Studying the transcriptome using RNA-seq

Cecilia Coimbra Klein

IBUB
Institut de Biomedicina de la Universitat de Barcelona
I→B=U
B
UNIVERSITAT DE BARCELONA

CRG
Centre for Genomic Regulation

UVIC UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA
Master in Omics Data Analysis

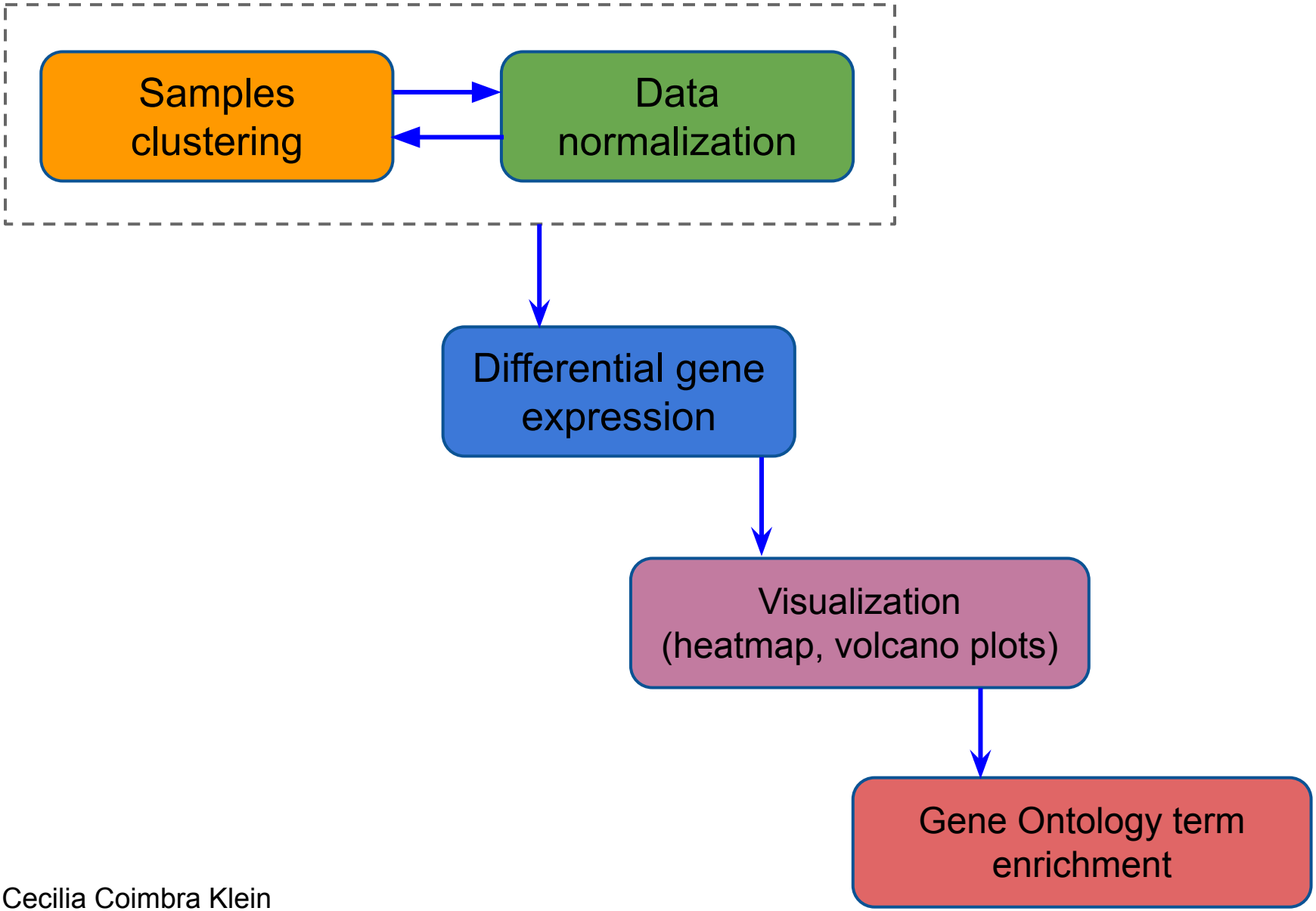
Outline

Outline

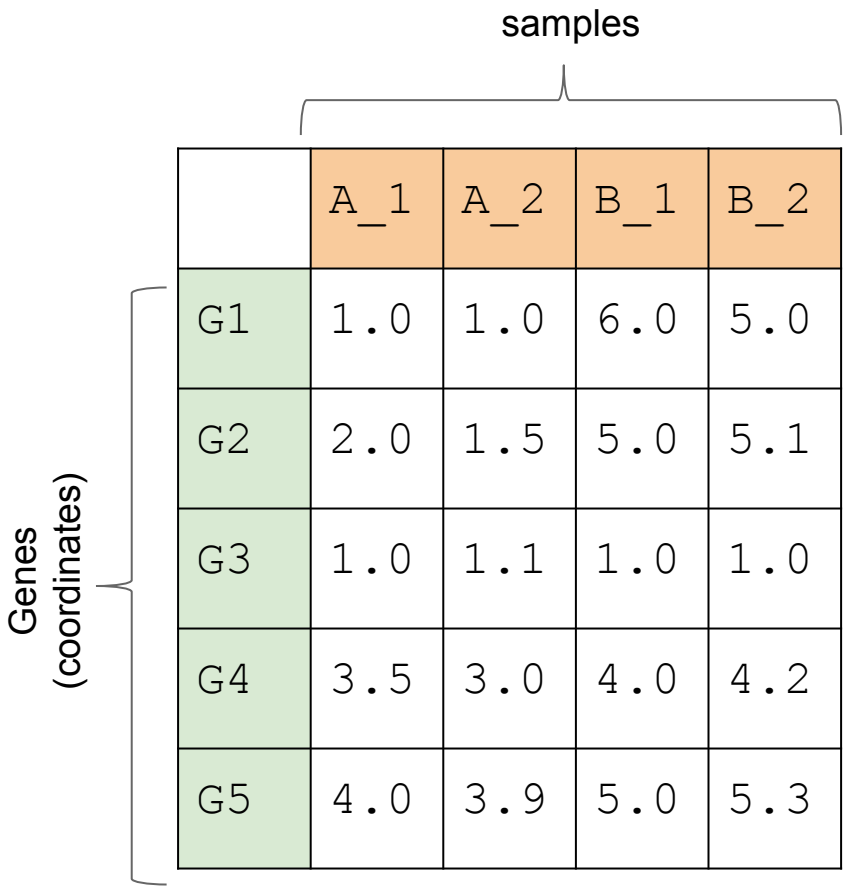
1. Introduction
2. Basic concepts
3. Short-read RNA-seq data processing
4. Gene level RNA-seq data analysis
   - 4.1. Sample clustering based on gene expression
   - 4.2. Differential gene expression
   - 4.3. Gene ontology (GO) term enrichment
5. Isoform level RNA-seq analyses
6. Regulation of gene expression

RNA-seq data analysis

Analysis pipeline



A practical example: Gene expression matrix



	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

- which samples are more alike and which are more different?
- which genes are more alike and which are more different?
- clustering: grouping genes and/or samples such that similar ones are closer to each other

Distance matrix calculation

	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

distance matrix

	A_1	A_2	B_1	B_2
A_1				
A_2				
B_1				
B_2				

Distance matrix calculation

	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

0.52

0.72

Euclidean distance:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Distance matrix calculation

