 12 :293.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.

Mapping and quantifying mammalian transcriptomes by RNA-seq.
Nature Methods, 2008 Jul ; 5(7) ; 621-628

Oshlack A and Wakefield MJ

Transcript length bias in RNA-seq confounds systems biology
Biology Direct 2009, 4 :14.

Pachter L

Models for transcript quantification from RNA-seq

eprint 2011 arXiv :1104.3889

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK.
Understanding mechanisms underlying human gene expression variation with RNA sequencing.
Nature letters, 2010, vol 464.

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D
Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data
Genome Biology 2013 ,14 :R95

Rau A, Marot G, Jaffrézic F
Differential meta-analysis of RNA-seq data from multiple studies.
BMC Bioinformatics 2014 ,15 :91

Risso D, Schwartz K, Sherlock G, Dudoit S.
GC-content normalization for RNA-Seq data
BMC Bioinformatics 2011, 17, 12 :480

Risso D, Ngai J, Speed T and Dudoit S
Normalization of RNA-seq data using factor analysis of control genes or samples
Nature Biotechnology 2014, 32 (9), 896-905

Robinson MD and Smyth, GK.
Moderated statistical tests for assessing differences in tag abundance.
Bioinformatics 23(21) ; 2881-2887

Robinson MD and Smyth, GK.
Small-sample estimation of negative binomial dispersion, with applications to SAGE data
Biostatistics (2008), 9, 2 ; 321-332

Robinson MD, McCarthy DJ, Smyth, GK.
edgeR : a Bioconductor package for differential expression analysis of digital gene expression data.
Bioinformatics 2009


Robinson MD, Oshlack A.
A scaling normalization method for differential expression analysis of RNA-seq data.
Genome Biology 2010, 11 :R25


Robles J.A., Qureshi S.E., Stephen S.J., Wilson S.R., Burden C.J., Taylor J.M.
Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing
BMC Genomics 2012, 13 :484


Schulze SK, Kanwar R, Gölzenleuchter M, Therneau TM, Beutler AS.
SERE : Single-parameter quality control and sample comparison for RNA-Seq
BMC Genomics 2012, 13 :524


Soneson C and Delorenzi M
A comparison of methods for differential expression analysis of RNA-seq data
BMC Bioinformatics 2013, 14 :91


Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB.
A gene atlas of the mouse and human protein-encoding transcriptomes.
Proc. Natl. Acad. Sci. USA, 101(16) :6062-6067.


A. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.
Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles.
PNAS 2005, 102(43) :15545–15550.


Trapnell C, Hendrickson D, Sauvageau M, Goff L, Rinn J and Pachter L
Differential analysis of gene regulation at transcript resolution with RNA-seq
Nature Biotechnology 2013, 31, 1


Varet H, Brillet-Gueguen L, Coppée JY, Dillies MA
SARTools : A DESeq2- and edgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data

PLoS One 2016, 11(6) :e0157022

Wu H, Wang C, Wu Z.

PROPER : comprehensive power evaluation for differential expression using RNA-seq.
Bioinformatics 2015, 15 ;31(2) :233-41

Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A.,

Gene ontology analysis for RNA-seq : accounting for selection bias
Genome Biology, 11, 2, Feb 2010, R14

Zhou X, Lindsay H and Robinson MD

Robustly detecting differential expression in RNA sequencing data using observation weights.
Nucl. Acids Res. 2014, 42 (11) : e91