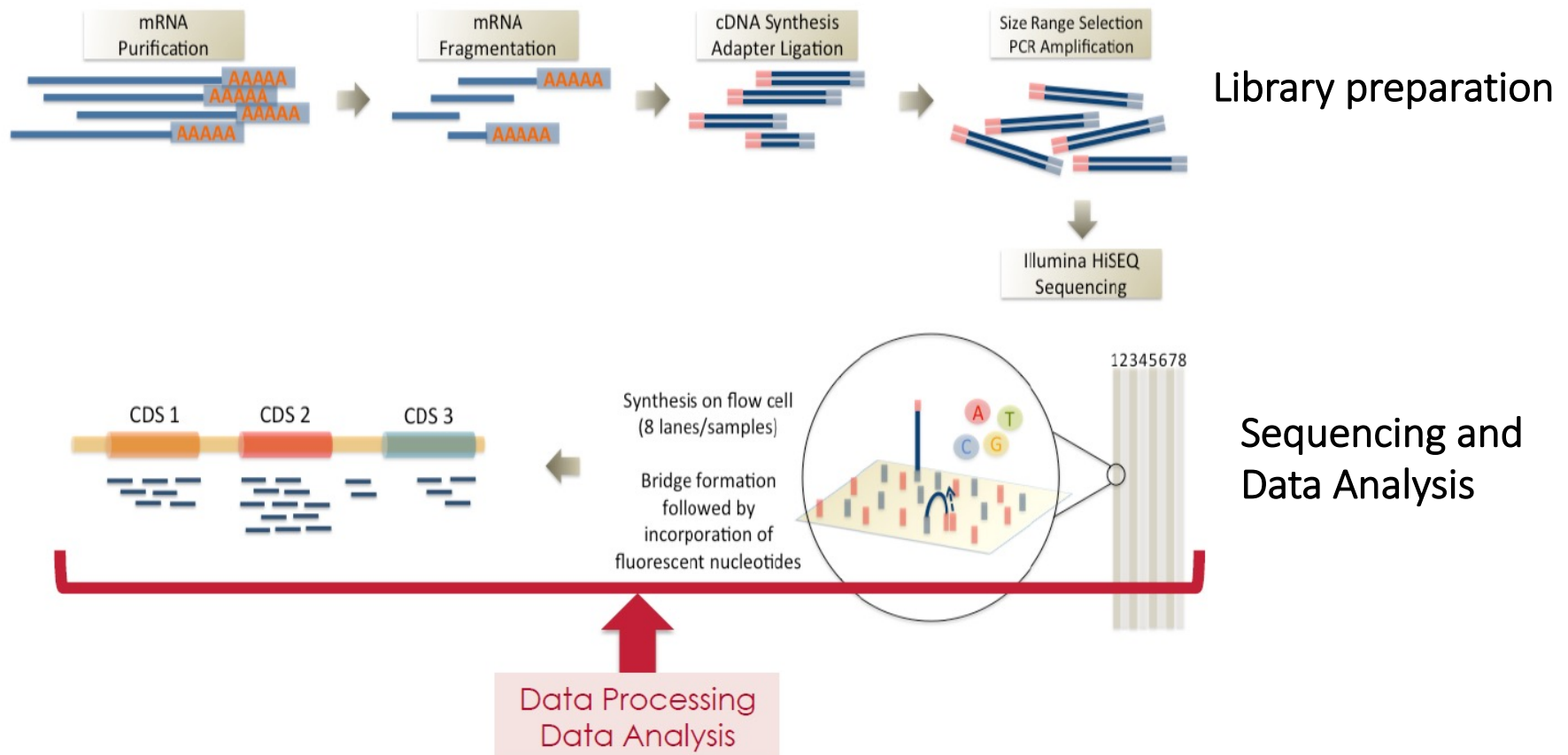


What is RNA-seq?

- RNA-seq is essentially **massively parallel sequencing of RNA** (or, in fact, the corresponding cDNA) and has heralded the second technical revolution in transcriptomics.
- It is **based on next-generation sequencing (NGS) platforms** that were initially developed for high-throughput sequencing of genomic DNA.
- Typically, **all the RNA molecules in a sample are reverse transcribed into cDNA**, and depending on the platform to be used, the **cDNA molecules may (amplification-based sequencing) or may not (single-molecule sequencing (SMS)) be amplified before deep sequencing**.
- After the sequencing reaction has taken place, **the obtained sequence stretches (reads) are mapped onto a reference genome** to deduce the structure and/or expression state of any given transcript in the sample.

RNA-Seq

The method



OVERVIEW OF RNA-Seq EXPERIMENT

Examples of experimental design:

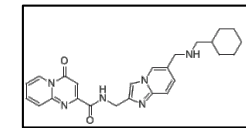
Treatment effect:

Un-Treated

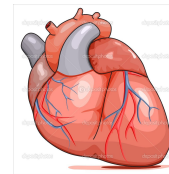


VS

Treated



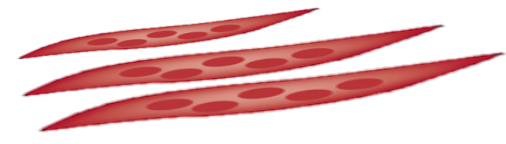
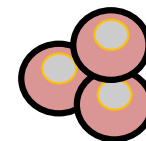
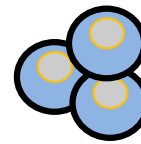
Tissue comparison:



VS



Cell Differentiation:



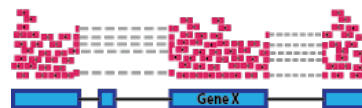
Gene function:

Myoblast

Myocyte

Myotube

siRNA / CRISPR-Cas9 /...



....

OVERVIEW OF RNA-Seq EXPERIMENT

RNA-Seq leads you to **Identify** and **quantify** RNAs that are present in your samples

- Qualitative
- Quantitative

Examples of RNA-Seq analysis:

Differential Expression Analysis (DEA):

- mRNAs (poly-A selection)
- RNAs (total RNAs - Ribominus)
- circRNAs
- small RNAs

Alternative Splicing

Alternative Poly-Adenylation

RNA enrichment in Precipitates or Pull-Down

...

EXPERIMENTAL DESIGN

Defining the technical details

Choice of sequencing depth

If we want to measure the expression of known genes, depth can be relatively low (e.g. 20 M reads for polyA+). If we want to discover new genes and transcripts, depth must be higher (e.g. 60 M for polyA+, 120 for total RNA).

Length and pairing of reads

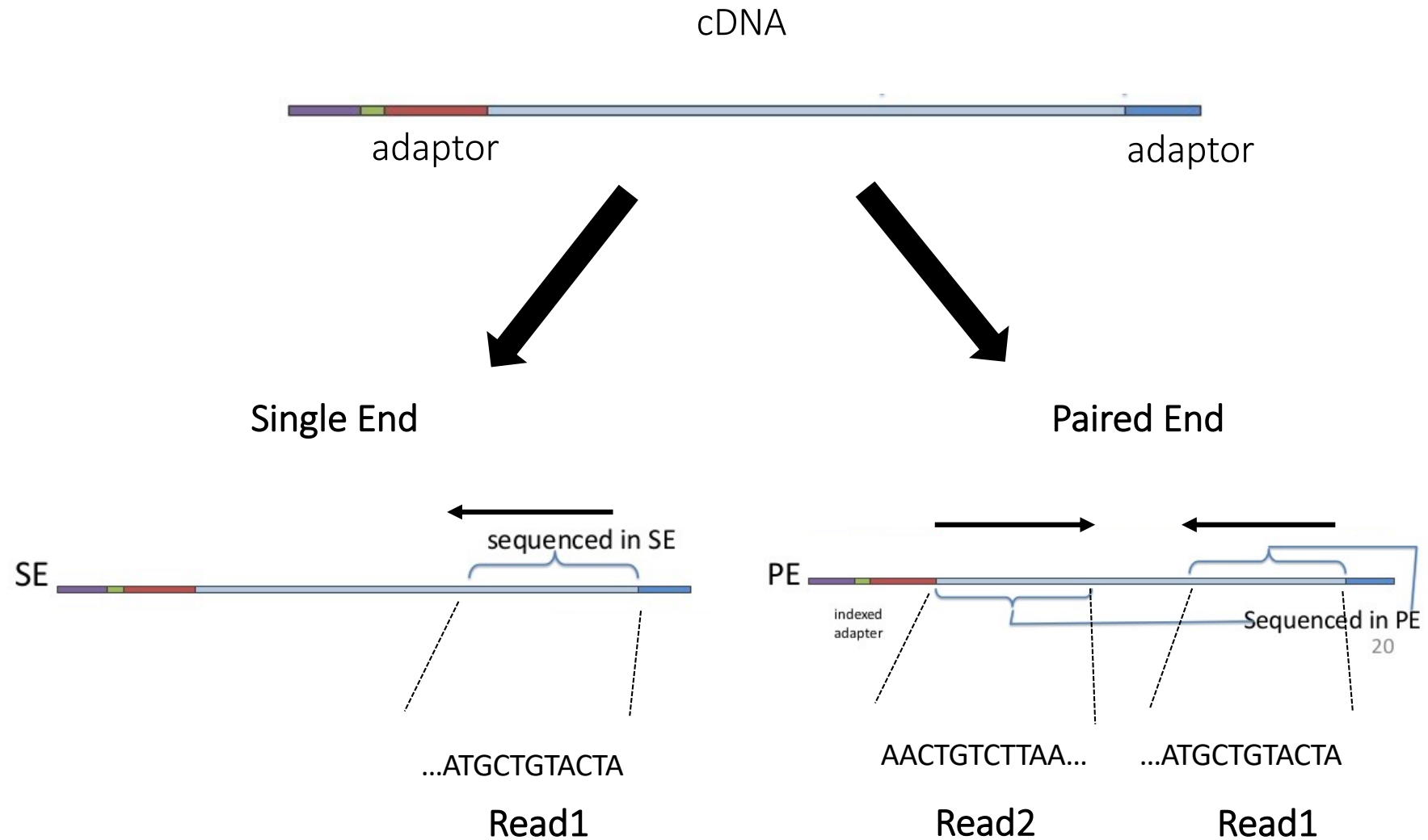
Theoretically speaking, read length should be > 20 bp (they usually are longer than 35 bp). PE reads are usually better (except for small RNA-Seq and Ribo-Seq), but they are more expensive.

Strandedness

It is usually better to have a directional (stranded) sequencing: it costs slightly more, but it is able to discriminate between antisense RNAs.



