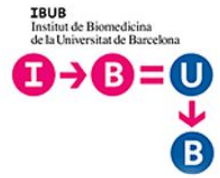
Studying the transcriptome using RNA-seq

Cecilia Coimbra Klein
Computational Biology of RNA Processing, CRG
Departament de Genètica, IBUB, UB

Master in Omics Data Analysis
Jan. 2019



Outline

Outline

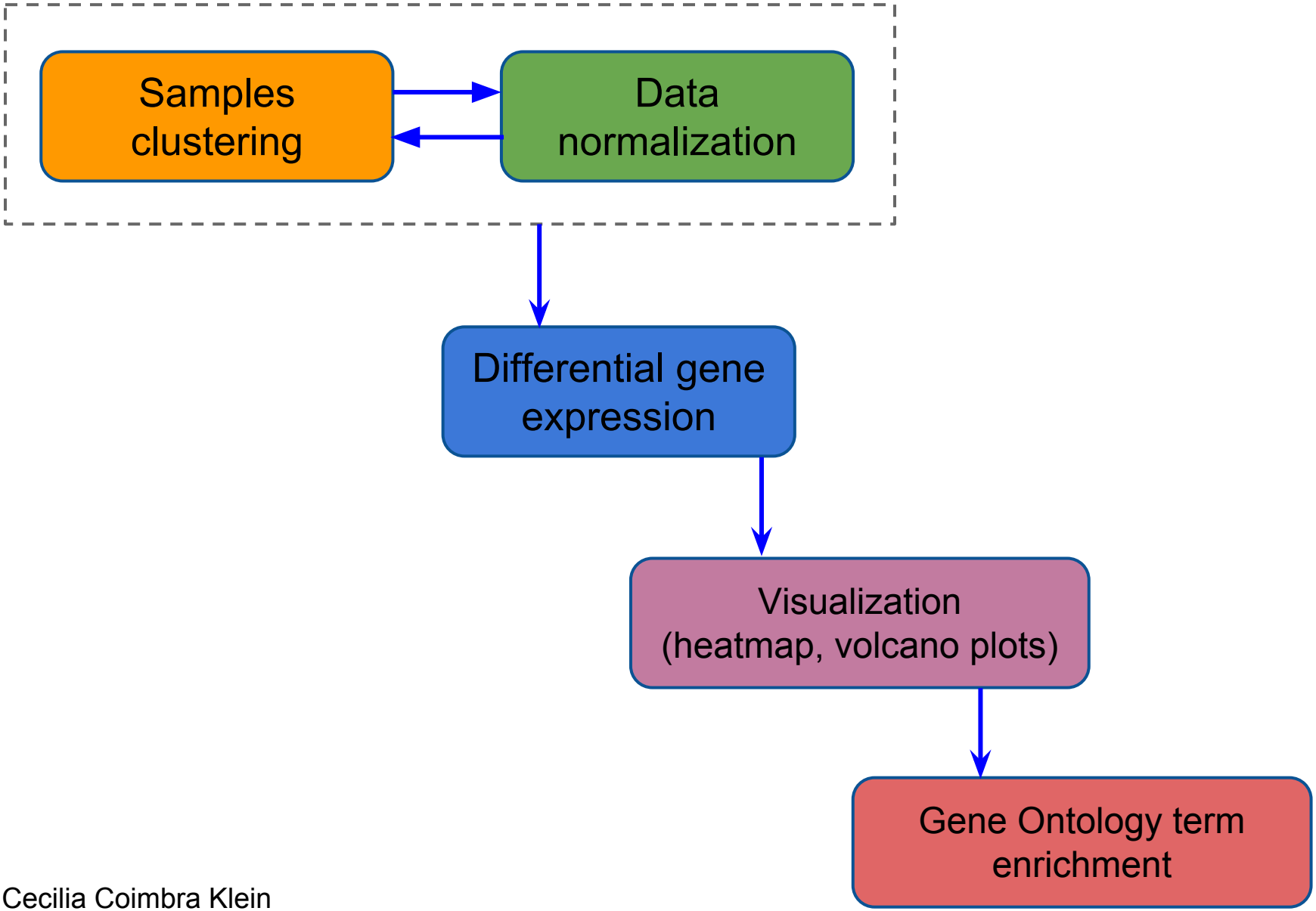
- Basic concepts
- Reference gene annotation
- Next generation sequencing
- RNA-seq experimental protocols
- Short-read RNA-seq data processing
  - mapping
  - visualisation of gene expression signal
  - gene expression quantification
- RNA-seq data analysis
  - sample clustering based on gene expression
  - differential gene expression
  - gene ontology (GO) term enrichment
  - differential splicing analysis

Outline

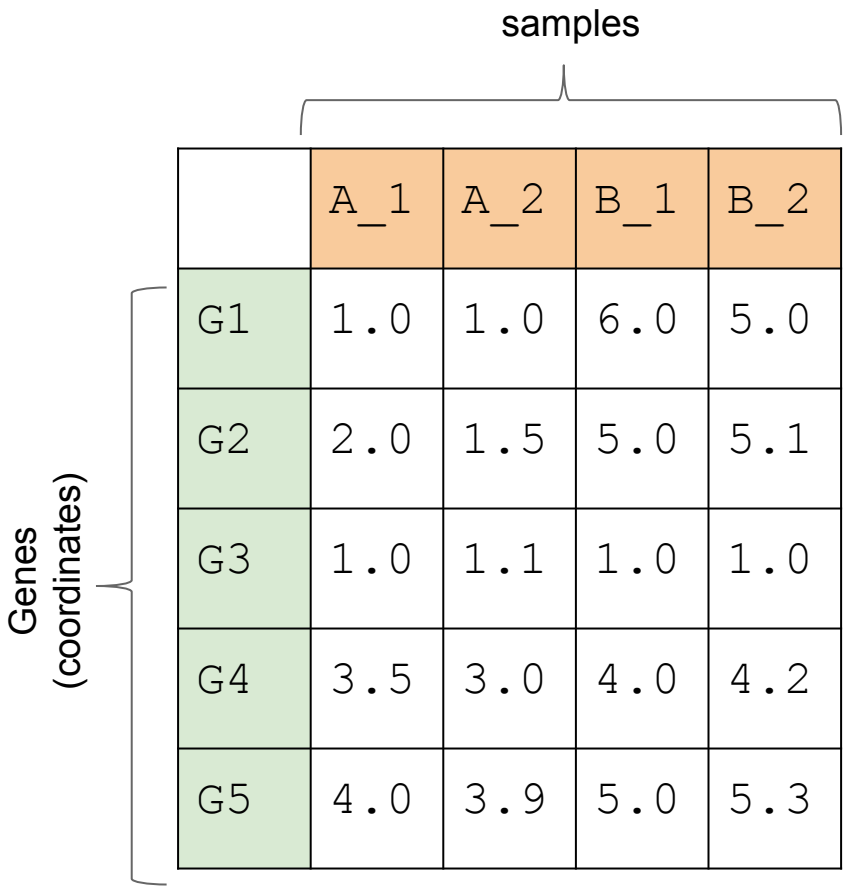
- ChIP-seq data processing
  - mapping
  - peak calling
  - visualisation of signal
- ChIP-seq data analysis
  - genomic locations
  - differential peaks per tissue
  - BED files in UCSC browser
- Integrative data analysis
  - promoter regions of differentially expressed genes
  - ATAC-seq signal in the UCSC genome browser
  - promoter regions of differentially spliced genes
  - omics portals

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.

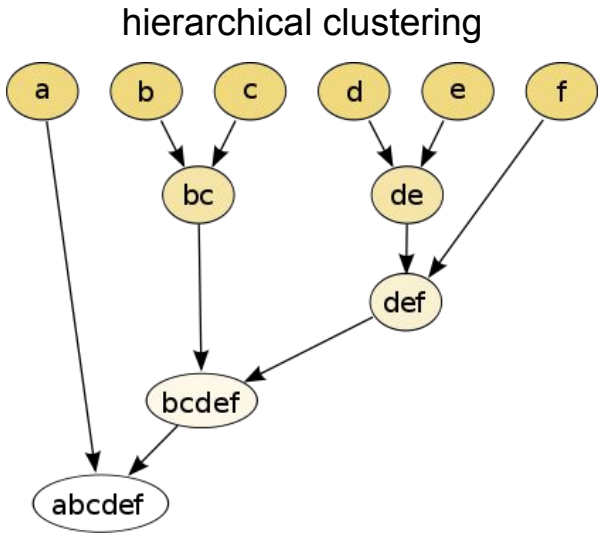
Hierarchical clustering

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

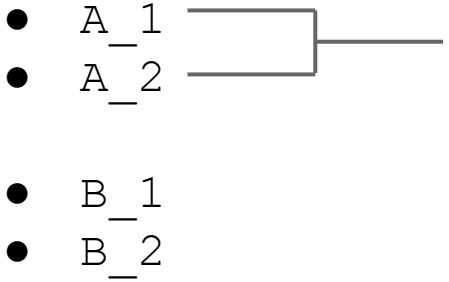
	A_1	A_2	B_1	B_2
A_1	0.0	0.72	5.9	5.27
A_2		0.0	6.28	5.69
B_1			0.0	1.07
B_2				0.0

- A_1
- A_2

- B_1
- B_2



Hierarchical clustering



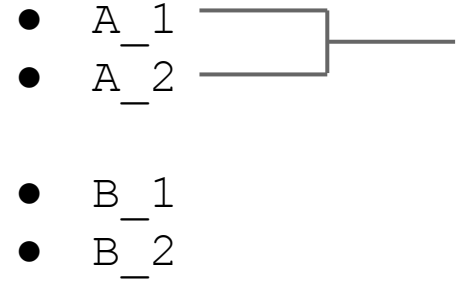
	A_1	A_2	B_1	B_2
A_1	0.0	0.72	5.9	5.27
A_2		0.0	6.28	5.69
B_1			0.0	1.07
B_2				0.0

	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



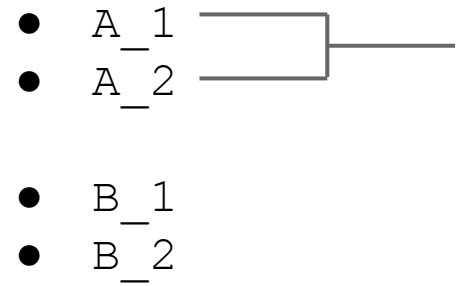
	A_1	A_2	B_1	B_2
A_1	0.0	0.72	5.9	5.27
A_2		0.0	6.28	5.69
B_1			0.0	1.07
B_2				0.0

	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

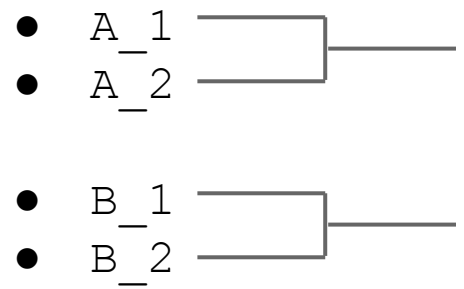
	A_12	B_1	B_2
A_12	0.0	6.28	5.69
B_1		0.0	1.07
B_2			0.0

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

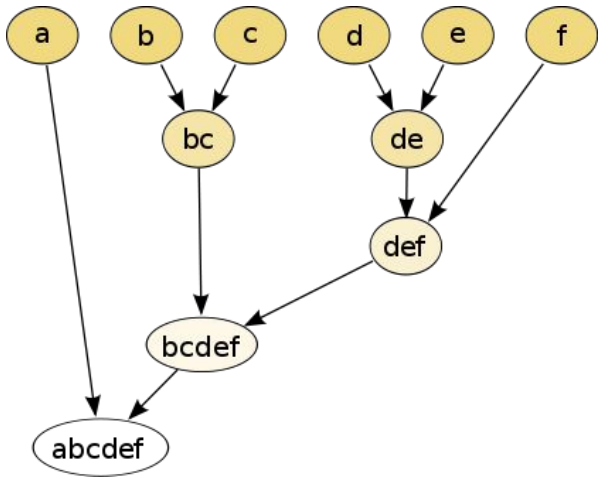
In complete linkage, the distance of this new cluster to other samples is filled by taking the max of the element of this cluster with respect to each sample.

Hierarchical clustering

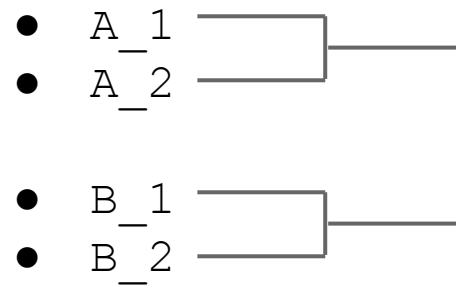
Now the merging is done, we find the smallest distance again.



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



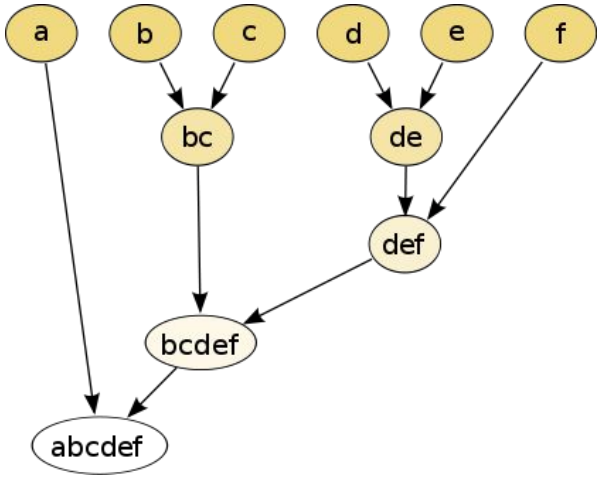
Hierarchical clustering



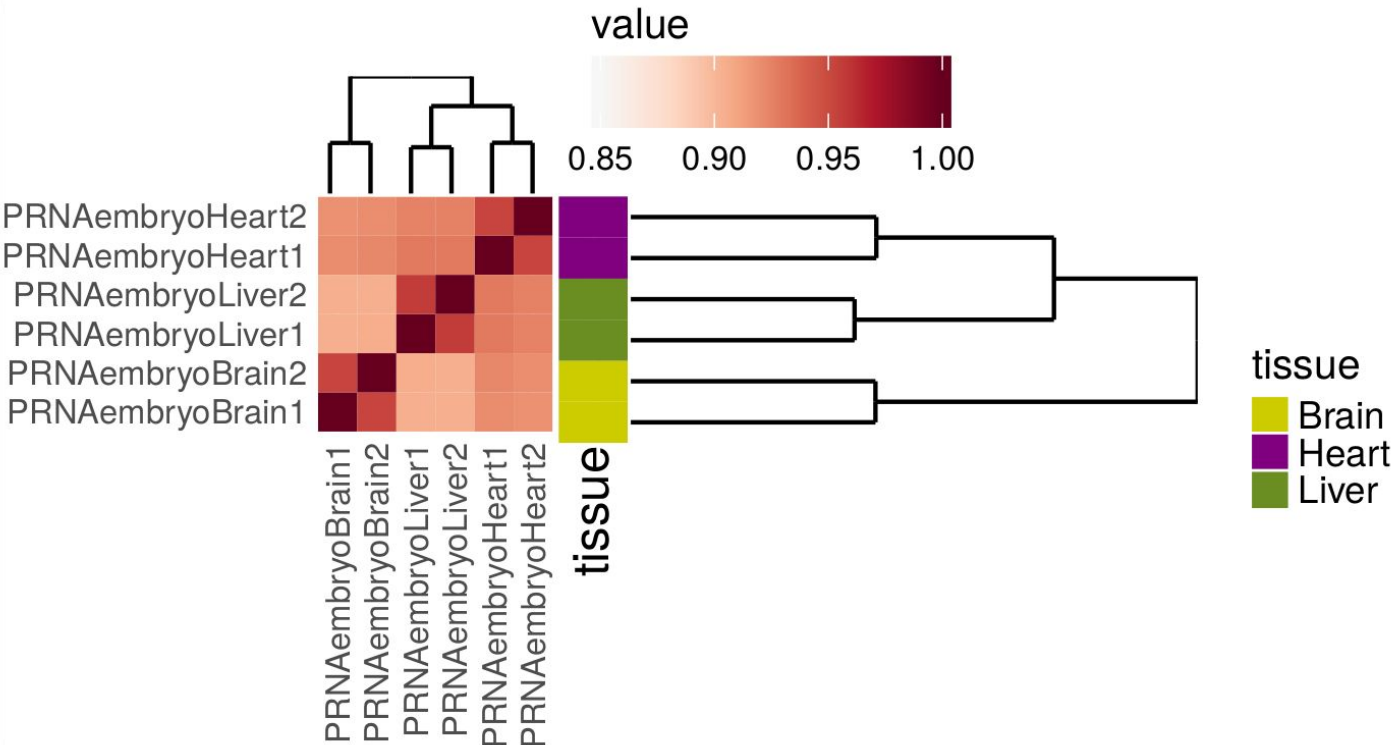
We recompute the distance matrix by selecting the maximum...

