Bioinformatics Week!

RNA-seq data analysis

...

Tamara Perteghella,
Silvia González López

Computational Biology of RNA Processing Lab
(Roderic Guigó)

25th October 2024

CRG
Centre
for Genomic
Regulation

Outline

Outline

- **Background**
  - RNA-seq experimental protocols
  - Short-read RNA-seq data processing
  - Reference gene annotation

- **RNA-seq data analysis**
  - Sample clustering based on gene expression
  - Differential gene expression
  - Functional enrichment

Background

What, How and why RNAseq?

- Set of techniques that employ sequencing to **measure the presence and quantity** of RNA molecules in a biological sample.

- Different applications:
  - **Characterising transcriptional landscape** of cells and their function.
  - **Dissect transcriptional complexity** (e.g., alternative splicing, start and termination sites).
  - **Annotate novel elements**.

RNA-seq experiment

Library prep

Cellular localisation

- Whole cell
- Nucleus
- Cytoplasm

Technology of choice

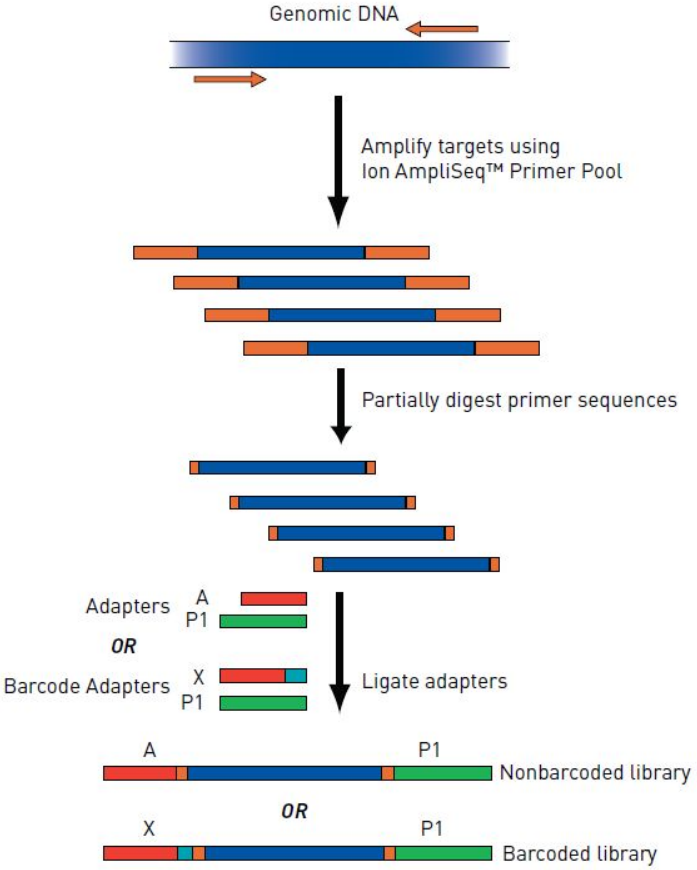
Single or paired sequencing

RNA Purification

- Ribo$^-$
- PolyA$^+$
- Total RNA

Strandedness

Size selection



RNA-seq experiment

Library prep

Sequencing

Short read

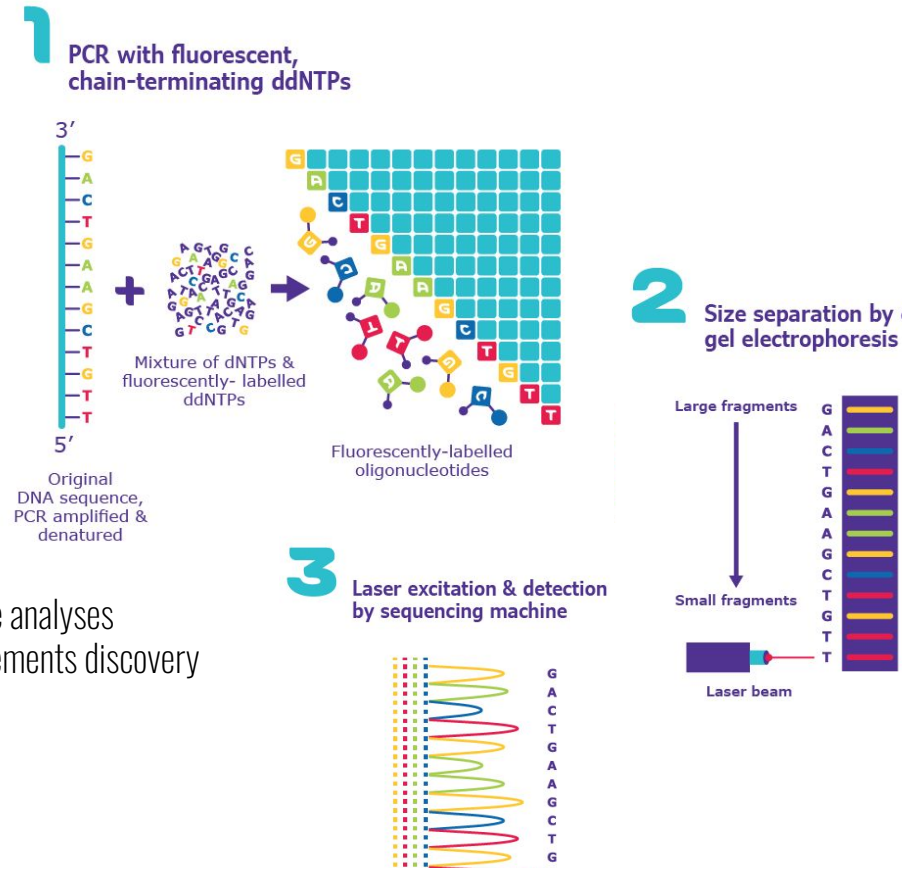
- Illumina

Long-read

- Nanopore
- PacBio

Depth

- >30M reads for simple analyses
- >100M reads novel elements discovery



RNA-seq experiment