	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	5.94	5.27
A_2	0.72	0.0		
B_1	5.94		0.0	
B_2	5.27			0.0

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Distance matrix calculation

	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	5.94	5.27
A_2	0.72	0.0		
B_1	5.94		0.0	
B_2	5.27			0.0

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Distance matrix calculation

	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	5.94	5.27
A_2	0.72	0.0		
B_1	5.94		0.0	
B_2	5.27			0.0

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Distance matrix calculation

	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	5.94	5.27
A_2	0.72	0.0		
B_1	5.94		0.0	
B_2	5.27			0.0

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Distance matrix calculation

5 x 4

	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

4 x 4

	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

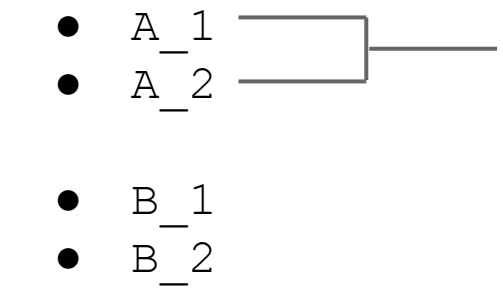
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

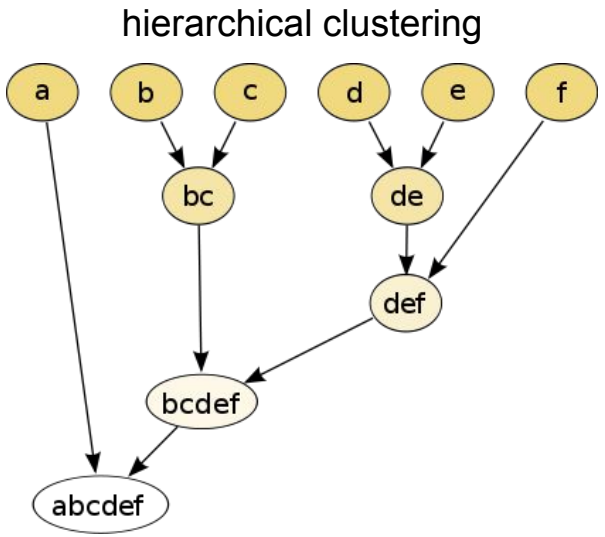
Euclidean distance is not the only way to define distance: manhattan distance, Lipschitz distance, correlation distance, etc.
They all measure distance from a different perspective.

Hierarchical clustering

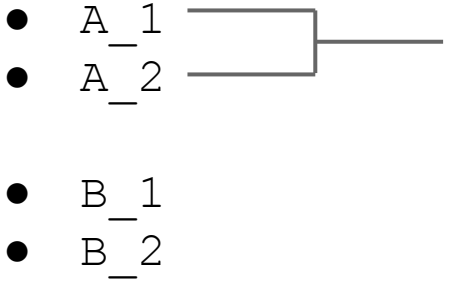
Start by finding the smallest non-diagonal element in the distance matrix. Merge these two samples together.