	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

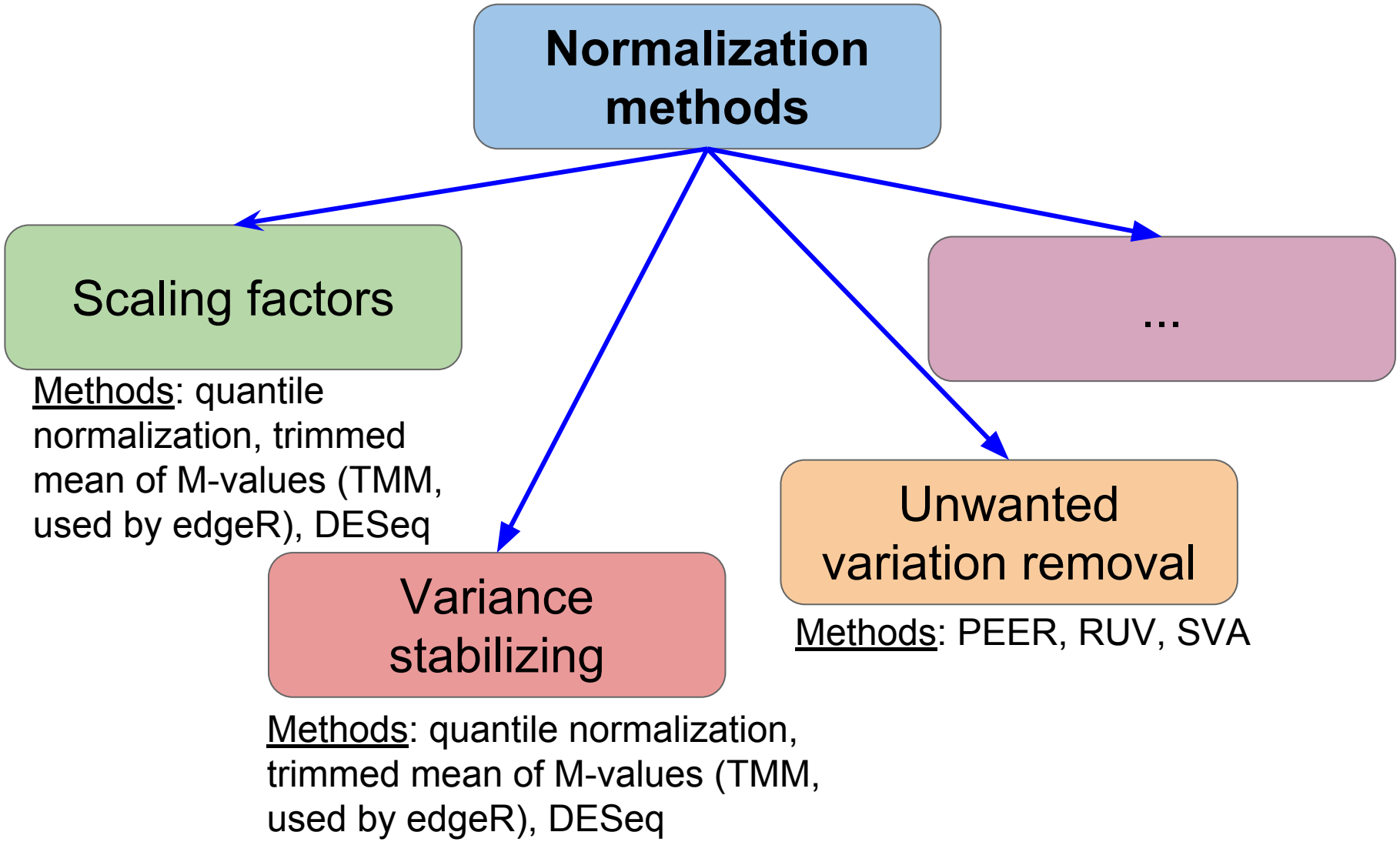


Data normalization

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



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.

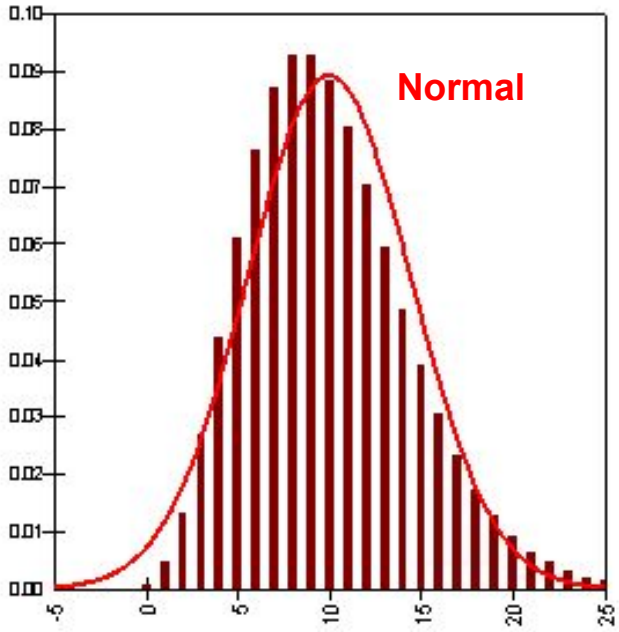
Basics of DGE

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

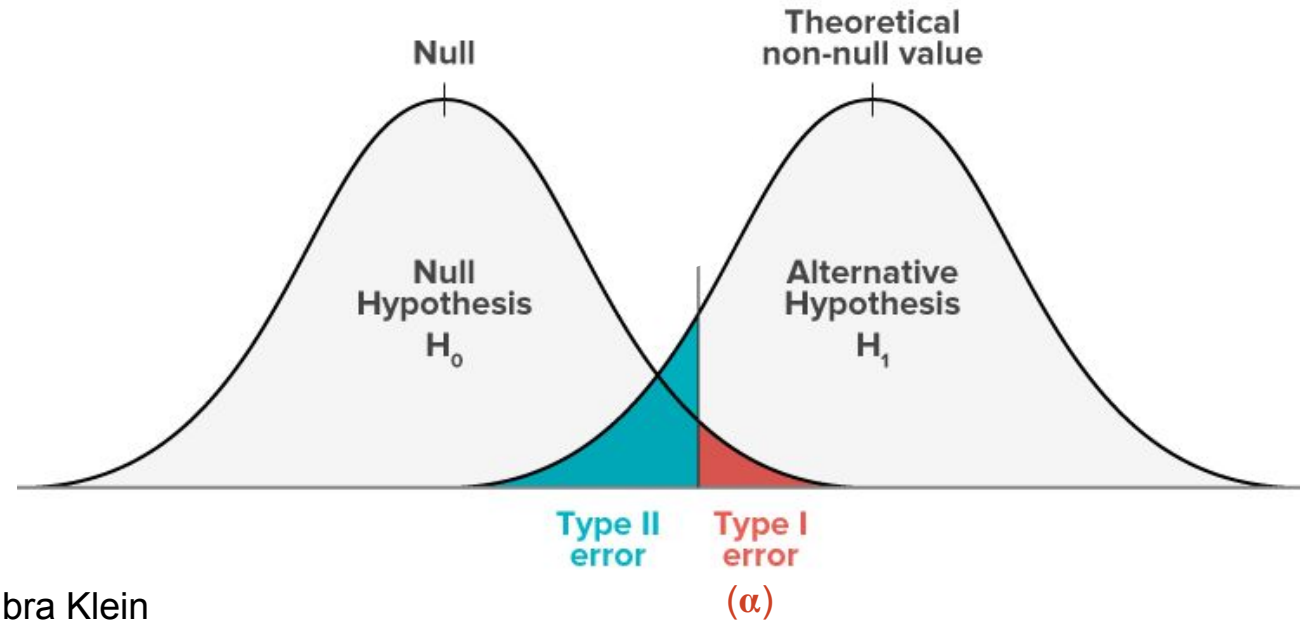


We need to estimate the **mean** and **variance** of the fitted distribution

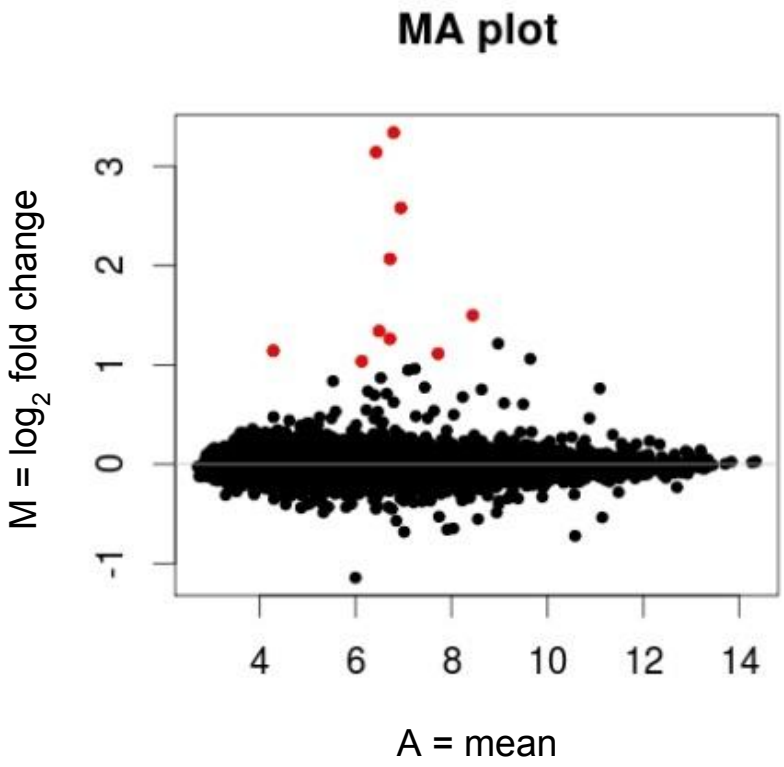
Hypothesis testing

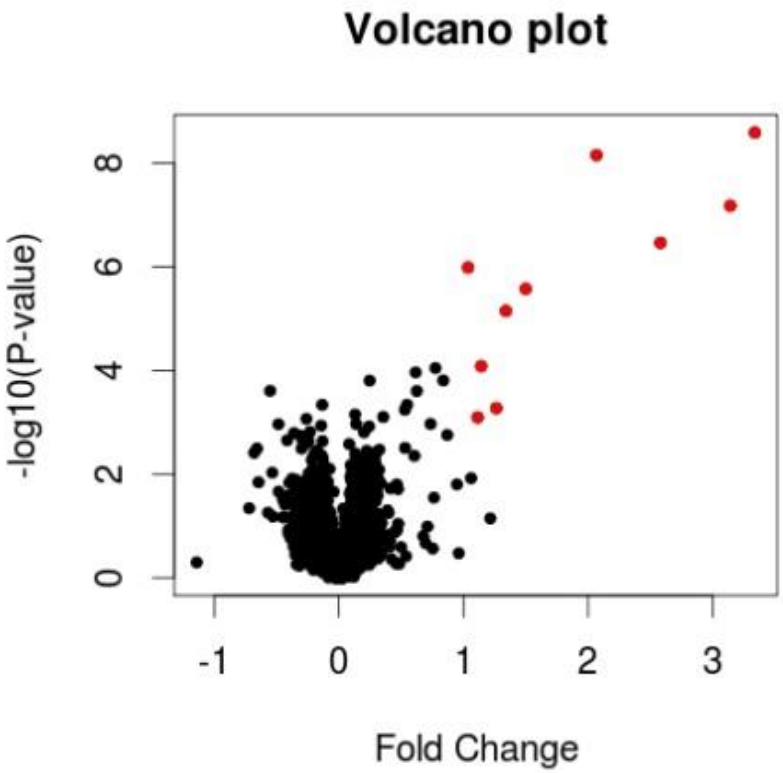
per gene

- 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



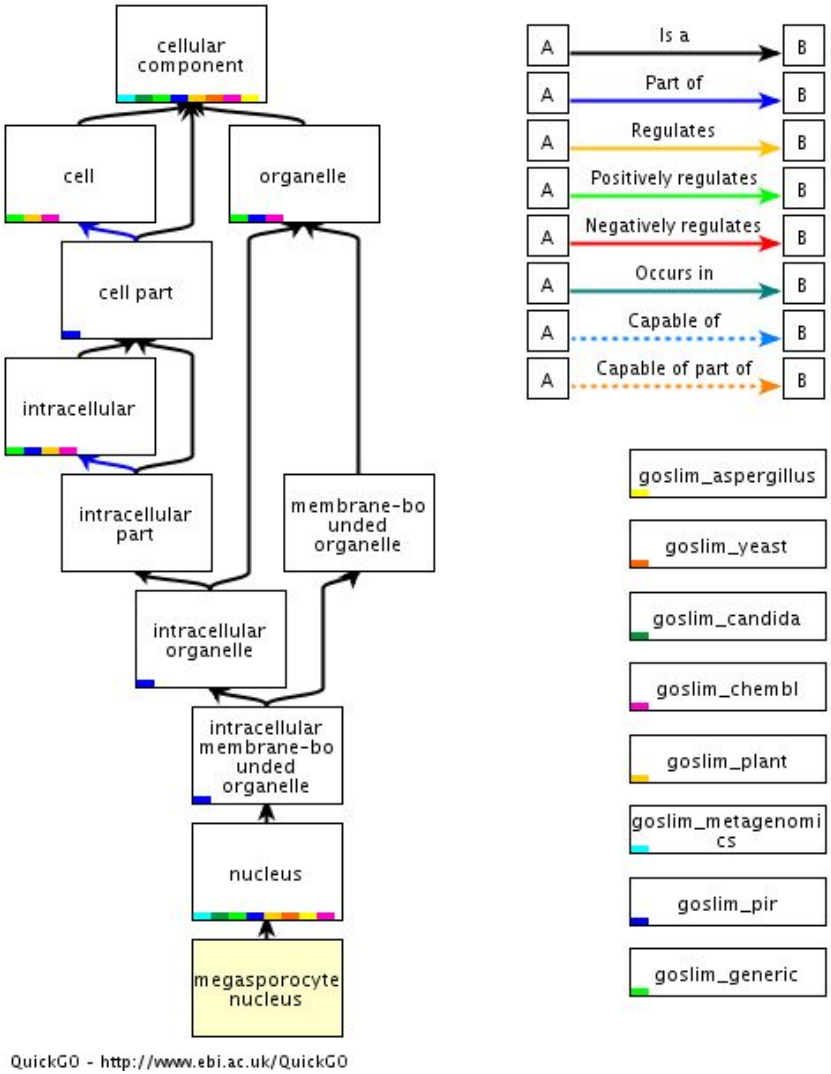


Gene Ontology Term Enrichment

Gene Ontology (GO)

