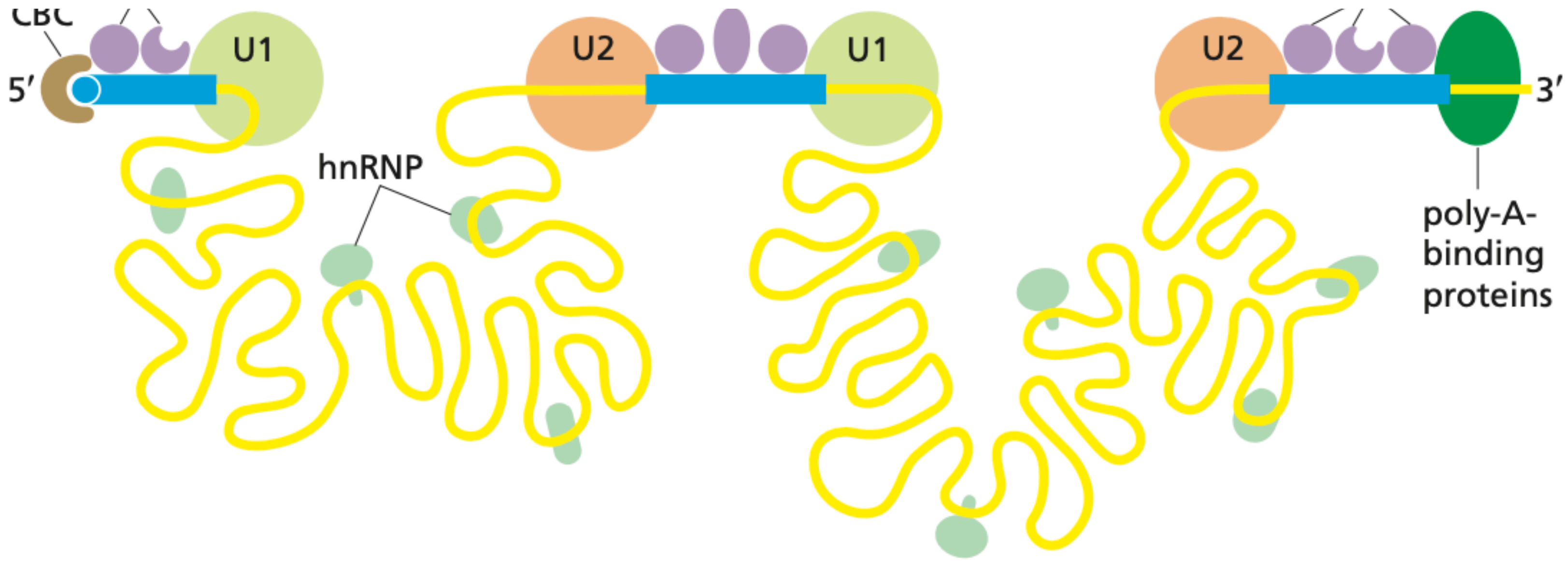
# RNA sequencing

2025-03-25

C.G.B.

