Differential analysis of RNA-Seq data: design, describe, explore and model

Ecole de Bioinformatique IFB/Inserm – Roscoff – Nov. 2024

Charlotte Berthelier – cberthelier@sb-roscoff.fr
Elise Jacquemet – elise.jacquemet@pasteur.fr

Bioinformatics and Biostatistics Hub - Institut Pasteur - Université de Paris, Paris, France

CNRS • SORBONNE UNIVERSITÉ Station Biologique de Roscoff 1872

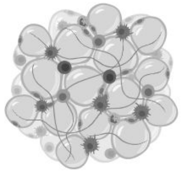
ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Inserm

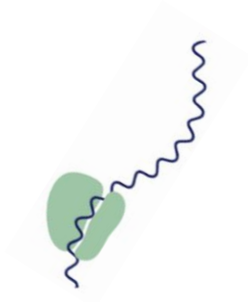
Main RNA-Seq steps

Designing the experiment and the sequencing

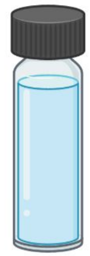
Tissue/cells



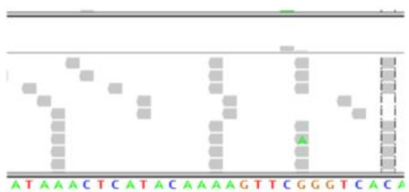
RNA



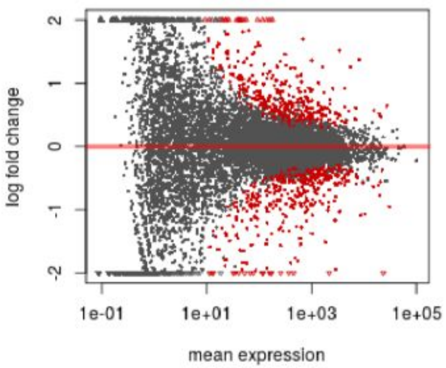
cDNA Library



Sequencing



Bioinformatics



Differential analysis

INSTITUT PASTEUR

Citations

"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, Atelier Inserm 145, 2003

Statistical modeling

Goal of an experiment: address **one** biological question

Result of an experiment: many numerical values

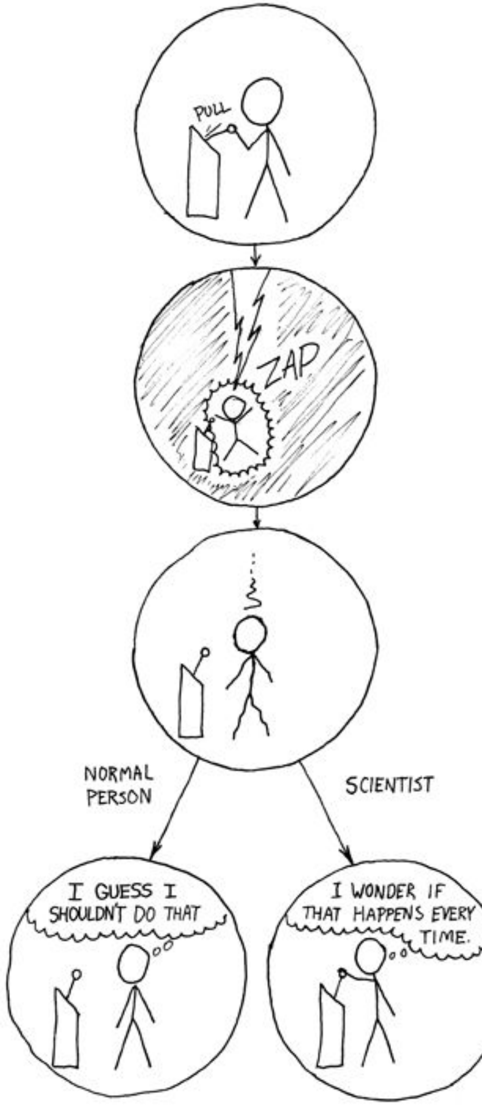
Statistical modeling consists in using a mathematical formula involving:

- Experimental conditions X
- Numerical values measured Y
- Parameters β linking X and Y (to be estimated), e.g.:

$$Y \sim X\beta + \varepsilon$$

- Some hypotheses on the data variability/law, e.g.:

$$\varepsilon \sim \text{Gaussian}(0, \sigma^2)$$



INSTITUT PASTEUR

Starting point of the differential analysis

	T0-1	T0-2	T0-3	T4-1	T4-2	T4-3	T8-1	T8-2	T8-3
gene1	151	131	183	31	35	44	19	31	18
gene2	142	134	153	650	629	783	136	241	151
gene3	157	147	166	7	10	20	8	10	8
gene4	275	249	342	70	44	91	75	64	62
gene5	4	5	2	0	0	1	2	2	3
gene6	2	0	1	0	1	2	7	3	3
gene7	4	7	3	0	0	0	0	0	0
gene8	10	16	10	28	12	10	16	33	23
gene9	12	20	24	74	84	77	10	10	9
gene10	269	262	379	112	132	138	44	33	48
gene11	10065	9593	11955	4076	3739	4137	2736	3311	2749
gene12	651	566	819	101	86	74	97	87	96
gene13	118	116	150	18	24	42	15	8	5
...	...	...	...	...	...	...	...	...	...
geneN	18	31	39	4	4	7	2	6	2

Goal: find **genes** differentially expressed between biological conditions

Outline

1. Introduction
2. **Designing the experiment**
3. Description/exploration
4. Normalization
5. Modeling
6. SARTools

Why an experimental design?

To control the variability during the experiment in order to be able to address the biological question:

1. What is the biological question?
2. How to estimate the associated biological variabilities?
3. How to control the technical variabilities (day, lane, run, etc.)?

Biological or technical uncontrolled effects could:

- Hide/cancel the biological effect of interest
- Wrongly increase the biological effect of interest

"Ensure that the right type of data, and enough of it, is available to answer the questions of interest as clearly and efficiently as possible"

http://www.stats.gla.ac.uk/steps/glossary/anova.html#expdes

INSTITUT PASTEUR

Why an experimental design?

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDITORIAL

Ten simple rules for providing effective bioinformatics research support

Judit Kumuthini, Michael Chimenti, Sven Nahnsen, Alexander Peltzer, Rebone Meraba, Ross McFadyen, Gordon Wells, Deanne Taylor, Mark Maienschein-Cline, Jian-Liang Li, Jyothi Thimmapuram, Radha Murthy-Karuturi, Lyndon Zass

Published: March 26, 2020 • https://doi.org/10.1371/journal.pcbi.1007531

"A good experimental design starts with a **well-defined hypothesis** [...]. The experimental design should aim to reduce the types and sources of variability, increase the **generalizability** of the experiment, and make it **replicable** and reusable. It is both easier and more cost efficient to identify and correct experimental design issues ahead of time than to address deficiencies thereafter. Thus, discussion between data-generating researchers and bioinformaticians is highly desirable and should occur as early as possible during project development and experimental design."

INSTITUT PASTEUR

Basic comparison

I want to study differences in the transcriptome of cystic fibrosis patients

id	state
h1	healthy
h2	healthy
h3	healthy
cf1	CF
cf2	CF
cf3	CF

- one **factor** of interest : the state of the patients
- this factor has two **levels**: healthy and CF

→ **mRNA sequencing of lung cells.**

Paired samples

I want to study differences in the transcriptome of cystic fibrosis patients

id	state	RNA extraction date
h1	healthy	June 12th, 2019
h2	healthy	June 20th, 2019
h3	healthy	June 25th, 2019
cf1	CF	June 12th, 2019
cf2	CF	June 20th, 2019
cf3	CF	June 25th, 2019

On the laboratory bench...

Time course experiment (paired)

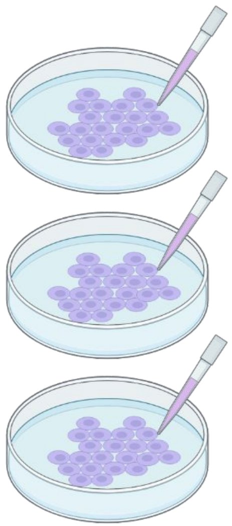
I want to find differentially expressed genes between time 0 and time 24h on cultures of E. Coli

Time 0h

Sample 1

Sample 2

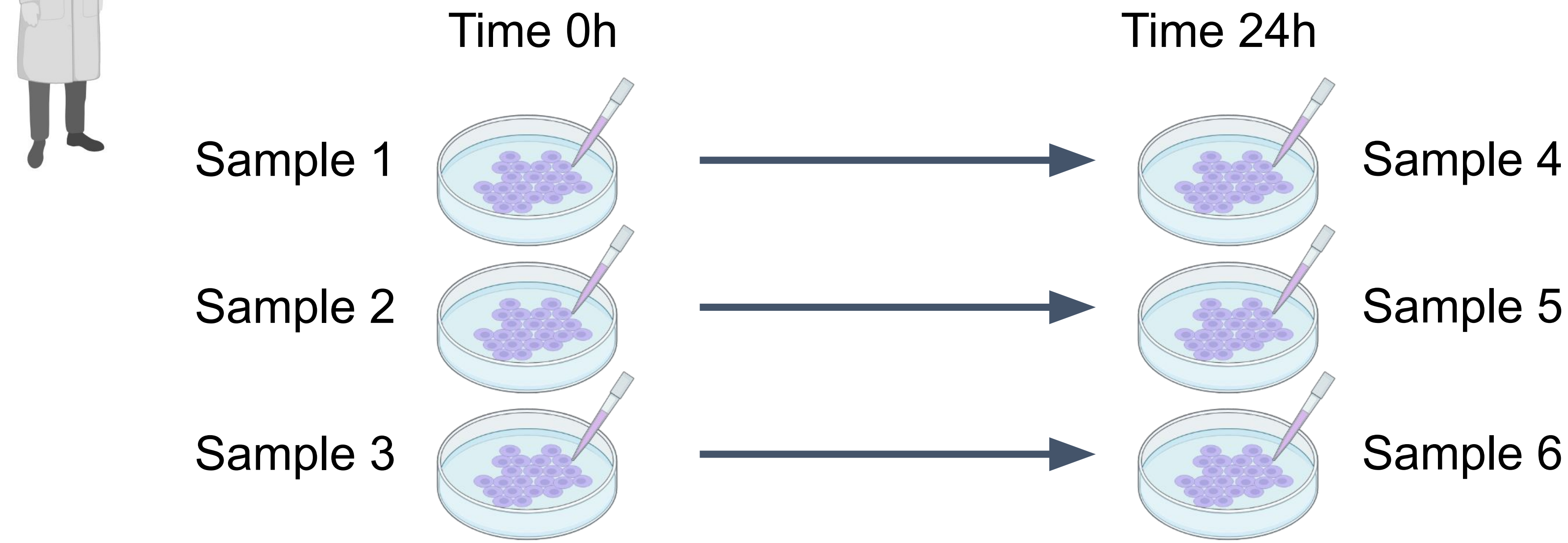
Sample 3



On the laboratory bench...

Time course experiment (paired)

I want to find differentially expressed genes between time 0 and time 24h on cultures of E. Coli



On the laboratory bench...

Time course experiment (unpaired)

I want to find differentially expressed genes between time 0 and time 24h on cultures of E. Coli

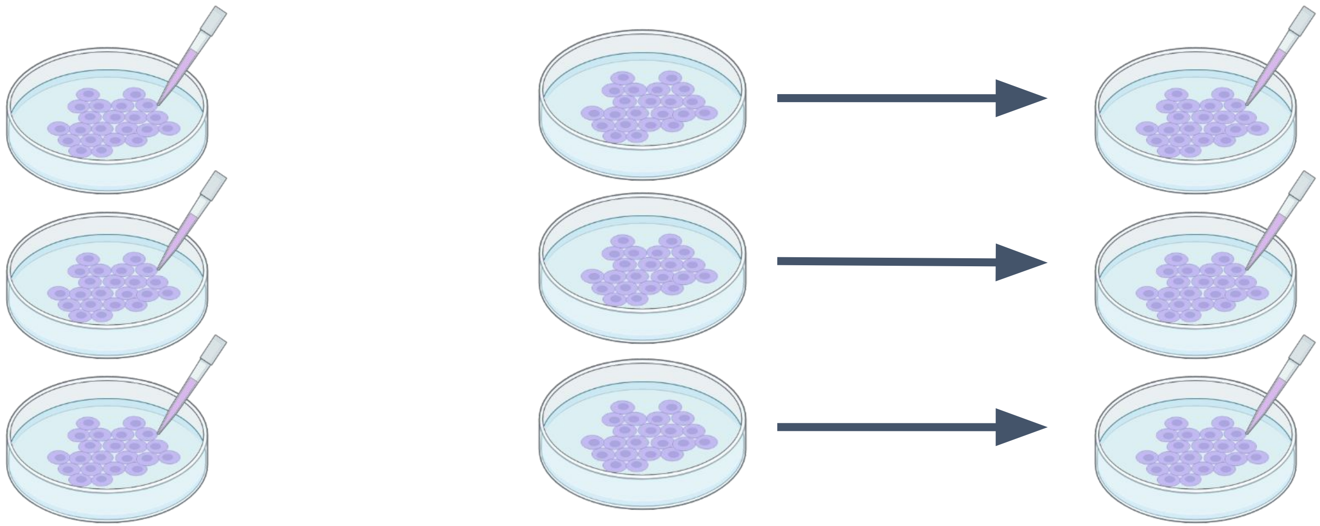
Time 0h

Time 24h

Sample 1

Sample 2

Sample 3



Sample 4

Sample 5

Sample 6

Complex design

I want to study the effect of a virus infection level (high vs. low) on the transcriptome of two mouse strains (B6 vs. SEG).

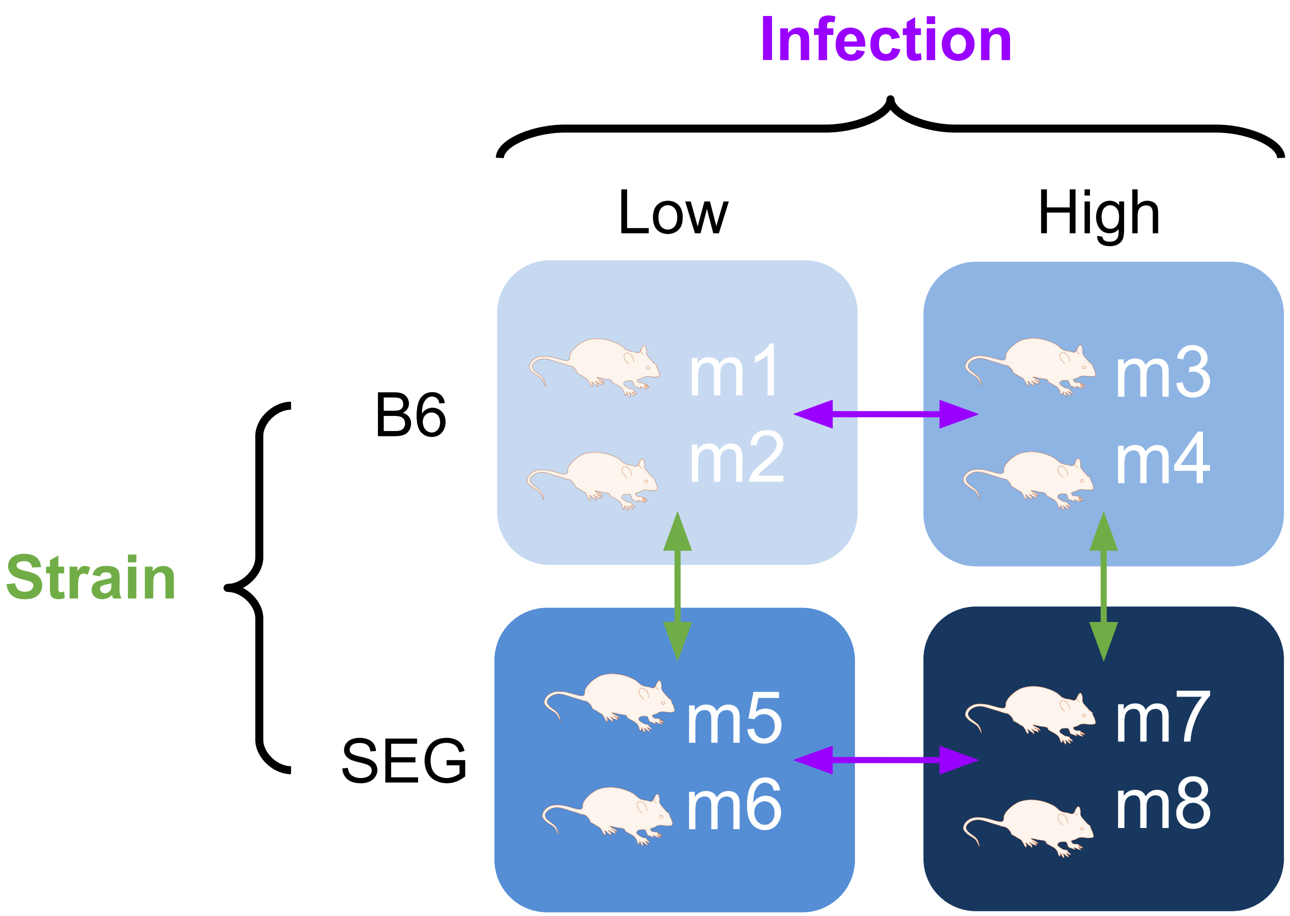
id	strain	infection
m1	B6	low
m2	B6	low
m3	B6	high
m4	B6	high
m5	SEG	low
m6	SEG	low
m7	SEG	high
m8	SEG	high

Two factors of interest with **two levels** each :

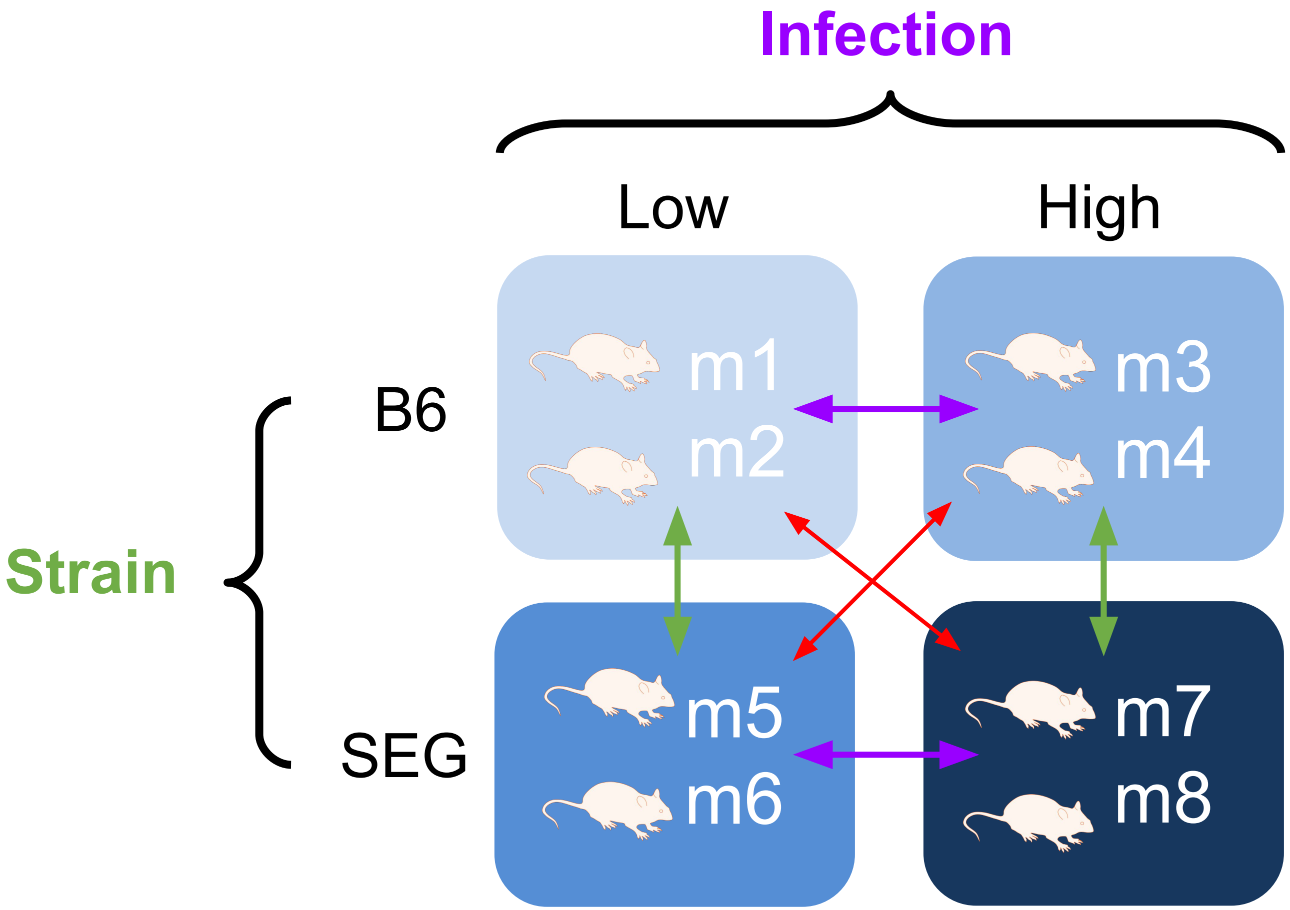
- the infection level of the patients (low or high)
- the mouse strain (SEG and B6)

INSTITUT PASTEUR

Interaction between two factors/variables



Interaction between two factors/variables

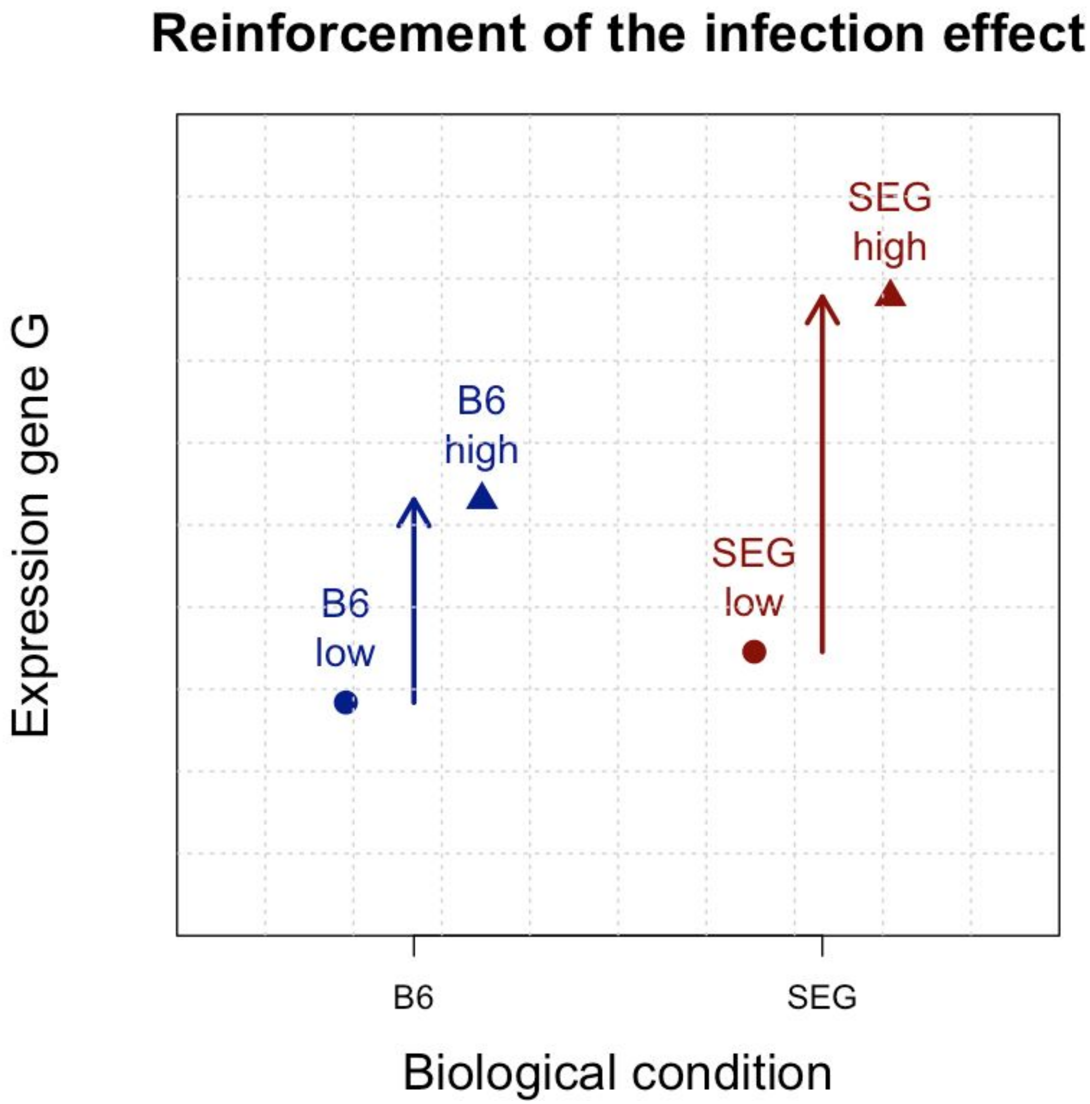


Interaction:

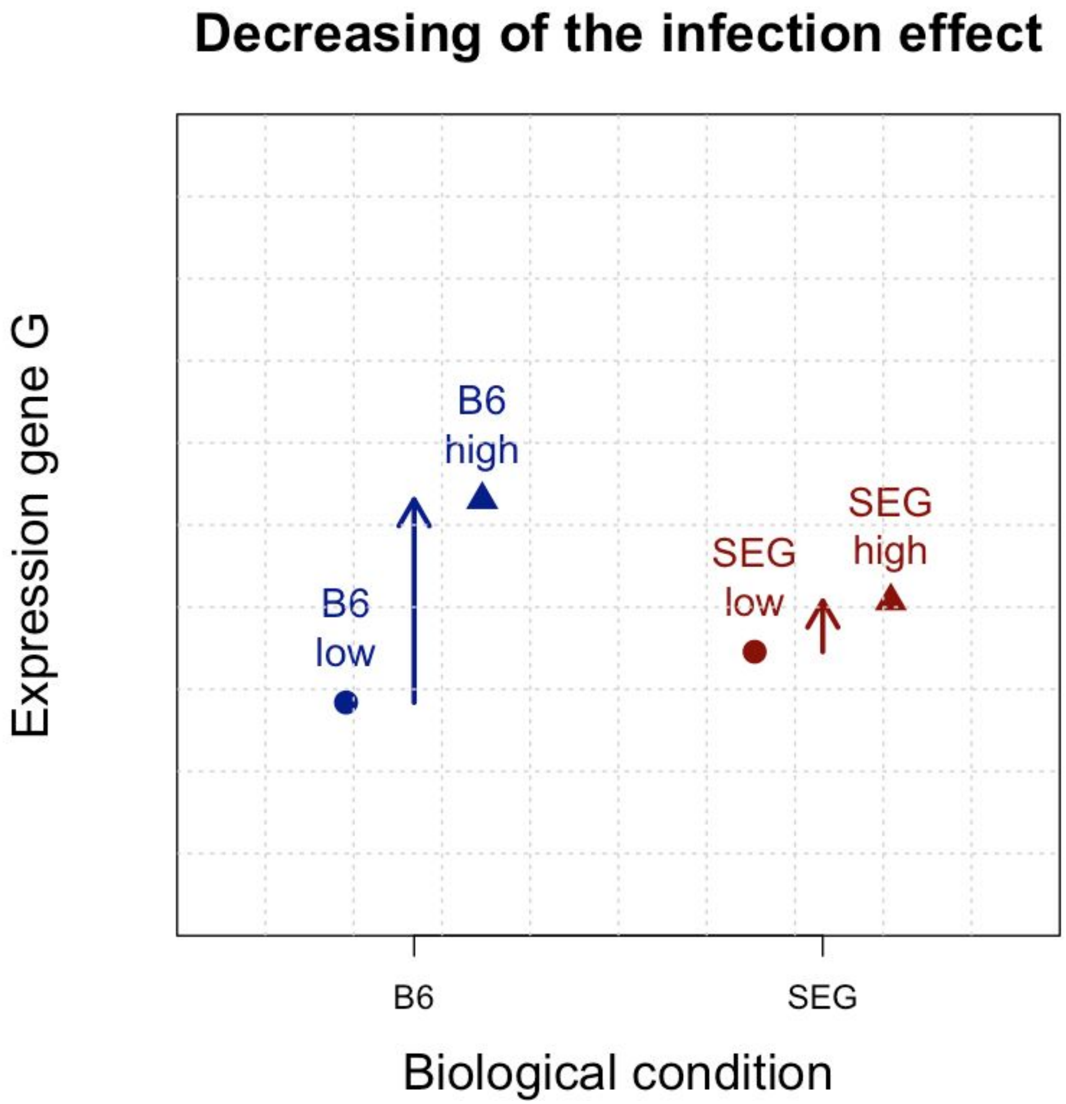
- Is the infection effect different between the two strains?
- Does the difference between the strains change according to the infection?

INSTITUT PASTEUR

Examples of interactions

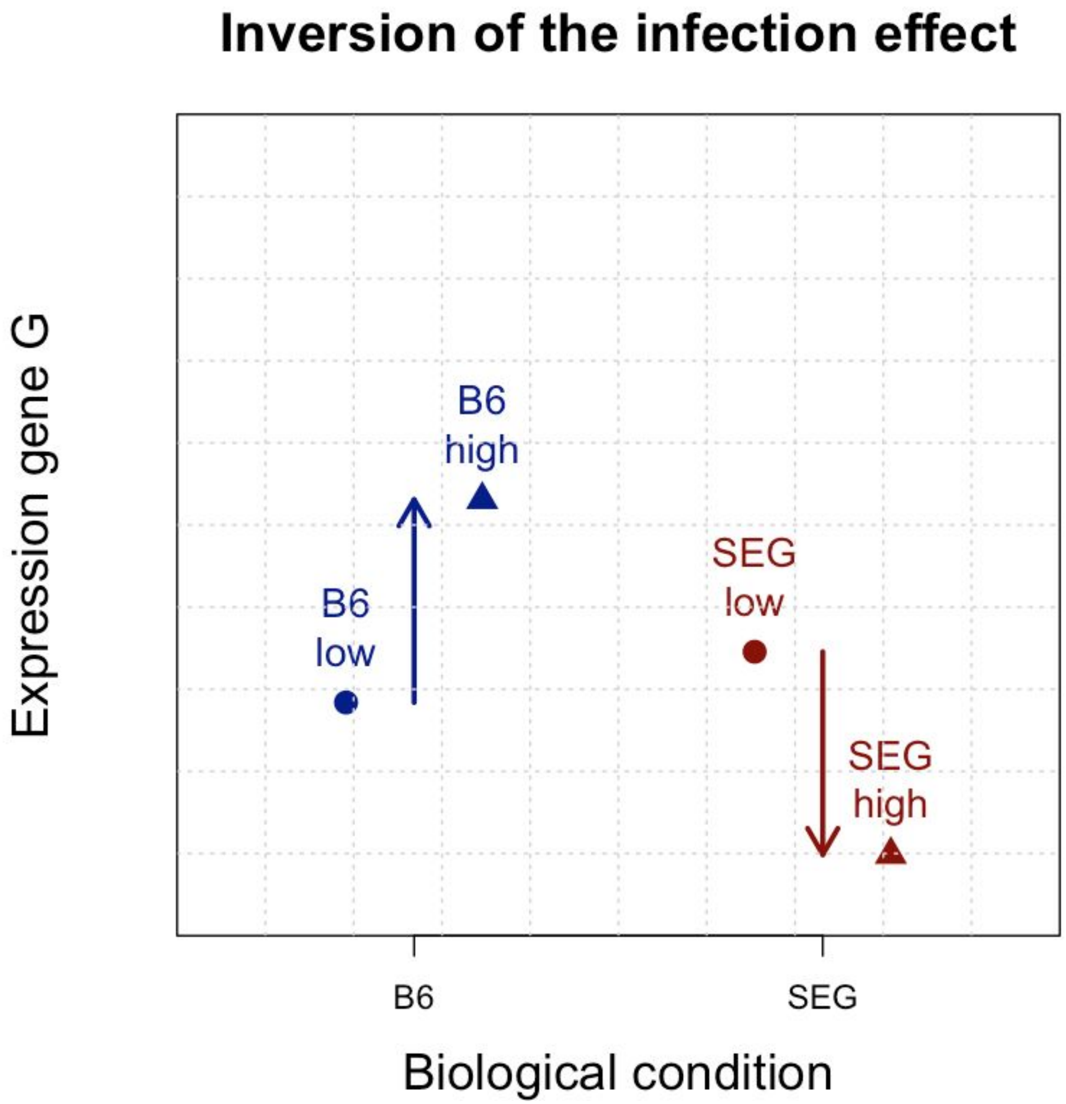


Examples of interactions



INSTITUT PASTEUR

Examples of interactions



Complex design with nested factors

A treatment T is applied to two CF patients and two healthy people. We study the initial transcriptome and after 4h of treatment.

id	state	time	patient
h1-0	healthy	0h	h1
h2-0	healthy	0h	h2
h1-4	healthy	4h	h1
h2-4	healthy	4h	h2
cf1-0	CF	0h	cf1
cf2-0	CF	0h	cf2
cf1-4	CF	4h	cf1
cf2-4	CF	4h	cf2

The "patient" effect need to be taken into account, but it is nested into the "state" effect.

Be careful with confounding effects !

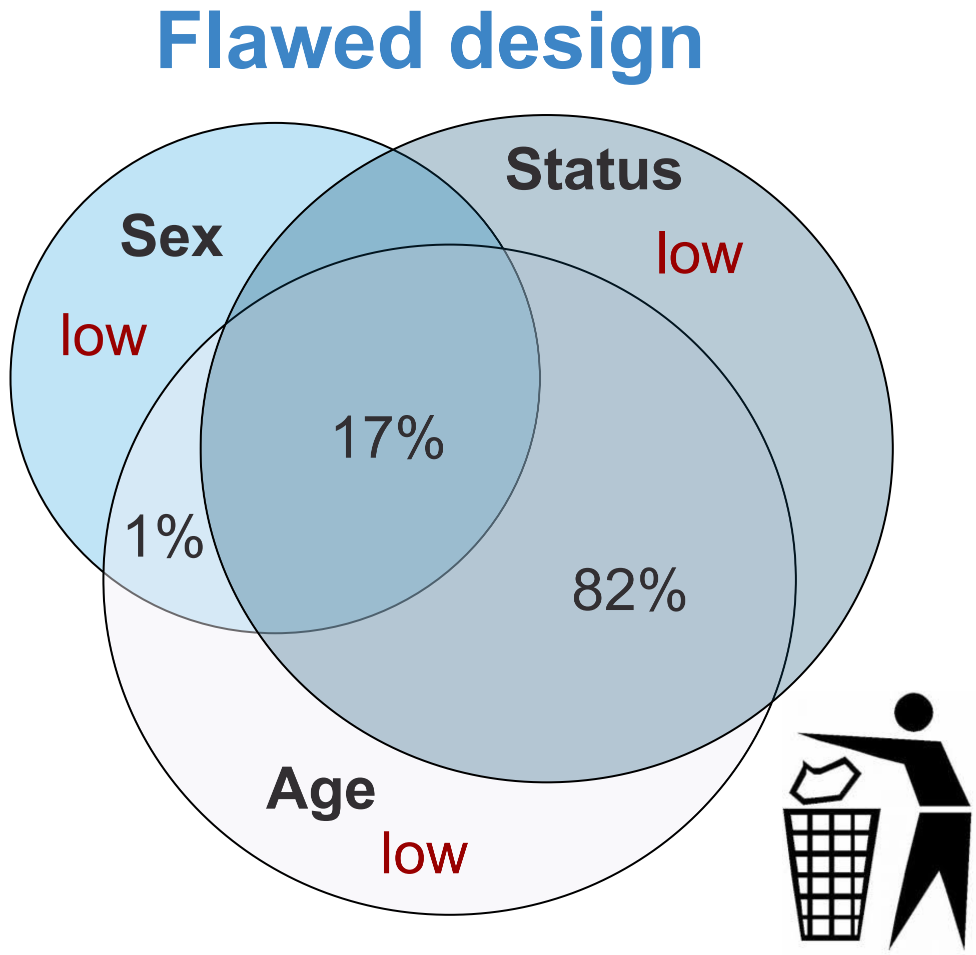
Comparison of lung cells in healthy and cystic fibrosis patients

id	state	age	sex	RNA extraction day	experimentalist
h1	healthy	45	female	July 9th, 2019	Louis
h2	healthy	52	female	July 12th, 2019	Louis
h3	healthy	48	female	July 15th, 2019	Louis
cf1	CF	31	male	Feb 20th, 2019	Françoise
cf2	CF	25	male	Feb 24th, 2019	Françoise
cf3	CF	27	male	Feb 29th, 2019	Françoise

Be careful with confounding effects !

A gene is detected as being differentially expressed between healthy and CF patients. Is it due to:

- The disease?
- The sex effect?
- The age effect?
- The date effect?
- The technician effect?

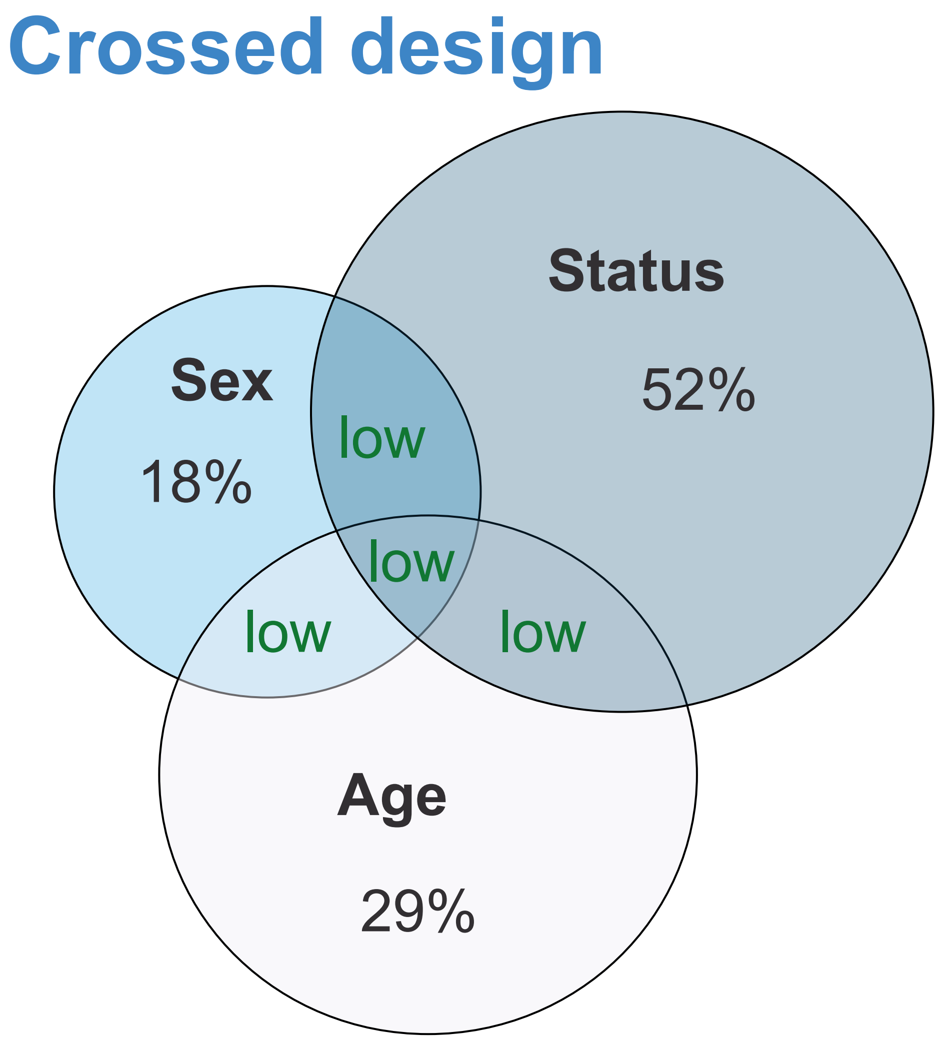


Be careful with confounding effects !