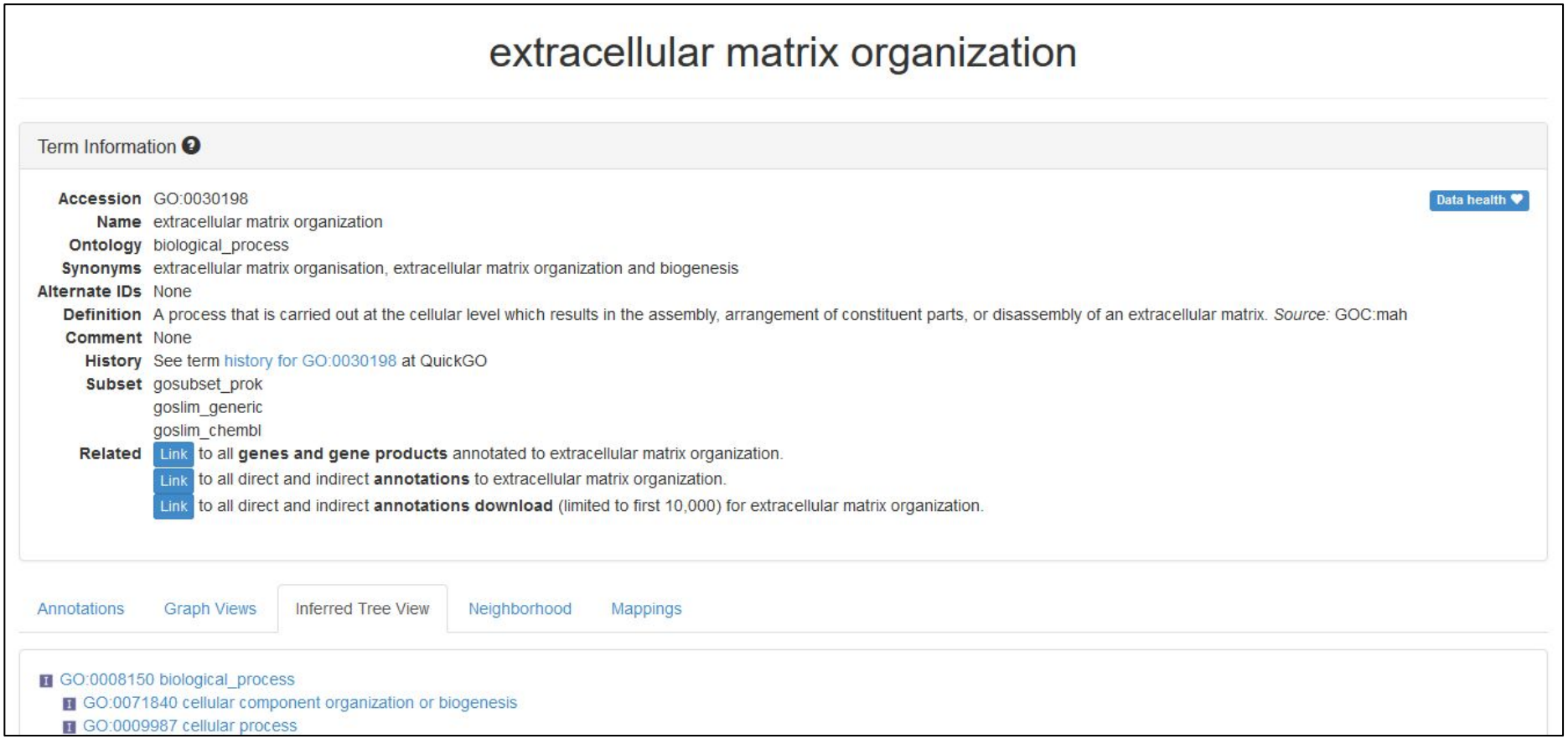
- Allows to capture biological knowledge in a written and computable form.
- Defines **concepts**/classes used to describe gene function, and **relationships** between these concepts.
- Controlled vocabulary
- 3 main categories:
  - ➔ Biological Process (BP)
  - ➔ Molecular Function (MF)
  - ➔ Cellular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

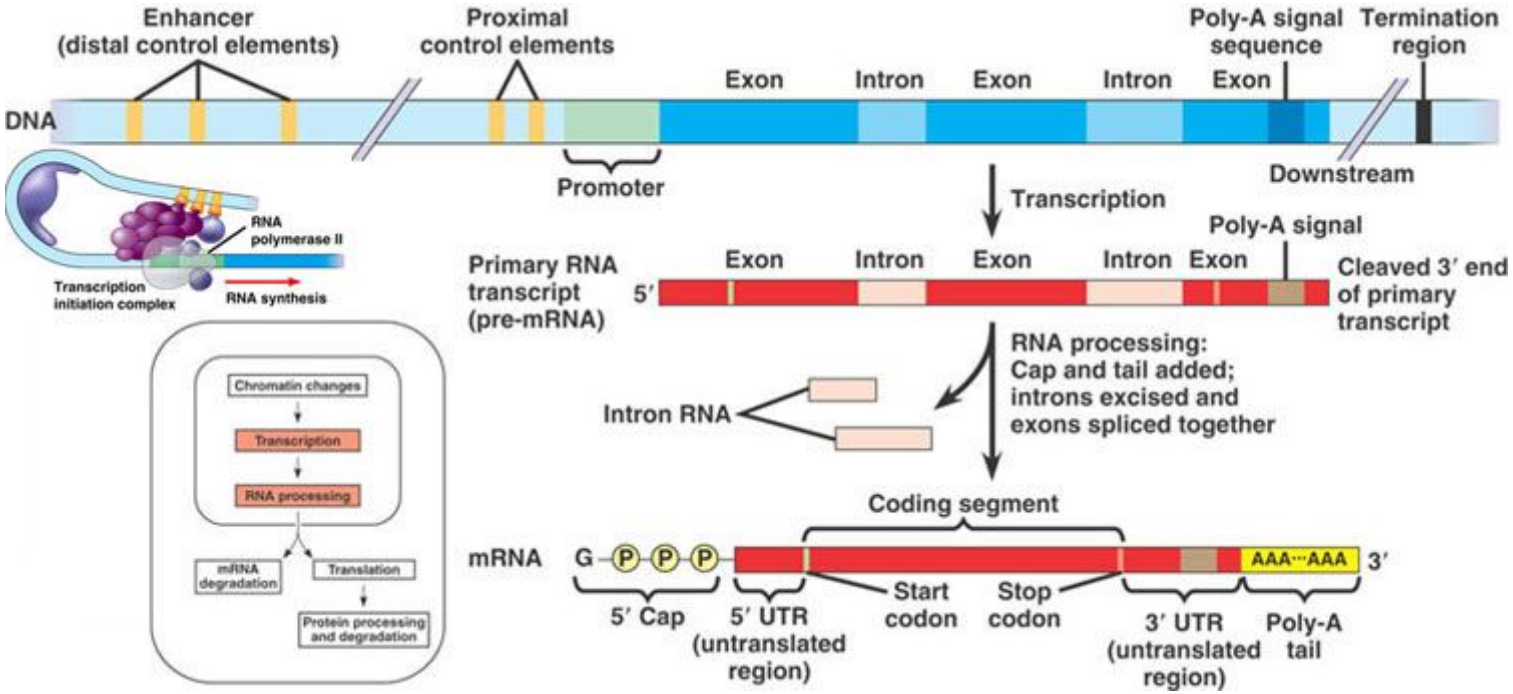
Other software: topGO, GOrilla, Metascape

Visualization: REVIGO

http://revigo.irb.hr/

Alternative splicing

RNA transcription and processing



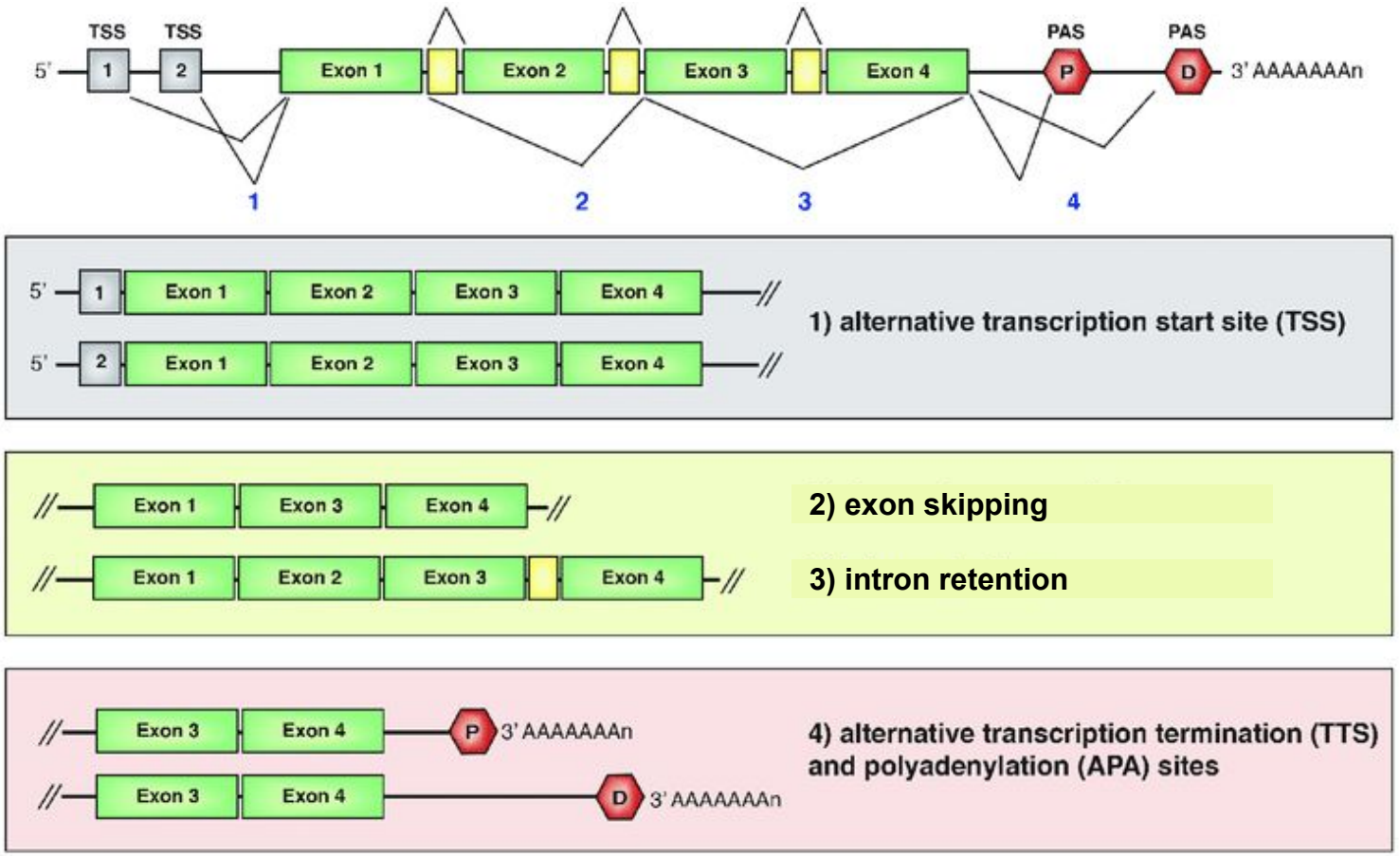
Primary RNA transcripts are extensively processed: capping, splicing, polyadenylation, editing

This process is highly regulated and results in a gene producing many distinct transcript isoforms: one gene, many transcripts

The transcriptome is distinct from and more complex than the genome

The transcriptome cannot be predicted from the genome sequence alone: it must be measured

Complexity arising from differential processing



These processing events can result in different protein products, differentially (post-) transcriptionally regulated mRNAs or non-protein coding isoforms.