UNIVERSITÀ DEGLI STUDI DI MILANO

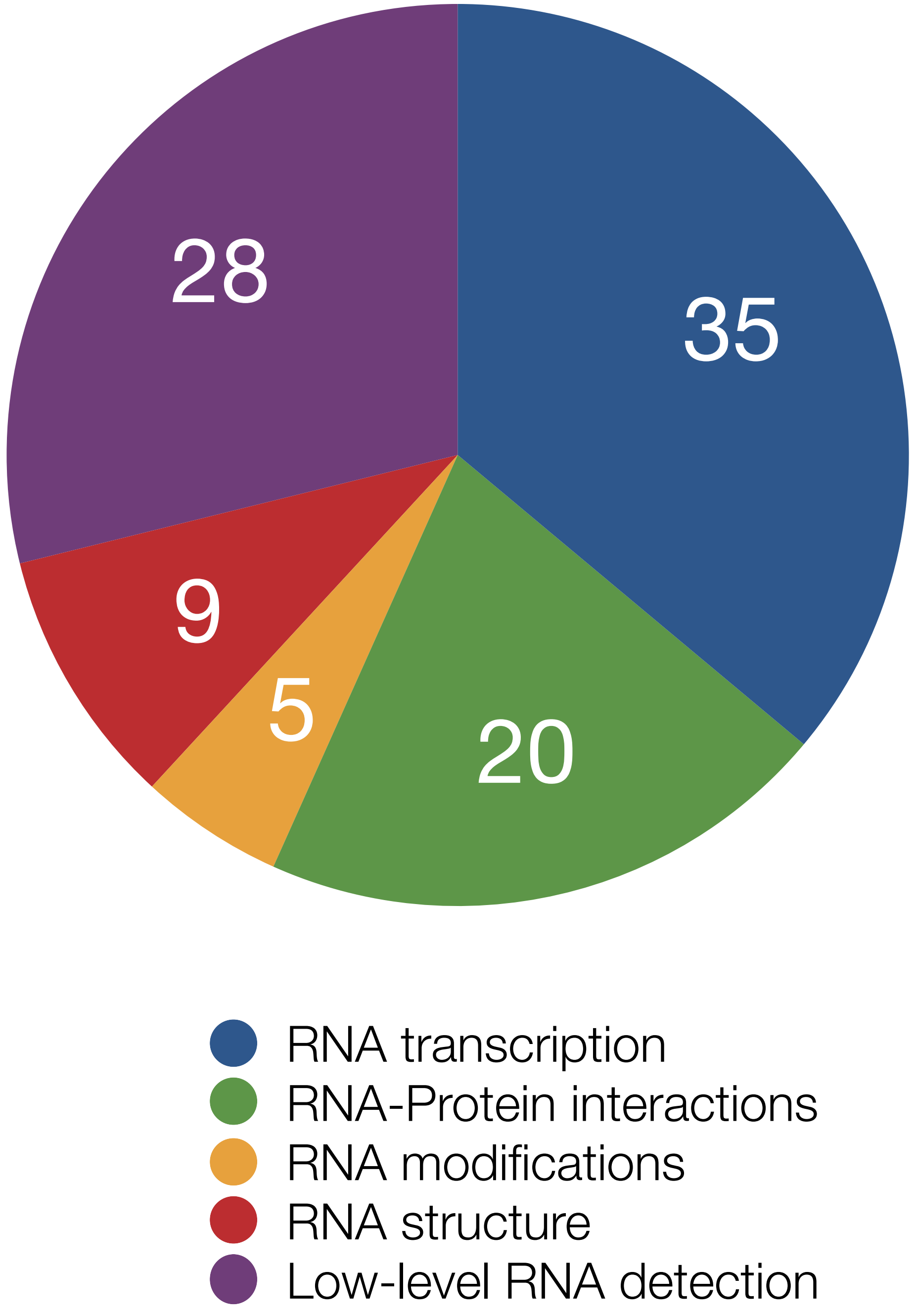
# Sequencing transcriptome

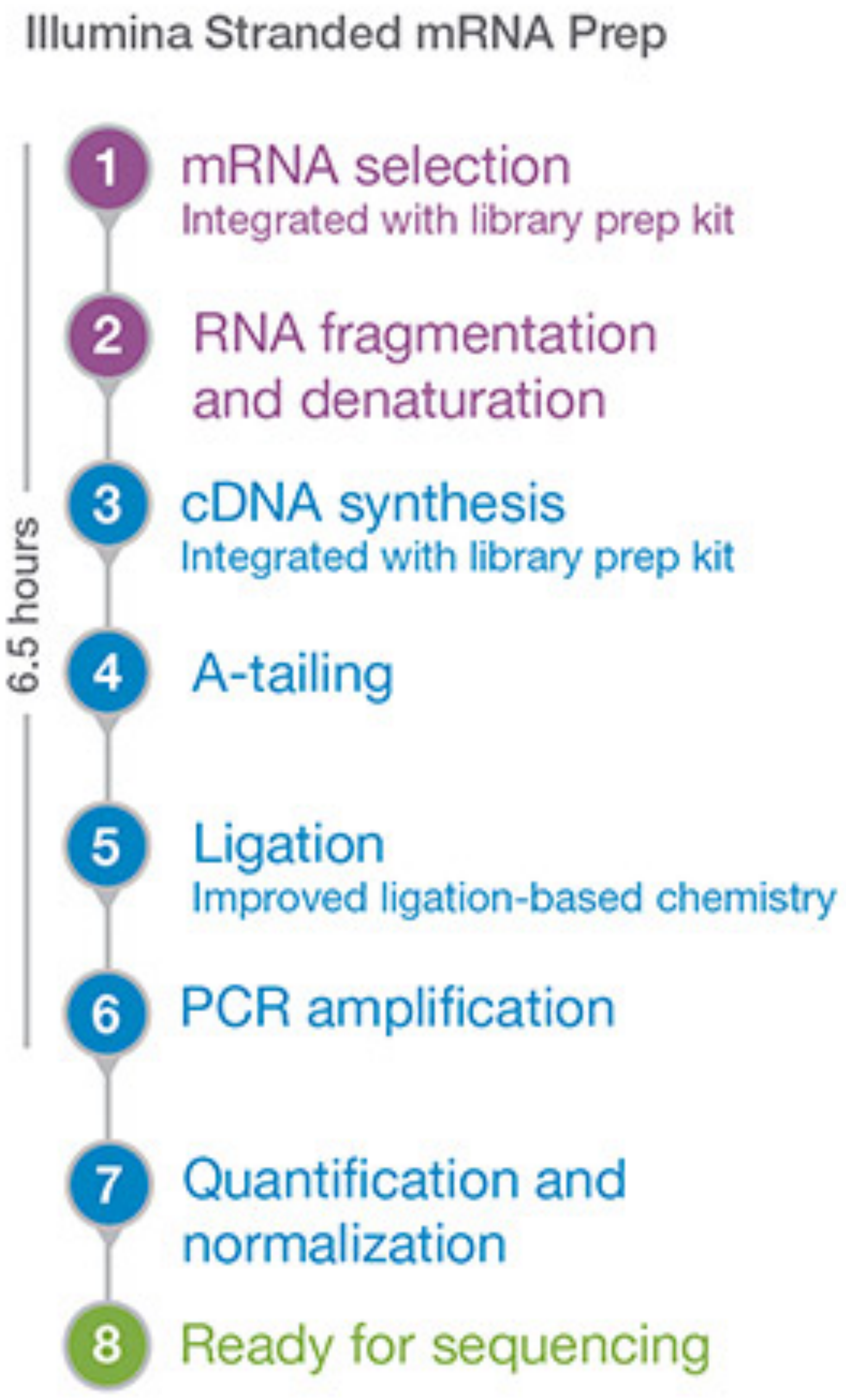


- Evaluate **expression** of genes/transcripts for:
  - All species of RNA
  - mRNA
  - small RNAs
- Evaluate expression levels of exons
  - Patterns of alternative splicing
- Evaluate transcriptional **alterations**
- Annotate **regions** and **functional elements**

| RNA Transcription | RNA-Protein Interactions | RNA Modifications | RNA Structure | Low-Level RNA Detection |
|---|---|---|---|---|
| RNA-Seq | Ribo-Seq | MeRIP-Seq | SHAPE-Seq | scRNA-Seq |
| CaptureSeq | RIP-Seq | miCLIP-m6A | icSHAPE | SUPeR-Seq |
| RASL-Seq | CLIP-Seq | PSI-Seq | CIRS-Seq | UMI |
| ClickSeq | Pol II CLIP | Pseudo-Seq | SHAPE-MaP | Digital RNA Sequencing |
| 3Seq | miR-CLIP | ICE | DMS-Seq | MARS-Seq |
| cP-RNA-Seq | eCLIP | | SPARE | Quartz-Seq |
| 3P-Seq | irCLIP | | PARS-Seq | DP-Seq |
| 2P-Seq | PAR-CLIP | | Cap-Seq | Smart-Seq |
| 3’-Seq | iCLIP | | CIP-TAP | FRISCR |
| TIF-Seq | BrdU-CLIP | | | CEL-Seq |
| PEAT | AGO-CLIP | | | STRT-Seq |
| SMORE-Seq | PIP-Seq | | | TCR Chain Pairing |
| TL-Seq | hiCLIP | | | TCR-LA-MC PCR |
| TATL-Seq | RBNS | | | CirSeq |
| RARseq | TRIBE | | | TIVA |
| TAIL-Seq | HiTS-RAP | | | PAIR |
| PAL-Seq | TRAP-Seq | | | CLaP |
| FRT-S wcell | DLAF | | | CytoSeq |
| ChIRP | miTRAP | | | Drop-Seq: |
| CHART | CLASH | | | Hi-SCL |
| RAP | | | | InDrop |
| GRO-seq | | | | snRNA-Seq |
| Bru-Seq | | | | Nuc-Seq |
| BruChase-Seq | | | | Div-Seq |
| 5’-GRO-Seq | | | | SCRB-Seq |
| BruDRB-Seq | | | | G&T-Seq |
| 4sUDRB-Seq | | | | scM&T-Seq |
| PRO-Seq | | | | scTrio-seq |
| PRO-Cap | | | | |
| CAGE | | | | |
| 3’NT Method | | | | |
| NET-Seq | | | | |
| mNET-Seq | | | | |
| PARE-Seq | | | | |
| GMUCT | | | | |

