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

Ecole de Bioinformatique AVIESAN/IFB – November 2016

Hugo Varet – `hugo.varet@pasteur.fr`

Institut Pasteur

Transcriptome & Epigenome Platform – Biomics Pole – Citech

Bioinformatics & Biostatistics Hub – C3BI & USR 3756 CNRS

aviesan
alliance nationale
pour les sciences de la vie et de la santé

ifb

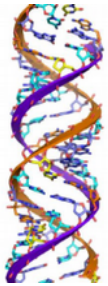
CNRS UPMC
Station Biologique
Roscoff

Main RNA-Seq steps

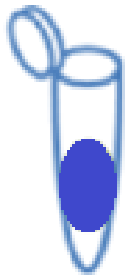
Designing the experiment and the sequencing

Individual

RNA

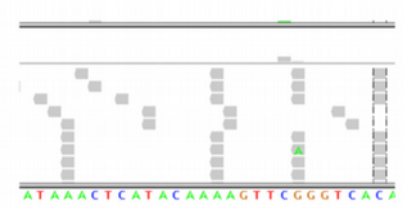


Library

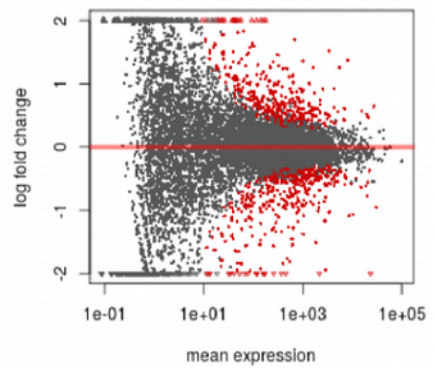


Sequencing

Bioinformatics



Differential analysis



Citation

"*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

Citation

“While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment”

Kathleen Kerr, Atelier Inserm 145, 2003

Vocabulary

Design file:

Samples	VariableV	FactorF
ReplicateA-1	levelA	biologicalConditionX
ReplicateA-2	levelA	biologicalConditionY
ReplicateB-1	levelB	biologicalConditionX
ReplicateB-2	levelB	biologicalConditionY

Example:

id	strain	day
WT-1	WT	d1
WT-2	WT	d2
WT-3	WT	d3
KO-1	KO	d1
KO-2	KO	d2
KO-3	KO	d3

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.:

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

Starting point of the differential analysis

	T0-1	T0-5	T0-6	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
gene14	288	238	304	6	6	8	2	7	3
gene15	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

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

Basic comparison

Transcriptome differences between Cystic Fibrosis (CF) patients and healthy people: mRNA sequencing of lung cells.

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

Paired samples

Transcriptome differences between Cystic Fibrosis (CF) patients and healthy people: mRNA sequencing of lung cells.

id	state	RNA extraction date
h1	healthy	June 12th, 2016
h2	healthy	June 20th, 2016
h3	healthy	June 25th, 2016
cf1	CF	June 12th, 2016
cf2	CF	June 20th, 2016
cf3	CF	June 25th, 2016

Paired samples

RNA-Seq of both lung and skin cells from three Cystic Fibrosis (CF) patients.

id	state	tissue	patient
cf1-s	CF	skin	cf1
cf2-s	CF	skin	cf2
cf3-s	CF	skin	cf3
cf1-l	CF	lung	cf1
cf2-l	CF	lung	cf2
cf3-l	CF	lung	cf3

Time course experiment

New treatment T applied to cultures of lung cells from 3 Cystic Fibrosis (CF) patients. Study of the initial transcriptome and after 4h and 8h of treatment.

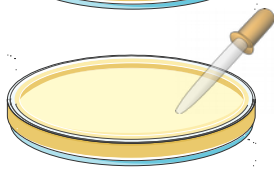
id	state	time	patient
cf1-0	CF	0h	cf1
cf2-0	CF	0h	cf2
cf3-0	CF	0h	cf3
cf1-4	CF	4h	cf1
cf2-4	CF	4h	cf2
cf3-4	CF	4h	cf3
cf1-8	CF	8h	cf1
cf2-8	CF	8h	cf2
cf3-8	CF	8h	cf3

On the laboratory bench...

Time 0h

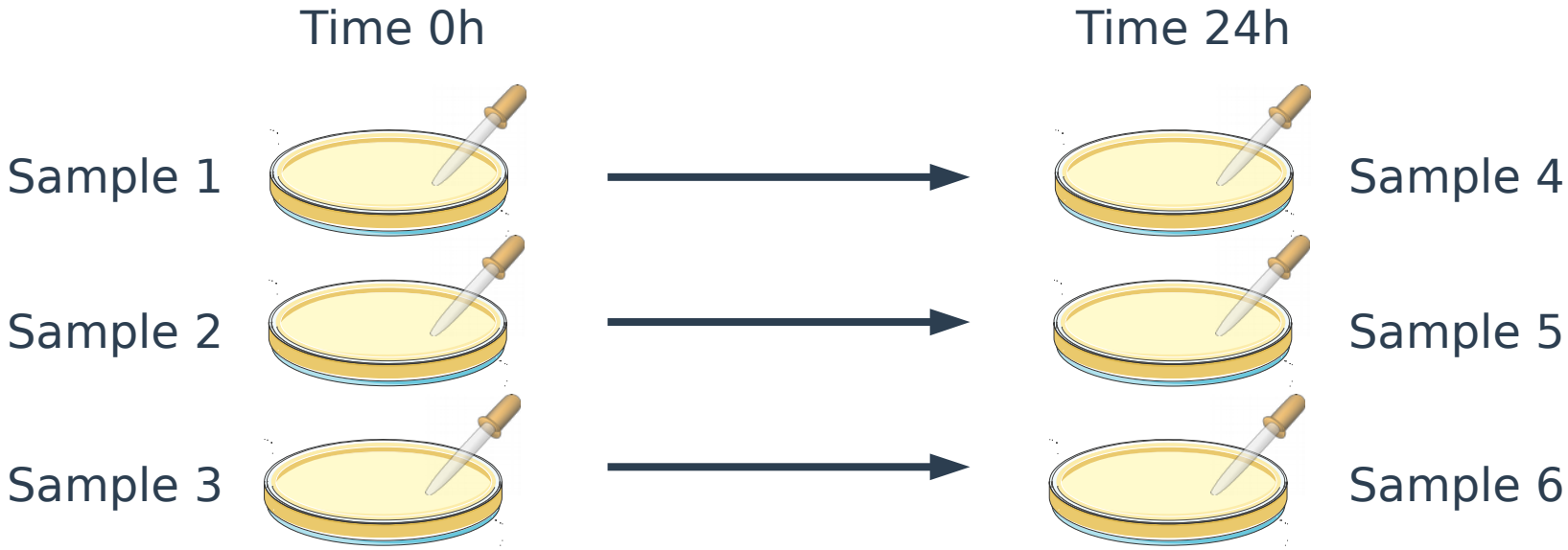
Sample 1

Sample 2

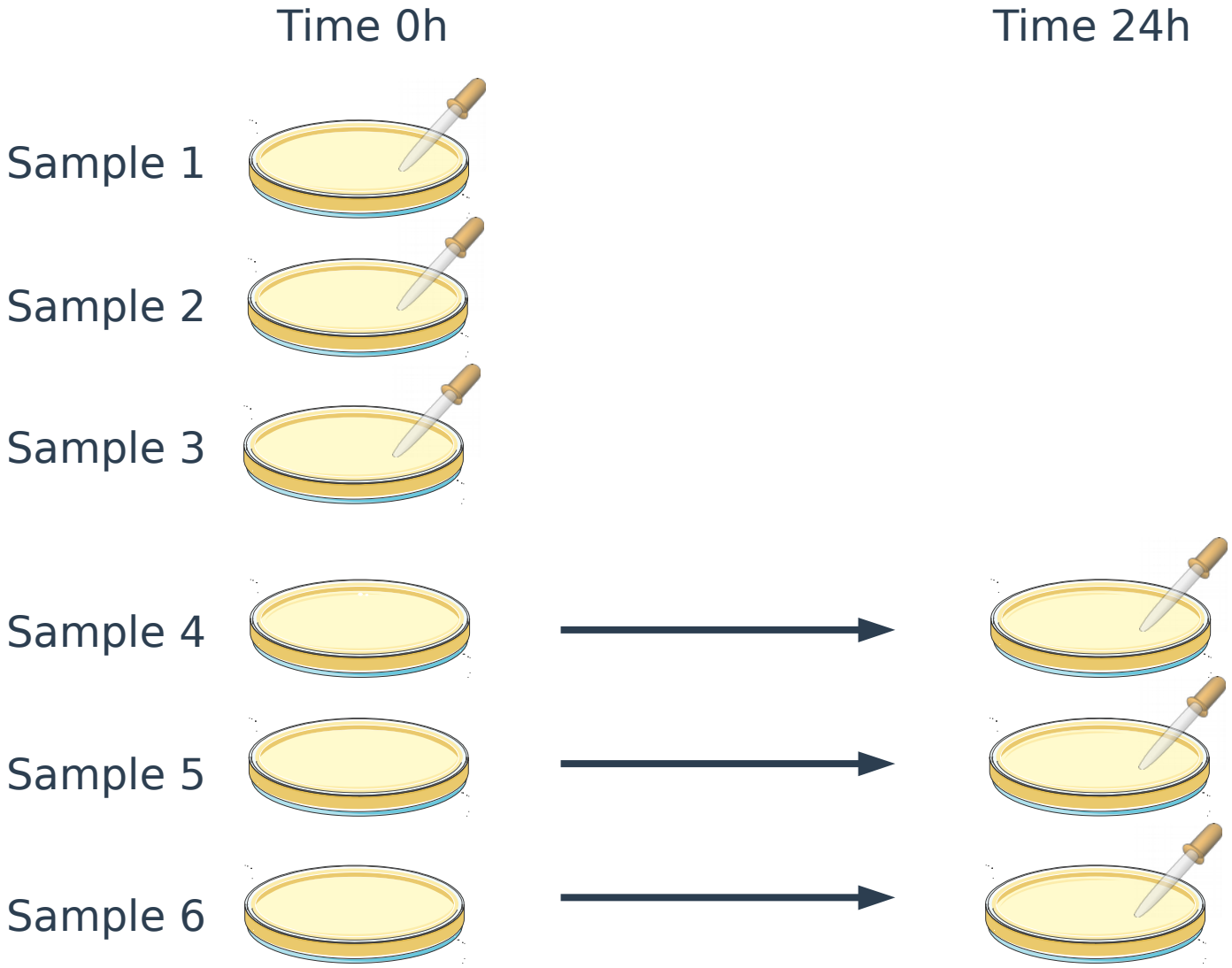


Sample 3

On the laboratory bench...



On the laboratory bench...

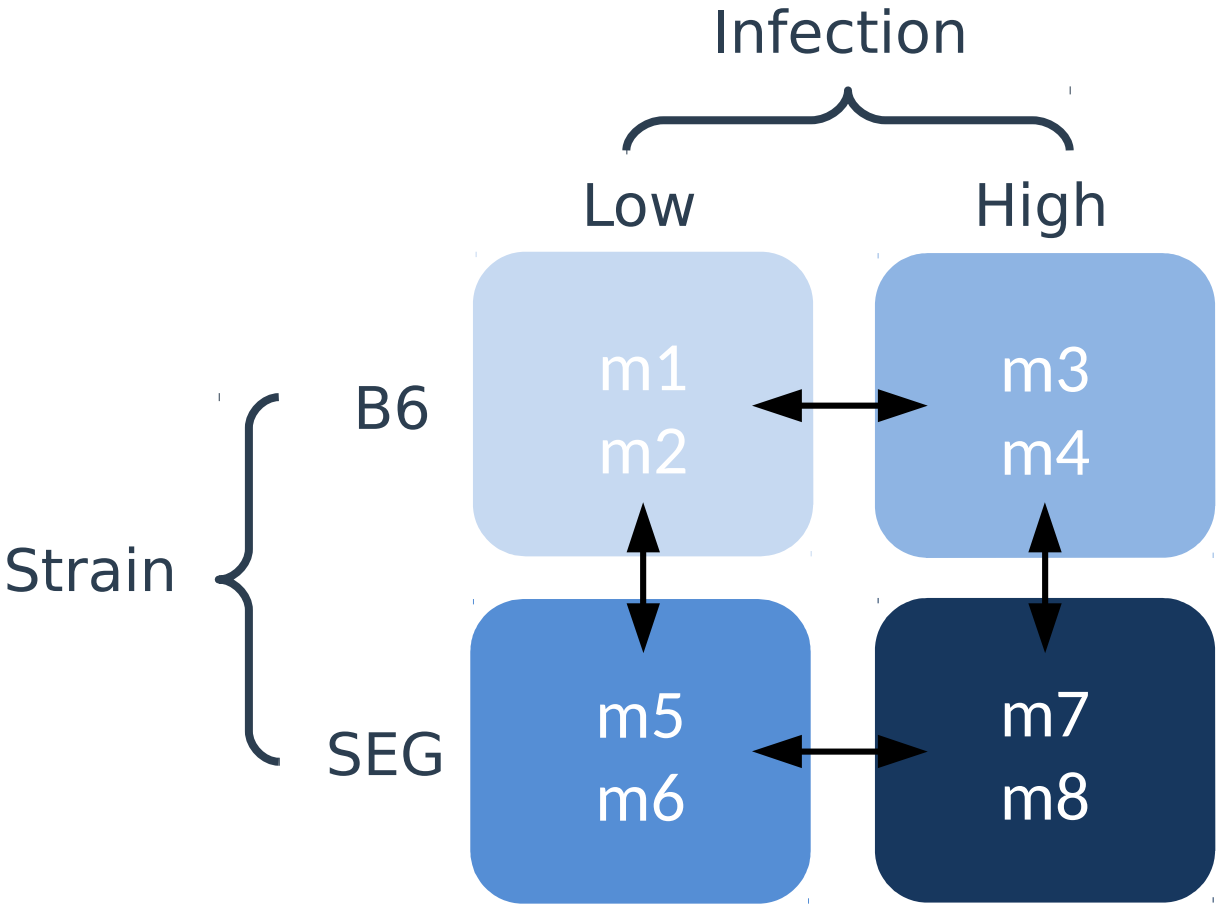


Complex design

Effect of the 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

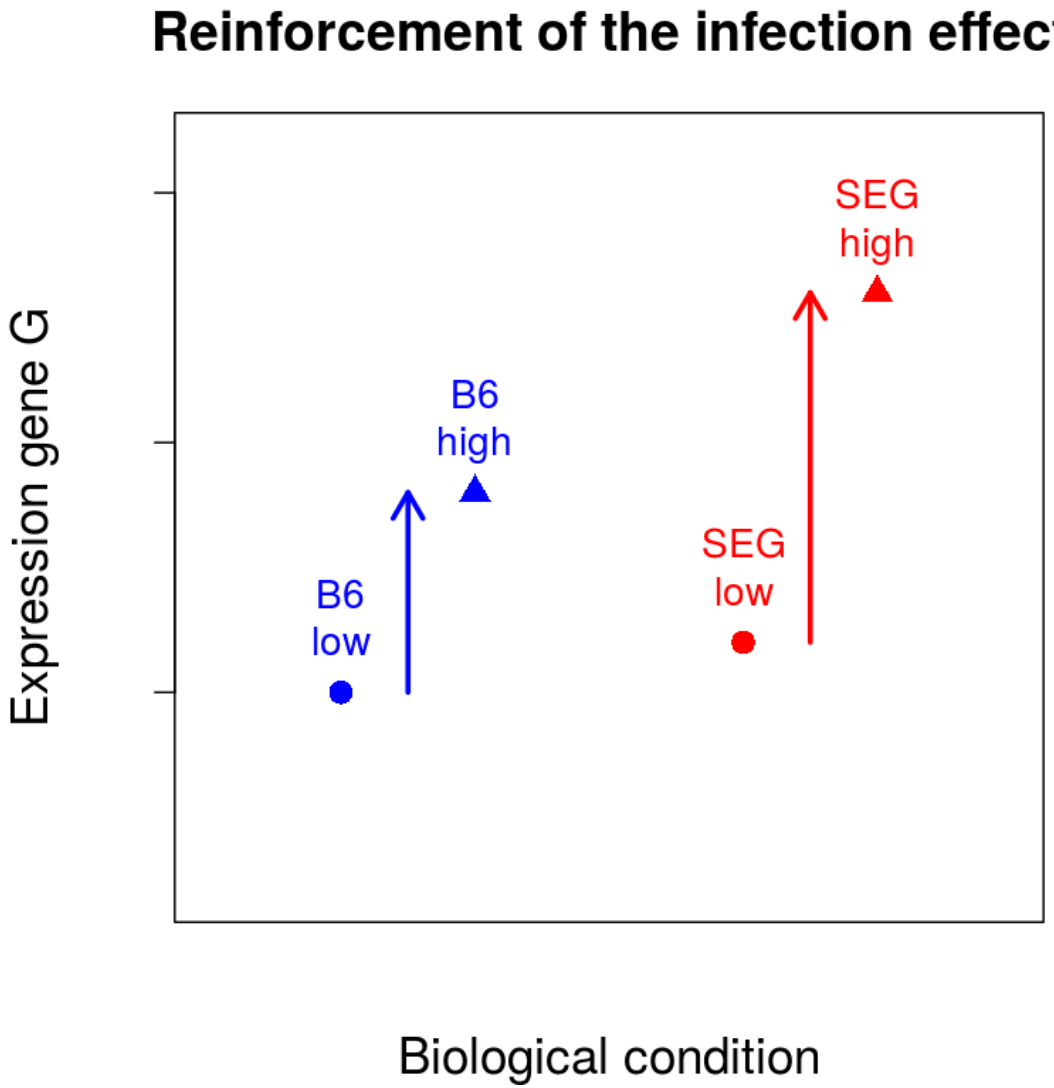
Interaction between two factors/variables



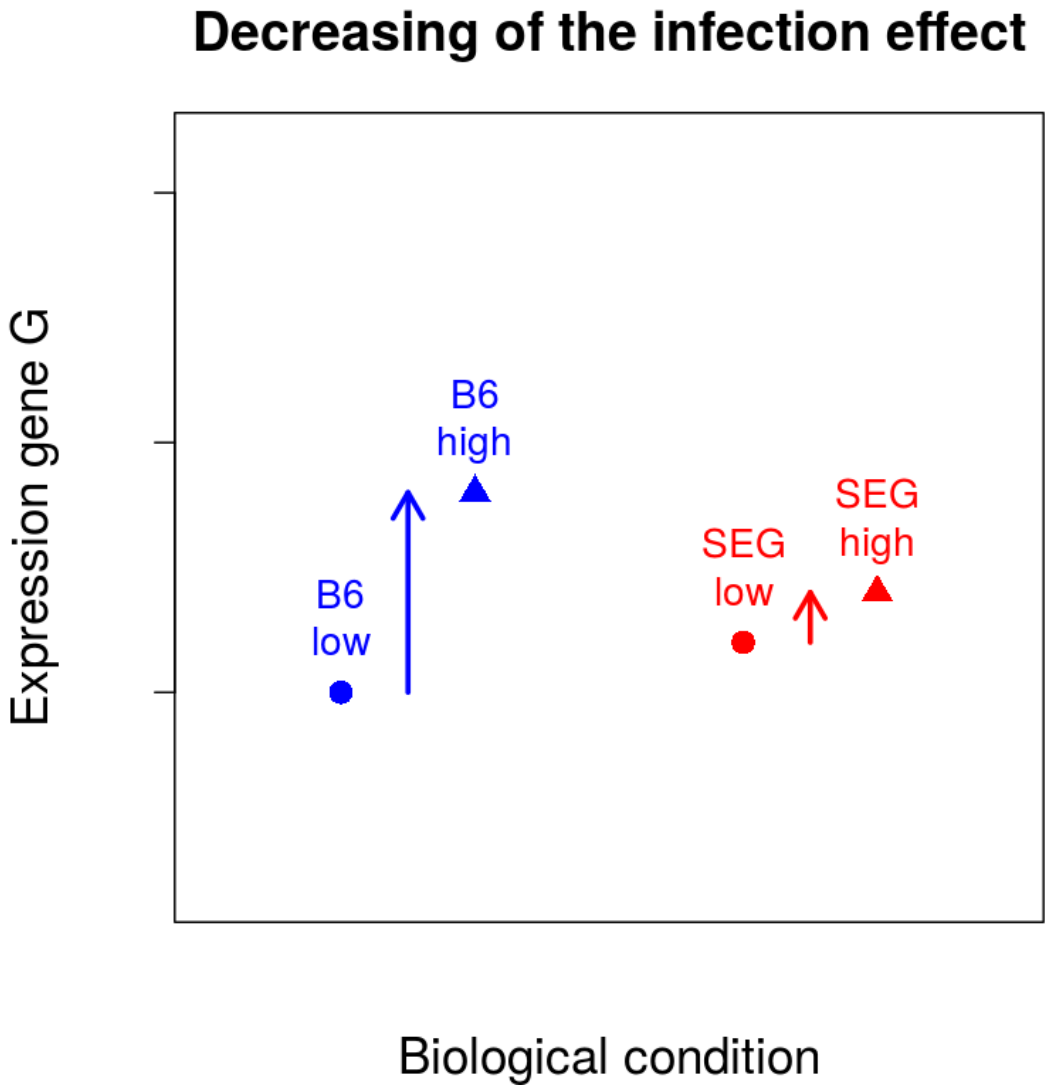
Interaction:

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

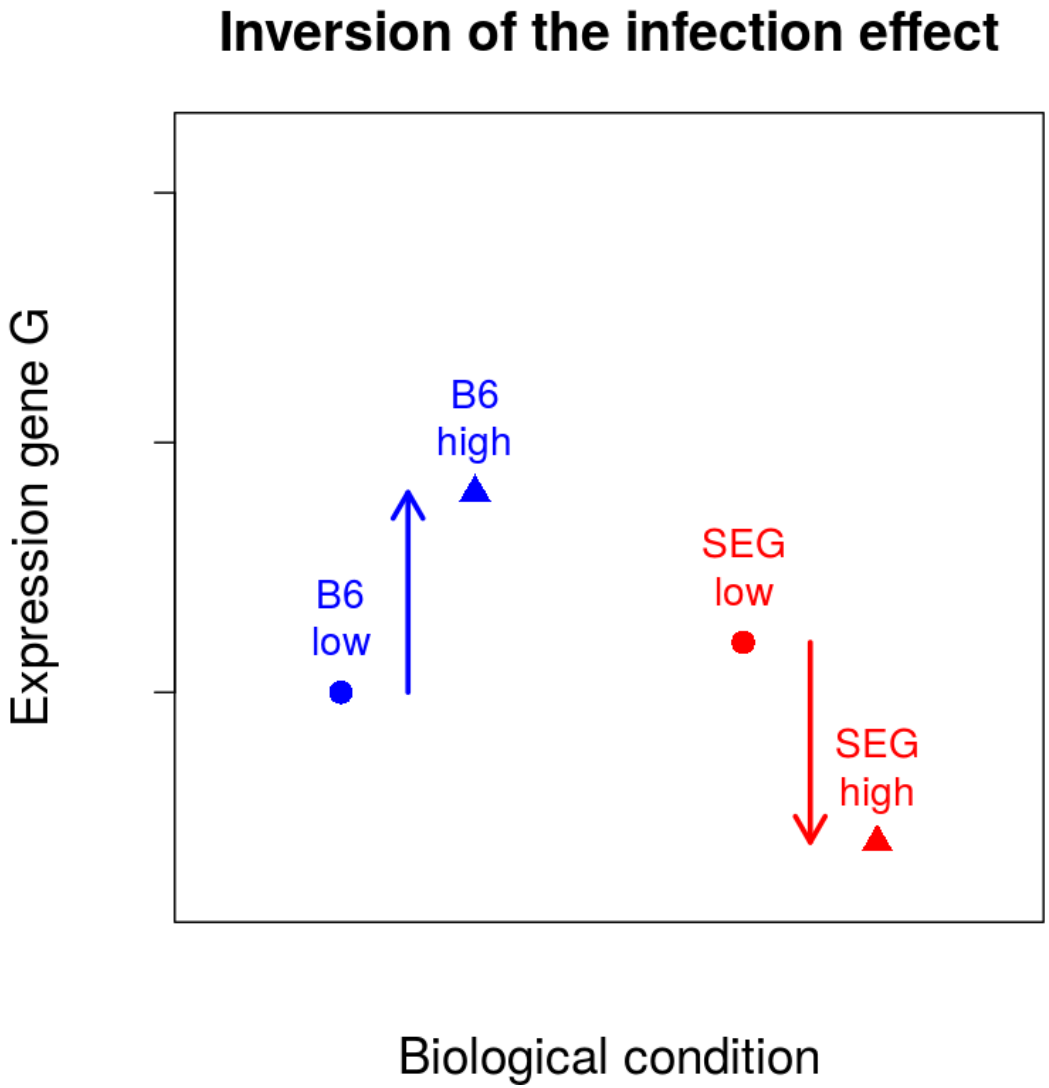
Examples of interactions



Examples of interactions



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.