Library prep

Sequencing

Analysis

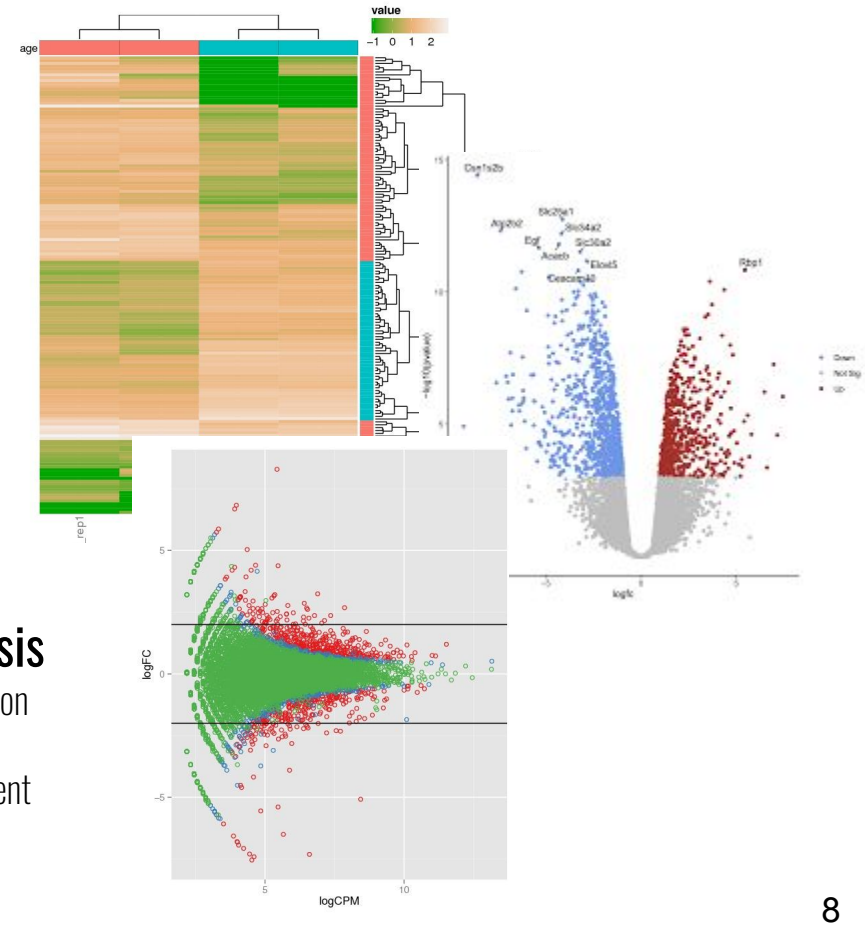
Read data processing

- Mapping
- Quantification

Quality evaluation

Downstream analysis

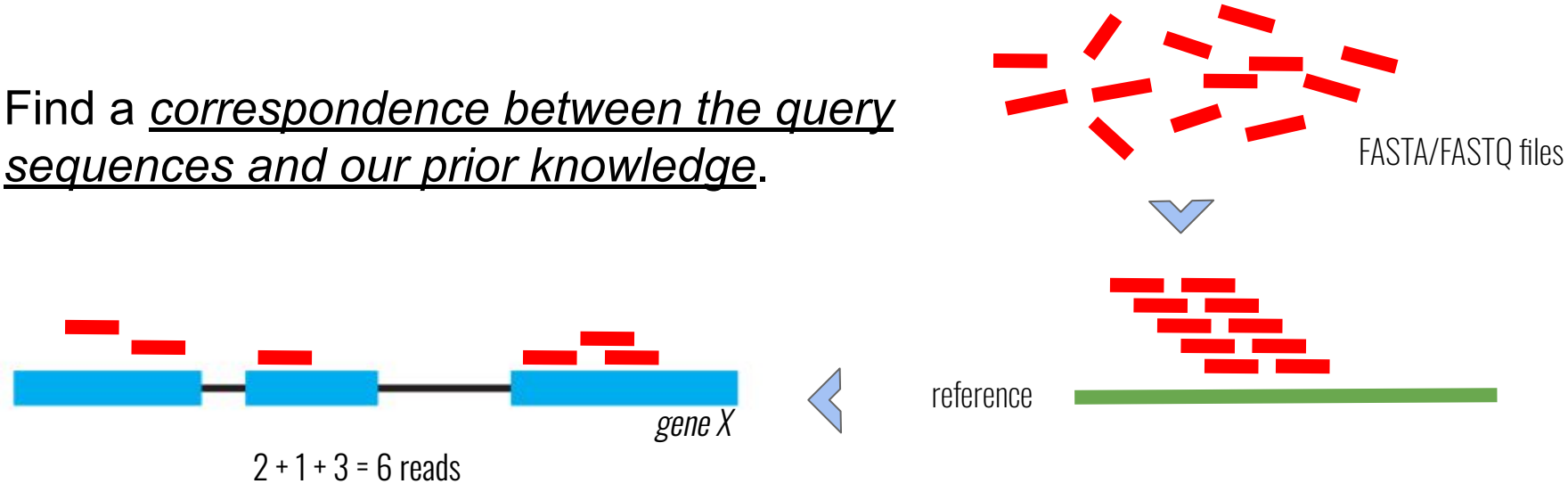
- Differential Expression
- Clustering
- Functional assessment



Mapping strategy

Mapping and Quantification

Find a *correspondence between the query sequences and our prior knowledge*.



This will then be used to **quantify the expression of a gene**, upon a simple idea to count the RNA-seq reads that fall within the exons of this gene.

Reference gene annotation

- For a *given species* and associated genome assembly, the **reference gene annotation** is the collection of all genes known for this species.

- Various completion stages (high-quality annotations are those of human, and main model organisms; e.g., mouse, D. melanogaster, C. elegans or yeast).

- The choice of annotation is extremely important as it will serve as **ground truth** against which the RNA-seq data will be compared.

https://www.gencodegenes.org/

GENCODE annotation

HUMAN

GENCODE 46 (May 2024)

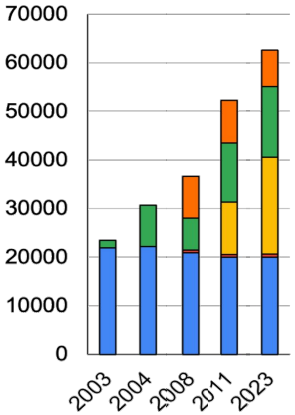
MOUSE

GENCODE M35 (May 2024)

International consortium which goal is to classify all gene features in the human and mouse genomes with high accuracy **<u>based on biological evidence</u>**.

- **Broad gene categories:**
- **GFF/GTF file format;** several features.