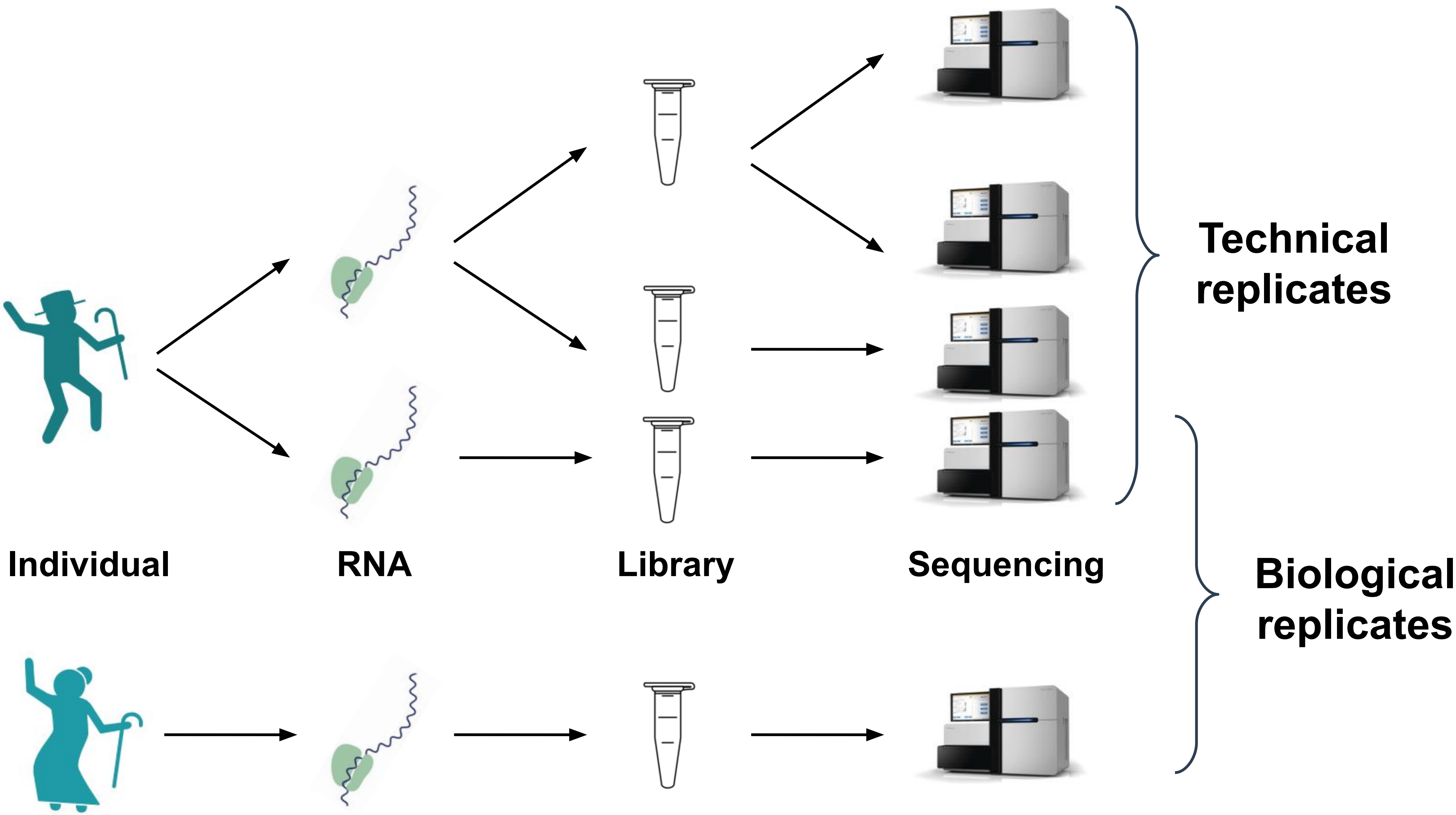
Re-doing the experiment but making sure all levels of all factors are **crossed** to avoid any confusion

Possibility to distinguish every source of variability & their interaction :

- The disease
- The sex effect
- The age effect
- The date effect
- The technician effect



Biological vs. technical replicates



Biological vs. technical replicates in RNAseq

Technical replicates:

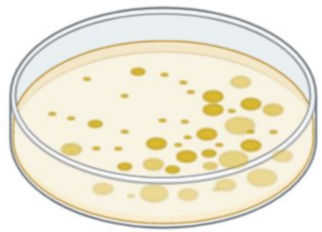
- Several extractions of the same RNA
- Several libraries built from the same RNA extraction
- A library sequenced several times

Allow to get more sequencing depth and a better coverage. Need to sum the counts associated to each technical replicates.

Biological replicates:

- Parallel measurements of biologically **distinct samples**
- Correspond to the variability visible in the real life

Comment: what happens when studying fungi/yeast?



Why replicate?

Perfect world:

No biological nor technical variability

↓

Only one sample from each condition to conclude!

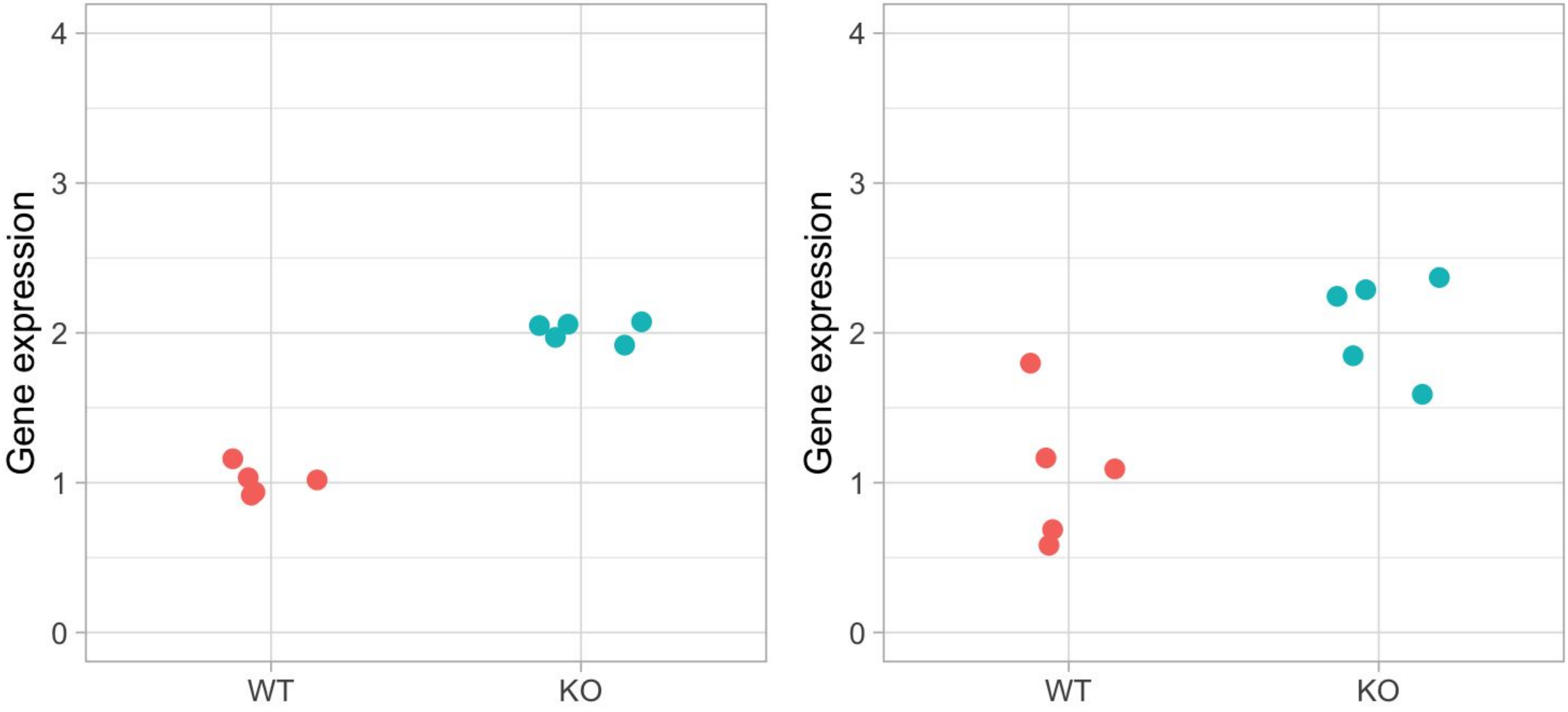
Our world:

Each individual has its own behavior

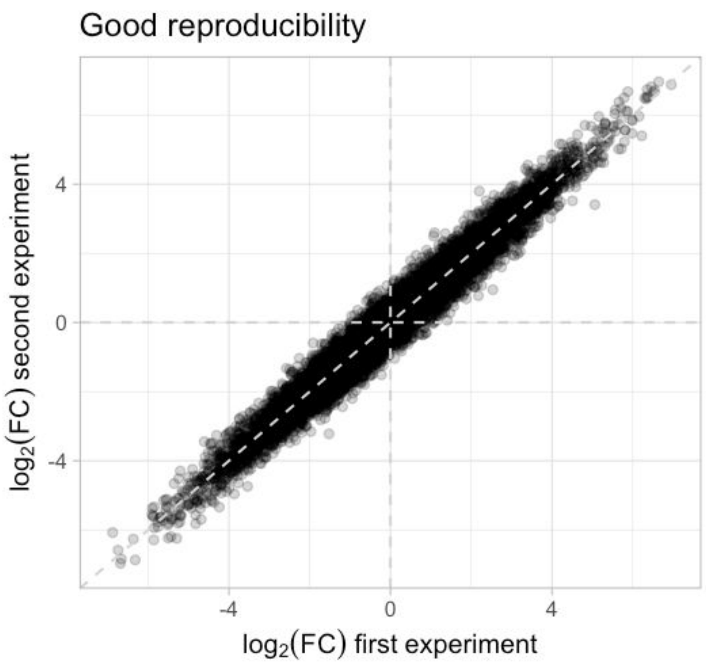
↓

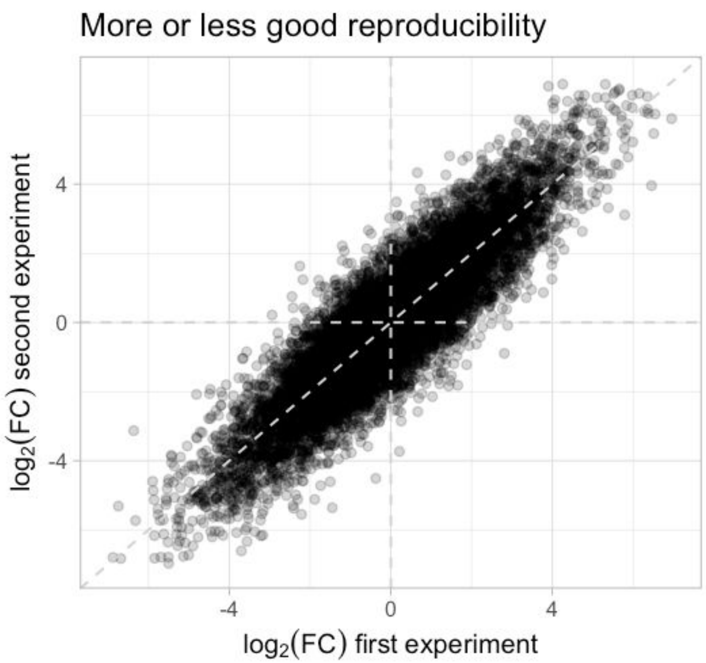
Need several biological replicates to handle variability

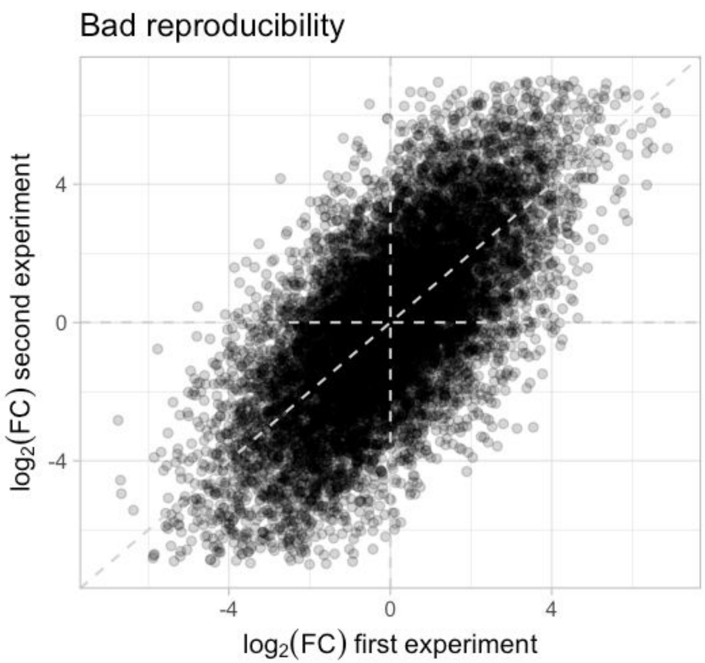
Why replicate?

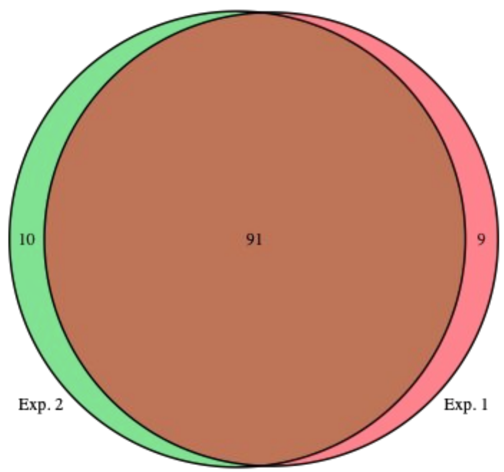


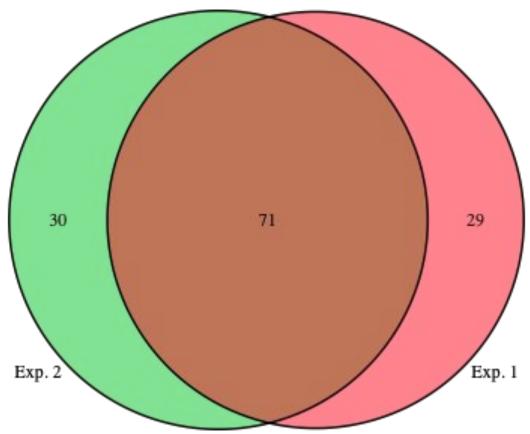
Reproducibility of an experiment: 3 KO vs 3 WT

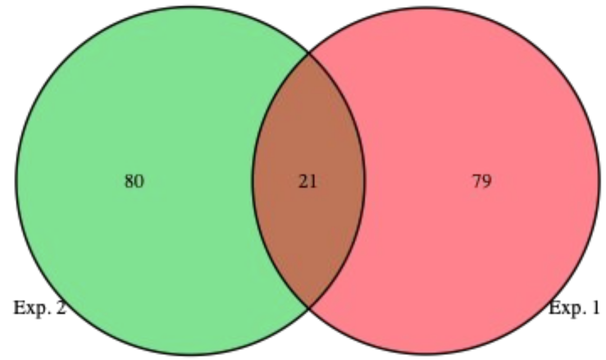






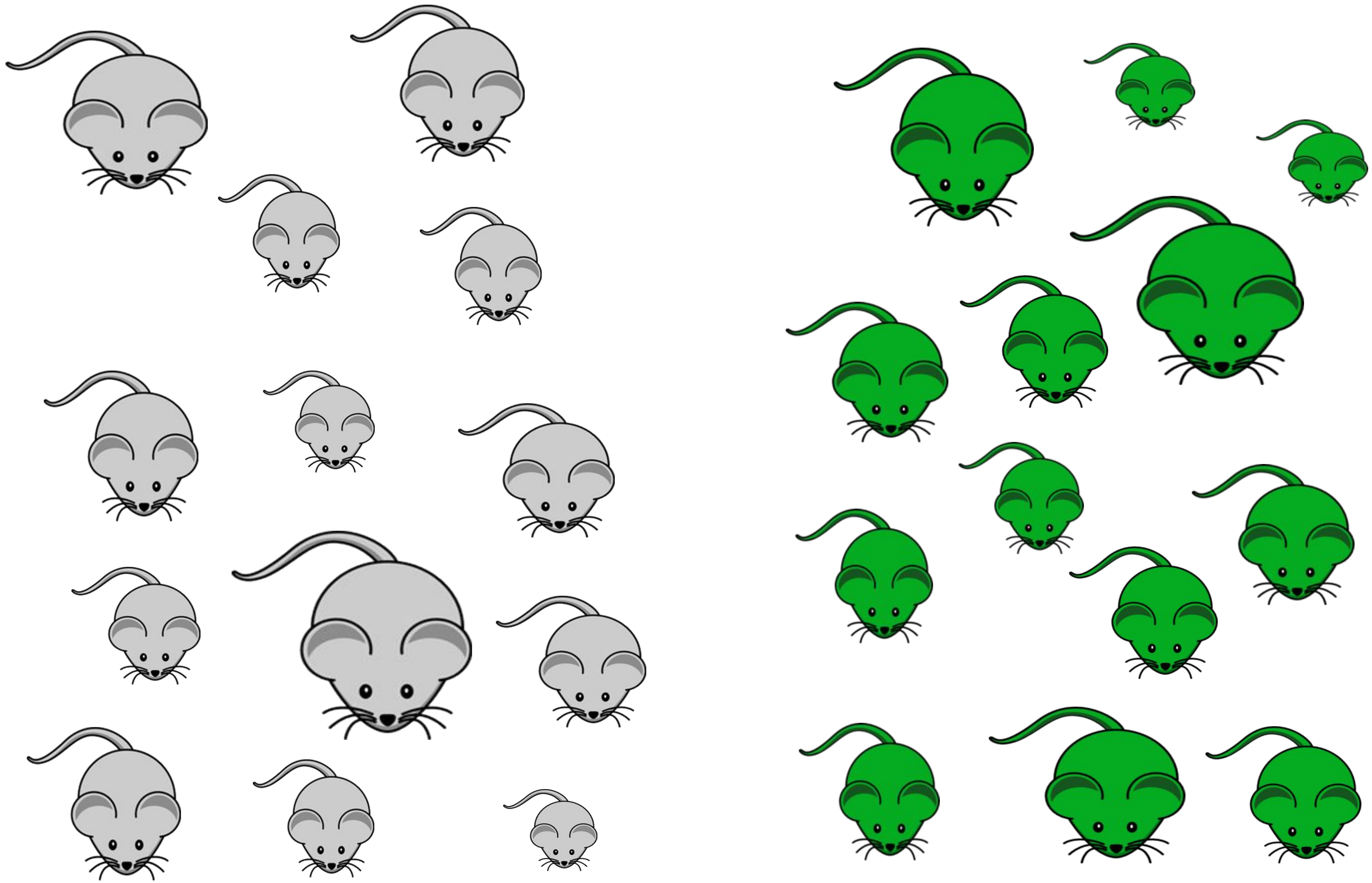




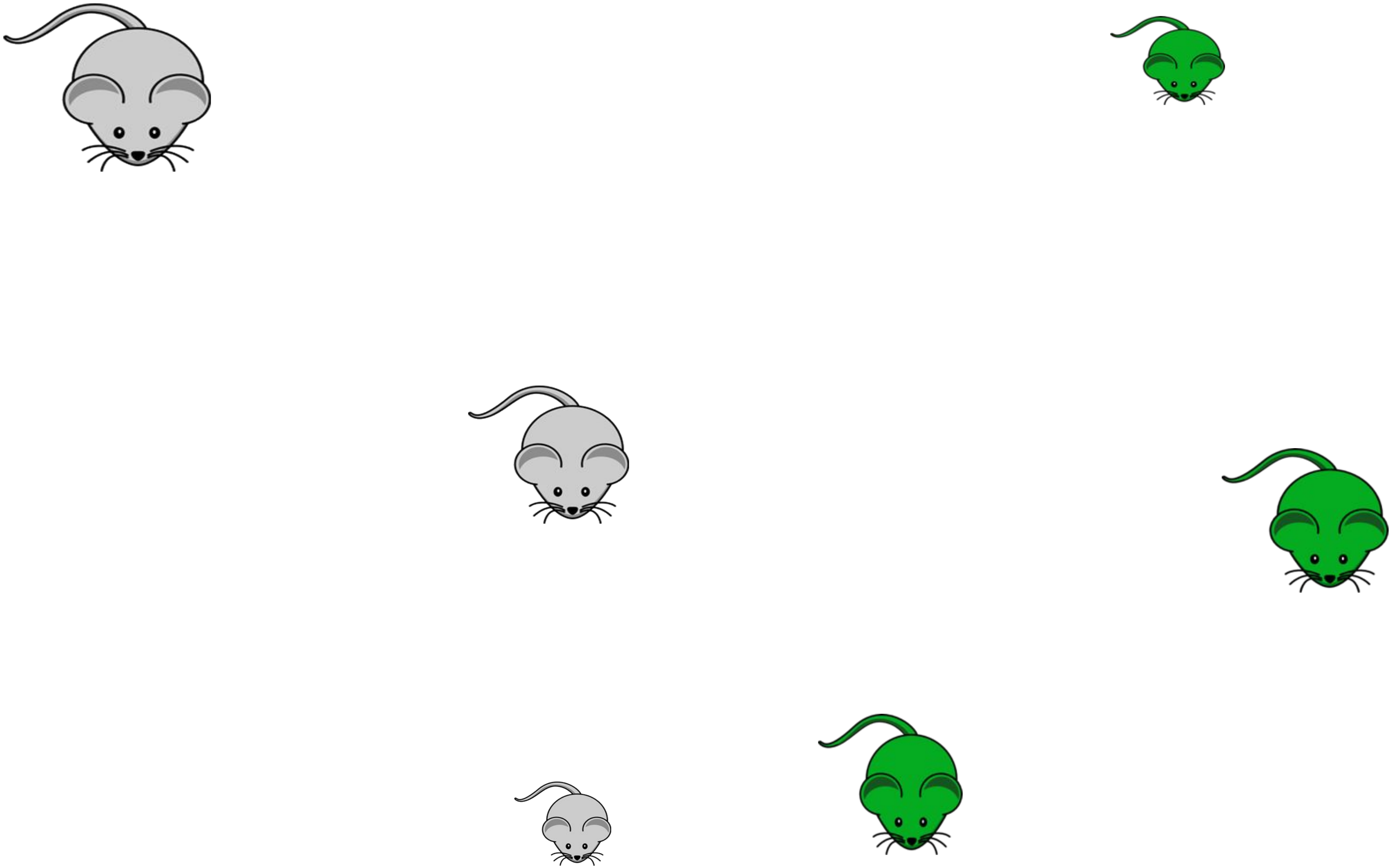


Population: set of all mice we could measure

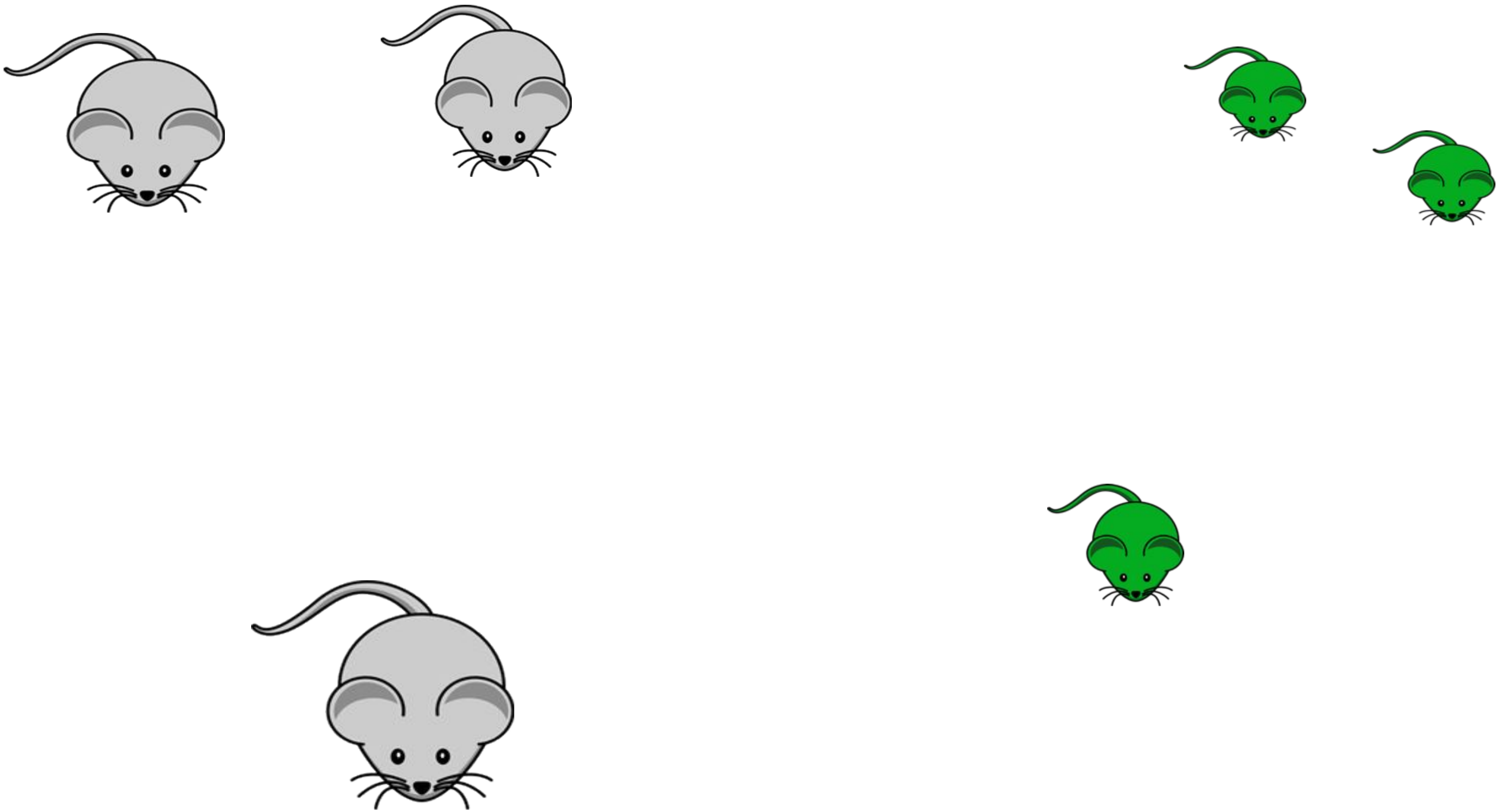
Sampling must be representative of the whole population under study !



Sampling 1: selection of 3 mice per condition

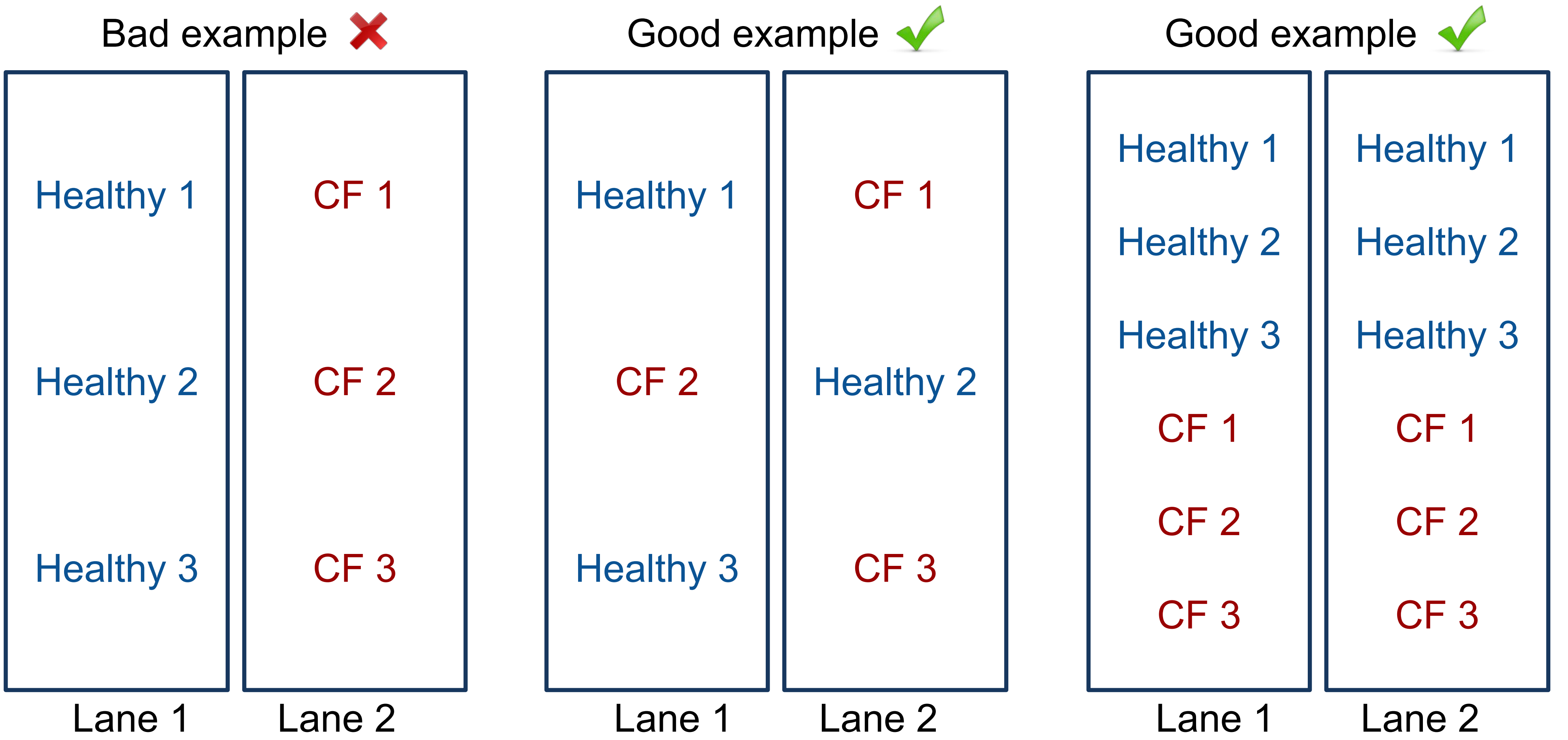


Sampling 2: non representative



Sequencing design

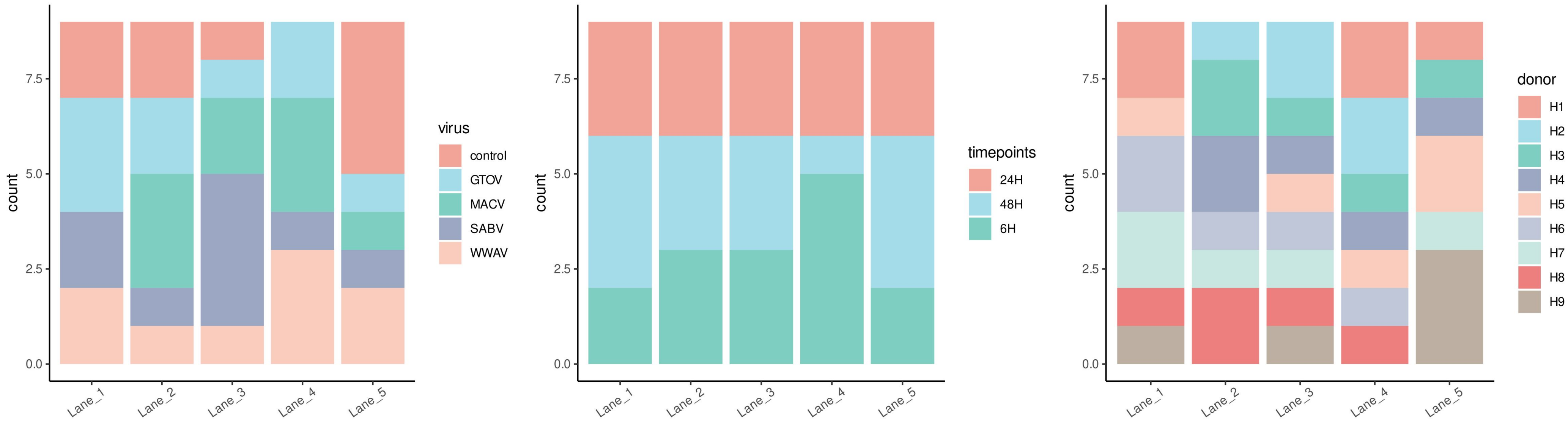
Goal:
Do not add any confounding technical effect (day, lane, run, etc.) to the factor of interest.



Sequencing design

Goal:
Do not add any confounding technical effect (day, lane, run, etc.) to the factor of interest.

Impossible to cross evenly all sources of technical variation → **Randomize !**



https://mixnpick.pasteur.fr/

Sequencing design

Technical variabilities:

- Lane
- Flowcell
- Run

lane effect < flowcell effect < run effect << biological variability

Use the same multiplexing rate for all the samples!

Experimental design : Take-home message

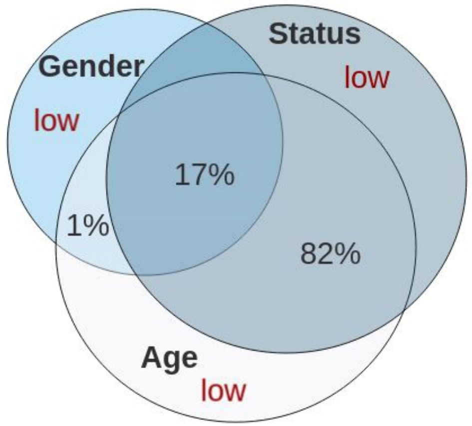
Express the biological question as accurately as possible to build an experimental design which will be able to address it.

The simpler, the better : If >2 factors, the results may be very difficult to interpret

Identify all the sources of variability to avoid confounding effects

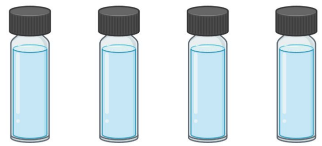
- Change of biological condition (e.g. KO vs WT)
- Within replicates variability (e.g. KO1 vs KO2 vs KO3)
- Experimentalist or day effect
- RNA: quality and extraction
- Library: PCR, concentration, random priming, rRNA removal
- Sequencing machine, flowcell and lane, …

Flawed design



Experimental design : Take-home message

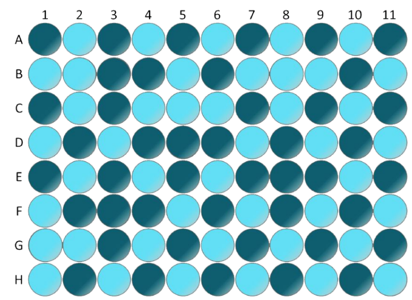
Experiments must be replicated to precisely measure the biological variability associated with the condition under study.



Sampling must be representative of the whole population under study

The higher the within group variability … the higher the number of biological replicates, in order to make sure that the whole range of variation is covered

Ideally, use blocking … to ensure that the biological conditions are evenly distributed among factors that are important unwanted) sources of variability.