Confounding effect

Comparison of CF vs healthy patients:

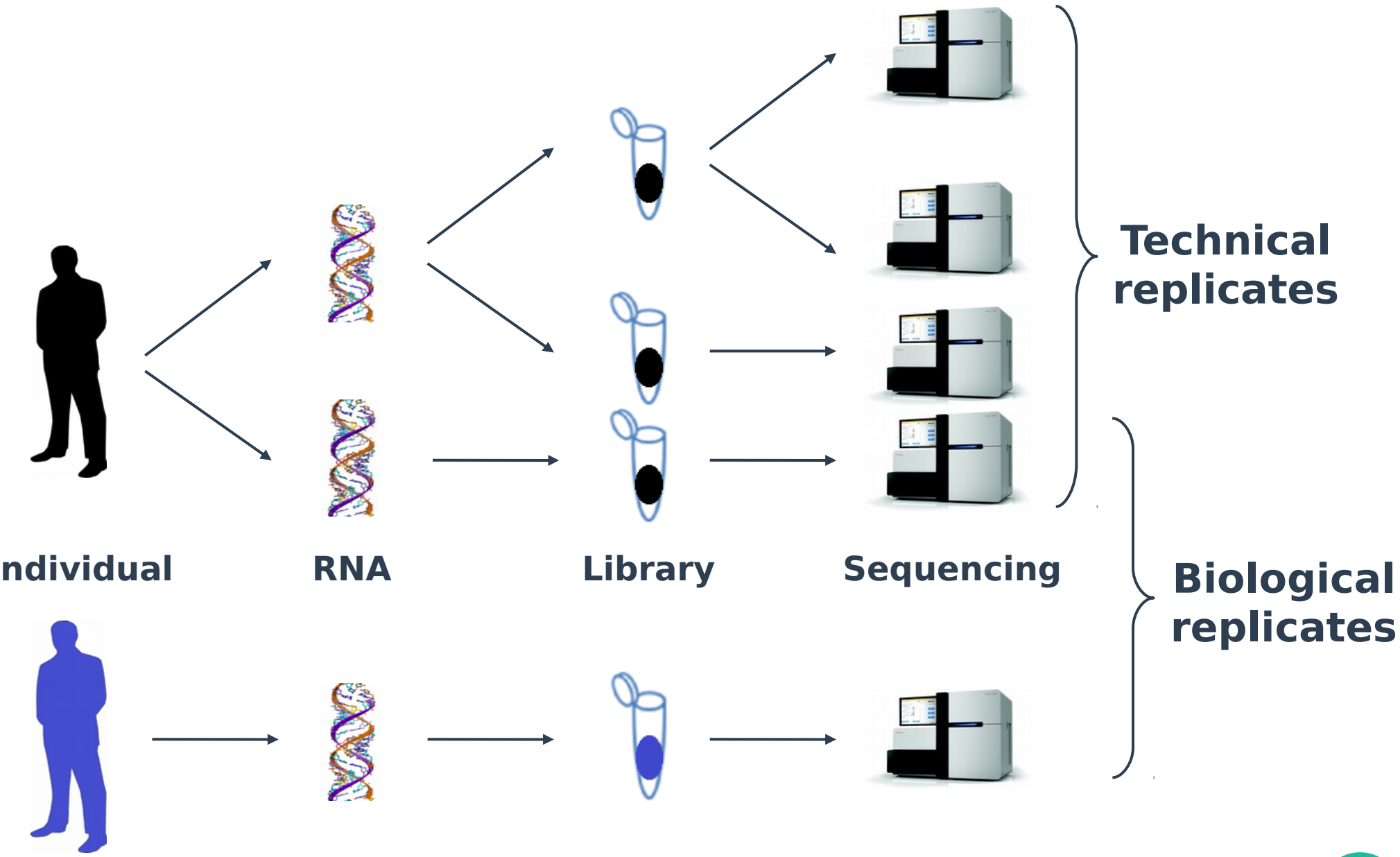
id	state	age	gender	RNA extraction day	experimentalist
h1	healthy	45	female	July 9th, 2015	Louis
h2	healthy	52	female	July 12th, 2015	Louis
h3	healthy	48	female	July 15th, 2015	Louis
cf1	CF	31	male	Feb 20th, 2016	François
cf2	CF	25	male	Feb 24th, 2016	François
cf3	CF	27	male	Feb 29th, 2016	François

Confounding effect

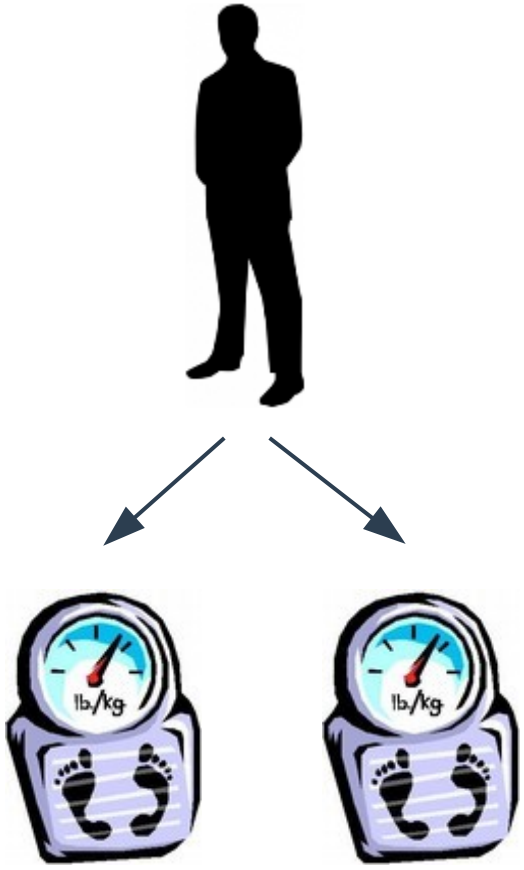
A gene is detected as being differentially expressed between healthy and CF patients. Is it due to:

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

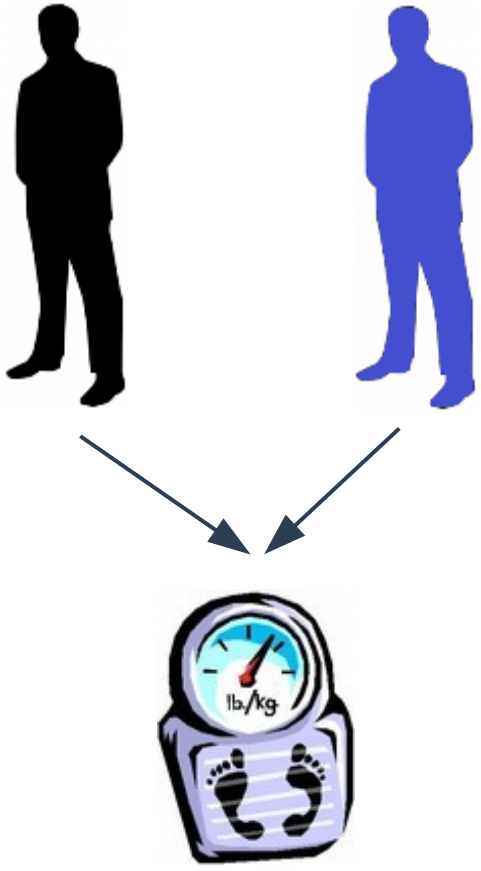
Biological vs. technical replicates



Biological vs. technical replicates



Technical



Biological

Biological vs. technical replicates

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:

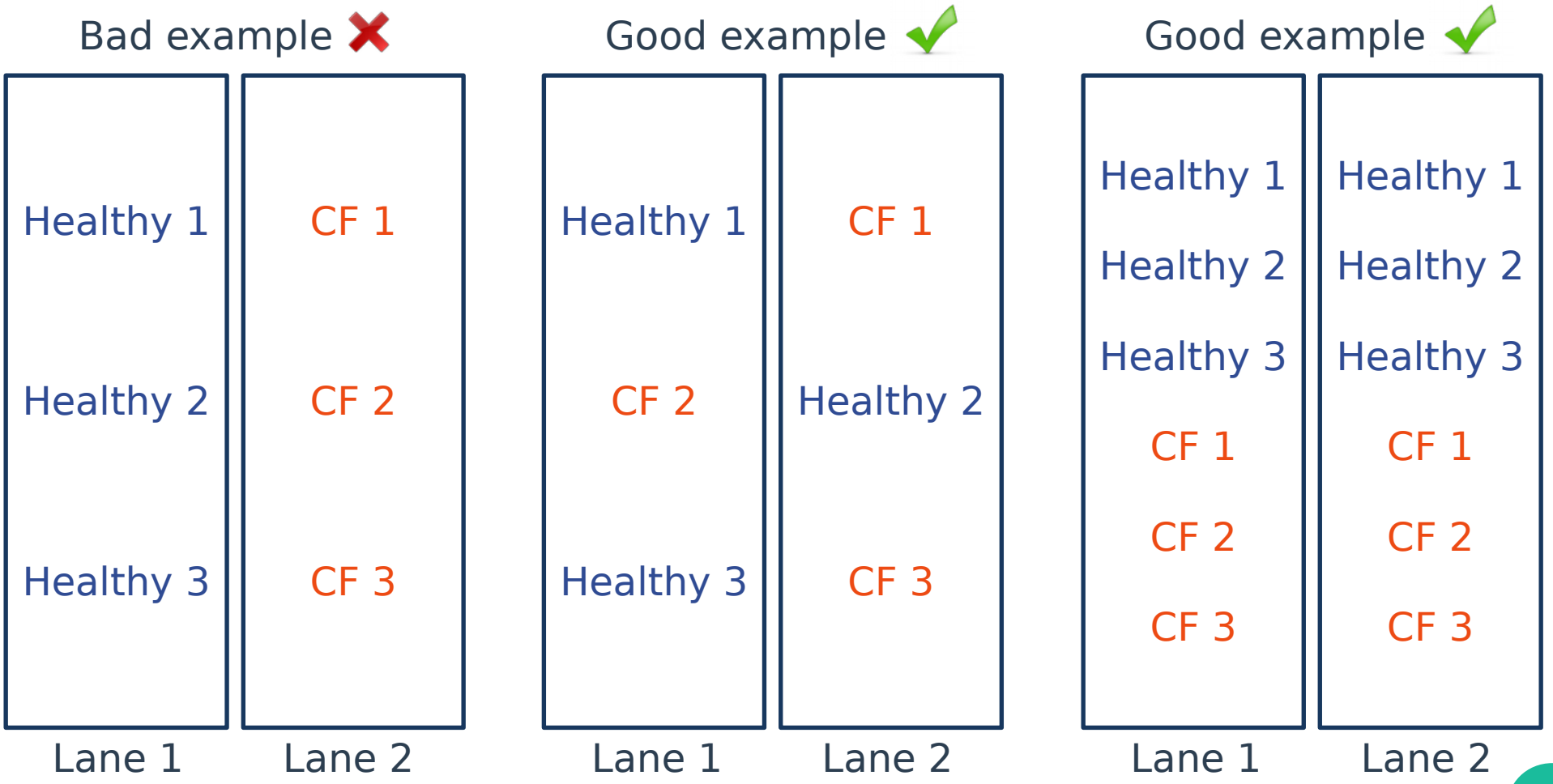
- Correspond to the variability visible in the real life

Comment: what happens when studying fungi/yeast?

Sequencing design

Goal:

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



Sequencing design

Technical variabilities:

- Lane
- Flowcell
- Run

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

Use the same multiplexing rate for all the samples!

Remember

The **biological question** must be well defined in order to build an experimental design which will be able to address it.

Identify all the sources of variability:

- 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
- And so on...

Outline

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

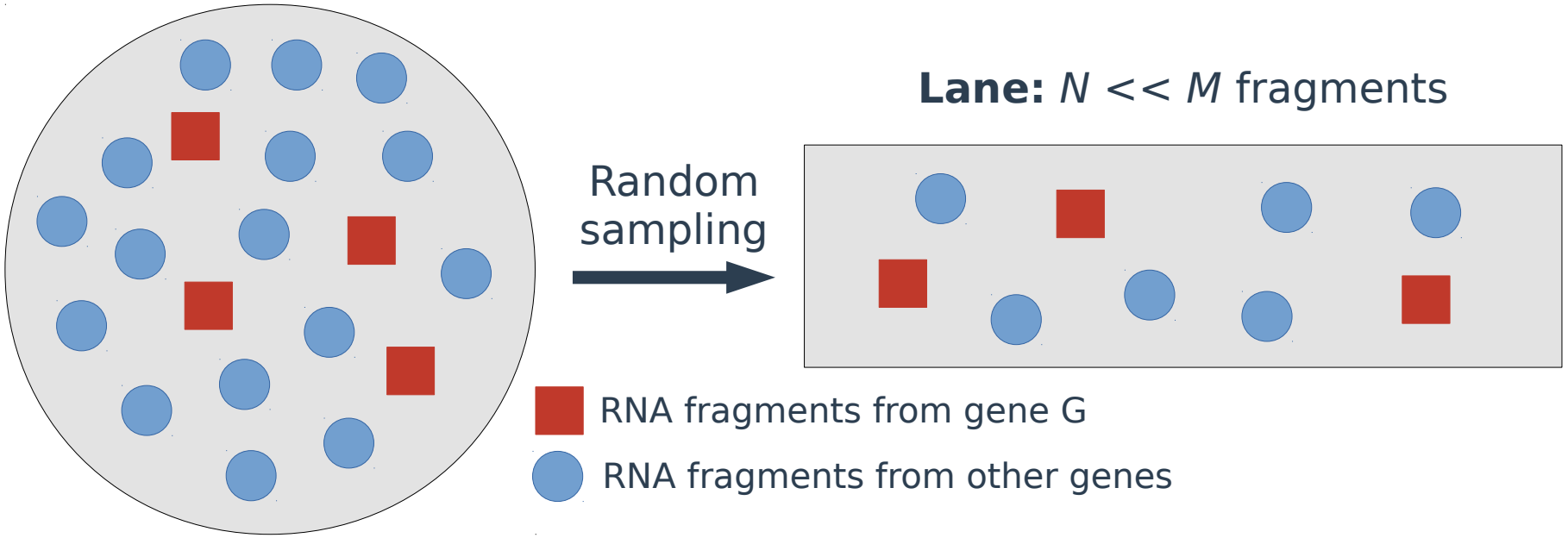
Reminder: count matrix

	T0-1	T0-5	T0-6	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
gene14	288	238	304	6	6	8	2	7	3
gene15	18	31	39	4	4	7	2	6	2

Goal: find genes differentially expressed between biological conditions

Distribution of counts data

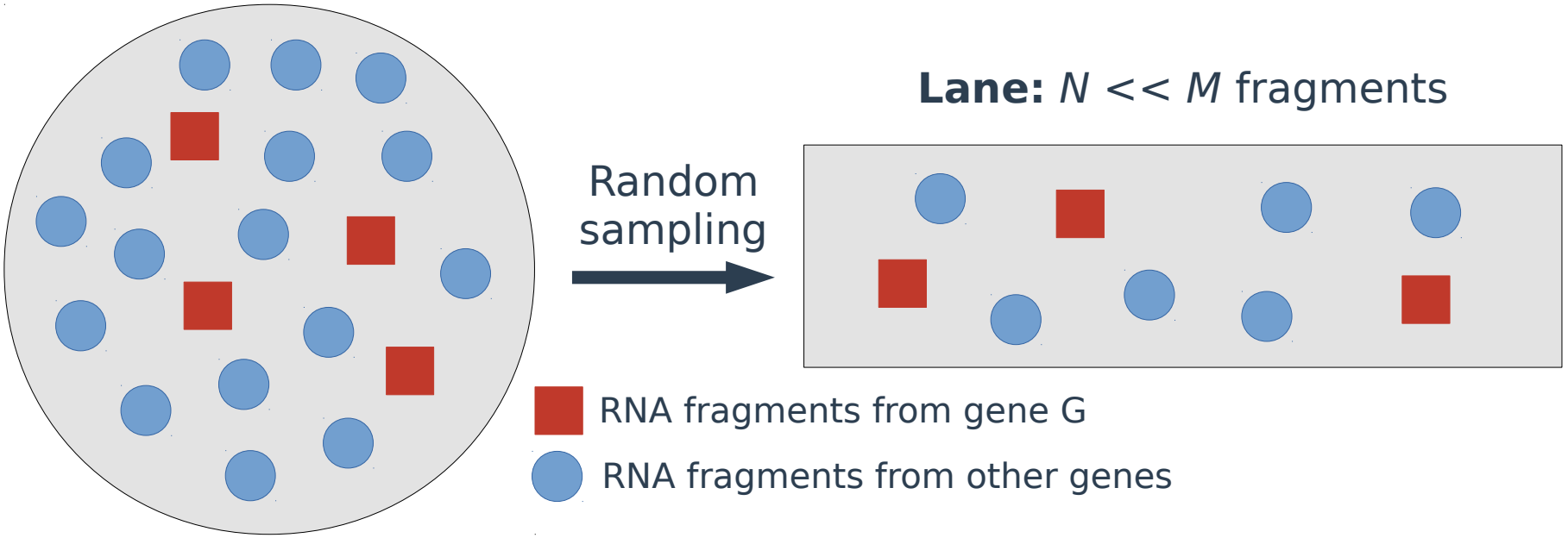
Library: *M* fragments of RNA



"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

Library: *M* fragments of RNA



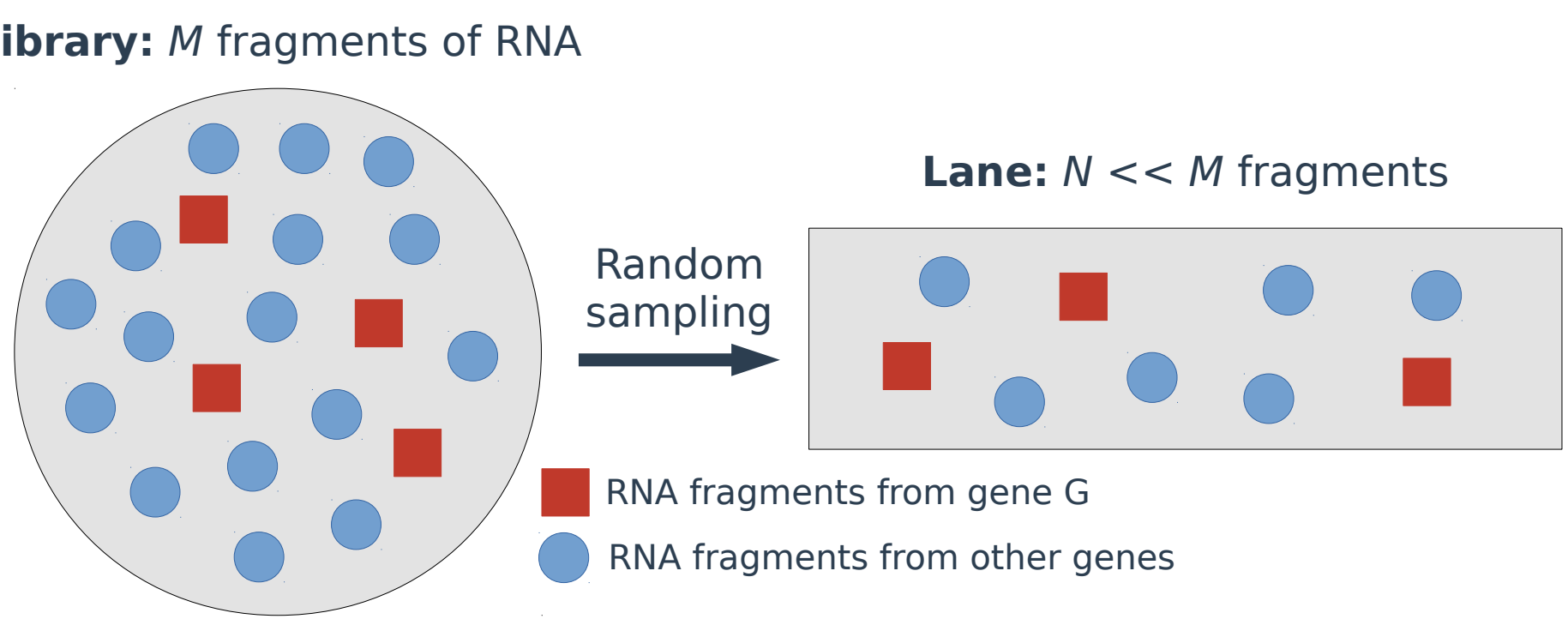
Let π_G = proportion of fragments of gene G:

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

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*):

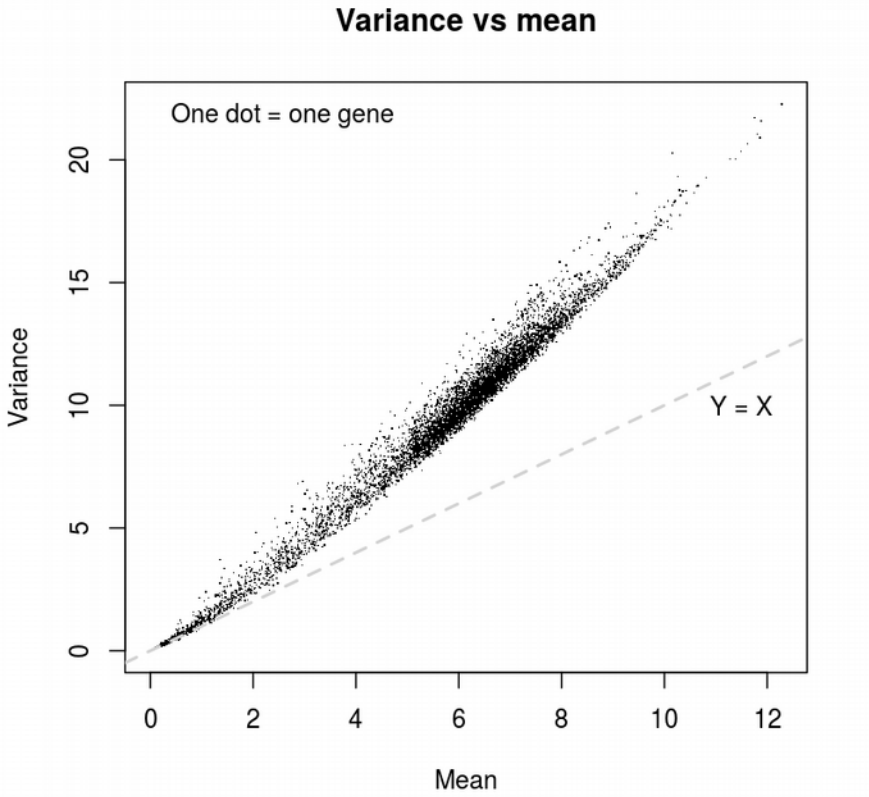
- Higher probability to catch lowly expressed genes
- Higher precision when estimating π_G

Distribution of counts data

If $X_G \sim \text{Poisson}(N\pi_G)$:

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

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).

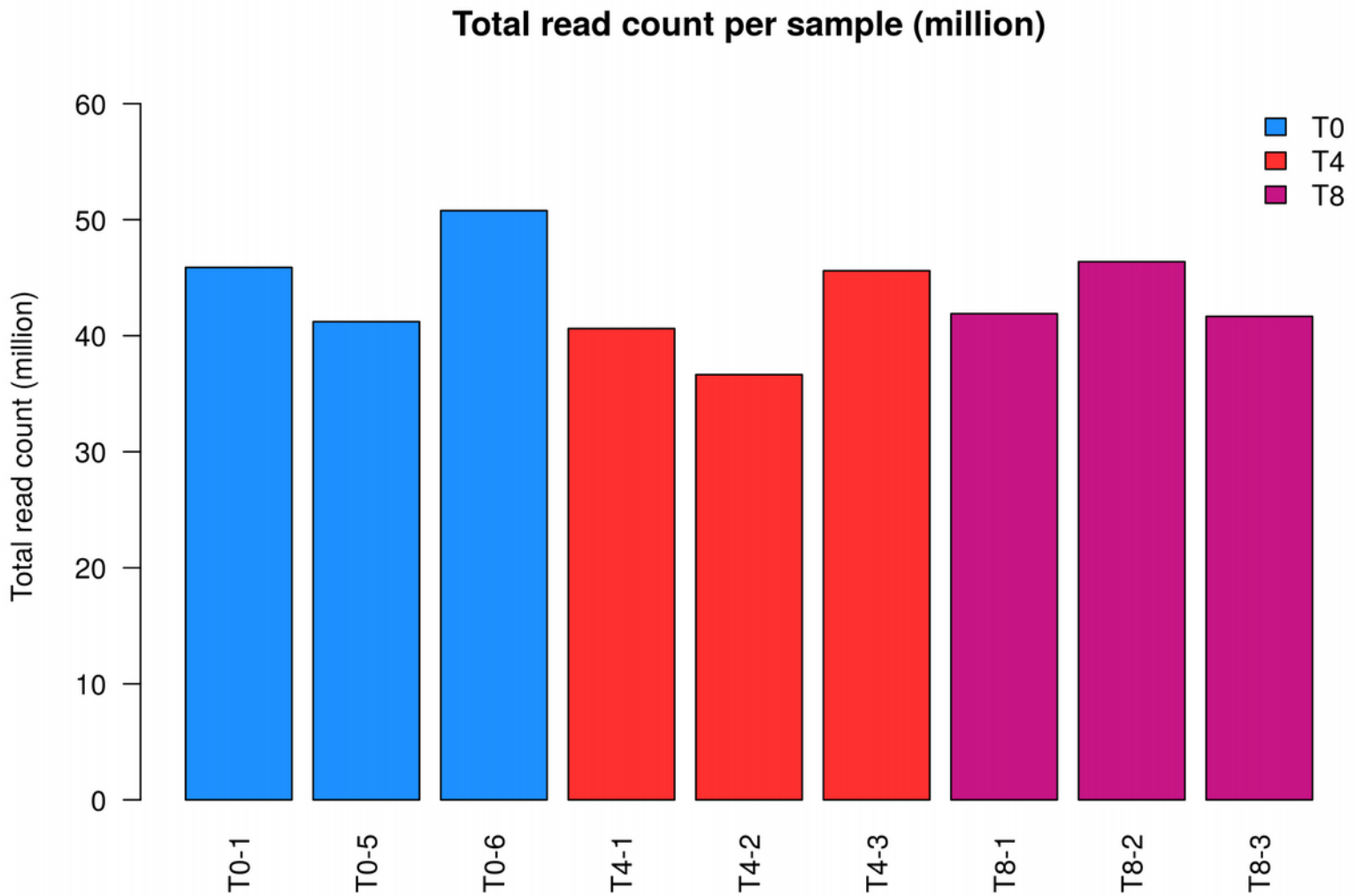
$$x_{ij} \sim \text{Negative-Binomial}(\text{mean} = \mu_{ij}, \text{variance} = \sigma_{ij}^2)$$

where:

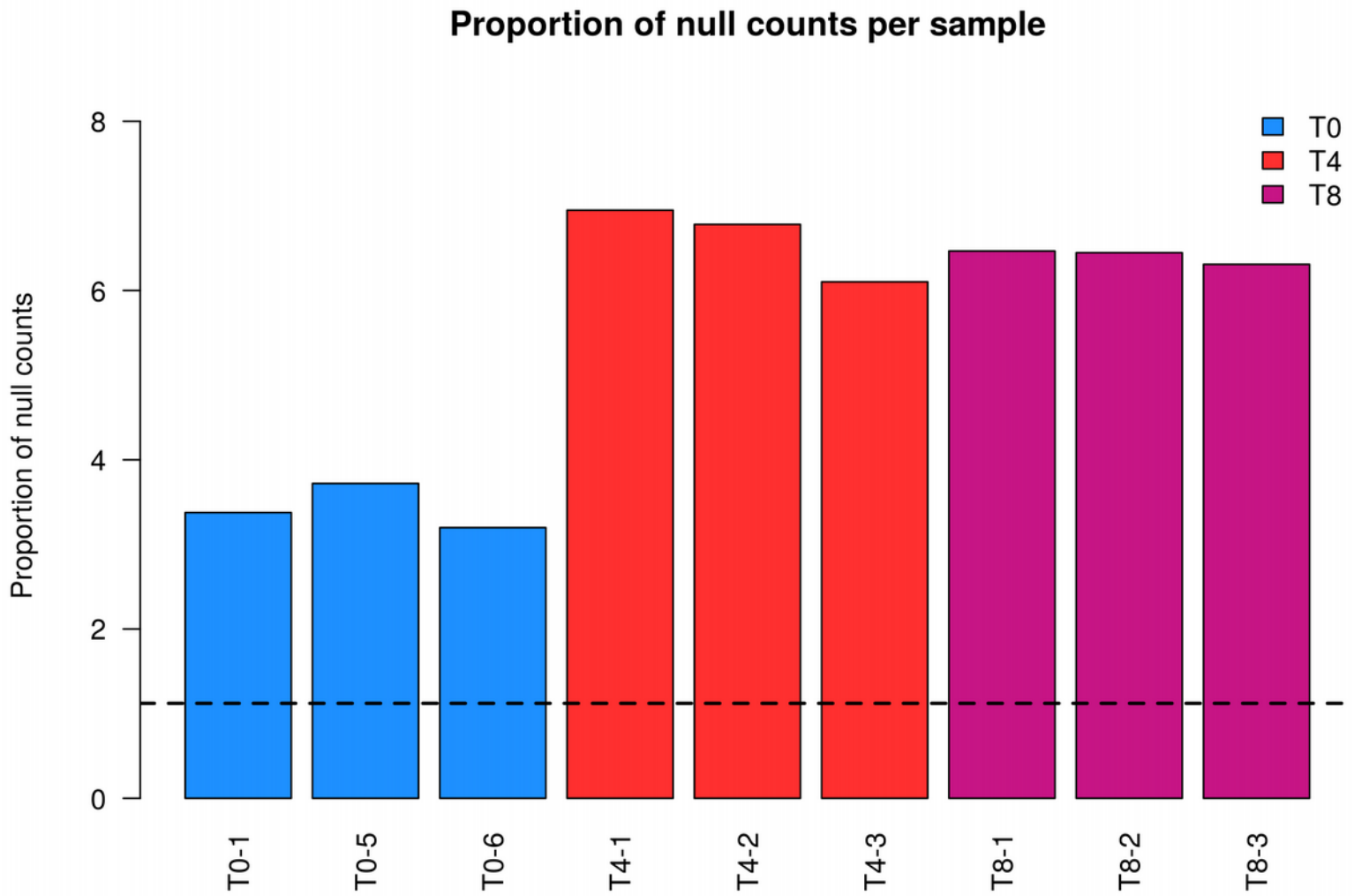
- $\sigma_{ij}^2 = \mu_{ij} + \varphi_i \, \mu_{ij}^2$
- φ_i : biological dispersion of gene i

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

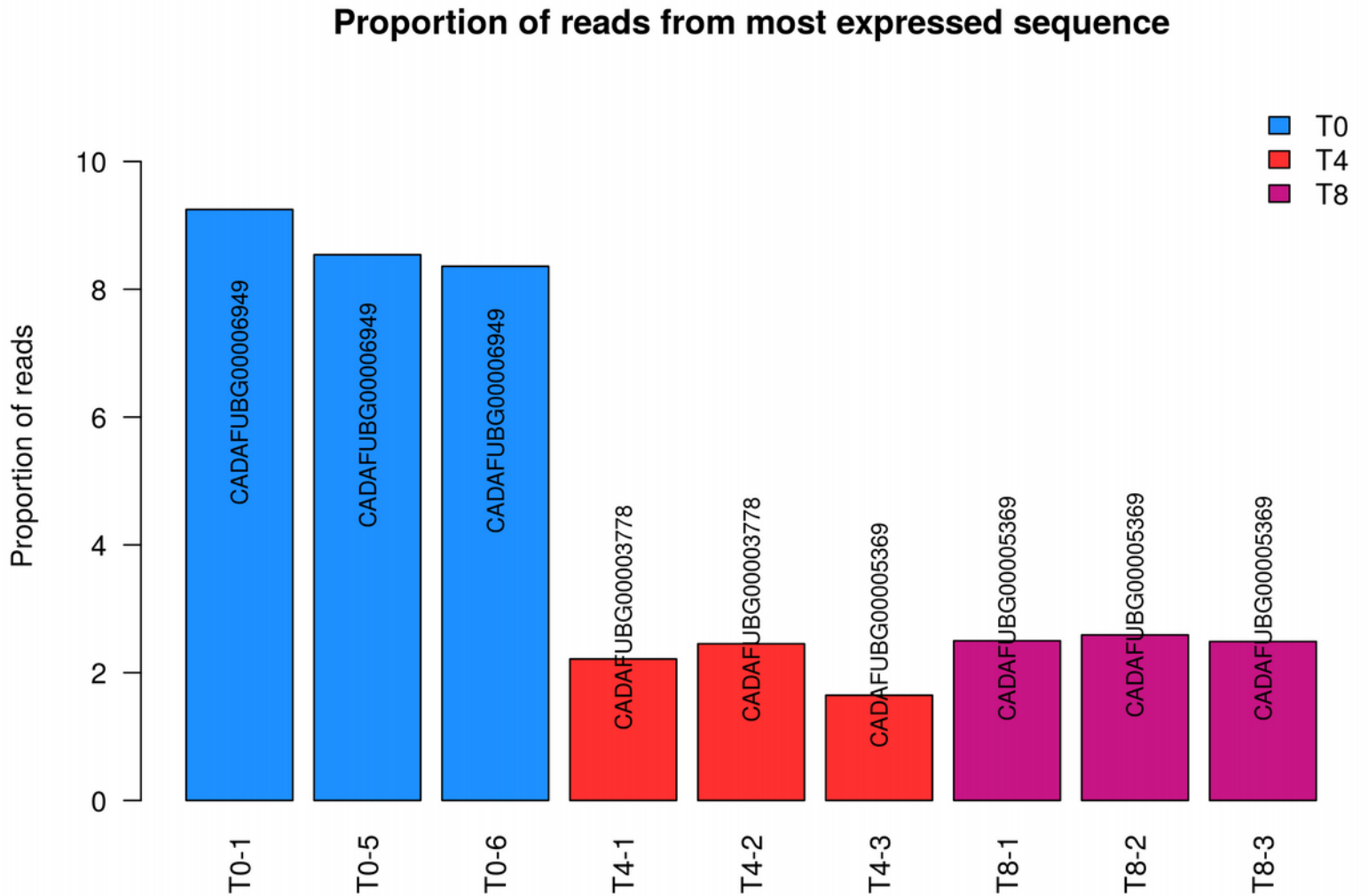
Total read count per sample



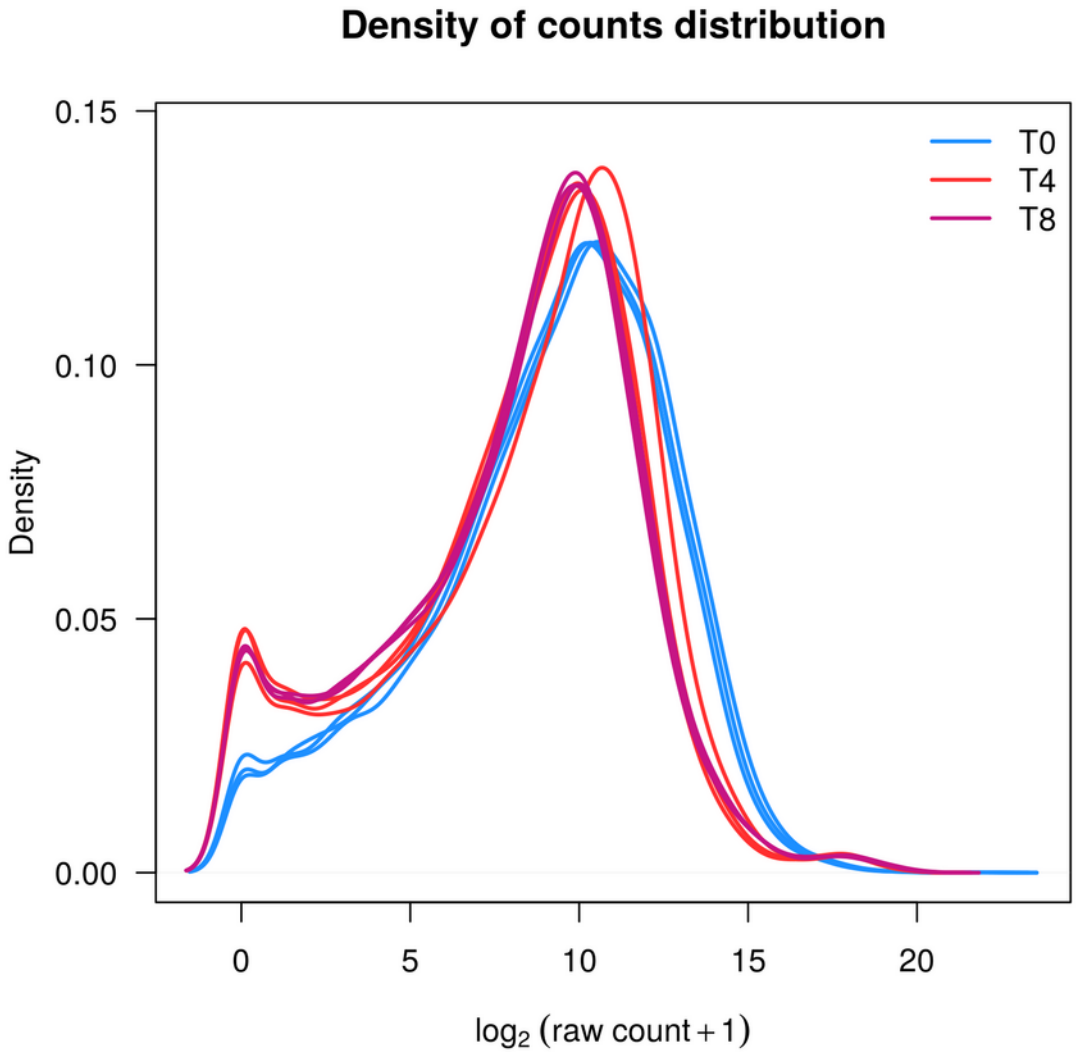
Proportion of null counts per sample



Prop. of reads from most expressed sequence



Distribution of the counts per sample



SERE coefficient [2]

Goal: assess the similarity/dissimilarity between samples

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: example

	T0-1	T0-5	T0-6	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
T0-5	2.97	0	3.00	72.21	70.03	72.33	72.94	75.15	72.32
T0-6	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.

Exploratory data analysis (EDA)

Two main tools:

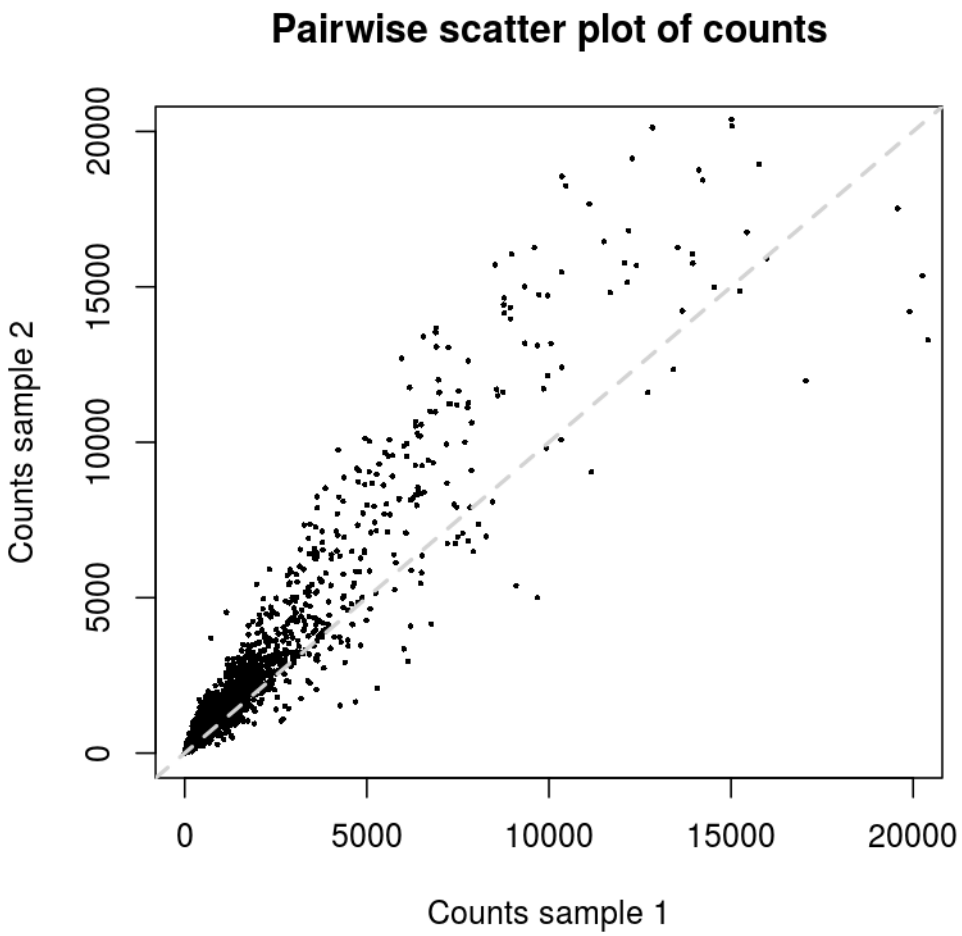
- Principal Component Analysis (PCA)
- Clustering

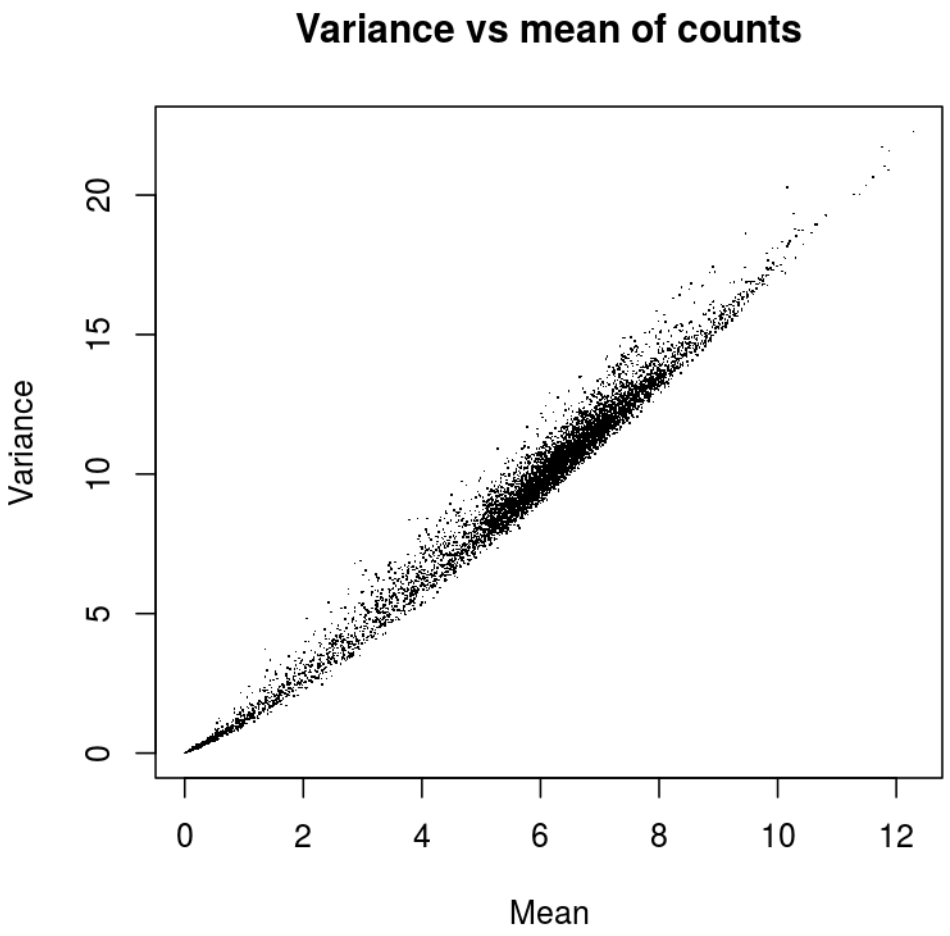
Pre-requisite:

- Notion of **distance** between the samples
- Make the data homoscedastic:

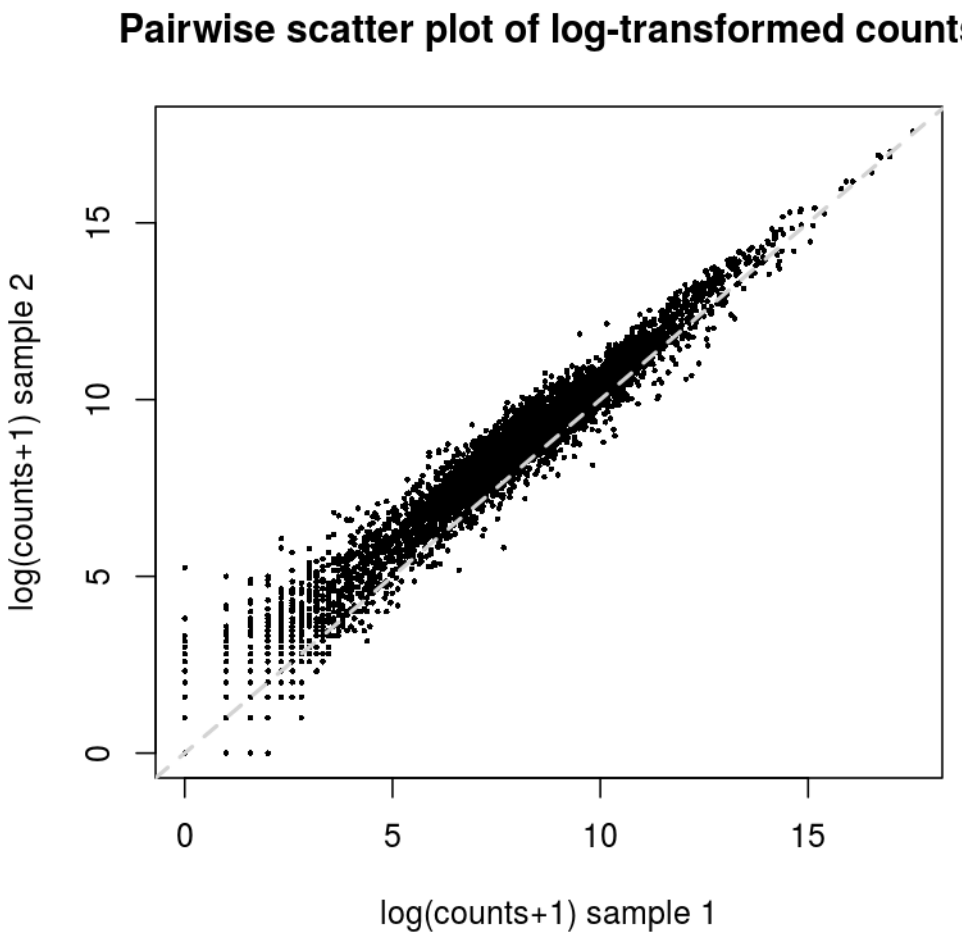
variance must be independent of the mean

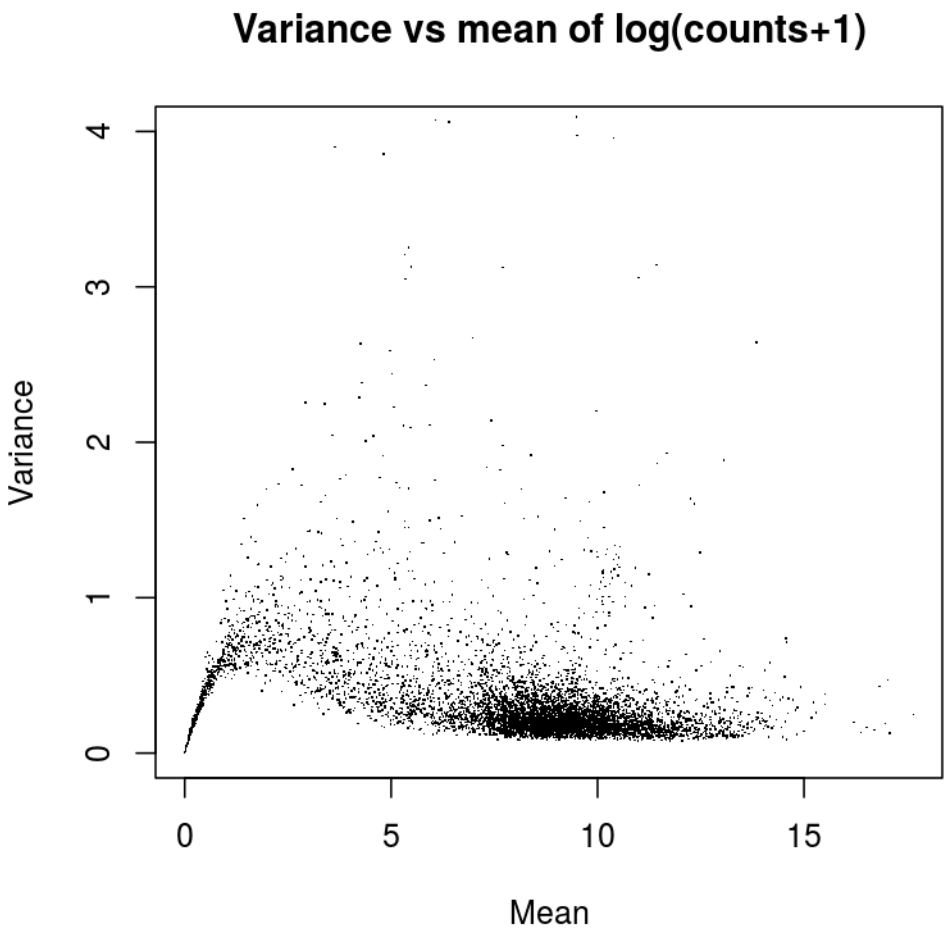
Variance increases with intensity



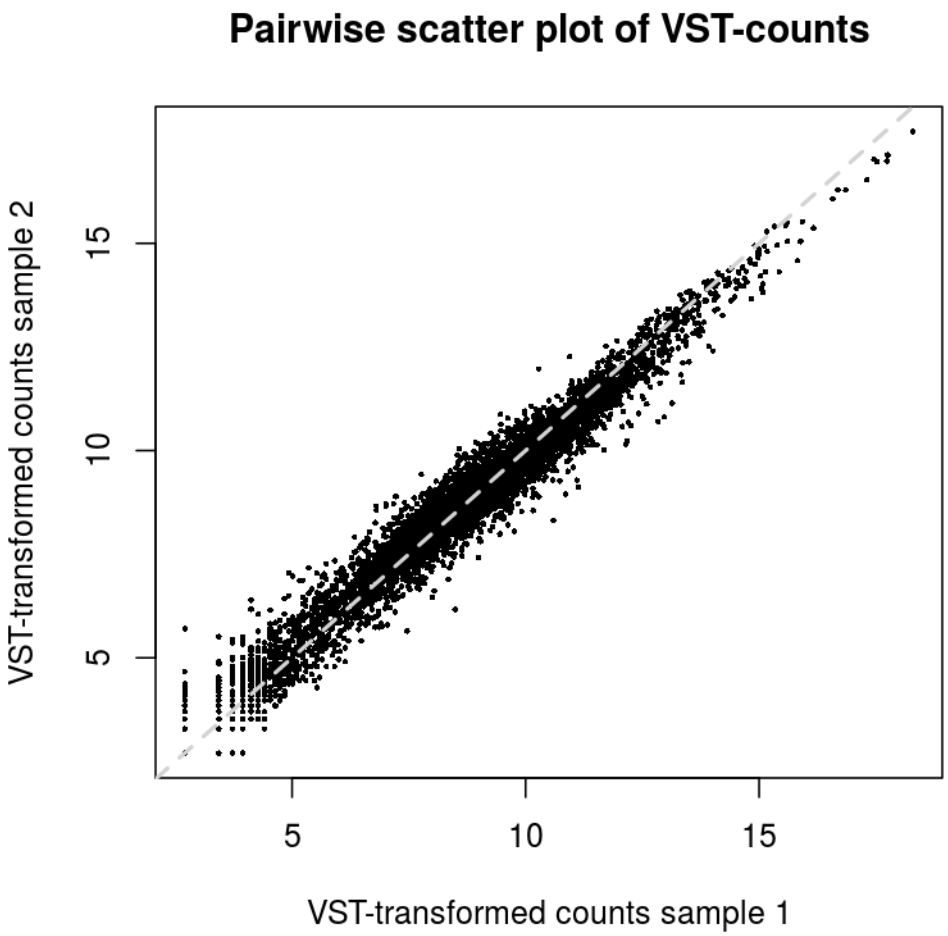


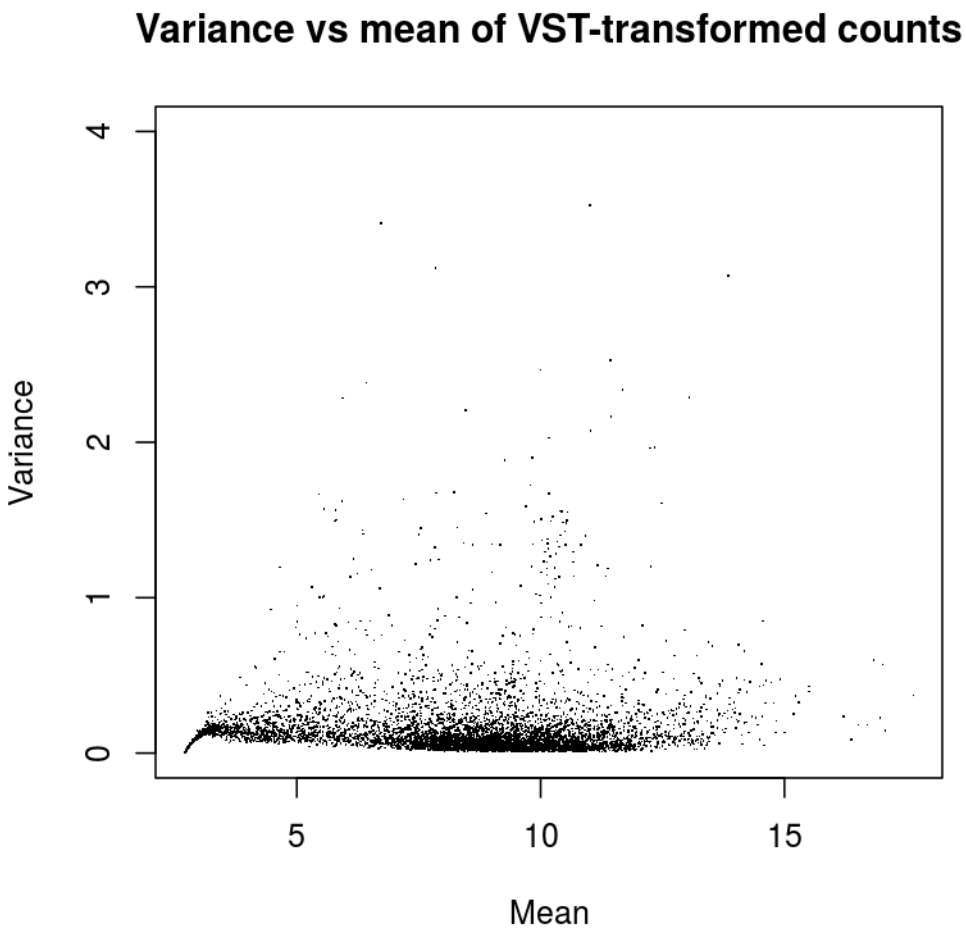
Log-transformation





Variance-Stabalizing Transformation [3]





Use these data to perform Exploratory Data Analysis!