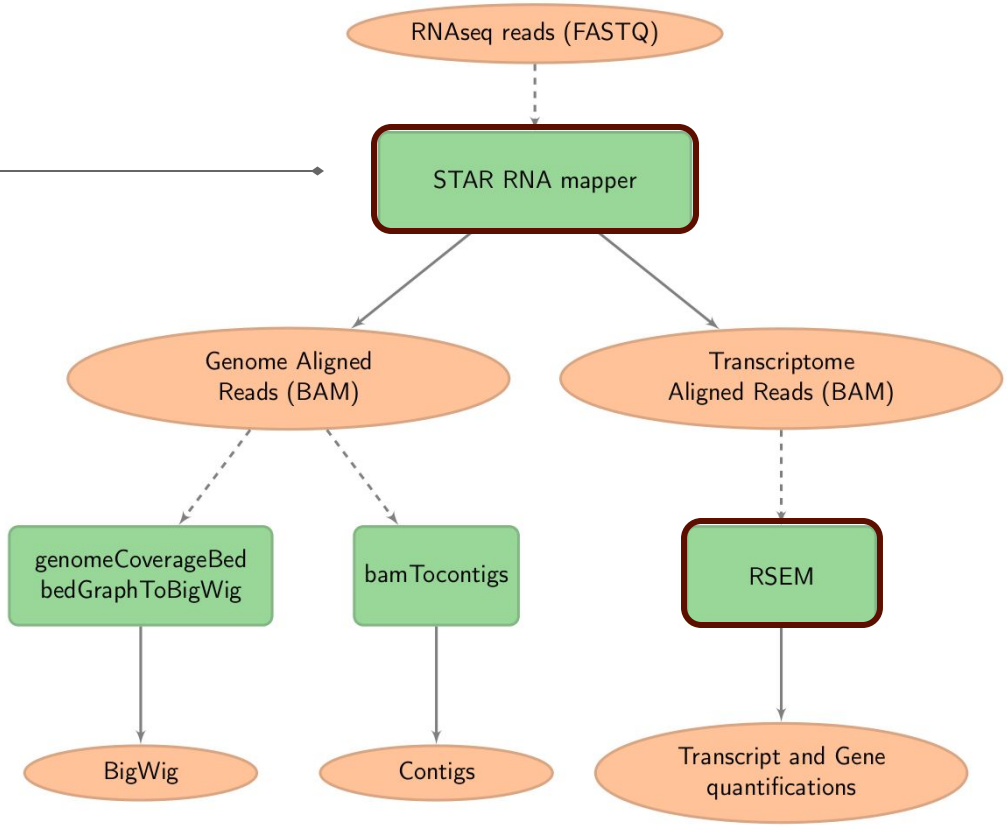
Human gene count throughout years



protein_coding IG/TR_segment lncRNA pseudogene small_RNA

Mapping and Quantification

1. Mapping
2. Quantification

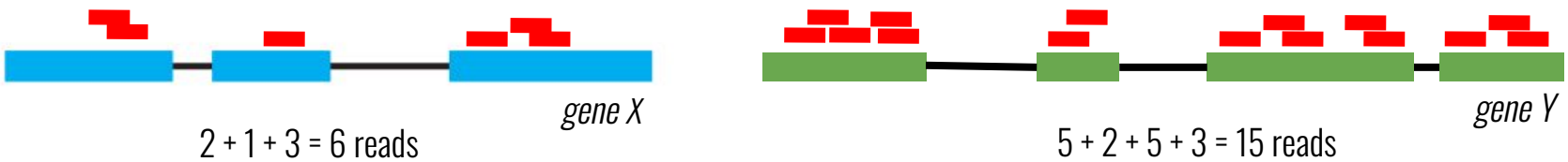


Emilio Palumbo

grape-nf: An automated RNA-seq pipeline using Nextflow

Gene expression quantification

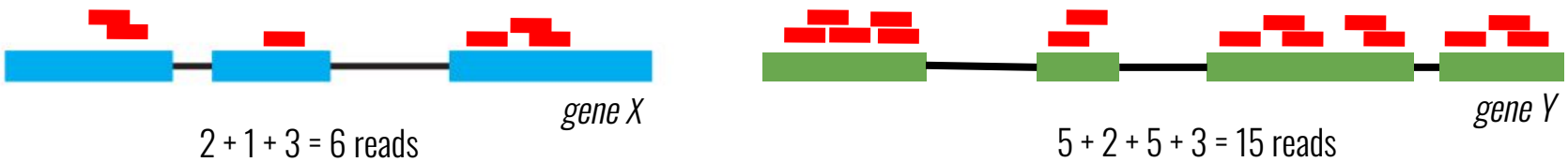
- Longer genes will get more reads than small genes



15 >> 6 → gene Y >> gene X

Gene expression quantification

- Longer genes will get more reads than small genes



15 >> 6 → gene Y >> gene X

REJECTED

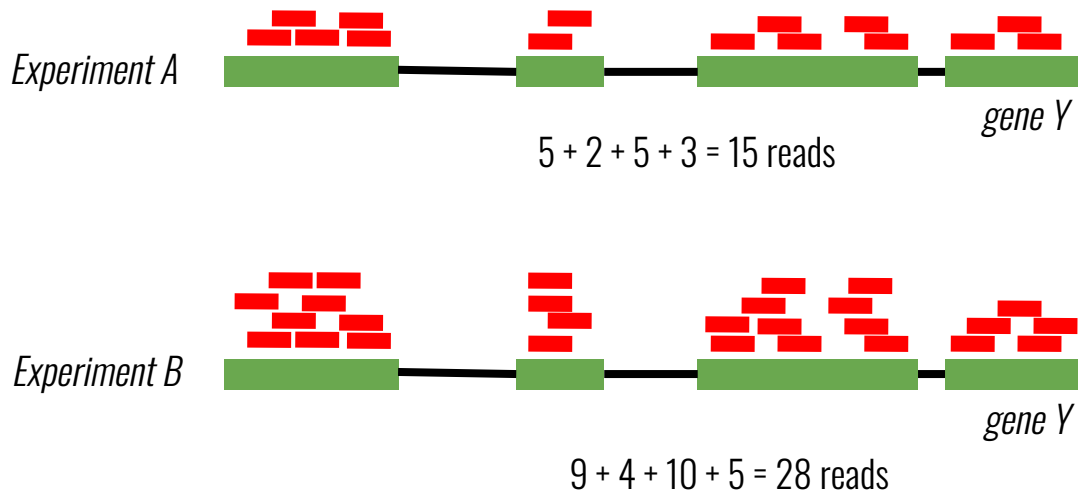
Gene expression quantification

- Higher depth of sequencing means we get higher number of mapped reads

28 >> 15

↓

gene Y_B >> gene Y_A

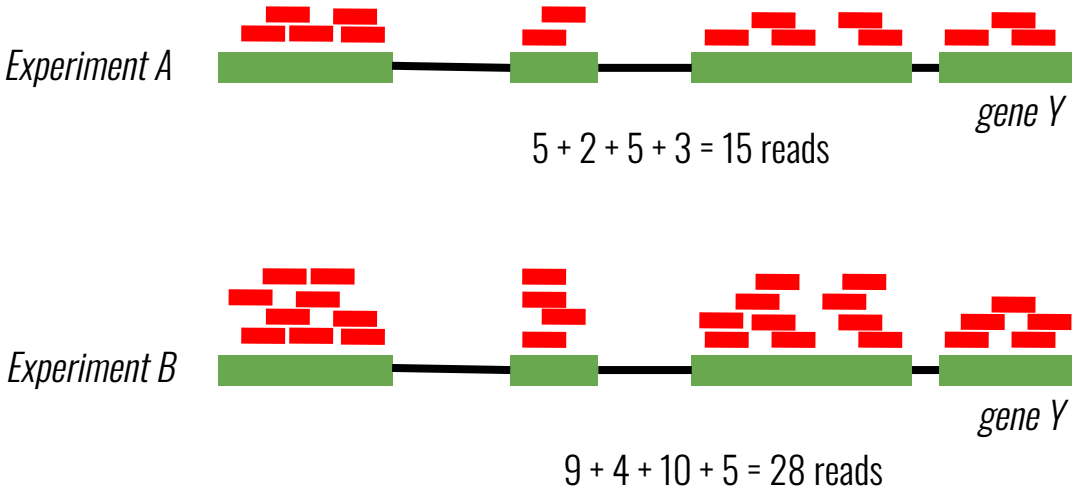


Gene expression quantification

- Higher depth of sequencing means we get higher number of mapped reads

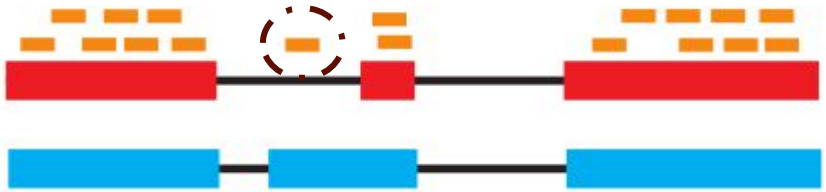
28 >> 15

gene Y_B >> gene Y_A



Transcripts expression quantification

Gene expression is indeed quite easy to compute, however estimating the expression of individual transcripts of each gene is a difficult problem:



Only one of this many reads is unequivocally assignable to an isoform!

<u>Read deconvolution</u> which is at the base of transcript quantification is possible via several methods, like RSEM, and Kallisto for example.