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





Hierarchical clustering



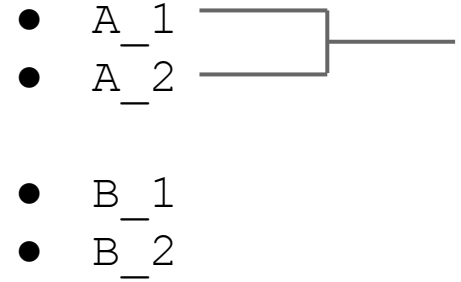
	A_1	A_2	B_1	B_2
A_1	0.0	0.72	5.9	5.27
A_2		0.0	6.28	5.69
B_1			0.0	1.07
B_2				0.0

	A_12	B_1	B_2
A_12	0.0		
B_1		0.0	1.07
B_2			0.0

Merge A_1 and A_2 into a new cluster “A_12”.

In complete linkage, the distance of this new cluster to other samples is filled by taking the max of the element of this cluster with respect to each sample.

Hierarchical clustering



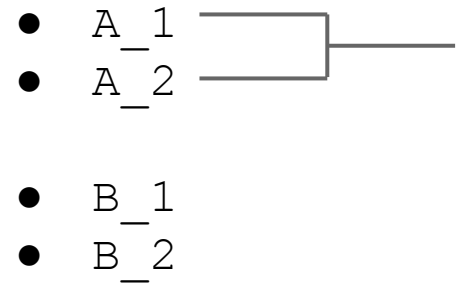
	A_1	A_2	B_1	B_2
A_1	0.0	0.72	5.9	5.27
A_2		0.0	6.28	5.69
B_1			0.0	1.07
B_2				0.0

	A_12	B_1	B_2
A_12	0.0	6.28	
B_1		0.0	1.07
B_2			0.0

Merge A_1 and A_2 into a new cluster “A_12”.

In complete linkage, the distance of this new cluster to other samples is filled by taking the max of the element of this cluster with respect to each sample.

Hierarchical clustering



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

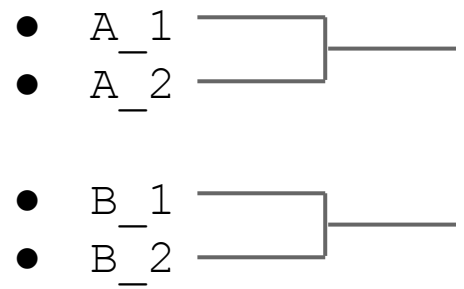
	A_12	B_1	B_2
A_12	0.0	6.28	5.69
B_1		0.0	1.07
B_2			0.0

Merge A_1 and A_2 into a new cluster “A_12”.

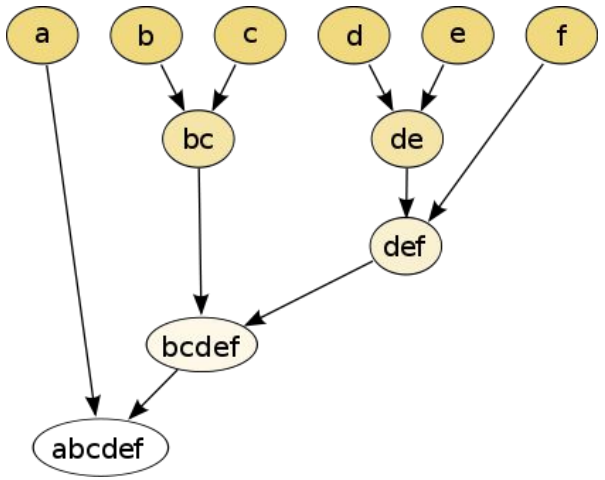
In complete linkage, the distance of this new cluster to other samples is filled by taking the max of the element of this cluster with respect to each sample.

Hierarchical clustering

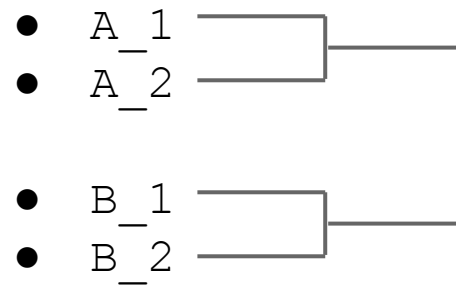
Now the merging is done, we find the smallest distance again.