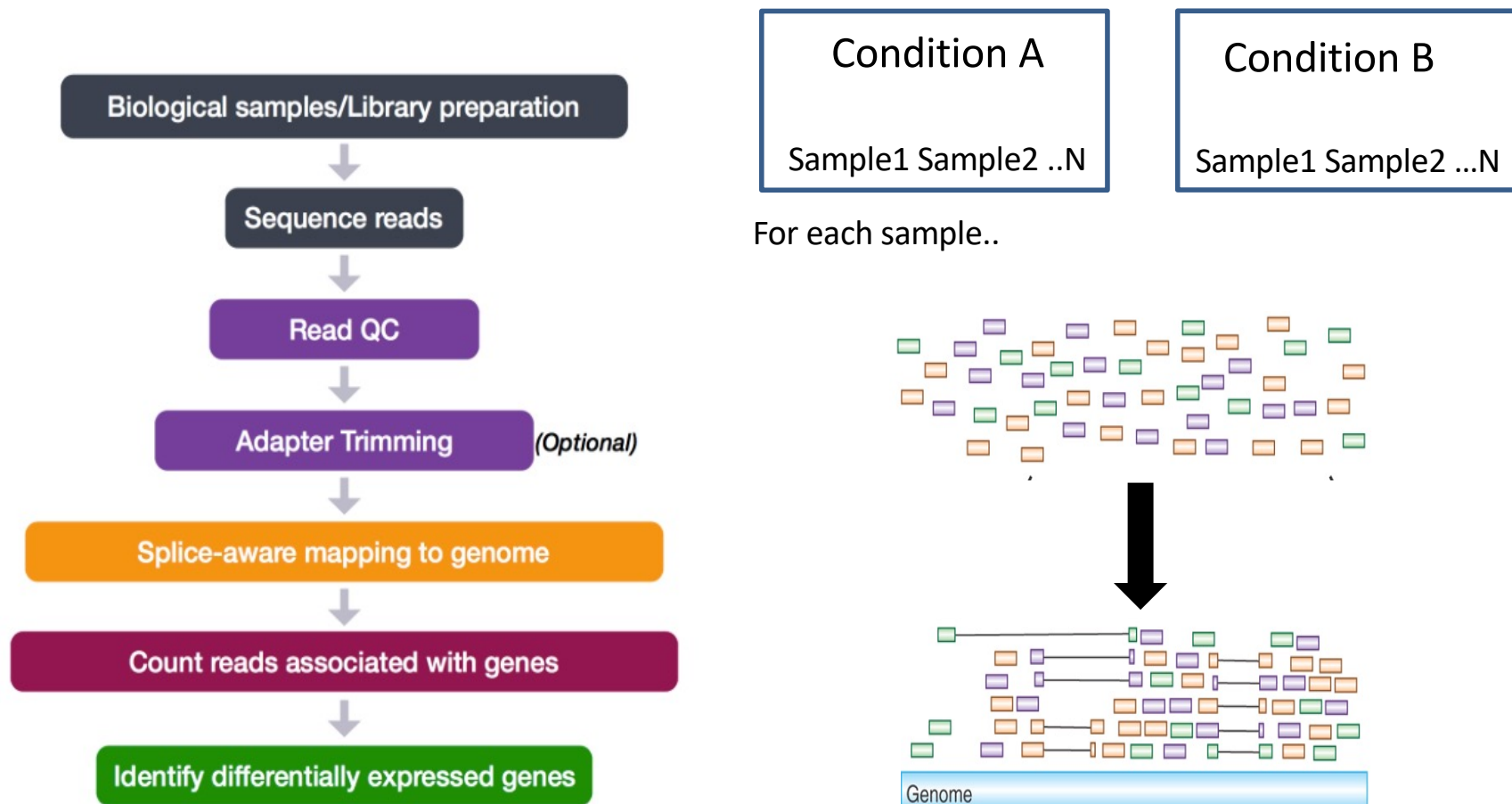
RNA-Seq: LIBRARY PREPARATION



DATA ANALYSIS



General RNA-Seq pipeline for Differential Expression



Can be compared also multiple conditions and also samples from many time points..

DATA ANALYSIS

Data format

Usually, the format of the file containing the sequence of the reads is FASTQ.

It is composed of **four-lines blocks**:

- the first line begins with @ and contains the ID of the read and optional information.
- the second line is the sequence
- the third line begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again
- the fourth line encodes the quality values for the sequence in Line 2.

For paired end reads, there are two FASTQ files (forward and reverse).



Example

```
@EAS54_6_R1_2_1_413_324
CCCTTCTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;7;;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;7;;;;;-;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;9;7;;.7;393333
```


DATA ANALYSIS

```
@SEQILMN03:128:HA5CBADXX:1:1101:1186:2059 2:N:0:GTCGTA  
NNNNNNGTTAAGATTATTGTCTATTGGCTAACTAAGCGCTACCAAGTACAAGTACAAATGC  
+  
#####0#0<BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB<<<<<<<<<<<<<<<<  
@SEQILMN03:128:HA5CBADXX:1:1101:1193:2104 2:N:0:GTCGTA  
CTATCTTCGTAACCCCAAATAAAATAAACTACTCTATTTCTTGTTAGGCAGGGTATTCC  
+  
BBBFFFFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIFB707BFFIIIIIIIIIIFFFFFFFFFFF<BBFFF  
@SEQILMN03:128:HA5CBADXX:1:1101:1227:2106 2:N:0:GTCGTA  
GGGGAGCATGACGGCCCACATCGGCGAAAACCCACTCTGGTGGGGTGAACCGGTATCCAN  
+  
BBBFFFFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIFFFFFBBFFFBFFBFFBFFBFF<BBFFO<BBFFBFBFFFFFB
```

A read is an inferred sequence of the fragment/molecule analyzed

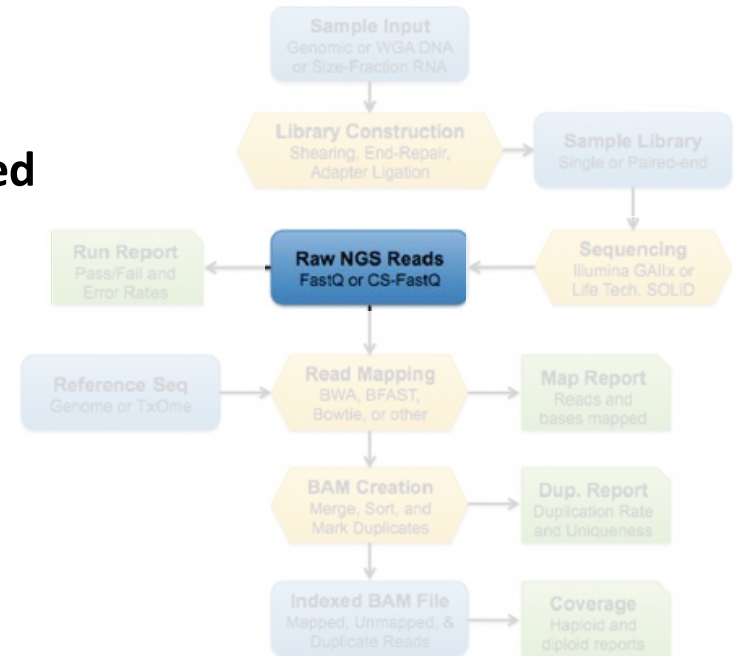
PHRED quality score

The quality score of a base, also known as a Phred or Q score, is an integer value representing the **estimated probability of an error**, i.e. that the base is incorrect.

$$Q = -10 \log_{10} P$$

A high quality score implies that a base call is more reliable and less likely to be incorrect. For example, for base calls with a quality score of Q40, one base call in 10,000 is predicted to be incorrect. For base calls with a quality score of Q30, one base call in 1,000 is predicted to be incorrect. Table 1 shows the relationship between the base call quality scores and their corresponding error probabilities.

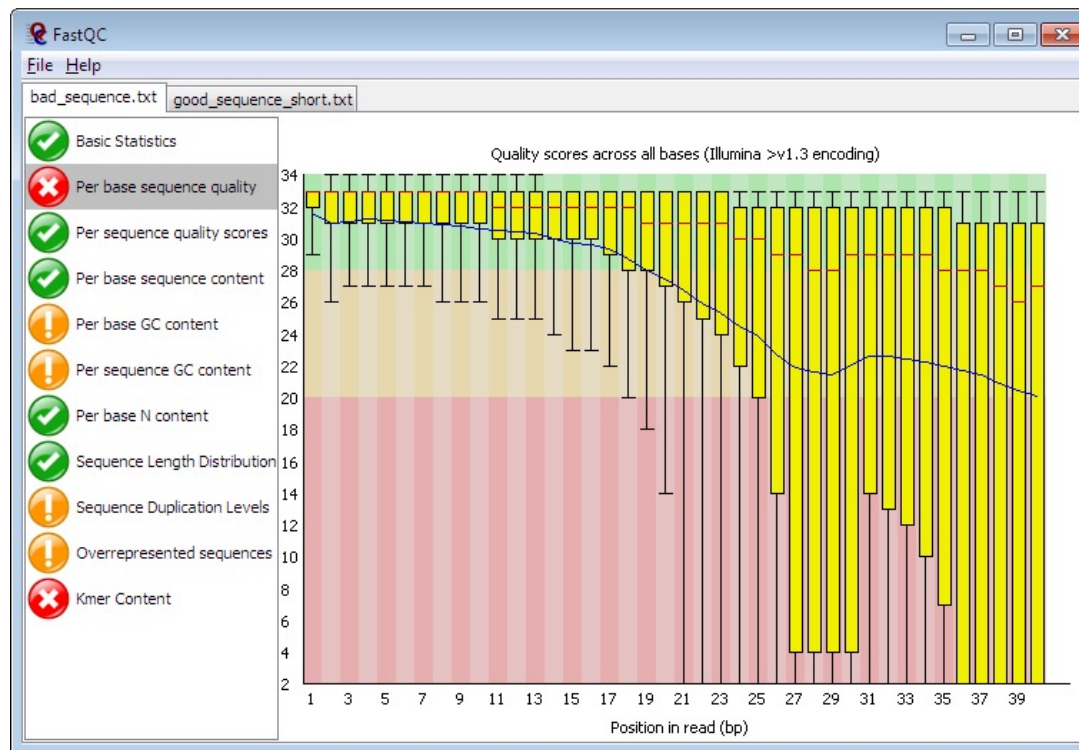
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%