Gene expression quantification

- **RPKM** (or FPKM in case of paired-end experiments*)

Read (Fragment) Per Kilobase of exon model per Million mapped reads; is the standardized read count of a gene in an experiment by

i) the length of the gene and
ii) the total number of mapped reads in the experiment *(Mortazavi, 2008)*.

$$RPKM = \frac{mapped\ reads * 10^9}{Tot\ mapped\ reads * Length}$$

However, it assumes that the absolute amount of total RNA in each cell is similar across different cell types or experimental conditions, which is not always the case *(Loven, 2012)*.

- **TPM**

Transcripts per Million *(Li, 2010)*

$$\mathrm{TPM} = 10^6 * \frac{\mathrm{RPKM}}{\mathrm{Sum(RPKM)}}$$

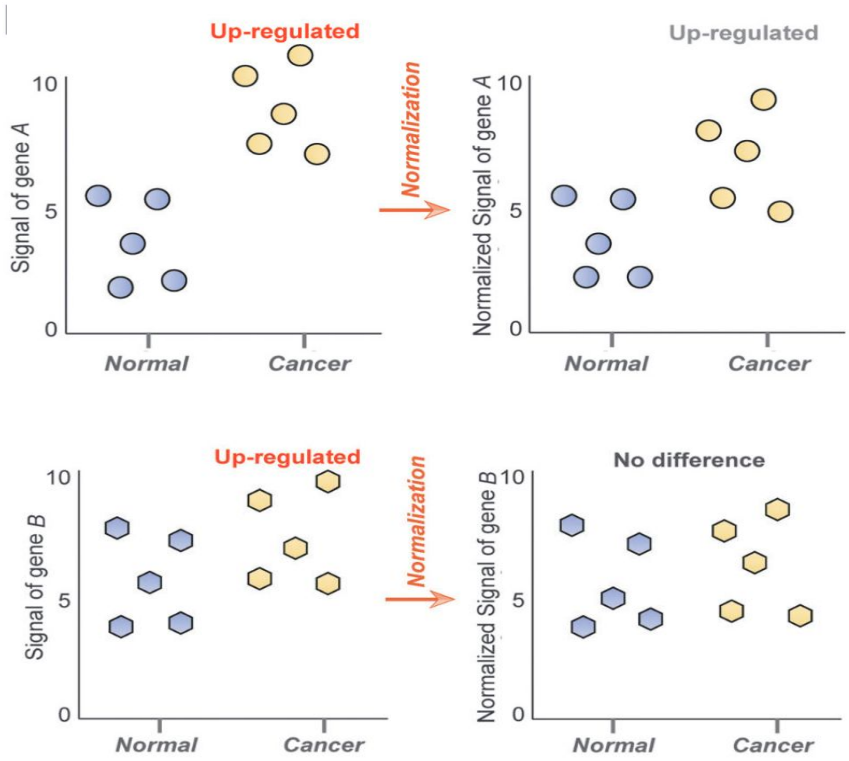
* Paired-end: 2 reads = 1 fragment.

Transformation methods

- <u>Scaling factors</u>,

Essentially a factor by which the raw counts for each sample are divided to account for sequencing depth, **making them directly comparable**.