Principal Component Analysis (PCA)

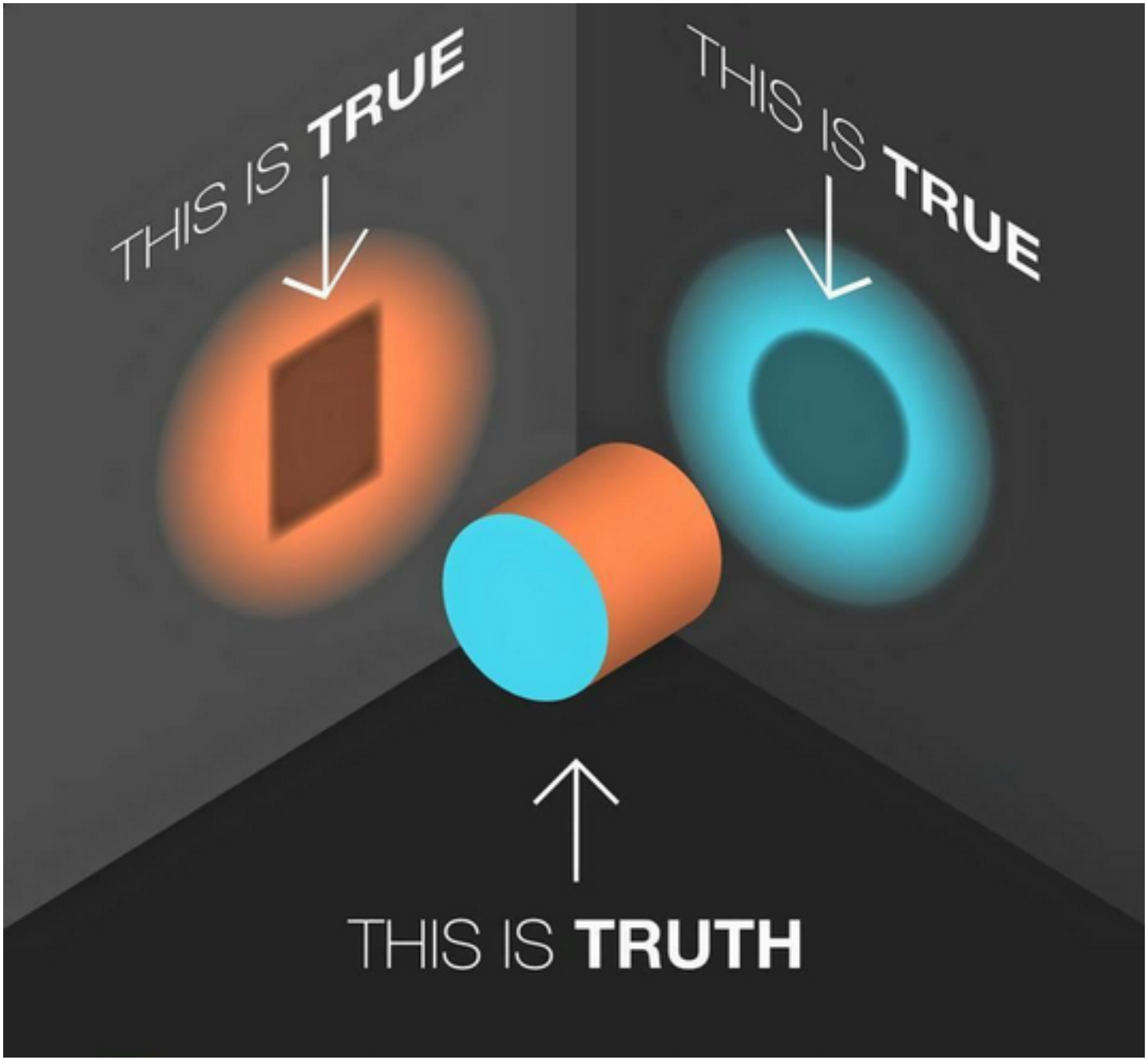
Goal:

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

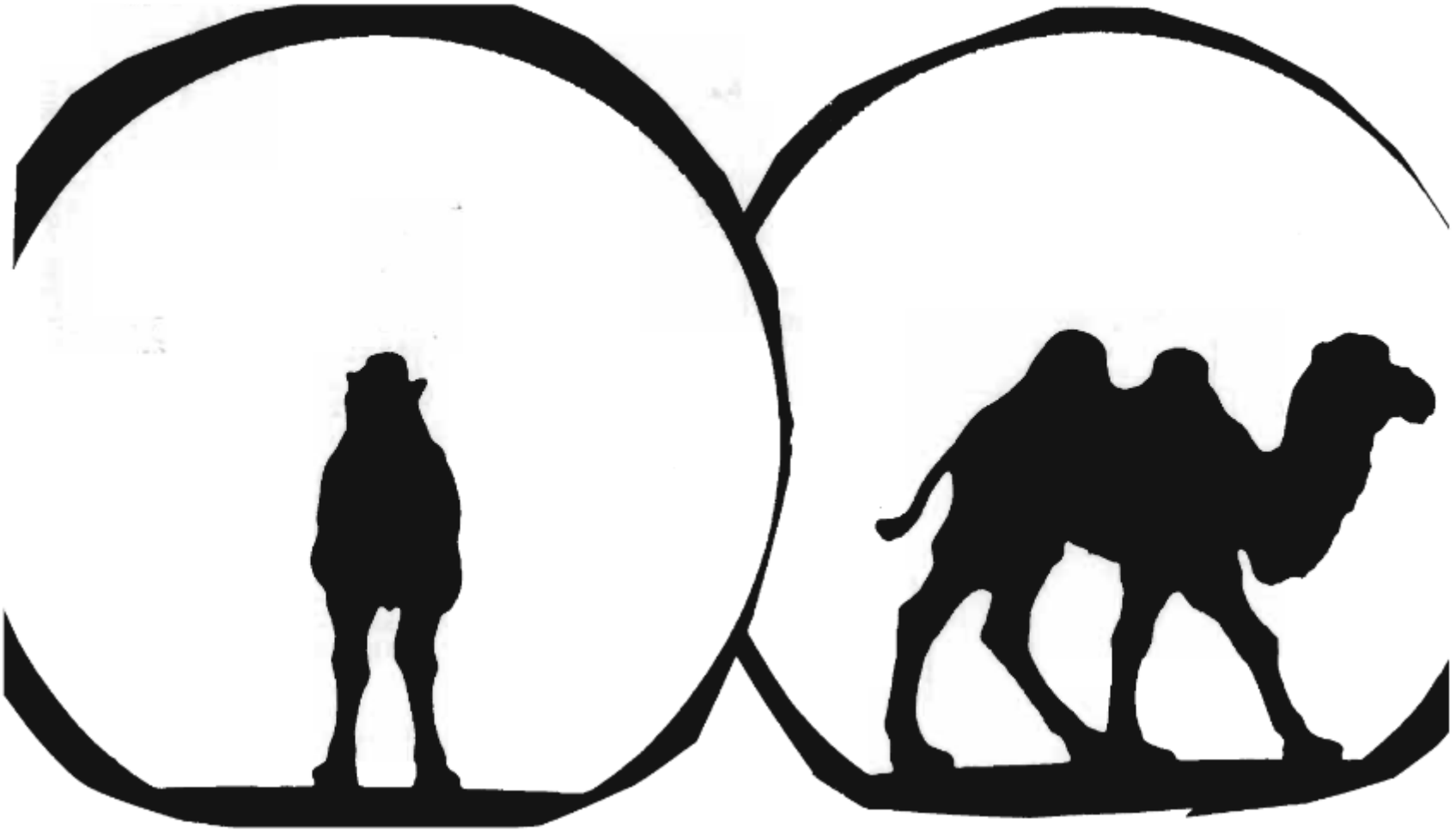
Method:

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

Projection: loss of information

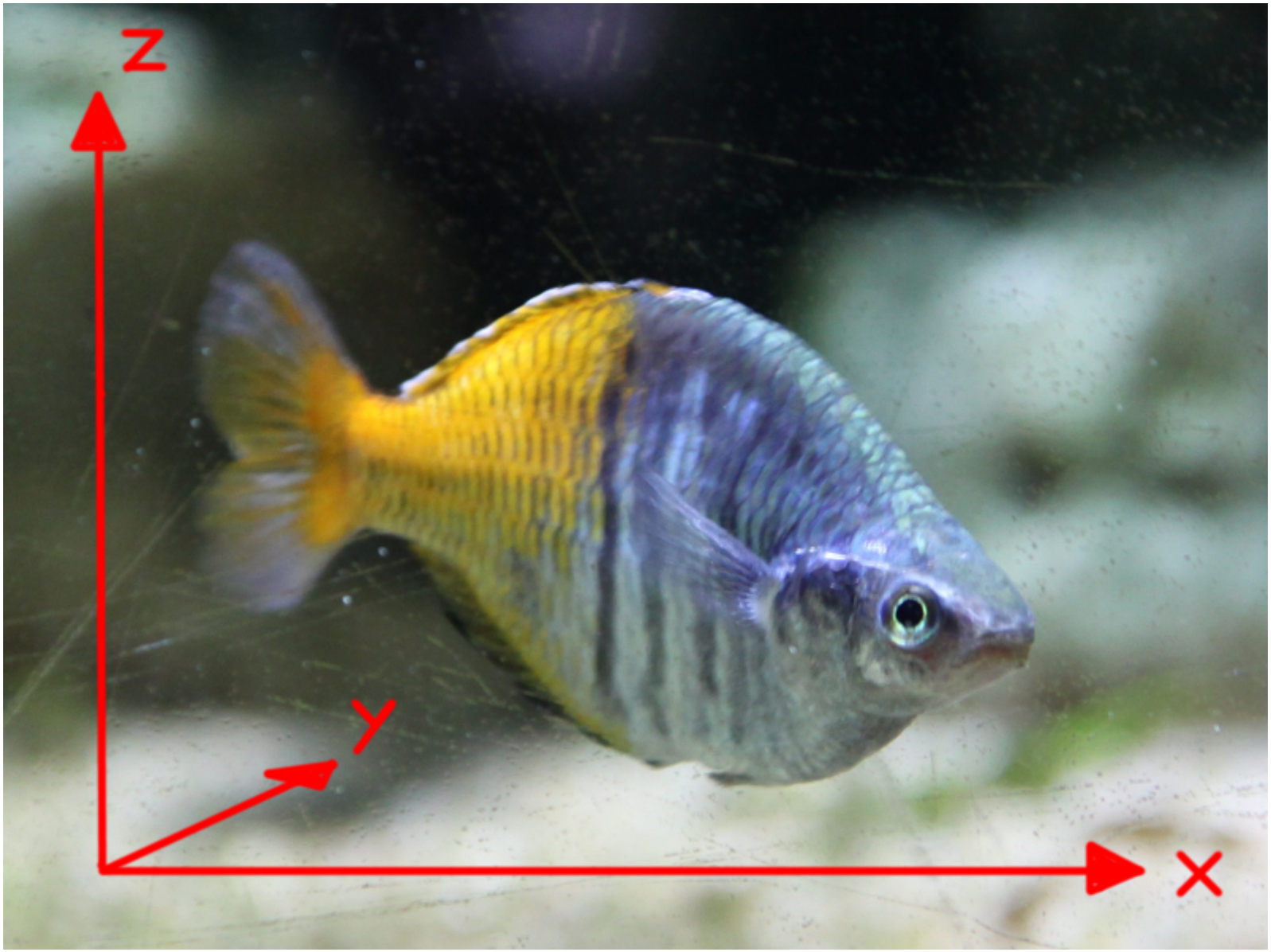


Projection: loss of information

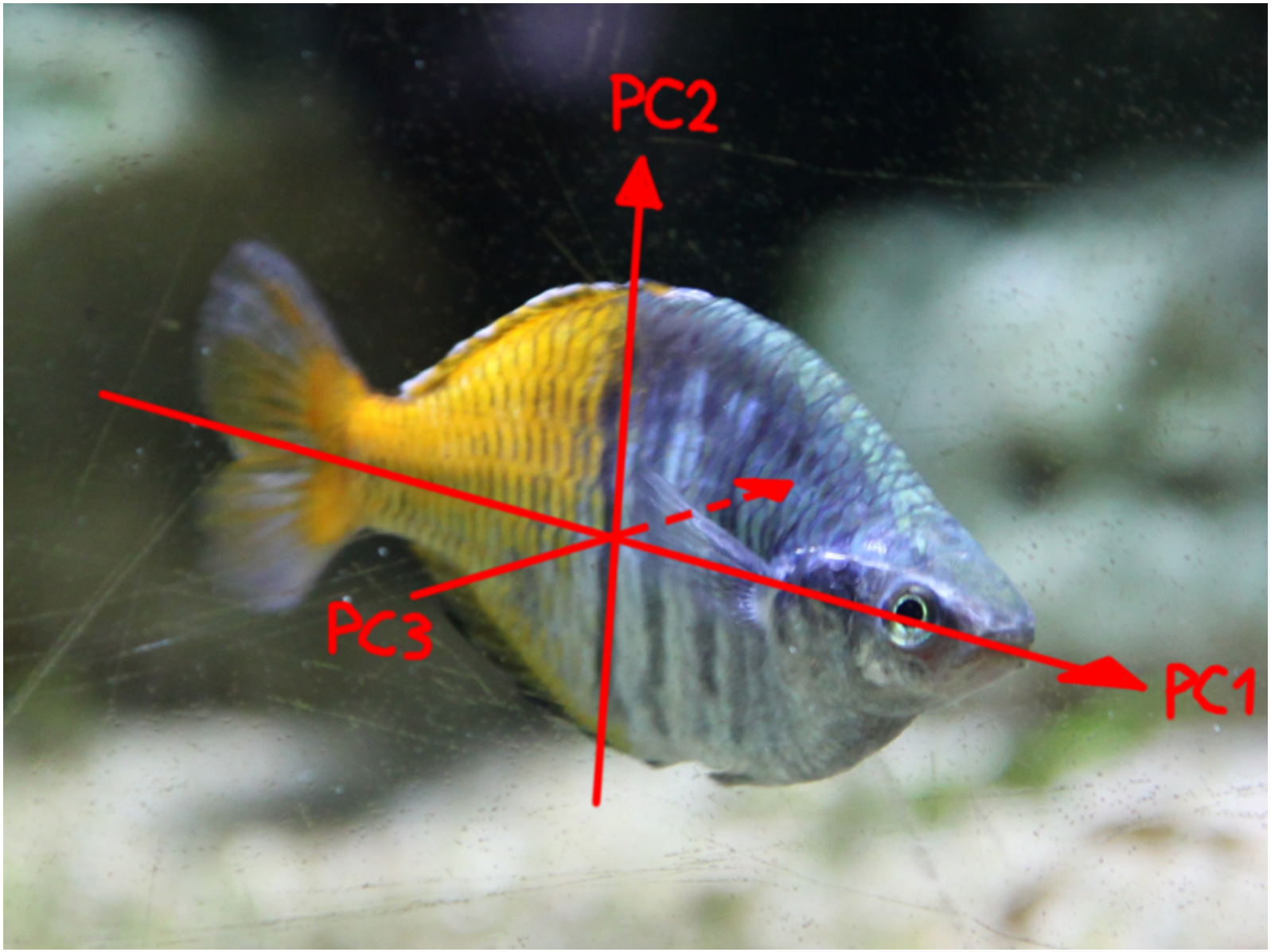


Camel vs. dromedary

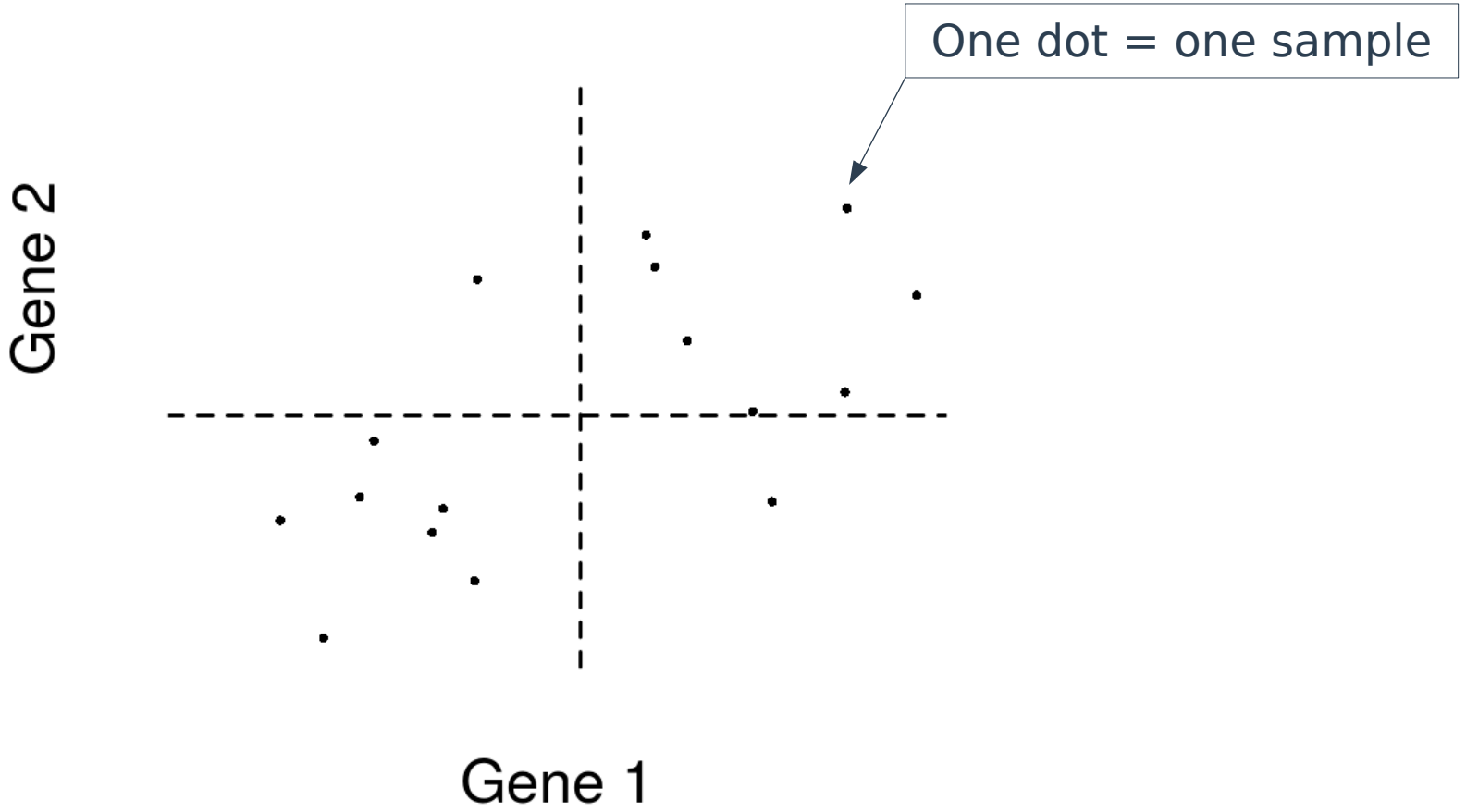
PCA on a fish (source: bioinfo-fr.net)



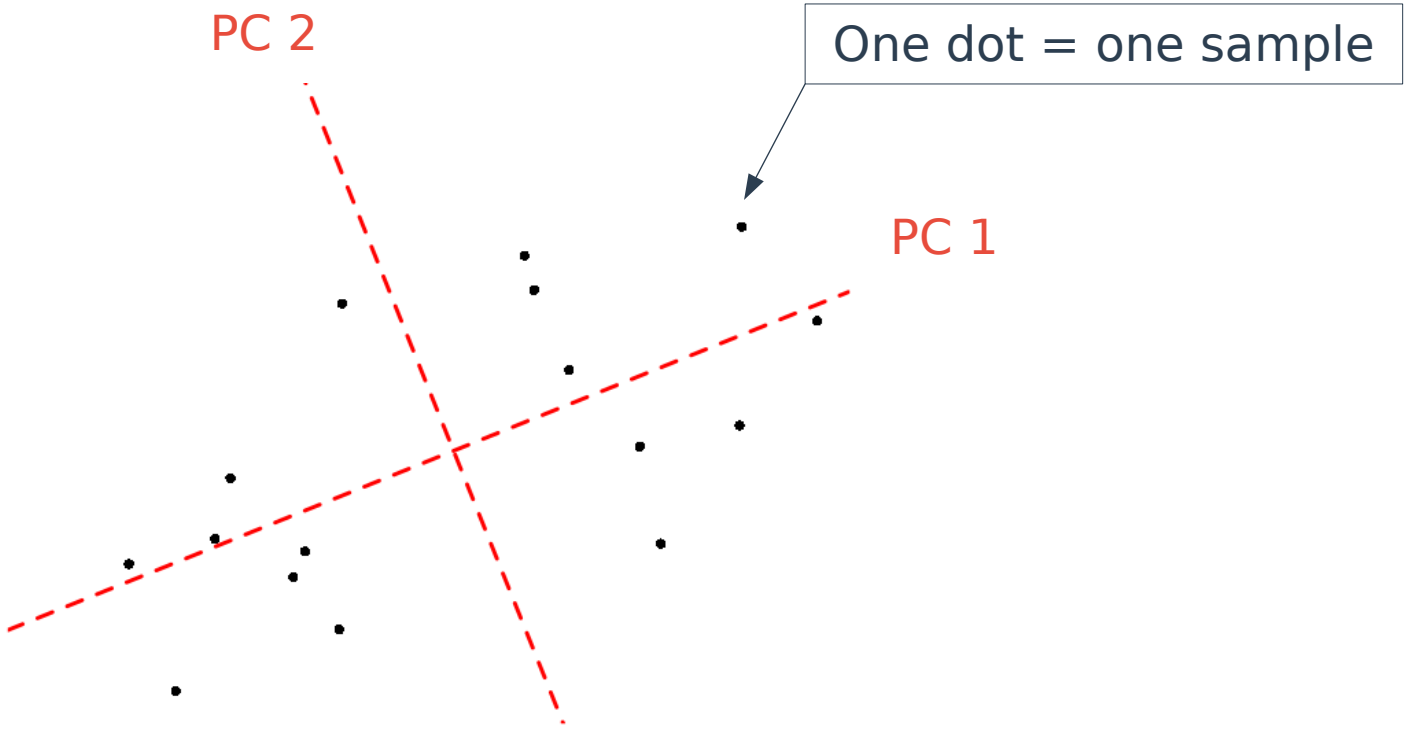
PCA on a fish (source: bioinfo-fr.net)



PCA of a small cloud (2 dimensions)



PCA of a small cloud (2 dimensions)



$$PC_1 = z^1_1 \, Gene_1 + z^1_2 \, Gene_2$$

$$PC_2 = z^2_1 \, Gene_1 + z^2_2 \, Gene_2$$

PCA: important scores

Percentage of inertia associated with an axis:

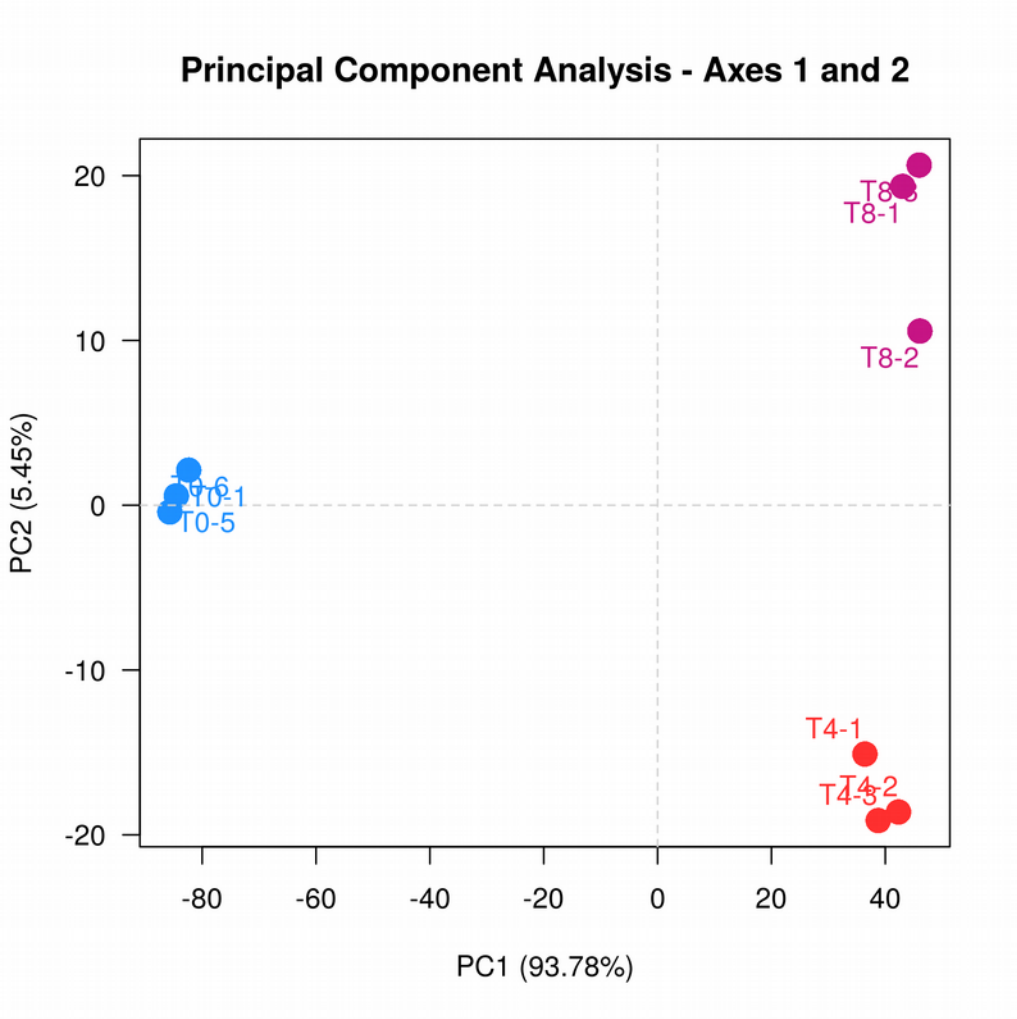
- Proportion of the total information supported by this axis
- Decreases with the axis rank (by construction)

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.