	A_12	B_1	B_2
A_12	0.0	6.28	5.69
B_1		0.0	1.07
B_2			0.0



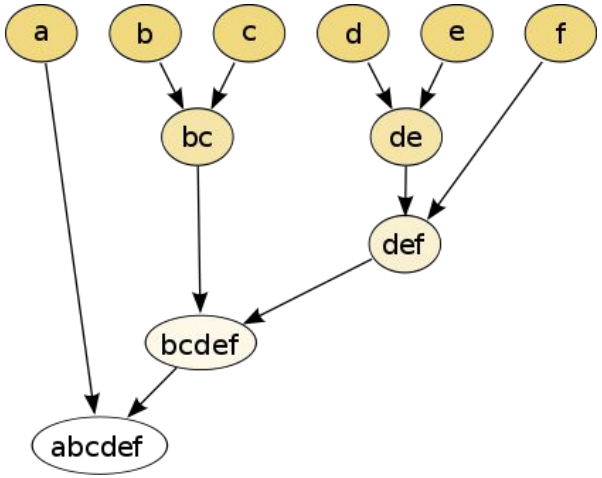
Hierarchical clustering



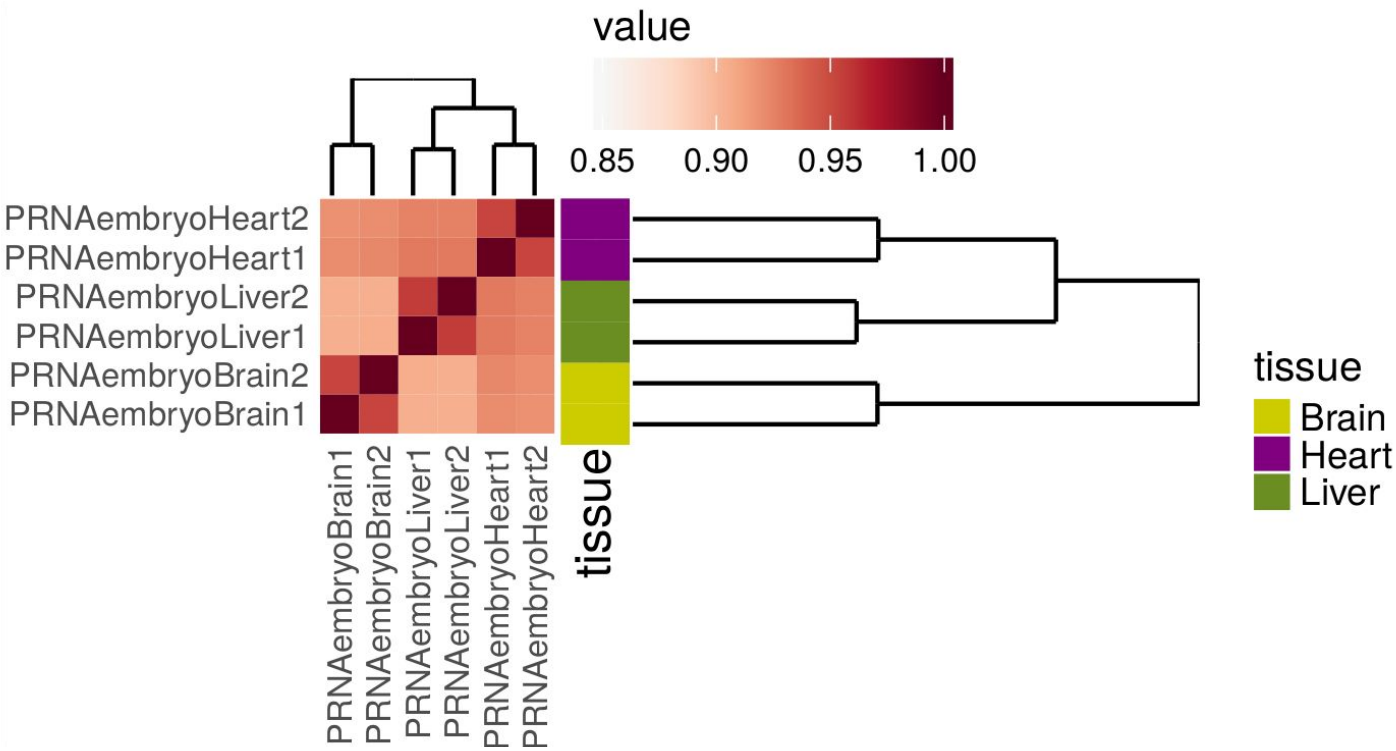
We recompute the distance matrix by selecting the maximum...