<u>Methods</u>:
quantile normalization,
trimmed mean of M-values (edgeR)
DESeq

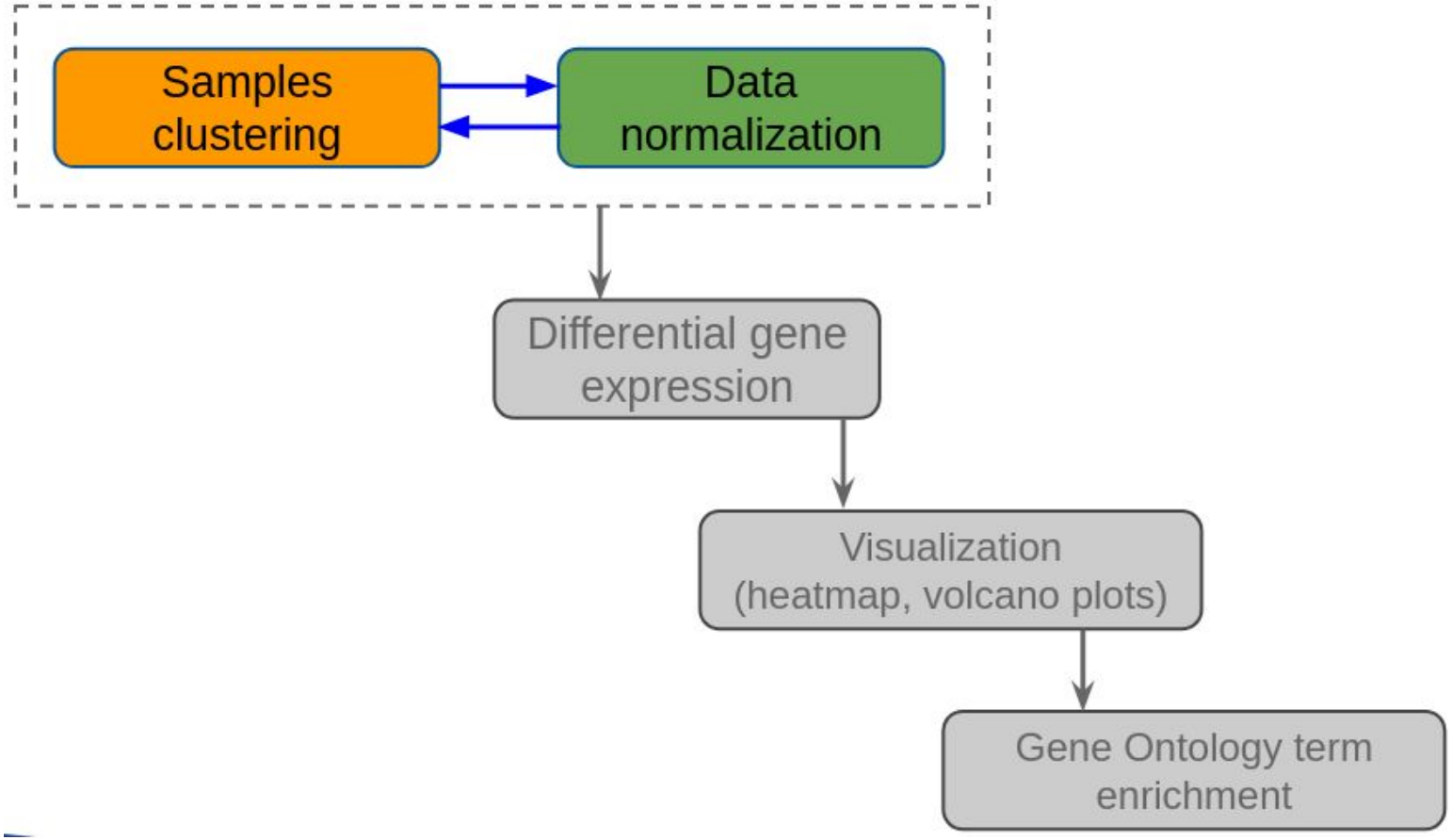


Hands-on I

- Overview
- Environment set-up
- Data preparation

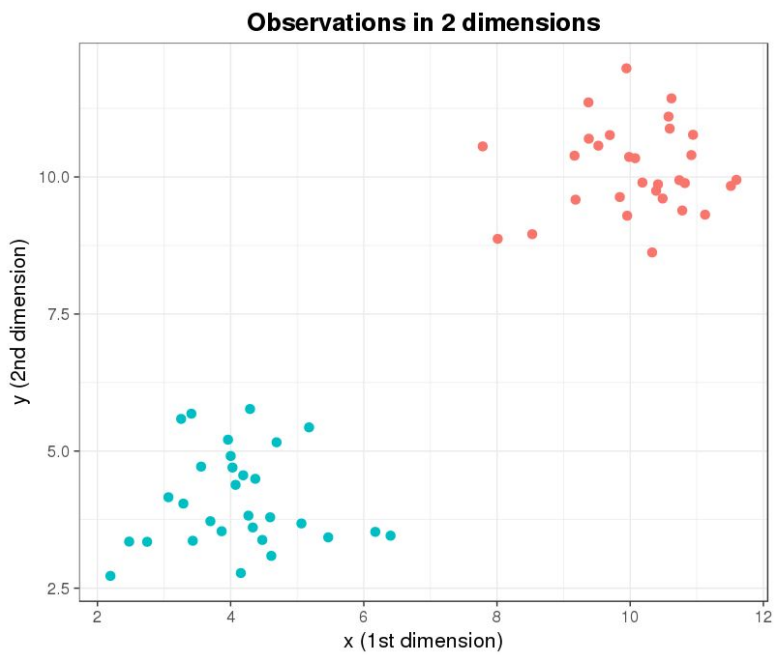
RNA-seq analysis

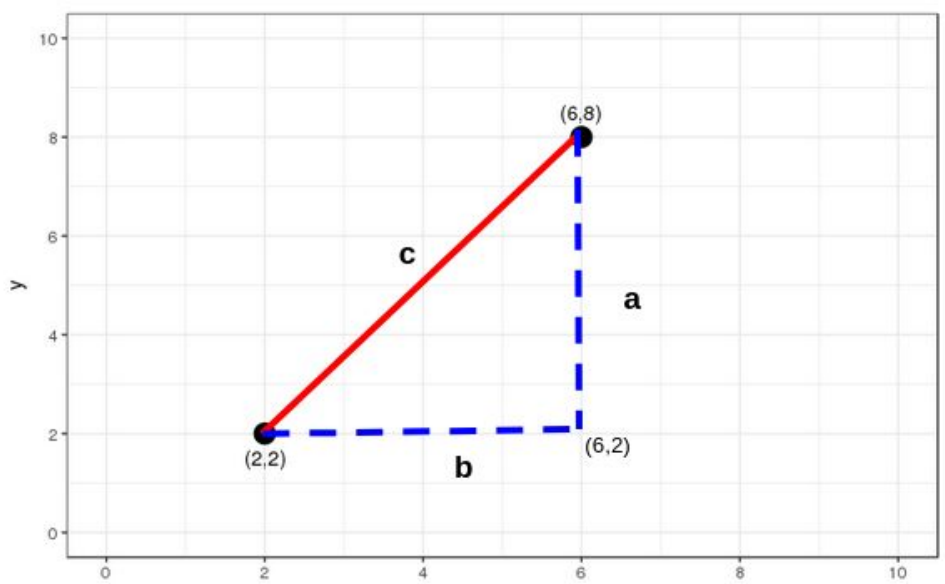
Analysis pipeline



A geometrical insight

x	y
10.82	9.89
3.26	5.59
5.18	5.43
10.58	11.10
8.01	8.87
10.39	9.75
10.74	9.94
...	...

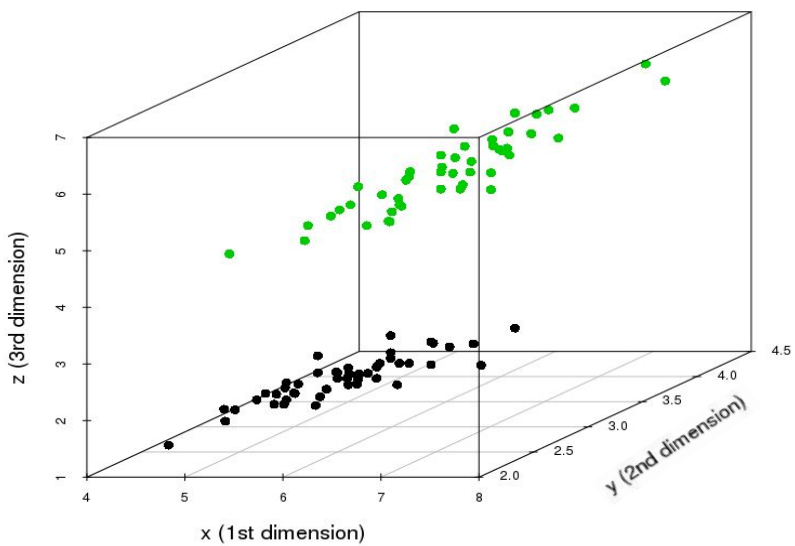


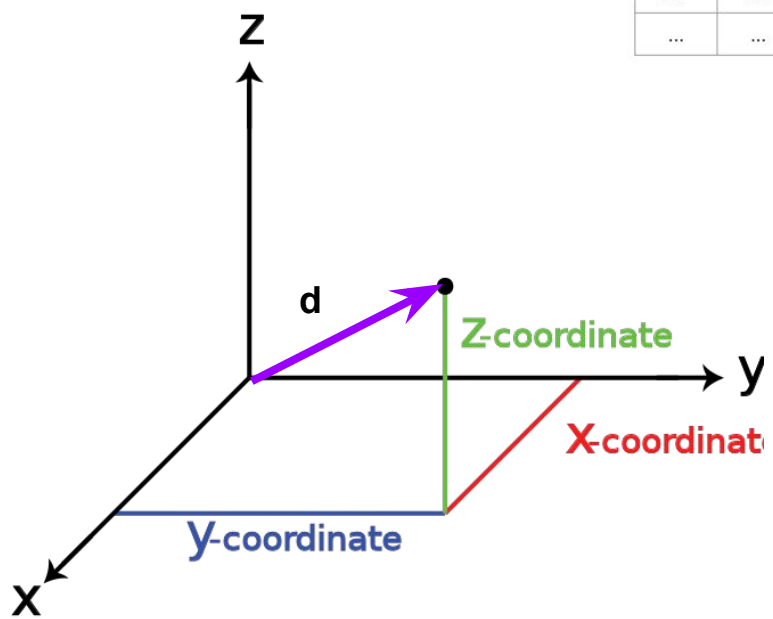


$$c^2 = a^2 + b^2 \rightarrow c = \sqrt{a^2 + b^2} \rightarrow c = \sqrt{(8-2)^2 + (6-2)^2}$$

A geometrical insight

x	y	z
6.2	2.8	4.8
5.8	2.7	5.1
5.1	3.8	1.6
6.7	2.5	5.8
6.5	3.0	5.2
5.4	3.7	1.5
5.1	3.3	1.7
6.7	3.0	5.2
...	...	...