… or randomization when blocking is not possible

INSTITUT PASTEUR

Outline

1. Introduction

2. Designing the experiment

3. **Description/exploration**

4. Normalization

5. Modeling

6. SARTools

Starting point of the differential analysis

	T0-1	T0-2	T0-3	T4-1	T4-2	T4-3	T8-1	T8-2	T8-3
gene1	151	131	183	31	35	44	19	31	18
gene2	142	134	153	650	629	783	136	241	151
gene3	157	147	166	7	10	20	8	10	8
gene4	275	249	342	70	44	91	75	64	62
gene5	4	5	2	0	0	1	2	2	3
gene6	2	0	1	0	1	2	7	3	3
gene7	4	7	3	0	0	0	0	0	0
gene8	10	16	10	28	12	10	16	33	23
gene9	12	20	24	74	84	77	10	10	9
gene10	269	262	379	112	132	138	44	33	48
gene11	10065	9593	11955	4076	3739	4137	2736	3311	2749
gene12	651	566	819	101	86	74	97	87	96
gene13	118	116	150	18	24	42	15	8	5
...	...	...	...	...	...	...	...	...	...
geneN	18	31	39	4	4	7	2	6	2

Goal: find **genes** differentially expressed between biological conditions

Many plots to produce

QC & diagnostics

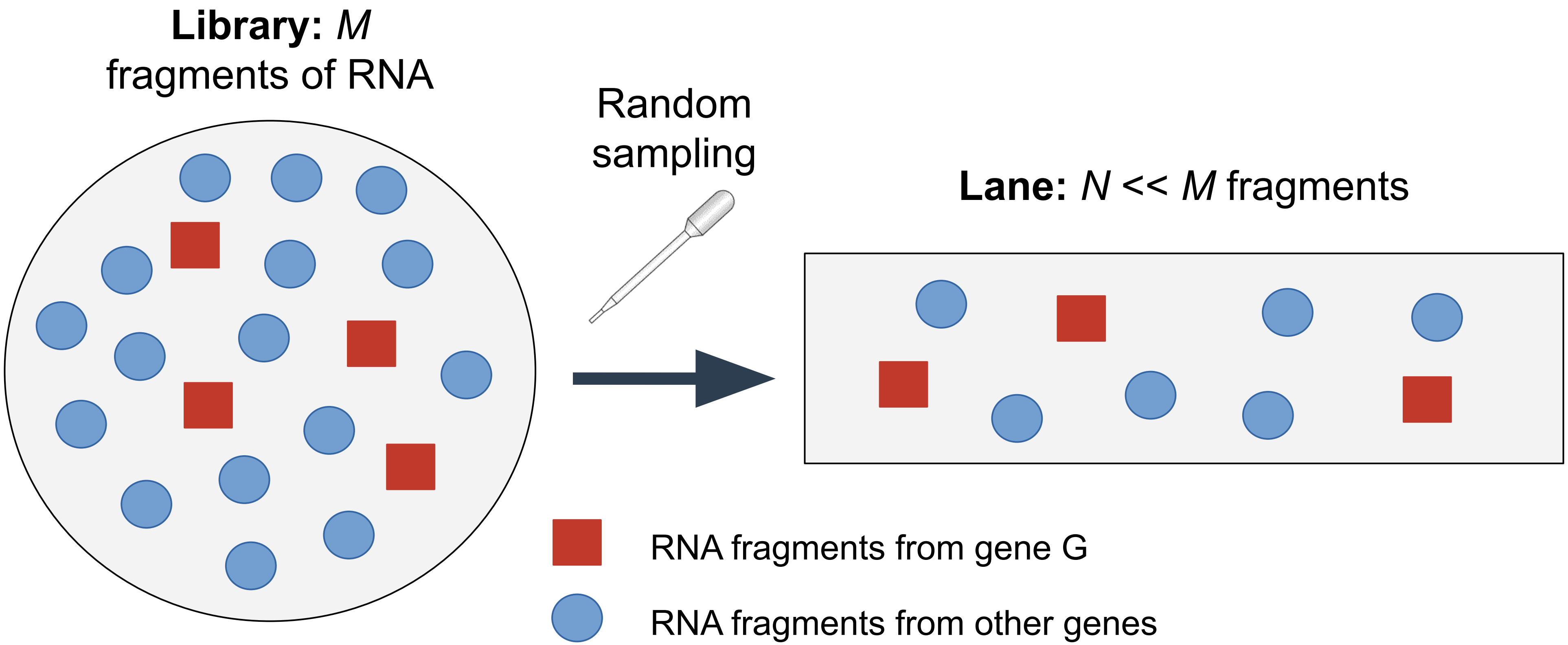
Description sample by sample:

- Total number of reads
- Percentage of null counts
- Percentage of reads caught by the most expressed gene
- Distribution of the counts

Multivariate description of the data:

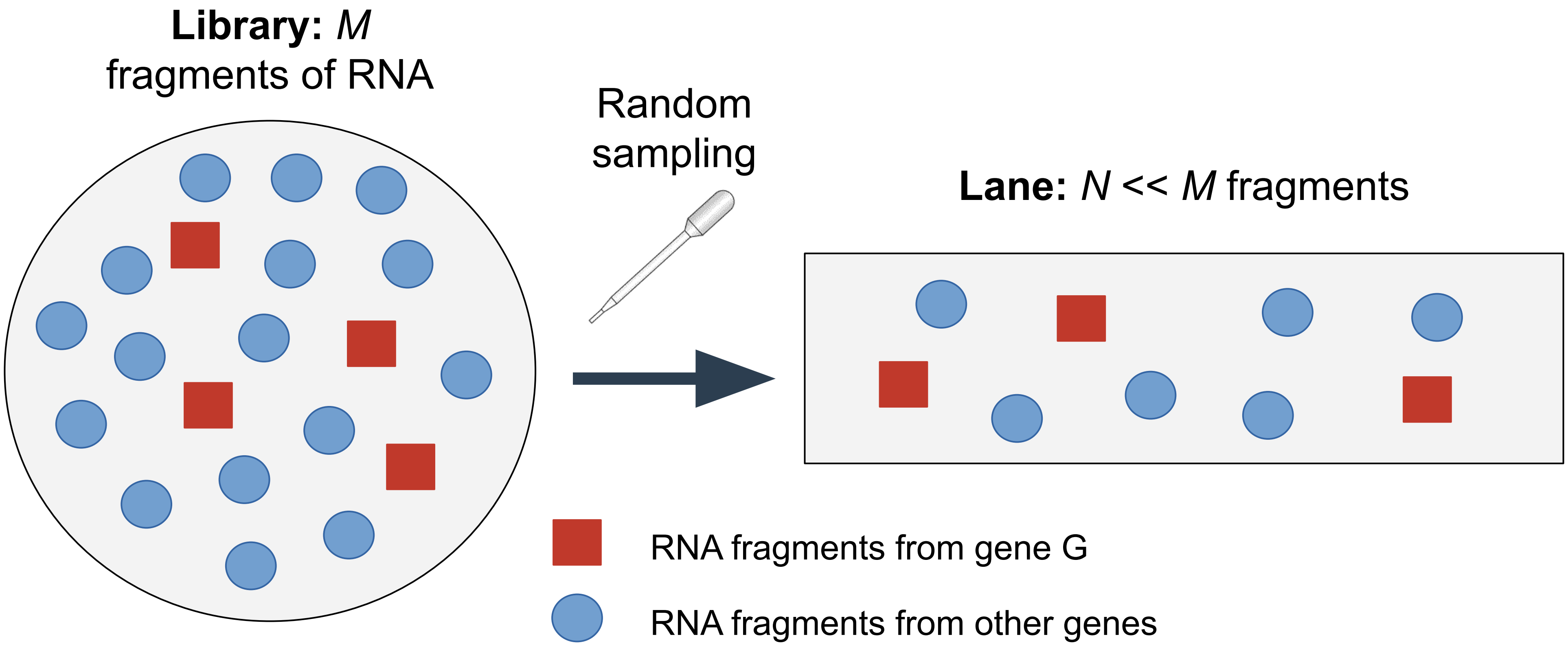
- SERE coefficient for each pair of samples [2]
- Principal Component Analysis
- Hierarchical clustering

Distribution of counts data



"*It is a good approximation to say that there is a linear relationship between read counts resulting from a sequencing experiment and the abundance of each sequence in the starting RNA material.*" [1]

Distribution of counts data



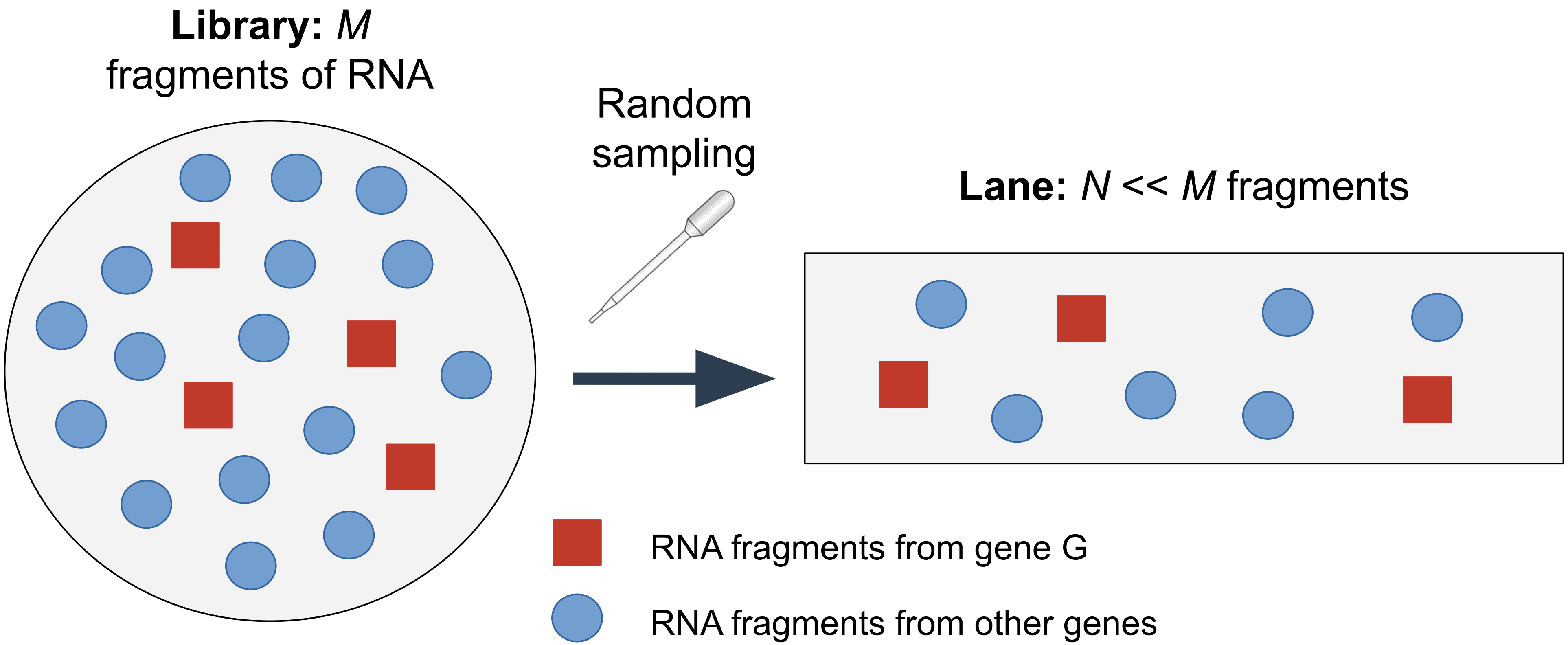
Let π_G = proportion of fragments of gene G:

$$\{\text{read } R \text{ comes from gene G}\} \sim \text{Bernoulli}(\pi_G)$$

Thus:

$$X_G = \text{nb. of reads from gene G} \sim \text{Binomial}(N, \pi_G) \approx \text{Poisson}(N\pi_G)$$

Distribution of counts data



With a deeper sequencing (i.e. larger N):

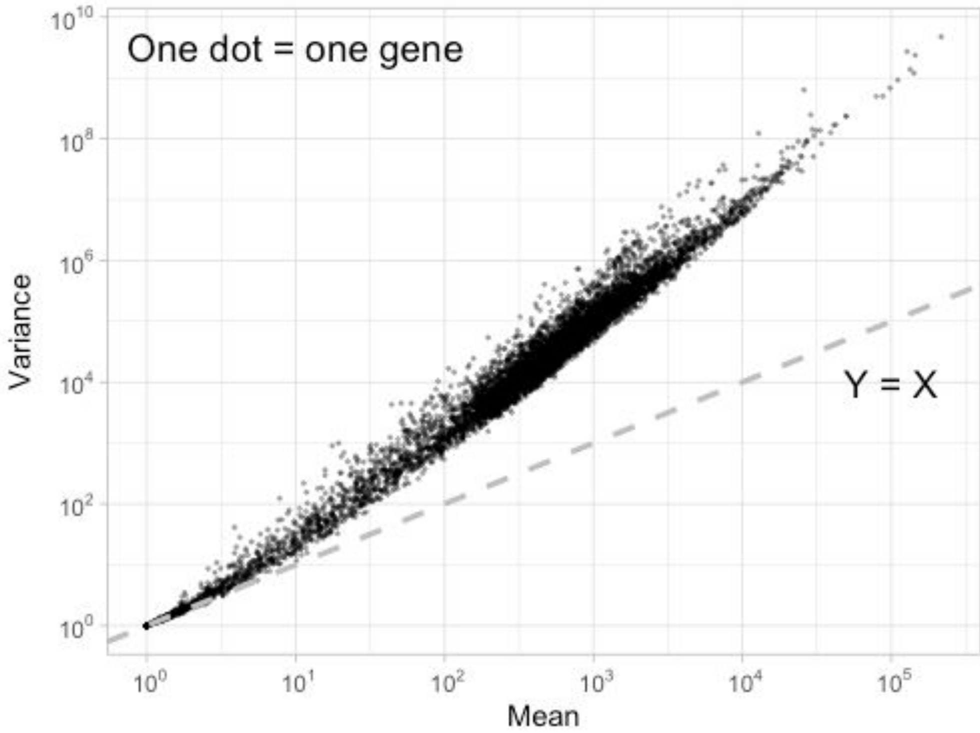
- Higher probability to catch lowly expressed genes
- Higher precision when estimating π_G

Distribution of counts data

If X_G ~ Poisson($N\pi_G$):

$$\text{mean}(X_G) = \text{variance}(X_G) = N\pi_G$$

Due to biological variability, we observe over-dispersion:



→ Need a statistical law with variance ≠ mean.

Distribution of counts data

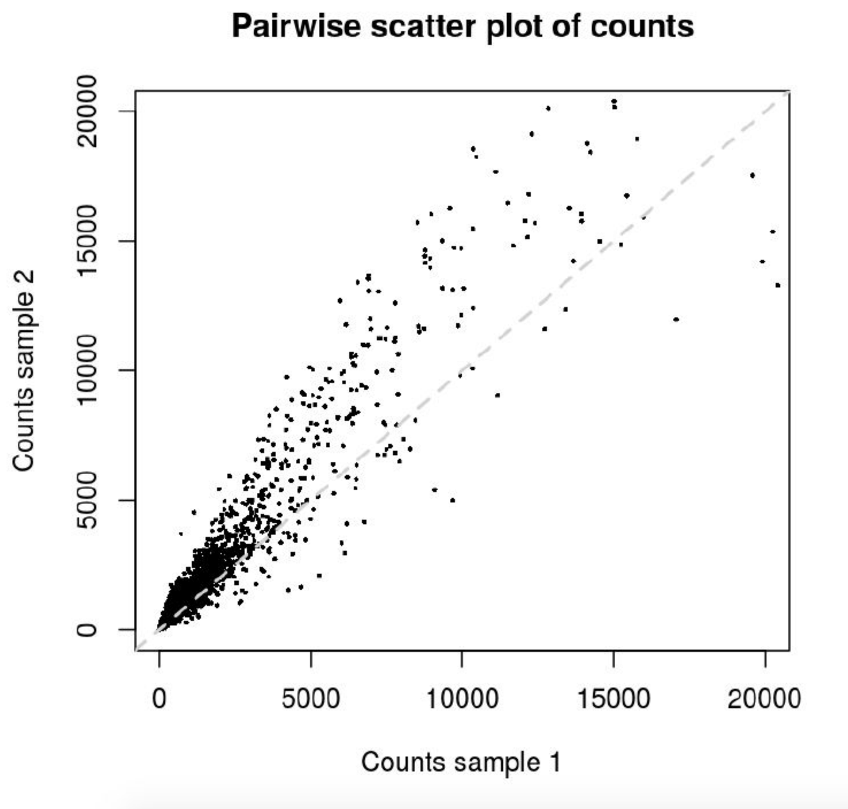
Let x_{ij} the number of reads that align on gene *i* for sample *j* (intersection row *i* - column *j* of the count matrix).

$$x_{ij} \sim \text{Negative-Binomial}(\text{mean} = \mu_{ij}, \text{variance} = \sigma_{ij}^2)$$

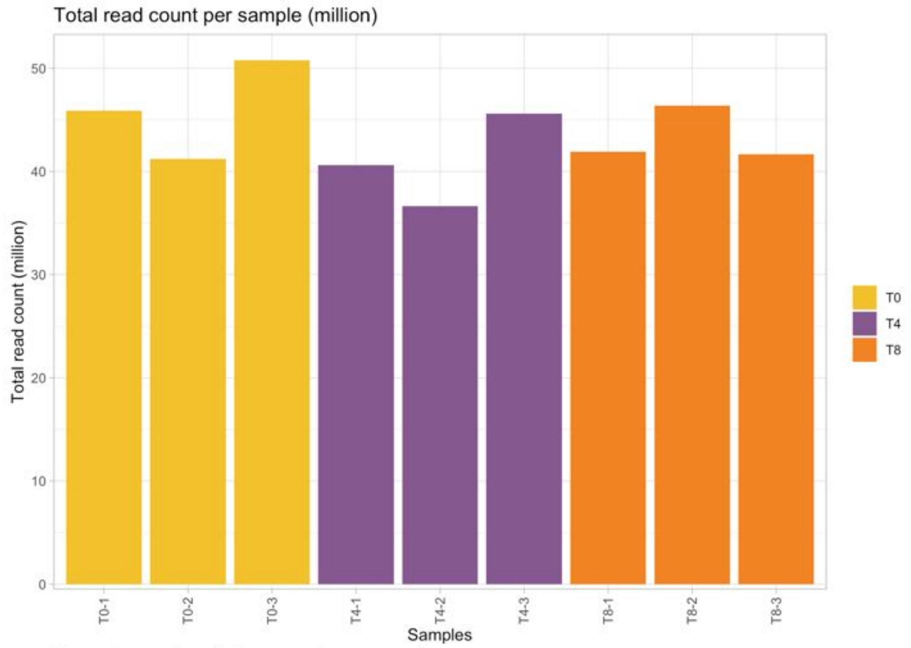
where:

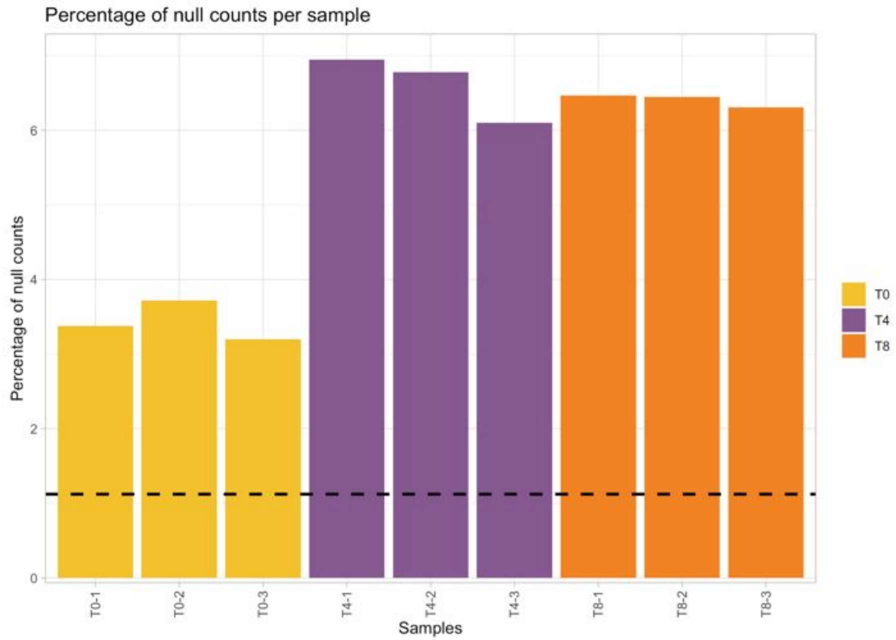
- $\sigma_{ij}^2 = \mu_{ij} + \varphi_i \mu_{ij}^2$
- φ_i : biological dispersion of gene *i*

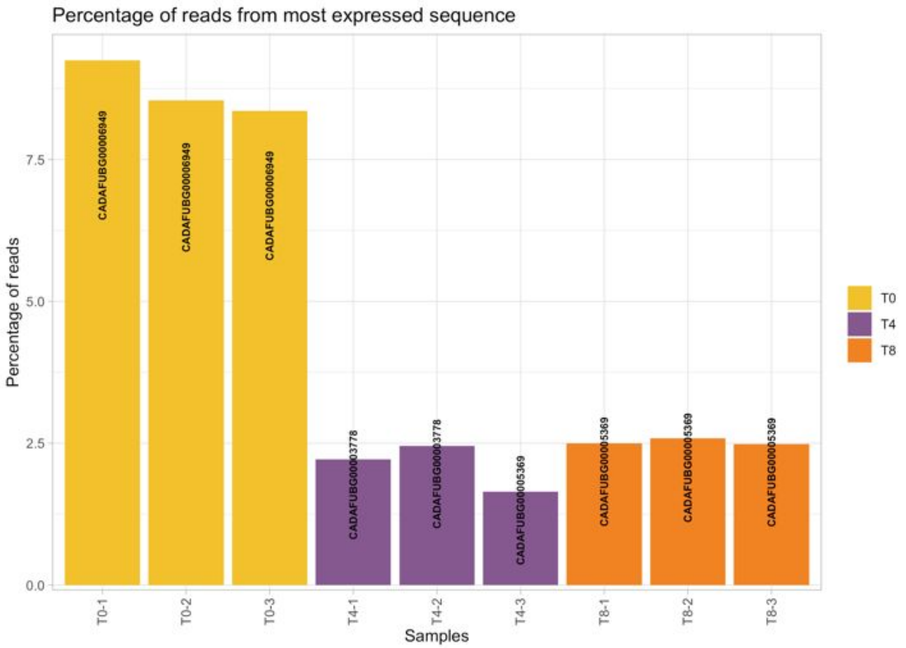
Particularity:
the x_{ij}'s are **null** or **positive integers**.

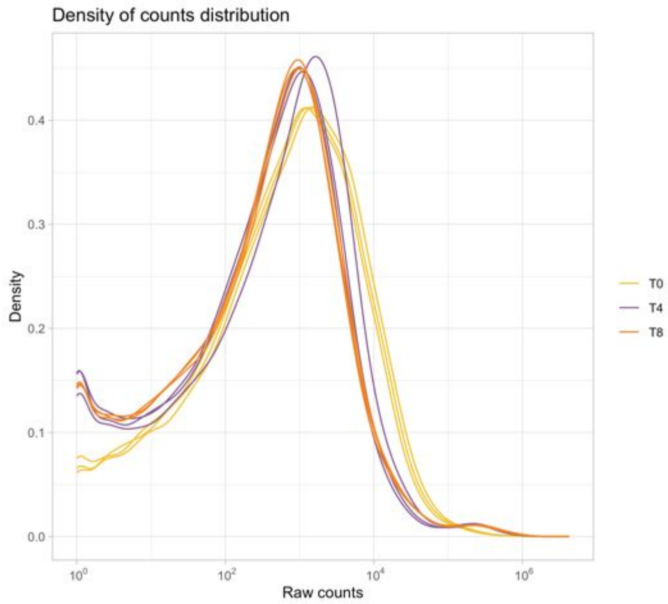


Descriptions sample by sample









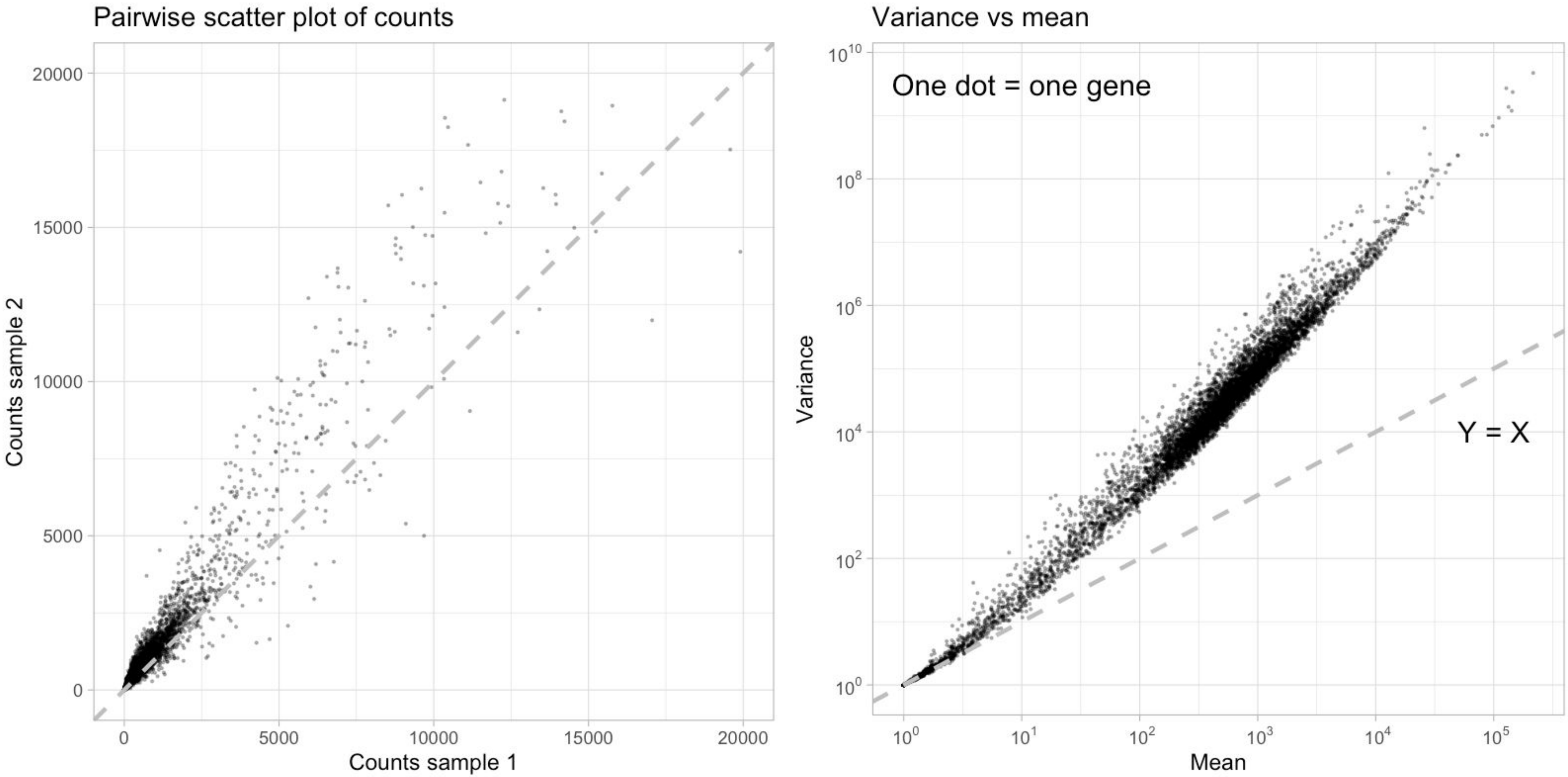
Exploratory data analysis (EDA)

Two main tools:
- Principal Component Analysis (PCA)
- Clustering

Pre-requisite:
- Notion of **distance** between the samples
- Make the data homoscedastic (=homogeneous variance)

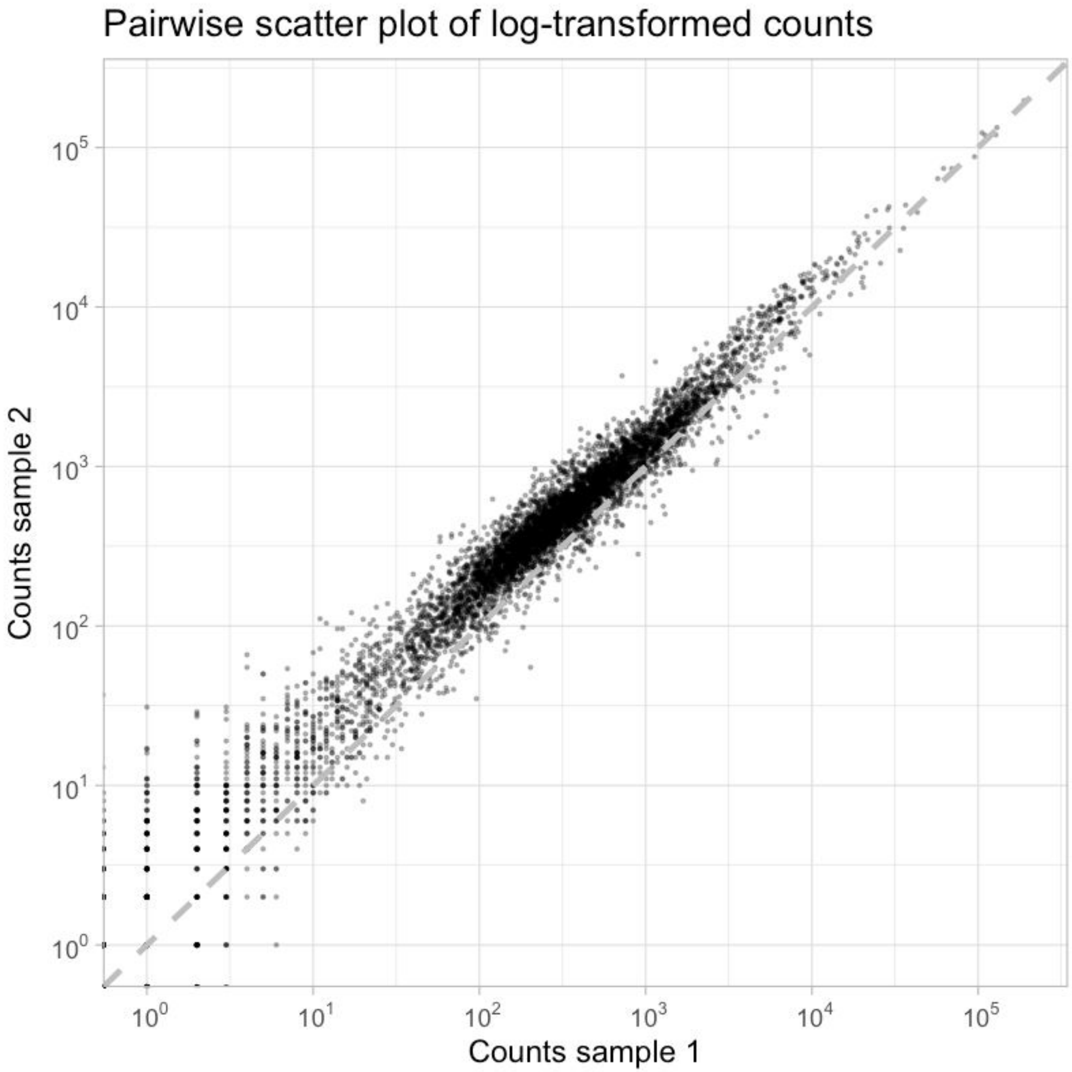
variance must be independent of the mean

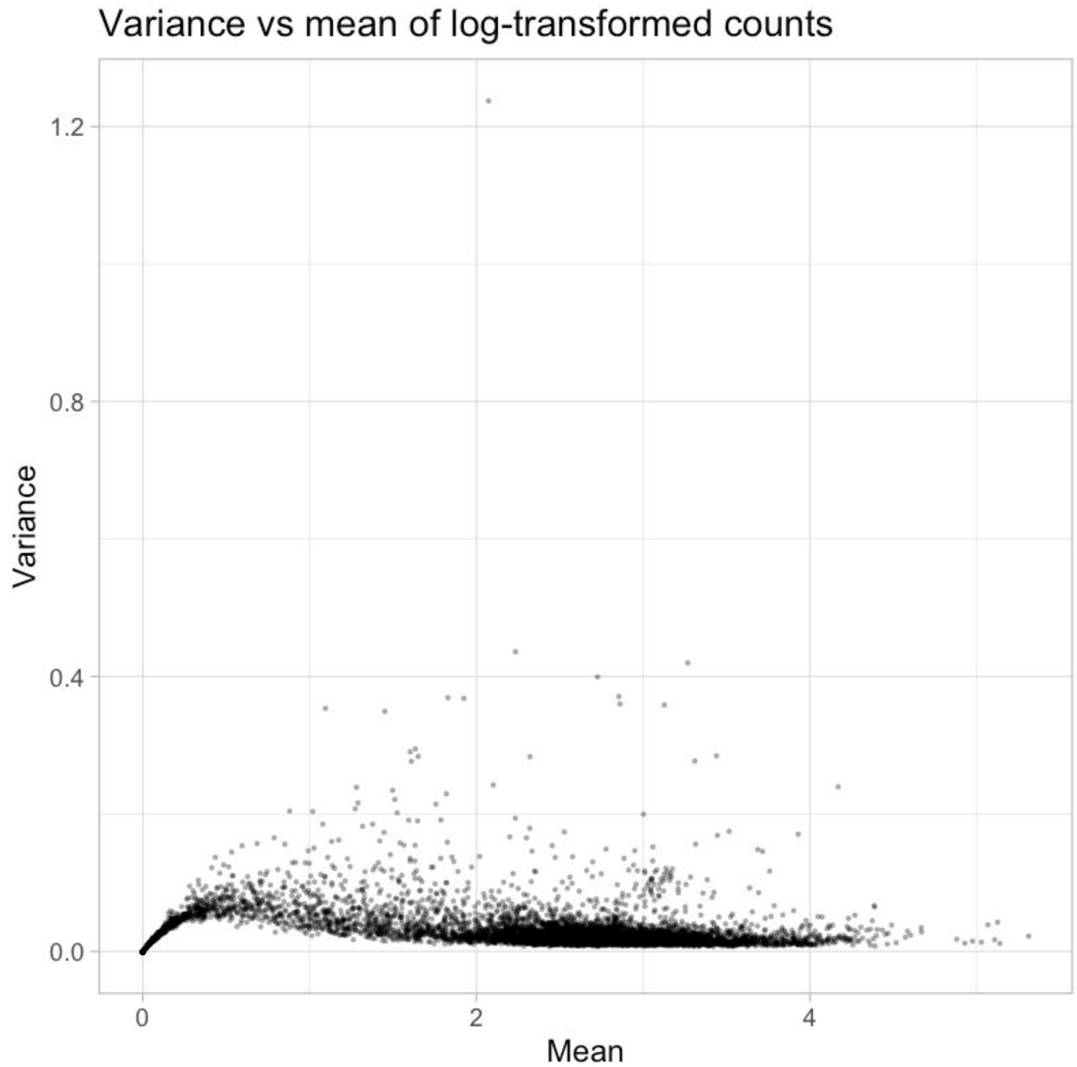
Variance increases with intensity



INSTITUT PASTEUR

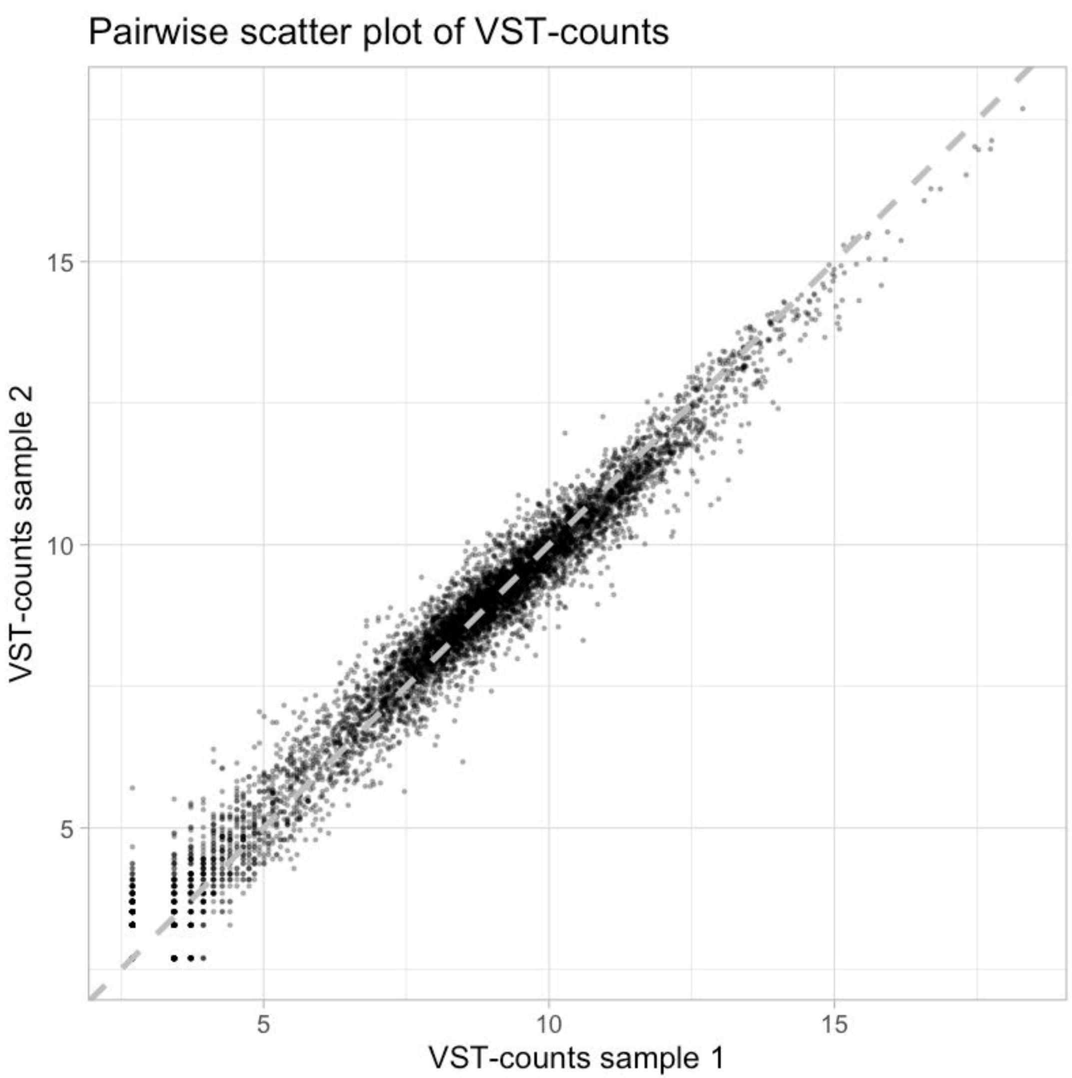
Log-transformation

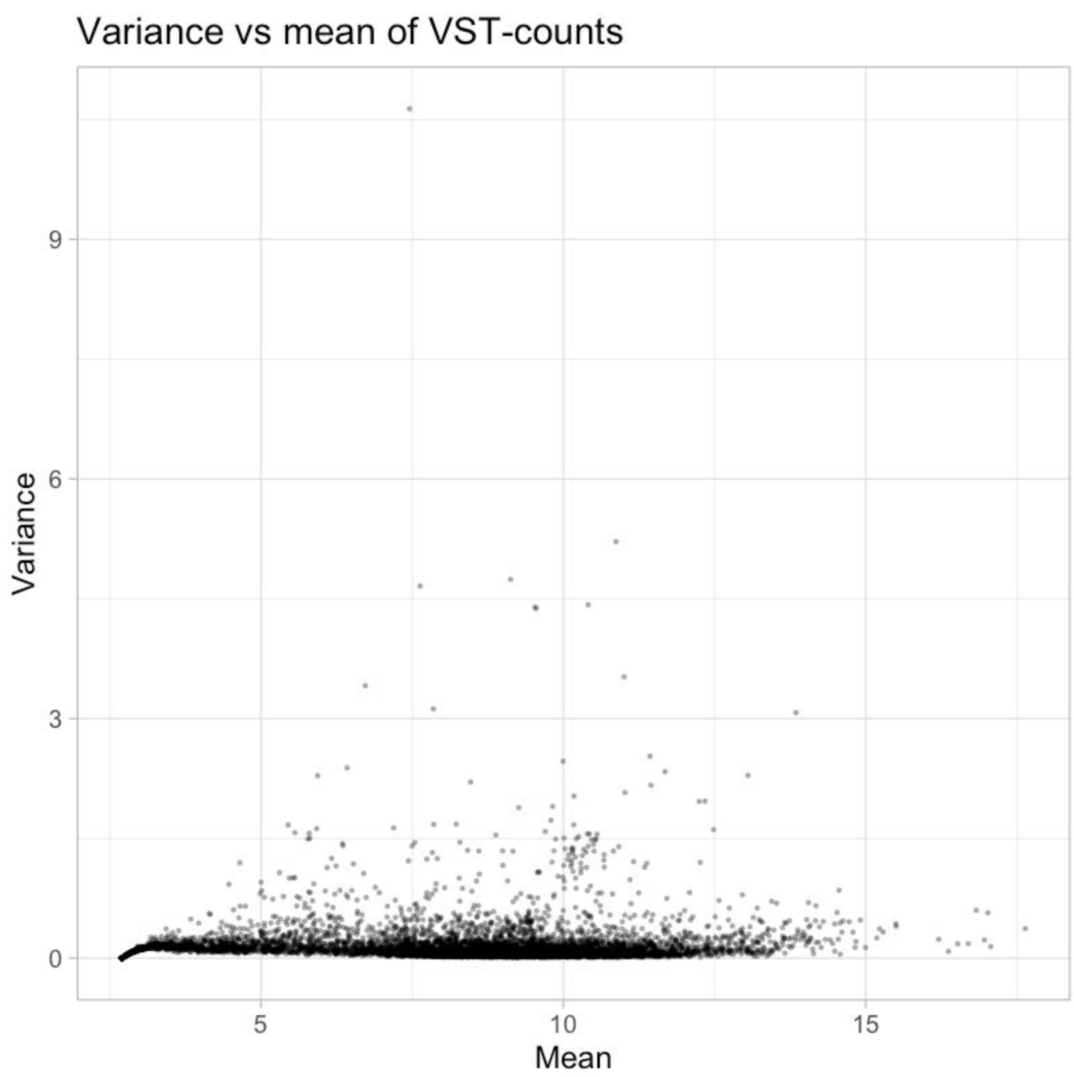




INSTITUT PASTEUR

Variance-Stabilizing Transformation [3]



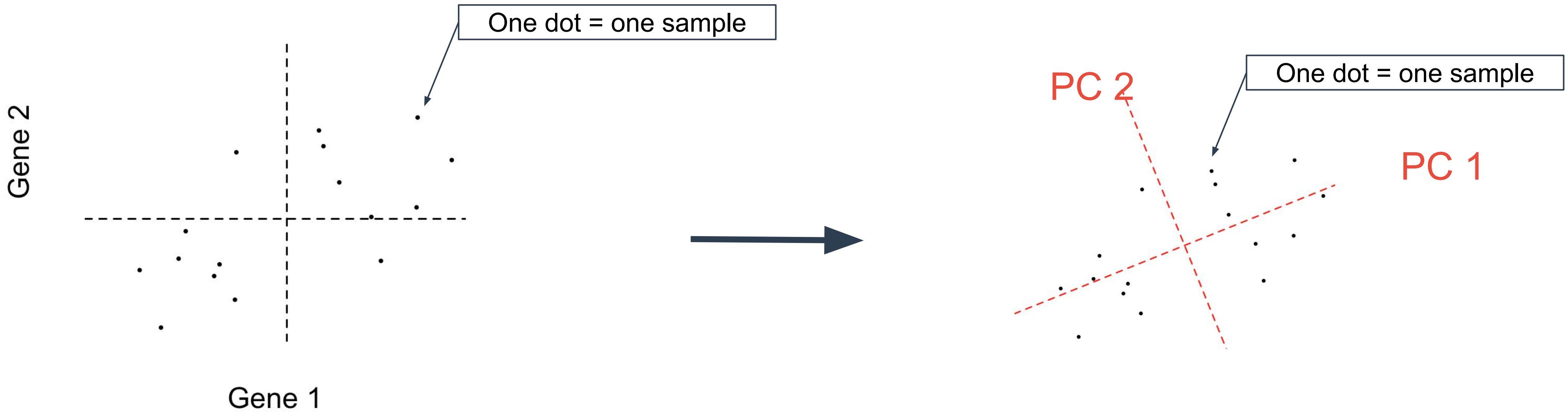


Use these data to perform Exploratory Data Analysis ONLY !