Andreassi, C. *et al.* (2018). doi: 10.3389/fnmol.2018.00304

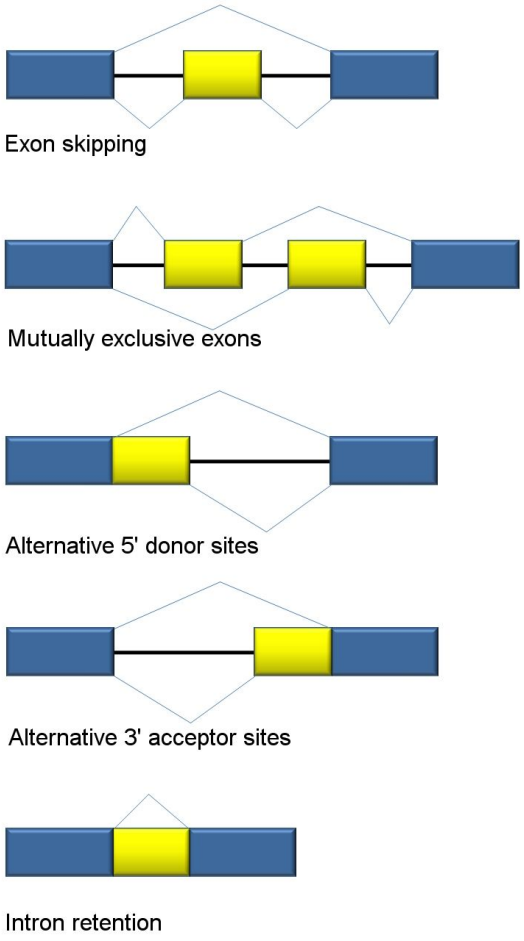
Complexity arising from differential processing

	Human[b]	Mouse[b]	Fly[c]	Worm[c]
Genome size	3,300 MB	3,300 MB	165 MB	100 MB
Protein-coding genes	22,180	22,740	13,937	20,541
Multiexonic genes (percentage with 2+ isoforms)	21,144 (88%)	19,654 (63%)	11,767 (45%)	20,008 (25%)
Isoforms (average number per gene)	215,170 (3.4)	94,929 (2.4)	29,173 (1.9)	56,820 (1.2)
Genes (all)	63,677	39,179	15,682	46,726

- pre-mRNA splicing scales with organismal complexity.
- Alternative pre-mRNA splicing occurs in ~88% of human genes, compared with ~63% of mouse genes.
- More recent deep RNA-seq data, 95% to 100% of human genes may encode two or more (2+) isoforms
- One function of alternative splicing is to significantly expand the form and function of the human proteome

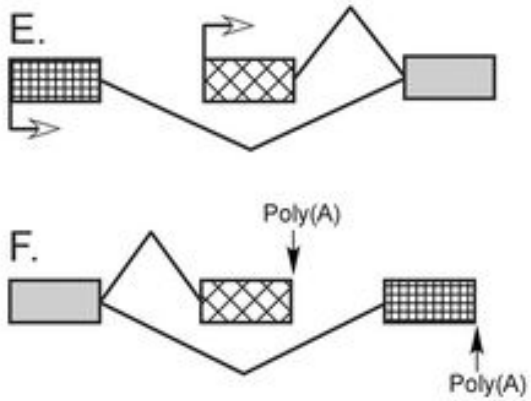
Lee & Rio (2015). doi:10.1146/annurev-biochem-060614-034316

Modes of AS



Exons are represented as blue and yellow blocks, introns as lines in between.

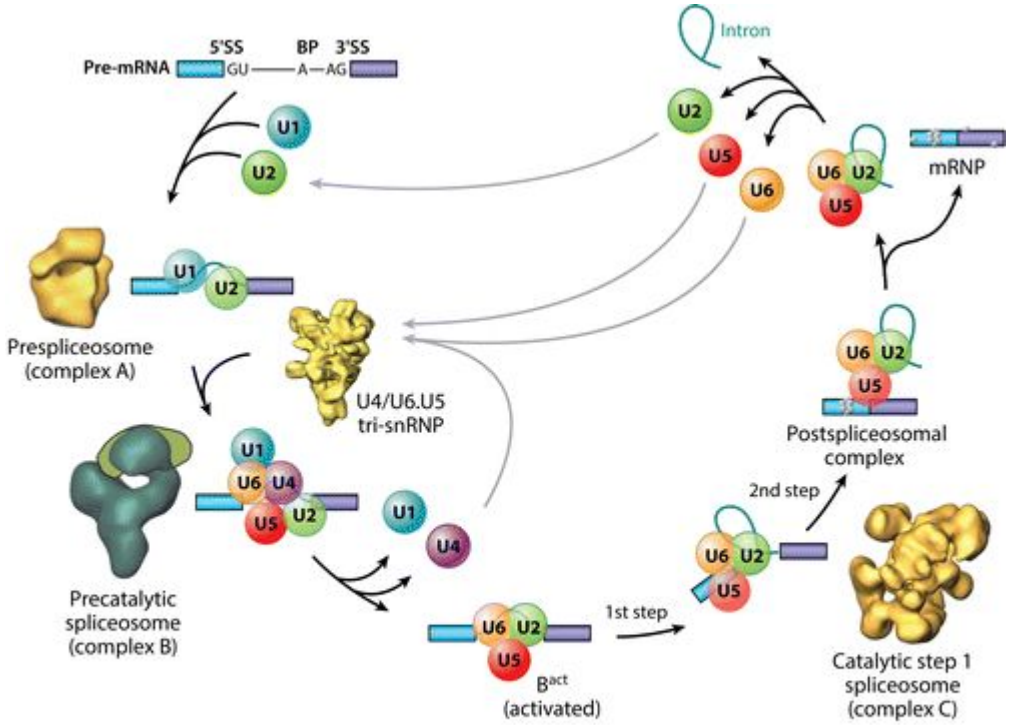
Alternative promoters and polyadenylation sites



Alternative promoters are primarily an issue of transcriptional control. Control of polyadenylation appears mechanistically similar to control of splicing.Both of these mechanisms are found in combination with alternative splicing and provide additional variety in mRNAs derived from a gene

Black (2003) doi: 10.1146/annurev.biochem.72.121801.161720
https://en.wikipedia.org/wiki/Alternative_splicing

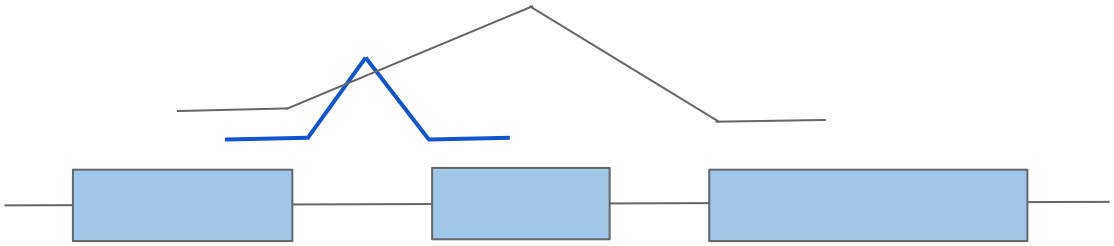
General splicing mechanism



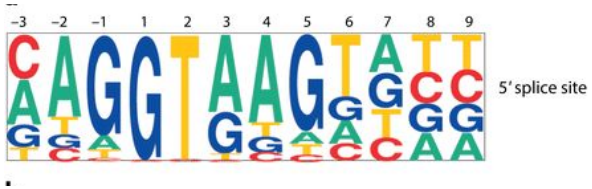
Lee Y, Rio DC. 2015.
Annu. Rev. Biochem. 84:291–323

Lee & Rio (2015). doi:10.1146/annurev-biochem-060614-034316

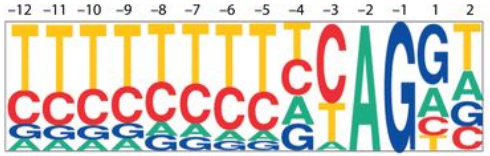
Junctions



Splice sites in the human genome:



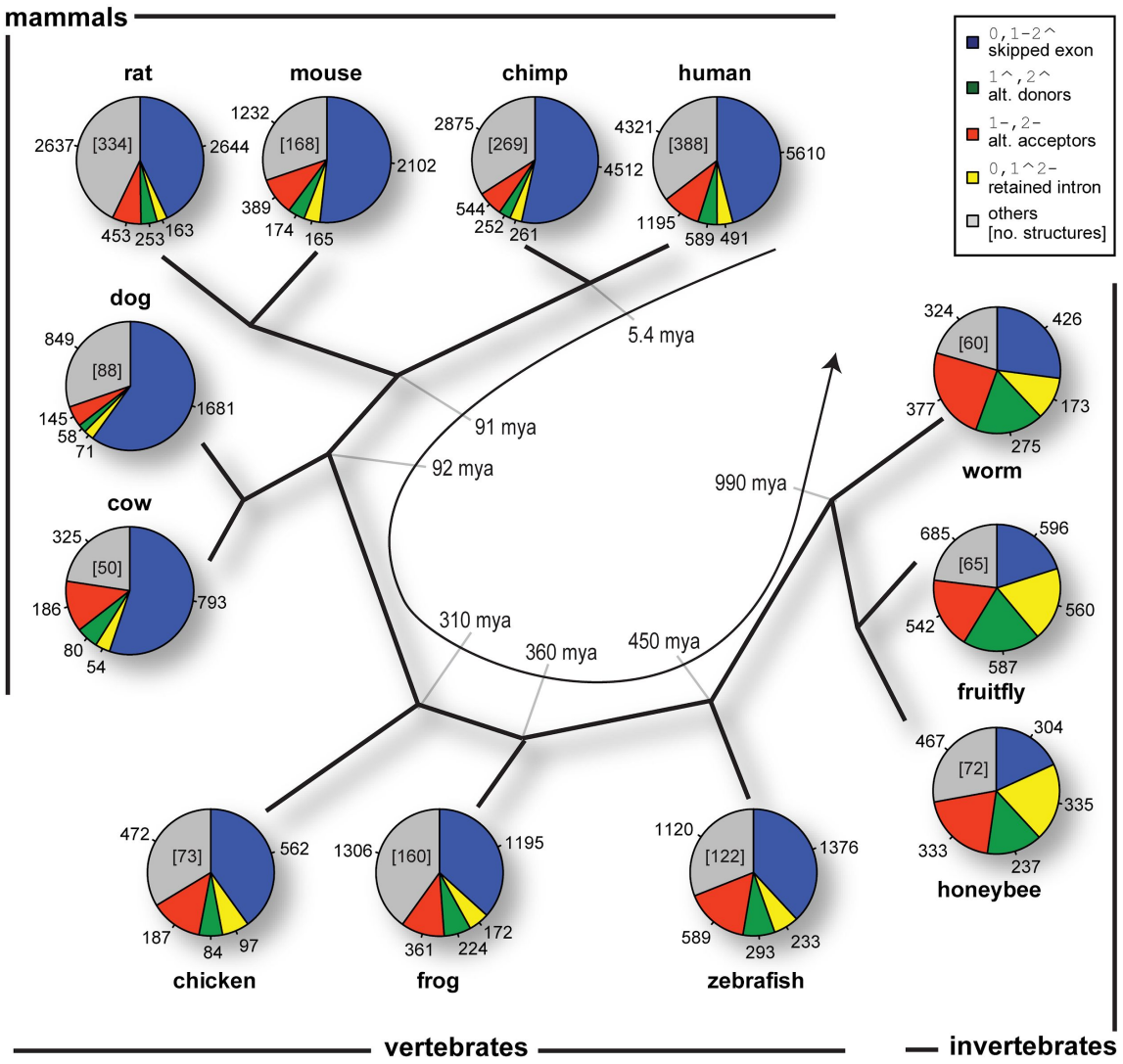
5′ splice site



3′ splice site

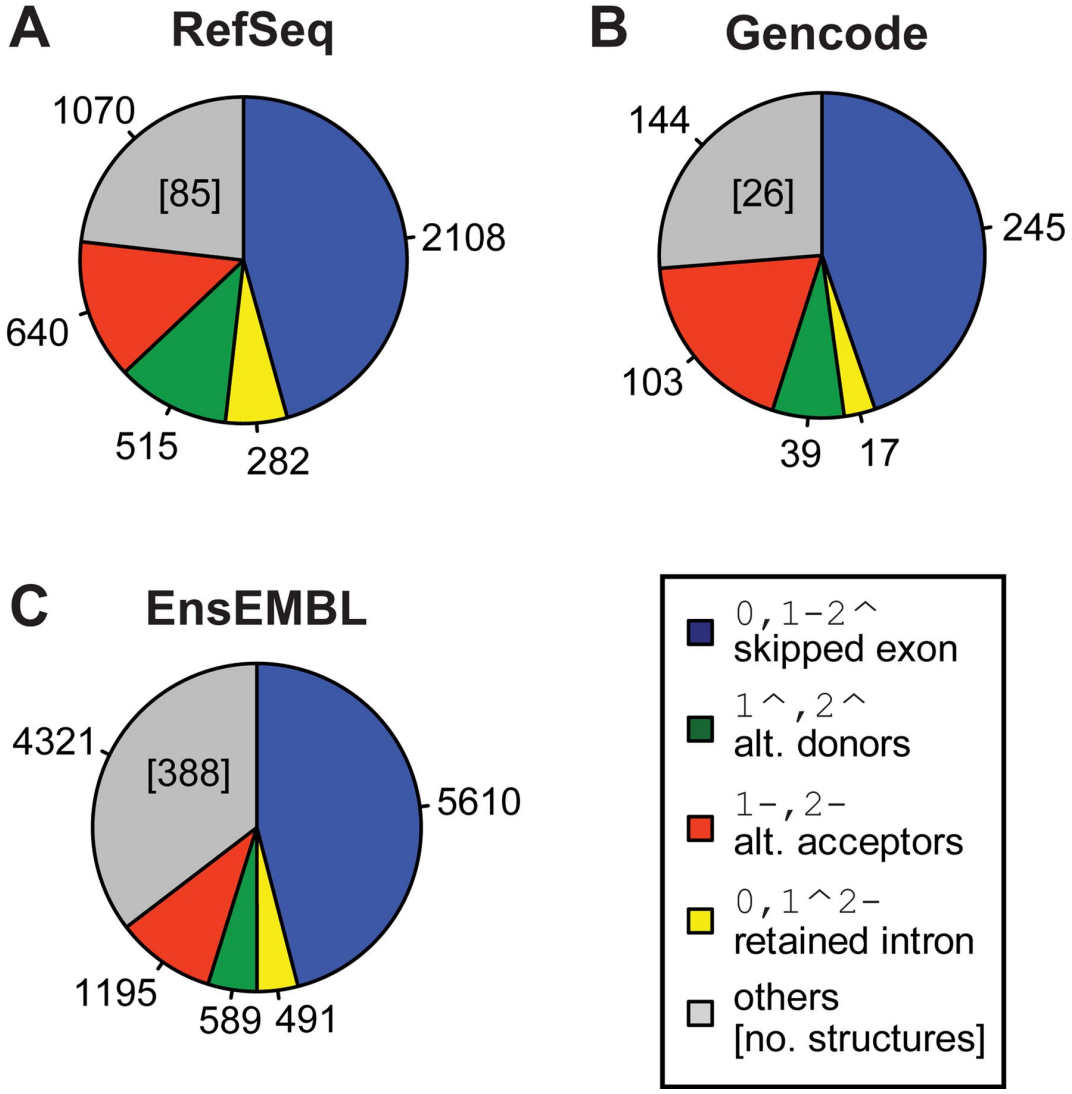
Lee Y, Rio DC. 2015.
Annu. Rev. Biochem. 84:291–323