$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

A geometrical insight

samples

Genes

	A_1	A_2	B_1	B_2
G1	1.0	1.0	6.0	5.0
G2	2.0	1.5	5.0	5.1
G3	1.0	1.1	1.0	1.0
G4	3.5	3.0	4.0	4.2
G5	4.0	3.9	5.0	5.3

A geometrical insight

Euclidean distance: $\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$

samples

Genes

	A_1	A_2	B_1	B_2
G1	1.0	1.0	6.0	5.0
G2	2.0	1.5	5.0	5.1
G3	1.0	1.1	1.0	1.0
G4	3.5	3.0	4.0	4.2
G5	4.0	3.9	5.0	5.3

D
0.0
0.5
-0.1
0.5
0.1

D^2
0.0
0.25
0.01
0.25
0.01

sum → 0.52

sqrt → 0.72

A geometrical insight

Euclidean distance: $\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$

samples

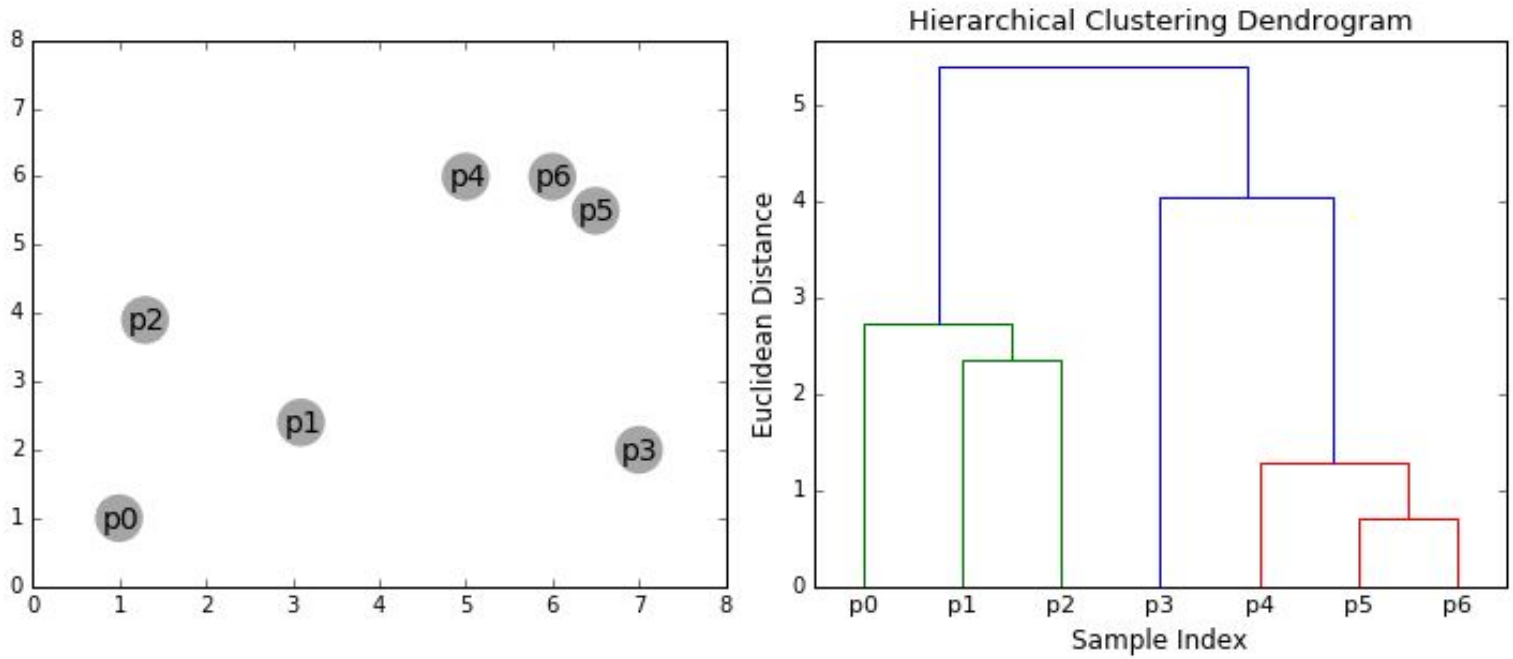
Genes

	A_1	A_2	B_1	B_2
G1	1.0	1.0	6.0	5.0
G2	2.0	1.5	5.0	5.1
G3	1.0	1.1	1.0	1.0
G4	3.5	3.0	4.0	4.2
G5	4.0	3.9	5.0	5.3

	A_1	A_2	B_1	B_2
A_1	0.0	0.72	5.9	5.27
A_2	0.72	0.0	6.28	5.69
B_1	5.94	6.28	0.0	1.07
B_2	5.27	5.69	1.07	0.0

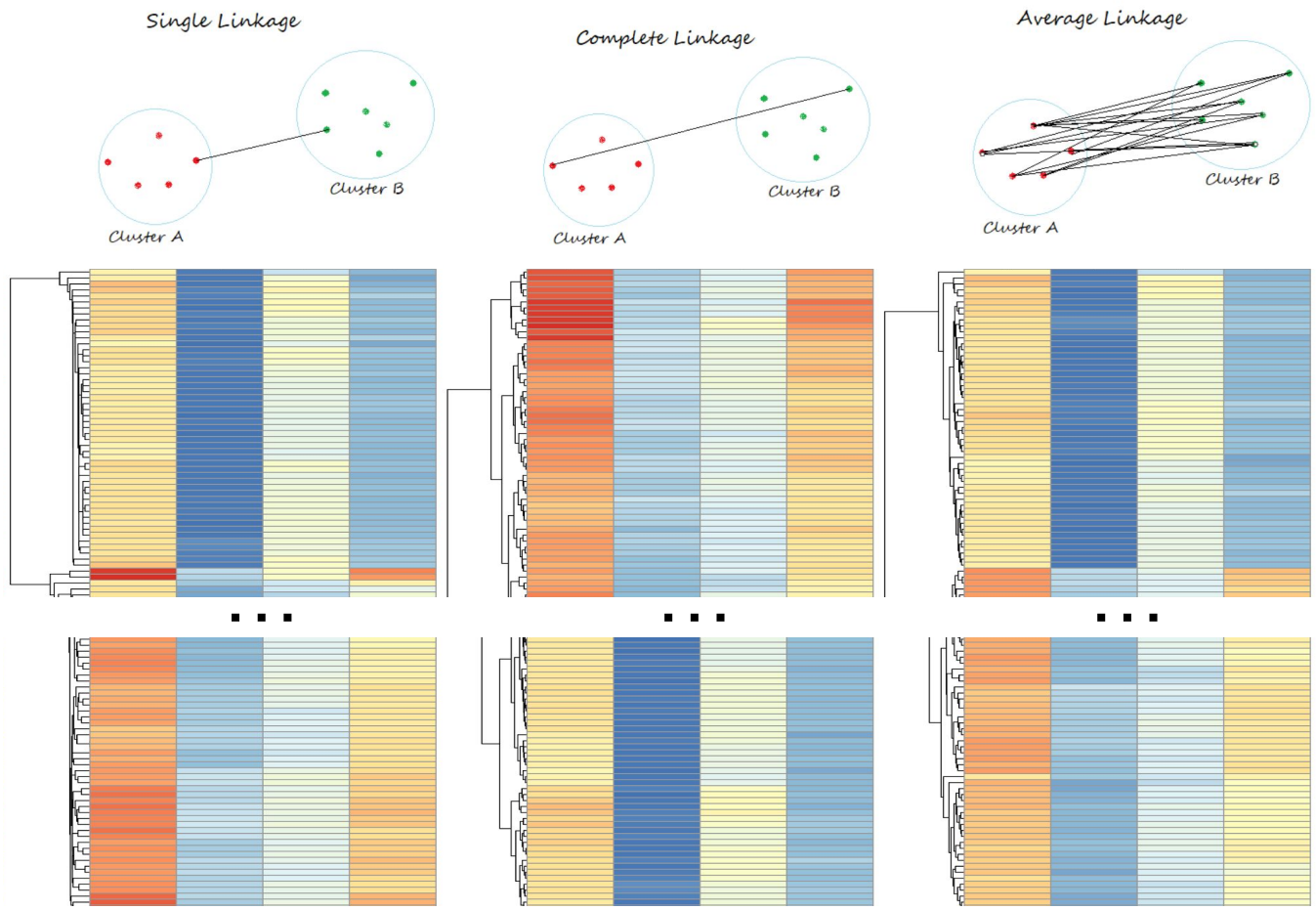
Clustering methods – Hierarchical

Starting from the distance matrix it repeatedly seeks for the two closest samples, bringing them together in a cluster.