	A_12	B_1	B_2
A_12	0.0	6.28	5.69
B_1		0.0	1.07
B_2			0.0

	A_12	B_12
A_12	0.0	6.28
B_12		0.0



Samples clustering

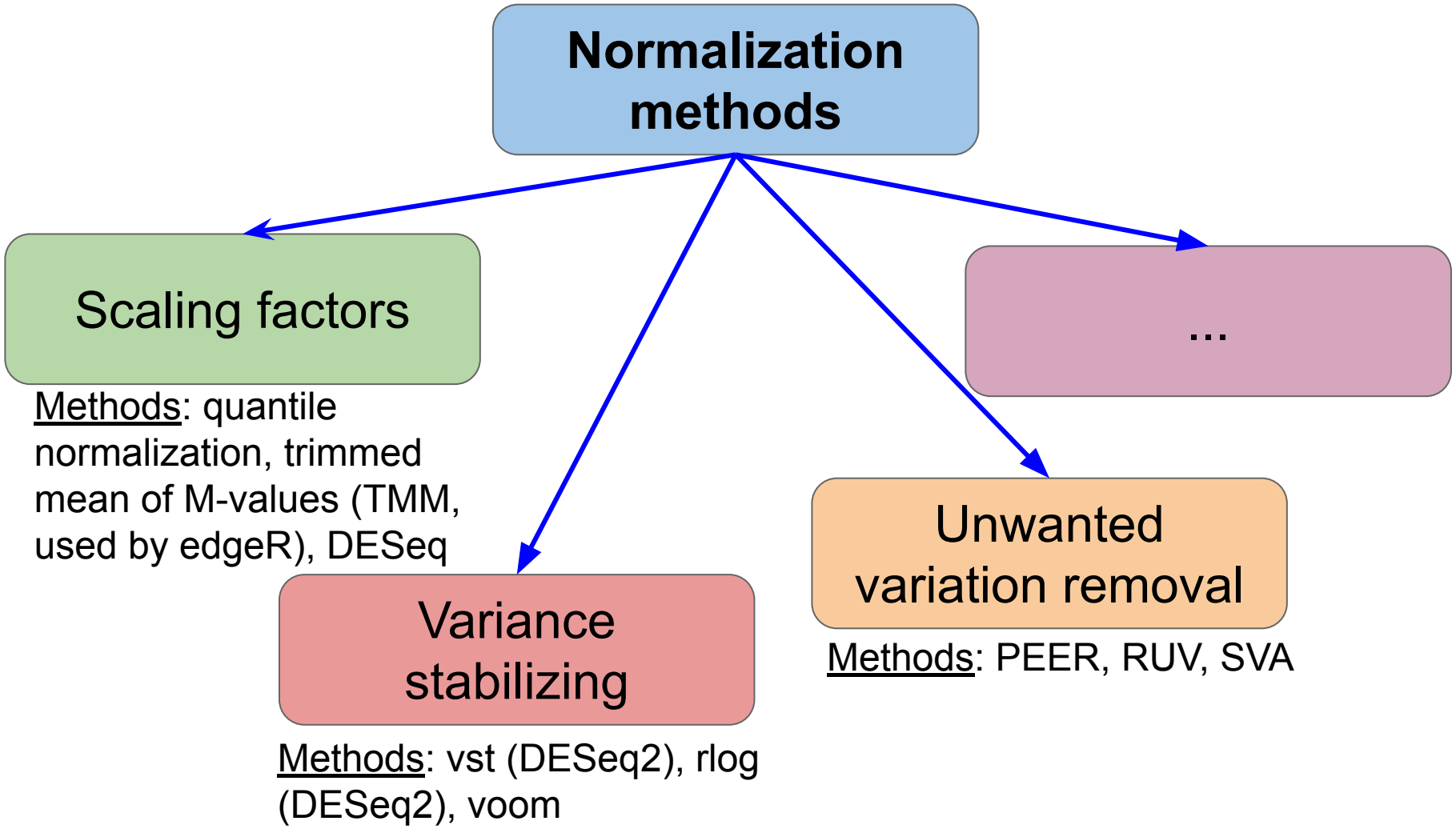


Data normalization

Raw read counts can not be compared directly: different library size, gene length, gene abundance, Normalization allows to:

- Compare different datasets
- Compare different genes
- Remove unwanted variation

Normalization methods



Methods: quantile normalization, trimmed mean of M-values (TMM, used by edgeR), DESeq

Methods: PEER, RUV, SVA

Methods: vst (DESeq2), rlog (DESeq2), voom

Differential gene expression (DGE)

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

Comparisons:

- pairwise with one factor (most common)
- pairwise with multiple factors
- among more than two samples
- time-series

Always better to have ≥ 2 replicates per sample

Soneson, Charlotte, and Mauro Delorenzi. "**A comparison of methods for differential expression analysis of RNA-seq data.**" *BMC bioinformatics* 14.1 (2013): 91.

Differential gene expression (DGE)

Sex	Sample	g_1	g_2	g_3	...
Male	A_1				
Male	A_2				
Male	A_3				
Male	A_4				
Female	B_1				
Female	B_2				
Female	B_3				
Female	B_4				

Software examples

- 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.
- DESeq (R package)
  - Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." *Genome biol* 11.10 (2010): R106.
- 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): 550.
- 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.
- Cuffdiff 2
  - Trapnell, Cole, et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." *Nature biotechnology* 31.1 (2013): 46-53.

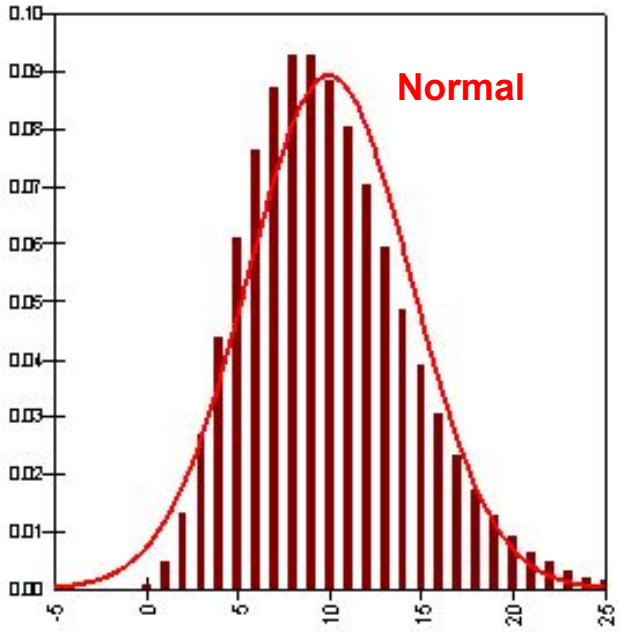
Basics of DGE

Normalization

Is it required?