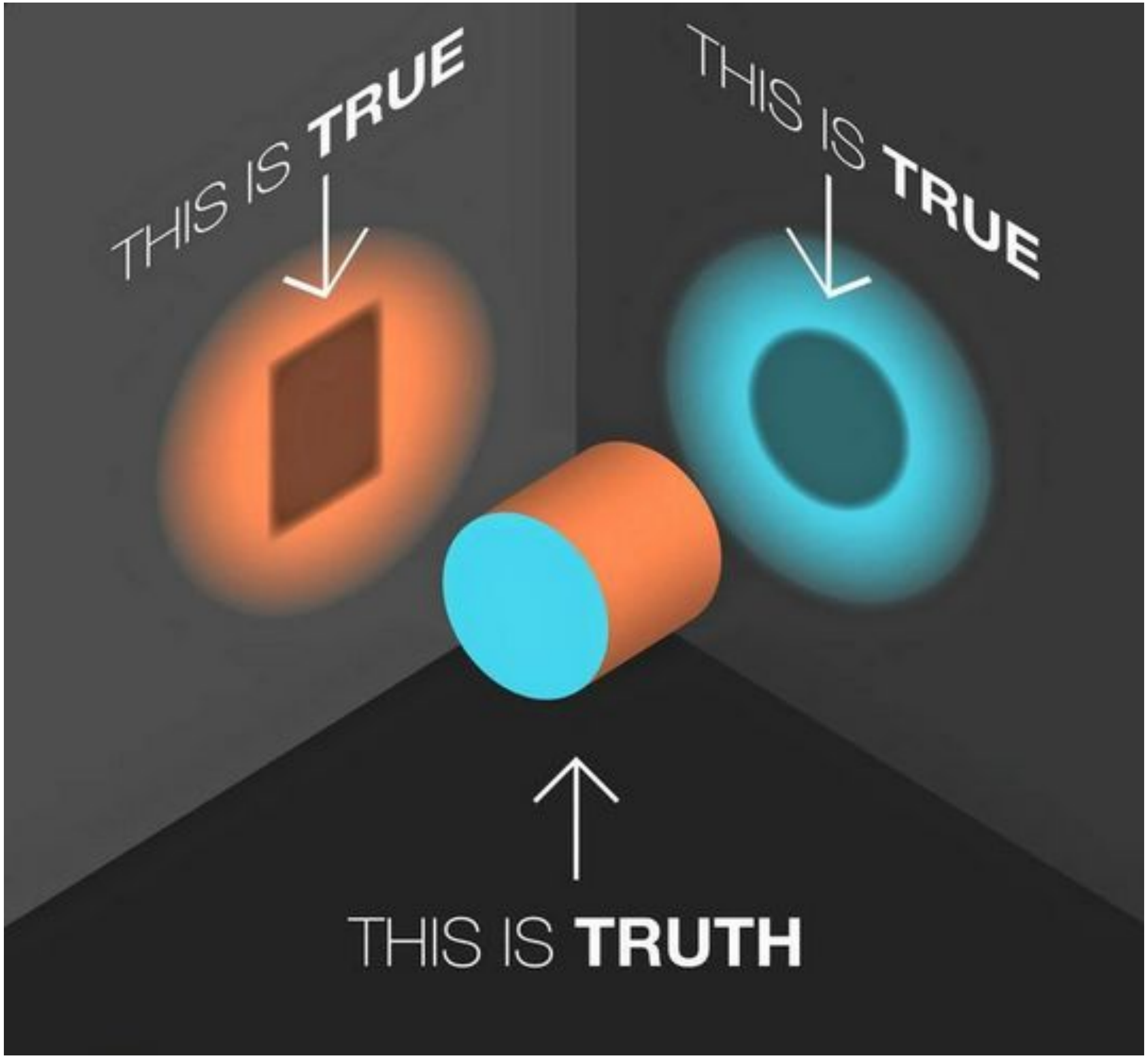
Principal Component Analysis (PCA)

Goal:
Facilitate the vision of a large (high dimensional) data set.

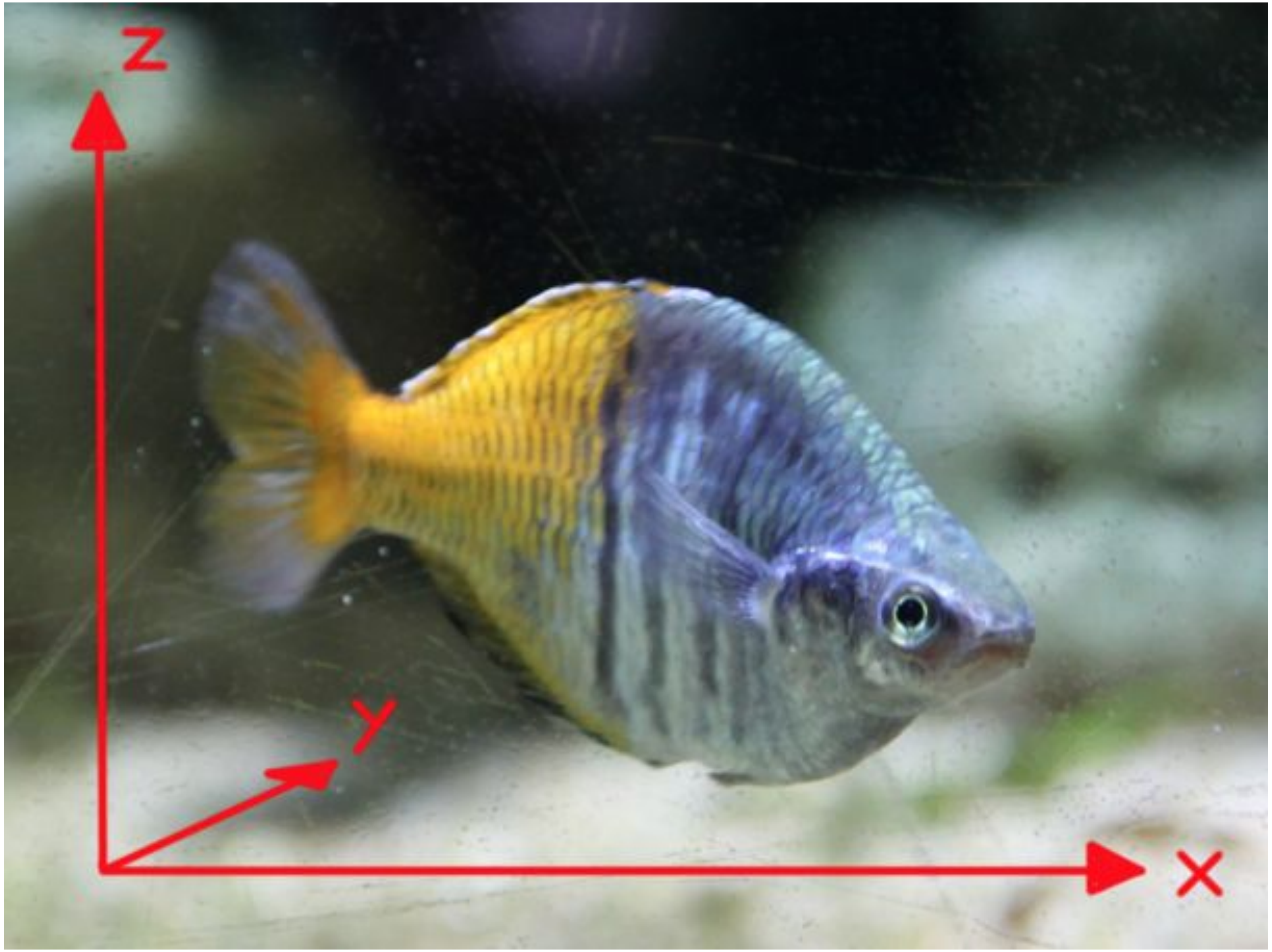
Method:
Project a cloud of *P* dots (samples) of dimension *N* (genes) on a subspace (e.g. a line or a plan) while conserving most of its structure.



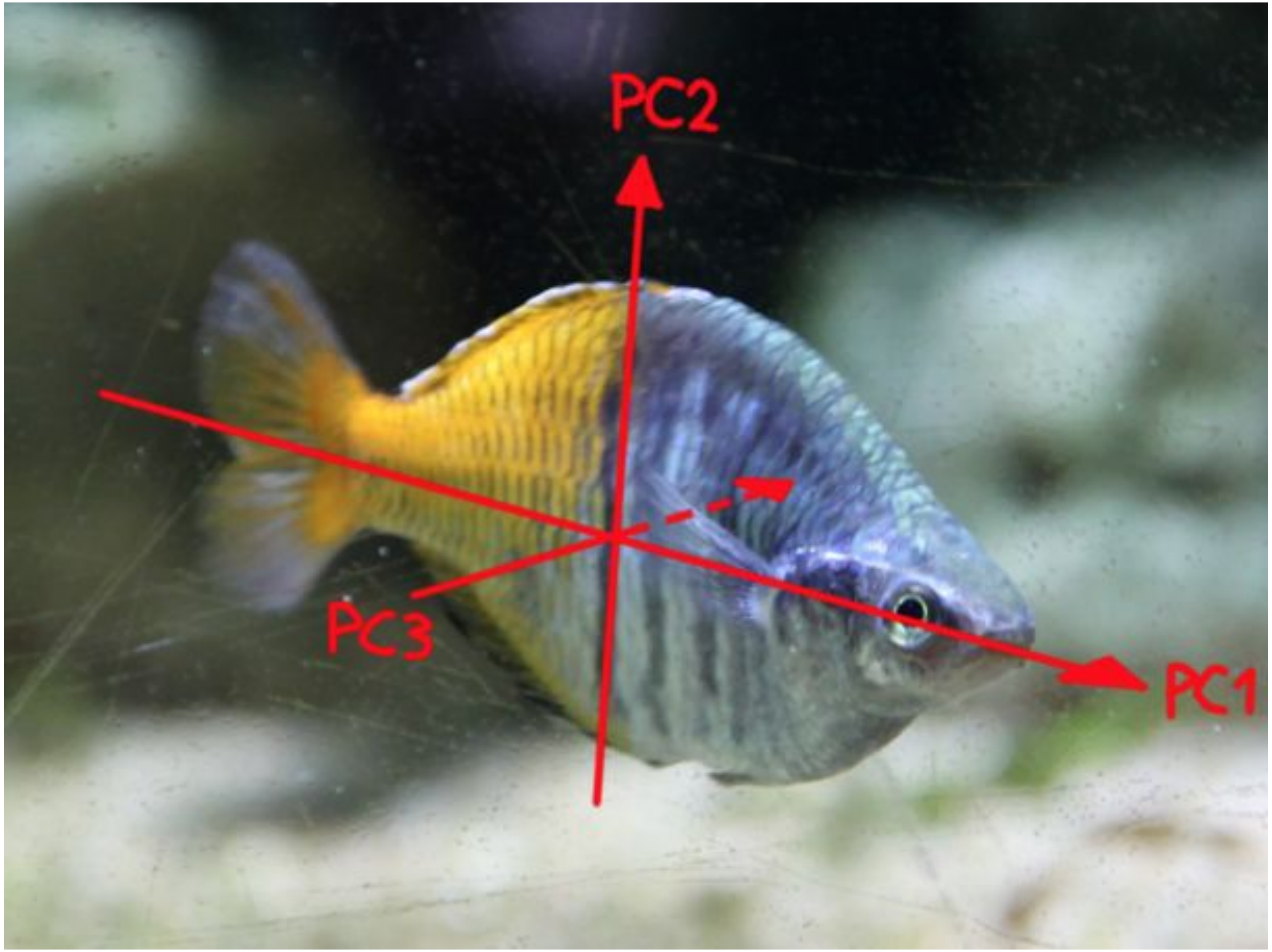
Projection: loss of information



PCA on a fish (source: bioinfo-fr.net)



PCA on a fish (source: bioinfo-fr.net)



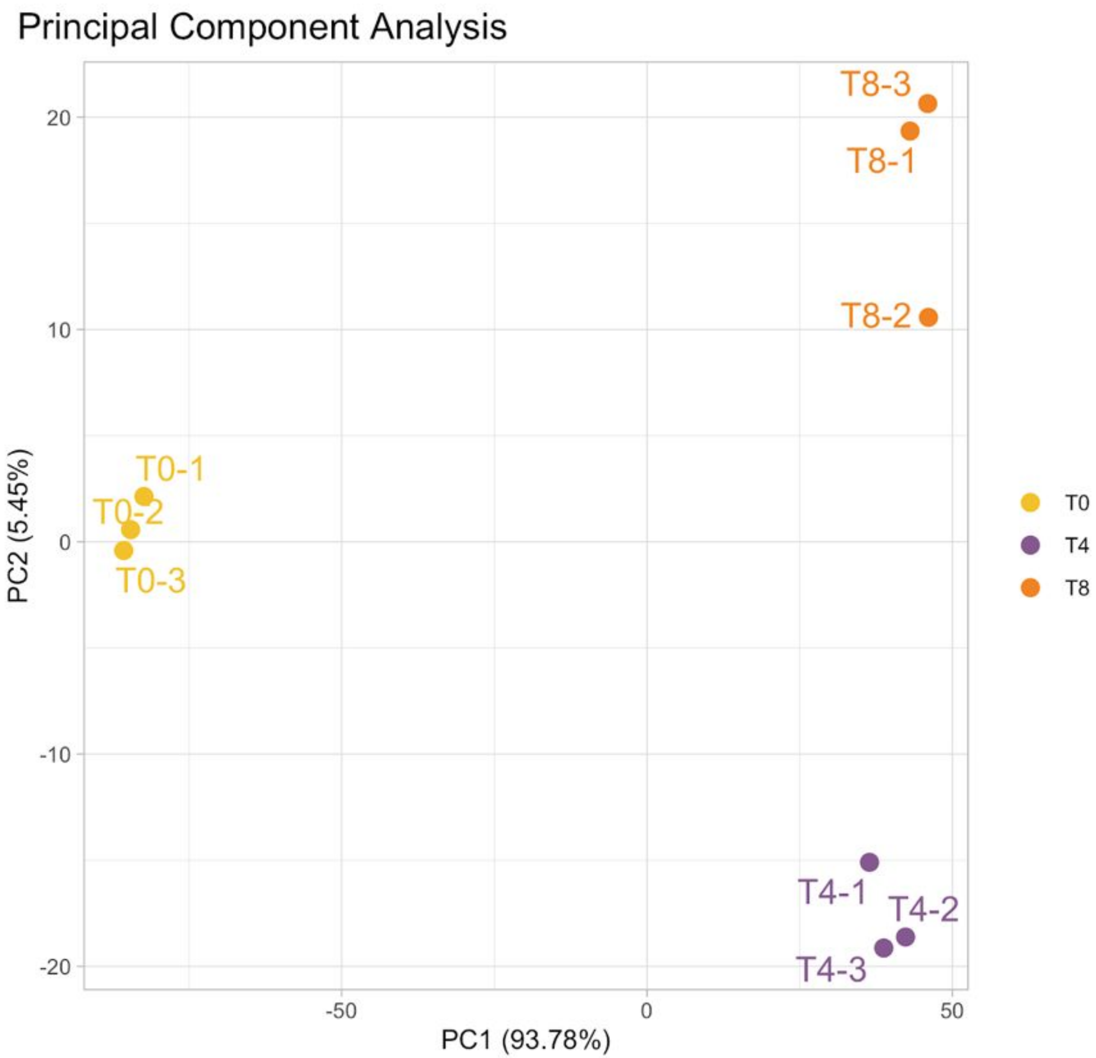
PCA: important scores

Percentage of inertia associated with an axis:
- Proportion of the total information supported by this axis
- Decreases with the axis rank (by construction)

Number of axes to interpret:
- Such as the sum of the percentages of inertia is ≥ x%
- Elbow criterion
- And many other methods

Comment: the data structure is (supposed to be) known in a differential analysis framework.

PCA: RNA-Seq example



Pre-requisite: counts must be transformed (made homoscedastic) before building the PCA.

PCA: dimensionality reduction

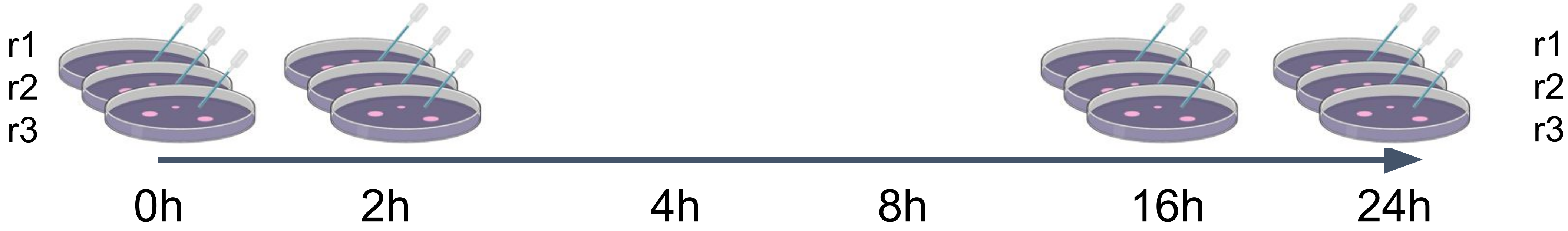
	T0-1	T0-2	T0-3	T4-1	T4-2	T4-3	T8-1	T8-2	T8-3
gene1	6.41	6.35	6.47	5.36	5.54	5.38	5.03	5.41	4.96
gene2	7.07	7.10	7.02	9.21	9.24	9.05	7.69	8.19	7.77
gene3	6.21	6.24	6.12	3.71	4.06	4.32	3.93	4.05	3.91
gene4	7.35	7.34	7.44	6.51	6.12	6.44	6.71	6.47	6.50
gene5	1.04	1.24	0.62	0.16	0.17	0.50	1.02	0.97	1.26
gene6	0.69	0.04	0.36	0.12	0.67	0.80	2.02	1.28	1.32
gene7	0.24	0.69	-0.01	-0.76	-0.74	-0.79	-0.72	-0.74	-0.72
...	3.29	3.76	3.18	4.74	3.98	3.47	4.31	4.95	4.65
geneN	3.65	4.17	4.13	5.96	6.17	5.65	4.09	4.02	3.98

From genes/variables to principal components

PC1	-60.1	-61.0	-61.5	25.9	30.4	28.8	31.0	33.1	33.3
PC2	1.3	0.5	-0.1	-11.9	-14.0	-15.0	15.1	7.9	16.3
PC3	0.4	0.3	0.1	-0.1	-0.2	-0.3	0.1	0	-0.1
PC4	-0.2	0	-0.1	0.1	0.1	0.2	-0.1	-0.2	0.2

PCA: confounding effect

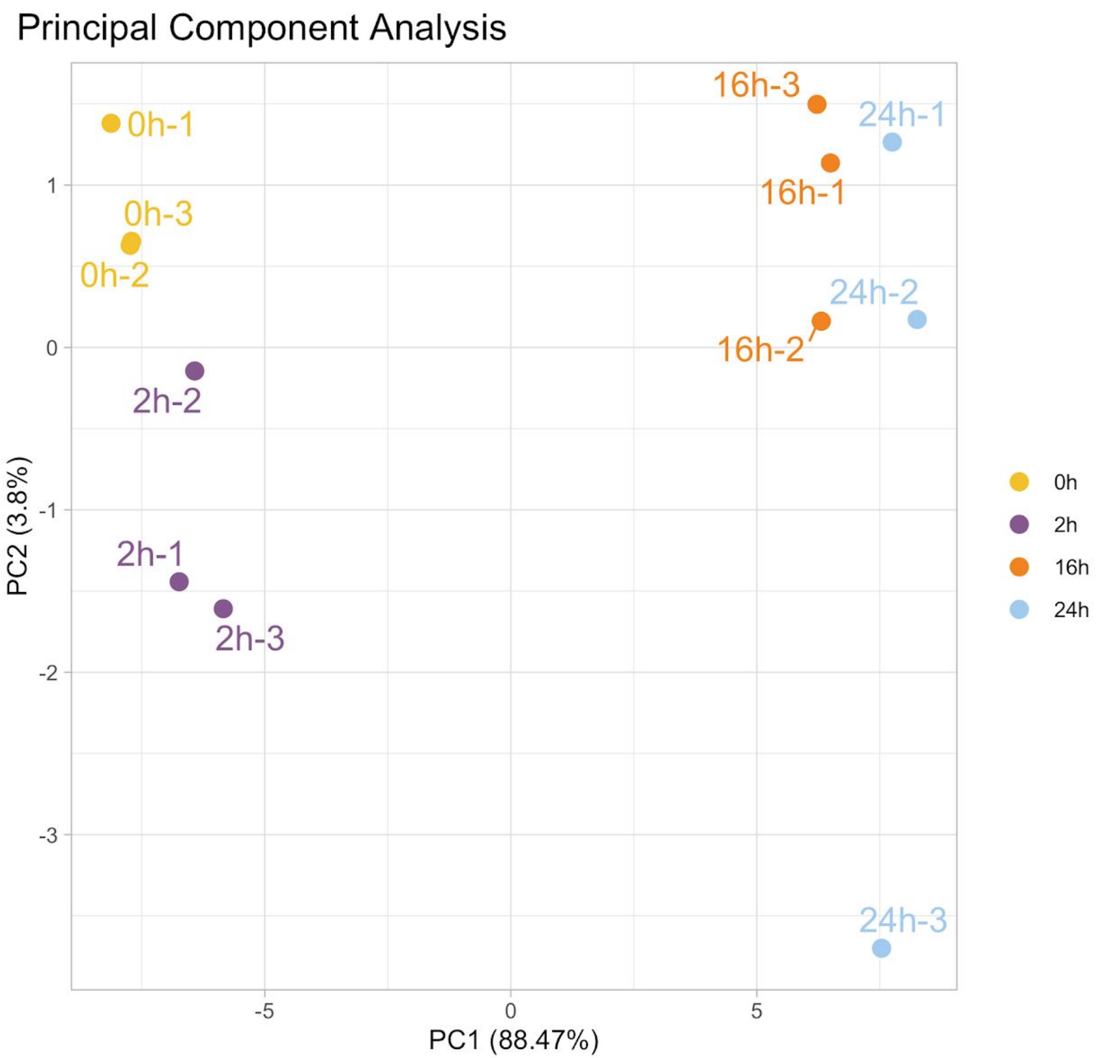
Transcriptome study of a bacteria at 0h, 2h, 16h and 24h:



label	time	replicate	date	libraries_method	libraries_exp	libraries_date
0h-1	0h	r1	oct18	robot	Bob	nov18
0h-2	0h	r2	oct18	robot	Bob	nov18
0h-3	0h	r3	oct18	robot	Bob	nov18
2h-1	2h	r1	oct18	robot	Bob	nov18
2h-2	2h	r2	oct18	robot	Bob	nov18
2h-3	2h	r3	oct18	robot	Bob	nov18
16h-1	16h	r1	oct18	robot	Bob	nov18
16h-2	16h	r2	oct18	robot	Bob	nov18
16h-3	16h	r3	oct18	robot	Bob	nov18
24h-1	24h	r1	oct18	robot	Bob	nov18
24h-2	24h	r2	oct18	robot	Bob	nov18
24h-3	24h	r3	oct18	robot	Bob	nov18

PCA: confounding effect

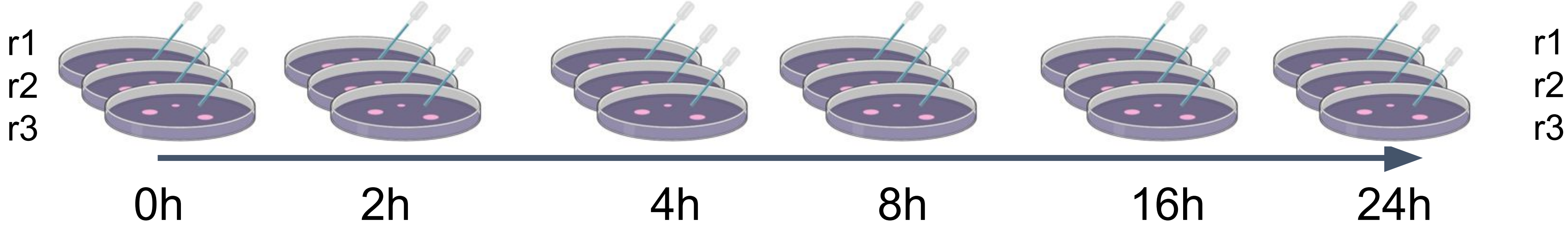
Transcriptome study of a bacteria at 0h, 2h, 16h and 24h:



INSTITUT PASTEUR

PCA: confounding effect

Add samples 4h and 8h **from the same cultures**:

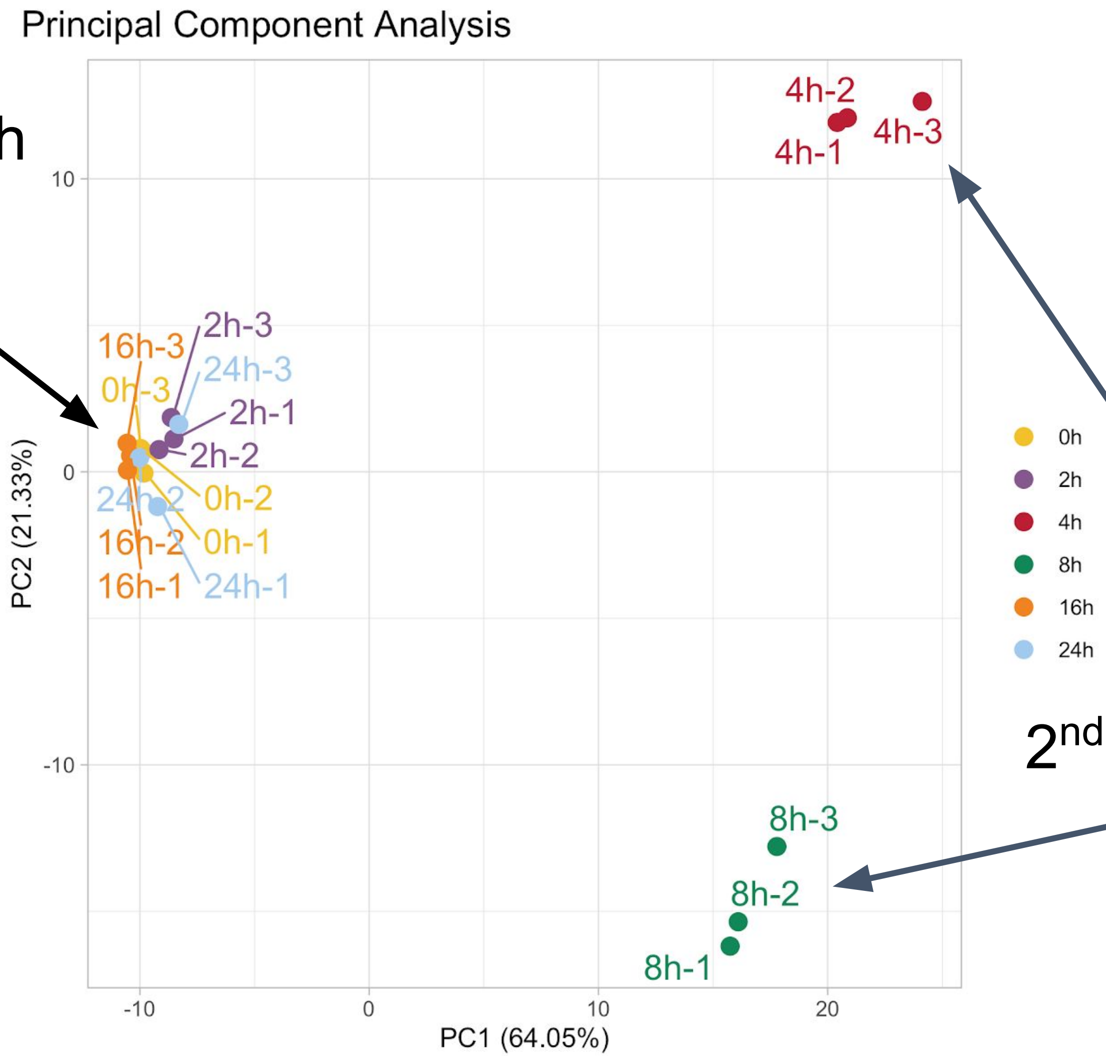


label	time	replicate	date	libraries_method	libraries_exp	libraries_date
0h-1	0h	r1	oct18	robot	Bob	nov18
0h-2	0h	r2	oct18	robot	Bob	nov18
0h-3	0h	r3	oct18	robot	Bob	nov18
2h-1	2h	r1	oct18	robot	Bob	nov18
2h-2	2h	r2	oct18	robot	Bob	nov18
2h-3	2h	r3	oct18	robot	Bob	nov18
4h-1	**4h**	**r1**	**oct18**	**manual**	**Donald**	**jun19**
4h-2	**4h**	**r2**	**oct18**	**manual**	**Donald**	**jun19**
4h-3	**4h**	**r3**	**oct18**	**manual**	**Donald**	**jun19**
8h-1	**8h**	**r1**	**oct18**	**manual**	**Donald**	**jun19**
8h-2	**8h**	**r2**	**oct18**	**manual**	**Donald**	**jun19**
8h-3	**8h**	**r3**	**oct18**	**manual**	**Donald**	**jun19**
16h-1	16h	r1	oct18	robot	Bob	nov18
16h-2	16h	r2	oct18	robot	Bob	nov18
16h-3	16h	r3	oct18	robot	Bob	nov18
24h-1	24h	r1	oct18	robot	Bob	nov18
24h-2	24h	r2	oct18	robot	Bob	nov18
24h-3	24h	r3	oct18	robot	Bob	nov18

PCA: confounding effect

Global analysis of times 0h, 2h, 4h, 8h, 16h and 24h:

1st sequencing batch



2nd sequencing batch

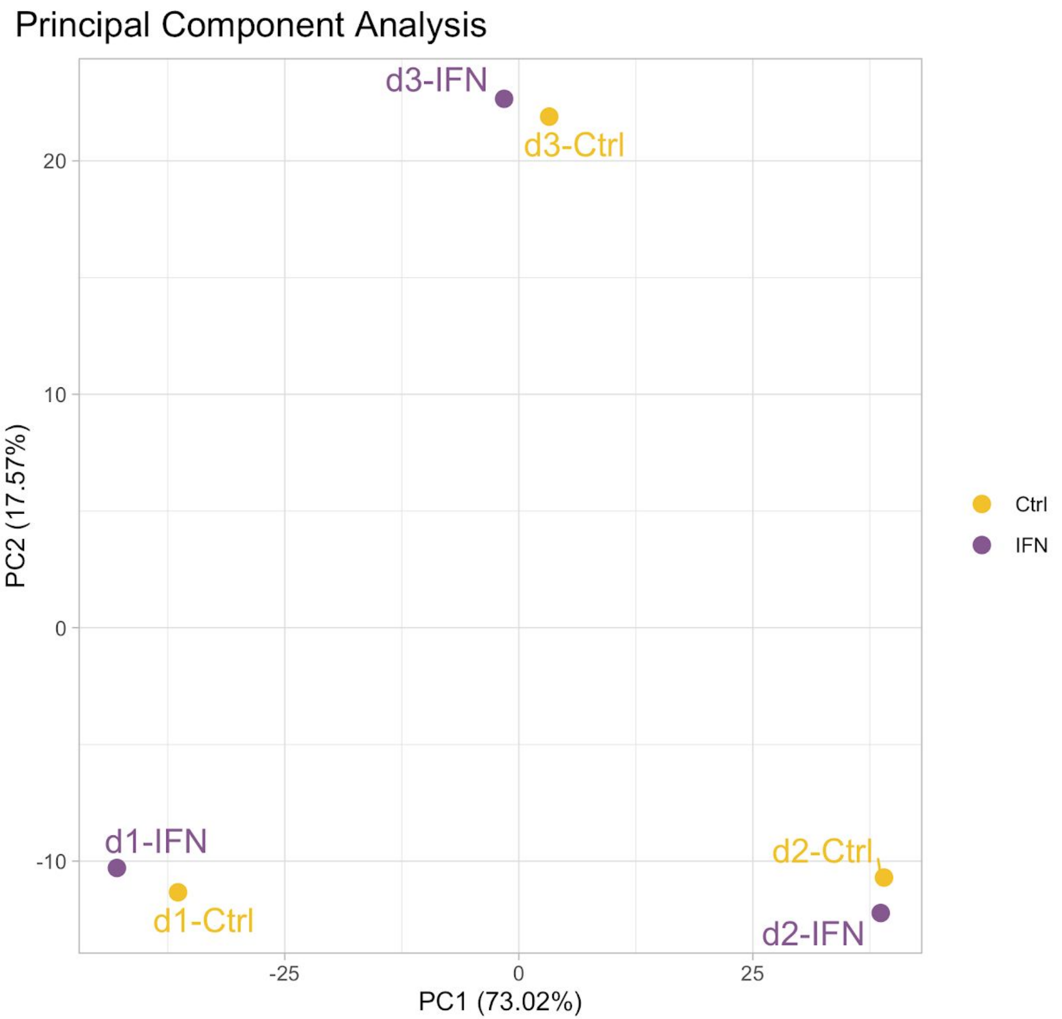
PCA: pairing factor

Two treatments applied to human cells coming from 3 donors:

label	treatment	donor
d1-IFN	IFN	d1
d1-Ctrl	Ctrl	d1
d2-IFN	IFN	d2
d2-Ctrl	Ctrl	d2
d3-IFN	IFN	d3
d3-Ctrl	Ctrl	d3

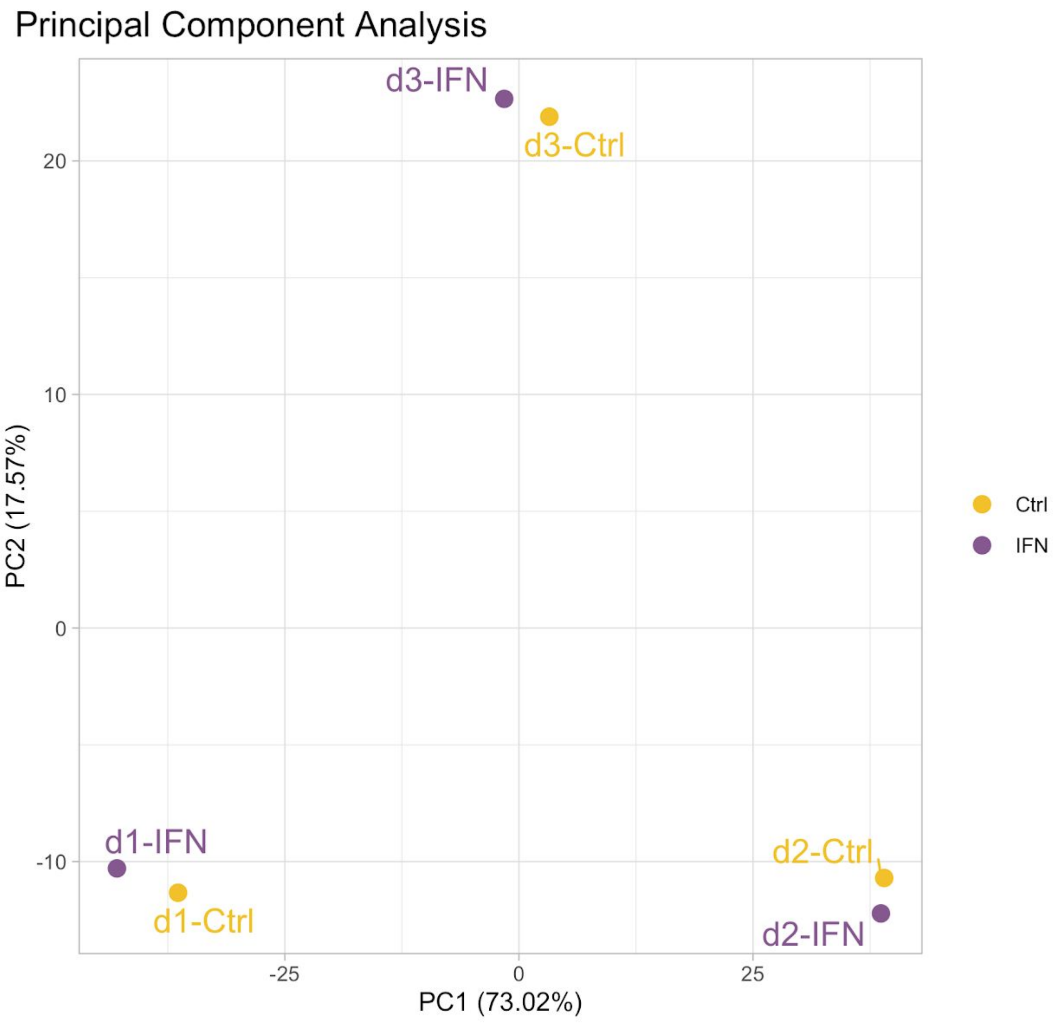
PCA: pairing factor

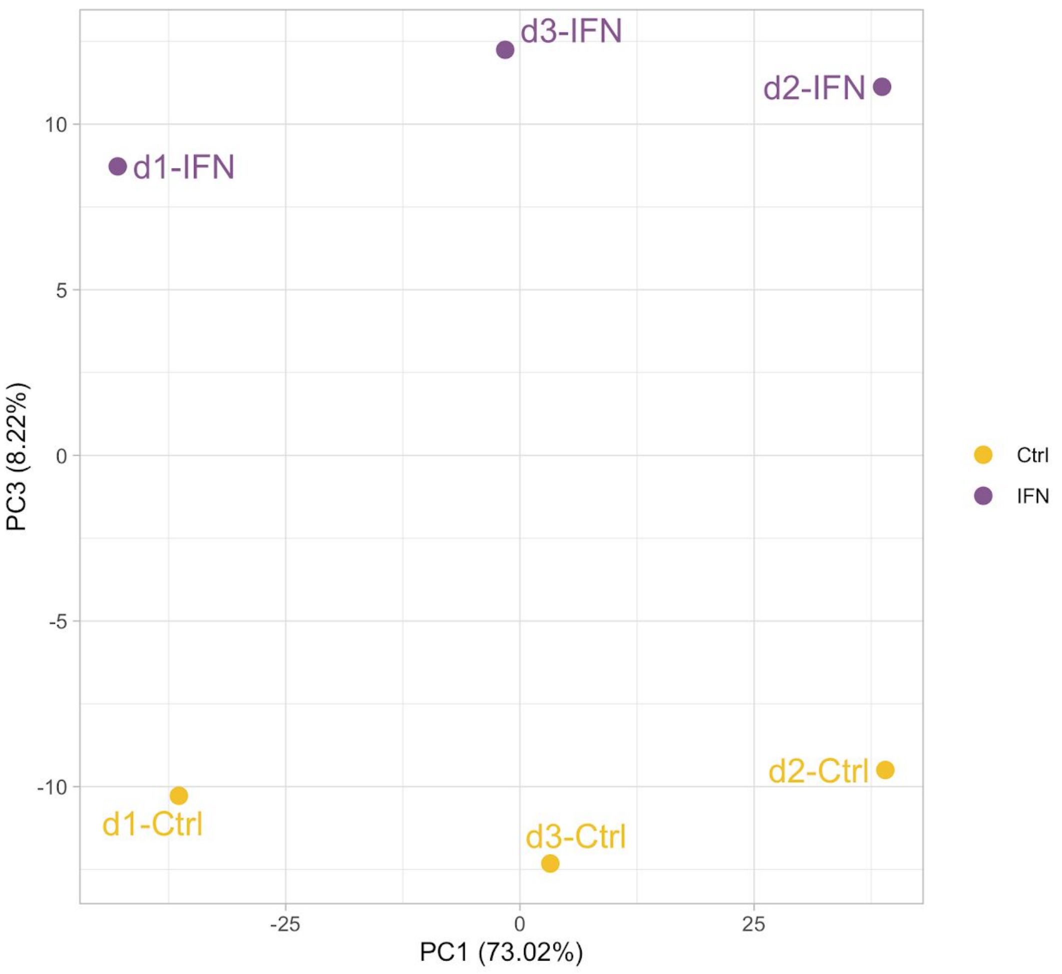
Two treatments applied to human cells coming from 3 donors:



PCA: pairing factor

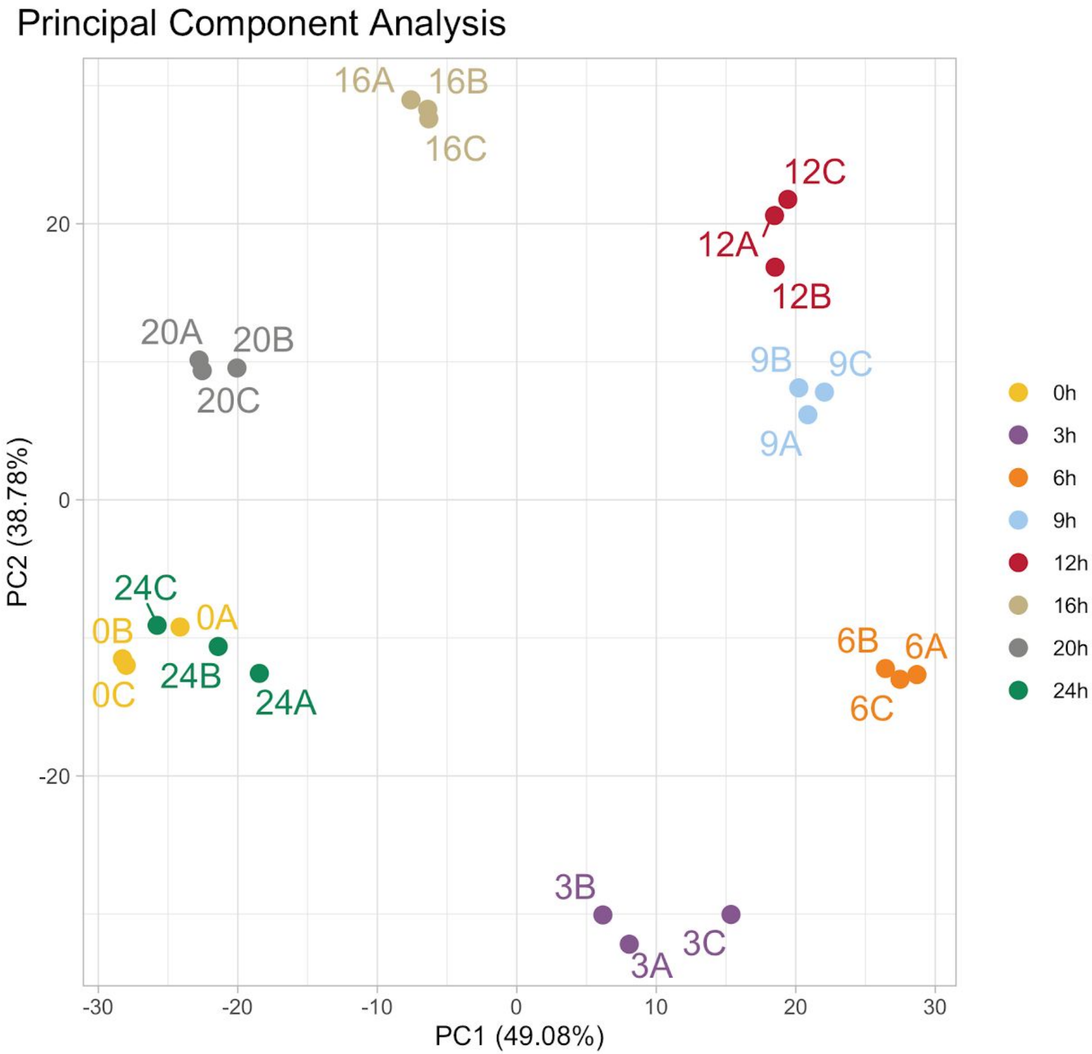
Two treatments applied to human cells coming from 3 donors:





PCA: most beautiful RNA-Seq example

Transcriptome study of a cyanobacteria at 8 time points from 0h to 24h:



INSTITUT PASTEUR

SERE coefficient [2]

Simple Error Ratio Estimate

Goal: assess the similarity/dissimilarity between samples

$$\text{SERE(A, B)} \begin{cases} = 0 \text{ if A = B} \\ \approx 1 \text{ if A and B are technical replicates} \\ > 1 \text{ if A and B are biological replicates} \\ >> 1 \text{ if A and B come from different bio. conditions} \end{cases}$$

More suited to RNA-Seq data than the Pearson/Spearman correlation coefficients.