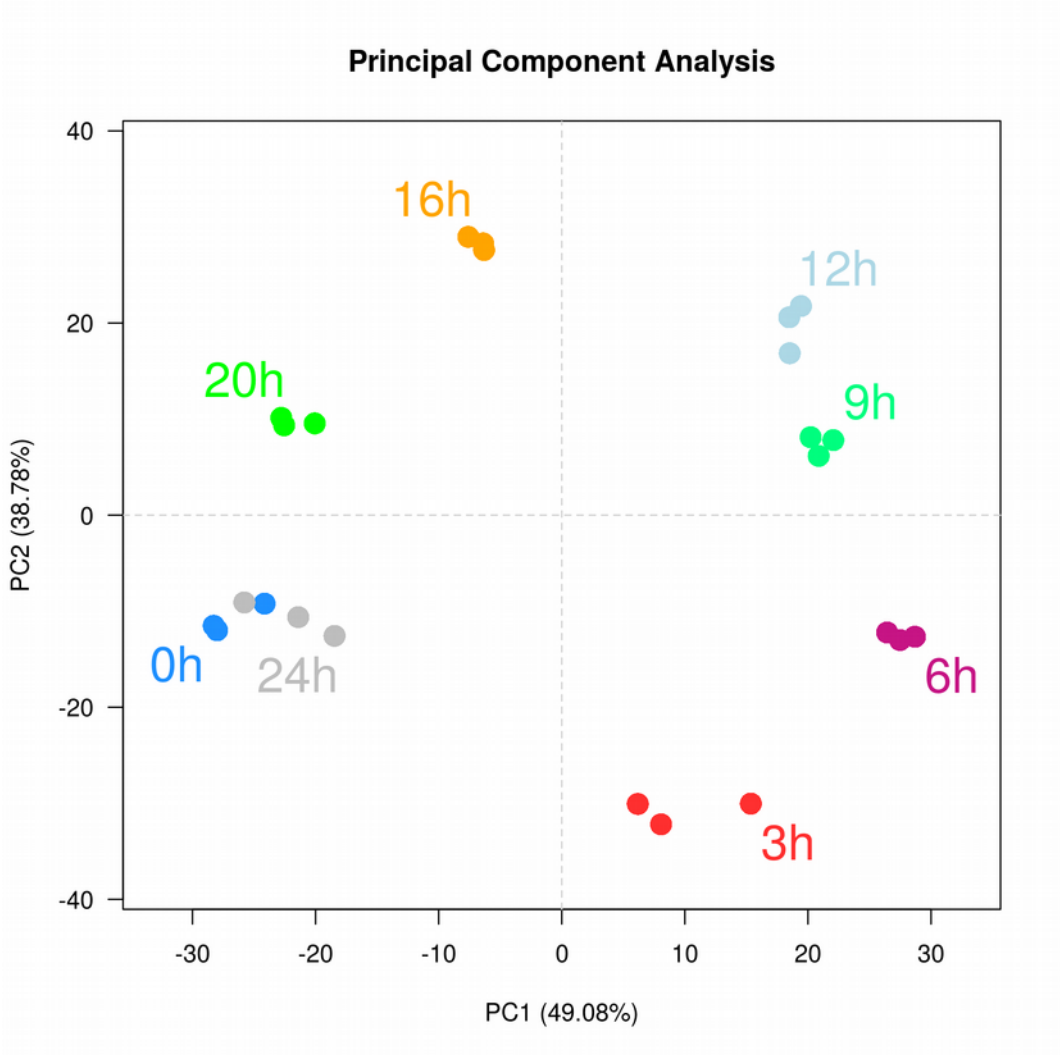
PCA: RNA-Seq example



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

PCA: RNA-Seq example

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



Clustering

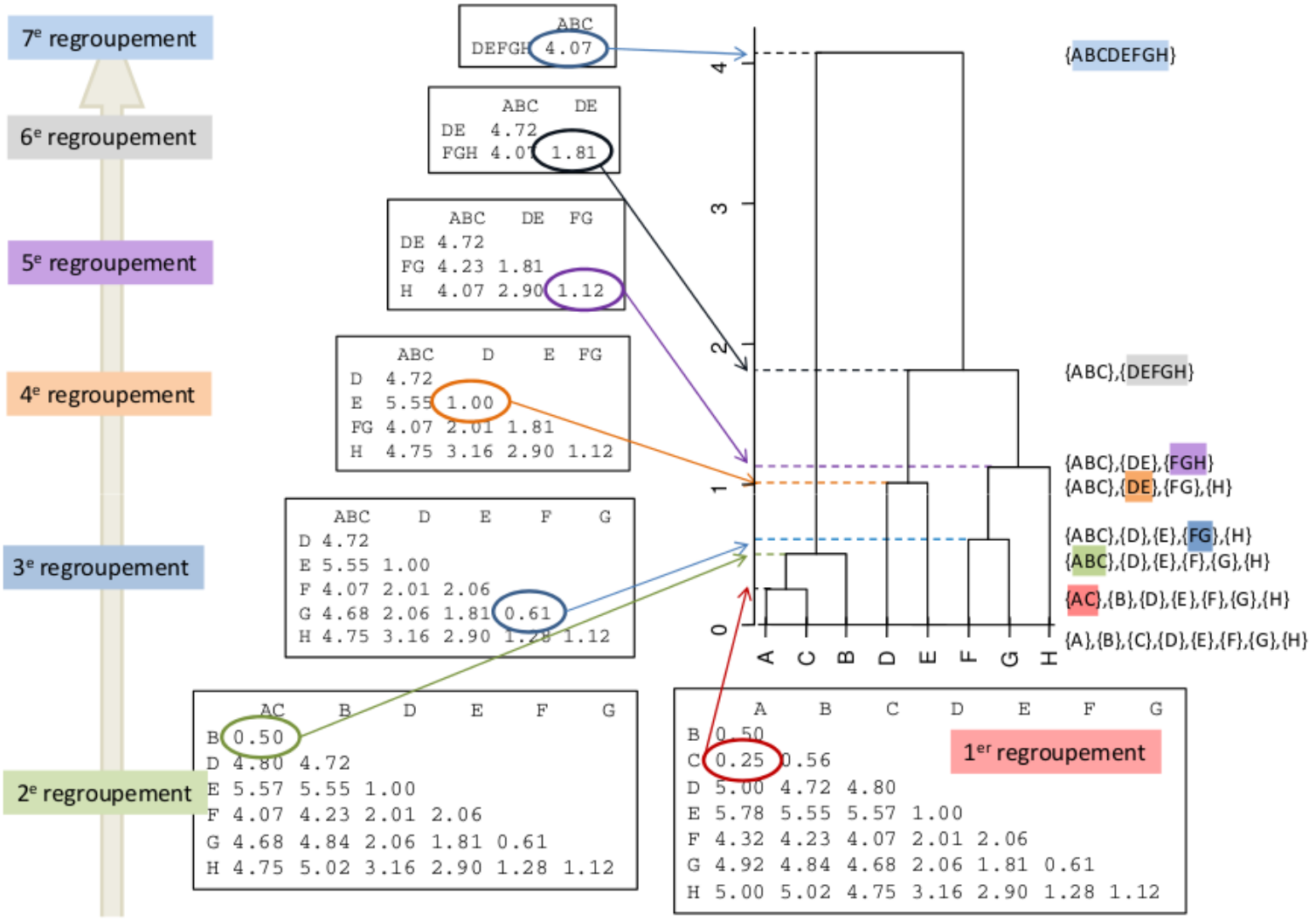
Goal: build groups of samples such that:

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

Method (ascendant clustering):

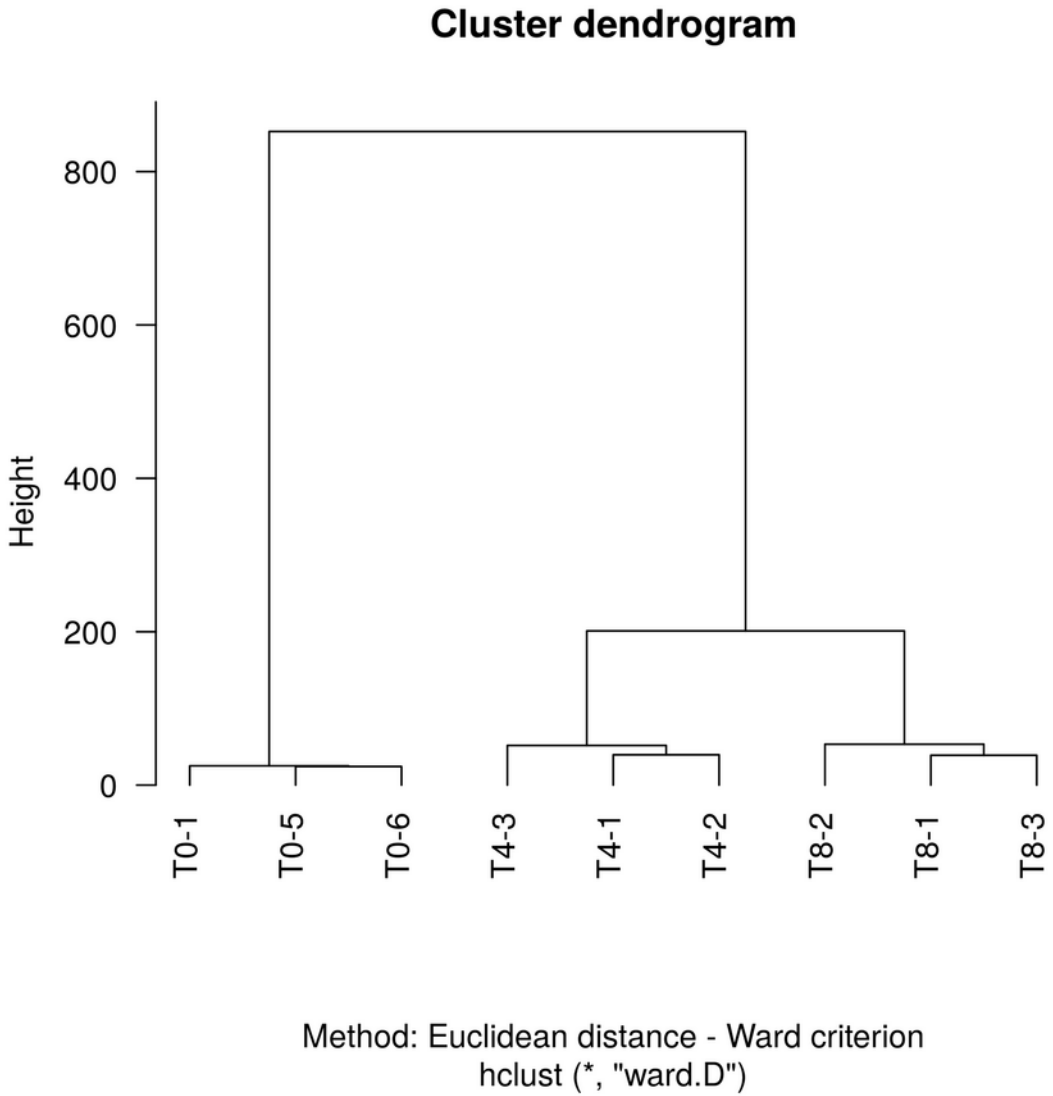
1. Calculate the distances between each pair of samples
2. Gather the two nearest samples into a cluster
3. Calculate the distance between this cluster and each sample
4. Gather the two nearest clusters/samples
5. Go back to step 3 until getting a single cluster

Hierarchical clustering: example



Source: MOOC FUN Analyse de données 2015 – Agrocampus Ouest

Hierarchical clustering: RNA-Seq example



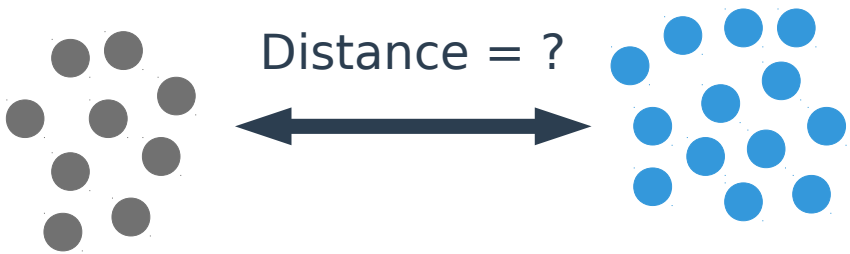
Pre-requisite: counts must be transformed (made homoscedastic) before building the PCA.

Clustering parameters

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

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**

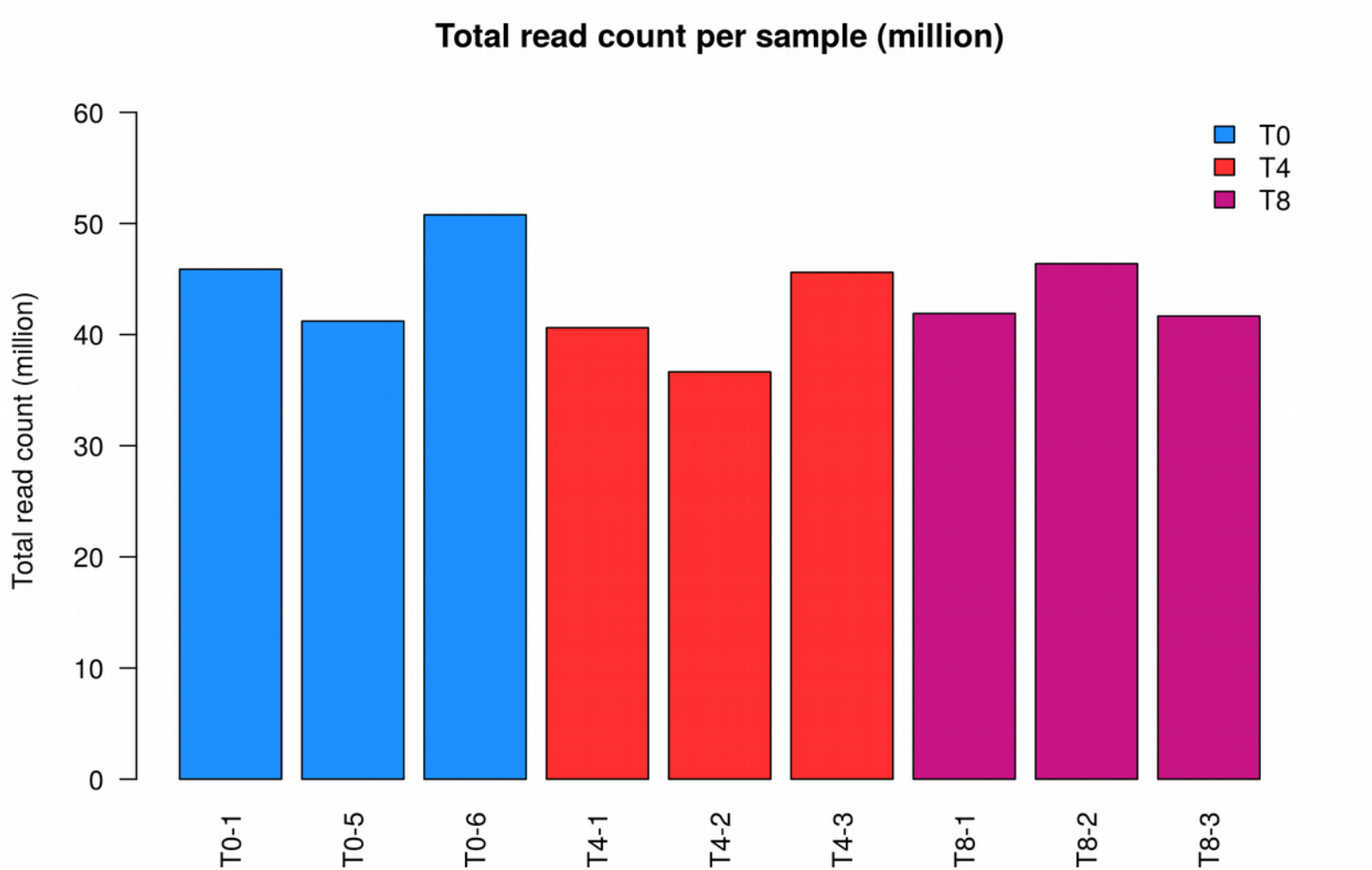


Outline

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).

Notations

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

DESeq2 normalization [3]

DESeq2 computes a size factor s_j per sample:

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

in order to normalize 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:

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

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

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

Assumptions: same than DESeq2.

Other normalization methods

Total number of reads:

$$s_j = \frac{N_j}{\frac{1}{n}\sum_k N_k} \quad \text{or} \quad \frac{N_j}{\sqrt[n]{\prod_k N_k}}$$

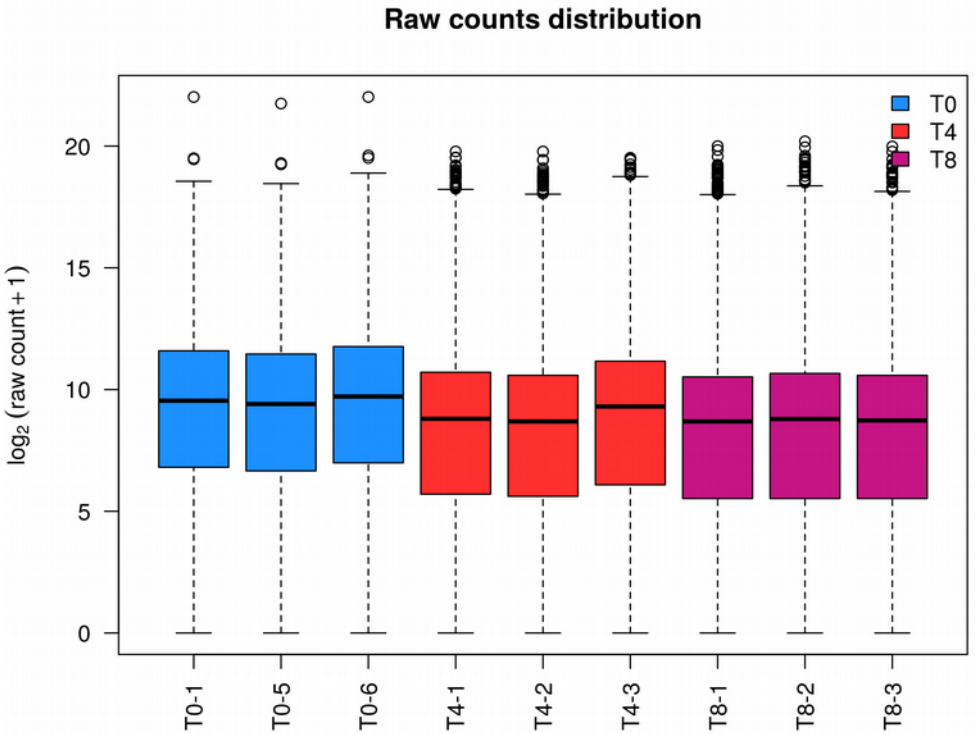
Robustness issue if a gene catches a very high number of reads.

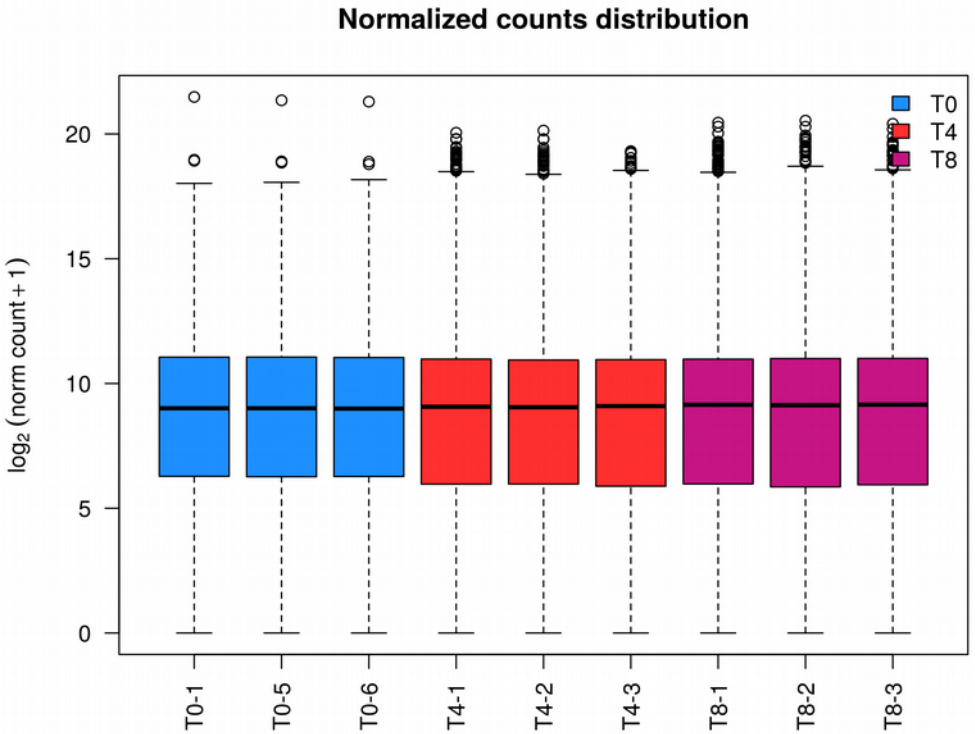
RPKM (Reads Per Kilobase per Million mapped reads):

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

- 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:

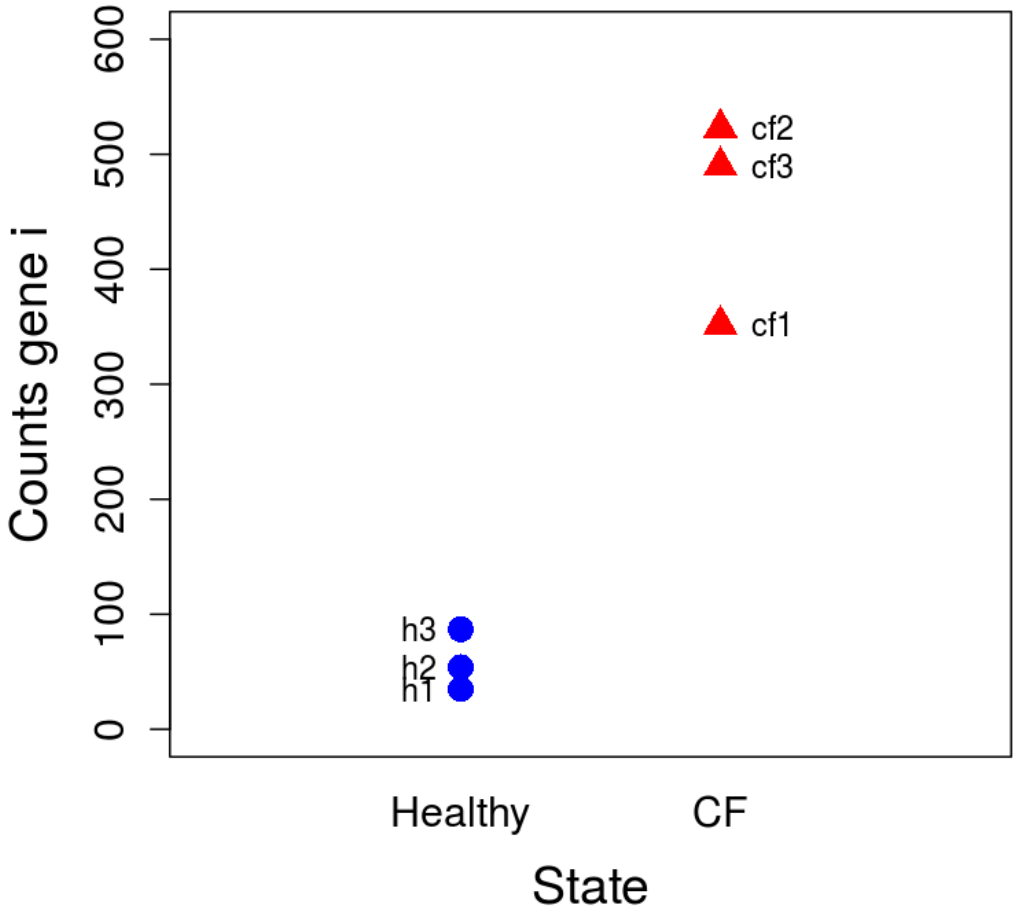
Explain a dependent variable Y thanks to a set a explicative variables $X = (X_1 , ..., X_n)$ using the model:

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

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 *i* thanks to the biological conditions.

Linear model: RNA-Seq example

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

Comparing 3 CF patients and 3 healthy people:

$$\log_2 \begin{pmatrix} 45 \\ 54 \\ 87 \\ 352 \\ 523 \\ 490 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \times \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ \epsilon_{i4} \\ \epsilon_{i5} \\ \epsilon_{i6} \end{pmatrix}$$

Here: $\hat{\beta}_{0i} = 5.95$ and $\hat{\beta}_{1i} = 2.91$

One model per gene → thousands of models!

Why replicate?