Fit a model to the data per gene

- Data (read counts) **discrete and positive**
- Which distribution do we select?
  Negative binomial

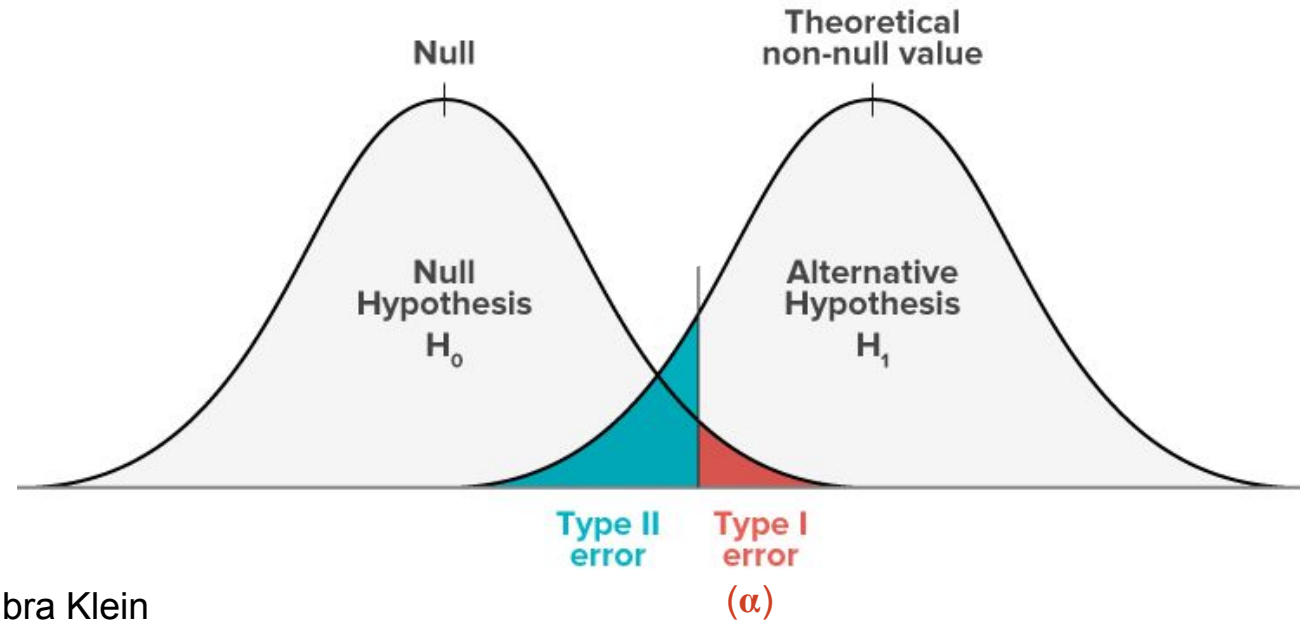


We need to estimate the **mean** and **variance** of the fitted distribution

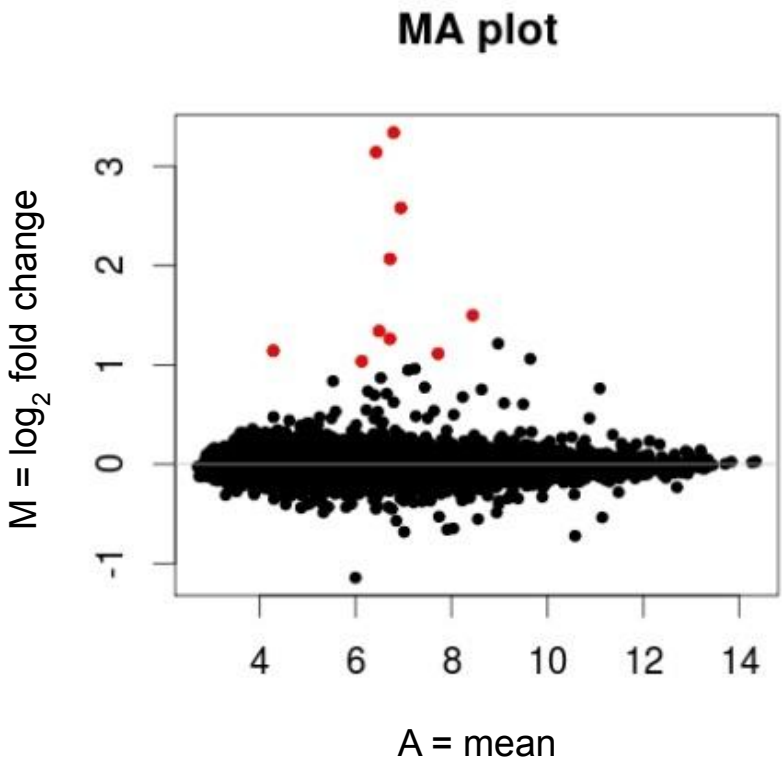