DATA ANALYSIS

FastQC

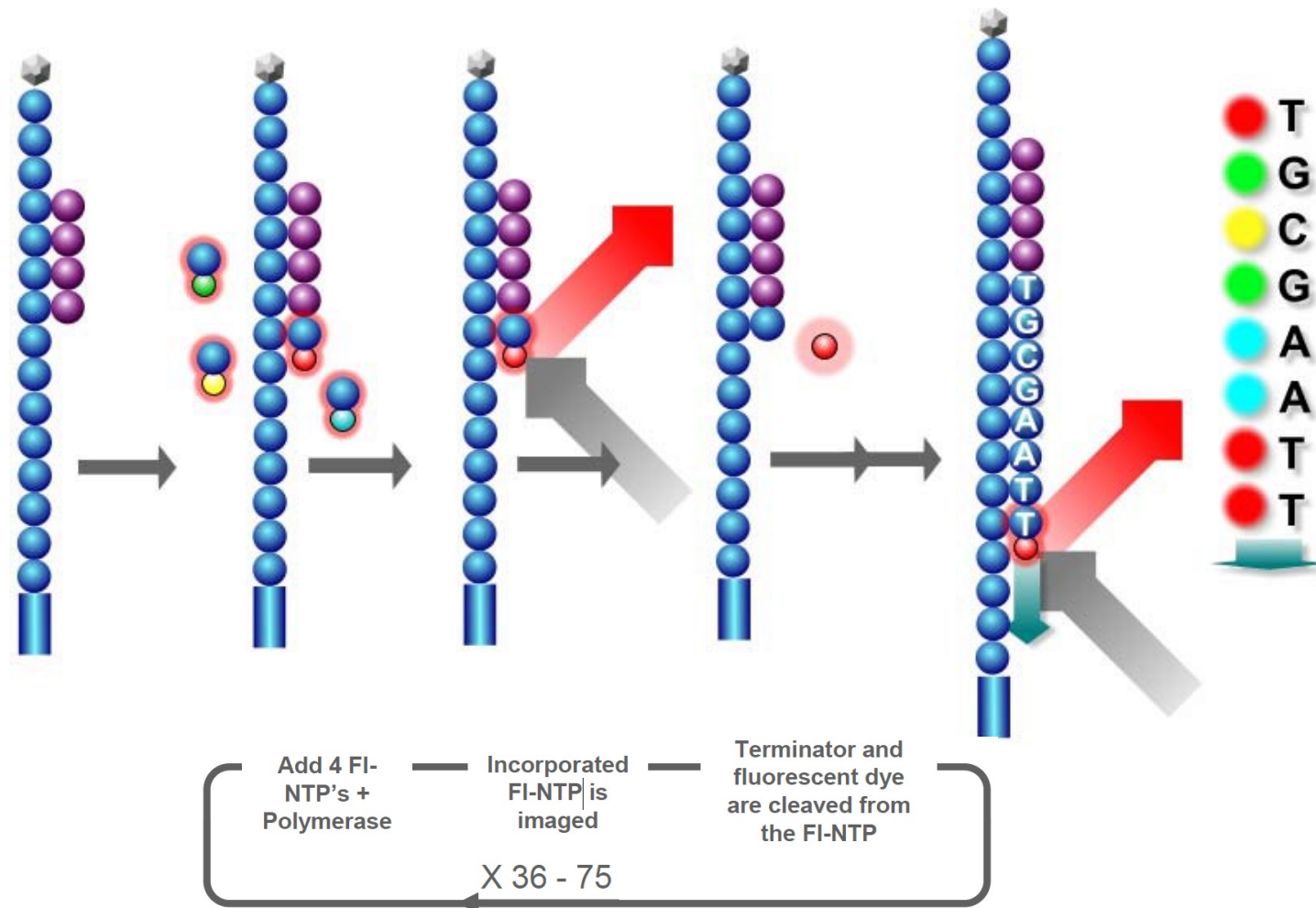
FastQC is a quality control tool for high throughput sequence data.



<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help>



RNA-Seq: SEQUENCING REACTION

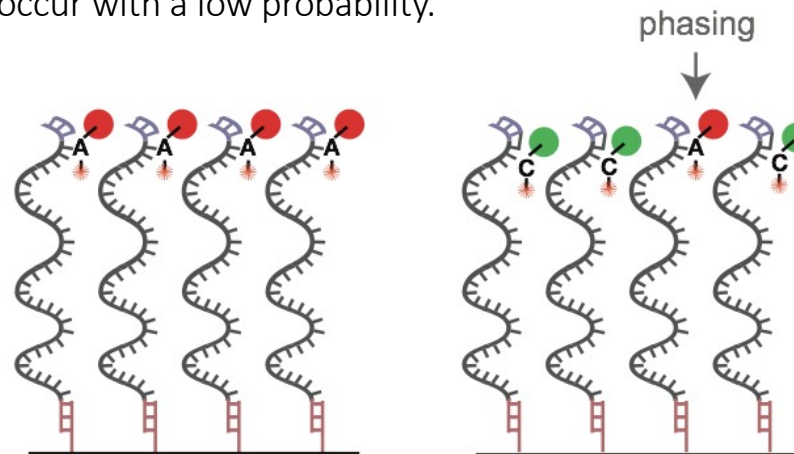
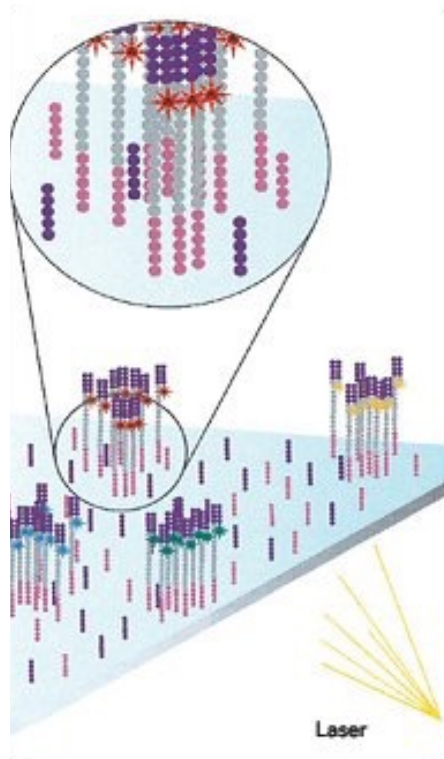


READ LENGTH = Number of reaction cycles

DATA ANALYSIS

Phasing means that the blocker of a nucleotide is not correctly removed after signal detection. In the next cycle no new nucleotide can bind on this DNA fragment and the old nucleotide is detected one more time whereby the fluorescence signal of this old nucleotide (probably) differs from the synchronous signal of the other nucleotides. From now on this DNA fragment will be 1 cycle behind the rest (out of phase), polluting the light signal that the sequencer's camera has to read.

A similar effect occurs if a nucleotide has a defect terminator cap (**prephasing**). In this case two nucleotides can bind in one cycle whereby the fragment will be 1 cycle before the rest. These errors occur with a low probability.



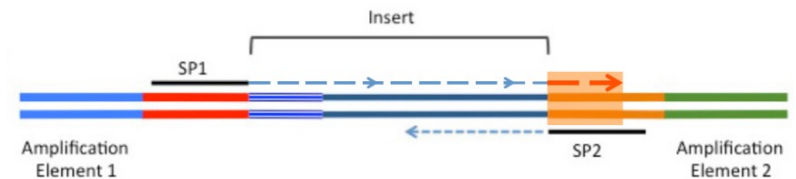
But over time (with increasing read length) they add up and pollute the light signal more and more. The signal gets more and more asynchronous. And since the light signal is used to calculate quality scores the asynchronous signal results in a decreasing sequence quality score.

DATA ANALYSIS: PREPROCESSING

Issues that can be addressed during pre-processing phase



If the read is longer than the insert (e.g. in Small RNA-Seq), its sequence will also contain part of the 3' adapter. This unwanted sequence must be removed.



If the overall quality of the read is low, it must be removed. A trimming is useful if quality decreases too much towards the end of the read.



Sometimes the read terminates with ambiguous (N) bases which must be removed.



Some of the most common preprocessing tool are FASTX-Toolkit, Cutadapt, Trimmomatic, Prinseq.

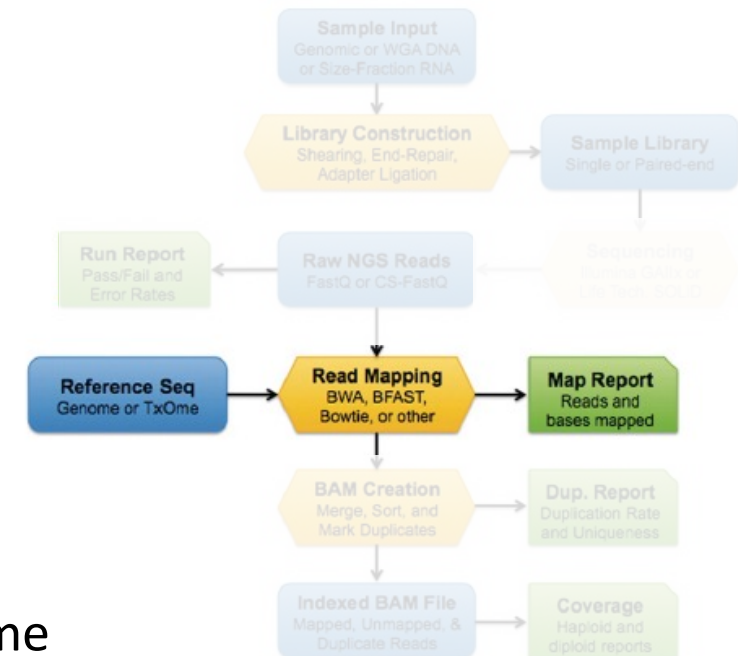
DATA ANALYSIS: ALIGNMENT

Read alignment

After pre-processing, we can align reads to a reference sequence.

- to align a read means finding the region of the genome to which it **belongs**.
- if the genome sequence of the organism is **known**, reads can be aligned to it.
- other approaches have to be used if the genome sequence is **not known** (de novo transcriptome assembly).

The accurate and fast alignment of millions of reads is not a simple task: many programs have been developed to address this issue.



DATA ANALYSIS: ALIGNMENT

Read alignment



After pre-processing, we can align reads to a reference sequence.