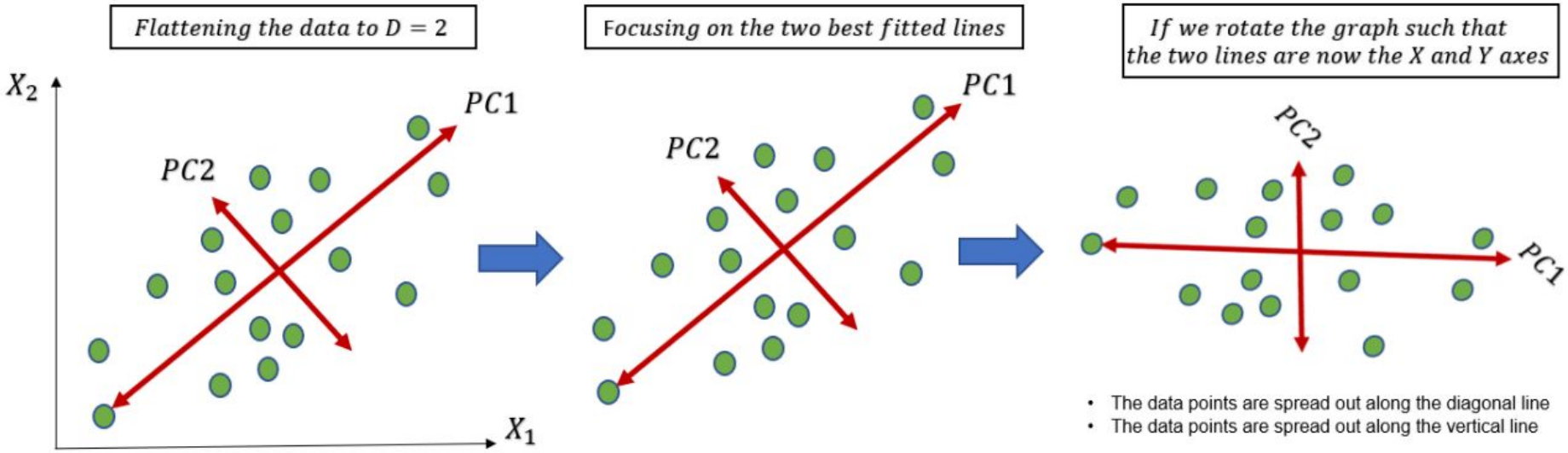
Animation from: https://dashee87.github.io

The linkage criterion



PCA

Principal Component Analysis consists in transforming the original dataset to decrease its dimensionality while preserving most of the variance, ultimately helping their interpretability while extracting the most significant features.



Hands-on II

- **Clustering and PCA**

RNA-seq analysis

Differential gene expression (DGE)

Aim: identify genes that are more (less) expressed in one condition than in the other

Batch	Sex	Sample	g_1	g_2	g_3	...
1	Male	A_1				
2	Male	A_2				
3	Male	A_3				
4	Male	A_4				
1	Female	B_1				
2	Female	B_2				
3	Female	B_3				
4	Female	B_4				

Many tools, what to chose?

- edgeR (R package)
  Robinson, McCarthy, Smyth, "EdgeR: a bioconductor package for for differential expression of digital gene expression data." Bioinformatics 26(1) (2010): 139-40.
- DESeq2 (R package)
  Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2."Genome biology 15.12 (2014).
- voom+limma (R package)
  Law, Charity W., et al. "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts." Genome Biol 15.2 (2014): R29.

Characteristic	edgeR	DESeq2	limma
Data Type	Count data (e.g., RNA-Seq)	Count data (RNA-Seq)	Count or continuous data
Normalization Method	TMM (Trimmed Mean of M-values)	Median of Ratios	Voom (variance modeling)
Data Size	Small to medium datasets	Small to medium datasets	Small to large datasets
Statistical Model	Negative binomial	Negative binomial	Linear models
Assumptions	Over-dispersed counts, biological variation	Over-dispersed counts, biological variation	Homoscedasticity, continuous data

More than 90% of the genes detected in each group were overlapped across these methods

Basics of DGE

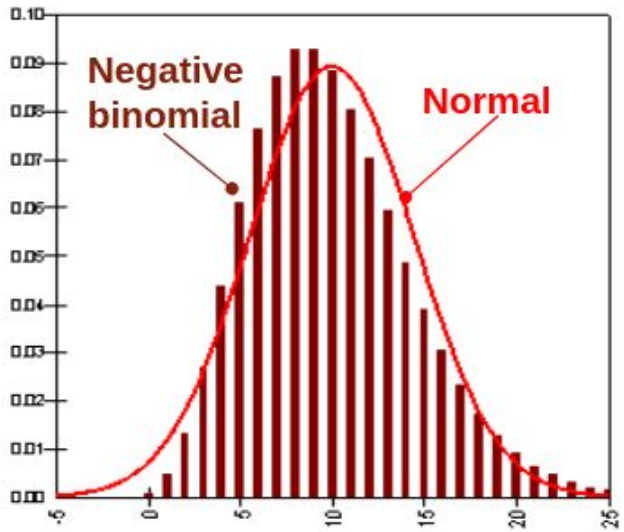
We will use **edgeR** to perform these steps.

Normalization

↓

Fit a model to the data per gene

- Data (read counts) **<u>discrete</u>** and **<u>positive</u>**
- **Negative binomial** is the most suitable choice