Comparative genomics of the AS landscape in 12 metazoa



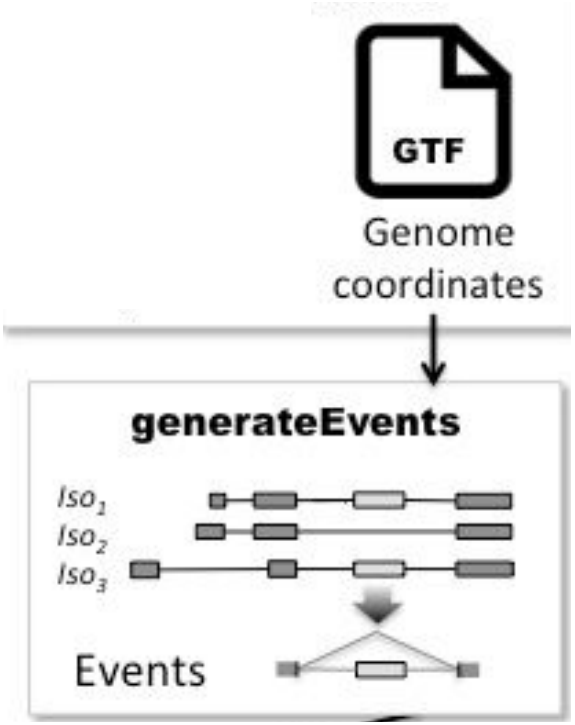
Sammeth, Foissac , Guigó (2008) PLoS Comput Biol 4(8): e1000147

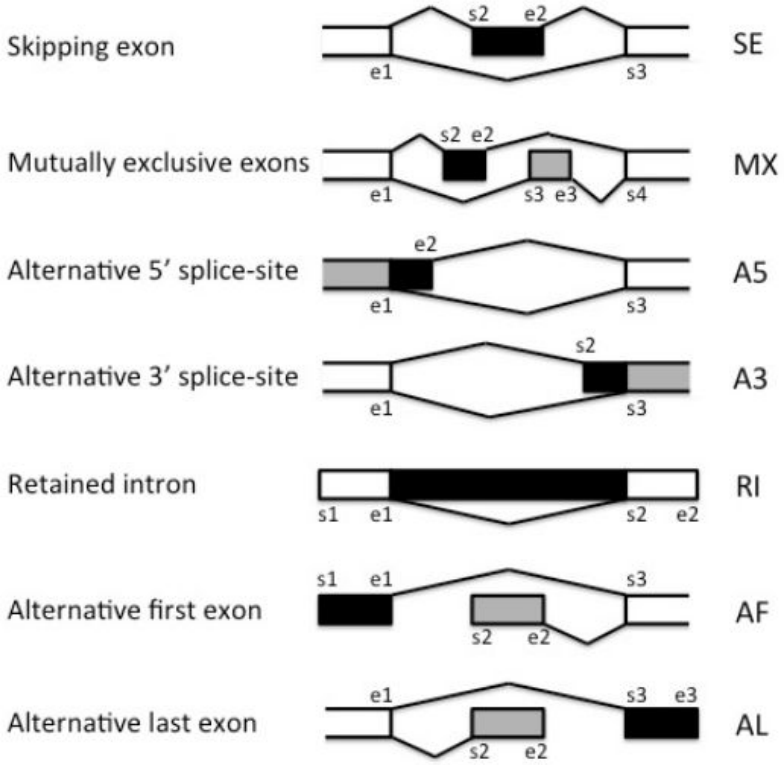
AS landscape in human reference annotations



Sammeth, Foissac , Guigó (2008) PLoS Comput Biol 4(8): e1000147

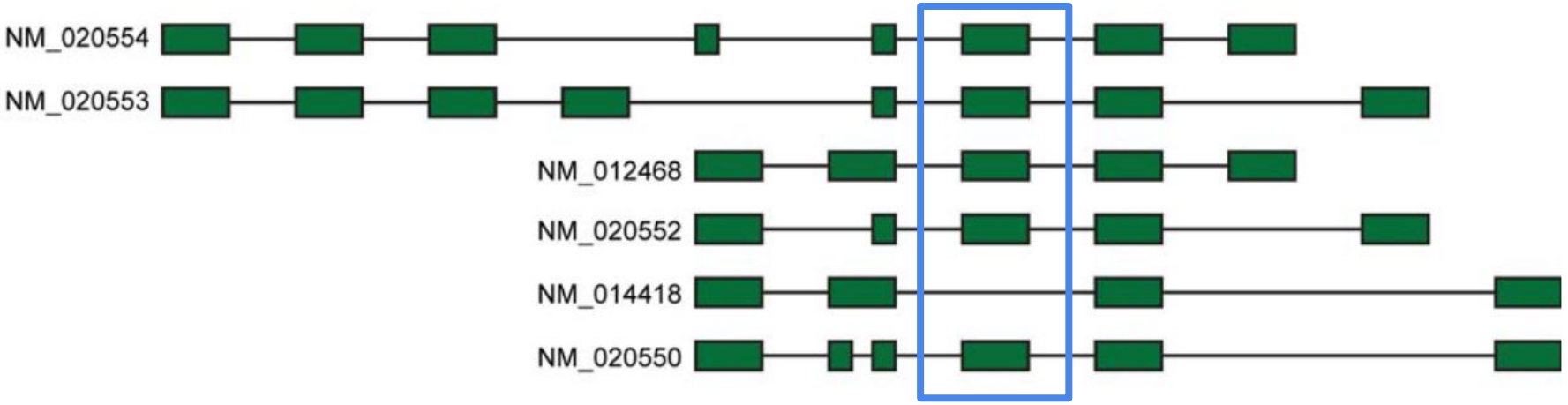
SUPPA: generate events based on gene annotation



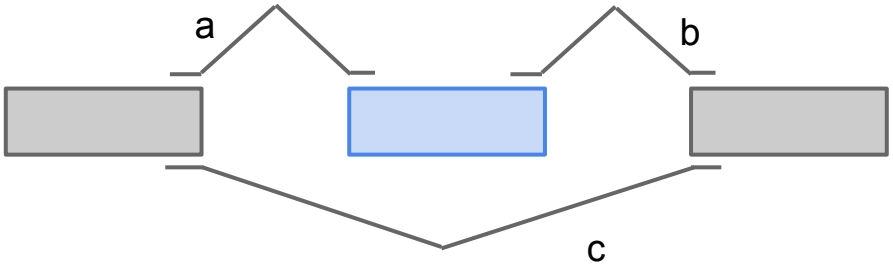


https://bitbucket.org/regulatorygenomicsupf/suppa

Alternative Splicing (AS)



PSI = percent-spliced-in = the number of transcripts in which the given exon is included as a fraction of the number of transcripts in which it is included or excluded

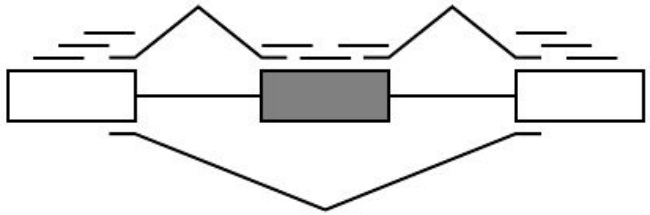


$$PSI = \frac{a+b}{a+b+2c}$$

More than one way to define PSI

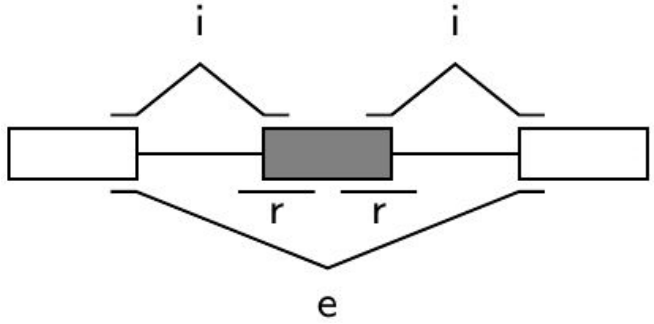
PSI = Percent-Spliced-In

Transcript-centric



$$\Psi = \frac{t_i}{t_i + t_e}$$

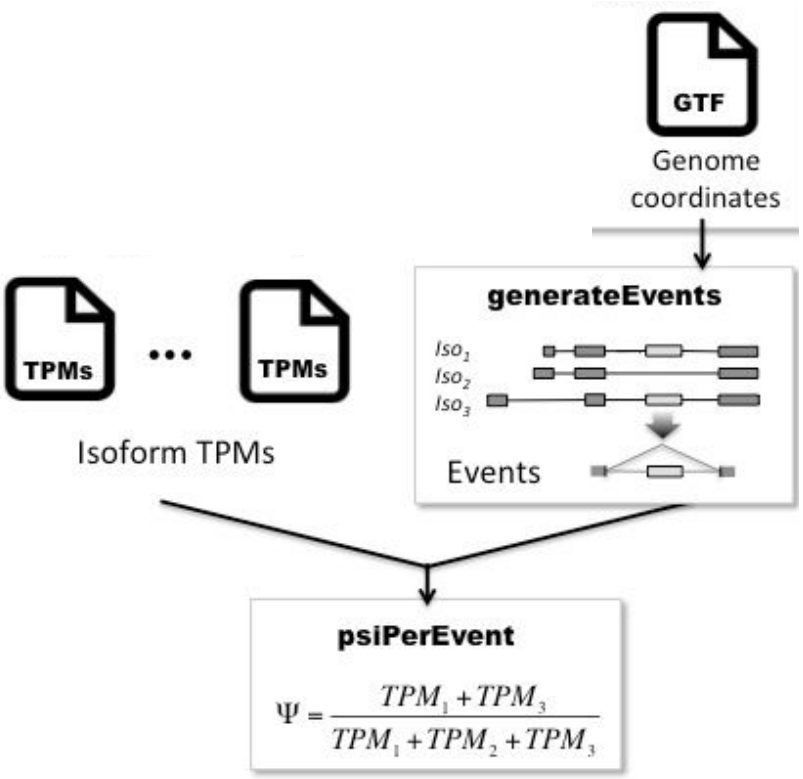
Exon-centric



$$\Psi = \frac{i}{i + e}$$

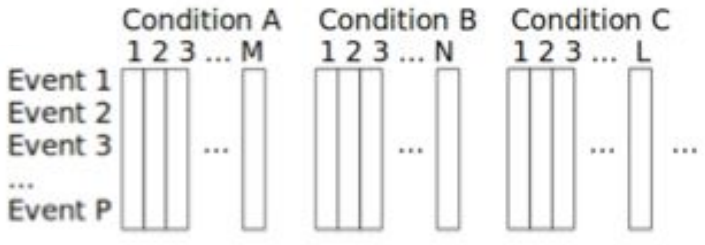
i = inclusion
e = exclusion
r = retention

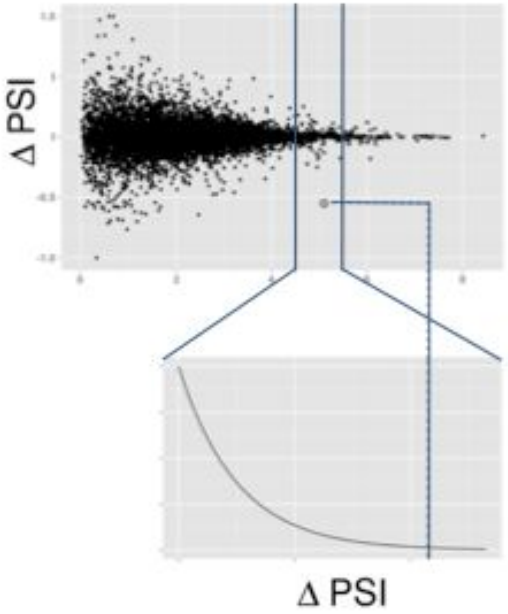
SUPPA: Quantify event inclusion levels (PSIs)



https://bitbucket.org/regulatorygenomicsupf/suppa

SUPPA: compare conditions





- SUPPA calculates the magnitude of splicing change (ΔPSI) and their significance across multiple biological conditions, using two or more replicates per condition.

- Statistical significance is calculated by comparing the observed ΔPSI between conditions with the distribution of the ΔPSI between replicates as a function of the gene expression (measured as the expression of the transcripts defining the events).

https://bitbucket.org/regulatorygenomicsupf/suppa

Hands-on

Setup environment 1

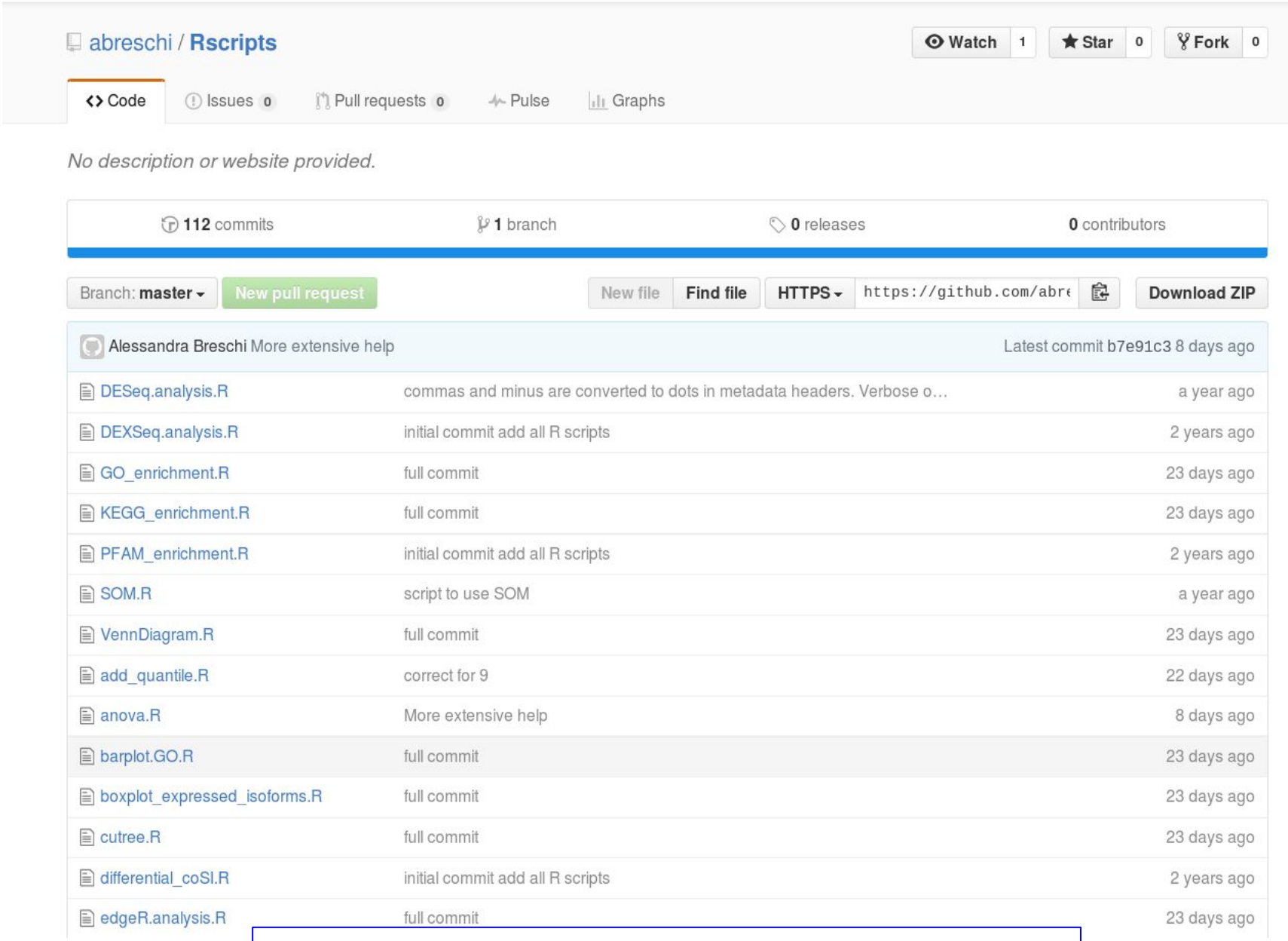
RNA-seq data analysis 4

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Hands-on

- Forebrain, heart and liver of 12.5 days mouse embryos
  - 2 bio replicates
  - RNA-seq, ChIP-seq and ATAC-seq
- References:
  - mouse genome – mm10 assembly
  - gene annotation – gencode vM4
- Processing:
  - References: a small sample of the genome and annotation (21 chromosomes, 1Mb long)
  - Data: one sample only (100,000 alignment-based pre-filtered reads)
- Analysis:
  - all samples