SERE coefficient: details

- 2 samples (A and B) and N genes under study
- y_{ij} = # of reads for gene i (1, …, N) and sample j (A or B)
- L_j = total # of reads (library size) for sample j
- $E_i = y_{iA} + y_{iB}$ = number of reads for gene i
- Expected # of reads for gene i and sample j:

$$\hat{y}_{ij} = E_i \times L_j / (L_A + L_B)$$

- **Expected variation** for each observation y_{ij} : $(y_{ij} - \hat{y}_{ij})^2$
- **Expected variation** under Poisson assumption: $\hat{y}_{ij}$
- Overdispersion for each gene i: $s_i^2 = (y_{iA} - \hat{y}_{iA})^2/\hat{y}_{iA} + (y_{iB} - \hat{y}_{iB})^2/\hat{y}_{iB}$

$$\text{SERE(A, B)} = \text{sqrt}((\Sigma_{i=1..N}\, s_i^2) / N)$$

SERE coefficient: details

Simple Error Ratio Estimate (SERE)

Given a set of N exons and M lanes, let y_{ij} denote the number of reads covering the i^{th} exon in the j^{th} lane. Let L_j be the total read count for lane j, E_i the total for exon i, and T the grand total count across all lanes and exons. Under the hypothesis that the lanes are simple technical replicates of each other, they will have a Poisson distribution with one parameter. This parameter can be thought of as the expected number of reads for the lane j and the exon i. Its estimate can be calculated using eq. 1.

$$\widehat{y}_{ij} = \frac{E_i L_j}{T}$$

The expected variation for each observation y_{ij} is $\left(y_{ij} - \widehat{y}_{ij}\right)^2$, and the expected variation under the Poisson assumption is $\widehat{y}_{ij}$. This gives a per exon overdispersion estimate of:

$$s_i^2 = \frac{1}{M-1} \sum_j \frac{\left(y_{ij} - \widehat{y}_{ij}\right)^2}{\widehat{y}_{ij}}$$

The denominator is $(M-1)$ due to the constraint that $\sum_j \left(y_{ij} - \widehat{y}_{ij}\right) = 0$ for each exon i.

Averaging over all N exons we have:

$$s^2 = \frac{1}{N} \sum_i s_i^2 \qquad (3)$$

The SERE estimate is $s = \sqrt{\left(s^2\right)}$.

SERE coefficient: example

	T0-1	T0-2	T0-3	T4-1	T4-2	T4-3	T8-1	T8-2	T8-3
T0-1	**0**	**2.97**	**3.88**	73.89	71.83	74.02	74.69	76.90	74.03
T0-2	**2.97**	**0**	**3.00**	72.21	70.03	72.33	72.94	75.15	72.32
T0-3	**3.88**	**3.00**	**0**	76.34	74.28	76.33	77.18	79.38	76.51
T4-1	73.89	72.21	76.34	**0**	**5.83**	**10.42**	17.27	14.93	17.99
T4-2	71.83	70.03	74.28	**5.83**	**0**	**10.89**	17.77	15.07	18.10
T4-3	74.02	72.33	76.33	**10.42**	**10.89**	**0**	19.86	18.25	20.07
T8-1	74.69	72.94	77.18	17.27	17.77	19.86	**0**	**6.72**	**4.04**
T8-2	76.90	75.15	79.38	14.93	15.07	18.25	**6.72**	**0**	**8.22**
T8-3	74.03	72.32	76.51	17.99	18.10	20.07	**4.04**	**8.22**	**0**

Drawback: not very easy to interpret with many samples.

Clustering

Goal: build groups of samples such that:

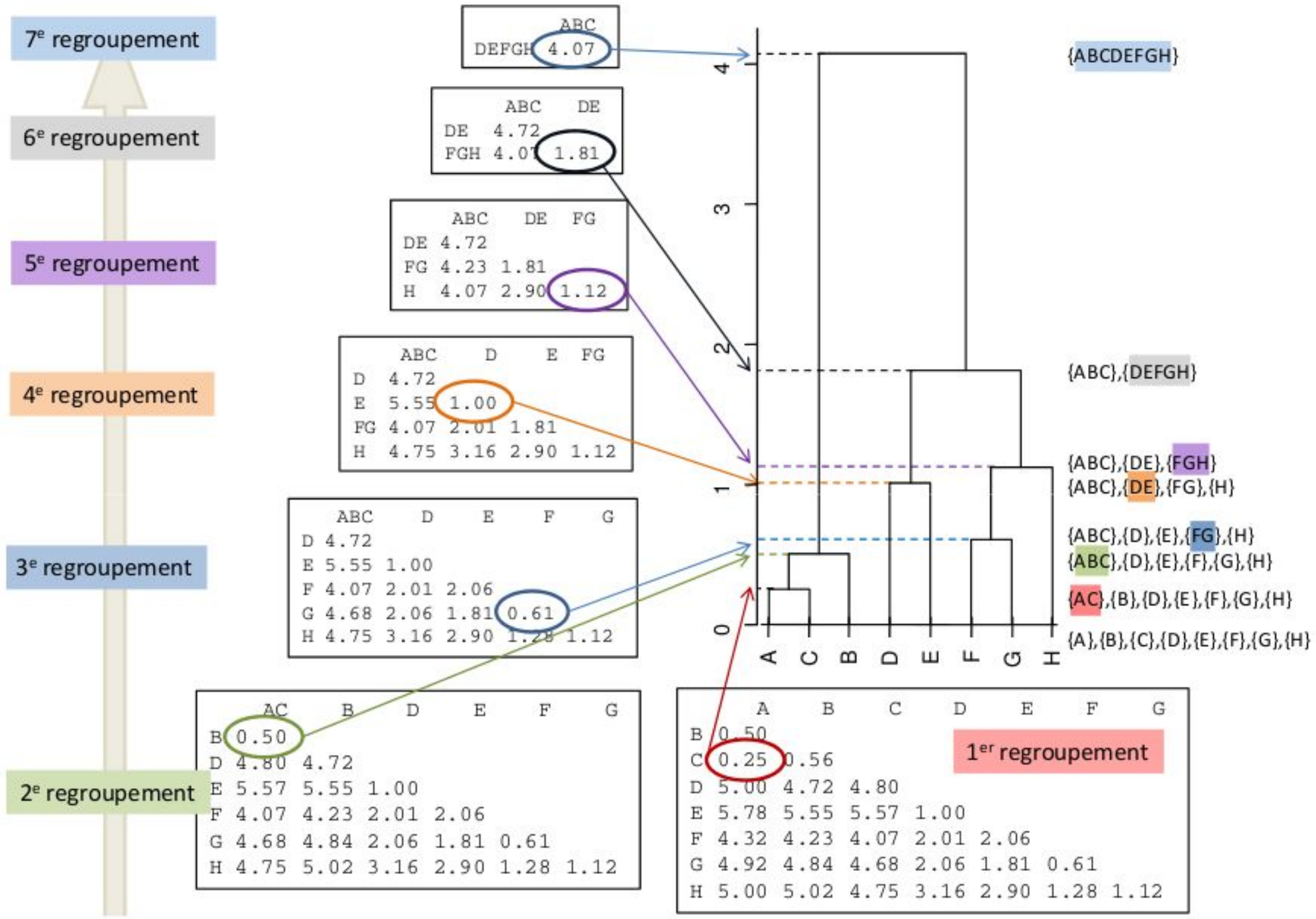
- samples within a group are similar
- samples from distinct groups are different

Method (ascendant clustering):

1. Calculate the distances between each pair of samples
2. Gather the two nearest samples into a cluster
3. Calculate the distance between this cluster and each sample
4. Gather the two nearest clusters/samples
5. Go back to step 3 until getting a single cluster

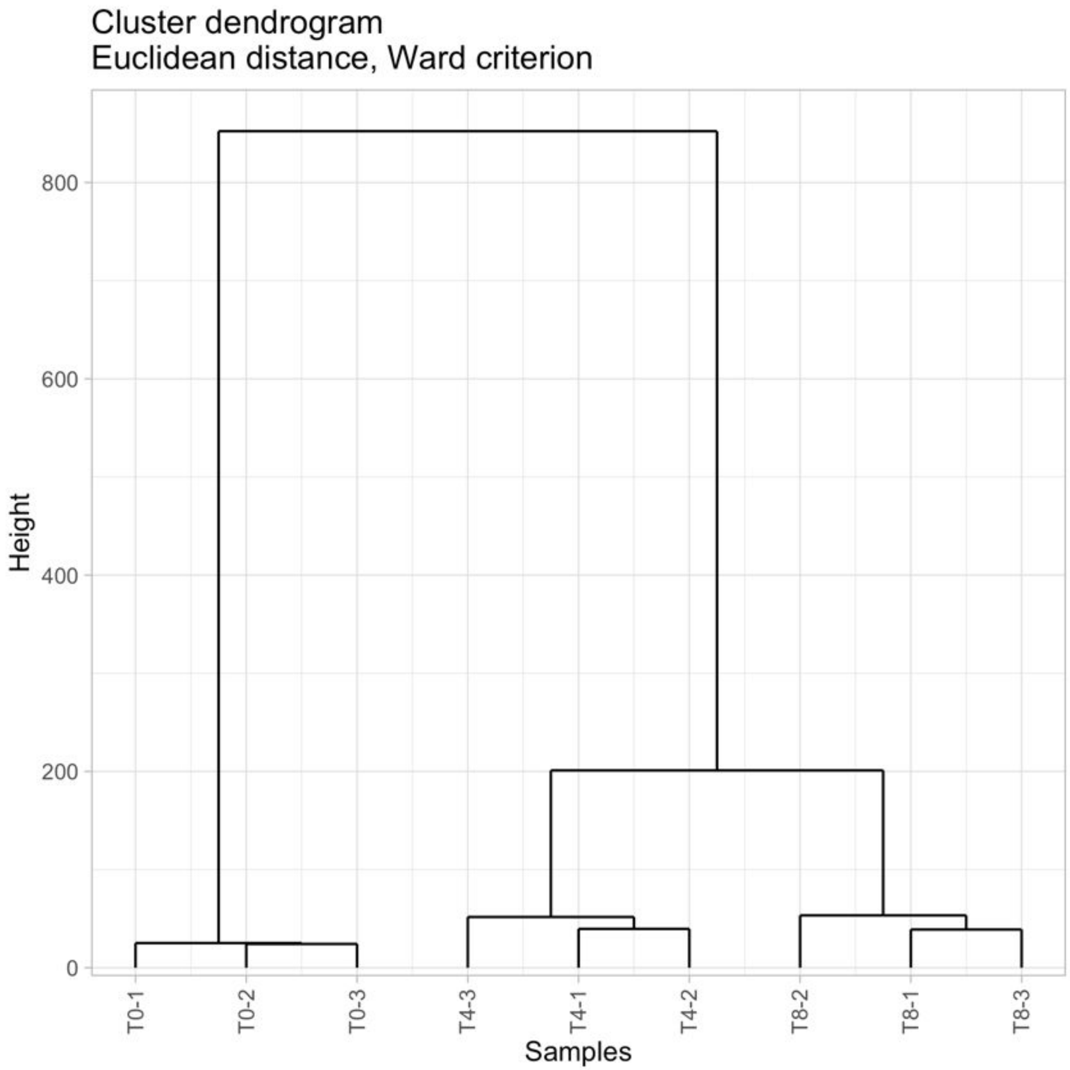
INSTITUT PASTEUR

Hierarchical clustering: example



Source: MOOC FUN Analyse de données 2015 – Agrocampus Ouest

Hierarchical clustering: RNA-Seq example



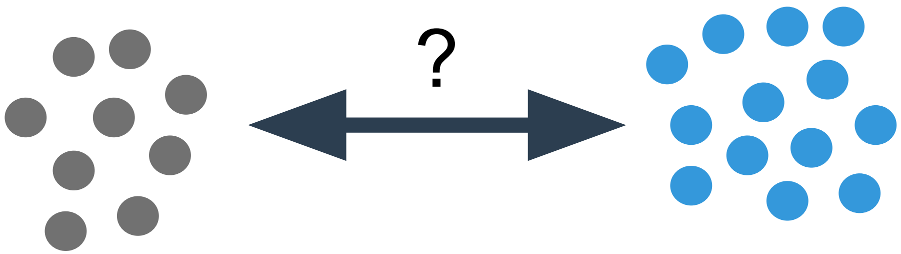
Pre-requisite: counts must be transformed (made homoscedastic) before building the PCA.

Clustering parameters

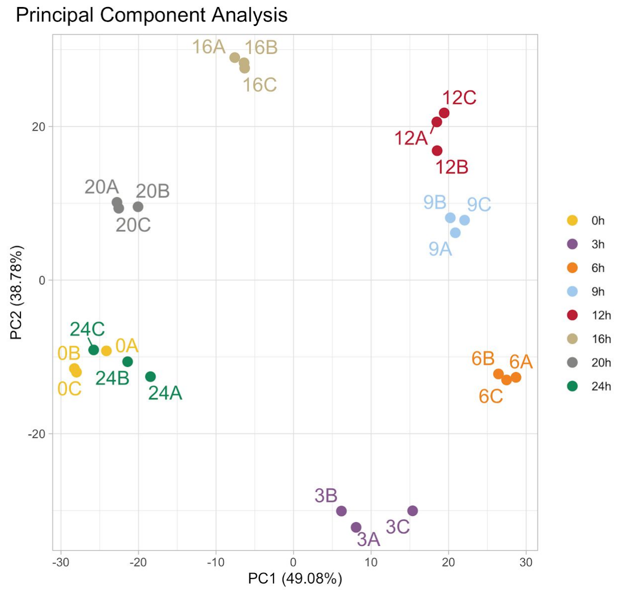
Distance between two samples: euclidean, correlation, Manhattan, SERE ...

Aggregation criterion (i.e. distance between two clusters):

- Average linkage: **average distance** between all the samples
- Single linkage: distance between the two **closest** samples
- Complete linkage: distance between the two **furthest** samples
- Ward: merge the clusters that lead to the cluster with **minimum variance**



Data exploration : Take-home message



Always visualize your data first !

To detect early on potential problems in the design

To guide you through the next steps of the analysis

To provide some biological interpretation

To communicate your results

Don't overlook potential breach of hypothesis for the analysis methods, or choices of parameters

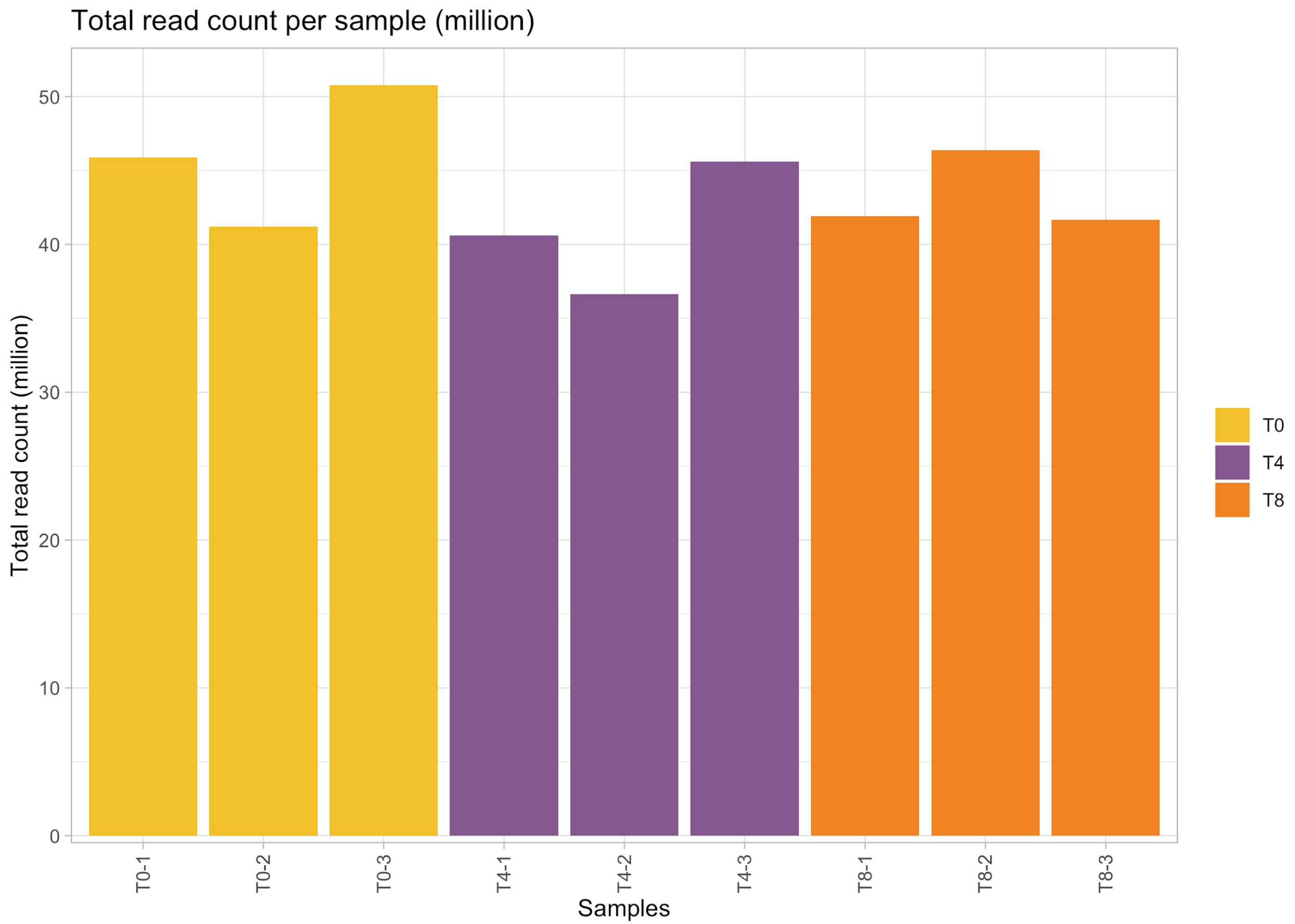
INSTITUT PASTEUR

Outline

1. Introduction
2. Designing the experiment
3. Description/exploration
4. **Normalization**
5. Modeling
6. SARTools

Goal

Identify and correct for systematic technical bias and make the counts comparable between samples.



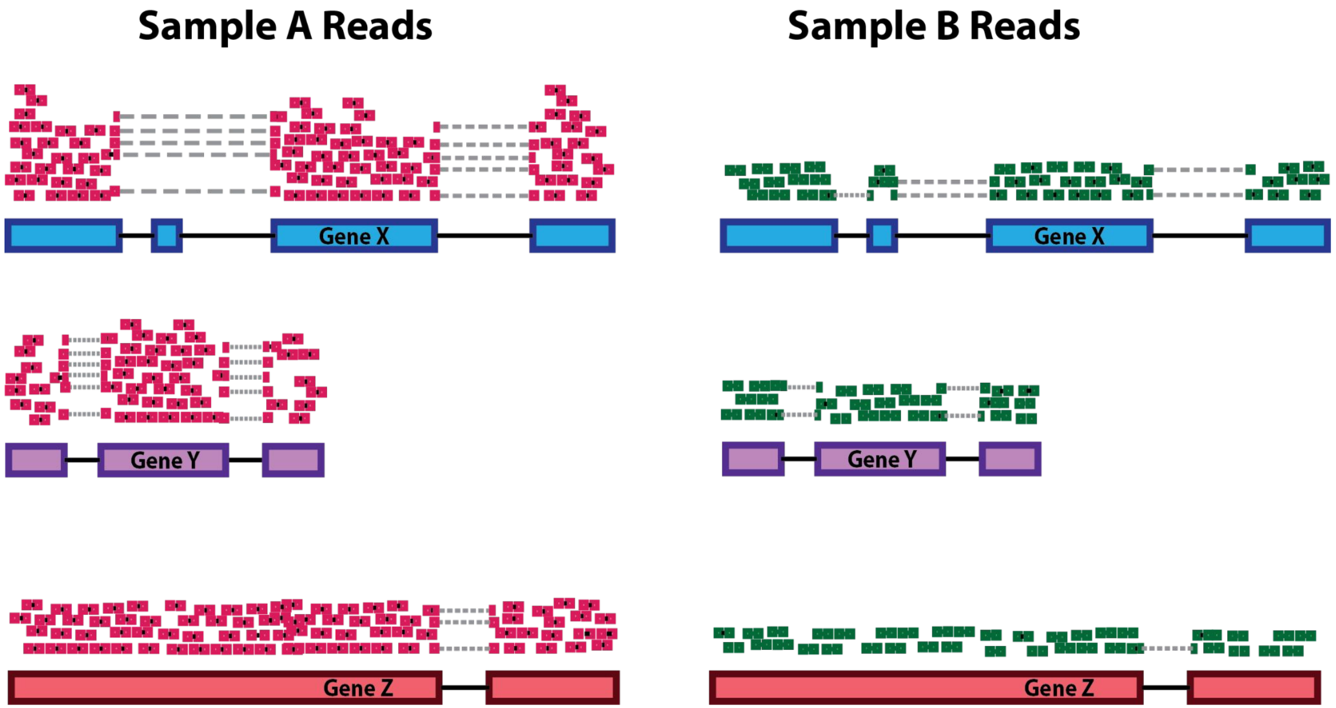
Framework

Normalization framework:
- RNA-seq data
- Differential expression experiment
- Counts data (positive integer values)

Total number of reads (library size): number of reads sequenced, mapped and counted for a given sample (sum over the rows for a given column of the count matrix).

Goal of the RNA-Seq normalizations

1. Correct for the differences of library sizes:

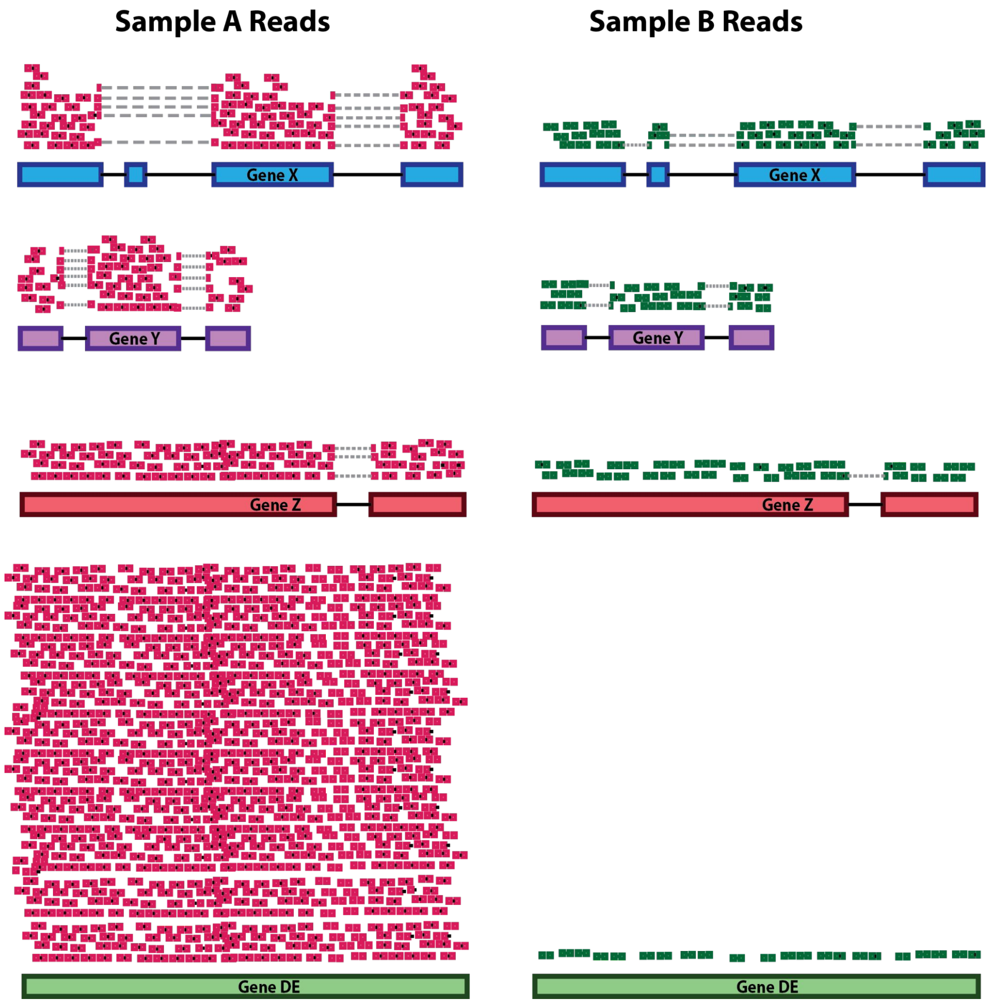


	Sample 1	Sample 2
gene1	30	60
gene2	50	100
gene3	20	40
gene4	100	200
Total	**200**	**400**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

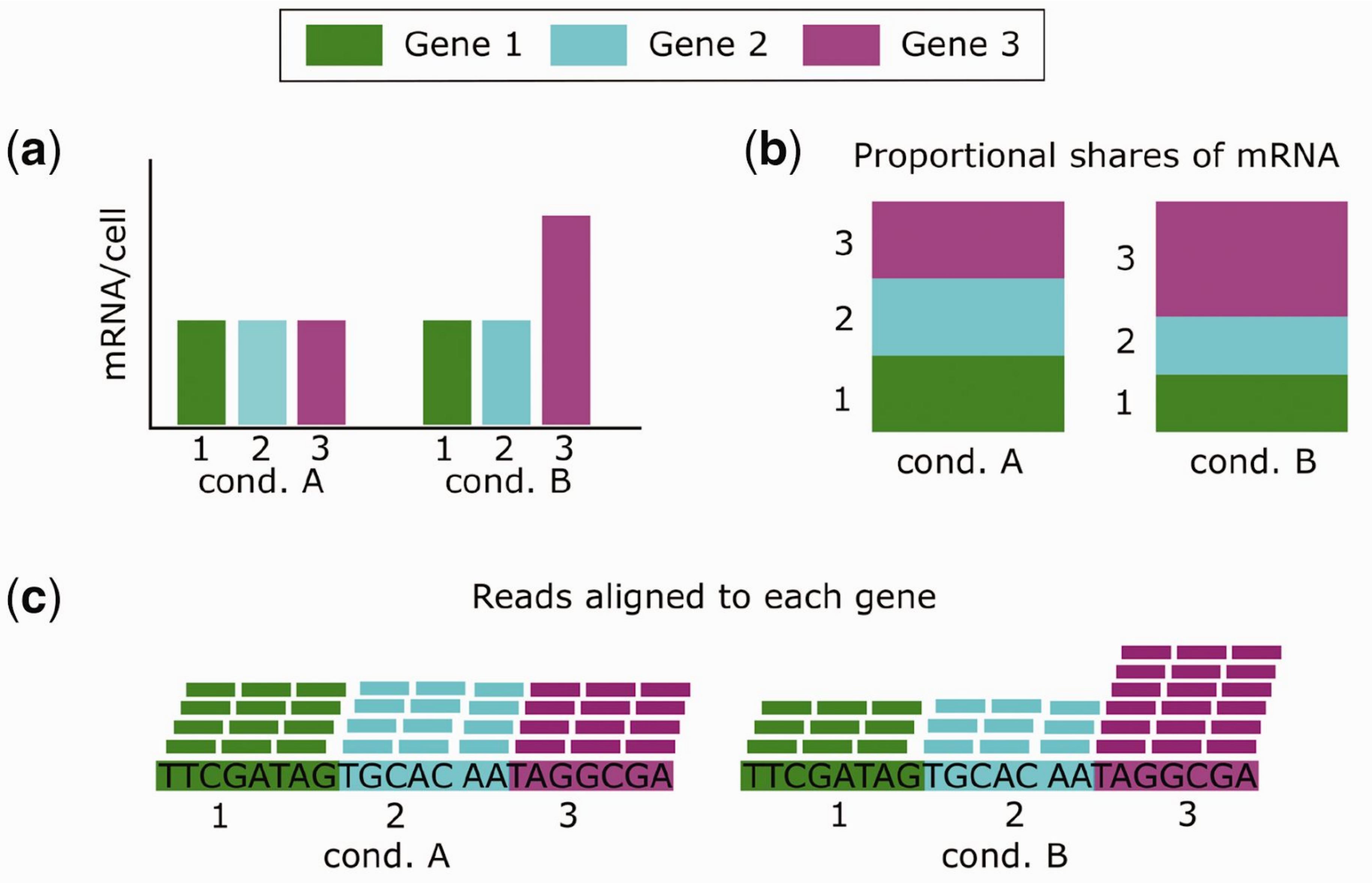
Goal of the RNA-Seq normalizations

2. Correct for the differences of RNA compositions:



	Sample 1	Sample 2
gene1	30	60
gene2	50	100
gene3	20	40
gene4	100	0
Total	**200**	**200**

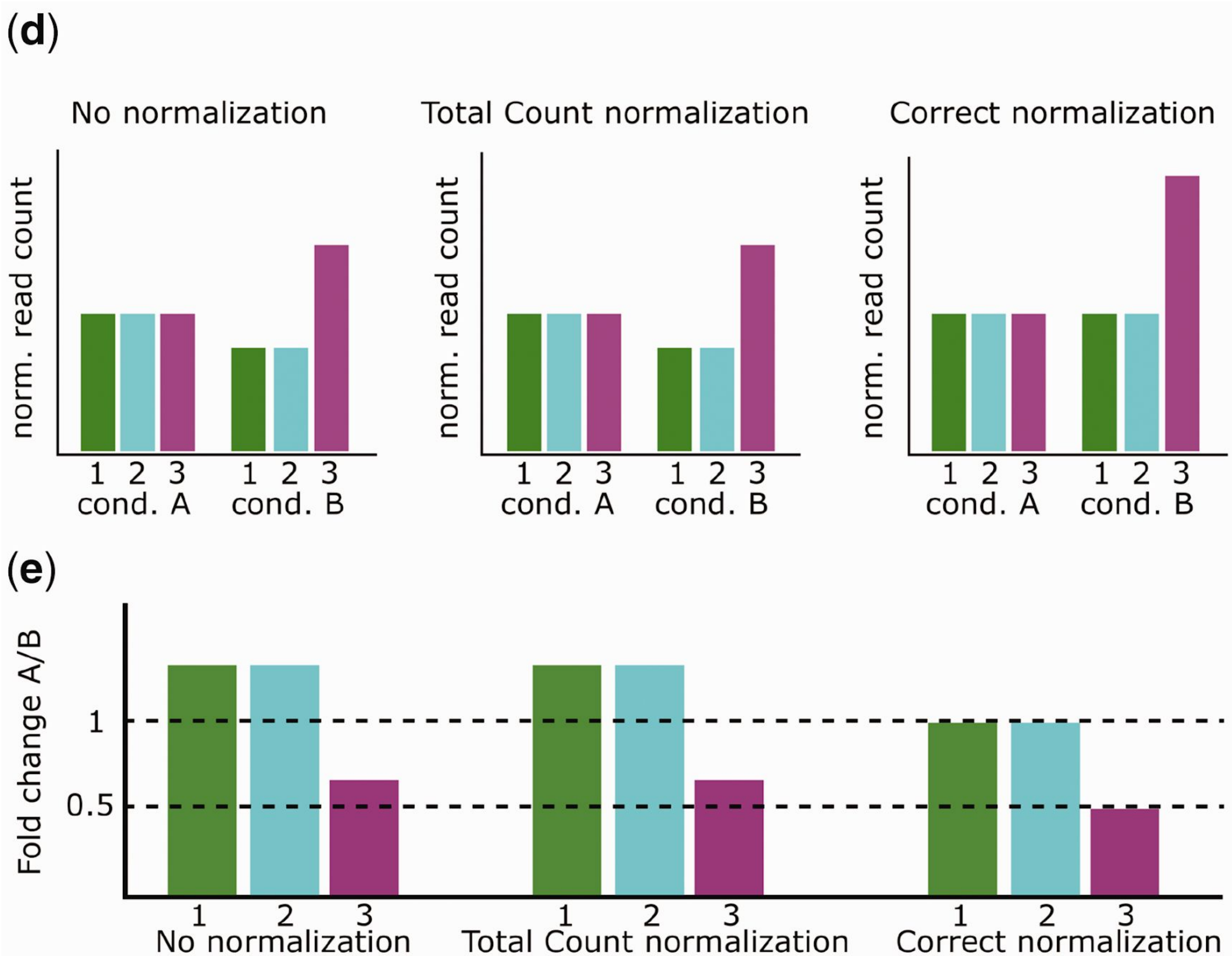
What is a differentially expressed gene? [10]



C. Evans et al. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics, 2017.

INSTITUT PASTEUR

What is a differentially expressed gene? [10]



C. Evans et al. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics, 2017.

DESeq2 normalization: computation of s_1

DESeq2 computes a **size factor** per sample:

Step 1 : Creating a pseudo-reference sample (row-wise *geometric mean*)

	T0-1	**T0-5**	...	**T8-3**
gene1	151	131	...	18
gene2	142	134	...	151
gene3	157	147	...	8
gene4	275	249	...	62
gene5	4	5	...	3
gene6	2	0	...	3
gene7	4	7	...	0
gene8	10	16	...	23
gene9	12	20	...	9
gene10	269	262	...	48
...	...	...	...	...
geneN	18	31	...	2

DESeq2 normalization: computation of s_1

DEseq2 computes a **size factor** per sample:

Step 1 : Creating a pseudo-reference sample (row-wise *geometric mean*)

	T0-1	T0-5	...	T8-3	pseudo-ref
gene1	(151 x	131 x	... x	18)$^{1/n}$	31
gene2	142	134	...	151	
gene3	157	147	...	8	
gene4	275	249	...	62	
gene5	4	5	...	3	
gene6	2	0	...	3	
gene7	4	7	...	0	
gene8	10	16	...	23	
gene9	12	20	...	9	
gene10	269	262	...	48	
...	...	...	...	...	
geneN	18	31	...	2	

INSTITUT PASTEUR

DESeq2 normalization: computation of s_1

DESeq2 computes a **size factor** per sample:

Step 1 : Creating a pseudo-reference sample (row-wise *geometric mean*)

	T0-1	T0-5	...	T8-3	pseudo-ref
gene1	151	131	...	18	31
gene2	142	134	...	151	650
gene3	157	147	...	8	7
gene4	275	249	...	62	70
gene5	4	5	...	3	2
gene6	2	0	...	3	1
gene7	4	7	...	0	5
gene8	10	16	...	23	28
gene9	12	20	...	9	74
gene10	269	262	...	48	112
...	...	...	...	...	...
geneN	18	31	...	2	4

INSTITUT PASTEUR

DESeq2 normalization: computation of s_1

DESeq2 computes a **size factor** per sample:

Step 2 : Comparing each sample to pseudo-reference (ratio)

	T0-1	T0-5	...	T8-3	pseudo-ref	T0-1 / ref
gene1	151	131	...	18	31	4.87
gene2	142	134	...	151	650	0.22
gene3	157	147	...	8	7	22.43
gene4	275	249	...	62	70	3.93
gene5	4	5	...	3	2	2.00
gene6	2	0	...	3	1	2.00
gene7	4	7	...	0	5	0.80
gene8	10	16	...	23	28	0.36
gene9	12	20	...	9	74	0.16
gene10	269	262	...	48	112	2.40
...	...	...	...	...	...	...
geneN	18	31	...	2	4	4.87

DESeq2 normalization: computation of s_1

DESeq2 computes a **size factor** per sample:

Step 3 : Final size factor (median)

	T0-1	T0-5	...	T8-3	pseudo-ref	T0-1 / ref
gene1	151	131	...	18	31	4.87
gene2	142	134	...	151	650	0.22
gene3	157	147	...	8	7	22.43
gene4	275	249	...	62	70	3.93
gene5	4	5	...	3	2	2.00
gene6	2	0	...	3	1	2.00
gene7	4	7	...	0	5	0.80
gene8	10	16	...	23	28	0.36
gene9	12	20	...	9	74	0.16
gene10	269	262	...	48	112	2.40
...	...	...	...	...	...	...
geneN	18	31	...	2	4	4.87

s_1 = median

INSTITUT PASTEUR

DESeq2 normalization: computation of s_1

Normalized count : $x'_{ij} = \frac{x_{ij}}{s_j}$

Step 1 : geometric mean of each gene

Step 2 : ratio between sample and reference

	T0-1	T0-5	...	T8-3	$(\prod_{k=1}^{n} x_{ik})^{\frac{1}{n}}$	$\frac{x_{ij}}{(\prod_{k=1}^{n} x_{ik})^{\frac{1}{n}}}$
gene1	151	131	...	18	31	4.87
gene2	142	134	...	151	650	0.22
gene3	157	147	...	8	7	22.43
gene4	275	249	...	62	70	3.93
gene5	4	5	...	3	2	2.00
gene6	2	0	...	3	1	2.00
gene7	4	7	...	0	5	0.80
gene8	10	16	...	23	28	0.36
gene9	12	20	...	9	74	0.16
gene10	269	262	...	48	112	2.40
...	...	...	...	...	...	...
geneN	18	31	...	2	4	4.87

Step 3 : median

s_1 = median

DESeq2 normalization [3]

Size factor s_j per sample:

- x_{ij}: number of reads for gene *i* in sample *j*
- *n*: number of samples studied
- s_j : normalization factor for sample *j*

$$s_j = \text{median}_i \frac{x_{ij}}{\left(\prod_{k=1}^{n} x_{ik}\right)^{\frac{1}{n}}}$$

Normalized counts:

$$x'_{ij} = \frac{x_{ij}}{s_j}$$

Assumptions:

1. The majority of the genes is not differentially expressed
2. As many down- as up-regulated genes

edgeR normalization [4]

edgeR computes a normalization factor f_j per sample and normalizes the total numbers of reads N_j:

- x_{ij}: number of reads for gene *i* in sample *j*
- N_j: total number of reads in sample *j* (lib size)
- n: number of samples studied
- s_j or f_j: normalization factor for sample *j*
- L_i: length of gene *i*

$$N'_j = f_j \times N_j$$

We can calculate DESeq2-like size factors s_j in order to normalize the counts:

$$s_j = \frac{N'_j}{\frac{1}{n}\sum_k N'_k} \quad \text{and so} \quad x'_{ij} = \frac{x_{ij}}{s_j}$$

Assumptions: same than DESeq2.