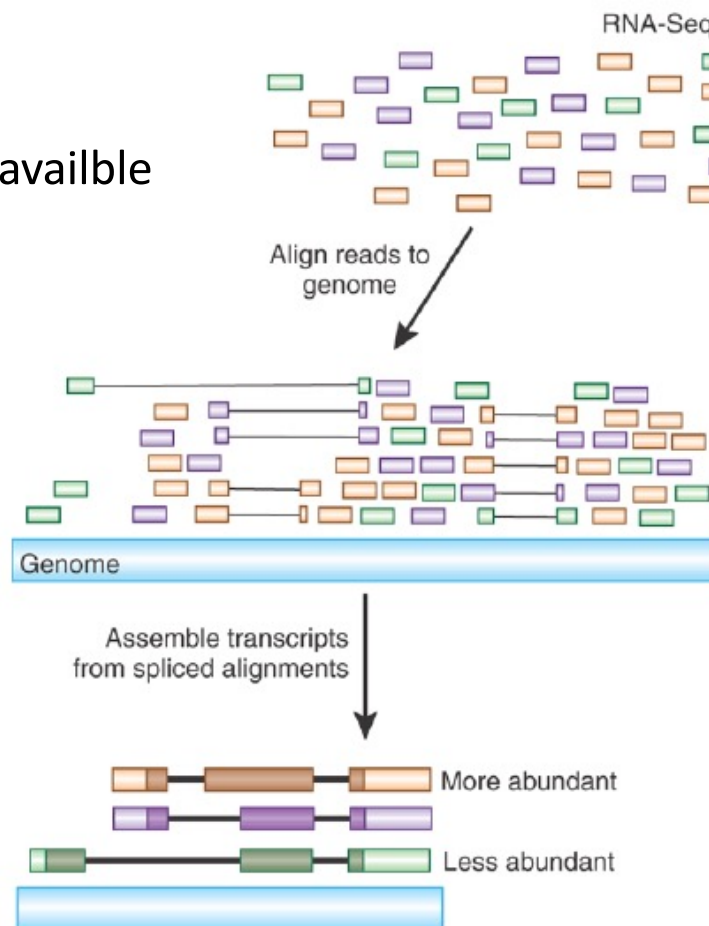


DATA ANALYSIS: ALIGNMENT

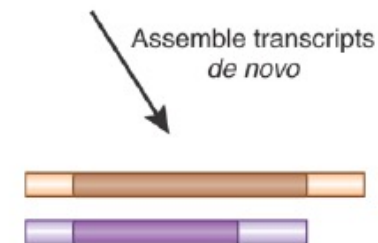
Main alignment strategies

- Reference available

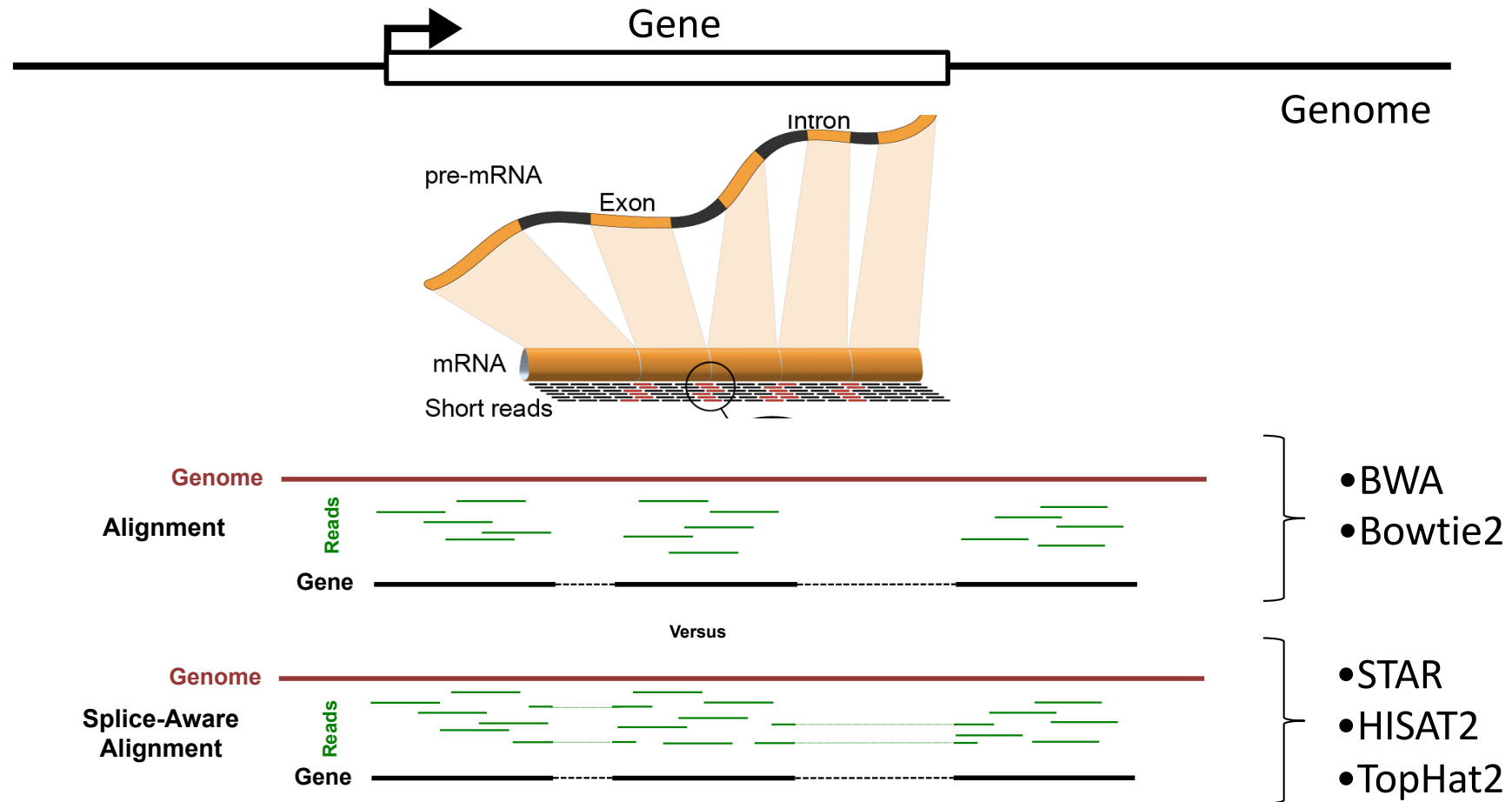
- transcriptome



- Reference not available



DATA ANALYSIS: ALIGNMENT



Splice-aware alignment

DATA ANALYSIS: ALIGNMENT

Alignment tools

BWA, Soap2 and Bowtie2 are based on the **Burrows-Wheeler Transform**, an indexing technique which allows to have reduced time required for the alignment compared to older tools like Maq (the alignment of 20M reads is done in few hours).

Table 3:
Selected mapping and alignment tools for massively parallel sequencing data

Aligner	Description	URL
Illumina platform		
ELAND	Vendor-provided aligner for Illumina data	http://www.illumina.com
Bowtie	Ultrafast, memory-efficient short-read aligner for Illumina data	http://bowtie-bio.sourceforge.net
Novoalign	A sensitive aligner for Illumina data that uses the Needleman-Wunsch algorithm	http://www.novocraft.com
SOAP	Short oligo analysis package for alignment of Illumina data	http://soap.genomics.org.cn/
MrFAST	A mapper that allows alignments to multiple locations for CNV detection	http://mrfast.sourceforge.net/
SOLiD platform		
Corona-lite	Vendor-provided aligner for SOLiD data	http://solidsoftwaretools.com
SHRIMP	Efficient Smith-Waterman mapper with colorspace correction	http://compbio.cs.toronto.edu/shrimp/
454 Platform		
Newbler	Vendor-provided aligner and assembler for 454 data	http://www.454.com
SSAHA2	SAM-friendly sequence search and alignment by hashing program	http://www.sanger.ac.uk/resources/software
BWA-SW	SAM-friendly Smith-Waterman implementation of BWA for long reads	http://bio-bwa.sourceforge.net
Multi-platform		
BFAST	BLAT-like fast aligner for Illumina and SOLiD data	http://bfast.sourceforge.net
BWA	Burrows-Wheeler aligner for Illumina, SOLiD, and 454 data	http://bio-bwa.sourceforge.net
Maq	A widely used mapping tool for Illumina and SOLiD; now deprecated by BWA	http://maq.sourceforge.net

DATA ANALYSIS: ALIGNMENT

Spliced aligners

- The algorithms discussed so far are not able to align reads on splicing junctions, unless we use the transcriptome sequence as a reference.
- There are several programs that are able to perform spliced alignments: Tophat2, STAR, Hisat2 ,Gsnap, MapSplice, PALMapper, ReadsMap etc.
- **Tophat** uses Bowtie as an alignment “engine”. The algorithm can be divided into two main steps:
 - Reads are aligned to the reference genome.
 - Reads that cannot be aligned directly to the reference are aligned to possible splicing junctions.

DATA ANALYSIS: ALIGNMENT

Main alignment programs

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners ^a	Seed methods	Short-read mapping package (SHRiMP) ⁴¹	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy ³⁹	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie ⁴³			
		BWA ⁴⁴	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice ⁵²	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap ⁵⁰			
		TopHat ⁵¹	Uses Bowtie alignments		
	Seed-extend methods	GSNAP ⁵³	Can use SNP databases		
		QPALMA ⁵⁴	Smith-Waterman for large gaps		
		Star	Superfast		

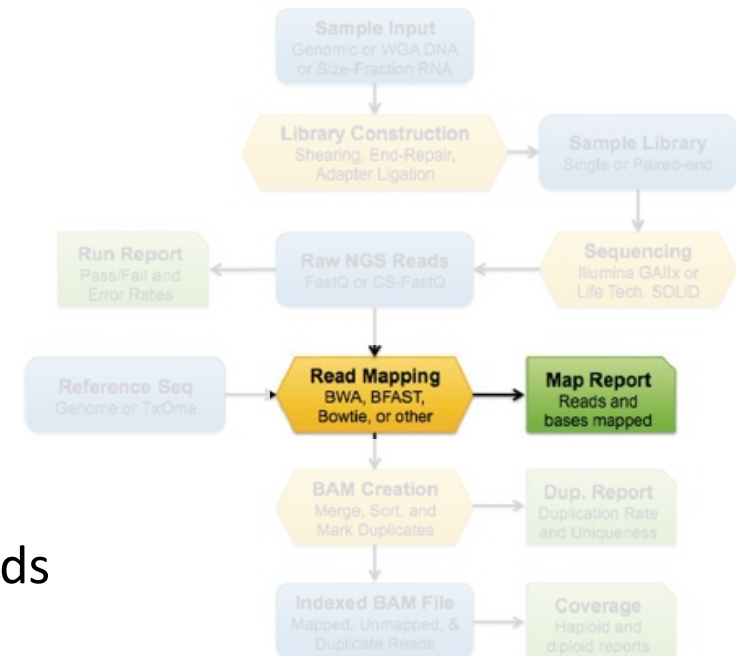
Gaber *et al.*, 2011, Nature Methods 8:469

DATA ANALYSIS: ALIGNMENT

Alignment output

After alignment, mapped and unmapped reads are usually exported in SAM/BAM format.

- **SAM** format specification (Sequence Alignment Map, <http://samtools.sourceforge.net/SAM1.pdf>) describes a generic format for the storing of reads sequence and their alignment on a reference.
- **BAM** is the binary equivalent of SAM.



A generic SAM/BAM file is composed of two parts:

-
- ```

graph TD
 A[Sample Input
Genomic or WGA DNA
or Size-Fraction RNA] --> B[Library Construction
Shearing, End-Repair,
Adapter Ligation]
 B --> C[Sample Library
Single or Paired-end]
 C --> D[Sequencing
Illumina GAIIx or
Life Tech. SOLiD]
 D --> E[Raw NGS Reads
FastQ or CS-FastQ]
 E --> F[Run Report
Pass/Fail and
Error Rates]
 E --> G[Read Mapping
BWA, BFAST,
Bowtie, or other]
 H[Reference Seq
Genome or TxOme] --> G
 G --> I[Map Report
Reads and
bases mapped]
 G --> J[BAM Creation
Merge, Sort, and
Mark Duplicates]
 J --> K[Dup. Report
Duplication Rate
and Uniqueness]
 J --> L[Indexed BAM File
Mapped, Unmapped, &
Duplicate Reads]
 L --> M[Coverage
Haploid and
diploid reports]

```
- The flowchart illustrates the NGS workflow. It begins with **Sample Input** (Genomic or WGA DNA or Size-Fraction RNA), which leads to **Library Construction** (Shearing, End-Repair, Adapter Ligation). This step produces a **Sample Library** (Single or Paired-end), which is then sequenced using **Sequencing** (Illumina GAIIx or Life Tech. SOLiD). The output is **Raw NGS Reads** (FastQ or CS-FastQ). From here, a **Run Report** (Pass/Fail and Error Rates) is generated, and the reads are processed for **Read Mapping** (BWA, BFAST, Bowtie, or other). The **Read Mapping** step also takes a **Reference Seq** (Genome or TxOme) as input. The mapping results are used to generate a **Map Report** (Reads and bases mapped) and to create a **BAM** (Binary Alignment Map). The **BAM** is then processed for **BAM Creation** (Merge, Sort, and Mark Duplicates), which produces a **Dup. Report** (Duplication Rate and Uniqueness) and an **Indexed BAM File** (Mapped, Unmapped, & Duplicate Reads). Finally, the **Indexed BAM File** is used to generate **Coverage** (Haploid and diploid reports).