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/



https://github.com/abreschi/Rscripts

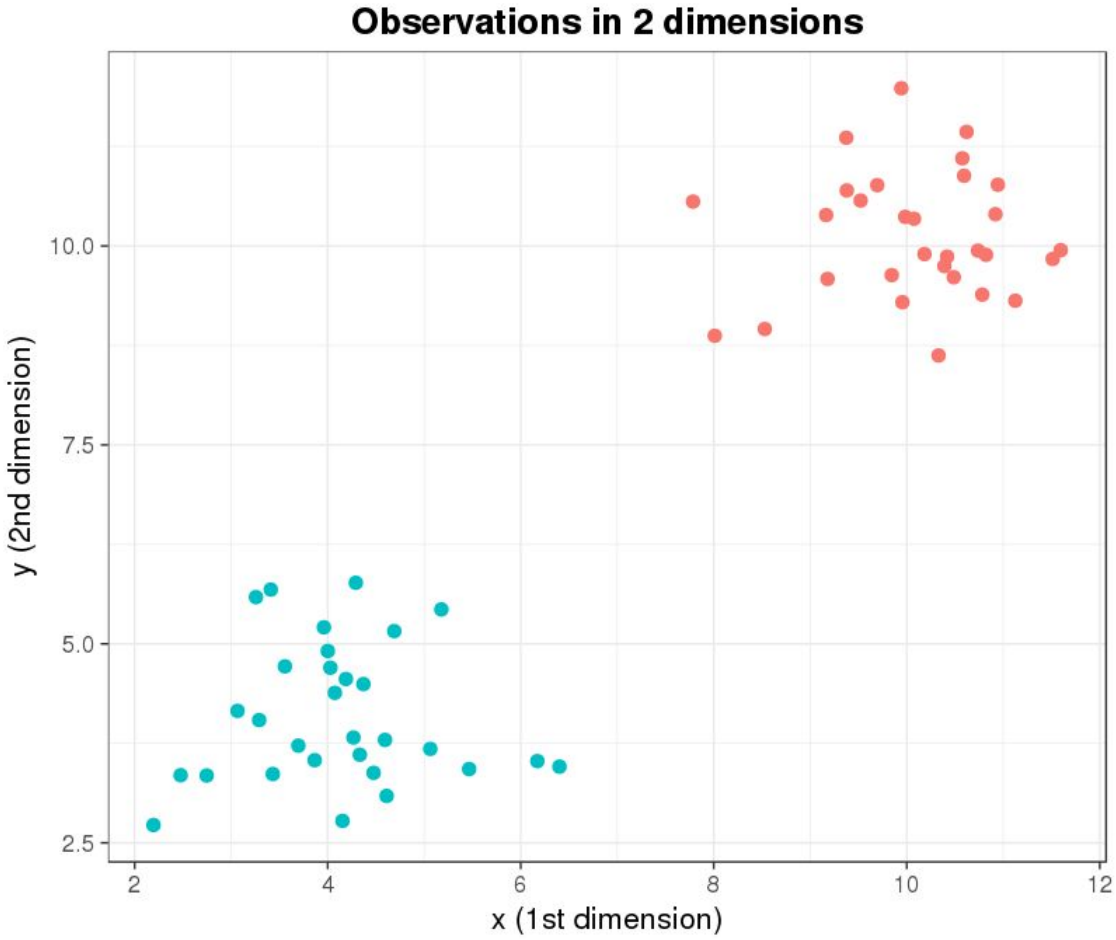
--help
will provide input/output parameters

```
Rscript rpkm_fraction.R --help

Usage: rpkm_fraction.R [options] file

Options:
    -i INPUT_MATRIX, --input_matrix=INPUT_MATRIX
            the matrix you want to analyze [default=stdin]
    -m METADATA, --metadata=METADATA
            tsv file with metadata on matrix experiment
    -o OUTPUT, --output=OUTPUT
            additional tags for otuput
    -c COLOR_BY, --color_by=COLOR_BY
            choose the color you want to color by. Leave empty for no color
    -y LINETYPE_BY, --linetype_by=LINETYPE_BY
            choose the factor you want the linetype by. Leave empty for no linetype
    -f FILE_SEL, --file_sel=FILE_SEL
            list of elements of which computing the proportion at each point
    --out_file=OUT_FILE
            store the coordinates in a file [default=NULL]
    -P PALETTE, --palette=PALETTE
            file with the colors
    -t TAGS, --tags=TAGS
            choose the factor by which grouping the lines [default=labExpId]
    -h, --help
            Show this help message and exit
```

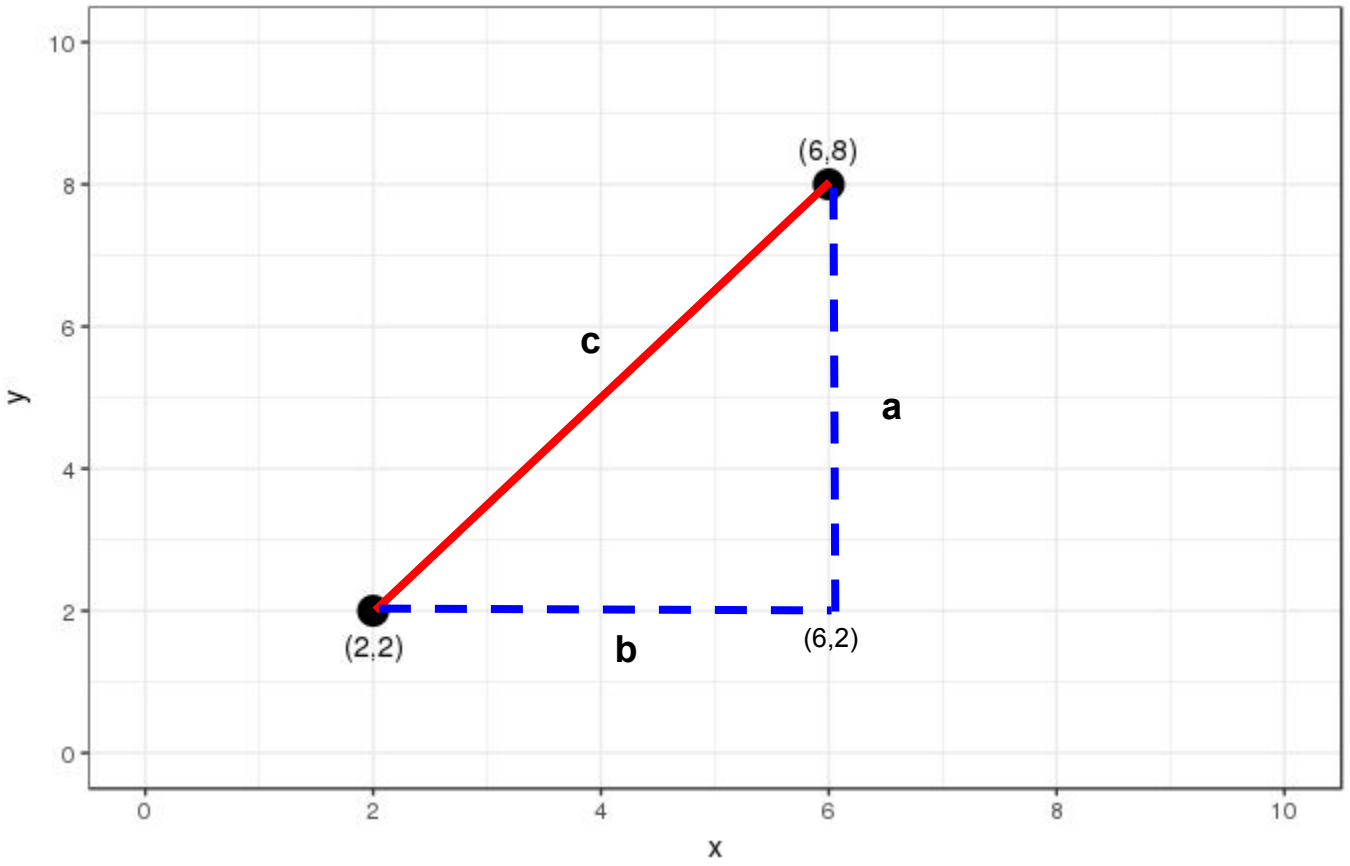
Additional slides

Understanding clustering: 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
4.33	3.61
10.74	9.94
...	...



Understanding clustering: a geometrical insight

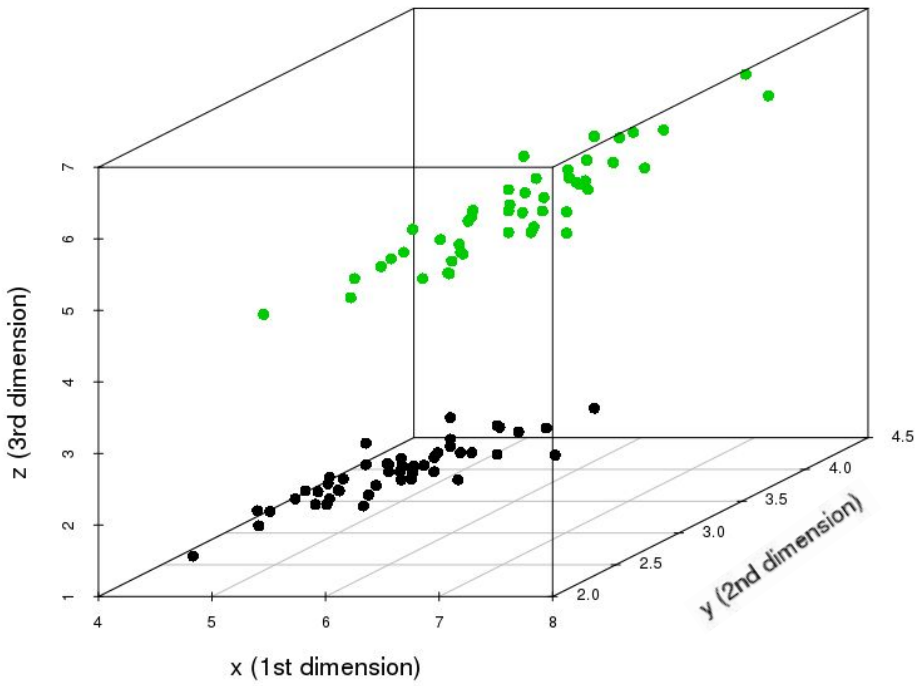


$$c^2 = a^2 + b^2 \rightarrow c = \sqrt{a^2 + b^2} \rightarrow c = \sqrt{(8-2)^2 + (6-2)^2}$$

Understanding clustering: a geometrical insight

Going one dimension higher ...

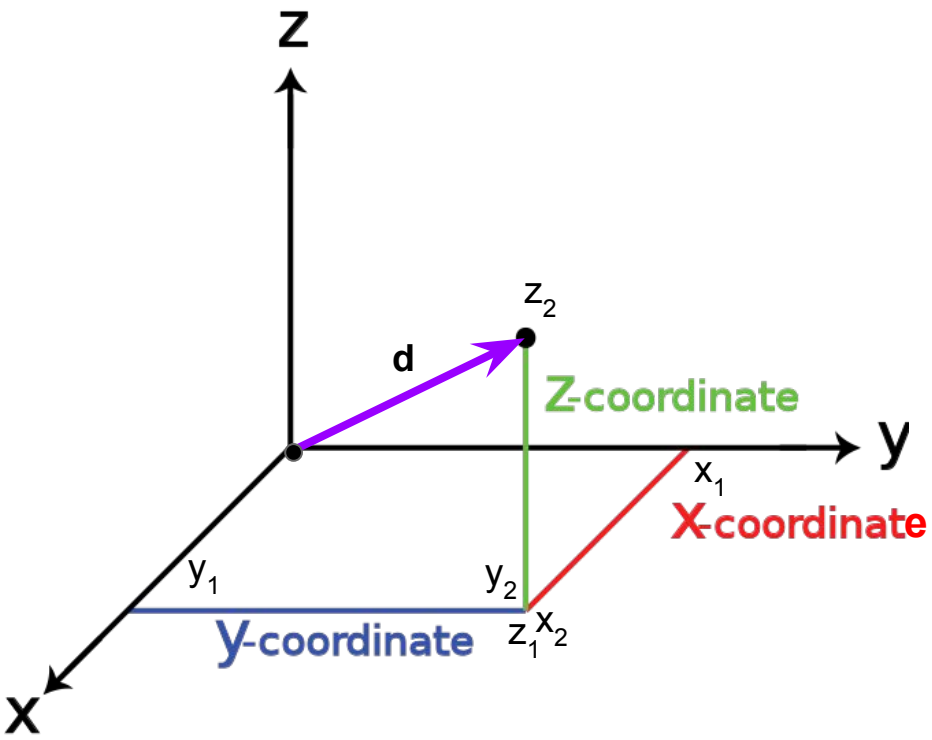
x	y	z
6.2	2.8	4.8
5.8	2.7	5.1
5.1	3.8	1.6
6.7	2.5	5.8
6.5	3.0	5.2
5.4	3.7	1.5
5.1	3.3	1.7
6.7	3.0	5.2
...	...	...



Understanding clustering: a geometrical insight

Going one dimension higher ...