INSTITUT PASTEUR

Other normalization methods

Total number of reads:

or

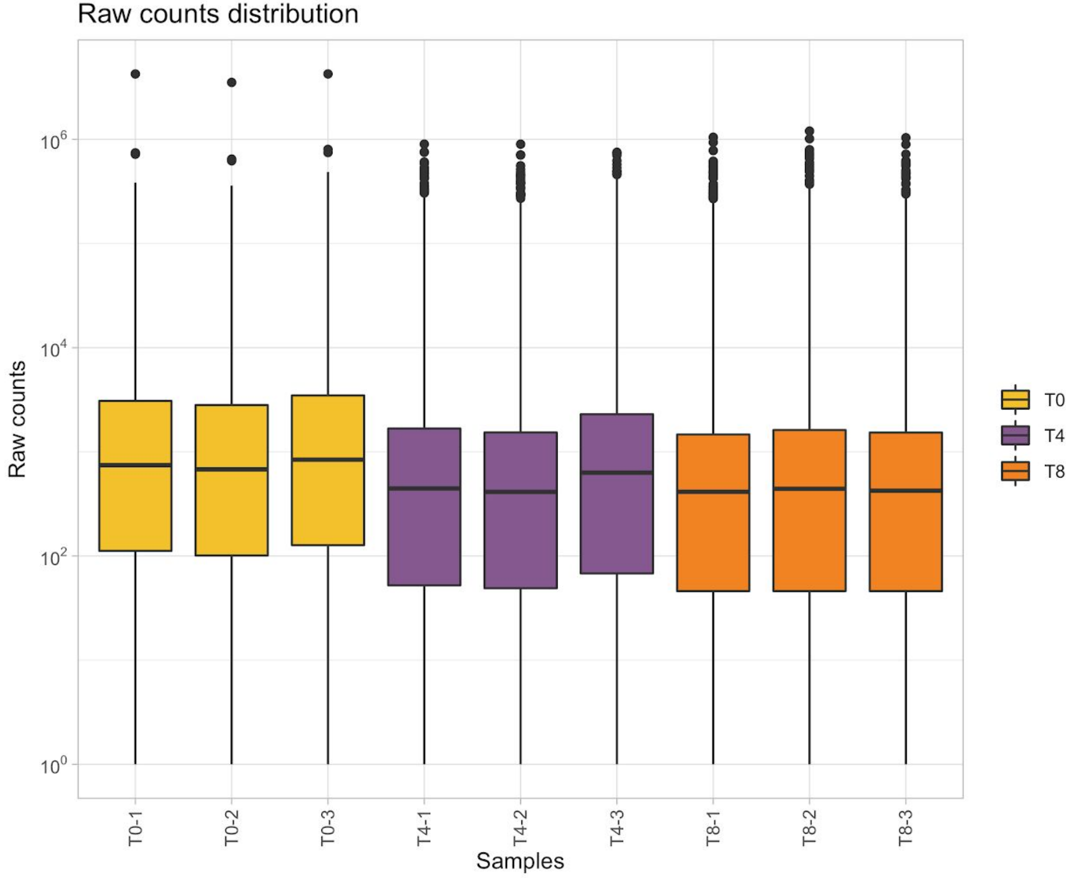
Robustness issue if a gene catches a very high number of reads.

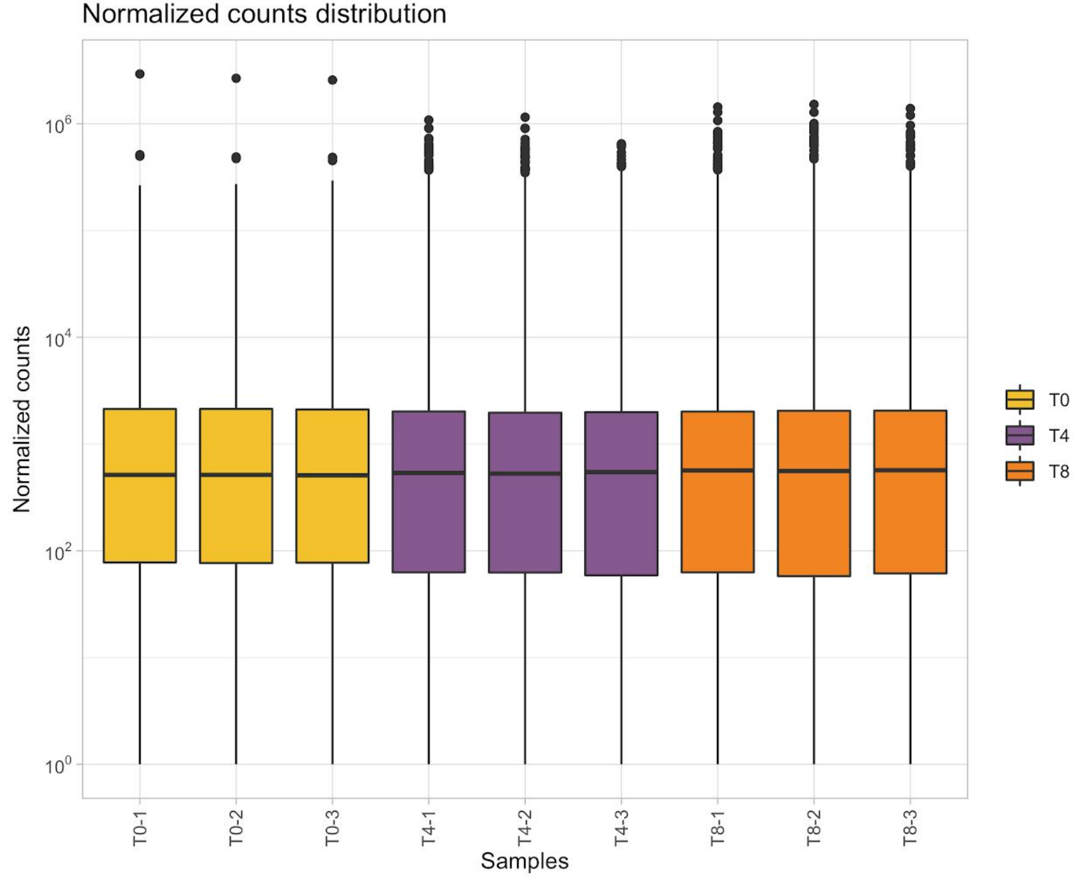
RPKM (Reads Per Kilobase per Million mapped reads):

$$x'_{ij} = \frac{x_{ij}}{N_j \times L_i} \times 10^6 \times 10^3$$

- Same issue than the total number of reads method
- Introduce other biases [5]
- No need to correct for the gene length since the gene is "fixed"

Effect of the normalization (DESeq2 or edgeR)





Outline

1. Introduction
2. Designing the experiment
3. Description/exploration
4. Normalization
5. **Modeling**
6. SARTools

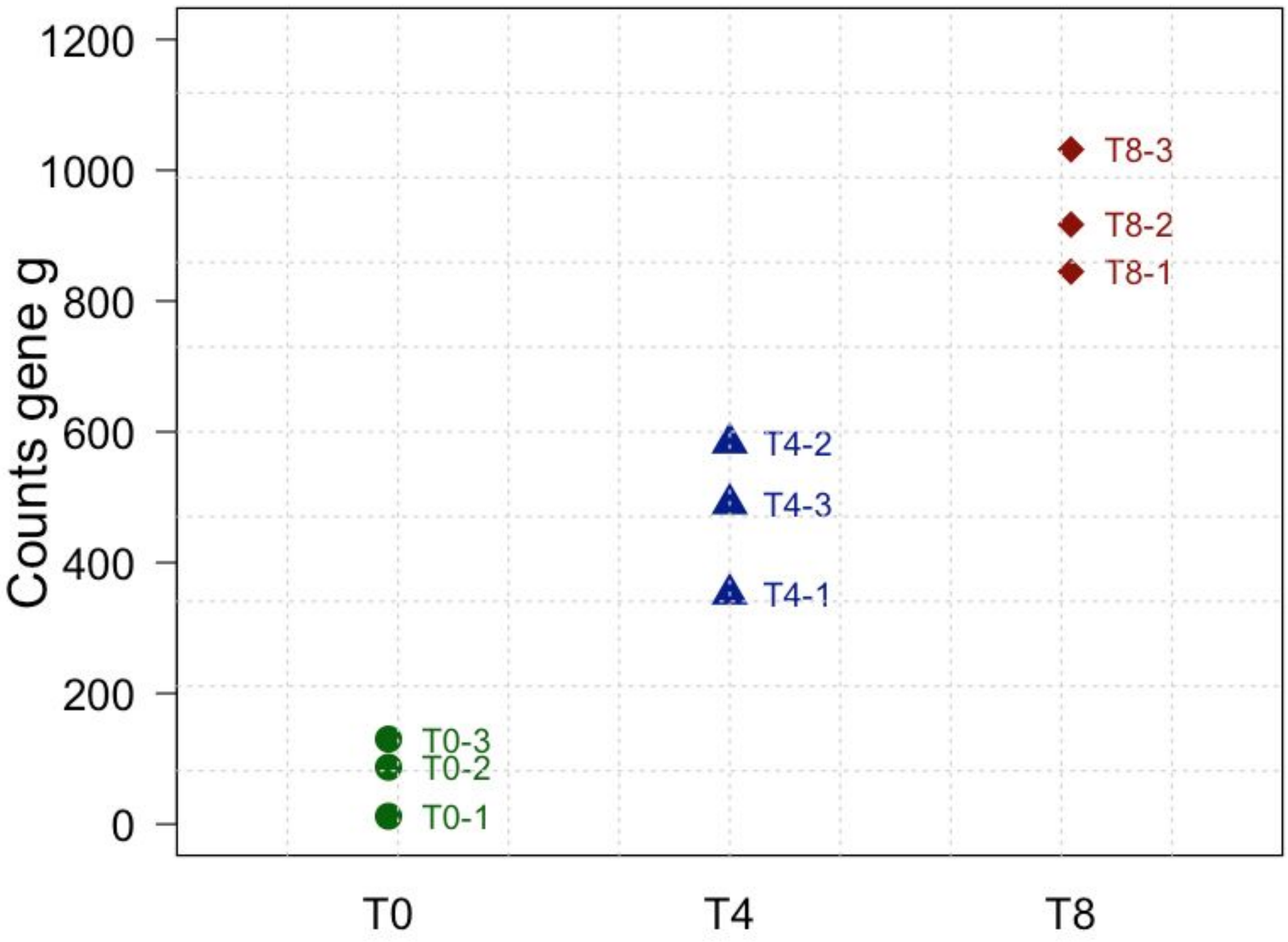
Classic linear model

Goal:
Explain a dependent variable Y thanks to a set a explicative variables $X = (X_1 , ..., X_n)$ using the model:

$$Y \sim X\beta + \varepsilon$$

Output of the model:
Estimations of $\beta_1, ..., \beta_n$: effect of each explicative variable on Y.

Linear model: RNA-Seq example



Goal: explain counts of gene *g* thanks to the biological conditions.

Linear model: RNA-Seq example

Goal: explain counts of gene *g* thanks to the bio. conditions (T0, T4 and T8).

$$\log_2 \begin{pmatrix} 12 \\ 87 \\ 130 \\ 352 \\ 583 \\ 490 \\ 845 \\ 917 \\ 1032 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \beta_{0g} \\ \beta_{1g} \\ \beta_{2g} \end{pmatrix} + \begin{pmatrix} \epsilon_{g1} \\ \epsilon_{g2} \\ \epsilon_{g3} \\ \epsilon_{g4} \\ \epsilon_{g5} \\ \epsilon_{g6} \\ \epsilon_{g7} \\ \epsilon_{g8} \\ \epsilon_{g9} \end{pmatrix}$$

Here: $\hat{\beta}_{0g} = 5.95$, $\hat{\beta}_{1g} = 2.91$ and $\hat{\beta}_{2g} = 3.57$

One model per gene → thousands of models!

Statistical testing

	Green1	Green2	Green3	Gray1	Gray2	Gray3
Gene g	151	131	183	135	184	122

Biological question:

Is gene *g* differentially expressed between green and gray mice?

Statistical testing

	Green1	Green2	Green3	Gray1	Gray2	Gray3
Gene g	151	131	183	135	184	122

Biological question:

Is gene *g* differentially expressed between green and gray mice?

Statistical notation

Let μ_1 the average expression of gene *g* for gray mice and μ_2 the expression of green mice. We wish to test the hypotheses:

$$H_0: \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1: \mu_1 \neq \mu_2$$

How to decide ?

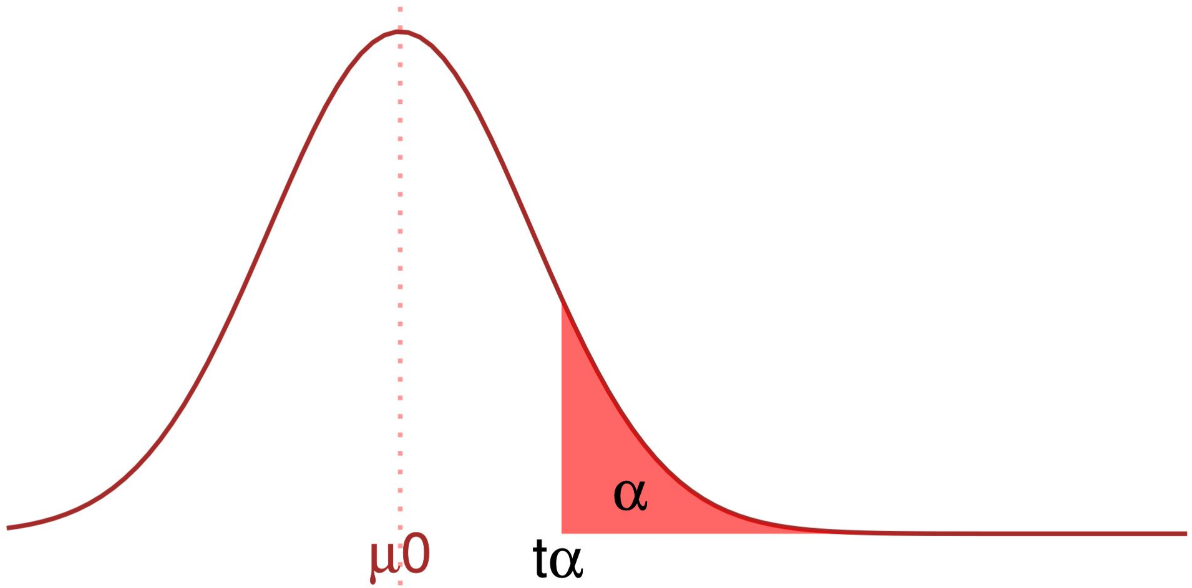
Type I error rate: α

Framework and goal:

We wish to show that the expression of gene *g* of gray mice is different from the expression of green mice.

Which **risk α** of being wrong do we allow when saying :

"**gene *g* is differentially expressed?**"



The risk α is chosen **before the analysis**

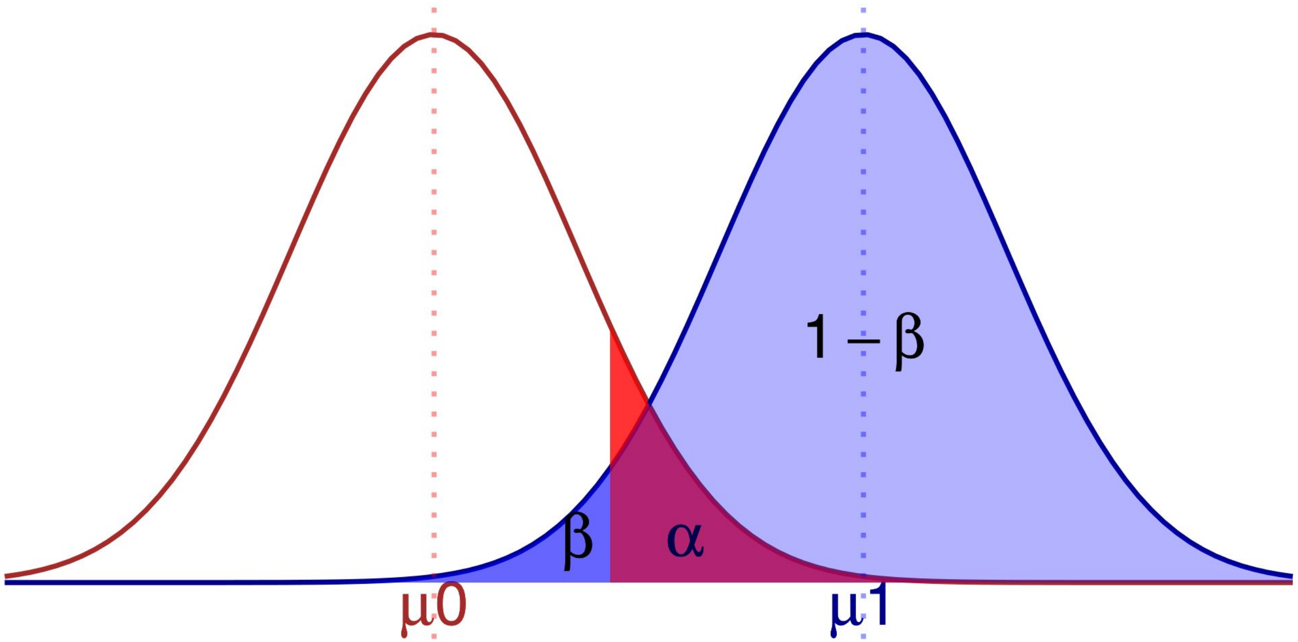
INSTITUT PASTEUR

Type II error rate: β

We assume that gene *g* is truly differentially expressed between gray and green mice.

- Which risk β of not discovering gene *g* do we allow?
- Which power $1 - \beta$ do we want?

We can theoretically control the risk β according to the risk α and the number of replicates.



Type I and type II errors

Hotdog classification

Type I error

True negative

False positive

False negative

True positive

Type II error

Formalization

Let μ_1 the average expression of gene *g* for gray mice and μ_2 the expression of green mice. We wish to test the hypotheses:

$$H_0: \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1: \mu_1 \neq \mu_2$$

The risks can be summarized in:

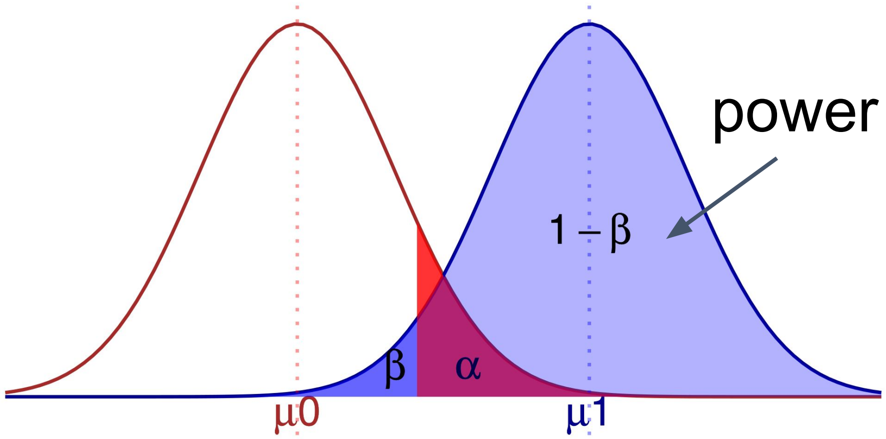
Decision

Unknown truth

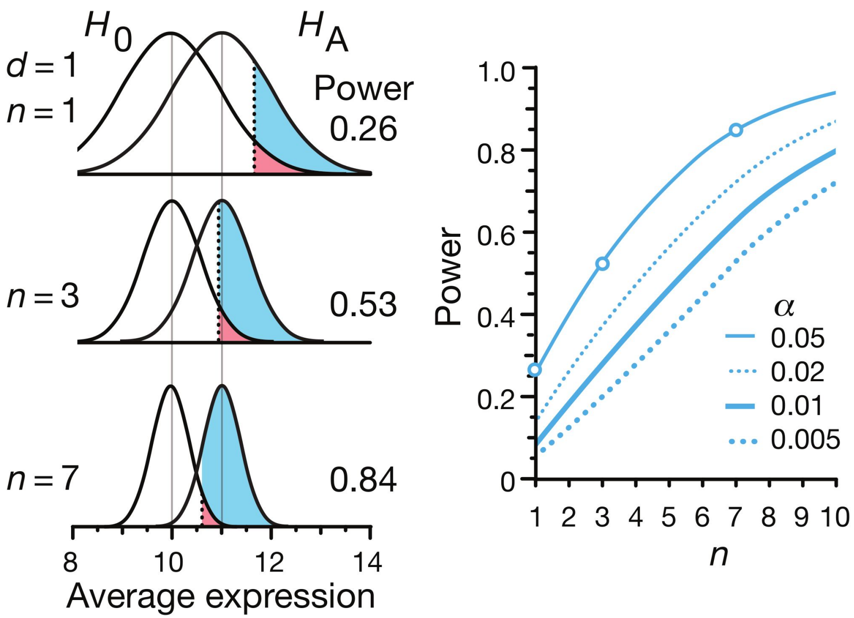
	Do not reject H_0	Reject H_0
H_0 true	$1 - \alpha$ TP	α FN
H_0 false	β FP	$1 - \beta$ TN

INSTITUT PASTEUR

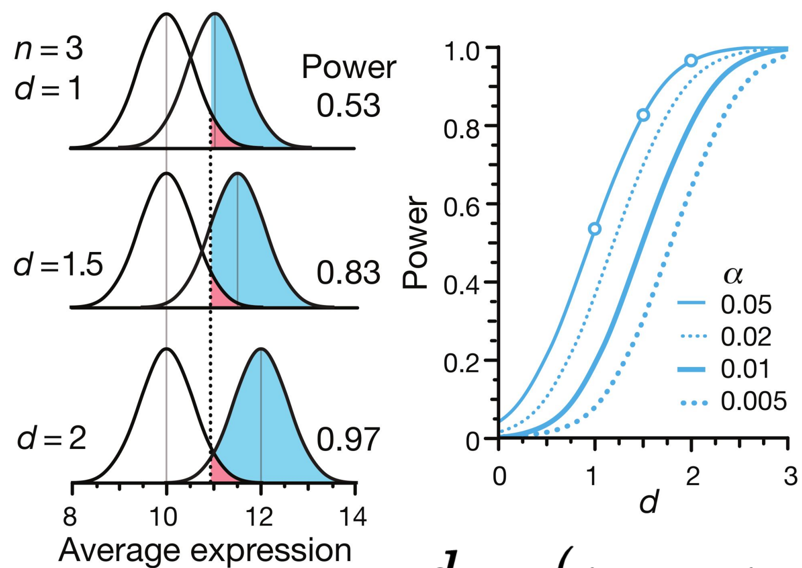
Statistical Power



power increases with sample size



and with effect size !



$$d = (\mu_1 - \mu_0)/(\sigma)$$

p-value and conclusion of the test

Definition:

p-value = Proba(reject H_0 | H_0 true)

= Proba(doing a mistake when rejecting H_0)

= Proba(observed difference is due to hazard)

Conclusion:

if *p*-value $\leq \alpha$ then we reject H_0

With a risk α, we can conclude that there is a significant difference in gene *g* expression between green and gray mice

INSTITUT PASTEUR

Equal Fold-Changes – different *p*-values

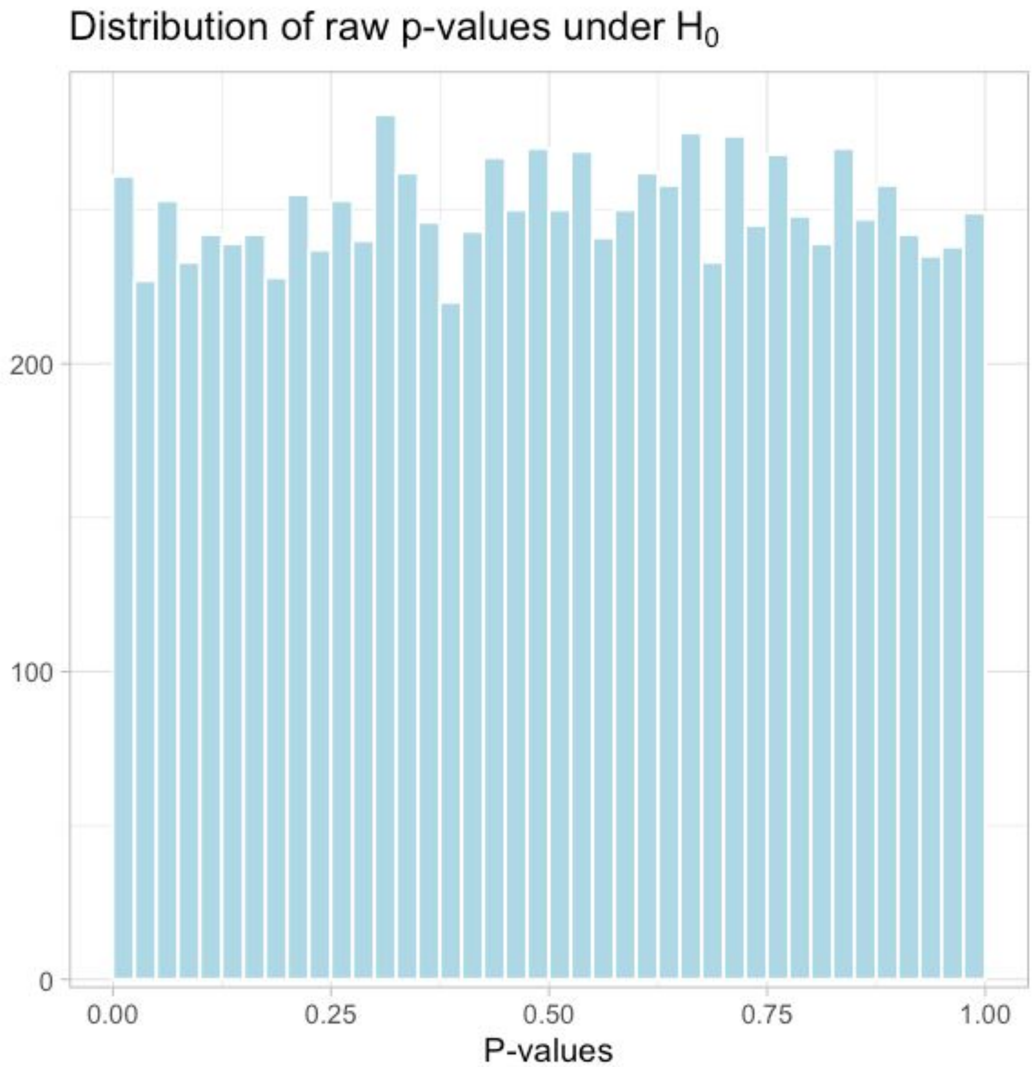
Reminder: Fold-Change definition:

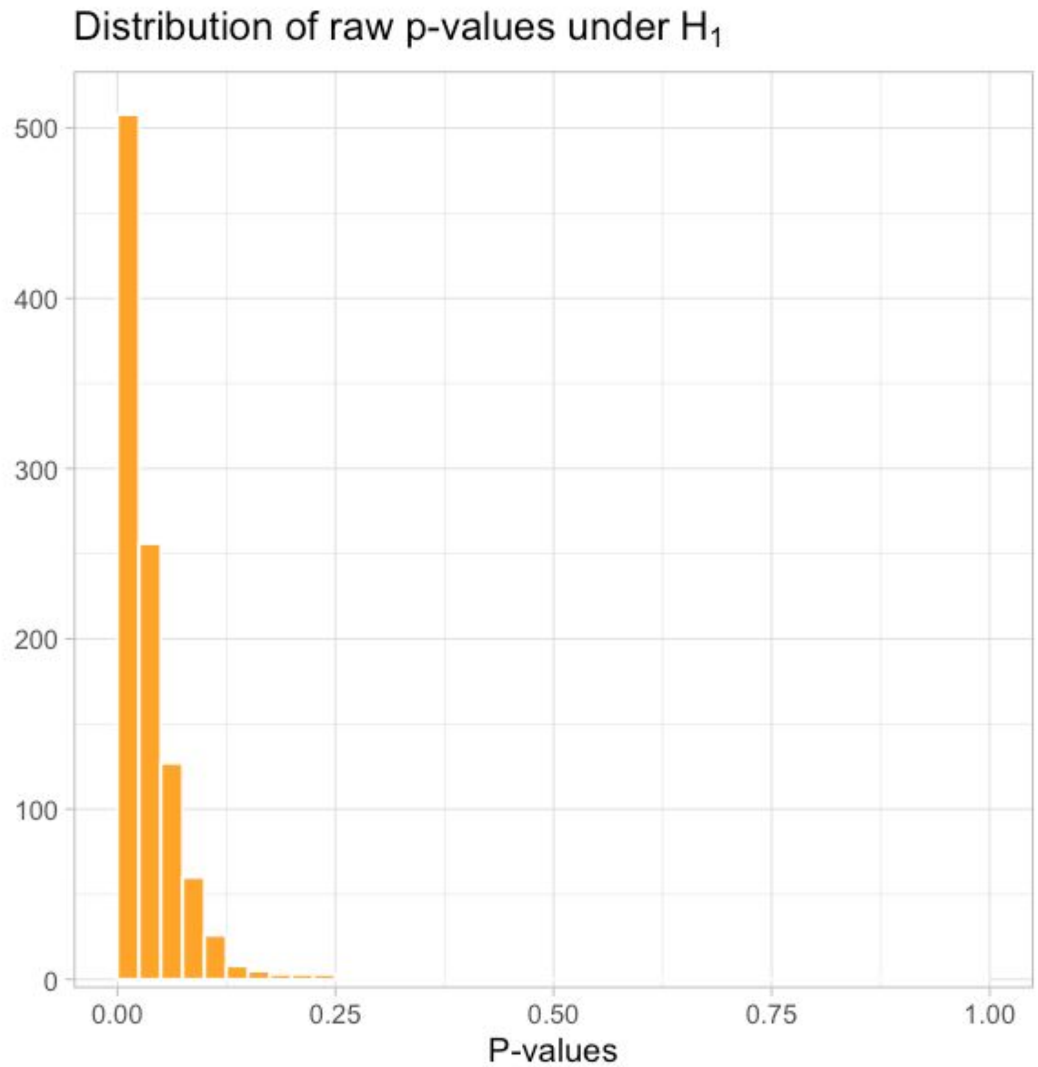
$$FC = \frac{\text{expression condition “green”}}{\text{expression condition “gray”}} = \frac{\mu_2}{\mu_1}$$

Gene	m1	m2	m3	m4	m5	m6	FC	*p*-value
gene1	5	7	6	2	2	2	3	0.06
gene2	800	1000	900	350	250	200	3	0.03
gene3	700	900	1100	350	200	250	3	0.10
gene4	900	500	1300	200	550	50	3	0.06
...	...	...	...	...	...	...	...	...

INSTITUT PASTEUR

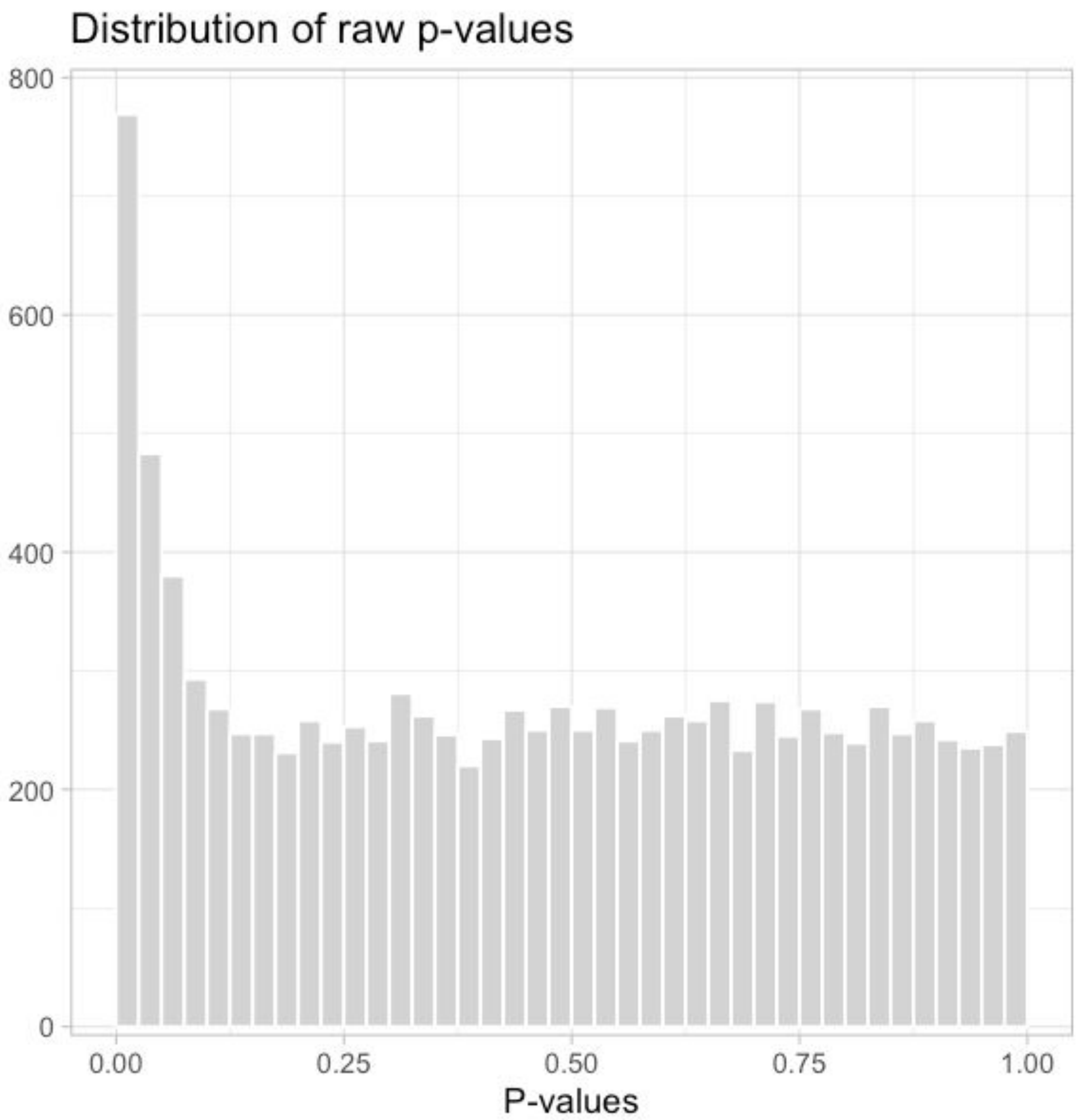
Distribution of raw *p*-values





INSTITUT PASTEUR

Distribution of raw *p*-values



Omics data: multiple testing issue

Context:

We perform a large number *N* of statistical tests for which we reject or not H_0.

Possible conclusions:

Decisions

Unknown truths

	Non rejects of H_0	Rejects of H_0
H_0 true	TN	FP
H_0 false	FN	TP

Among all the genes told differentially expressed, the False Discovery Rate (FDR) is:

$$\frac{FP}{FP + TP}$$

Example of the multiple testing issue

We perform N = 10000 statistical tests and we get the following conclusions:

	Non rejects of H_0	Rejects of H_0	Total
H_0 true	8550	450	9000
H_0 false	200	800	1000
Total	8750	1250	10000

$$\frac{FP}{FP + TP} = \frac{450}{450 + 800} = 36\% \text{ of falsely discovered genes!}$$

Control of the FDR

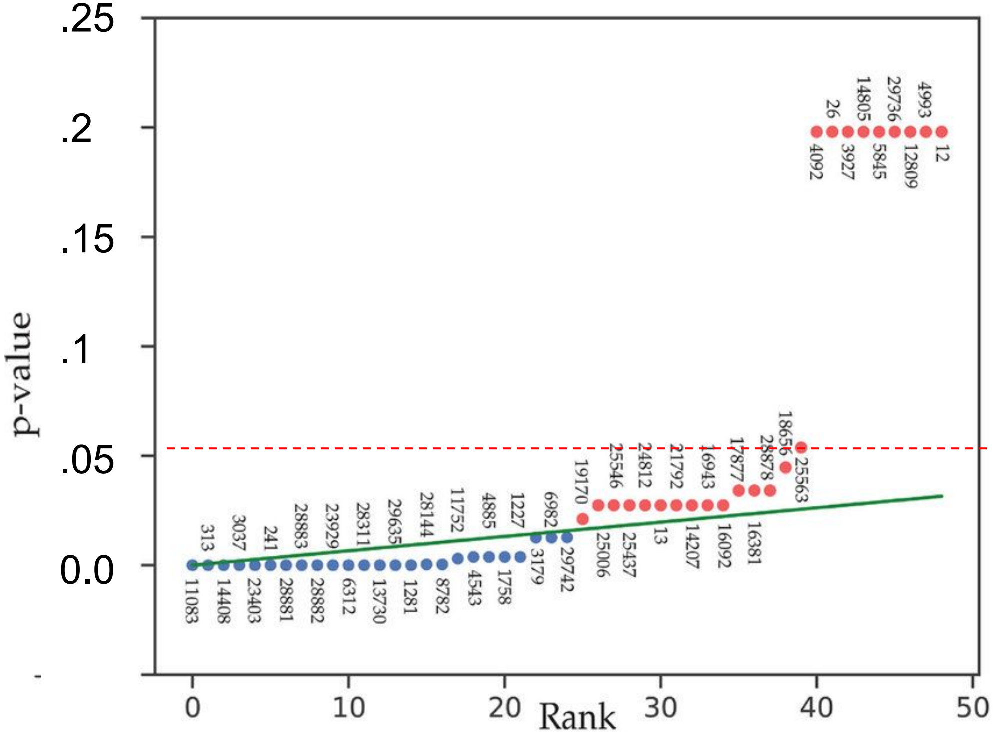
Goal: control the FDR among the list of differentially expressed genes.

(Very strong) assumption: all the *N* statistical tests are independent.

Procedure: The Benjamini & Hochberg [6] algorithm transforms the *N* raw *p*-values in *N* adjusted p-values.