GENE EXPRESSION REGULATION IN EUKARYOTES – LM-GBM a.a. 2023-2024  
Università La Sapienza di Roma

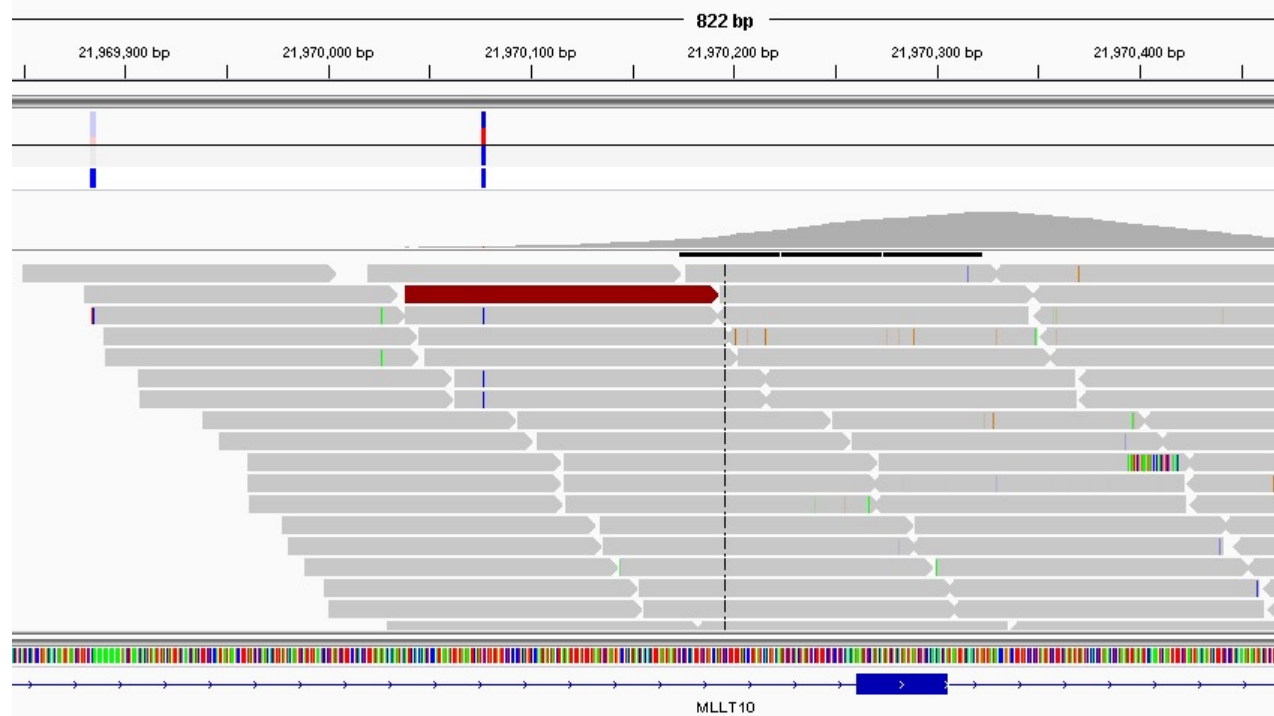


# DATA ANALYSIS: ALIGNMENT

## BAM file visualization

### IGV

IGV is a standalone program which allows a highly interactive visualization of BAM files (and other genomic annotation formats).



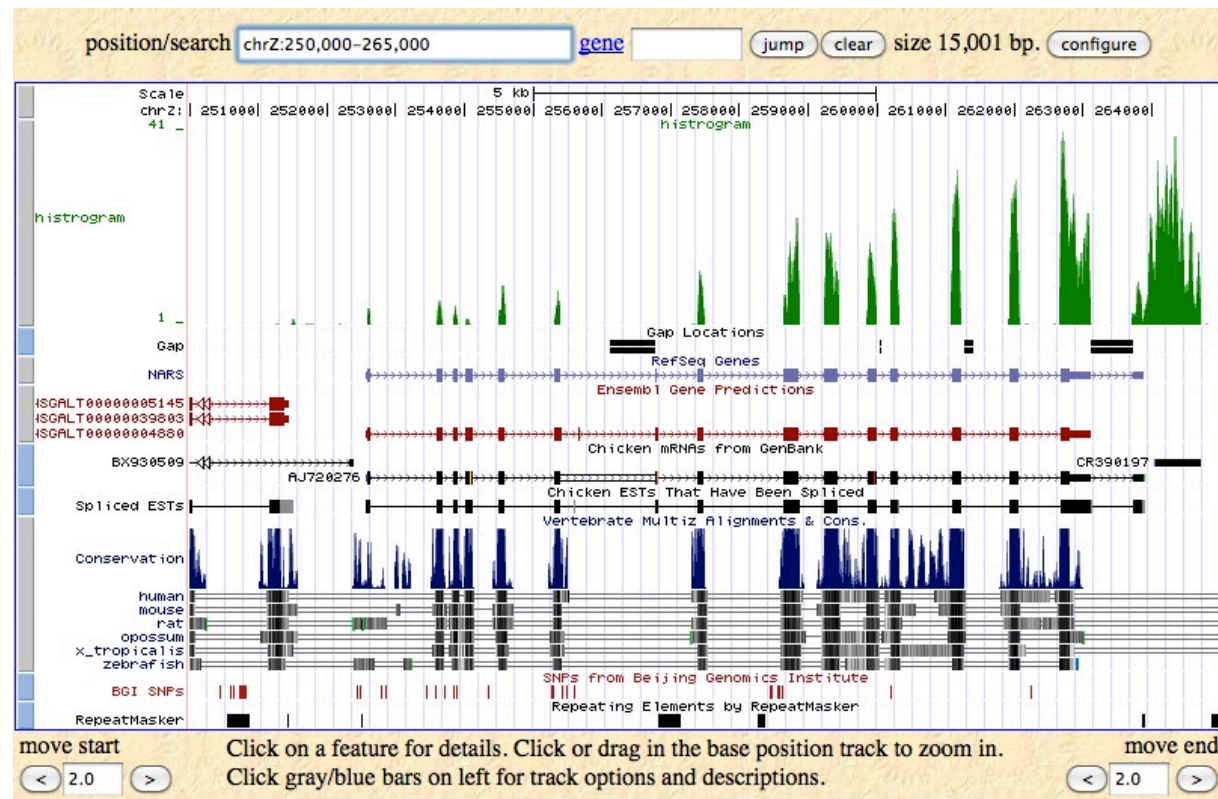


# DATA ANALYSIS: ALIGNMENT

## BAM file visualization

### Genome Browser (UCSC)

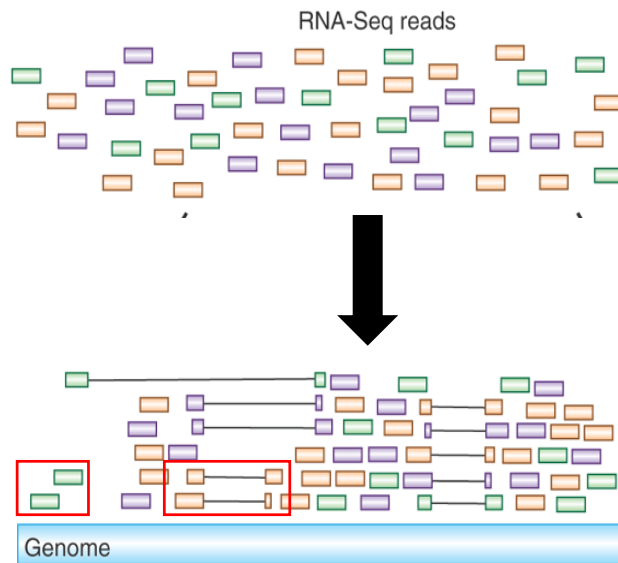
Visualization is less interactive, but many supplementary tracks are available.





# DATA ANALYSIS: ALIGNMENT

## How to assign reads to genes:



- htseq count
- featureCount
- STAR



After reads mapping, gene annotation (gtf) where used in order to quantify the expression of each gene in each sample



# DATA ANALYSIS: ALIGNMENT

## How to assign reads to genes:

Each column is a sample

Count matrix

Each row is a gene

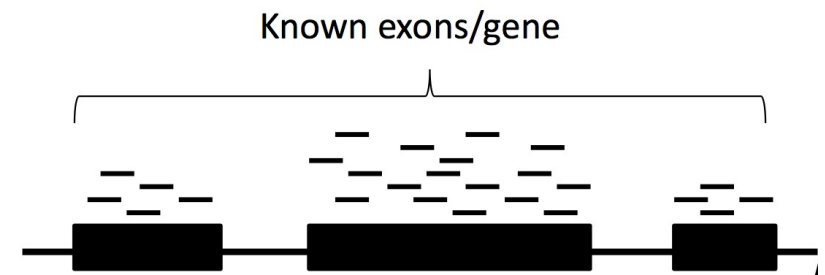
| GENE ID     | KD.2 | KD.3 | OE.1 | OE.2 | OE.3 | IR.1 | IR.2 | IR.3 |
|-------------|------|------|------|------|------|------|------|------|
| 1/2-SBSRNA4 | 57   | 41   | 64   | 55   | 38   | 45   | 31   | 39   |
| A1BG        | 71   | 40   | 100  | 81   | 41   | 77   | 58   | 40   |
| A1BG-AS1    | 256  | 177  | 220  | 189  | 107  | 213  | 172  | 126  |
| A1CF        | 0    | 1    | 1    | 0    | 0    | 0    | 0    | 0    |
| A2LD1       | 146  | 81   | 138  | 125  | 52   | 91   | 80   | 50   |
| A2M         | 10   | 9    | 2    | 5    | 2    | 9    | 8    | 4    |
| A2ML1       | 3    | 2    | 6    | 5    | 2    | 2    | 1    | 0    |
| A2MP1       | 0    | 0    | 2    | 1    | 3    | 0    | 2    | 1    |
| A4GALT      | 56   | 37   | 107  | 118  | 65   | 49   | 52   | 37   |
| A4GNT       | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    |
| AA06        | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| AAA1        | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0    |
| AAAS        | 2288 | 1363 | 1753 | 1727 | 835  | 1672 | 1389 | 1121 |
| AACS        | 1586 | 923  | 951  | 967  | 484  | 938  | 771  | 635  |
| AACSP1      | 1    | 1    | 3    | 0    | 1    | 1    | 1    | 3    |
| AADAC       | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| AADACL2     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| AADACL3     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| AADACL4     | 0    | 0    | 1    | 1    | 0    | 0    | 0    | 0    |
| AADAT       | 856  | 539  | 593  | 576  | 359  | 567  | 521  | 416  |
| AAGAB       | 4648 | 2550 | 2648 | 2356 | 1481 | 3265 | 2790 | 2118 |
| AAK1        | 2310 | 1384 | 1869 | 1602 | 980  | 1675 | 1614 | 1108 |
| AAMP        | 5198 | 3081 | 3179 | 3137 | 1721 | 4061 | 3304 | 2623 |
| AANAT       | 7    | 7    | 12   | 12   | 4    | 6    | 2    | 7    |
| AARS        | 5570 | 3323 | 4782 | 4580 | 2473 | 3953 | 3339 | 2666 |
| AAPC3       | 4451 | 2737 | 3381 | 3131 | 1340 | 3480 | 2874 | 1657 |



# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

## Measures of gene expression

- “The number of read counts mapping to the biological feature of interest (gene, transcript, exon etc.) is considered to be linearly related to the abundance of the target feature.”  
(Tarazona, 2011)



- The raw number of reads mapping on a gene (**read count**) requires a normalization.