Perfect world:

No biological nor technical variability

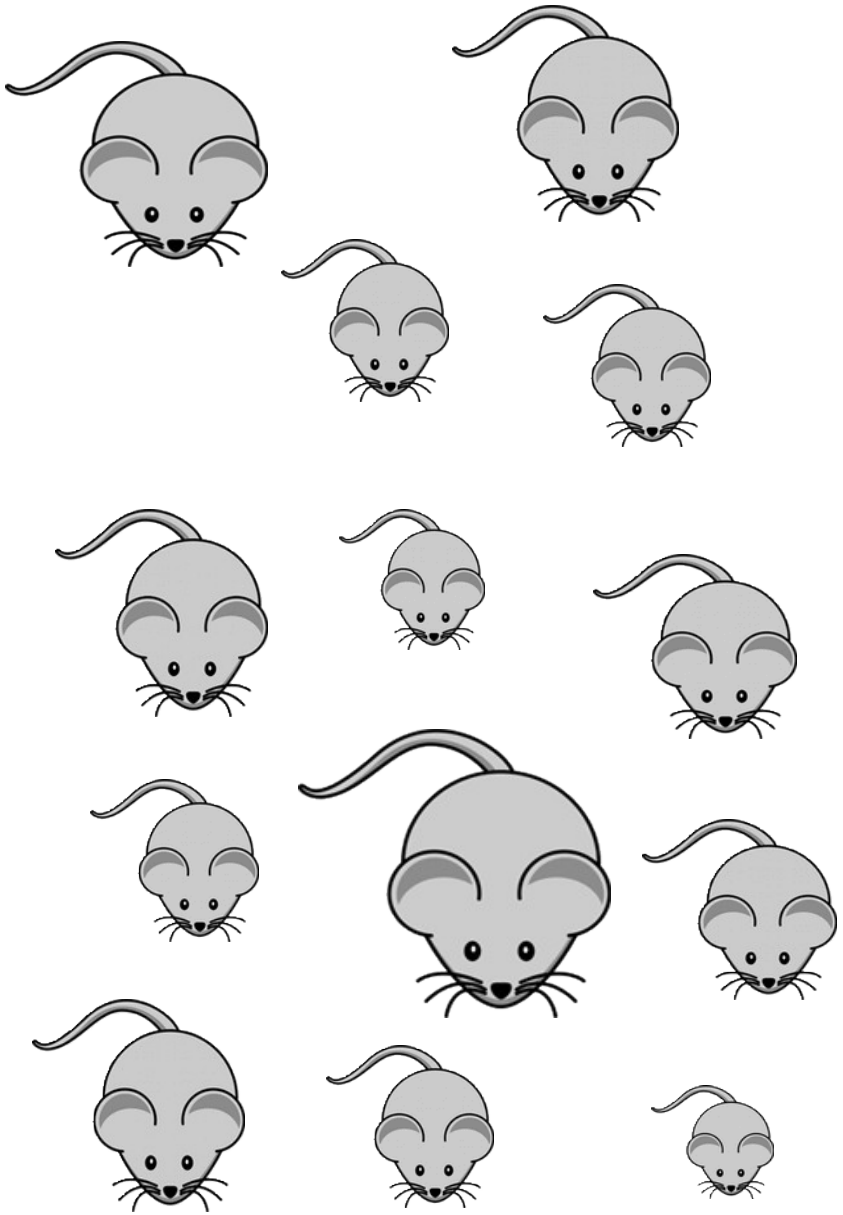
Only one sample from each condition to conclude!

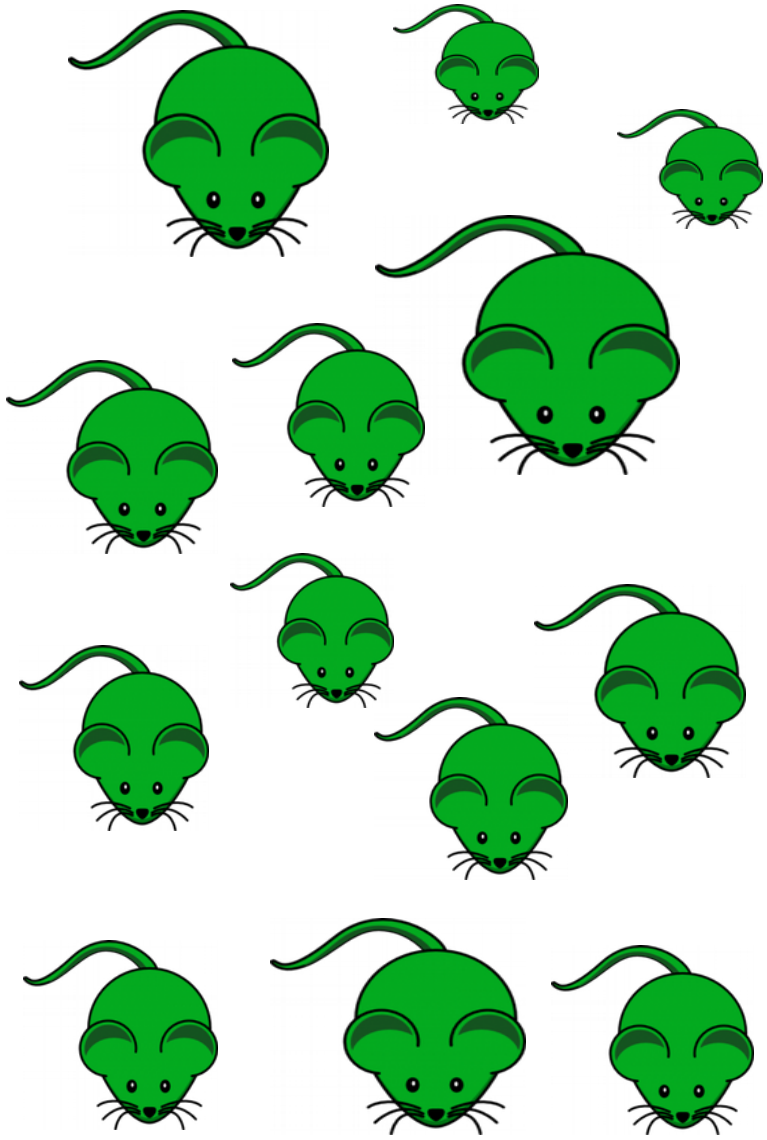
Our world:

Each individual has its own behavior

Need several biological replicates to handle variability

Population: set of all mice we could measure



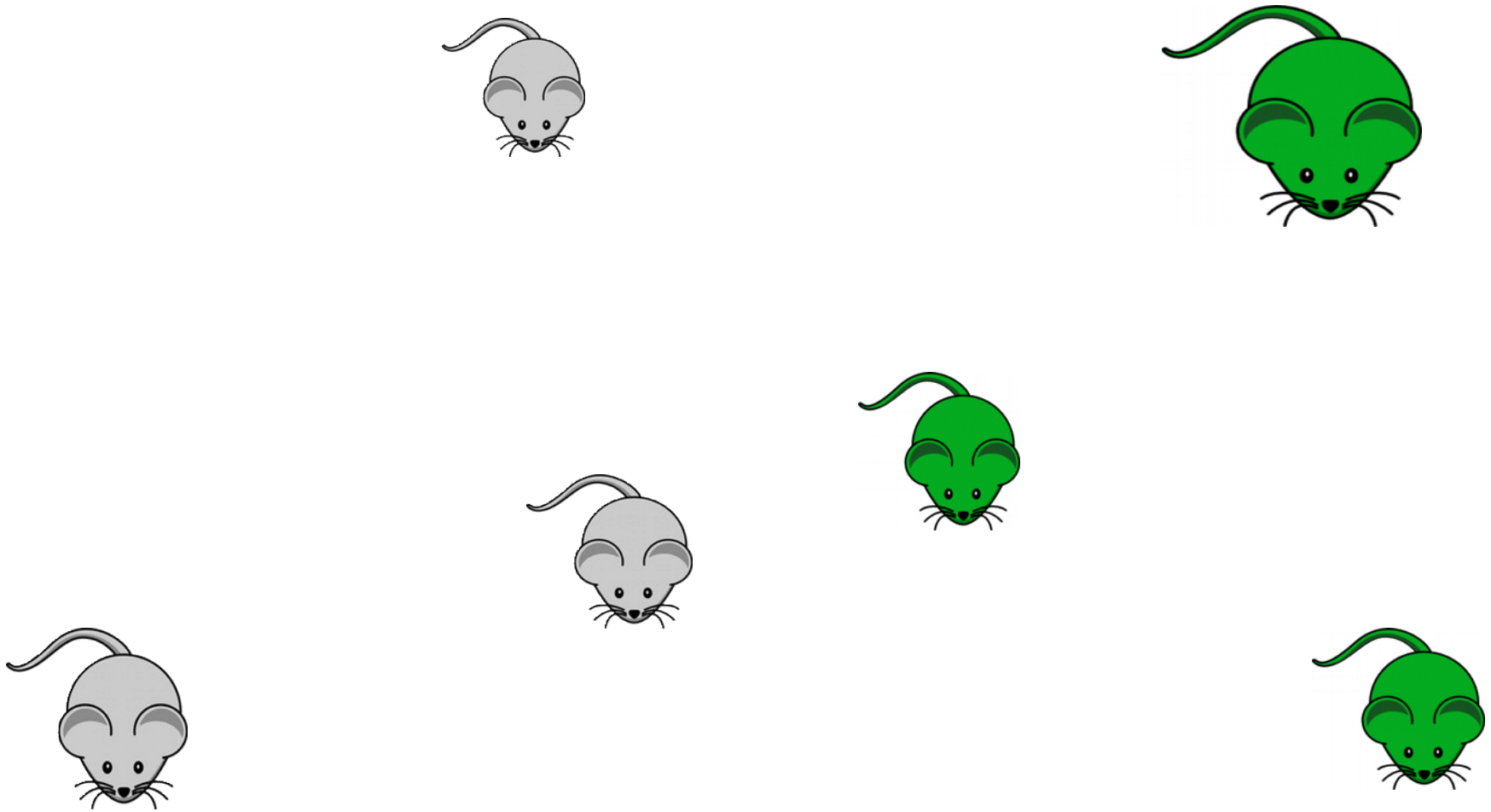


Sampling 1: selection of 3 mice per condition

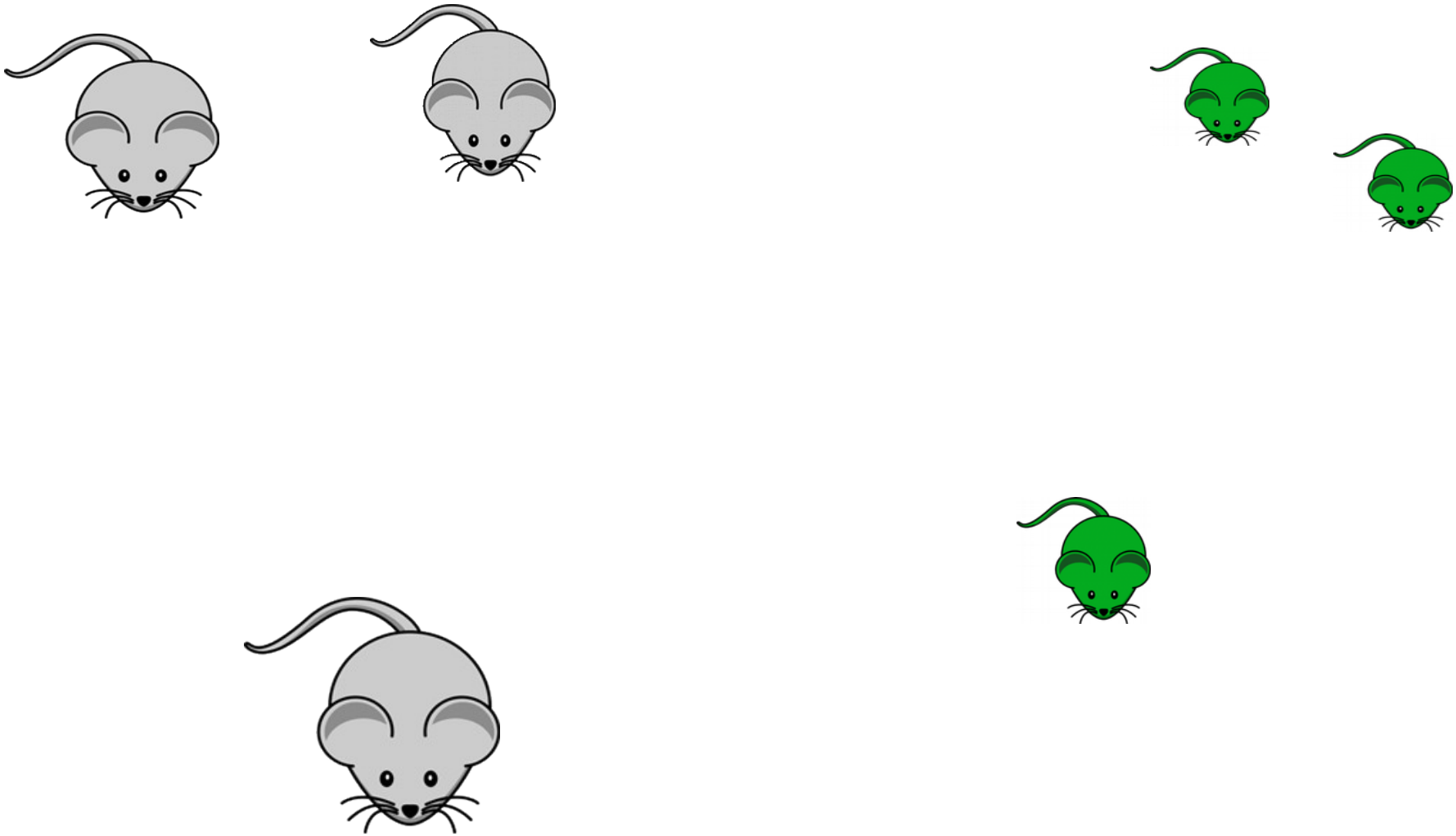





Sampling 2: selection of 3 mice per condition



Sampling 3: non representative



Statistical testing

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

Question:

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

Type I error rate: α

Framework and goal:

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

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

"gene *i* is differentially expressed?"

The risk α is chosen by the statistician before the analysis.

Type II error rate: β

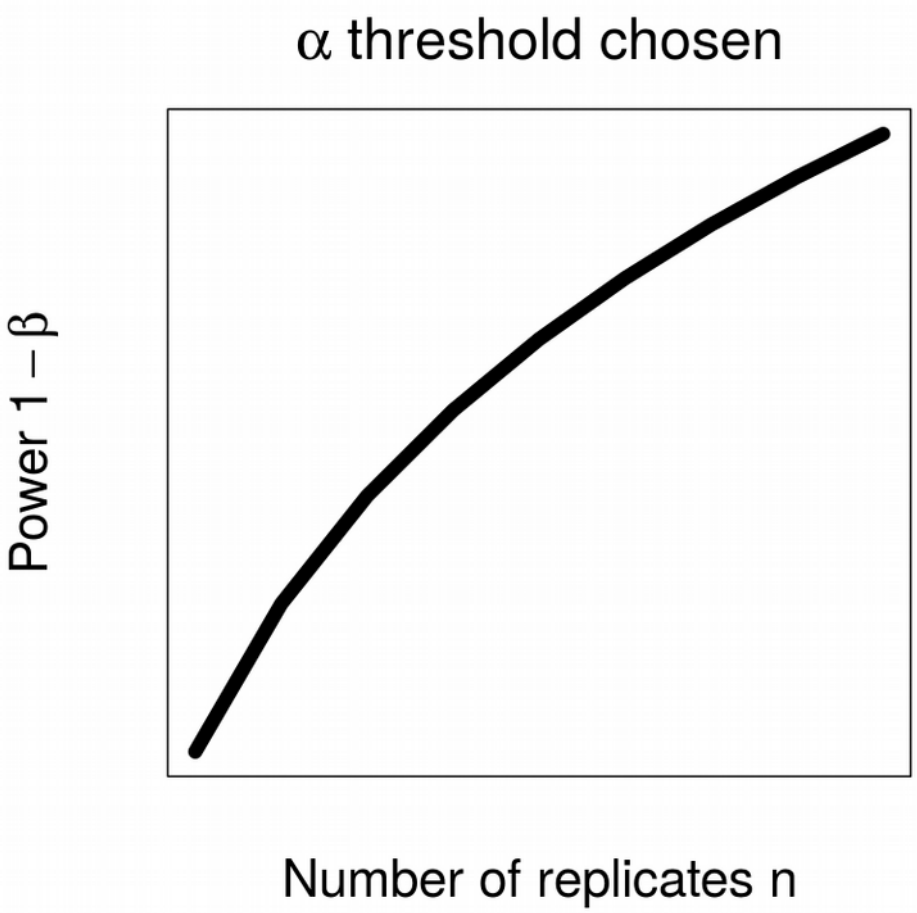
Context:

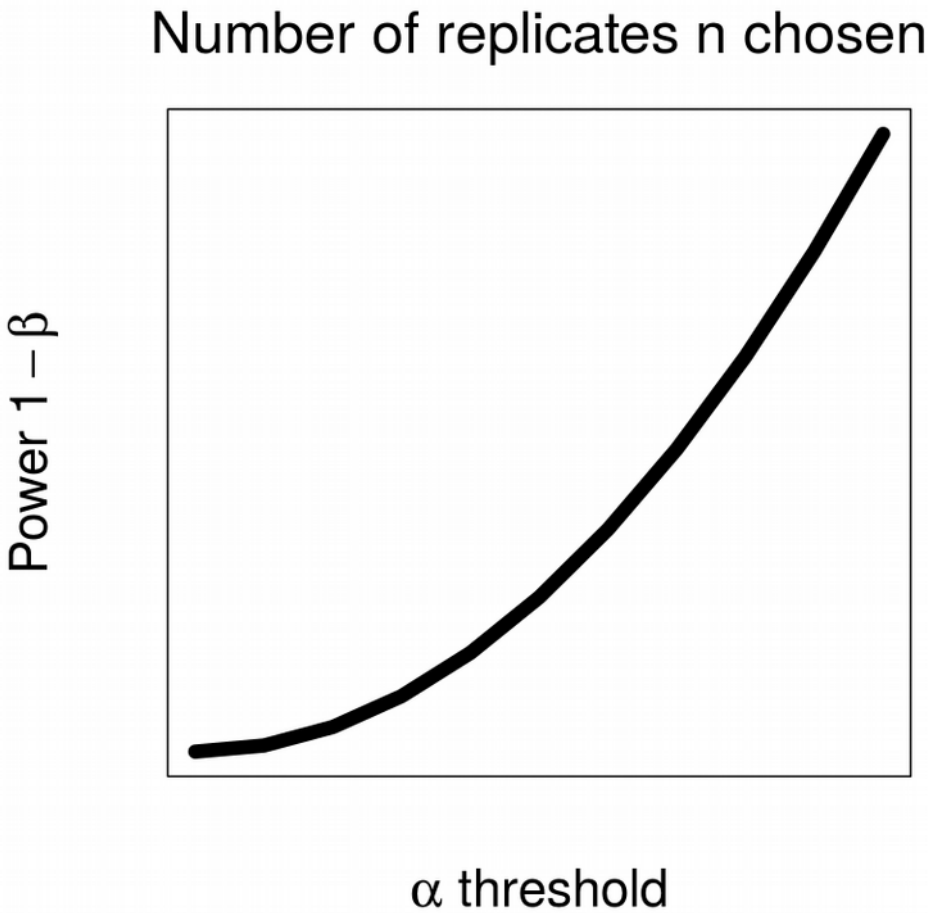
We assume that gene *i* is truly differentially expressed between gray and green mice.

- Which risk β of not discovering gene *i* 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.

α, β and number of replicates n





Formalization

Let μ_1 the average expression of gene *i* 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:

Unknown truth	Decision: Do not reject H_0	Decision: Reject H_0
H_0 true	$1 - \alpha$	α
H_0 false	β	$1 - \beta$

p-value and conclusion of the test

Definition:

$$
\begin{aligned}
p\text{-value} &= \text{Proba}(\text{reject } H_0 \mid H_0 \text{ true}) \\
&= \text{Proba}(\text{doing a mistake when rejecting } H_0) \\
&= \text{Proba}(\text{observed difference is due to hazard})
\end{aligned}
$$

Conclusion:

$$\text{if } p\text{-value} \leq \alpha \text{ then we reject } H_0$$

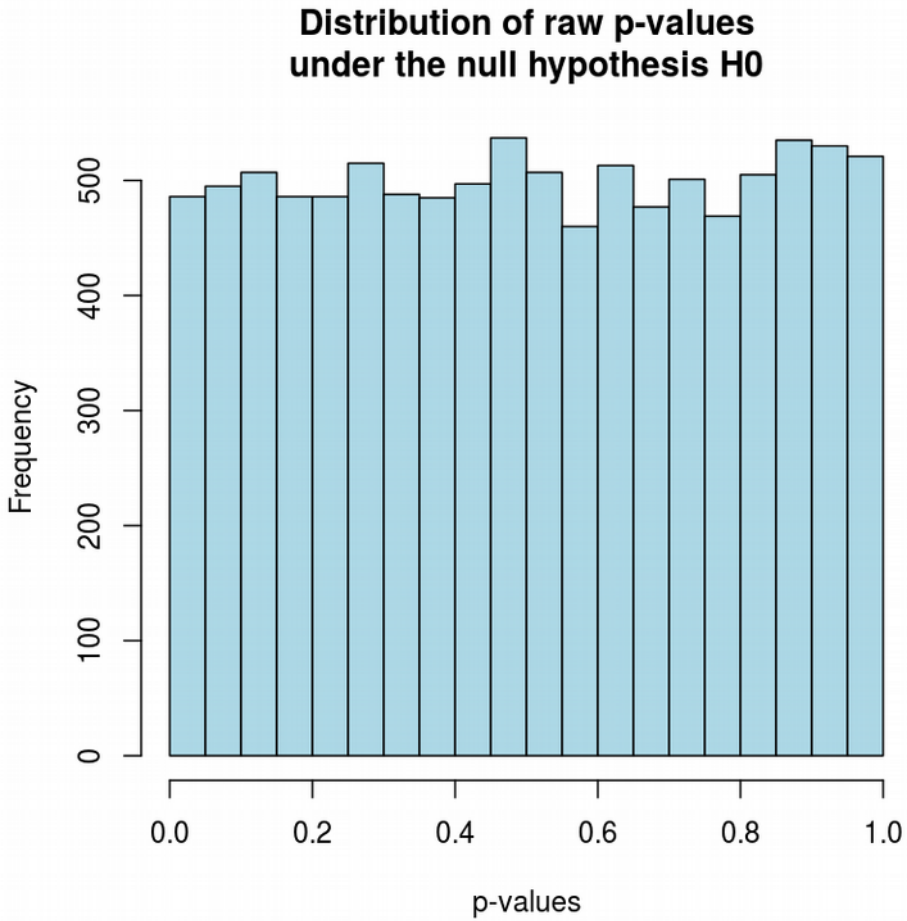
Equal Fold-Changes – different *p*-values

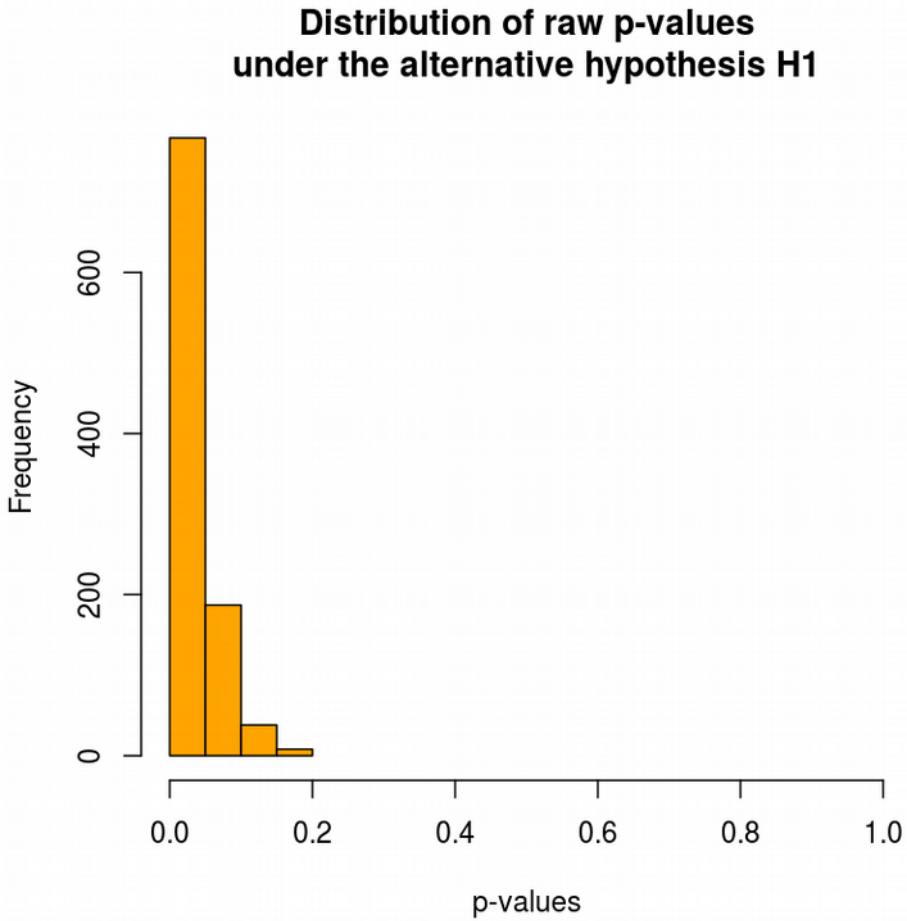
Reminder: Fold-Change definition:

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

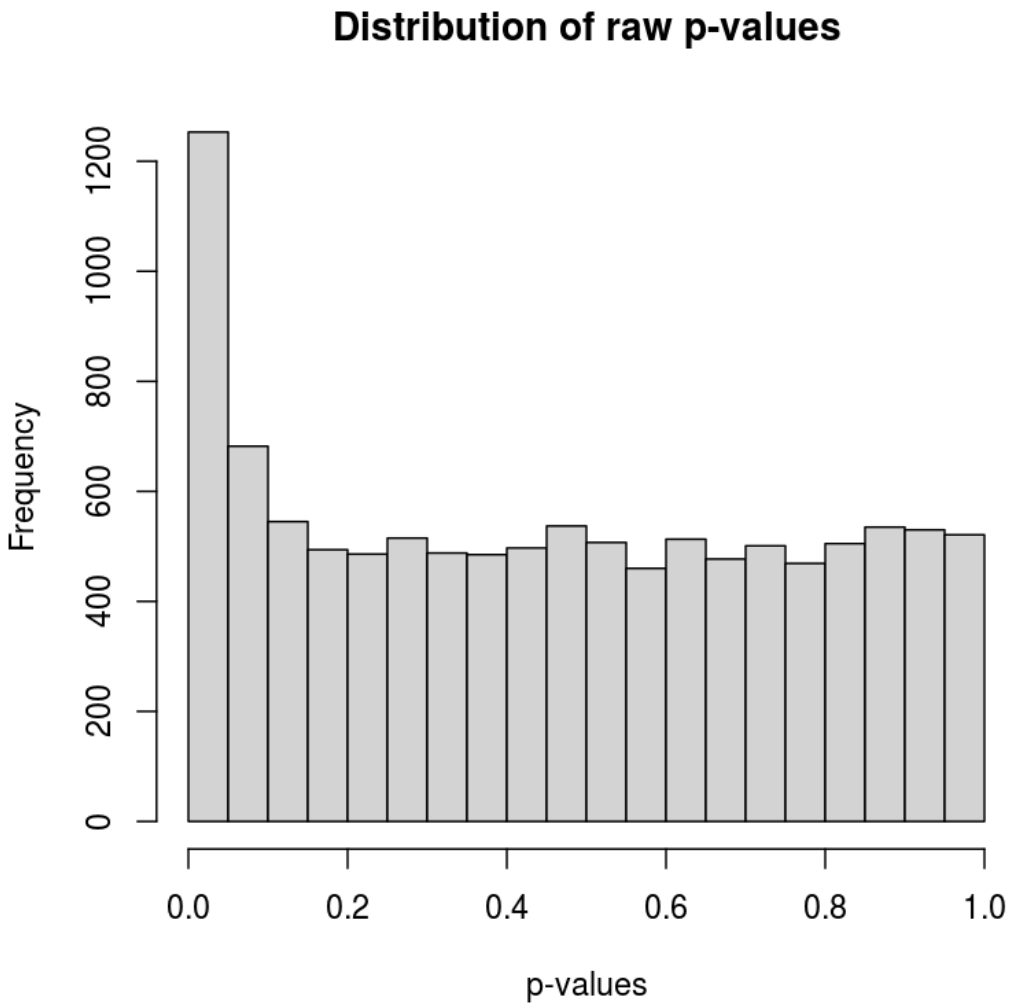
Gene	m1	m2	m3	m4	m5	m6	FC	*p*-value
gene1	5	7	6	2	2	2	3	0.06
gene2	800	1000	900	350	250	200	3	0.03
gene3	700	900	1100	350	200	250	3	0.10
gene4	900	500	1300	200	550	50	3	0.06
...	...	...	...	...	...	...	...	...