# Protocolli di sequenziamento





Illumina Stranded mRNA Prep

6.5 hours

1 mRNA selection
Integrated with library prep kit

2 RNA fragmentation and denaturation

3 cDNA synthesis
Integrated with library prep kit

4 A-tailing

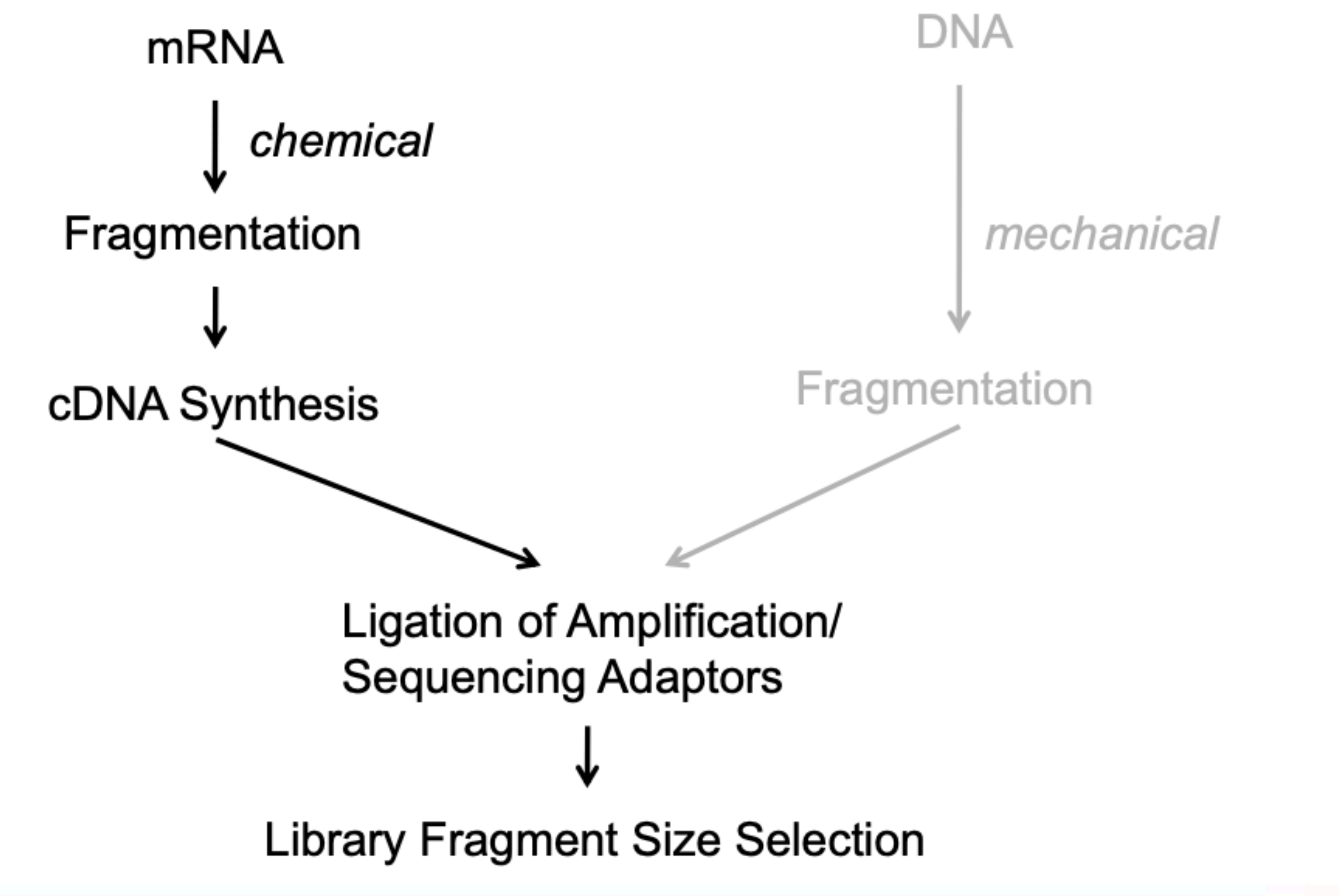
5 Ligation
Improved ligation-based chemistry

6 PCR amplification

7 Quantification and normalization

8 Ready for sequencing

# Key steps in sequencing



mRNA

*chemical*

Fragmentation

cDNA Synthesis

DNA

*mechanical*

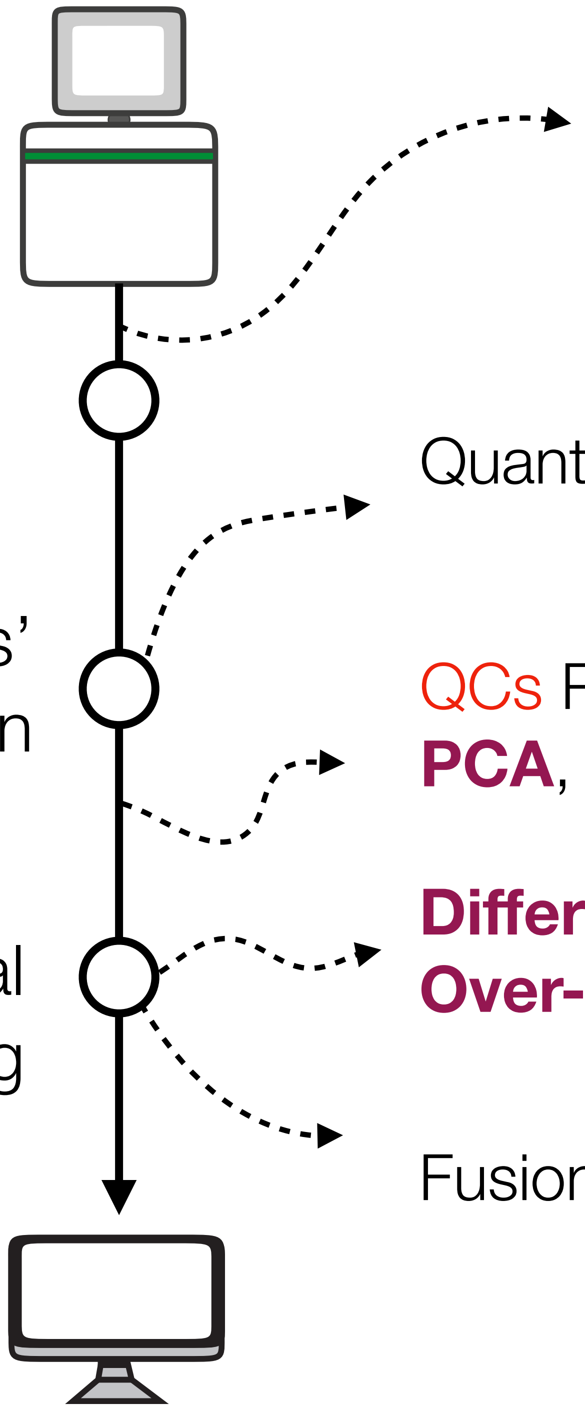
Fragmentation

Ligation of Amplification/
Sequencing Adaptors

Library Fragment Size Selection

# Deciphering gene expression

RNA-seq data analysis workflow:



**A**lignment

**T**ranscripts' quantification

**F**unctional profiling

QCs : sequence quality, GC content, duplicates

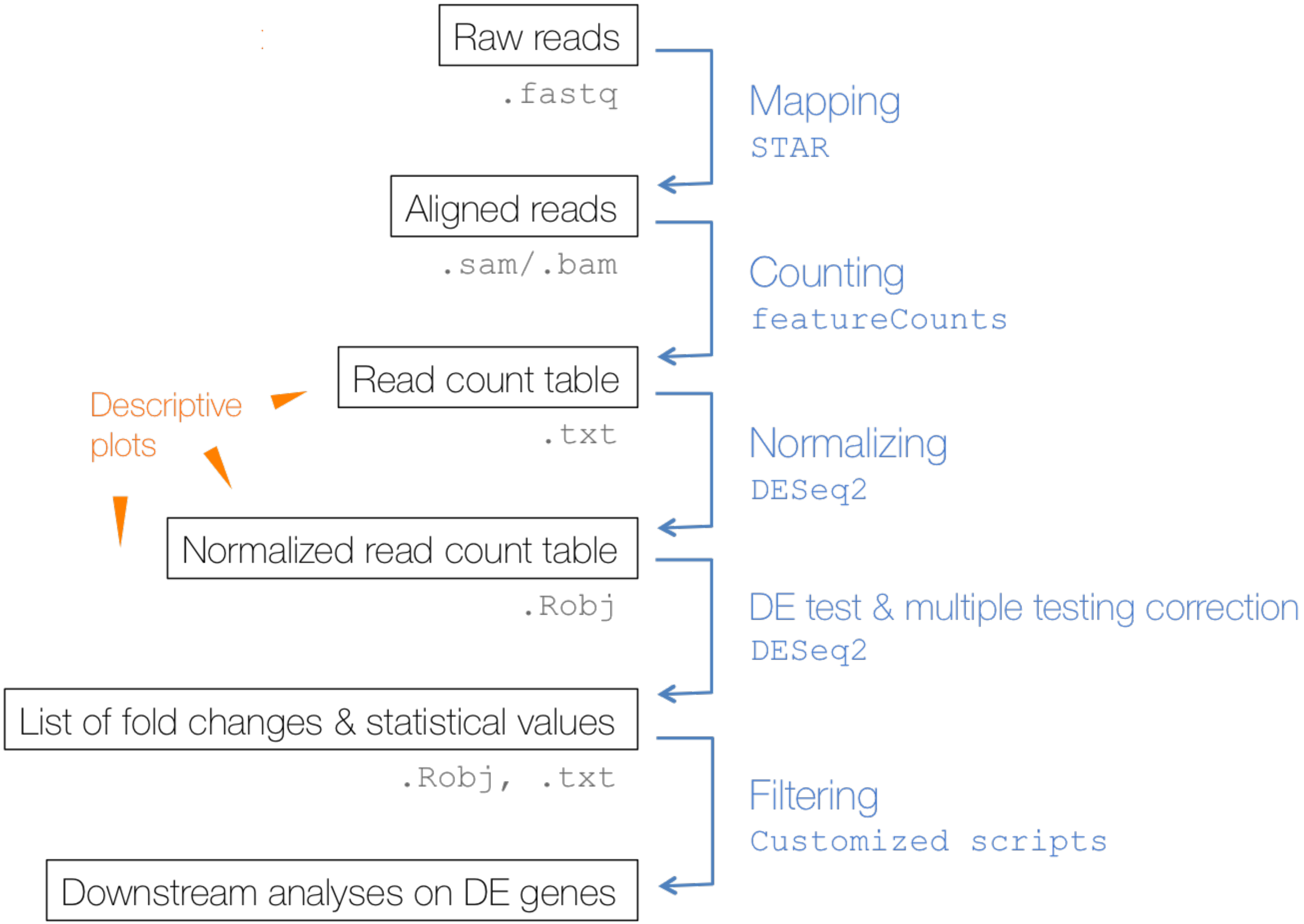
Quantification of gene/transcript expression

QCs Reproducibility : correlation of expression values, **PCA**, **batch effect correction**