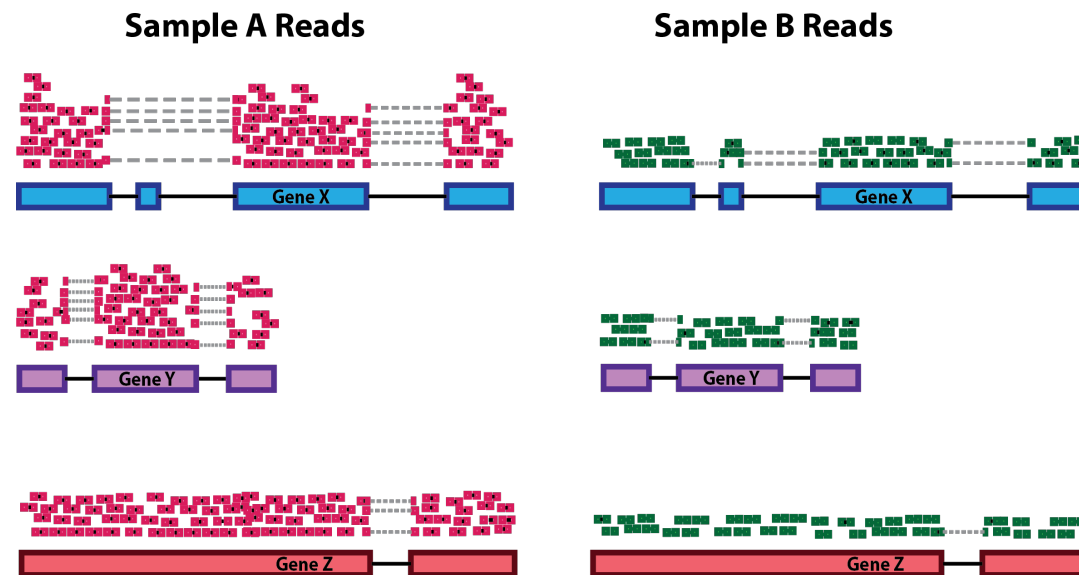
Why?



# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Why normalization is required before DE analysis?

- Sequencing Depth



**the number of reads mapped on a gene depends on sequencing depth:** to normalize for the total number of mapped reads is important to compare the expression levels of the same gene obtained from two different sequencing experiments.

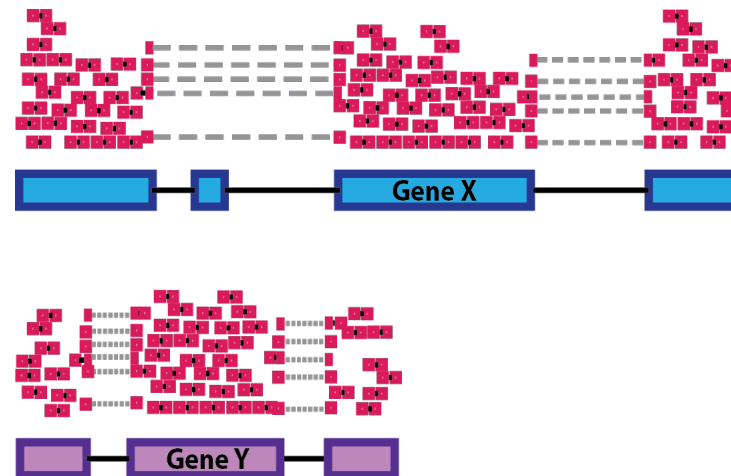


# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Why normalization is required before DE analysis?

- Gene length

## Sample A Reads



**longer genes will have a greater number of reads mapped on them compared to equally expressed shorter genes:** to normalize for gene length is important to compare the expression of distinct genes.



# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

## Measures of gene expression: RPKM

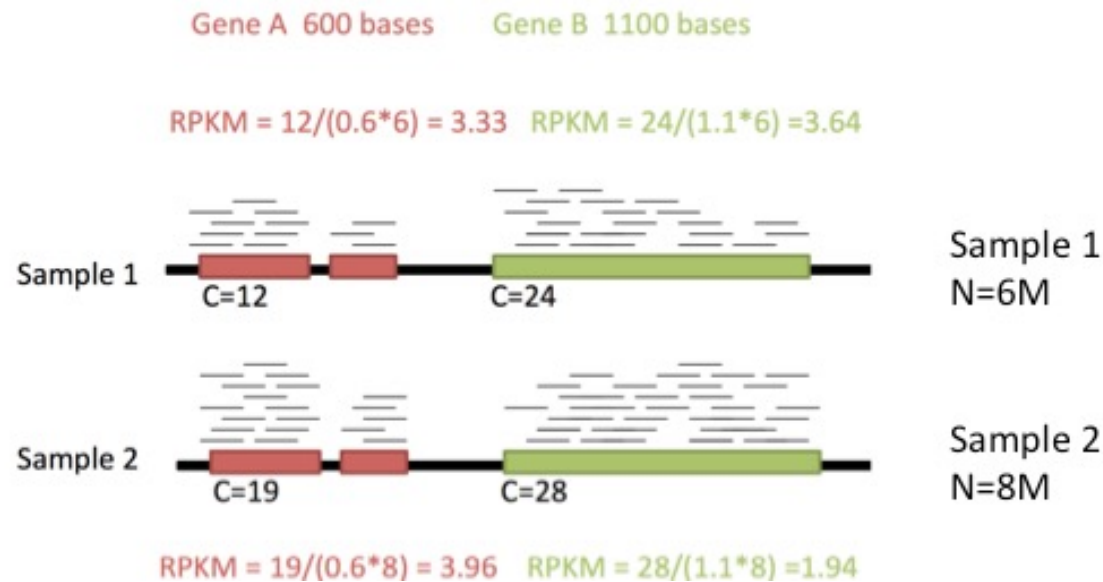
- RPKM stands for “Reads per Kilobase of exon per Million mapped reads”

$$\text{RPKM} = \frac{C}{LN}$$

➤ C : Number of mappable reads on a feature (eg. transcript, exon, etc.)

➤ L: Length of feature (in kb)

➤ N: Total number of mappable reads (in millions)





# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

## Measures of gene expression: FPKM

- FPKM stands for “Fragments per Kilobase of exon per Million mapped fragments”
- The unit used for quantification is no longer the single read, but the fragment. In single-end sequencing, each read represents a fragment, so  $FPKM = RPKM$ . In paired-end sequencing, each fragment is represented by a read pair: this way, each read pair is not counted twice.



RPKM = 1



RPKM = 2

FPKM = 1



# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

| Normalization method                                                                       | Description                                                                                                                  | Accounted factors                                  | Recommendations for use                                                                                                    |
|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|
| <b>CPM</b> (counts per million)                                                            | counts scaled by total number of reads                                                                                       | sequencing depth                                   | gene count comparisons between replicates of the same samplegroup; <b>NOT for within sample comparisons or DE analysis</b> |
| <b>TPM</b> (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; <b>NOT for DE analysis</b>             |
| <b>RPKM/FPKM</b> (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; <b>NOT for between sample comparisons or DE analysis</b>             |
| DESeq2's <b>median of ratios</b>                                                           | 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 <b>DE analysis</b> ; <b>NOT for within sample comparisons</b>               |
| EdgeR's <b>trimmed mean of M values (TMM)</b>                                              | 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 <b>DE analysis</b>                                               |



## Tools for de novo discovery of transcripts

- **genome-guided** programs use the alignment of reads to the genome to assemble novel transcripts and genes.
- **genome-independent** programs use the overlap between reads to assemble transcripts; alignment to the genome is not required. They are thus useful in the absence of a reference genome, but also to find transcripts coming from genes which underwent structural variations (indels, fusions etc.). These programs are usually slower.

### Transcriptome reconstruction

|                                   |                                               |                                                                |                                                                              |                                                                                  |                                |
|-----------------------------------|-----------------------------------------------|----------------------------------------------------------------|------------------------------------------------------------------------------|----------------------------------------------------------------------------------|--------------------------------|
| Genome-guided reconstruction      | Exon identification<br>Genome-guided assembly | G.Mor.Se<br>Scripture <sup>28</sup><br>Cufflinks <sup>29</sup> | Assembles exons<br>Reports all isoforms<br>Reports a minimal set of isoforms | Identifying novel transcripts using a known reference genome                     | Alignments to reference genome |
| Genome-independent reconstruction | Genome-independent assembly                   | Velvet <sup>61</sup><br>TransABYSS <sup>56</sup><br>Trinity    | Reports all isoforms                                                         | Identifying novel genes and transcript isoforms without a known reference genome | Reads                          |



# DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

## What is differential expression (DE) analysis?

DE analysis allows to find **genes** (or other genomic features like transcripts and exons) **that are expressed at significantly different levels between two groups of samples** (conditions): patients treated with drugs VS controls, healthy VS sick individuals, different tissues and different differentiation states. There could also be more than two conditions (e.g. time series).