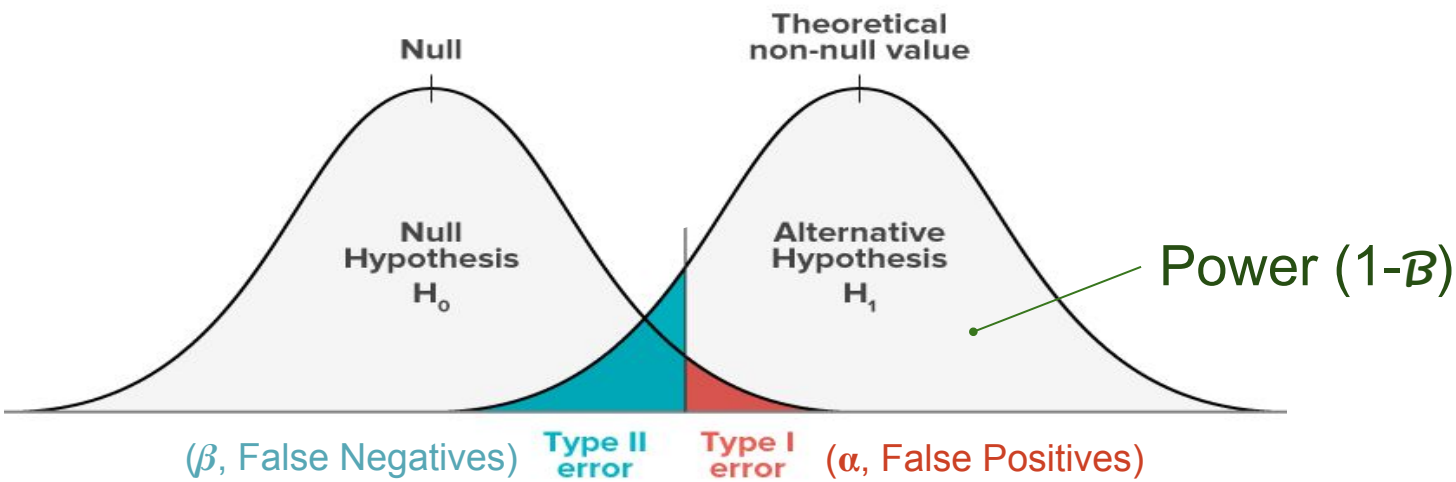


We need to estimate the **mean** and **variance** of the fitted distribution

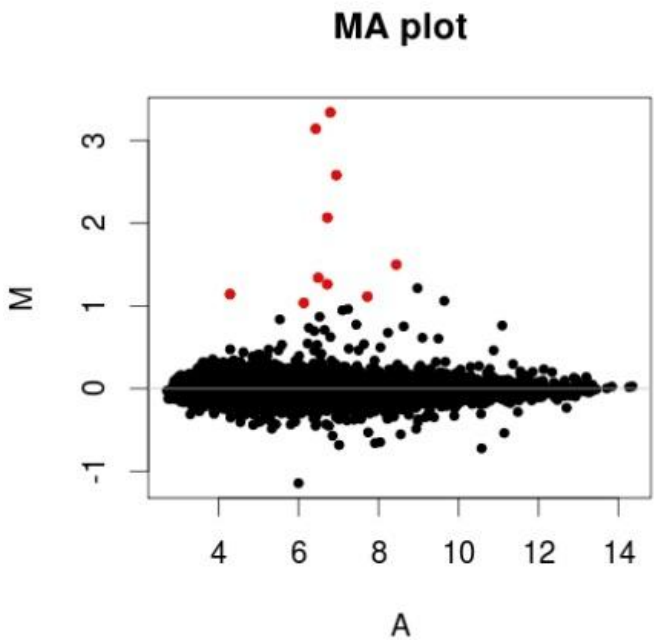
Hypothesis testing

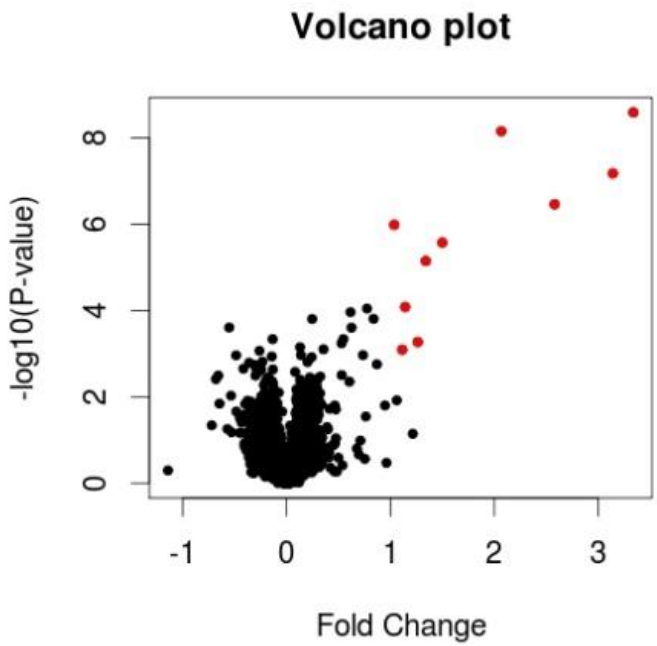
per gene

- The **null hypothesis (H_0)**: gene expression is the same in both conditions
- Calculate a **p-value**
- Adjust for **multiple testing** (e.g. FDR)



MA and volcano plots





MA plots shows the relationship between red and green channels.

Volcano plots FC against p-value of differences between samples.

Functional Enrichment

Connect the changes in the level of expression to changes in higher-level biological functions, checking if differentially expressed genes are **enriched** for specific functional terms.

Are the functional terms associated to the genes in my set **overrepresented** with respect to a **background** set of genes?

Functional Enrichment

Connect the changes in the level of expression to changes in higher-level biological functions, checking if differentially expressed genes are **enriched** for specific functional terms.

Are the functional terms associated to the genes in my set **overrepresented** with respect to a **background** set of genes?

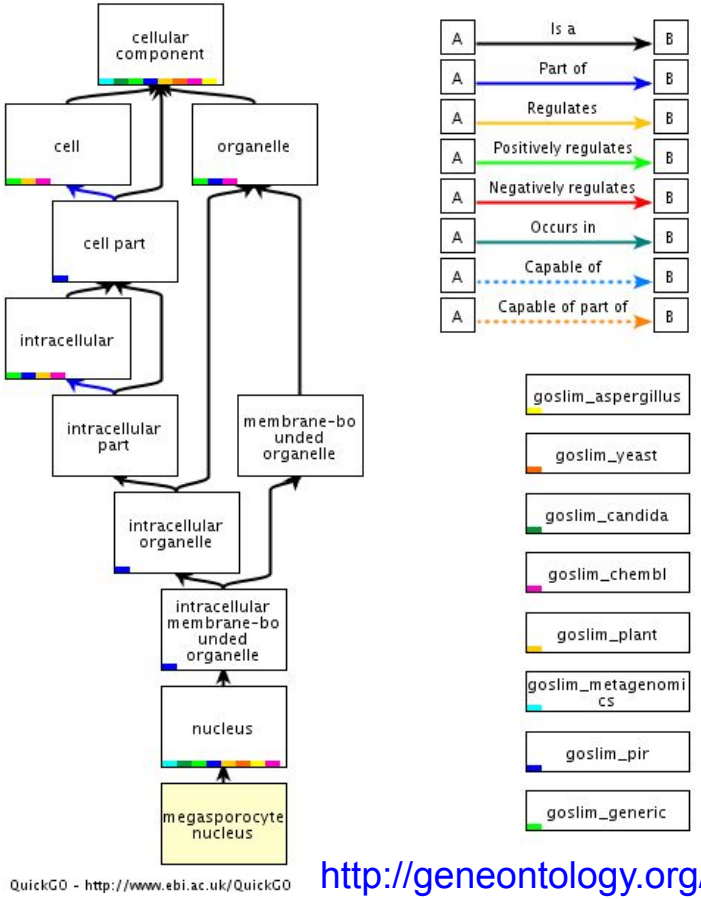
a. Databases of annotated functional terms:
   - Gene Ontology
   - KEGG pathways
   - Reactome
   - Human Phenotype Ontology
   - WikiPathways

a. Statistical test:
   - Over-representation analysis (hypergeometric test)
   - Gene Set Enrichment Analysis

Functional Enrichment

Gene Ontology (GO)

- Allows to capture biological knowledge in a written and computable form.
- Defines **classes** used to describe gene function, and relationships between these.
- 3 Main controlled vocabularies:
  - Biological Process (BP)
  - Molecular Function (MF)
  - Cellular Component (CC)



GO:0043076

http://geneontology.org/

Bioinformatics Week!

Thank you!

…

Tamara Perteghella,
Silvia González López

Hands-on III

- Differential Gene Expression
- Data Visualisation
- Functional enrichment

Hands-on 2024