**Differential Expression**, Alternative splicing analysis, **Over-representation Analyses/Gene Ontology**

Fusion identification

# Deciphering gene expression



Raw reads

.fastq

Mapping

STAR

Aligned reads

.sam/.bam

Counting

featureCounts

Read count table

.txt

Descriptive plots

Normalizing

DESeq2

Normalized read count table

.Robj

DE test & multiple testing correction

DESeq2

List of fold changes & statistical values

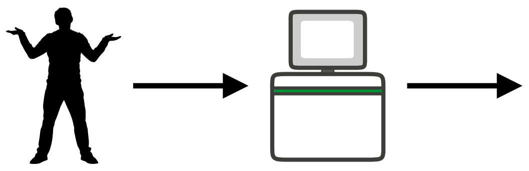
.Robj, .txt

Filtering

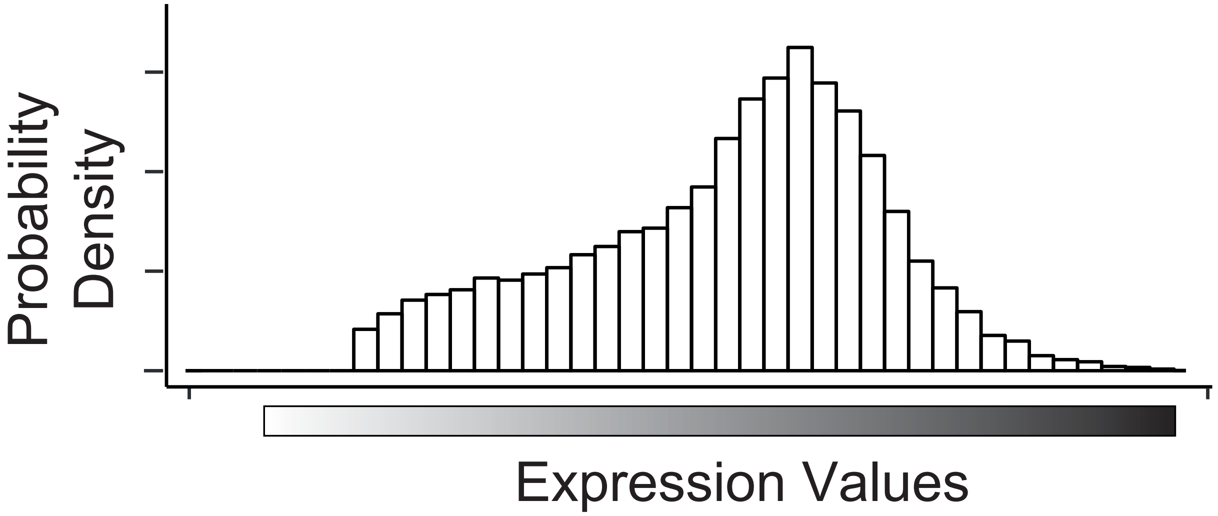
Customized scripts

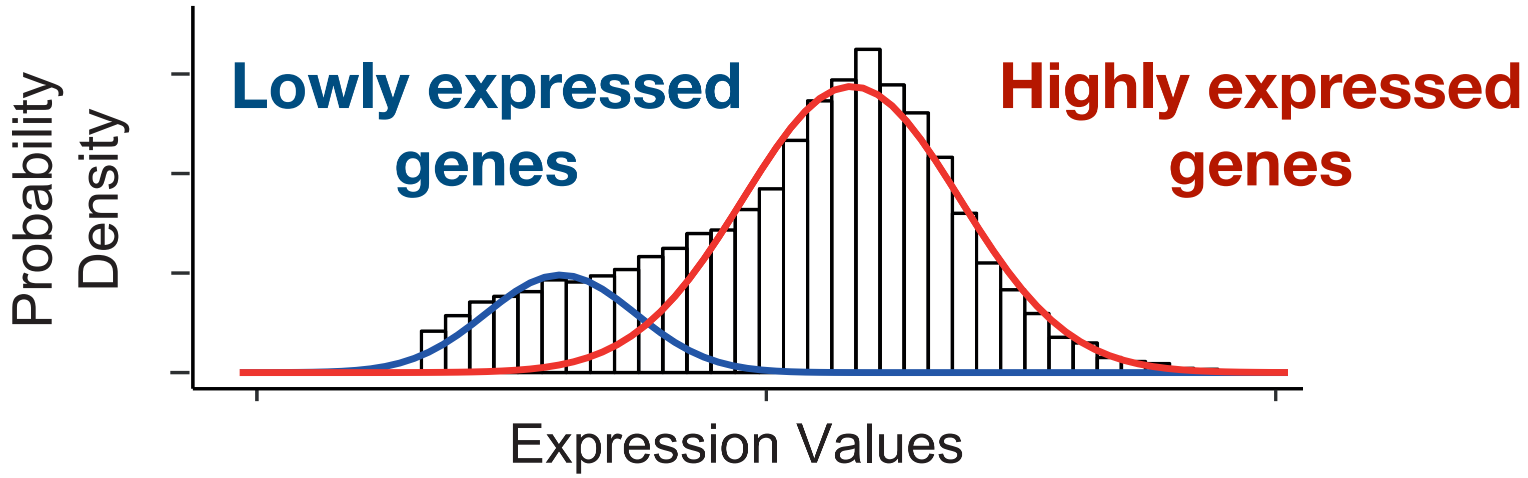
Downstream analyses on DE genes

# Gene expression level distribution



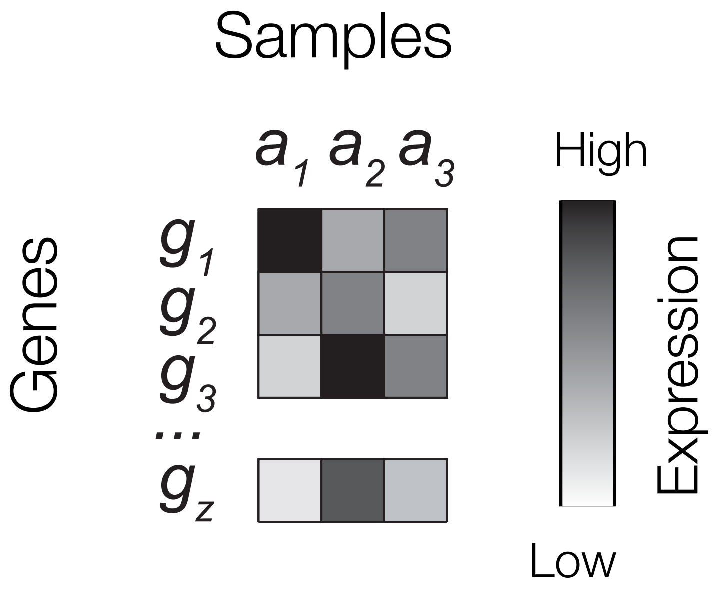
RNA-seq





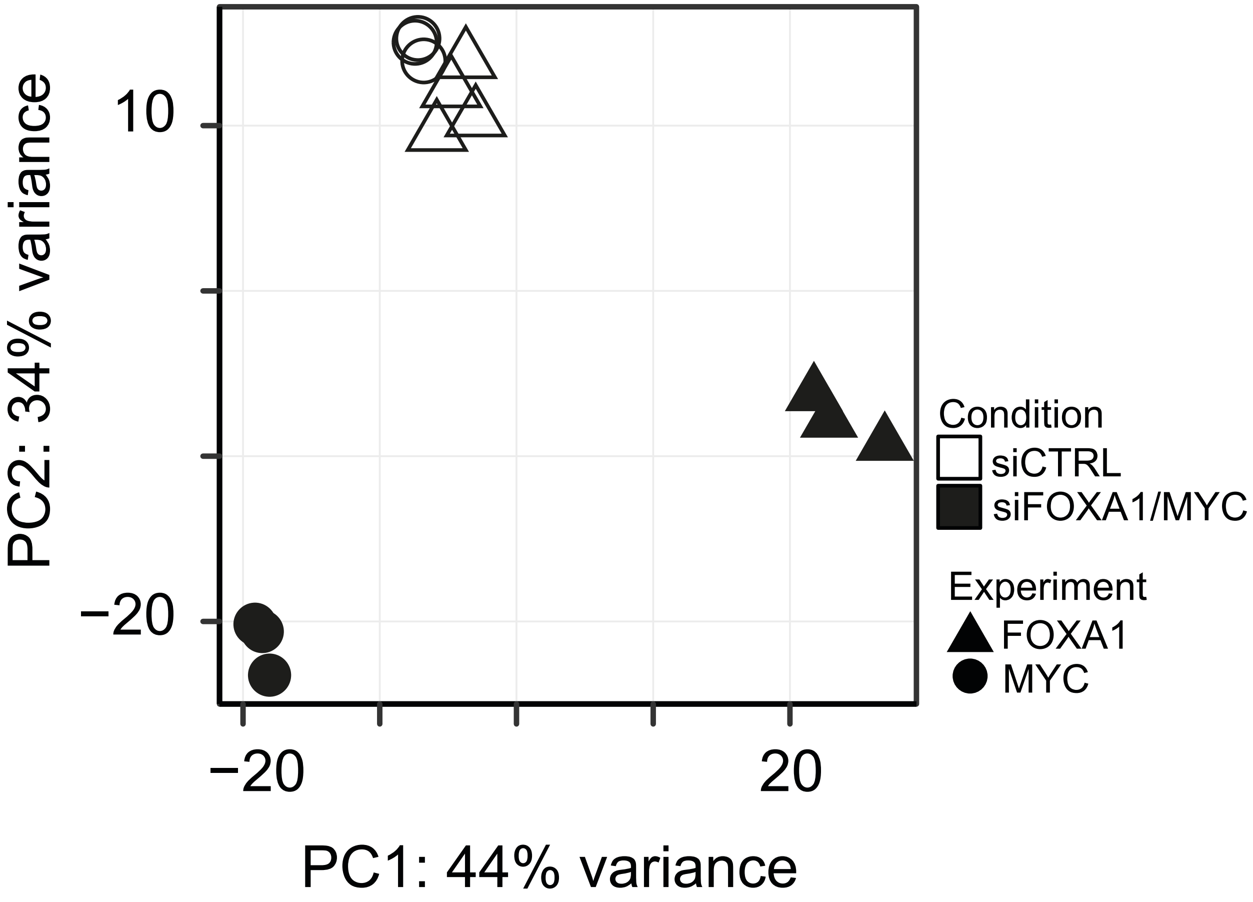
https://academic.oup.com/nar/article/48/4/1730/5691219