Distribution of raw *p*-values





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 $\leq \alpha$ 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.

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

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

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

Similarities and differences

Similarities:

- Negative Binomial distribution
- Generalized Linear Model (GLM)

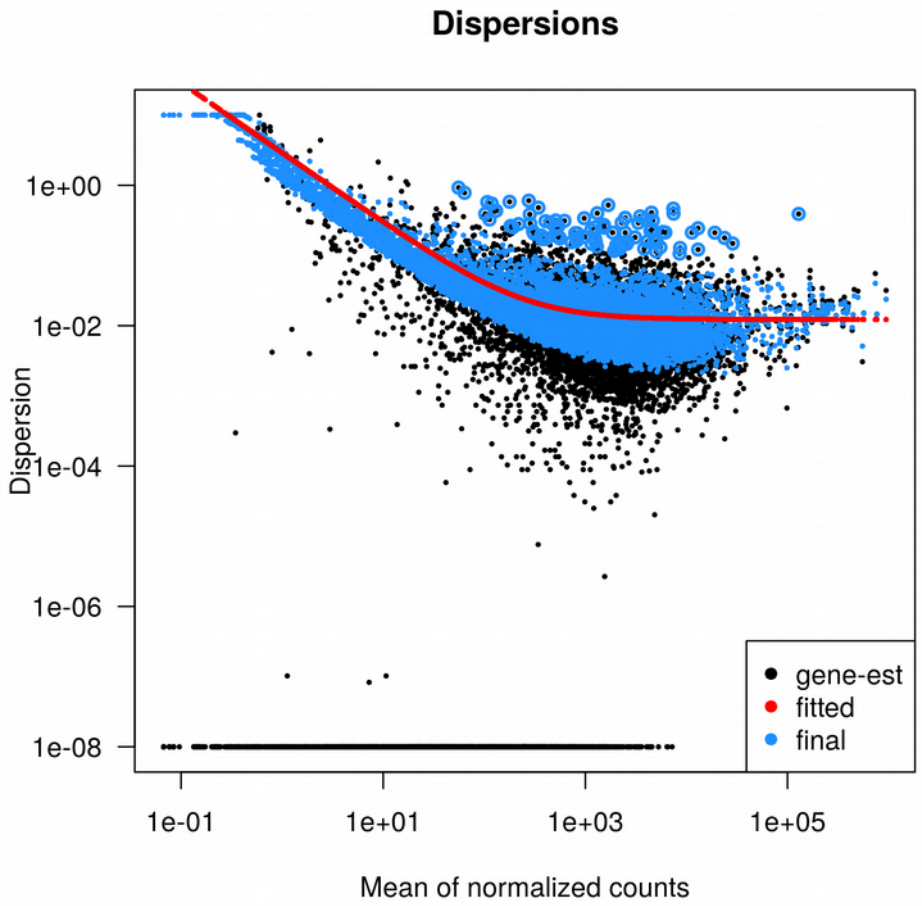
Differences:

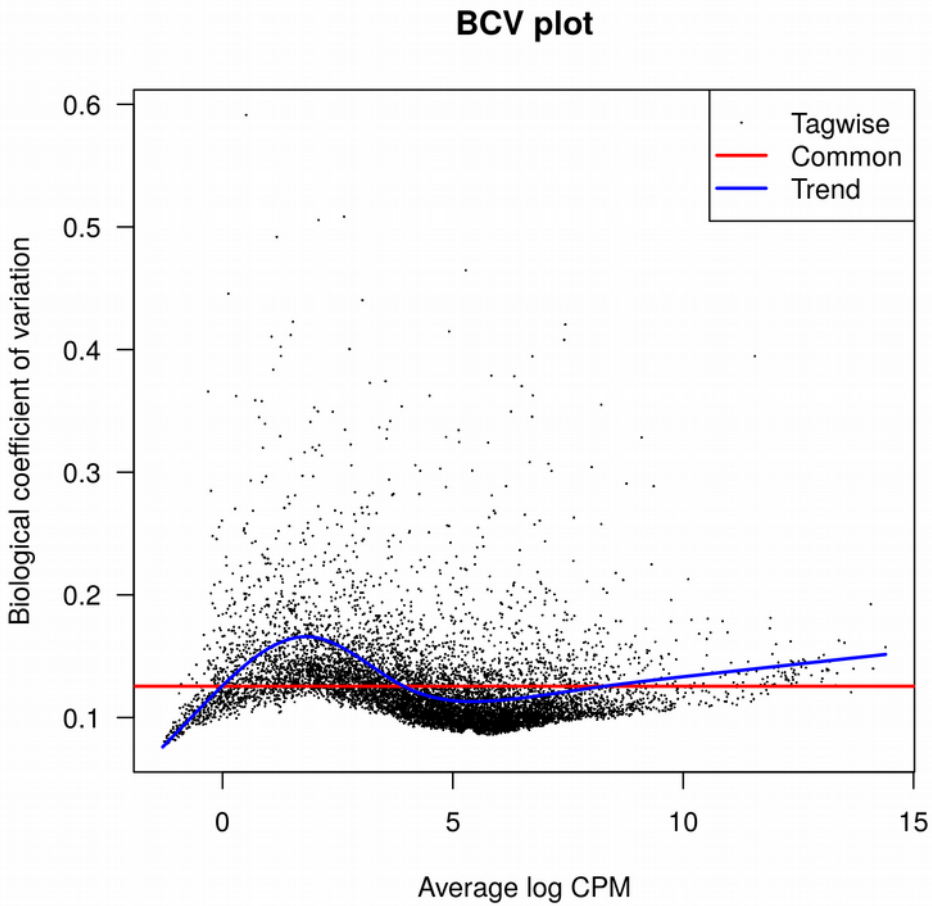
- Dispersion estimation
- Way of dealing with outlier counts
- Low counts filtering

Dispersion estimation φ_i: DESeq2 vs edgeR

Reminder:

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





Statistical testing

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

- an estimation of $\beta_i = \log_2(FC_i)$
- the precision of this estimation (standard error)
- so the *p*-value associated with gene *i*

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

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

Input files

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

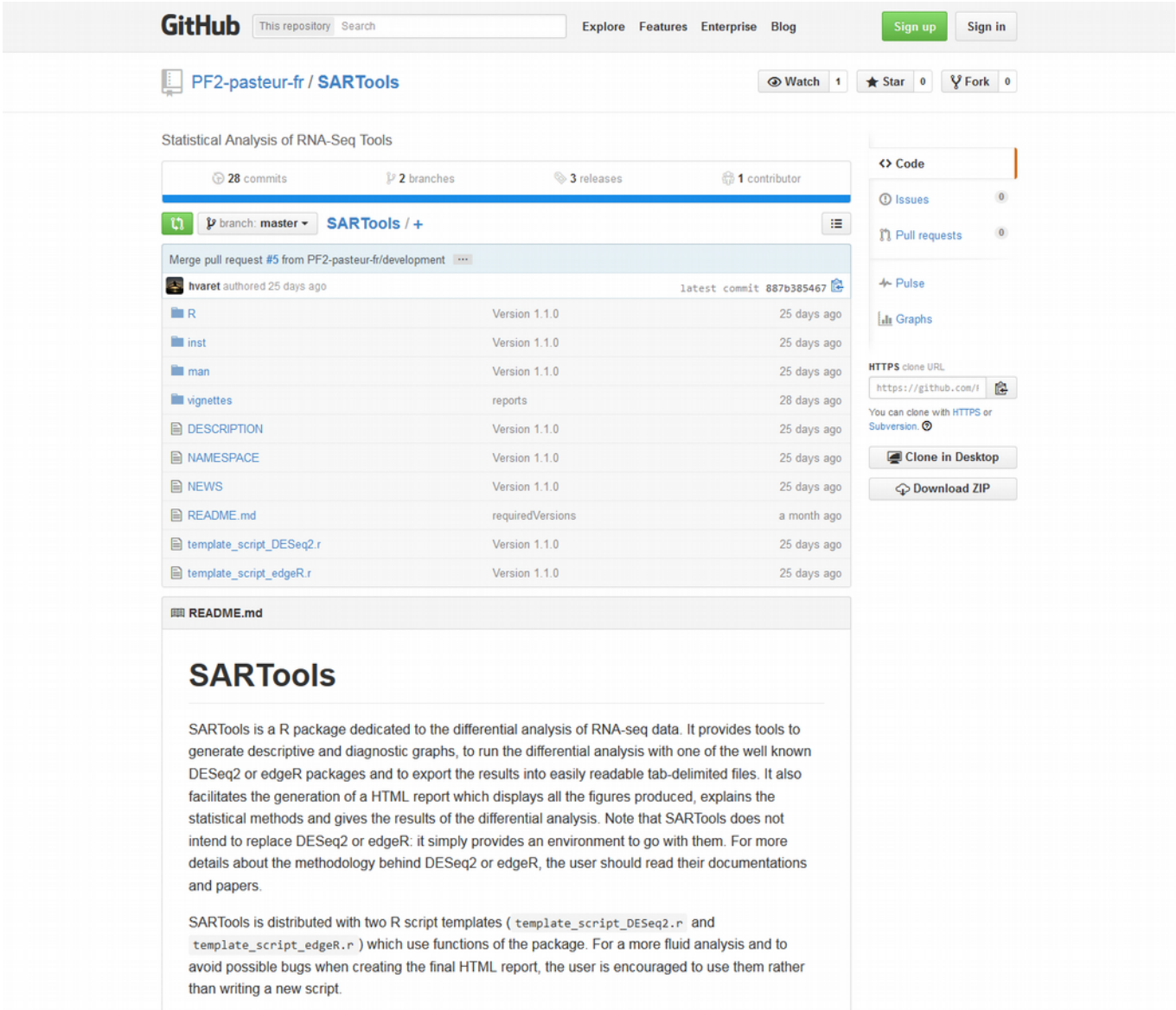
```
label   files             condition
WT1     WT1.counts.txt    WT
WT2     WT2.counts.txt    WT
KO1     KO1.counts.txt    KO
KO2     KO2.counts.txt    KO
```

Counts: one tab-delimited text file per sample:

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

Source code available on GitHub

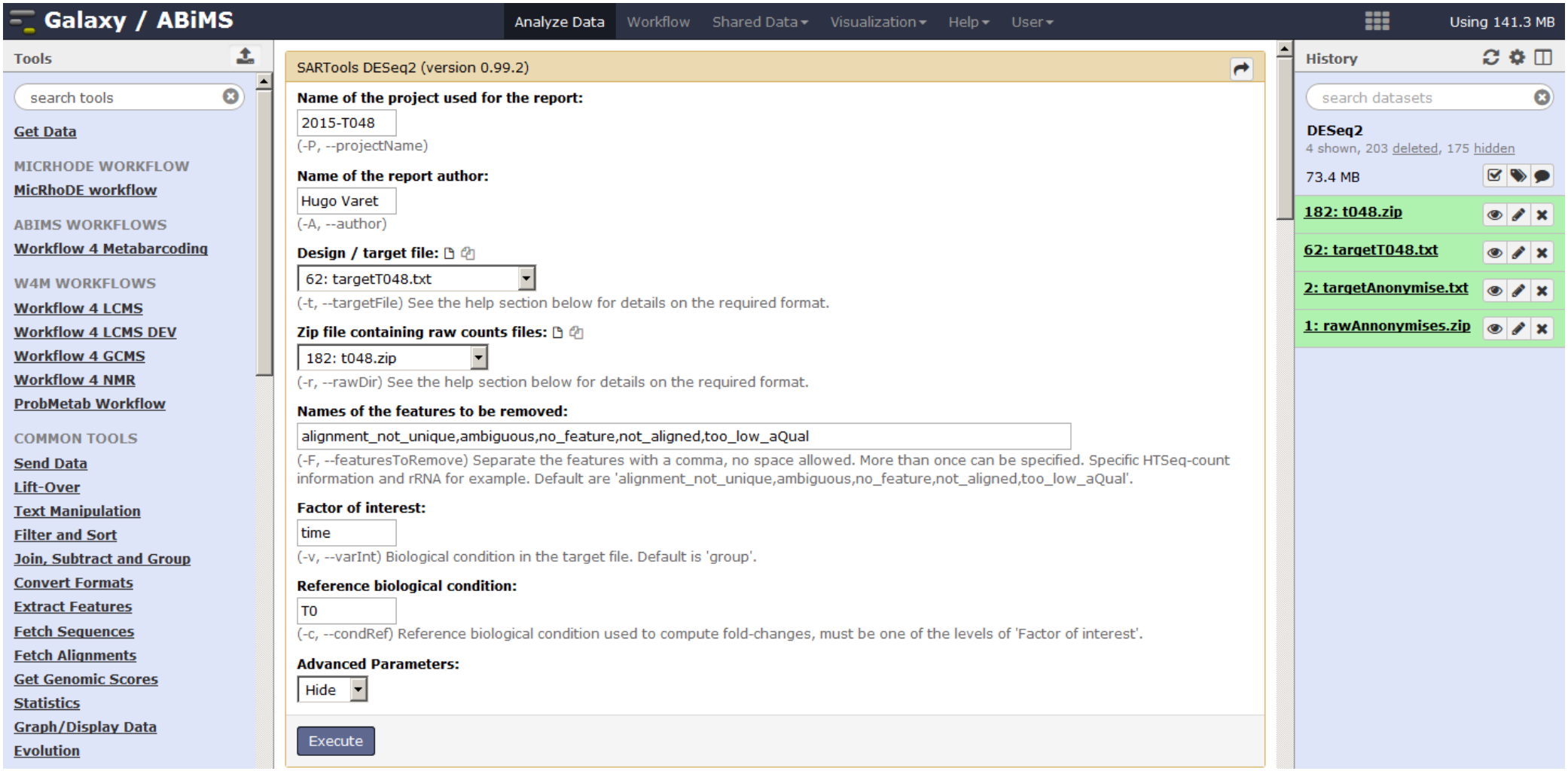
`github.com/PF2-pasteur-fr/SARTools/`



Utilization: with 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")
```

Utilization: with Galaxy



Output: HTML report

Statistical report of project T048: pairwise comparison(s) of conditions with DESeq2

Author: Hugo Varet

Date: 2016-07-27

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.

Table of contents

1 Introduction

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 3.3.1 (2016-06-21), 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
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: Biobase 2.32.0, BiocGenerics 0.18.0, DESeq2 1.12.3, edgeR 3.14.0, GenomeInfoDb 1.8.3, GenomicRanges 1.24.2, IRanges 2.6.1, limma 3.28.17, S4Vectors 0.10.2, SARTools 1.3.2, SummarizedExperiment 1.2.3, xtable 1.8-2
- Loaded via a namespace (and not attached): acepack 1.3-3.3, annotate 1.50.0, AnnotationDbi 1.34.4, BiocParallel 1.6.3, chron 2.3-47, cluster 2.0.4, codetools 0.2-14, colorspace 1.2-6, data.table 1.9.6, DBI 0.4-1, digest 0.6.9, evaluate 0.9, foreign 0.8-66, formatR 1.4, Formula 1.2-1, genefilter 1.54.2, geneplotter 1.50.0, ggplot2 2.1.0, grid 3.3.1, gridExtra 2.2.1, gtable 0.2.0, Hmisc 3.17-4, knitr 1.13, lattice 0.20-33, latticeExtra 0.6-28, locfit 1.5-9.1, magrittr 1.5, Matrix 1.2-6, munsell 0.4.3, nnet 7.3-12, plyr 1.8.4, RColorBrewer 1.1-2, Rcpp 0.12.6, rpart 4.1-10, RSQLite 1.0.0, scales 0.4.0, splines 3.3.1, stringi 1.1.1, stringr 1.0.0, survival 2.39-4, tools 3.3.1, XML 3.98-1.4, XVector 0.12.1, zlibbioc 1.18.0

Parameter values used for this analysis are:

- workDir: .
- projectName: T048
- author: Hugo Varet
- targetFile: target.txt
- rawDir: raw
- featuresToRemove: alignment_not_unique, ambiguous, no_feature, not_aligned, too_low_aQual

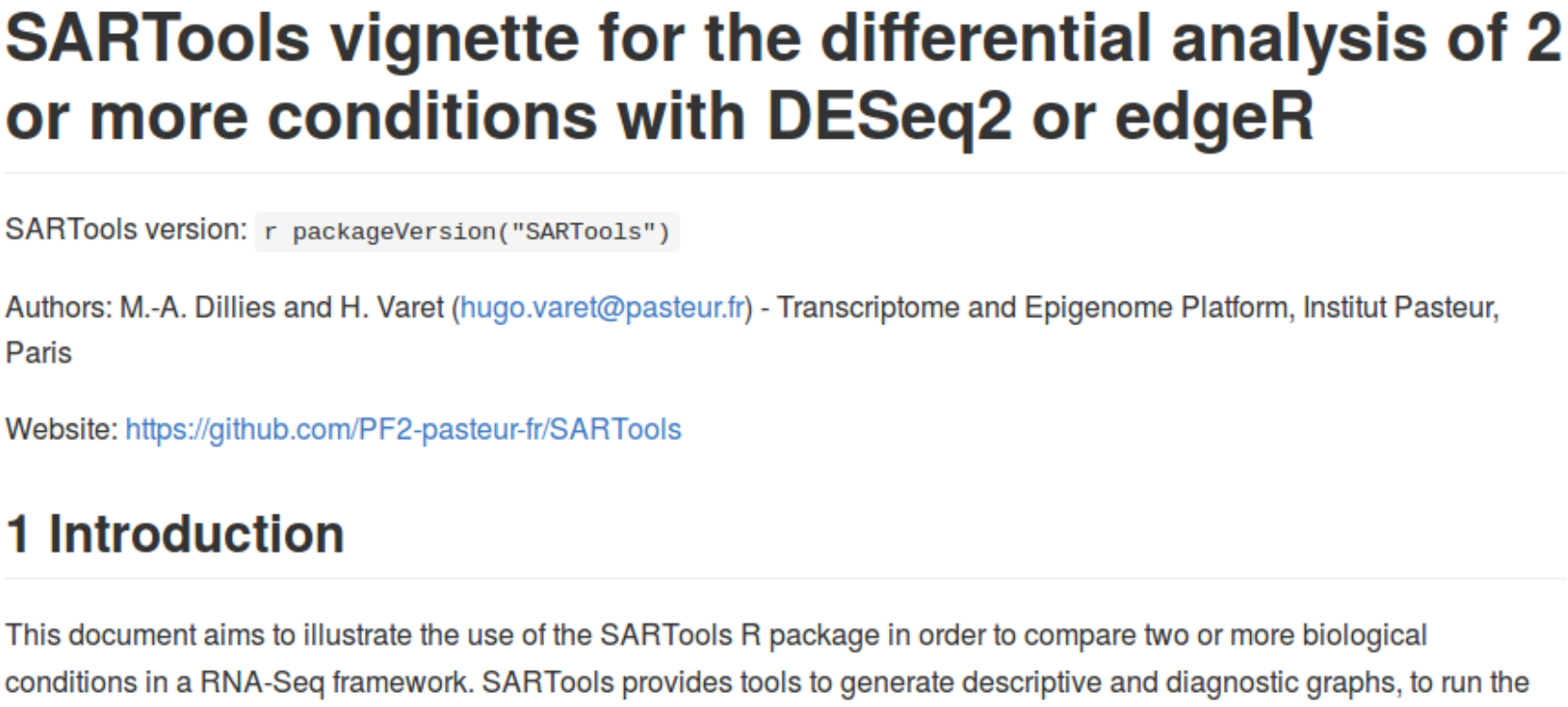
Output: lists of differentially expressed genes

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

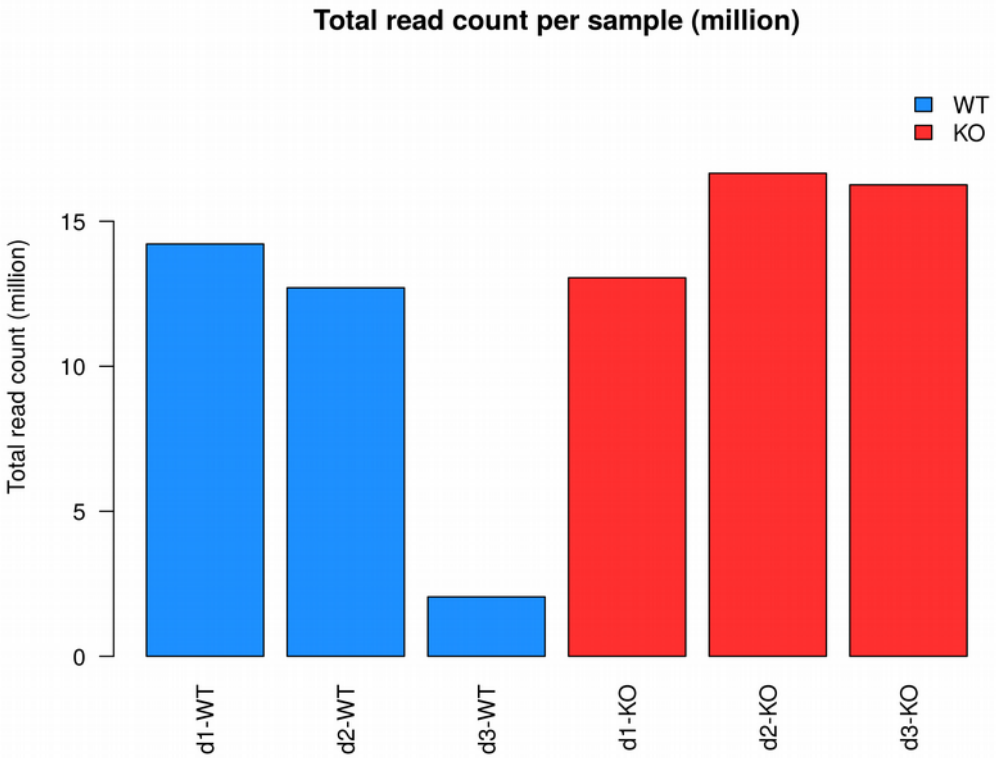
Website: https://github.com/PF2-pasteur-fr/SARTools

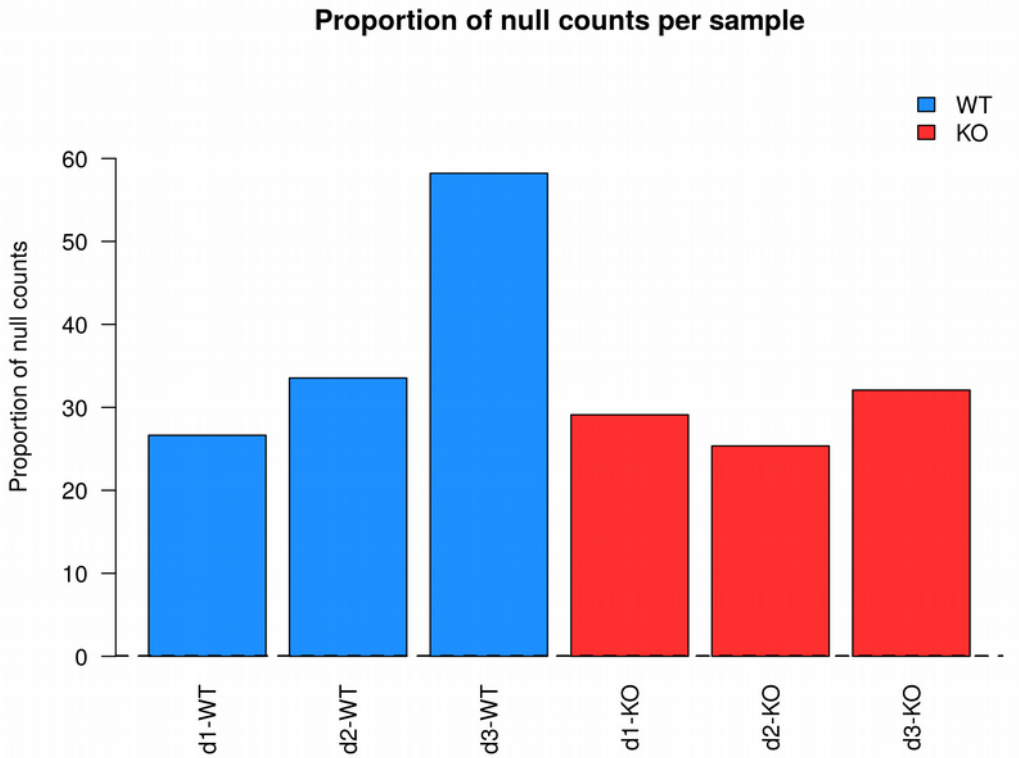
1 Introduction

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

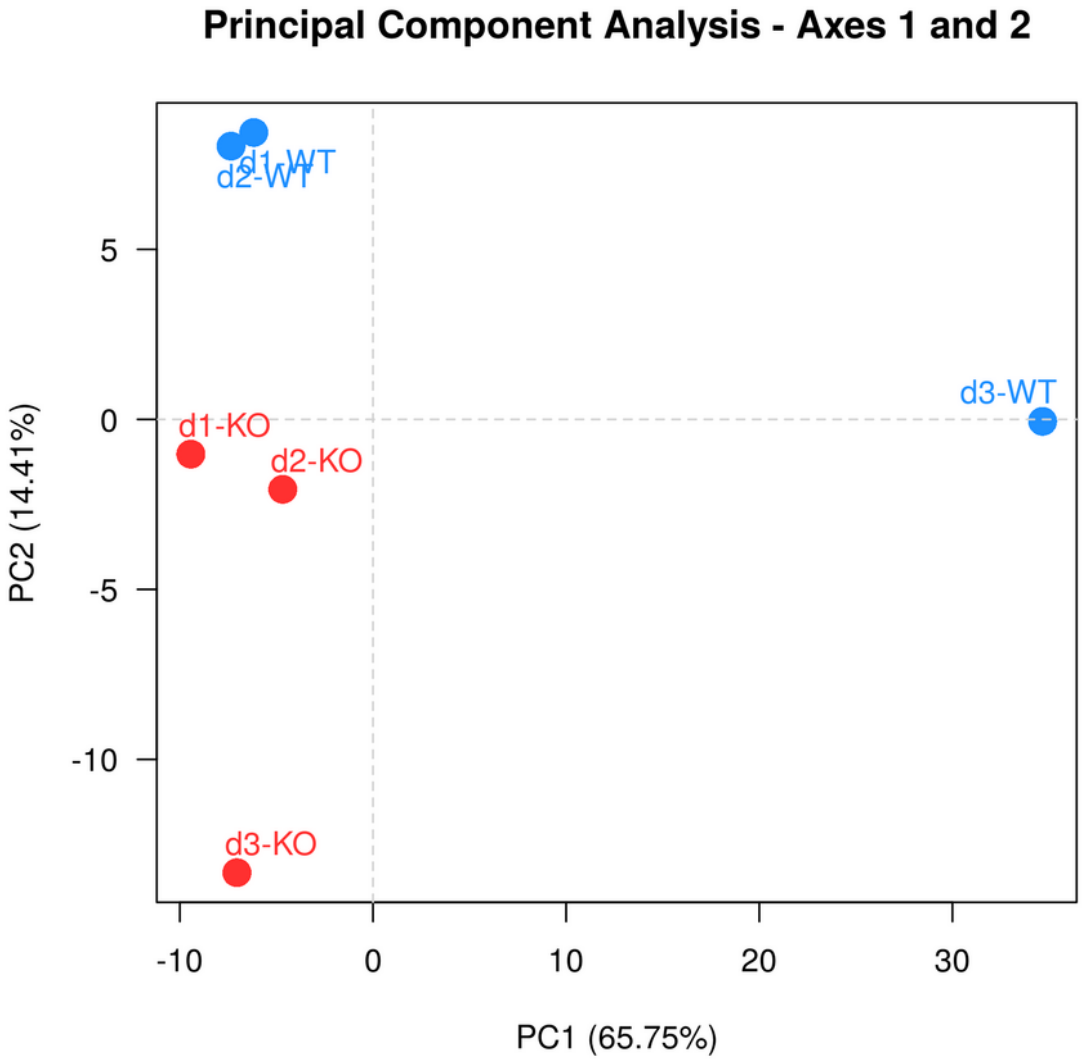
- Installation
- Input files
- Definition of the parameters
- Potential issues: technical problems, inversion of samples, batch effects, outliers...

Potential issue: detecting outliers

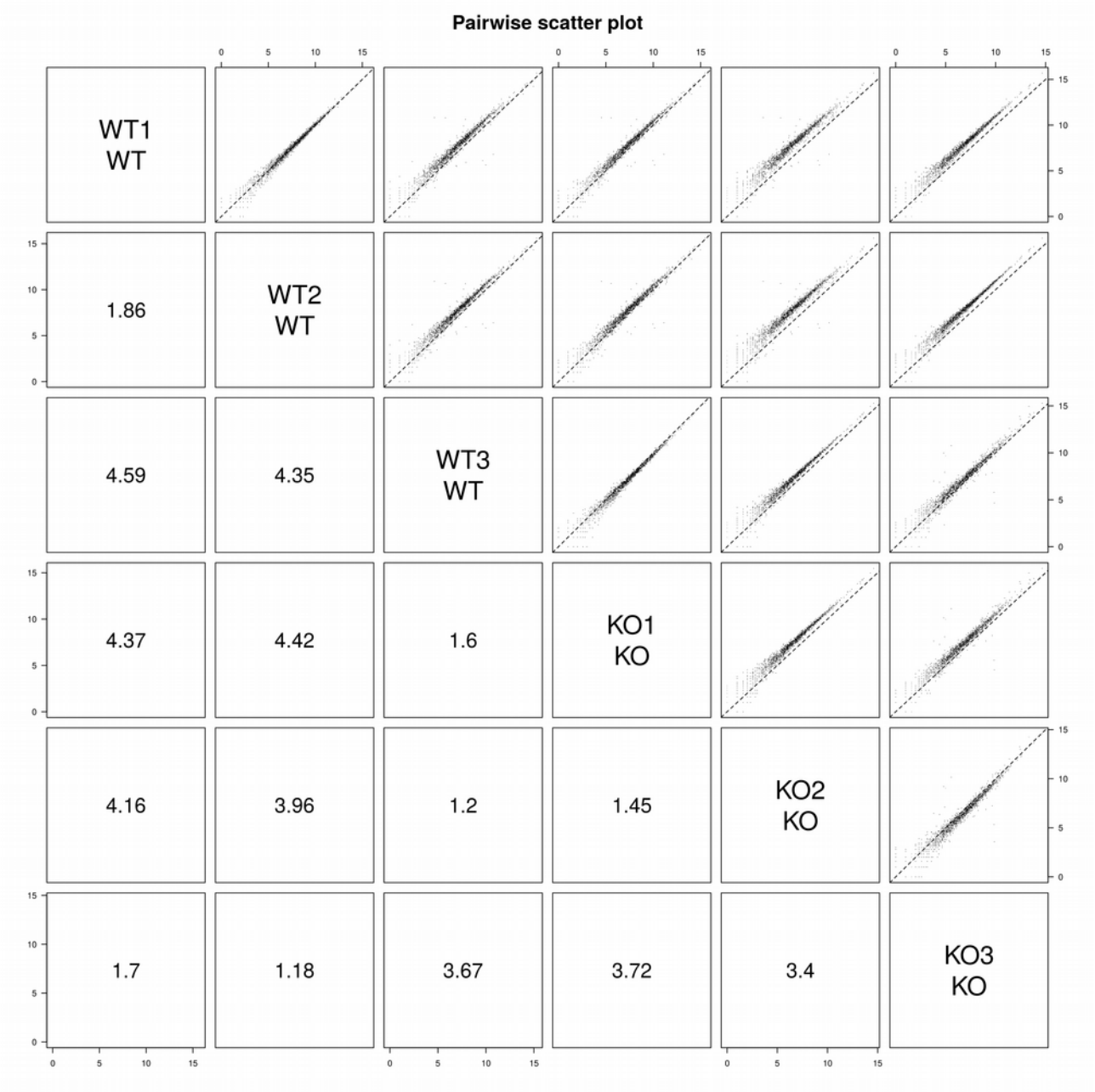




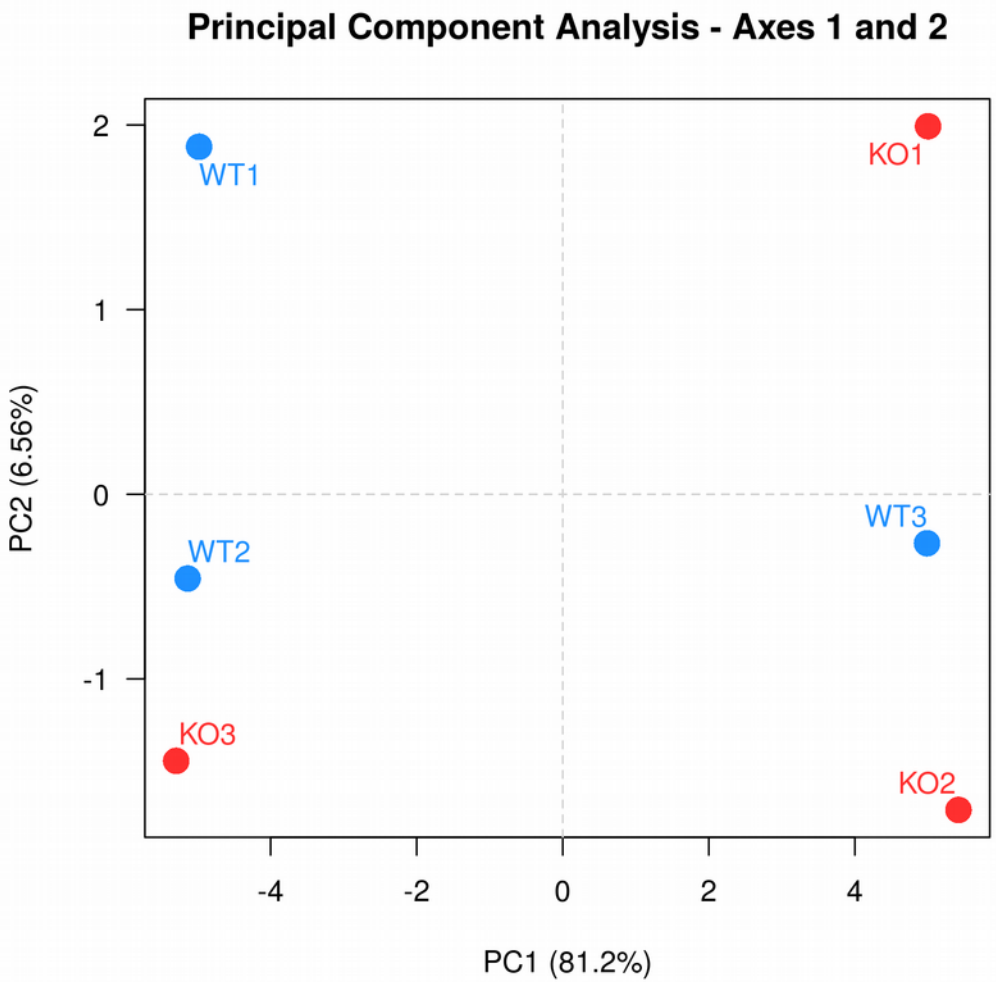
Potential issue: detecting outliers

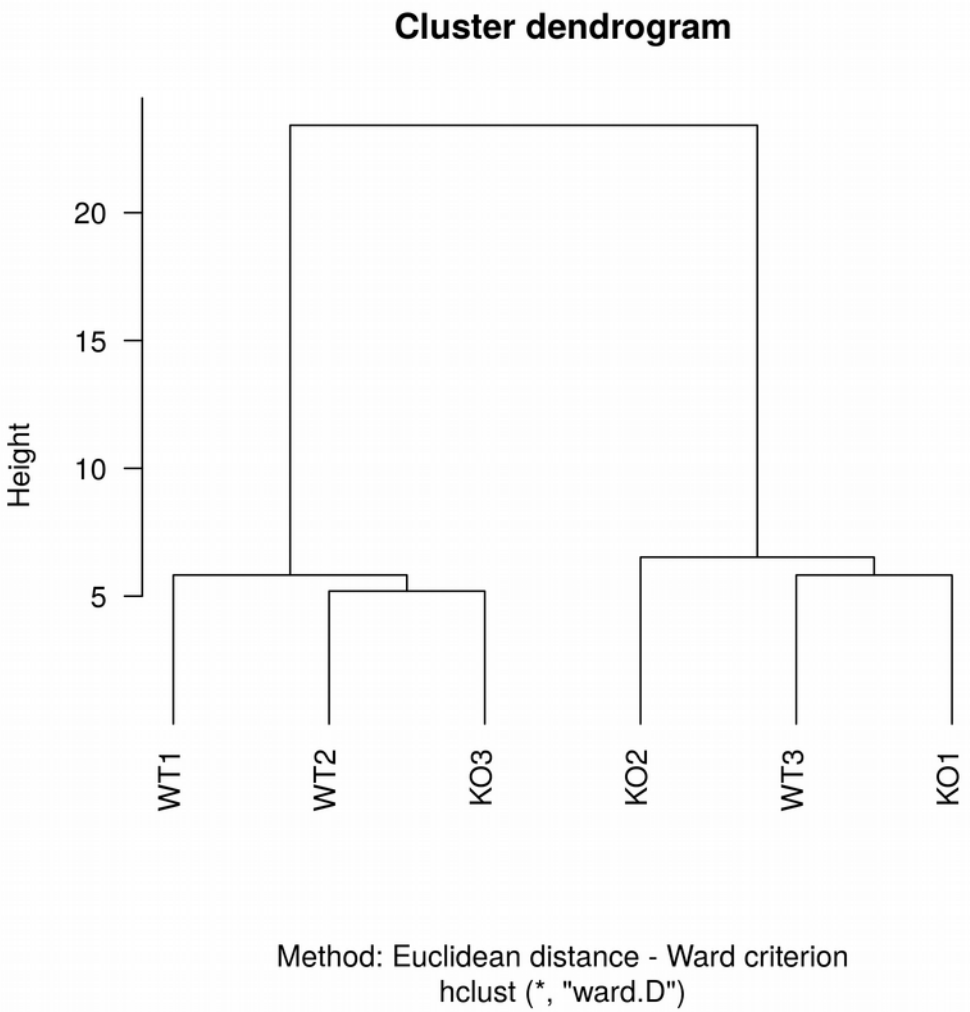


Potential issue: inversion of samples

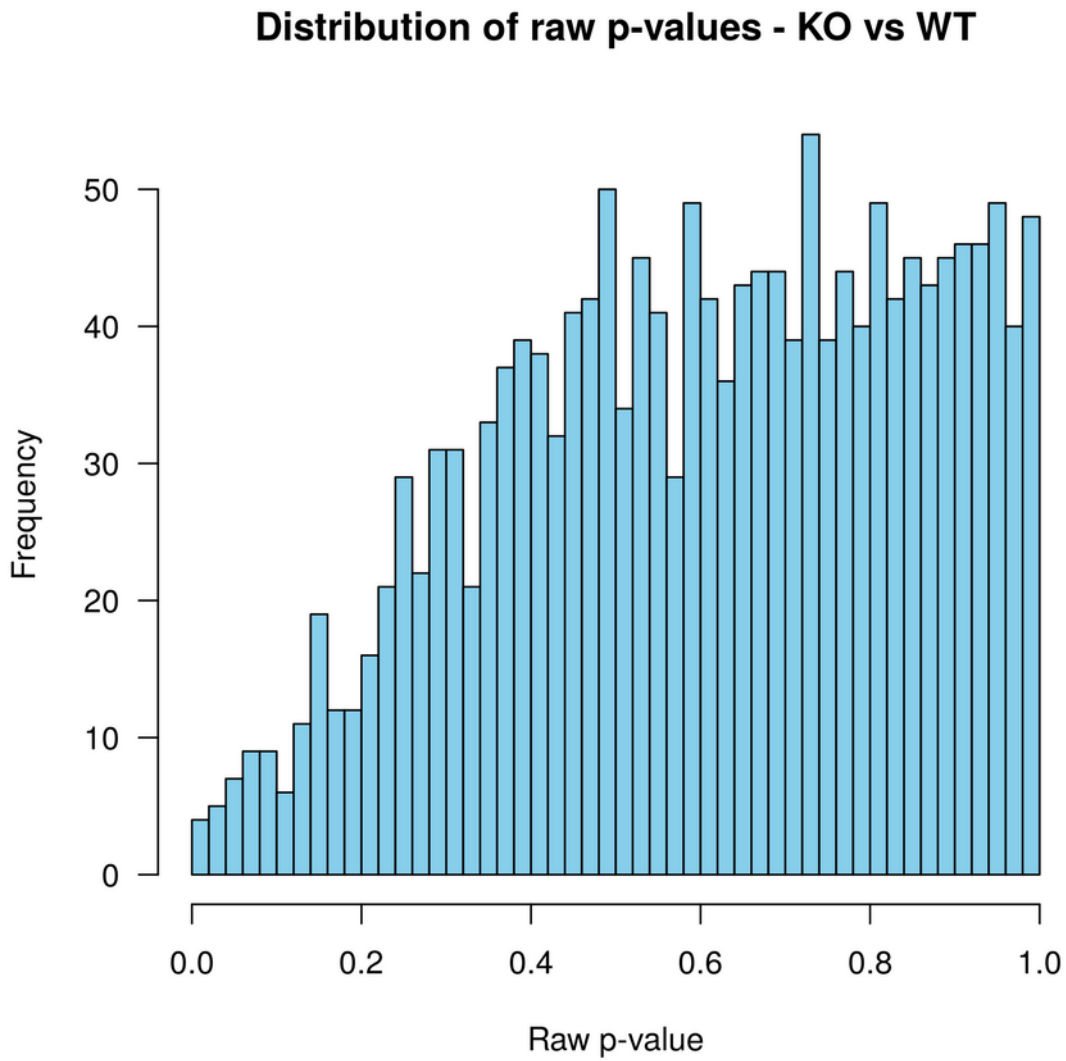


Potential issue: inversion of samples

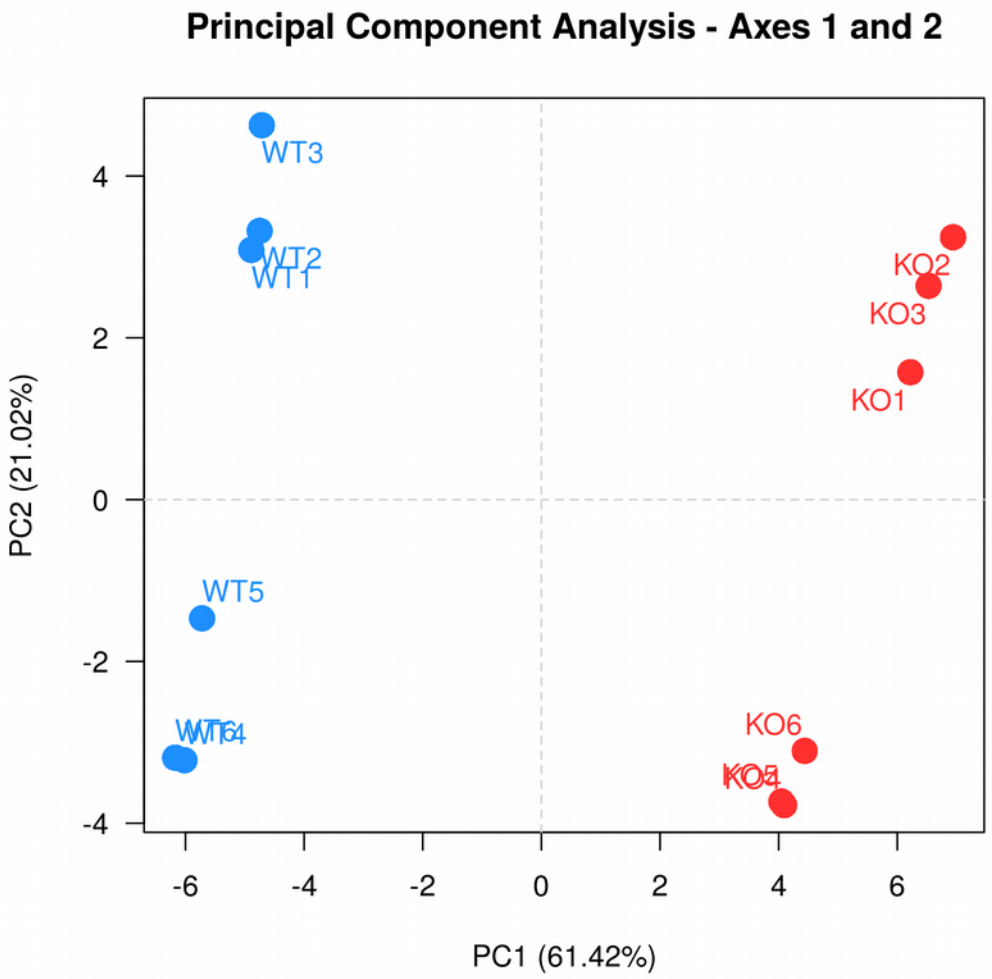


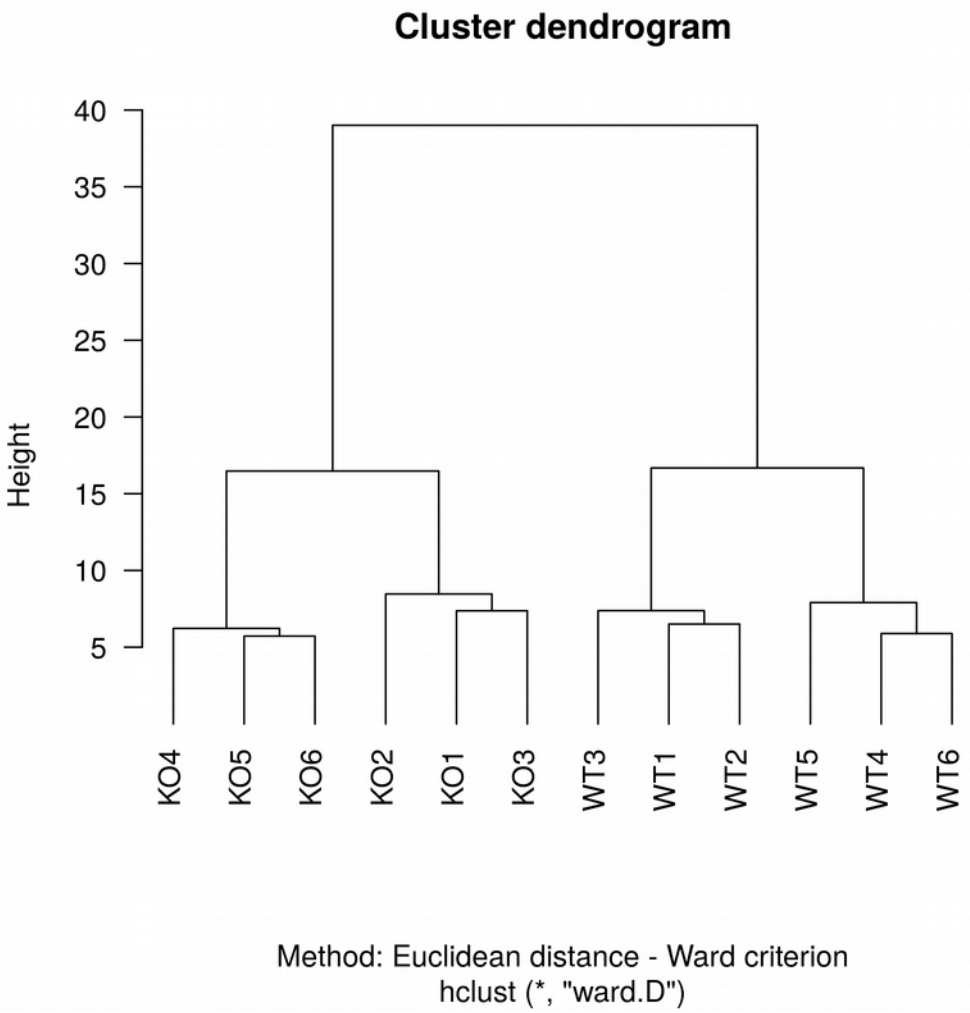


Potential issue: inversion of samples



Potential issue: batch effect





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!

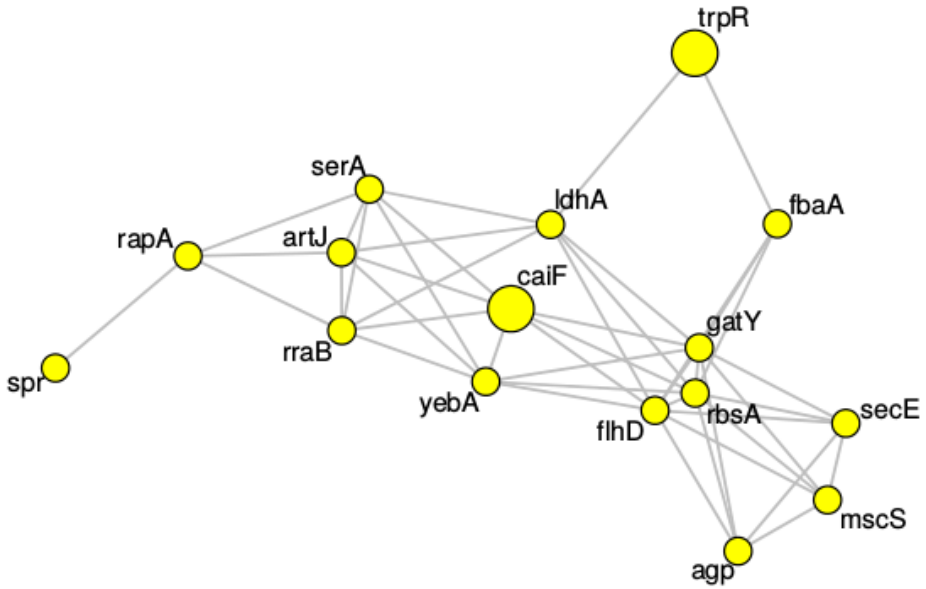
Interpreting the lists of DE genes

Gene Ontology (GO) terms:

Statistical tests based on the hyper-geometric law: are there over-represented categories among the differentially expressed genes?

Genes networks:

Link genes which interact from a statistical point of view: need a large number of replicates.



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

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.