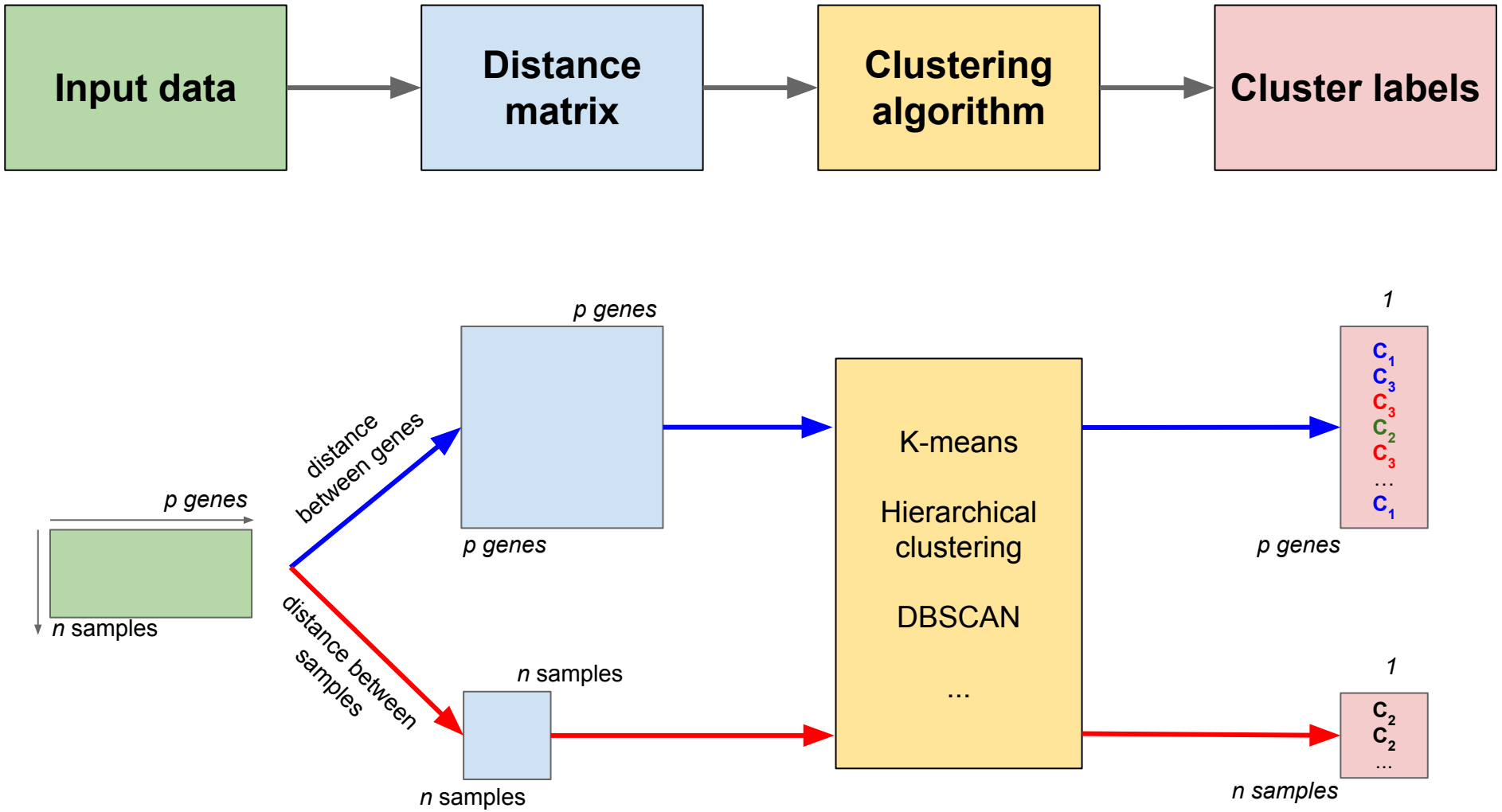
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$



Understanding clustering: a geometrical insight

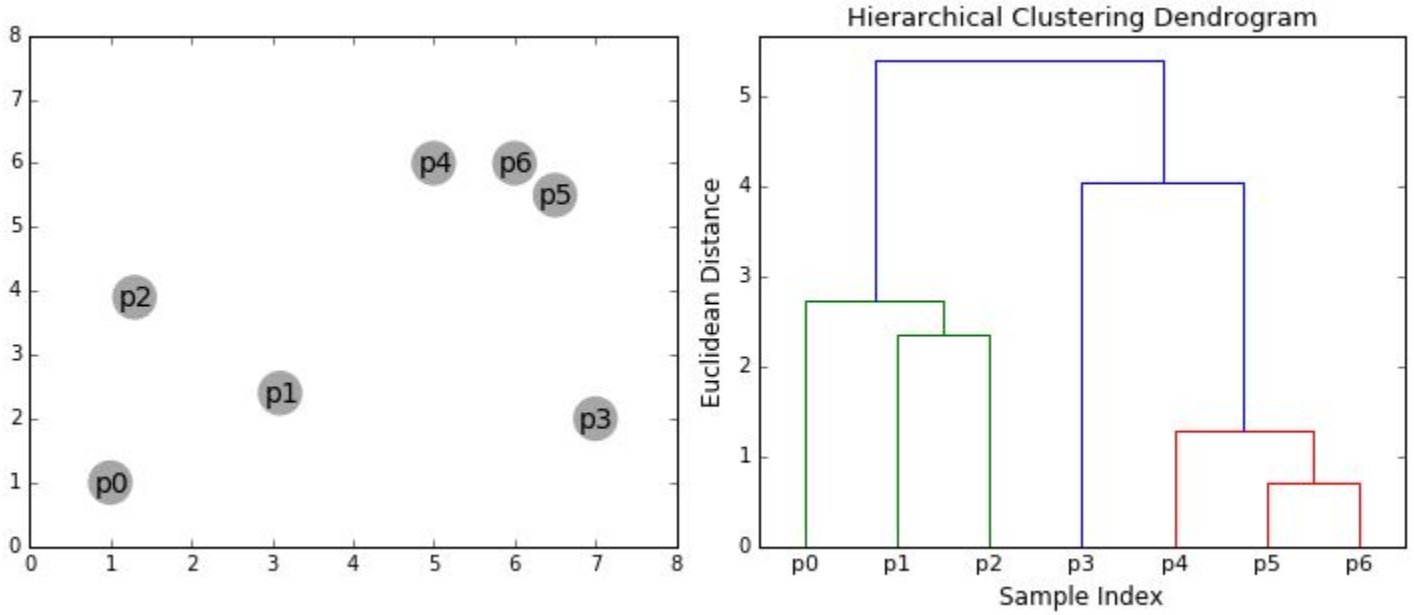
If we go up to 4, 5, …, *n*-dimensional space, how can we know which points (observations) are close to each other?

Clustering overview



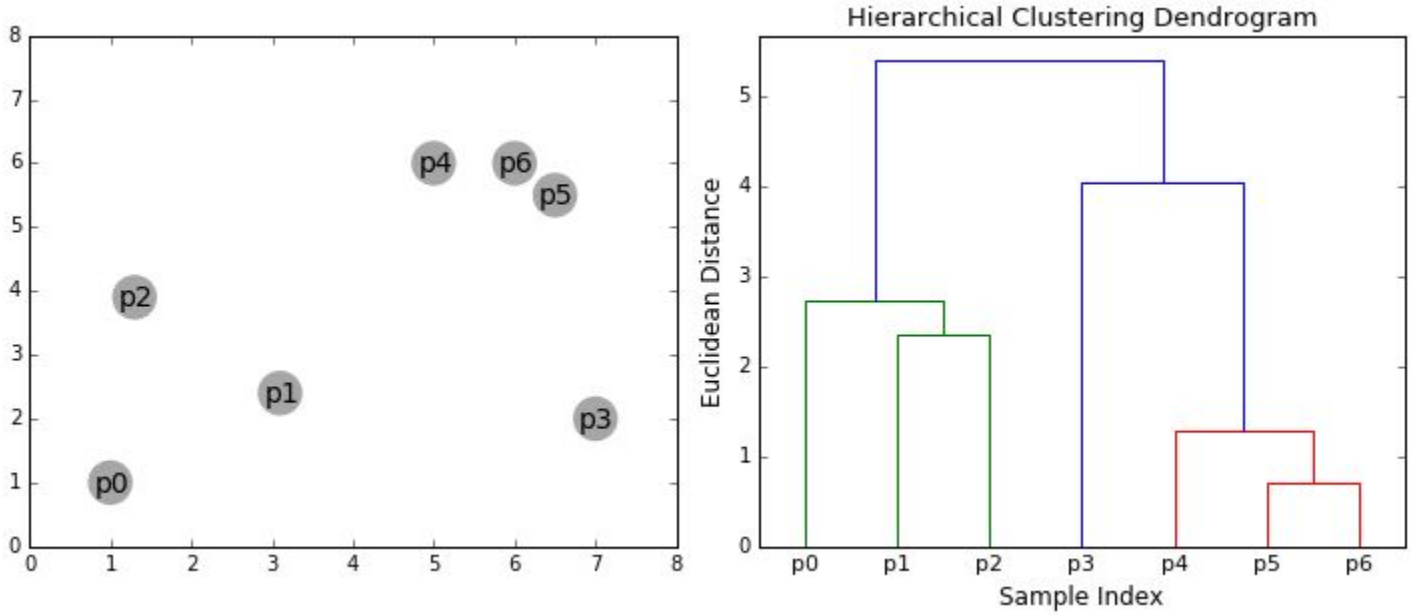
Clustering methods

Hierarchical clustering: Seeks to build a hierarchy of clusters. To generate the "class label" for each sample, we cut the tree at a certain height.



Clustering methods

Hierarchical clustering: Seeks to build a hierarchy of clusters. To generate the "class label" for each sample, we cut the tree at a certain height.



Clustering methods

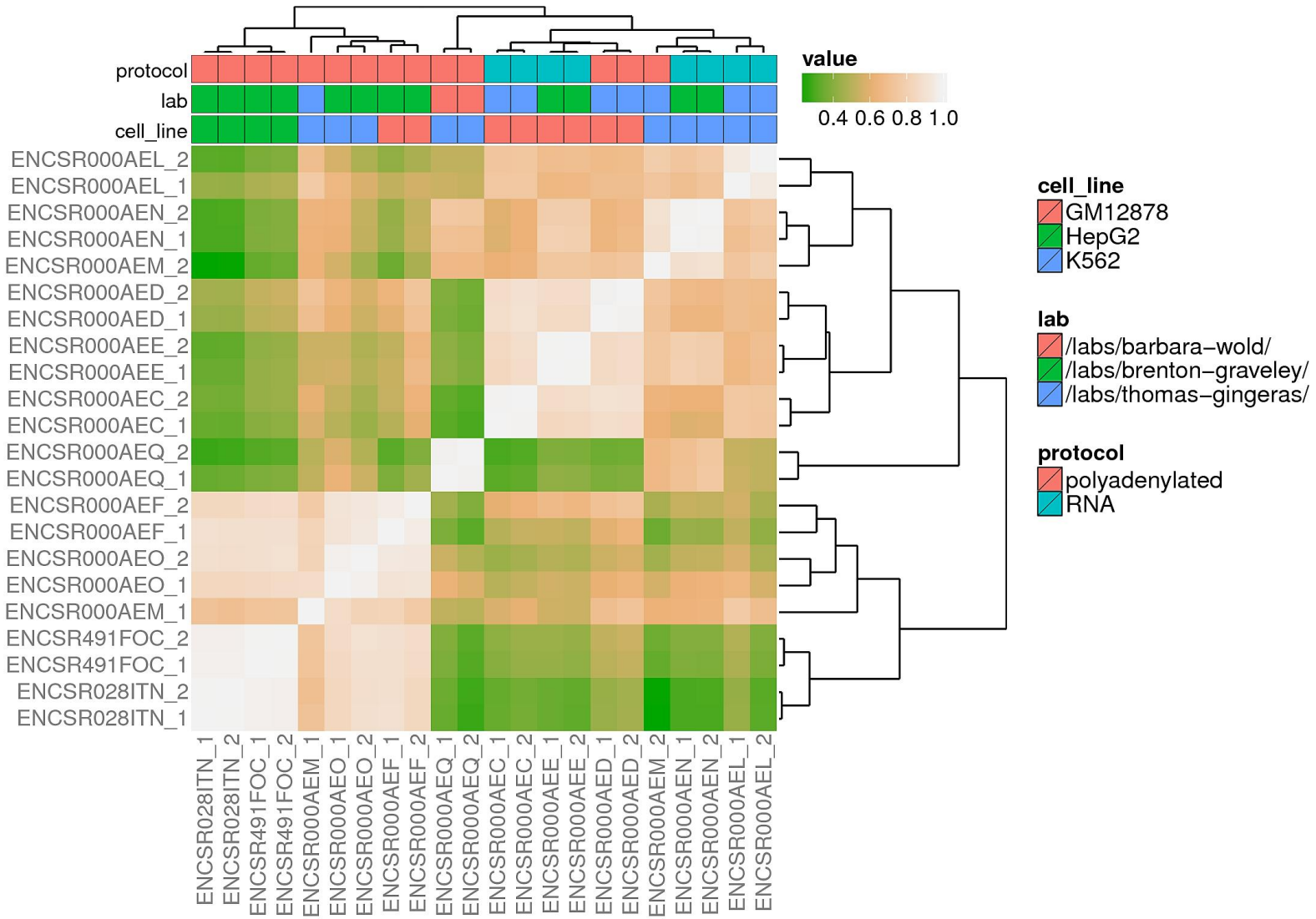
k-means: Partitions *n* observations into *k* clusters. Each observation will be assigned to the cluster with the nearest mean.