# Gene expression level distribution



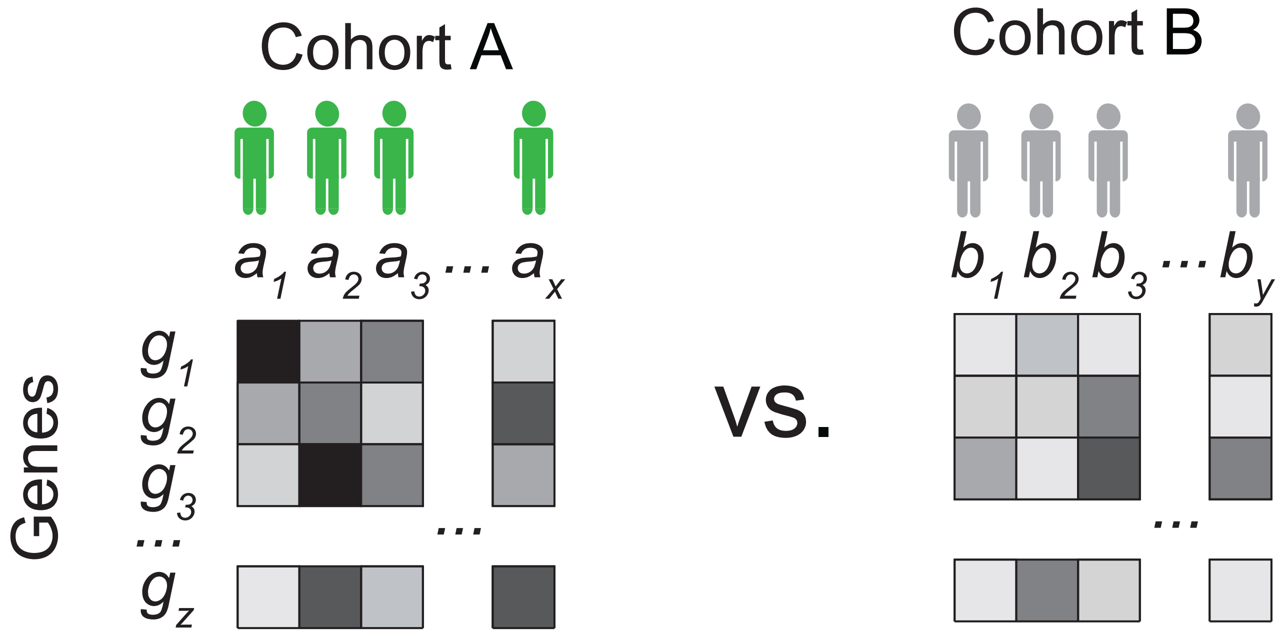
QCs **Reproducibility**

Principal Component Analysis (PCA)



PCA is a statistical technique for dimensionality reduction. We use PCA when a dataset presents a high number of features (genes in this case). It is like compressing information about ~20,000 in two dimensions or some more if we need it.

# Differential expression

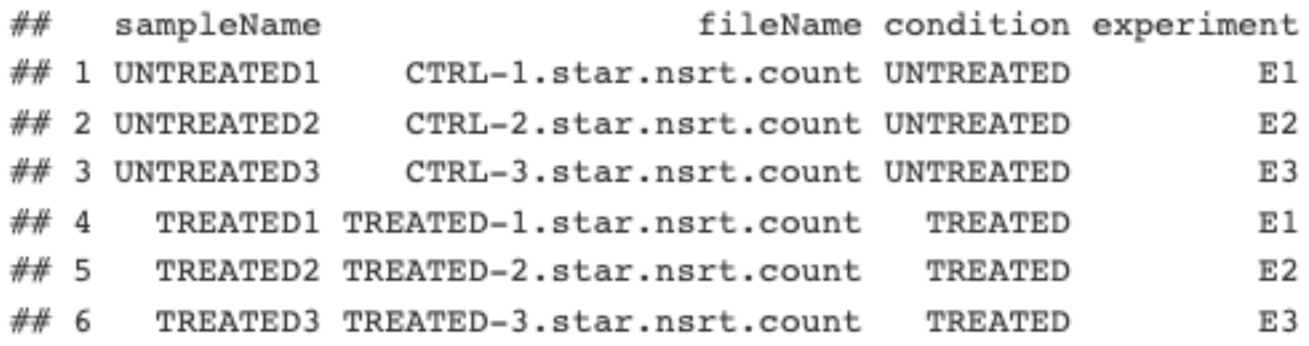


Two are the main goals of a differential expression (DE) analysis:

1. Estimate the **entity of variation** between the two conditions, i.e. calculate Fold Change (FC)

2. Estimate the **significance of the difference**, i.e. p-value, and correct it for multiple testing (p-adjusted).

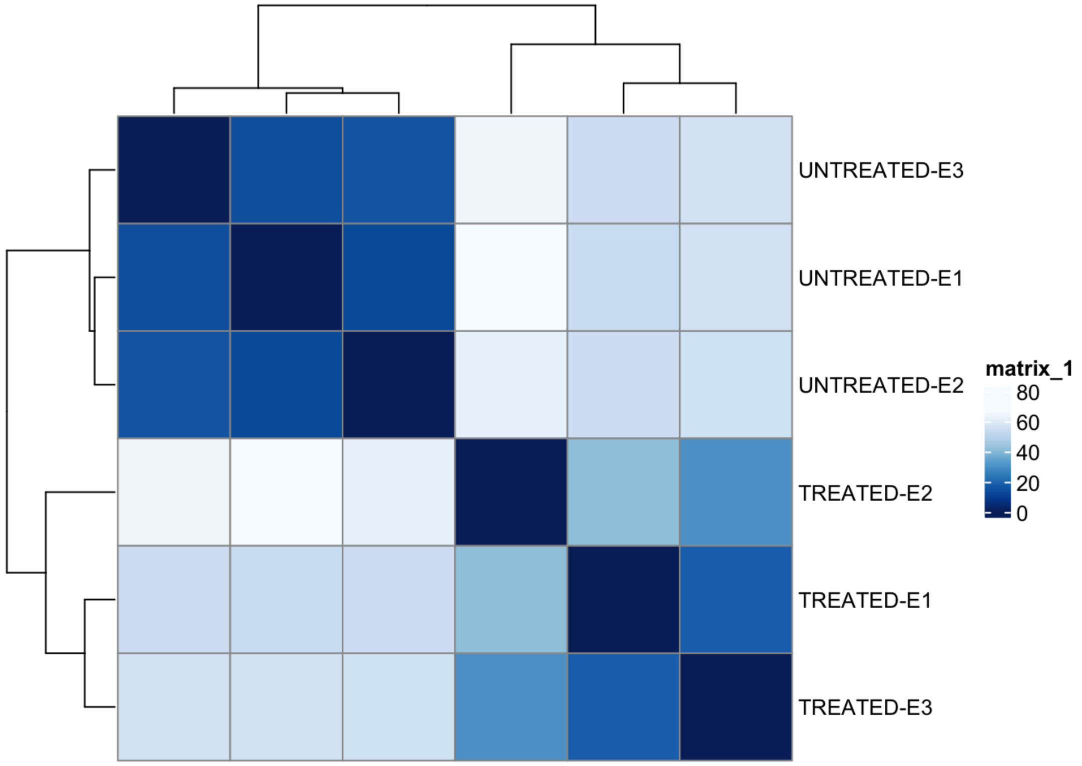
# DESeq2

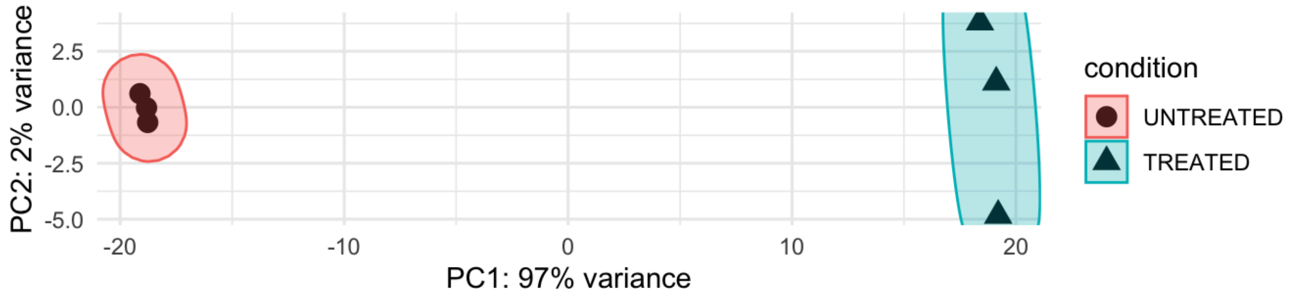


```
##   sampleName                  fileName condition experiment
## 1 UNTREATED1     CTRL-1.star.nsrt.count UNTREATED         E1
## 2 UNTREATED2     CTRL-2.star.nsrt.count UNTREATED         E2
## 3 UNTREATED3     CTRL-3.star.nsrt.count UNTREATED         E3
## 4   TREATED1 TREATED-1.star.nsrt.count   TREATED         E1
## 5   TREATED2 TREATED-2.star.nsrt.count   TREATED         E2
## 6   TREATED3 TREATED-3.star.nsrt.count   TREATED         E3
```