Hypothesis testing

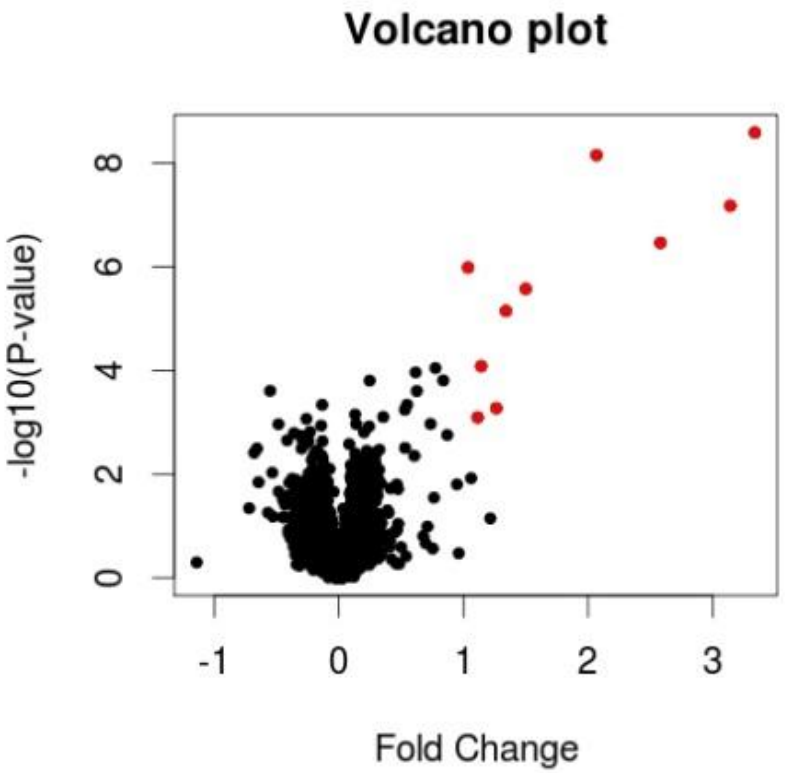
per gene

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



Visualization: MA and volcano plots



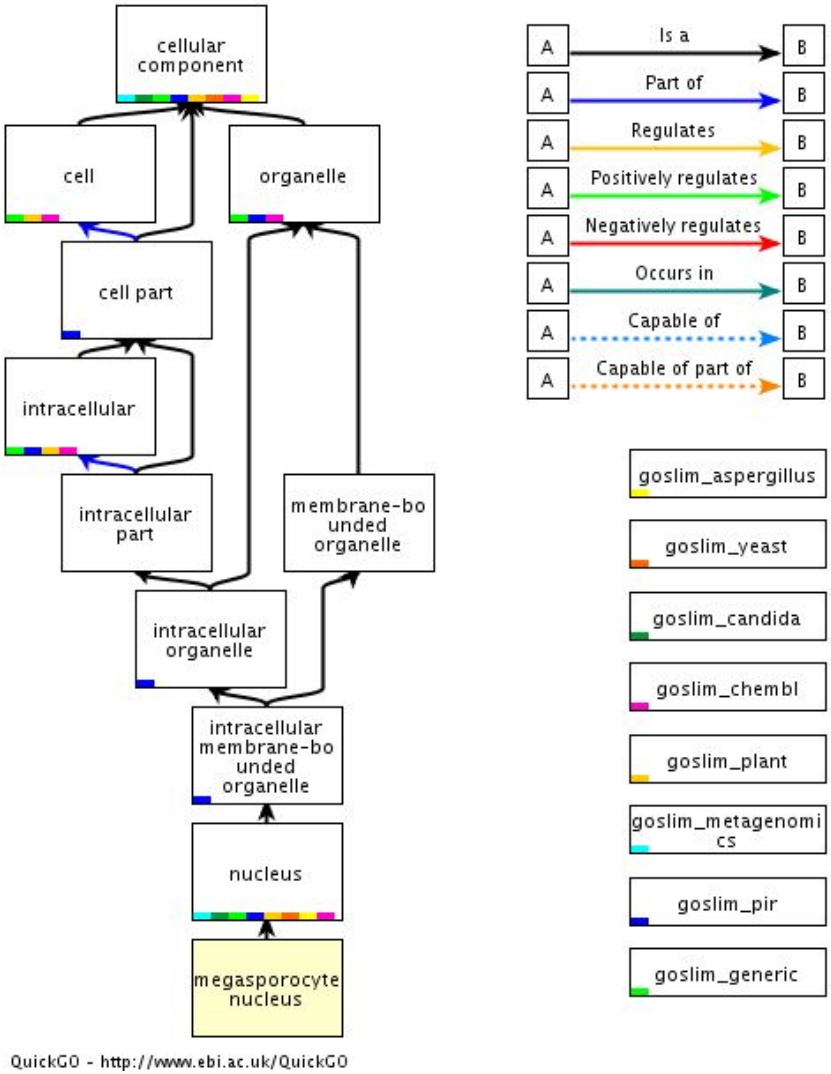


Gene Ontology Term Enrichment

Gene Ontology (GO)