Conclusion:

if adjusted *p*-value ≤ α then we reject H_0



Importance of the # of biological replicates

RNA-Seq specificity: often 2 or 3 replicates because of the high cost of the experiment … But it's not ideal !

With more biological replicates...

- Better estimation of:
  - the variability present in the populations studied
  - the difference between the biological conditions
- Better control of the FDR: bad control with only 2 replicates [7]
- Higher statistical power: we detect more easily genes which are truly differentially expressed

At the very least : 3 replicates !

DESeq2 [3] and edgeR [4,8]

Three main steps:

1. Normalization
2. Dispersion (i.e. variability) estimation: crucial step
3. Statistical tests and adjustment for multiple testing

Advantages:

- User friendly and very well documented
- Good performances
- Authors are reactive on web forums and mailing lists

Similarities:

- Negative Binomial distribution
- Generalized Linear Model (GLM)

Differences:

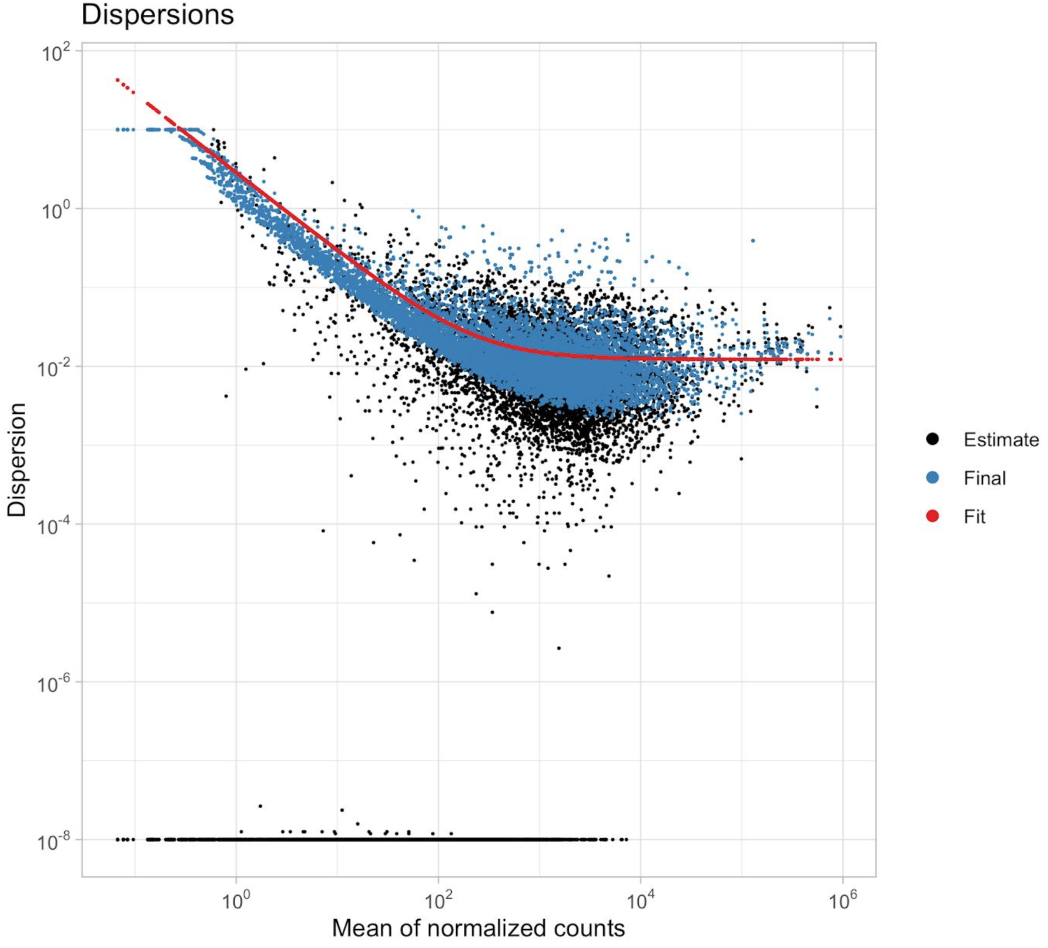
- Dispersion estimation
- Way of dealing with outlier counts
- Low counts filtering

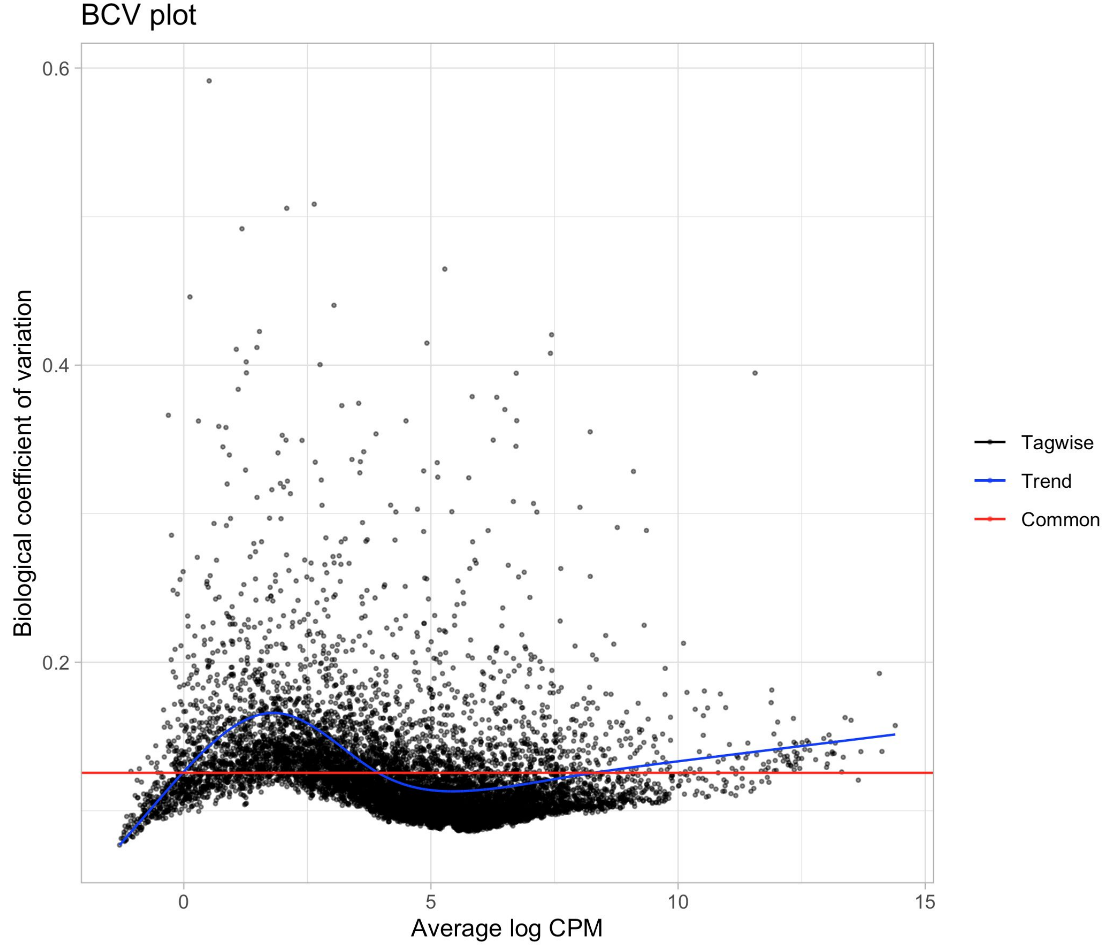
Many other tools exist: NBPSeq, TSPM, baySeq, EBSeq, NOISeq, SAMseq, ShrinkSeq, voom(+limma)

Dispersion estimation φ_i: DESeq2 vs edgeR

Reminder:

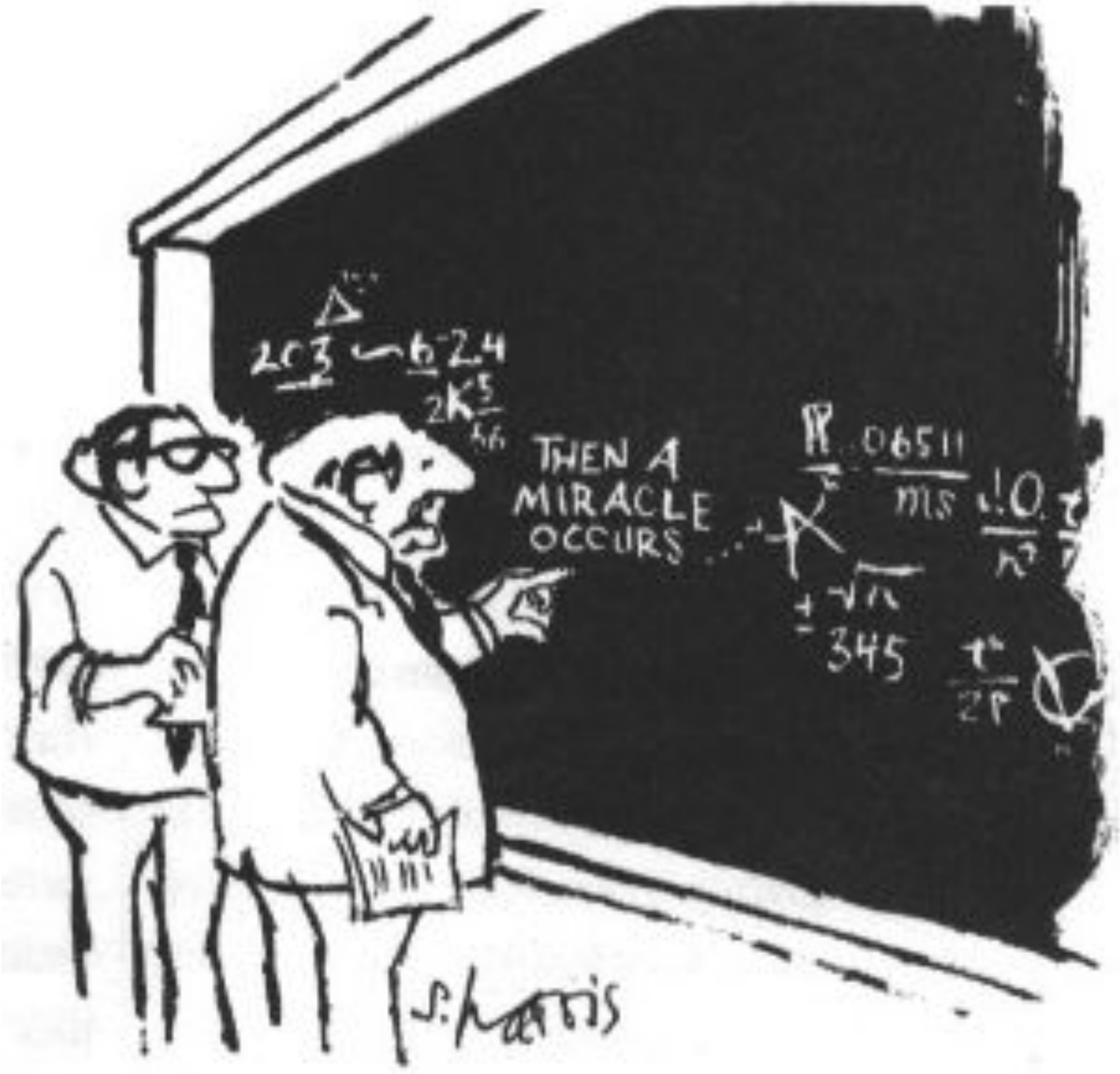
$$x_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2 = \mu_{ij} + \boldsymbol{\varphi_i}\, \mu_{ij}^2)$$





INSTITUT PASTEUR

Statistical theory and parameters tuning



"I think you should be more explicit here in step two."

INSTITUT PASTEUR

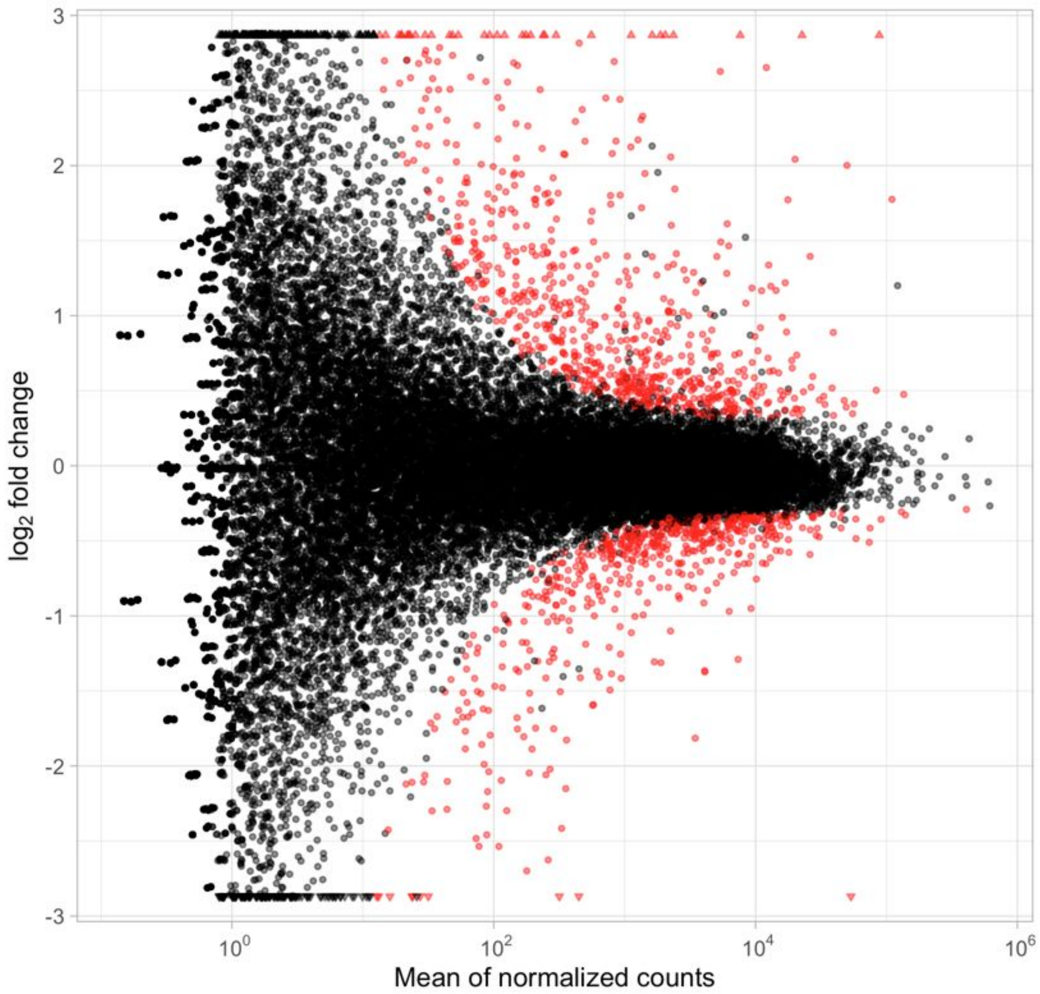
Statistical testing

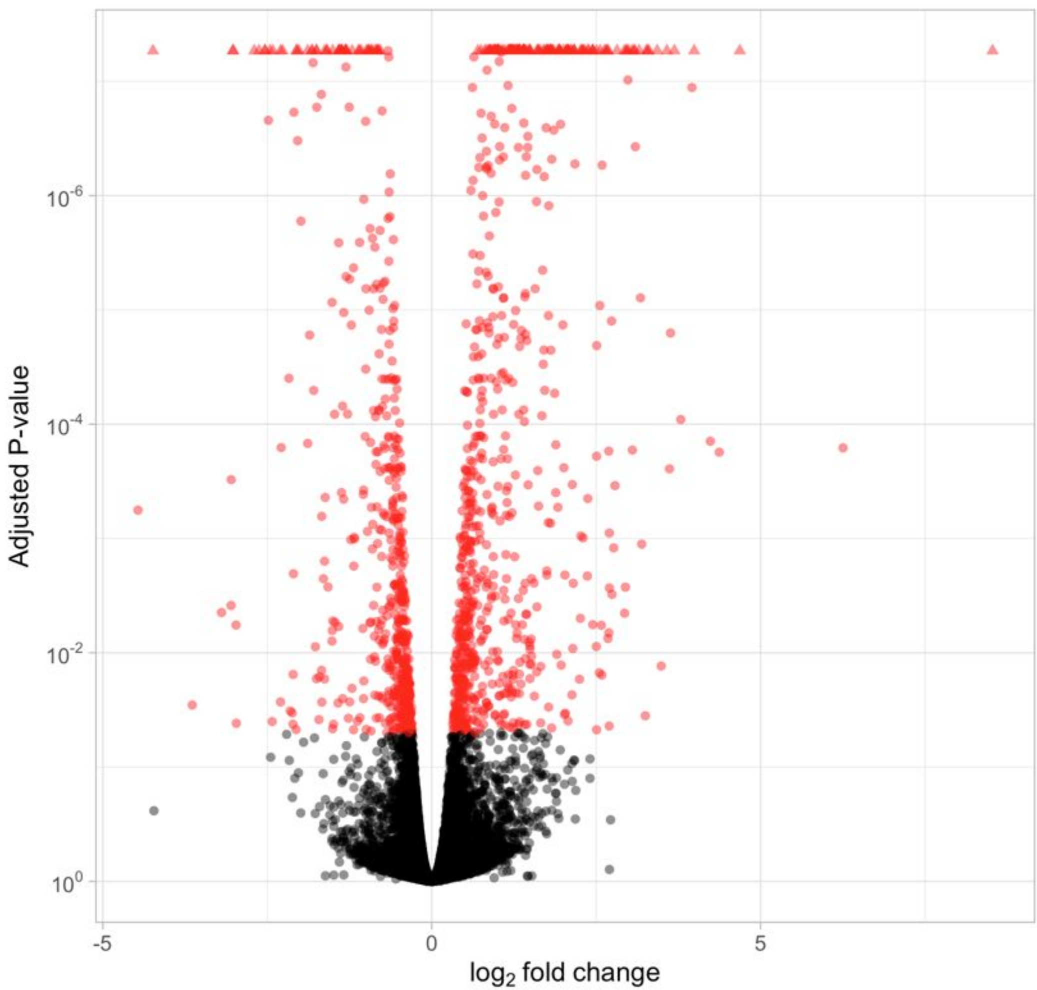
For each gene *g*, DESeq2 and edgeR give:

- an estimation of $\beta_g = \log_2(FC_g)$
- the precision of this estimation (standard error)
- so the *p*-value associated with gene *g*

The set of the *N* *p*-values is adjusted in order to conclude.

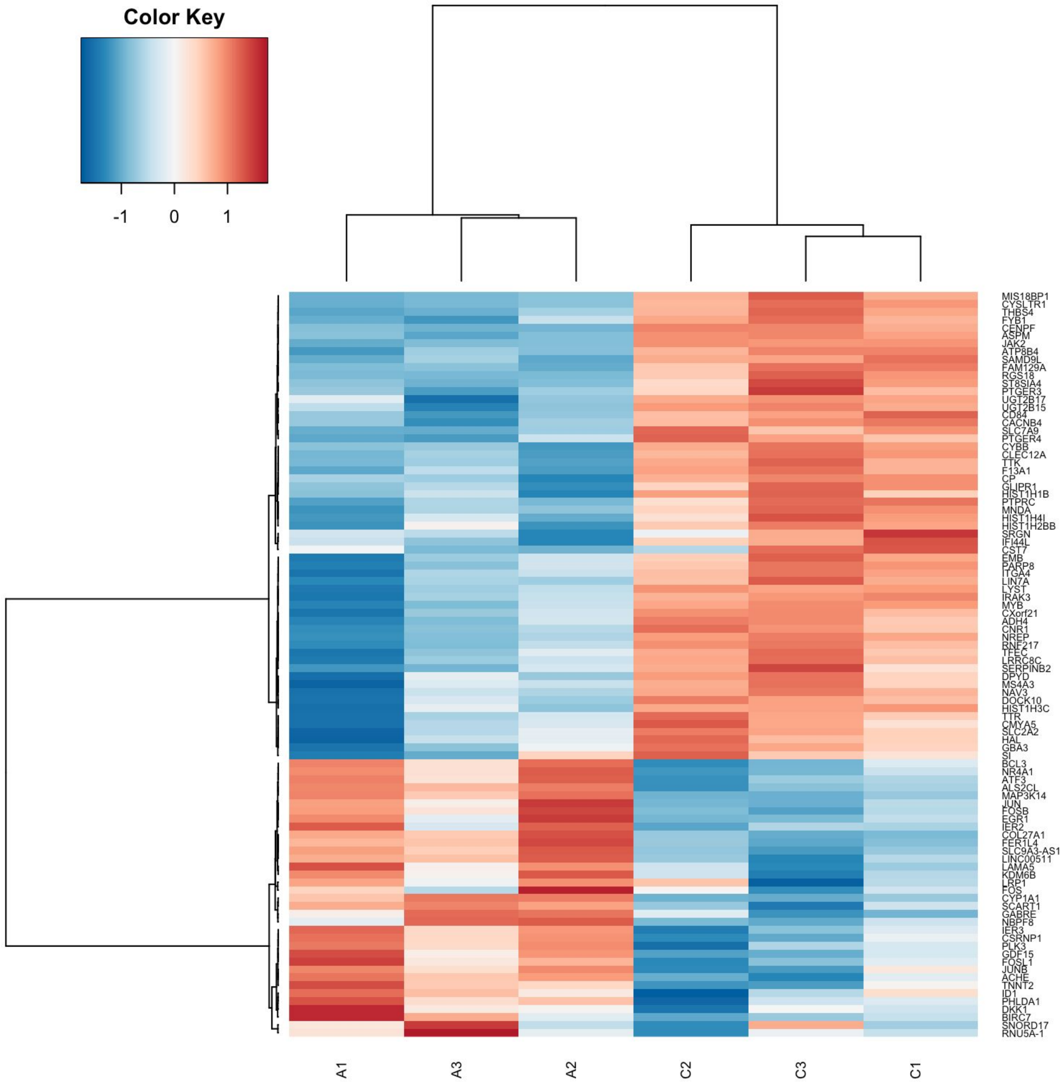
Description of the results: MA-plot and volcano-plot





INSTITUT PASTEUR

Description of the results: heatmap



Much more complex than it appears:

- Use expression data or $\log_2(FC)$?
- Which genes to display?

- Expression data transformation:
  - Homoscedasticity?
  - Row centering and scaling?

- Row/column clustering method?

- Average data by condition?
- Batch/replicate effect removal?

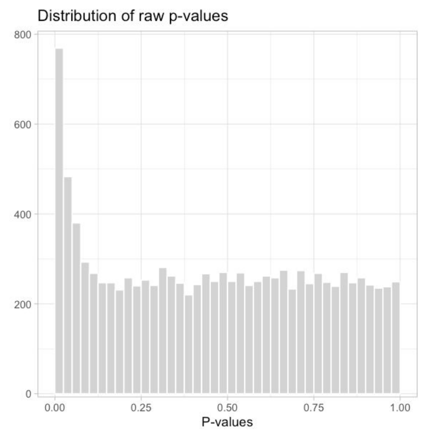
Data normalization & modelling : Take-home message

Data normalization is crucial to make sure you are really testing your biological question by removing systematic bias. Specific RNASeq methods must take into account library size & composition.

Multiple testing must be corrected using FDR as many tests are done simultaneously

Replicate your measures according to the expected variability in the data and the differences you want to highlight

Visualize your results and use diagnostic plots to check that the model / test you chose was adapted to your data.



Outline

1. Introduction
2. Designing the experiment
3. Description/exploration
4. Normalization
5. Modeling
6. **SARTools**

Why SARTools?

SARTools = Statistical Analysis of RNA-Seq Tools [9]

1. Perform a systematic quality control of the data
2. Avoid misusing the DESeq2 or edgeR packages
3. Keep track of all the parameters used: reproducible research
4. Provide a HTML report containing all the results of the analysis

`github.com/PF2-pasteur-fr/SARTools/`

Input files

Target: tab-delimited text file describing the experimental design:

label	files	condition
WT1	WT1.counts.txt	WT
WT2	WT2.counts.txt	WT
KO1	KO1.counts.txt	KO
KO2	KO2.counts.txt	KO

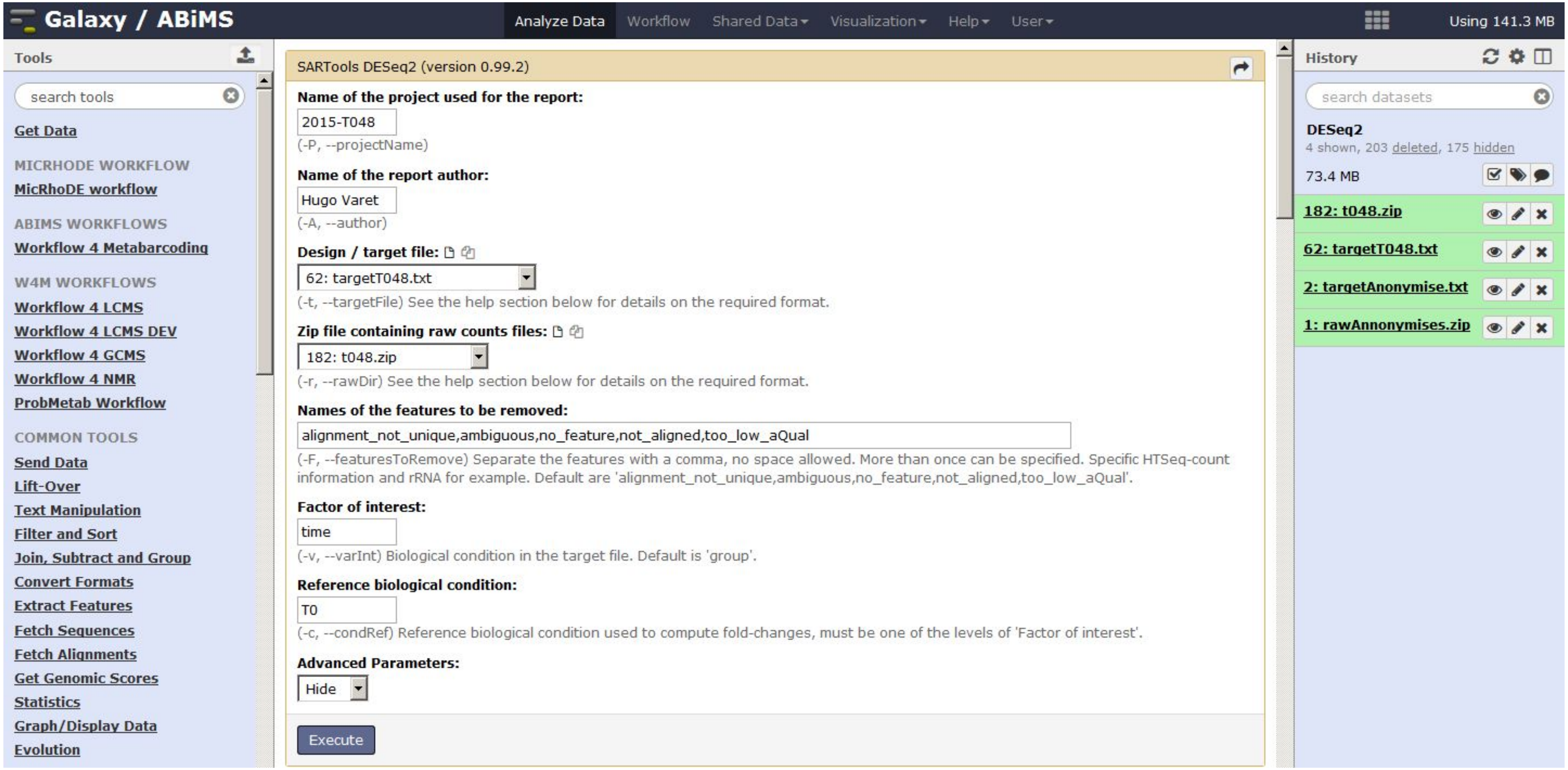
Counts: one tab-delimited text file per sample (from HTSeq-count or featureCounts):

```
gene1    23
gene2    355
gene3    0
...      ...
gene4    3643
```

Usage: with R

```r
################################################################################
###                 parameters: to be modified by the user                   ###
################################################################################
rm(list=ls())                                          # remove all the objects from the R session

workDir <- "C:/path/to/your/working/directory/"      # working directory for the R session

projectName <- "projectName"                           # name of the project
author <- "Your name"                                  # author of the statistical analysis/report

targetFile <- "target.txt"                             # path to the design/target file
rawDir <- "raw"                                        # path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",         # names of the features to be removed
                      "ambiguous", "no_feature",       # (specific HTSeq-count information and rRNA for example)
                      "not_aligned", "too_low_aQual")

varInt <- "group"                                      # factor of interest
condRef <- "WT"                                        # reference biological condition
batch <- NULL                                          # blocking factor: NULL (default) or "batch" for example

fitType <- "parametric"                                # mean-variance relationship: "parametric" (default) or "local"
cooksCutoff <- TRUE                                    # TRUE/FALSE to perform the outliers detection (default is TRUE)
independentFiltering <- TRUE                           # TRUE/FALSE to perform independent filtering (default is TRUE)
alpha <- 0.05                                          # threshold of statistical significance
pAdjustMethod <- "BH"                                  # p-value adjustment method: "BH" (default) or "BY"

typeTrans <- "VST"                                     # transformation for PCA/clustering: "VST" or "rlog"
locfunc <- "median"                                    # "median" (default) or "shorth" to estimate the size factors

colors <- c("dodgerblue","firebrick1",                 # vector of colors of each biological condition on the plots
            "MediumVioletRed","SpringGreen")
```

Usage: with Galaxy



Output: HTML report

Statistical report of project testdeseq2: pairwise comparison(s) of conditions with DESeq2

Hugo Varet

2017-12-11

The SARTools R package which generated this report has been developped at PF2 - Institut Pasteur by M.-A. Dillies and H. Varet (hugo.varet@pasteur.fr). Thanks to cite H. Varet, L. Brillet-Guéguen, J.-Y. Coppee and M.-A. Dillies, *SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data*, PLoS One, 2016, doi: http://dx.doi.org/10.1371/journal.pone.0157022 when using this tool for any analysis published.

1 Introduction

The analyses reported in this document are part of the testdeseq2 project. The aim is to find features that are differentially expressed between T0, T4 and T8. The statistical analysis process includes data normalization, graphical exploration of raw and normalized data, test for differential expression for each feature between the conditions, raw p-value adjustment and export of lists of features having a significant differential expression between the conditions.

The analysis is performed using the R software [1], Bioconductor [2] packages including DESeq2 [3,4] and the SARTools package developed at PF2 - Institut Pasteur. Normalization and differential analysis are carried out according to the DESeq2 model and package. This report comes with additional tab-delimited text files that contain lists of differentially expressed features.

For more details about the DESeq2 methodology, please refer to its related publications [3,4].

2 Description of raw data

The count data files and associated biological conditions are listed in the following table.

label	files	group	batch
T0-1	sampleT0-1-htseq.out	T0	1
T0-5	sampleT0-5-htseq.out	T0	2
T0-6	sampleT0-6-htseq.out	T0	3
T4-1	sampleT4-1-htseq.out	T4	1
T4-2	sampleT4-2-htseq.out	T4	2
T4-3	sampleT4-3-htseq.out	T4	3
T8-1	sampleT8-1-htseq.out	T8	1
T8-2	sampleT8-2-htseq.out	T8	2
T8-3	sampleT8-3-htseq.out	T8	3

Table 1: Data files and associated biological conditions.

Output: HTML report

6 R session information and parameters

The versions of the R software and Bioconductor packages used for this analysis are listed below. It is important to save them if one wants to re-perform the analysis in the same conditions.

- R version 3.4.1 (2017-06-30), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=fr_FR.UTF-8, LC_NUMERIC=C, LC_TIME=fr_FR.UTF-8, LC_COLLATE=fr_FR.UTF-8, LC_MONETARY=fr_FR.UTF-8, LC_MESSAGES=fr_FR.UTF-8, LC_PAPER=fr_FR.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=fr_FR.UTF-8, LC_IDENTIFICATION=C
- Running under: Ubuntu 16.04.3 LTS
- Matrix products: default
- BLAS: /usr/lib/libblas/libblas.so.3.6.0
- LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: Biobase 2.38.0, BiocGenerics 0.24.0, DelayedArray 0.4.1, DESeq2 1.18.1, edgeR 3.20.1, GenomeInfoDb 1.14.0, GenomicRanges 1.30.0, IRanges 2.12.0, limma 3.34.1, matrixStats 0.52.2, S4Vectors 0.16.0, SARTools 1.5.2, SummarizedExperiment 1.8.0, xtable 1.8-2
- Loaded via a namespace (and not attached): acepack 1.4.1, annotate 1.56.1, AnnotationDbi 1.40.0, backports 1.1.1, base64enc 0.1-3, BiocParallel 1.12.0, bit 1.1-12, bit64 0.9-7, bitops 1.0-6, blob 1.1.0, checkmate 1.8.5, cluster 2.0.6, colorspace 1.3-2, compiler 3.4.1, data.table 1.10.4-3, DBI 0.7, digest 0.6.12, evaluate 0.10.1, foreign 0.8-69, Formula 1.2-2, genefilter 1.60.0, geneplotter 1.56.0, GenomeInfoDbData 0.99.1, ggplot2 2.2.1, grid 3.4.1, gridExtra 2.3, gtable 0.2.0, Hmisc 4.0-3, htmlTable 1.9, htmltools 0.3.6, htmlwidgets 0.9, knitr 1.17, lattice 0.20-35, latticeExtra 0.6-28, lazyeval 0.2.1, locfit 1.5-9.1, magrittr 1.5, Matrix 1.2-10, memoise 1.1.0, munsell 0.4.3, nnet 7.3-12, plyr 1.8.4, RColorBrewer 1.1-2, Rcpp 0.12.13, RCurl 1.95-4.8, rlang 0.1.4, rmarkdown 1.8, rpart 4.1-11, rprojroot 1.2, RSQLite 2.0, scales 0.5.0, splines 3.4.1, stringi 1.1.6, stringr 1.2.0, survival 2.41-3, tibble 1.3.4, tools 3.4.1, XML 3.98-1.9, XVector 0.18.0, yaml 2.1.14, zlibbioc 1.24.0

Parameter values used for this analysis are:

- workDir: .
- projectName: testdeseq2
- author: Hugo Varet
- targetFile: target.txt
- rawDir: raw
- featuresToRemove: alignment_not_unique, ambiguous, no_feature, not_aligned, too_low_aQual
- varInt: group
- condRef: T0
- batch: NULL
- fitType: parametric
- cooksCutoff: TRUE
- independentFiltering: TRUE
- alpha: 0.05
- pAdjustMethod: BH
- typeTrans: VST
- locfunc: median
- colors: dodgerblue, firebrick1, MediumVioletRed, SpringGreen

Output: lists of differentially expressed genes

Three tab-delimited text files per comparison:

- `*.complete.txt`: all the genes
- `*.up.txt`: up-regulated genes ordered by adj. *p*-value
- `*.down.txt`: down-regulated genes ordered by adj. *p*-value

Columns: gene id, $\log_2$(Fold-Change), adjusted *p*-value, ...

HTML tutorial

SARTools vignette for the differential analysis of 2 or more conditions with DESeq2 or edgeR

SARTools version: `r packageVersion("SARTools")`

Authors: M.-A. Dillies and H. Varet (hugo.varet@pasteur.fr) - Transcriptome and Epigenome Platform, Institut Pasteur, Paris

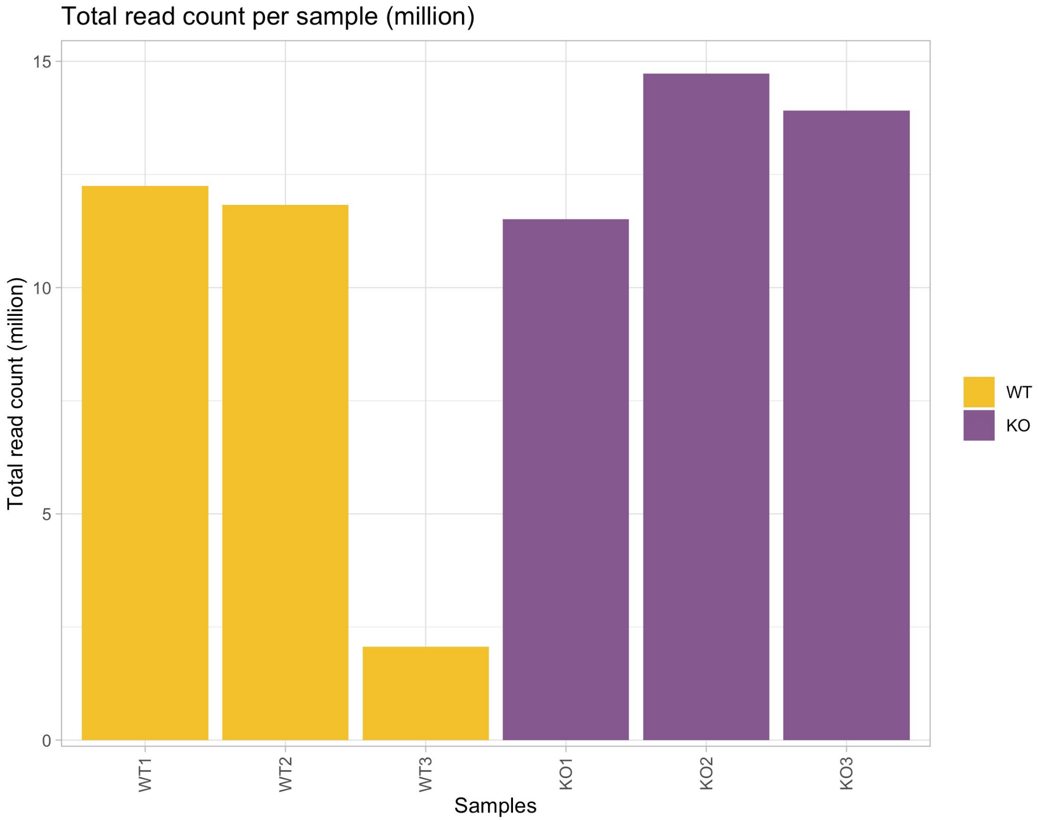
Website: https://github.com/PF2-pasteur-fr/SARTools

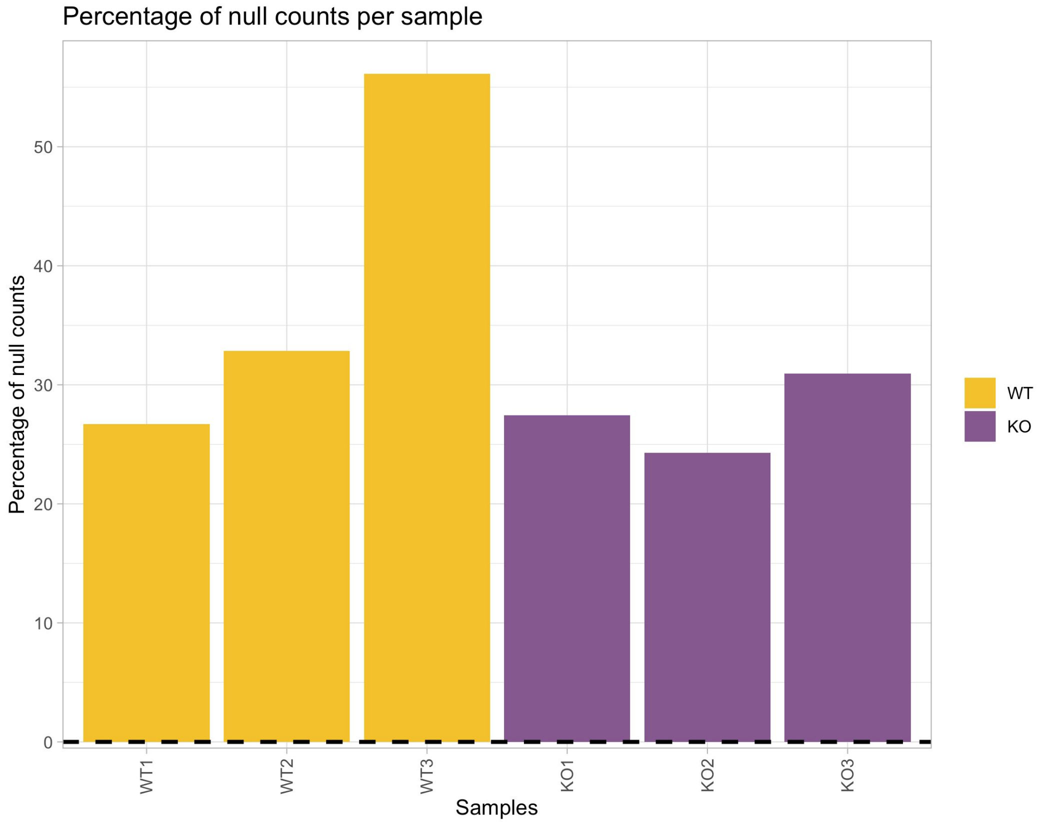
1 Introduction

This document aims to illustrate the use of the SARTools R package in order to compare two or more biological conditions in a RNA-Seq framework. SARTools provides tools to generate descriptive and diagnostic graphs, to run the

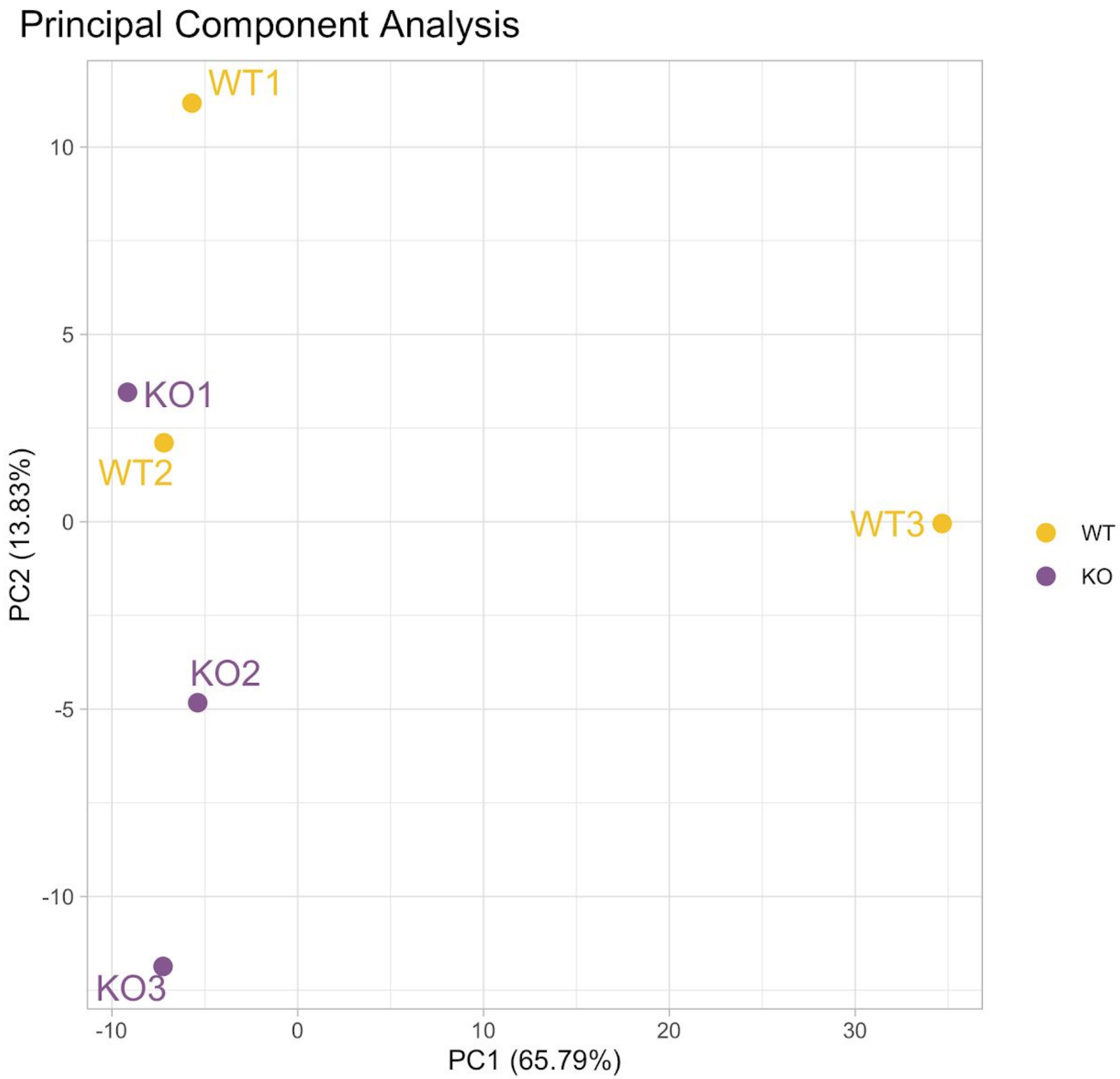
- Installation
- Input files
- Definition of the parameters
- Potential issues: technical problems, inversion of samples, batch effects, outliers...