Raw Count Normalisation

QC reproducibility

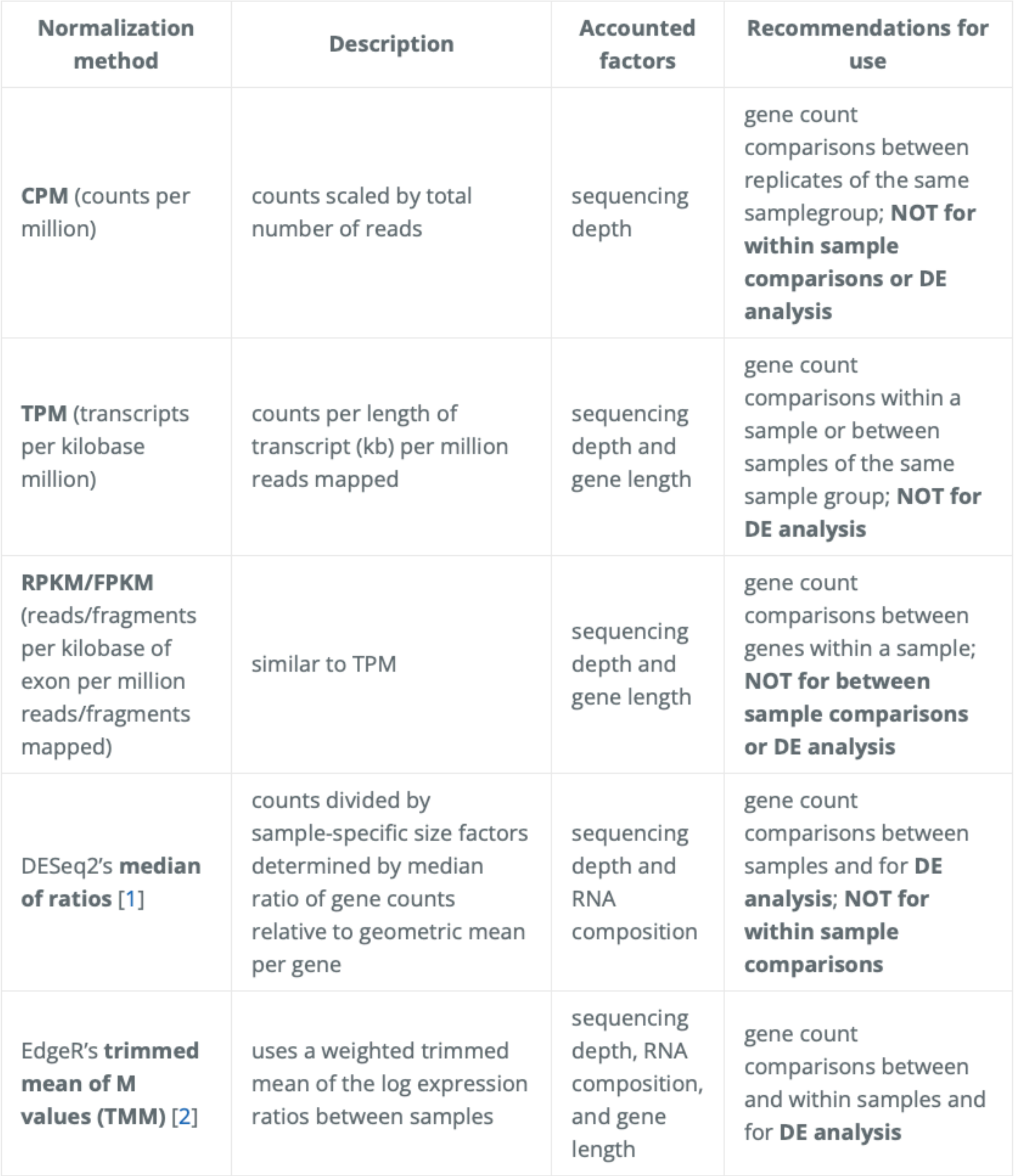




Differential expression analysis

# Normalization

Normalising data is fundamental. If we skip this step we introduce biases in our analysis.



| Normalization method | Description | Accounted factors | Recommendations for use |
|---|---|---|---|
| **CPM** (counts per million) | counts scaled by total number of reads | sequencing depth | gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis** |
| **TPM** (transcripts per kilobase million) | counts per length of transcript (kb) per million reads mapped | sequencing depth and gene length | gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis** |
| **RPKM/FPKM** (reads/fragments per kilobase of exon per million reads/fragments mapped) | similar to TPM | sequencing depth and gene length | gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis** |
| DESeq2's **median of ratios** [1] | counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene | sequencing depth and RNA composition | gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons** |
| EdgeR's **trimmed mean of M values (TMM)** [2] | uses a weighted trimmed mean of the log expression ratios between samples | sequencing depth, RNA composition, and gene length | gene count comparisons between and within samples and for **DE analysis** |

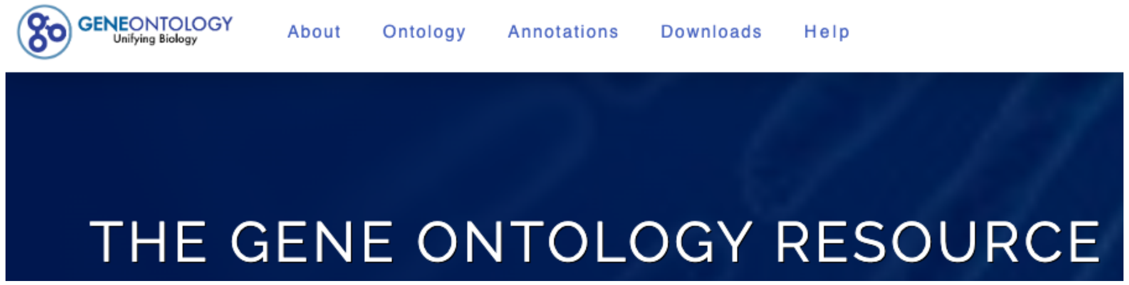
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#mrna-expression-transformation