Clustering methods

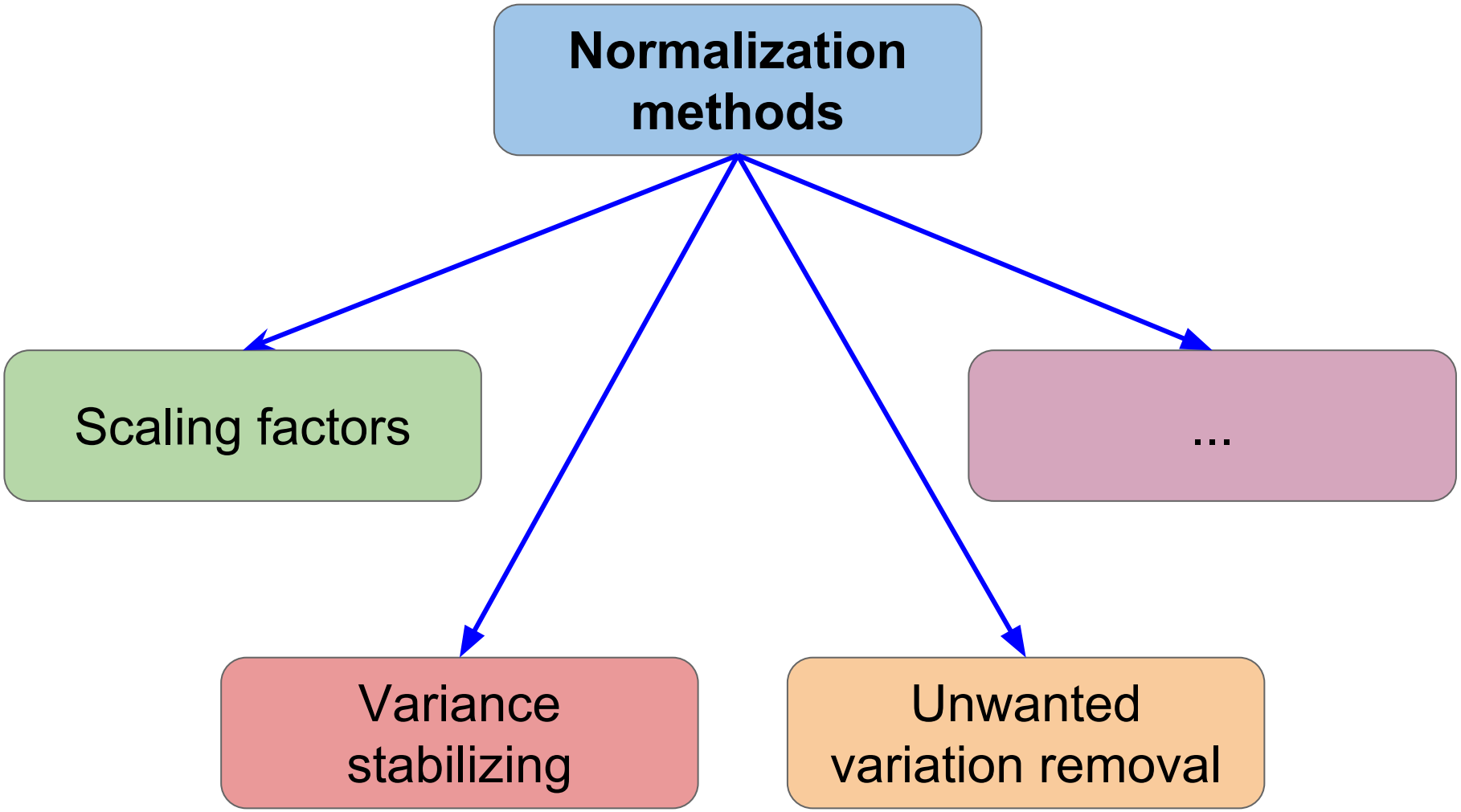
k-means: Partitions *n* observations into *k* clusters. Each observation will be assigned to the cluster with the nearest mean.

Animation adapted from: http://shabal.in/visuals/kmeans/1.html

Samples clustering

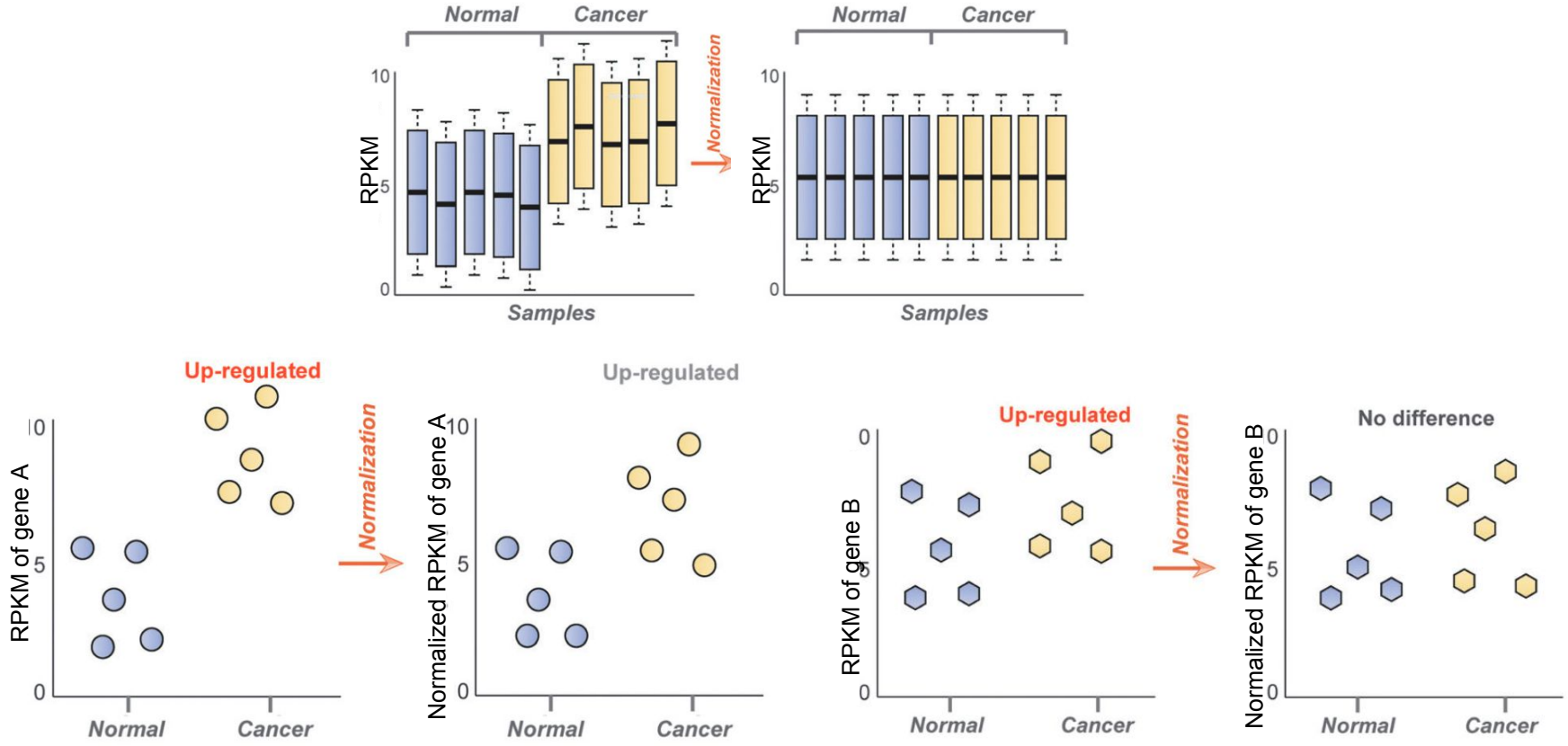


Normalization methods



Normalization methods

A) Scaling factors

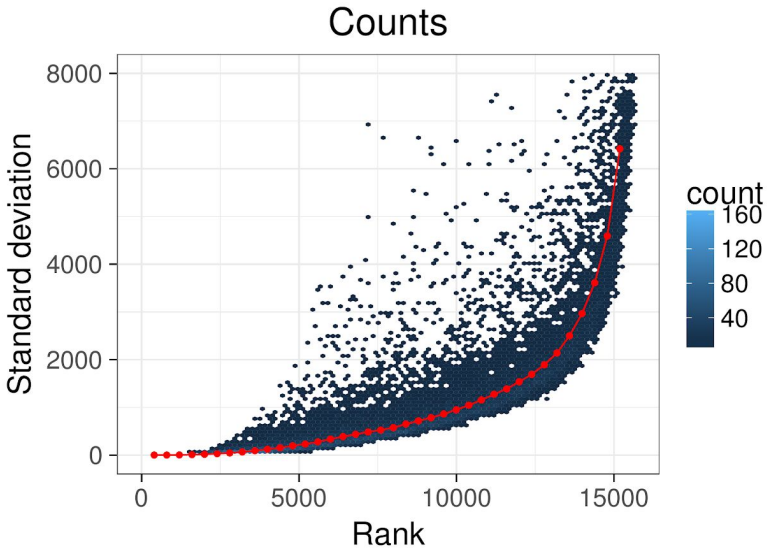


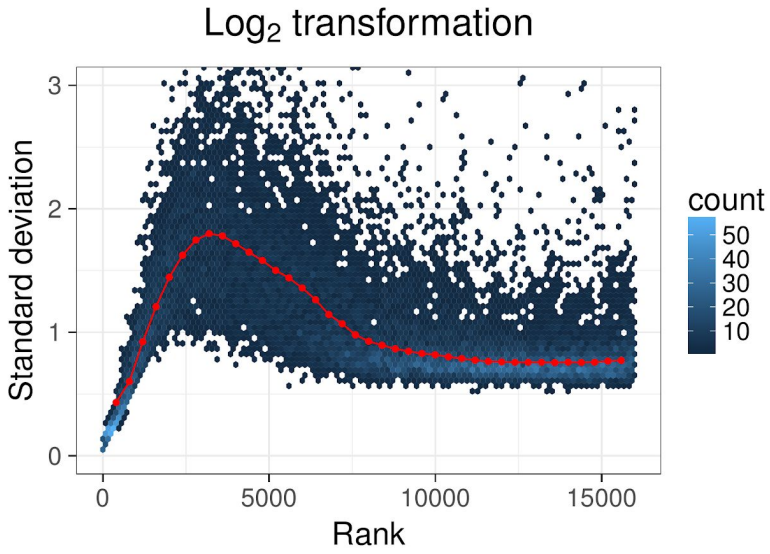
Adapted from: Wu et al (2014). Deciphering global signal features of high-throughput array data from cancers. Molecular Biosystems.

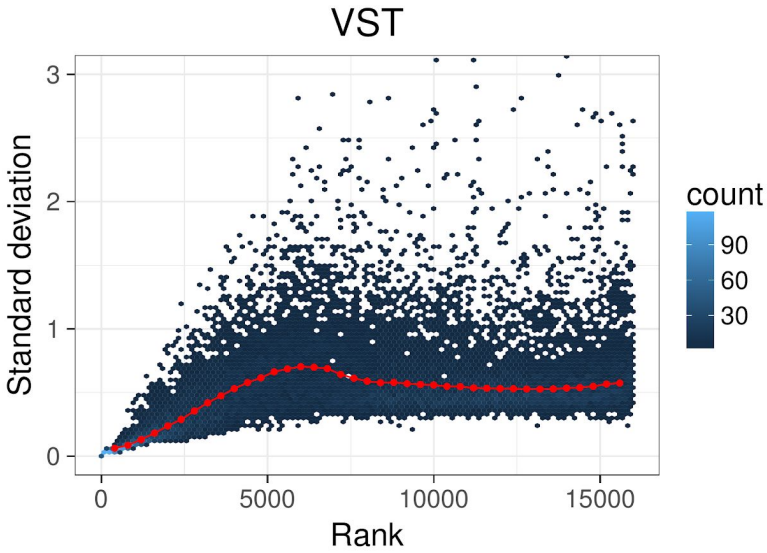
Methods: quantile normalization, trimmed mean of M-values (TMM, used by edgeR), DESeq

Normalization methods

B) Variance stabilizing







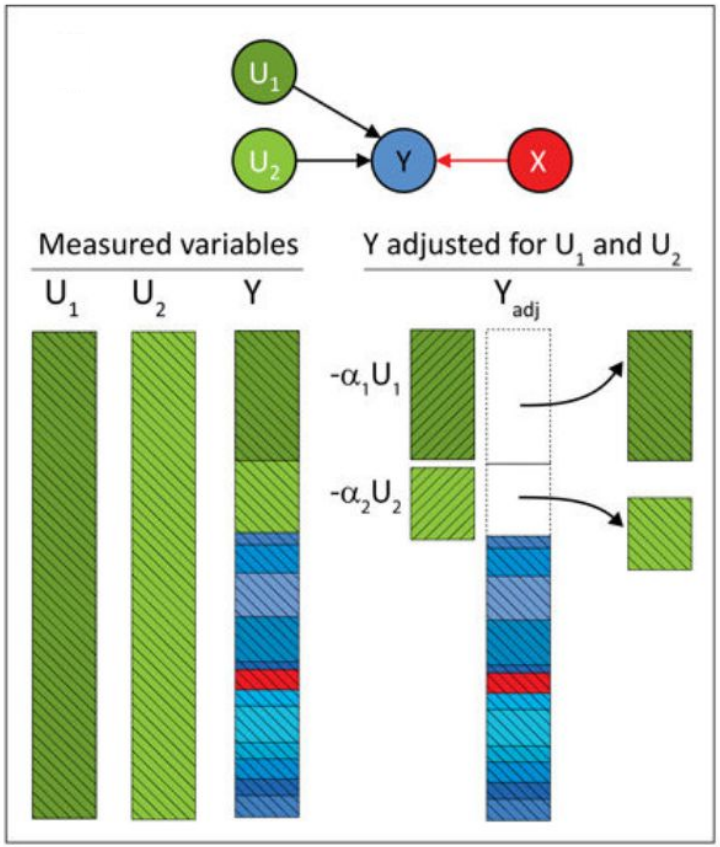
Increasing expression

Mainly for visualization and ML purposes

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

Normalization methods

C) Unwanted variation removal

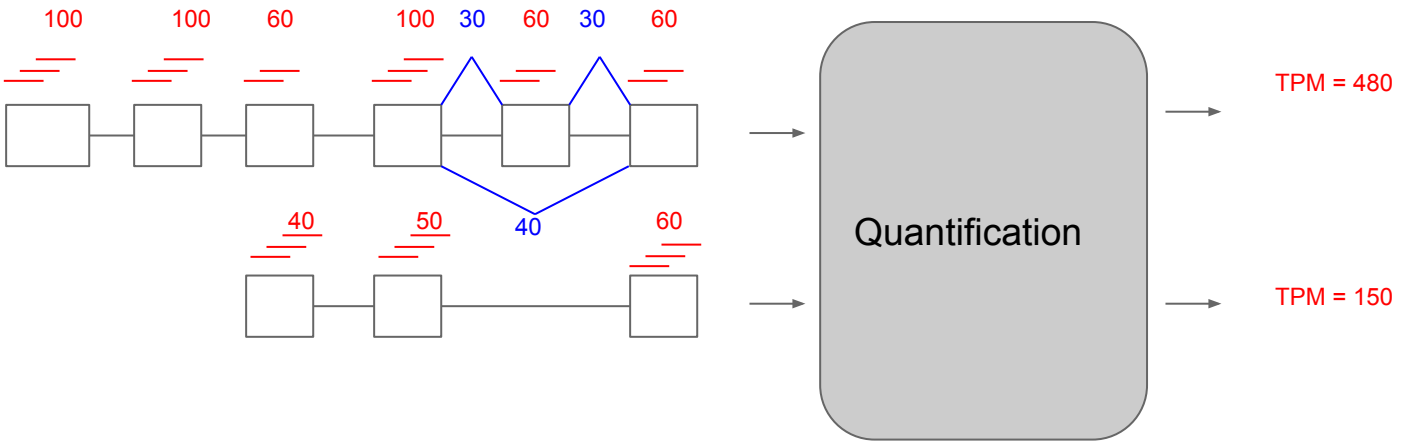


Adapted from: Nat Genet. 2017 December; 49(12): 1789–1795. doi:10.1038/ng.3975

Methods: PEER, RUV, SVA

More than one way to define PSI

PSI = Percent-Spliced-In



PSI = (30+30)/(30+30+80)=0.43

$$PSI = \frac{a+b}{a+b+2c}$$

PSI-tx = 480/(480+150)=0.76

Look at the gene expression distribution

To spot possible biases, detect outliers and assess the similarity among samples

- Look at the RPKM/FPKM/TPM distribution for individual samples (min, max, mean, median)
- Compare distributions among samples
- Look at the samples clustering