Potential issue: detecting outliers



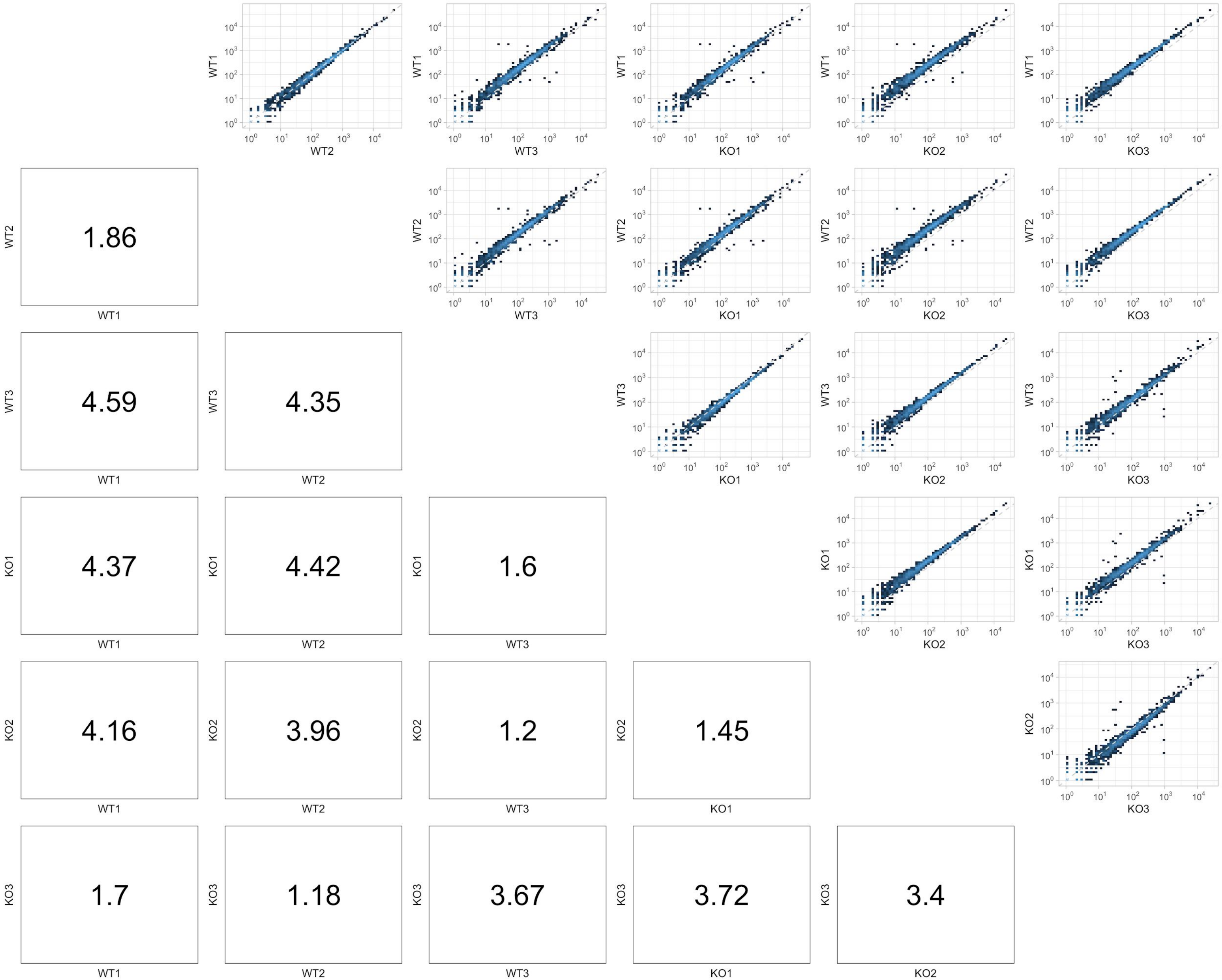


Potential issue: detecting outliers



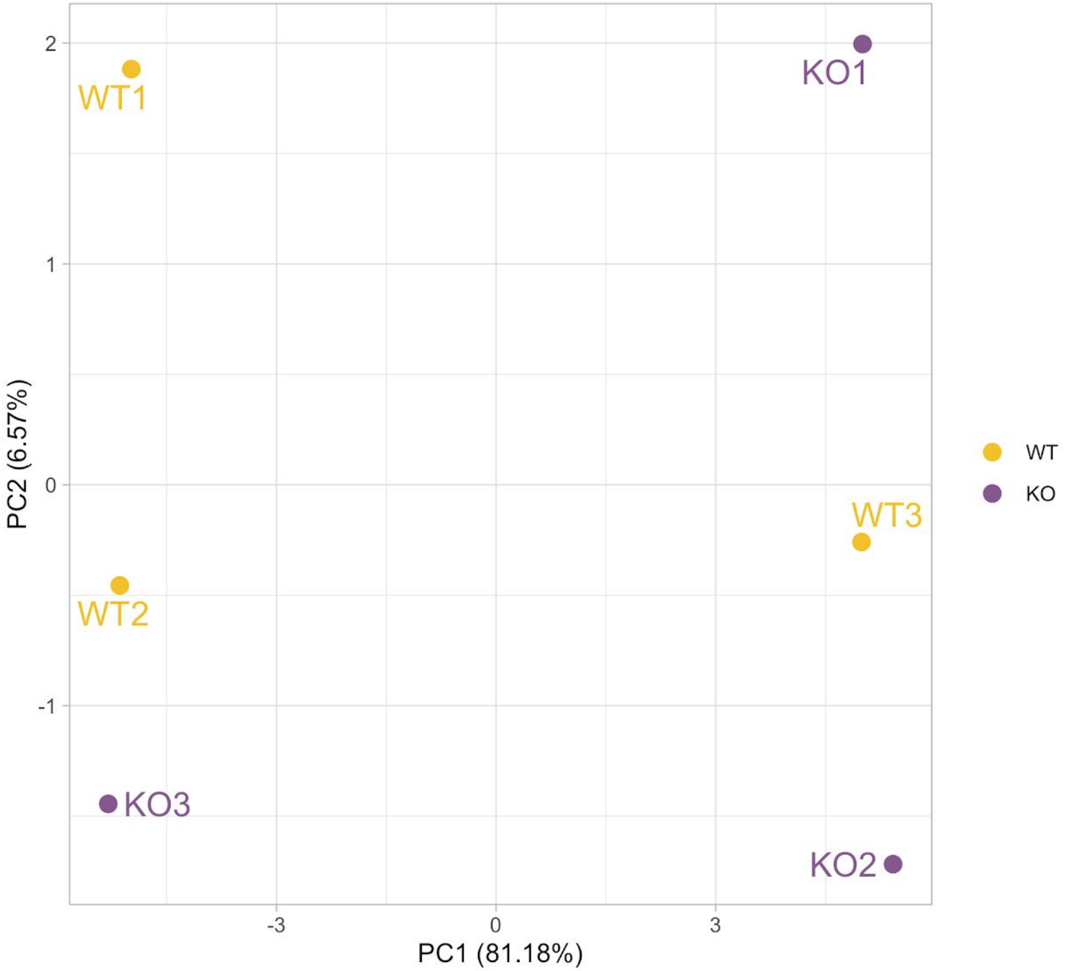
Potential issue: inversion of samples

Pairwise scatter plot

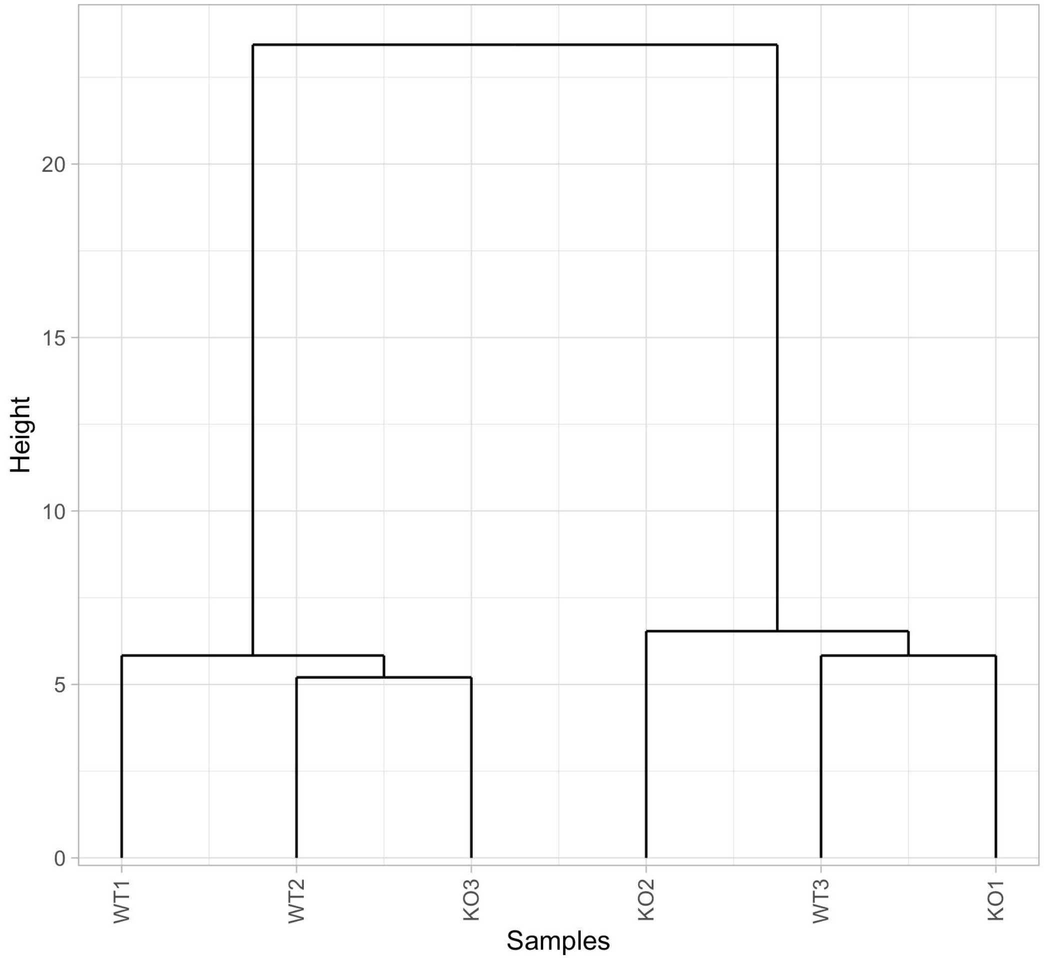


Potential issue: inversion of samples

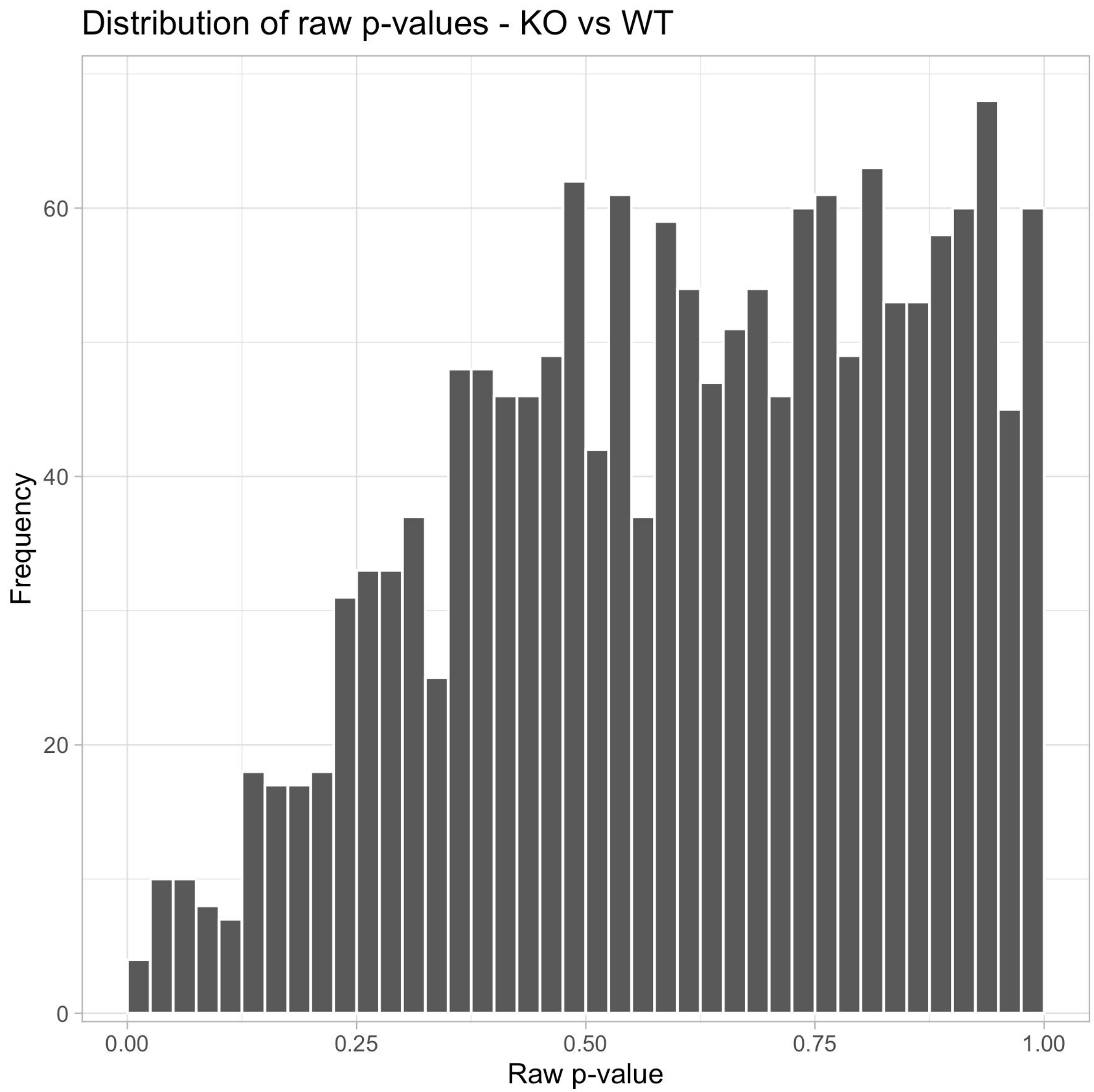
Principal Component Analysis



Cluster dendrogram
Euclidean distance, Ward criterion

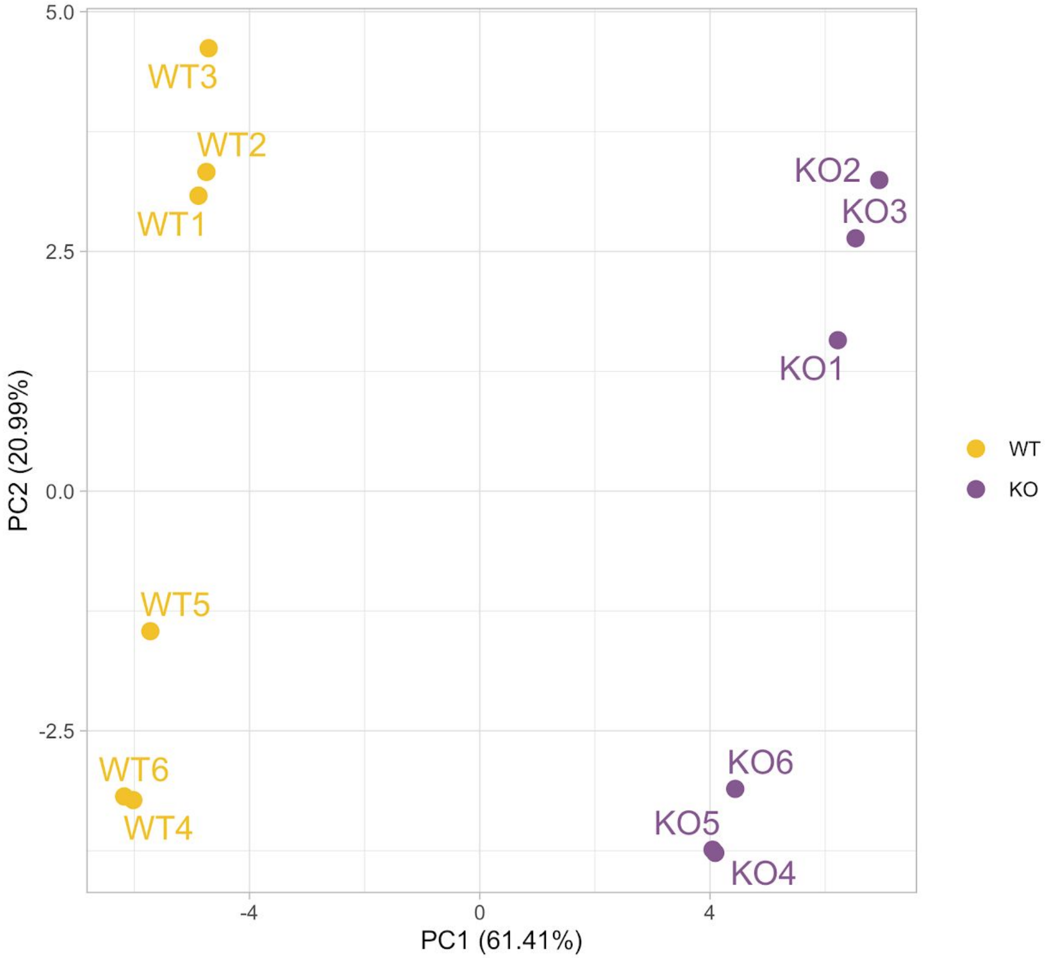


Potential issue: inversion of samples

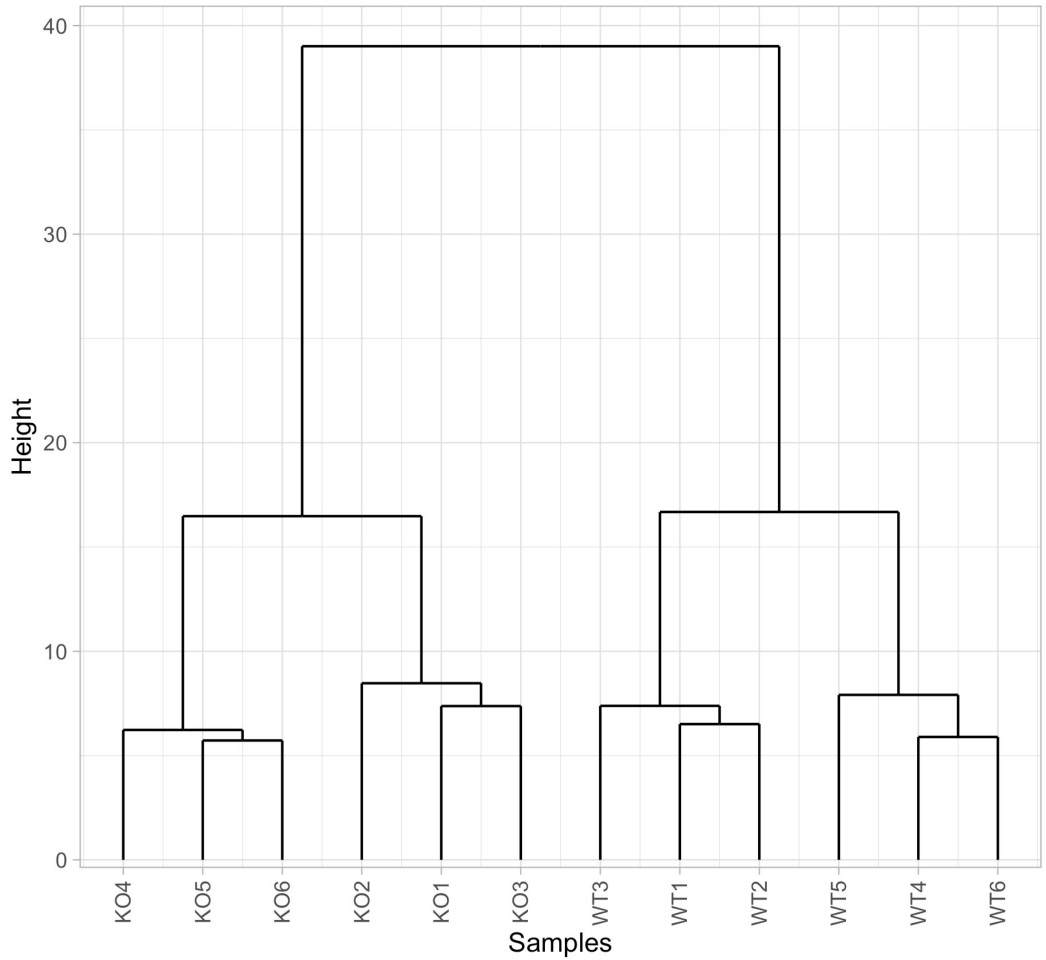


Potential issue: batch effect

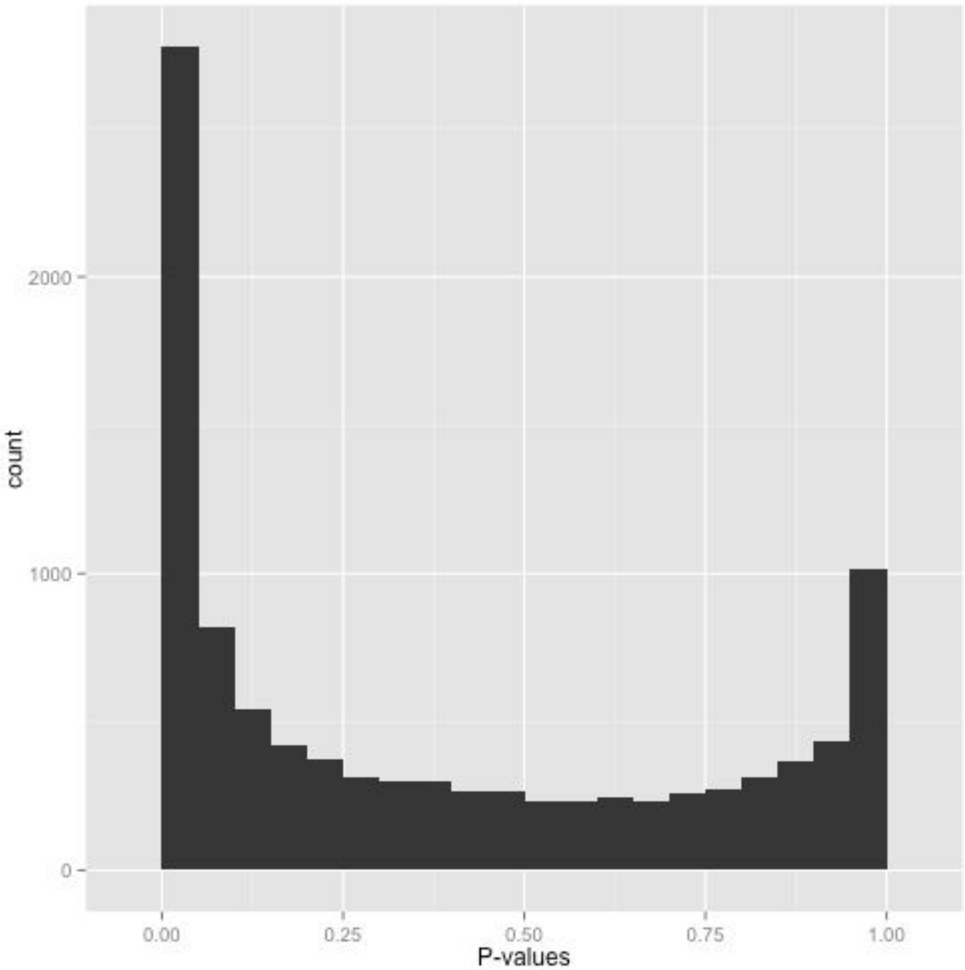
Principal Component Analysis

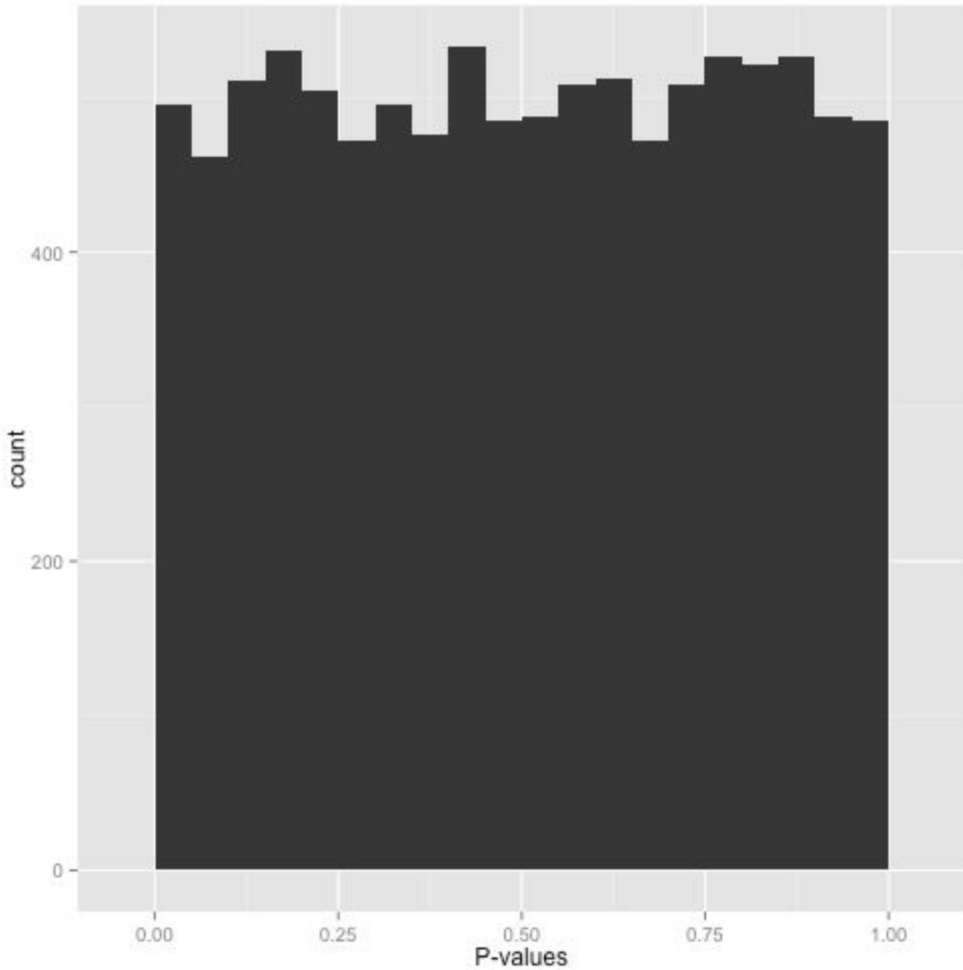


Cluster dendrogram
Euclidean distance, Ward criterion



Other cases :


count
2000
1000
0
0.00
0.25
0.50
0.75
1.00
P-values


count
400
200
0
0.00
0.25
0.50
0.75
1.00
P-values

DESeq2 and edgeR common parameters

- Project and author names
- Target and count files paths
- Rows of the count files to remove
- Factor of interest and the reference biological condition
- Adjustment variable (batch effect, pairing) in the target file
- Multiple testing adj. method and significance threshold α
- Colors for the graphics

DESeq2-specific parameters

- **fitType:** type of link to model the intensity-dispersion relationship, `parametric` (by default) or `local`
- **cooksCutoff:** `TRUE` (by default) to detect genes having outlier counts
- **independentFiltering:** `TRUE` (by default) to filter out lowly expressed genes and gain power on the others
- **typeTrans:** `VST` (by default) or `rlog` to make the data homoscedastic to perform exploratory data analysis (PCA, clustering, heatmaps)
- **locfunc:** `median` (by default) or `shorth`. `shorth` allows to improve the normalization for some cases

edgeR-specific parameters

- **cpmCutoff:** low counts filtering threshold (in counts per million of reads)
- **gene.selection:** genes selection method for the MDS-plot (`pairwise` by default)
- **normalizationMethod:** `TMM` by default, `RLE` (DESeq2), or `upperquartile`

Conclusion

SARTools...

- facilitates the utilization of DESeq2 and edgeR
- performs quality control and helps to detect potential problems
- fits the reproducible research criteria

Take time to interpret each figure/table in the HTML report!

Interpreting lists of DE genes: gene-set level analysis

What is a gene-set?

→ Any group of genes having a biological meaning

Note: some genes can belong to several sets and others to none

Two main approaches:

- **Competitive** null hypothesis: genes in the set are “as DE as” genes not in the set
- **Self-contained** null hypothesis: genes in the set are not DE

Several methods:

- Over-Representation Analysis (competitive): are genes in the set more DE than genes not in the set? → Fisher’s hypergeometric test
- Linear models using limma R package’s functions:
  - competitive: `camera()` and `romer()`
  - self-contained: `roast()` and `fry()`

Interpreting lists of DE genes: gene-set level analysis

Several issues/options to deal with:

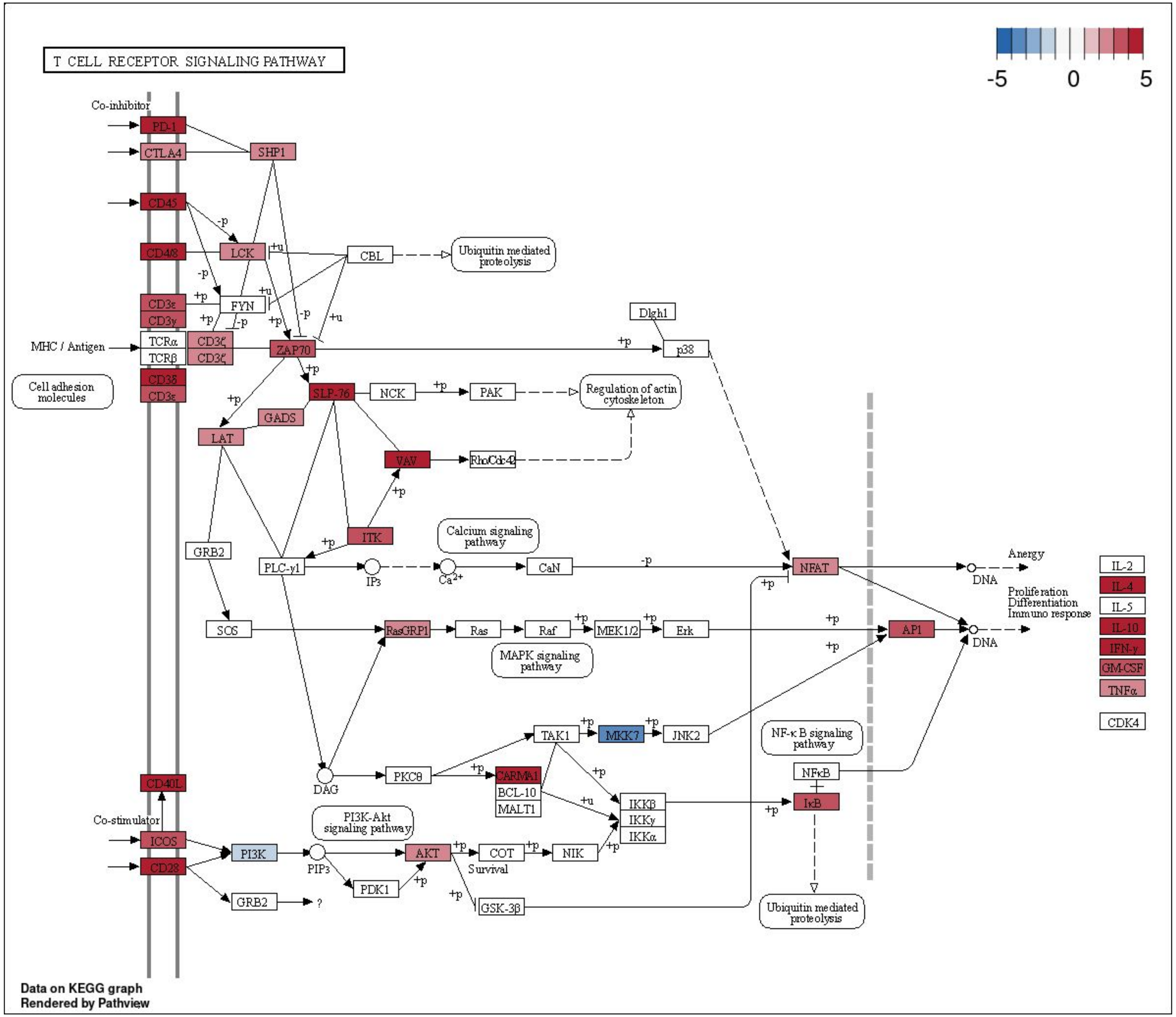
- Make gene IDs compatible with the gene-sets by converting diff. analysis **Ensembl** IDs (for instance) into **ENTREZ** IDs: no perfect matching and be careful with the annotation version(s) used

- Which gene-sets to test?
    - depends on the **biological question**
    - will impact the p-value adjustment for multiple testing
    - restrict the **background** to genes belonging to at least one set?

- Separate down- and up-regulated genes?

- Import gene-sets into R and make them ready for the analysis: from MSigDB or R packages… but there may be some differences

Interpreting lists of DE genes: gene-set level analysis



INSTITUT PASTEUR

General conclusion

- RNA-Seq project = discussions between biologists, bioinformaticians and biostatisticians... as soon as the project starts!

- Statistical needs during all the project, not only for the differential analysis
  - Normalization step is critical: the assumptions have to be checked
  - No magic recipe: need to choose the statistical model according to your biological question
  - Statistical analysis must not be a black box!

- Data visualization is a crucial tool along all the steps of the analysis

Complex experimental design → difficult interpretation of the results

The end

Thank you for your attention!

Bibliography

[1] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer and B. Wold. *Mapping and quantifying mammalian transcriptomes by RNA-Seq*. Nature Methods. 2008.
[2] S.-K. Schulze, R. Kanwar, M. Gölzenleuchter, T.-M. Therneau and A.-S. Beutler. *SERE: Single-parameter quality control and sample comparison for RNA-Seq*. BMC Genomics, 2012.
[3] M. Love, W. Huber and S. Anders. *Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2*. Genome Biology, 15, 2014.
[4] M.-D. Robinson and A. Oshlack. *A scaling normalization method for differential expression analysis of RNA-seq data*. Genome Biology 2010, 11:R25, 11(R25), 2010.
[5] M.-A. Dillies, A. Rau, J. Aubert and others. *A comprehensive evaluation of normalization methods for Illumina RNA-seq data analysis*. Briefings in Bioinformatics, 2012.
[6] Y. Benjamini and Y. Hochberg. *Controlling the false discovery rate : A practical and powerful approach to multiple testing*. Journal of the Royal Statistical Society, 57(1):289–300, 1995.
[7] C. Soneson and M. Delorenzi. *A comparison of methods for differential expression analysis of RNA-seq data*. BMC Bioinformatics, 14, 2013.
[8] M.-D. Robinson, D.-J. McCarthy and G.-K. Smyth. *edgeR : a bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics, 2009.
[9] H. Varet, L. Brillet-Guéguen, J.-Y. Coppée and M.-A. Dillies. *SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data*. PloS One, 2016.
[10] C. Evans, J. Hardin and D.-M. Stoebel. *Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions*. Briefings in Bioinformatics, 2017.