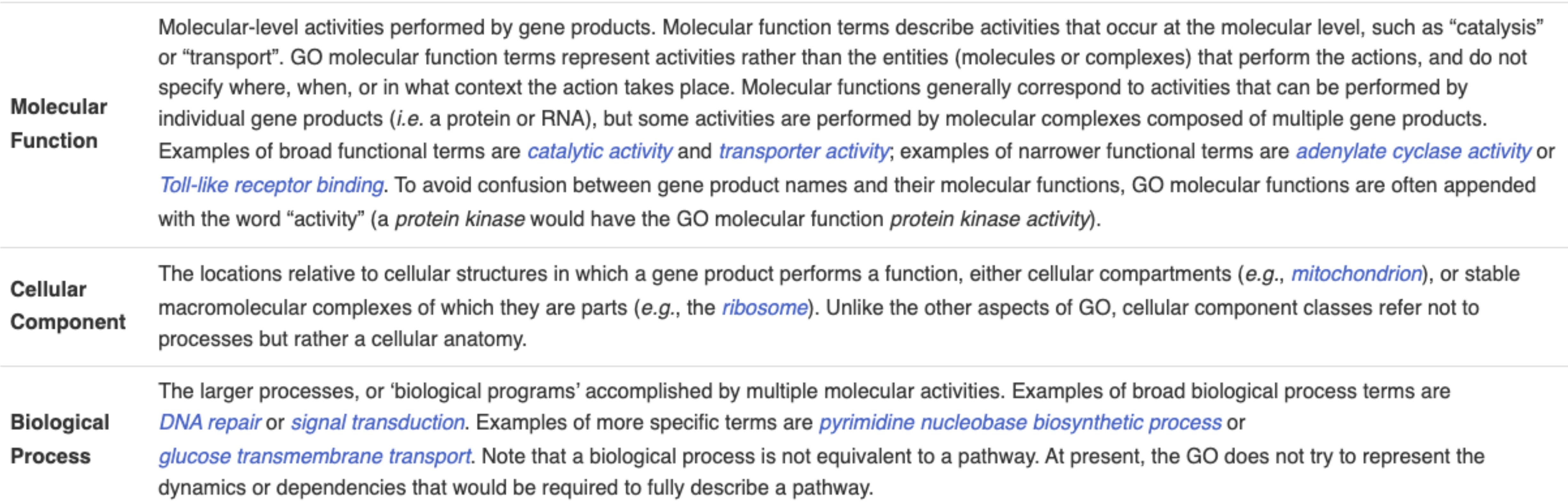
https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

# Functional annotation

Once identified differentially expressed genes, we can ask if they belong to some particular groups of genes, i.e. if they have common functionalities.

We can perform a gene ontology/over-representation analysis/gene set enrichment analysis





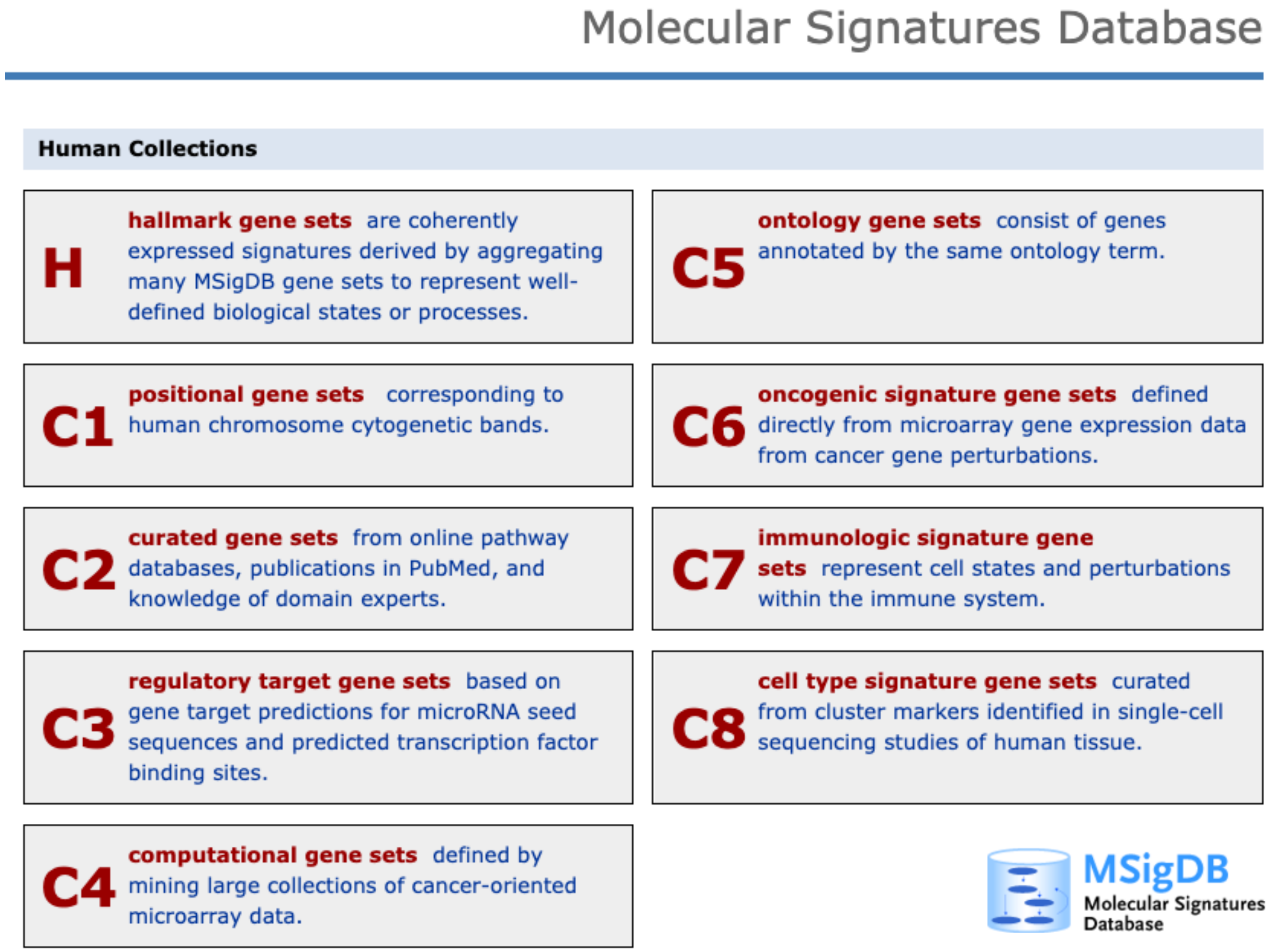
| | |
|---|---|
| **Molecular Function** | Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as “catalysis” or “transport”. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *Toll-like receptor binding*. To avoid confusion between gene product names and their molecular functions, GO molecular functions are often appended with the word “activity” (a *protein kinase* would have the GO molecular function *protein kinase activity*). |
| **Cellular Component** | The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (*e.g.*, *mitochondrion*), or stable macromolecular complexes of which they are parts (*e.g.*, the *ribosome*). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy. |
| **Biological Process** | The larger processes, or ‘biological programs’ accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*. Examples of more specific terms are *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*. Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway. |

http://geneontology.org/docs/ontology-documentation/

# Functional annotation

Once identified differentially expressed genes, we can ask if they belong to some particular groups of genes, i.e. if they have common functionalities.

We can perform a gene ontology/over-representation analysis/gene set enrichment analysis



Molecular Signatures Database

**Human Collections**

**H** **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

**C1** **positional gene sets** corresponding to human chromosome cytogenetic bands.

**C2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

**C3** **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

**C4** **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

**C5** **ontology gene sets** consist of genes annotated by the same ontology term.

**C6** **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

**C7** **immunologic signature gene sets** represent cell states and perturbations within the immune system.

**C8** **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

MSigDB
Molecular Signatures Database

https://www.gsea-msigdb.org/gsea/msigdb/