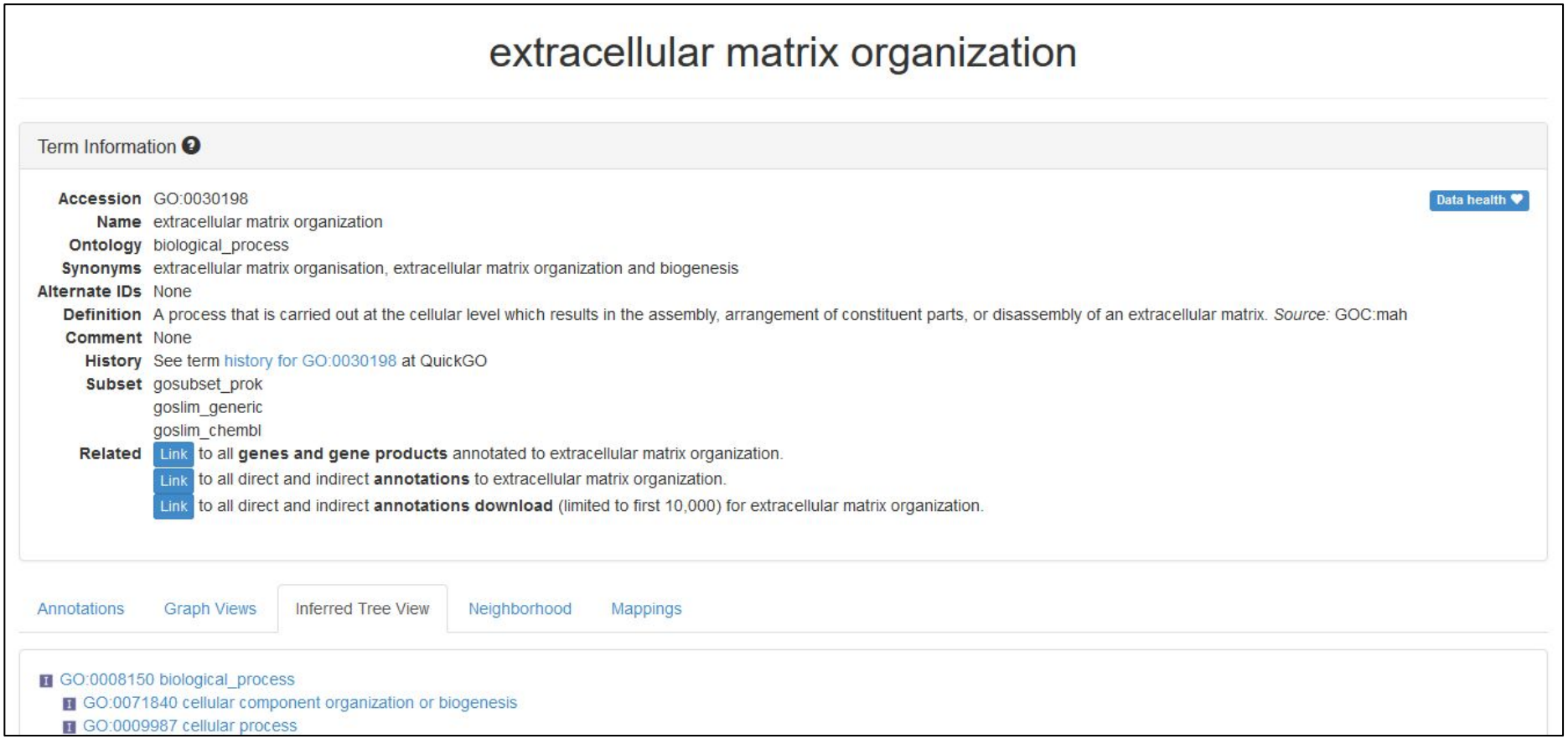
- Allows to capture biological knowledge in a written and computable form.
- Defines **concepts**/classes used to describe gene function, and **relationships** between these concepts.
- Controlled vocabulary
- 3 main categories:
  - ➔ Biological Process (BP)
  - ➔ Molecular Function (MF)
  - ➔ Cellular Component (CC)
- The same gene can have more than oneGO terms

The annotation is both manual and automatic



Not good for lncRNAs!

Gene Ontology Term Enrichment



http://amigo.geneontology.org/amigo

Gene Ontology Term Enrichment

Aim: Does my set of genes (identified as differentially expressed) have characteristic GO terms associated to it?

Enrichment: we should look whether GO terms associated to the genes in my set are overrepresented with respect to a background set of genes.

There are many ways to statistically test this, and multiple software available online. One example is the R package GOstats, which can be run locally. It uses a hypergeometric test to assess the enrichment.

Other software: topGO, GOrilla, Metascape

Visualization: REVIGO

http://revigo.irb.hr/

Hands-on

Gene level RNA-seq data analysis 4

https://public-docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/#_gene_level_rna_seq_data_analysis