For each analyzed gene, the result will be:

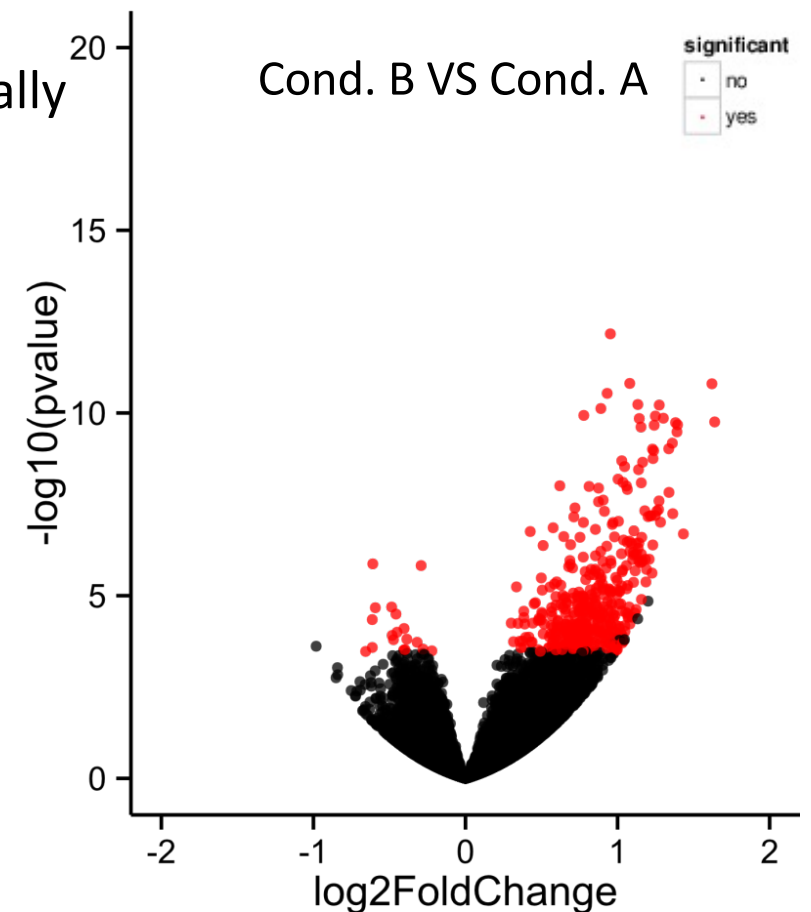
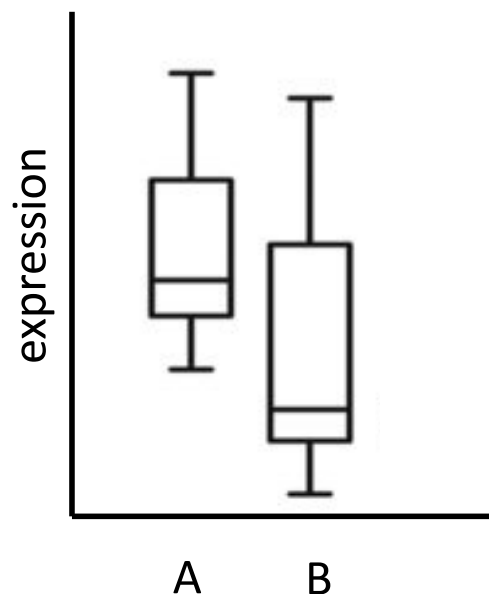
- **Fold Change (FC)**: the **ratio** of the average expression of gene in condition A to the average expression in condition B. log2 transformed fold changes are nicer to work with because the transform is **symmetric for reciprocals** (positive values for up-regulation, negative for down-regulation).
- **P-value**: it measures the statistical significance of the observed differential expression. The **lower the p-value**, the **higher the probability** that the gene underwent a significant **deregulation**. Goes from 0 to 1, usual cutoff is 0.05. It is often normalized to account for multiple testing.



# DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

## FC vs p-value

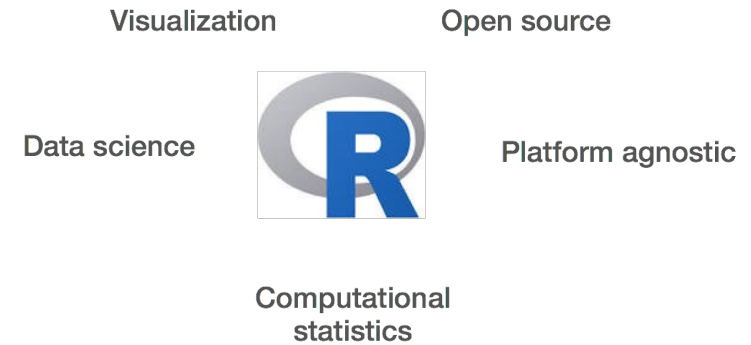
High absolute FC values are not necessarily associated with significant P-values, especially when the expression of the gene is highly variable.





# DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

Some Tools for DE analysis:



| tool         | input        | language |
|--------------|--------------|----------|
| • EdgeR      | Count-matrix | R        |
| • limma-voom | Count-matrix | R        |
| • DEseq2     | Count-matrix | R        |
| • Cuffdiff   | BAM files    | python   |



# DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

## • Benchmark of DE genes analysis tools:

|               |                                                                                                                                                                         |
|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DESeq         | - Conservative with default settings. Becomes more conservative when outliers are introduced.                                                                           |
|               | - Generally low TPR.                                                                                                                                                    |
|               | - Poor FDR control with 2 samples/condition, good FDR control for larger sample sizes, also with outliers.                                                              |
|               | - Medium computational time requirement, increases slightly with sample size.                                                                                           |
| edgeR         | - Slightly liberal for small sample sizes with default settings. Becomes more liberal when outliers are introduced.                                                     |
|               | - Generally high TPR.                                                                                                                                                   |
|               | - Poor FDR control in many cases, worse with outliers.                                                                                                                  |
|               | - Medium computational time requirement, largely independent of sample size.                                                                                            |
| NBPSeg        | - Liberal for all sample sizes. Becomes more liberal when outliers are introduced.                                                                                      |
|               | - Medium TPR.                                                                                                                                                           |
|               | - Poor FDR control, worse with outliers. Often truly non-DE genes are among those with smallest p-values.                                                               |
|               | - Medium computational time requirement, increases slightly with sample size.                                                                                           |
| TSPM          | - Overall highly sample-size dependent performance.                                                                                                                     |
|               | - Liberal for small sample sizes, largely unaffected by outliers.                                                                                                       |
|               | - Very poor FDR control for small sample sizes, improves rapidly with increasing sample size. Largely unaffected by outliers.                                           |
|               | - When all genes are overdispersed, many truly non-DE genes are among the ones with smallest p-values. Remedied when the counts for some genes are Poisson distributed. |
|               | - Medium computational time requirement, largely independent of sample size.                                                                                            |
| voom /<br>vst | - Good type I error control, becomes more conservative when outliers are introduced.                                                                                    |
|               | - Low power for small sample sizes. Medium TPR for larger sample sizes.                                                                                                 |
|               | - Good FDR control except for simulation study $B_0^{4000}$ . Largely unaffected by introduction of outliers.                                                           |
|               | - Computationally fast.                                                                                                                                                 |



# DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

## Extracting biological meaning from DE gene lists

Once we have obtained a list of differentially expressed genes, we would like to search for a statistically significant association between:



| ID     | symbol  | description                                                     | logFC     | AvLog2     | t          | P.Value  | adj.P.Val | B          |
|--------|---------|-----------------------------------------------------------------|-----------|------------|------------|----------|-----------|------------|
| 8480   | NUO5    | regulator of G protein signaling 5                              | 4.7979759 | 18.6808641 | 16.5010550 | 4.67E-18 | 2.47E-14  | 11.020298  |
| 8305   | ACAD2   | acyl-CoA oxidase 2, long-chain fatty acid                       | 3.641077  | 8.3097120  | 26.321912  | 5.58E-18 | 2.67E-14  | 11.050376  |
| 3661   | FRS37   | G protein-coupled receptor 37 (proteinase receptor type B-like) | 4.1331868 | 1.9397611  | 26.1012487 | 6.68E-18 | 2.67E-14  | 10.864034  |
| 21584  | CDNA9   | chaperone, molecular chaperone, alpha 9                         | 5.0486021 | 7.6584121  | 24.9512396 | 7.79E-17 | 2.58E-14  | 10.890101  |
| 64133  | MYL7    | myosin heavy chain 7                                            | 3.5094042 | 4.5326031  | 24.3919724 | 2.78E-17 | 6.68E-14  | 10.243605  |
| 2370   | CEACAM3 | cell adhesion molecule, type 3, alpha 3                         | 6.4978605 | 7.7885245  | 23.8075445 | 5.54E-17 | 1.11E-13  | 10.890911  |
| 6006   | SPPL1   | secreted phosphatase 1                                          | 3.0384272 | 2.0100740  | 23.4210149 | 6.52E-17 | 1.12E-13  | 10.736026  |
| 64005  | FEAL    | FEAL, F5A1 sequence effector                                    | 5.2402228 | 4.6584612  | 23.3026781 | 7.49E-17 | 1.29E-13  | 10.736026  |
| 824    | CANX2   | catenin 2, E-cadherin large subunit                             | 3.7884402 | 19.3575017 | 22.6349065 | 1.34E-16 | 1.78E-13  | 10.051074  |
| 8871   | TYRO1   | tyrosinase 1                                                    | 5.2821008 | 4.6381374  | 22.405051  | 4.65E-16 | 1.98E-13  | 10.244740  |
| 6285   | SLC38   | SLC38, sodium-binding protein 8                                 | 2.5376669 | 8.7993101  | 22.3541889 | 1.73E-16 | 1.89E-13  | 10.780106  |
| 20991  | ITPR9   | inositol triphosphate receptor 9                                | 4.008212  | 1.5312009  | 22.3492507 | 3.46E-16 | 8.42E-13  | 10.722174  |
| 1368   | CPN1    | carboxypeptidase Y, polypeptide 1                               | 3.6252678 | 8.0687409  | 21.1702959 | 5.38E-16 | 4.97E-13  | 10.710654  |
| 5349   | PRX1    | PRX domain-containing protein regulator 1                       | 1.5412762 | 8.6386761  | 20.7377744 | 7.92E-16 | 8.79E-13  | 10.336443  |
| 80762  | TRPS    | trans-acting protein, inhibitor of transcription factor 5       | 1.1211044 | 4.1121822  | 20.4427686 | 1.11E-15 | 6.95E-13  | 10.007662  |
| 4342   | CPH4    | cytoplasmic alpha 4                                             | 1.5722059 | 7.2998807  | 20.1849335 | 1.40E-15 | 1.06E-12  | 10.750499  |
| 114889 | CTDTRF3 | C-terminal domain protein 3                                     | 1.5848493 | 7.7384644  | 20.1613013 | 1.50E-15 | 1.56E-12  | 10.7148971 |
| 7301   | TYRO3   | TYRO3 protein tyrosine kinase                                   | 1.4831703 | 7.4601687  | 20.0516354 | 1.65E-15 | 1.59E-12  | 10.614345  |
| 2068   | TNFR1   | tumor necrosis factor receptor 1                                | 2.1108441 | 6.1010889  | 19.8817031 | 2.08E-15 | 1.29E-12  | 10.421005  |
| 51857  | PLXN1   | plexin kinase 1 subunit 1                                       | 1.4009185 | 6.6122413  | 19.8014792 | 2.13E-15 | 1.29E-12  | 10.377494  |
| 64152  | MRJ2    | MRJ2, myosin 13 subunit protein ligase 2                        | 1.6489344 | 10.2465551 | 19.7650737 | 2.34E-15 | 1.29E-12  | 10.332109  |
| 1366   | PLA2G1B | phospholipase A2, group IIB, member 1                           | 3.1241311 | 7.8657603  | 19.7438599 | 2.28E-15 | 1.29E-12  | 10.307114  |
| 6848   | PC2     | PC2, procathepsin B, member 1                                   | 1.1348713 | 10.388421  | 19.6300582 | 2.50E-15 | 1.48E-12  | 10.310769  |
| 222    | ALDH3A3 | aldehyde dehydrogenase 3, family, member A3                     | 2.7208743 | 10.2462209 | 19.4742747 | 3.01E-15 | 1.48E-12  | 10.010786  |
| 4208   | MYO2C   | myosin II, heavy chain 2C                                       | 1.6277211 | 2.0922787  | 19.4031955 | 3.09E-15 | 1.48E-12  | 10.212200  |
| 5815   | MYO10   | myosin 10                                                       | 1.479642  | 2.5450821  | 19.3071876 | 4.00E-15 | 1.86E-12  | 10.751792  |
| 4685   | ATP2B2  | ATP2B2, heavy chain 2B2                                         | 1.5171727 | 1.9108917  | 18.9014460 | 5.57E-15 | 2.46E-12  | 10.464081  |
| 1222   | PLA2G1A | phospholipase A2, group I, member 1                             | 1.612614  | 8.8961091  | 18.8771049 | 5.76E-15 | 4.94E-12  | 10.401814  |
| 64397  | EPH2    | ephrin type 2 receptor (mouse)                                  | 1.8655451 | 1.8912617  | 18.6122948 | 7.52E-15 | 3.23E-12  | 10.432322  |
| 19477  | ETN1    | ectonucleoside triphosphate diphosphate phosphatase 1           | 1.5043277 | 1.1080041  | 18.5751566 | 8.00E-15 | 3.29E-12  | 10.020062  |
| 56304  | PDGFR   | platelet-derived growth factor receptor                         | 2.7181405 | 7.9311477  | 18.3642768 | 1.02E-14 | 3.95E-12  | 10.364257  |
| 811    | IGF1R   | insulin-like growth factor receptor 1                           | 1.2111013 | 1.0681979  | 18.3441349 | 1.16E-14 | 4.19E-12  | 10.710140  |
| 3680   | ITGA8   | integrin, alpha 8                                               | 1.8906099 | 6.0707256  | 18.2110244 | 1.13E-14 | 4.23E-12  | 10.890940  |
| 57550  | MTSL1   | metastasis-associated tumor suppressor 1                        | 2.210664  | 7.2042605  | 18.1113875 | 1.34E-14 | 4.58E-12  | 10.751705  |
| 57228  | SLC6A6  | solute carrier family 6, member 6                               | 1.8818055 | 6.2363808  | 18.101025  | 1.35E-14 | 4.58E-12  | 10.754166  |
| 1214   | PL2     | proline 2, long-chain fatty acid                                | 2.014036  | 1.0389123  | 18.0912430 | 1.37E-14 | 4.58E-12  | 10.348513  |
| 12332  | ACCT1   | acyl-CoA oxidase 1, long-chain fatty acid                       | 1.1238108 | 7.7226842  | 17.8700566 | 1.76E-14 | 5.79E-12  | 10.310761  |
| 5955   | IGFBP1  | insulin-like growth factor binding protein 1                    | 2.282046  | 1.8354051  | 17.7719416 | 2.09E-14 | 6.50E-12  | 10.333554  |
| 584    | CTN     | catenin                                                         | 3.181879  | 1.4471444  | 17.7144436 | 1.11E-14 | 6.99E-12  | 10.335104  |
| 10276  | MYO11   | myosin 11, heavy chain 11                                       | 2.112020  | 6.4098401  | 17.6513817 | 2.27E-14 | 6.82E-12  | 10.013372  |
| 6728   | ADAM19  | ADAM, metalloprotease domain 19                                 | 2.2847679 | 2.0262241  | 17.6332588 | 2.82E-14 | 6.82E-12  | 10.280280  |
| 1710   | EDNRB   | melanin-concentrating hormone receptor 1                        | 1.9751918 | 4.7434091  | 17.4219735 | 2.94E-14 | 8.42E-12  | 10.7972215 |
| 150    | ANKK1   | ankyrin repeat domain 1                                         | 3.3474794 | 1.8751881  | 17.3641881 | 3.40E-14 | 1.21E-11  | 10.368010  |
| 2012   | EMPH1   | epithelial membrane protein 1                                   | 3.390447  | 2.2466807  | 17.3877582 | 4.42E-14 | 1.21E-11  | 10.395160  |
| 2314   | SPN     | spinophilin                                                     | 2.4290644 | 10.2209460 | 16.8010212 | 5.26E-14 | 2.06E-11  | 10.221000  |
| 6838   | WDR5    | WD repeat domain 5                                              | 1.6623518 | 8.7250617  | 16.7613413 | 5.32E-14 | 1.36E-11  | 10.711684  |
| 10642  | MYO17   | myosin 17, heavy chain 17                                       | 2.824040  | 1.0389123  | 16.9136493 | 6.33E-14 | 1.36E-11  | 10.280960  |
| 54898  | ILKAP2  | ILKAP2, beta and epsilon 2                                      | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 4907   | WDR5    | WD repeat domain 5                                              | 1.5452379 | 4.4272213  | 16.6880712 | 7.14E-14 | 1.74E-11  | 10.280960  |
| 5447   | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 110372 | WDR5    | WD repeat domain 5                                              | 1.5452379 | 4.4272213  | 16.6880712 | 7.14E-14 | 1.74E-11  | 10.280960  |
| 61442  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 7262   | PRKDA2  | protein kinase domain 2                                         | 1.8877483 | 7.2751473  | 16.4691308 | 9.26E-14 | 2.06E-11  | 10.663863  |
| 6440   | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 6299   | SALL1   | sall-like 1 (Drosophila)                                        | 3.0392668 | 6.2115148  | 16.4110484 | 9.95E-14 | 2.19E-11  | 10.590215  |
| 8112   | PCYT1   | phosphatidylcholine transferase 1                               | 1.0670262 | 1.0389123  | 16.4011011 | 1.01E-13 | 1.01E-11  | 10.000000  |
| 875    | CB      | cytochrome b5                                                   | 1.0670262 | 1.0389123  | 16.4011011 | 1.01E-13 | 1.01E-11  | 10.000000  |
| 935    | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 9445   | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 29135  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 54993  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 64666  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 70002  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 1376   | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 24122  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 1409   | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 8134   | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 21336  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 1675   | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |
| 11411  | WDR5    | WD repeat domain 5                                              | 1.1991701 | 1.8027912  | 16.7514718 | 6.78E-14 | 1.59E-11  | 10.107179  |



And now what ?



## Extracting biological meaning from DE gene lists



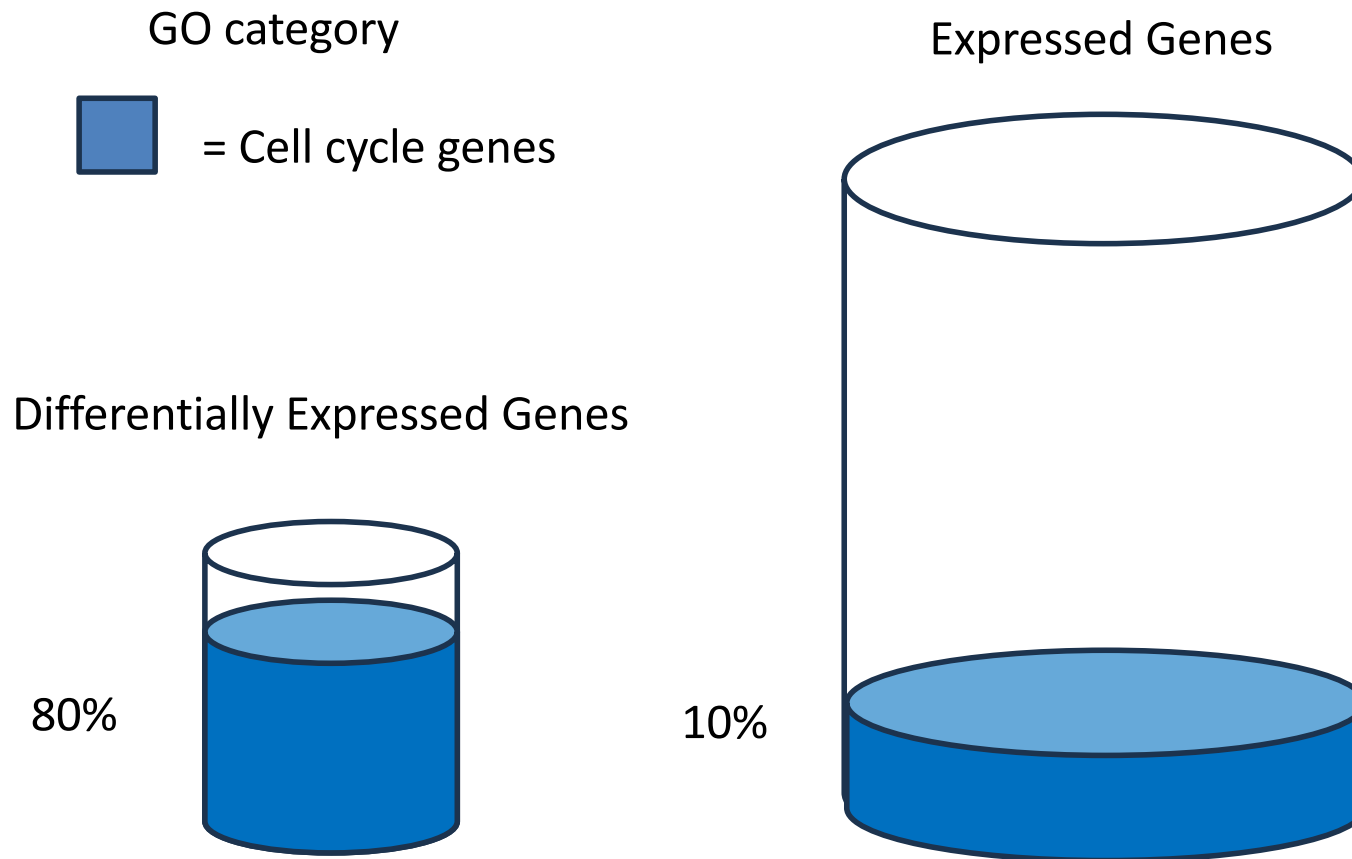
What do we need to perform a functional enrichment analysis?

- A list of “interesting” genes.
- A background gene list, representing the “universe” of possible genes that could be called as significantly regulated in the experiment. This list should contain only genes that are “called” as expressed (to avoid biological bias) in the experiment.
- Functional categories into which we can classify genes.
- A test which is able to tell what categories are significantly over or under-represented in our list compared to background.



# DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

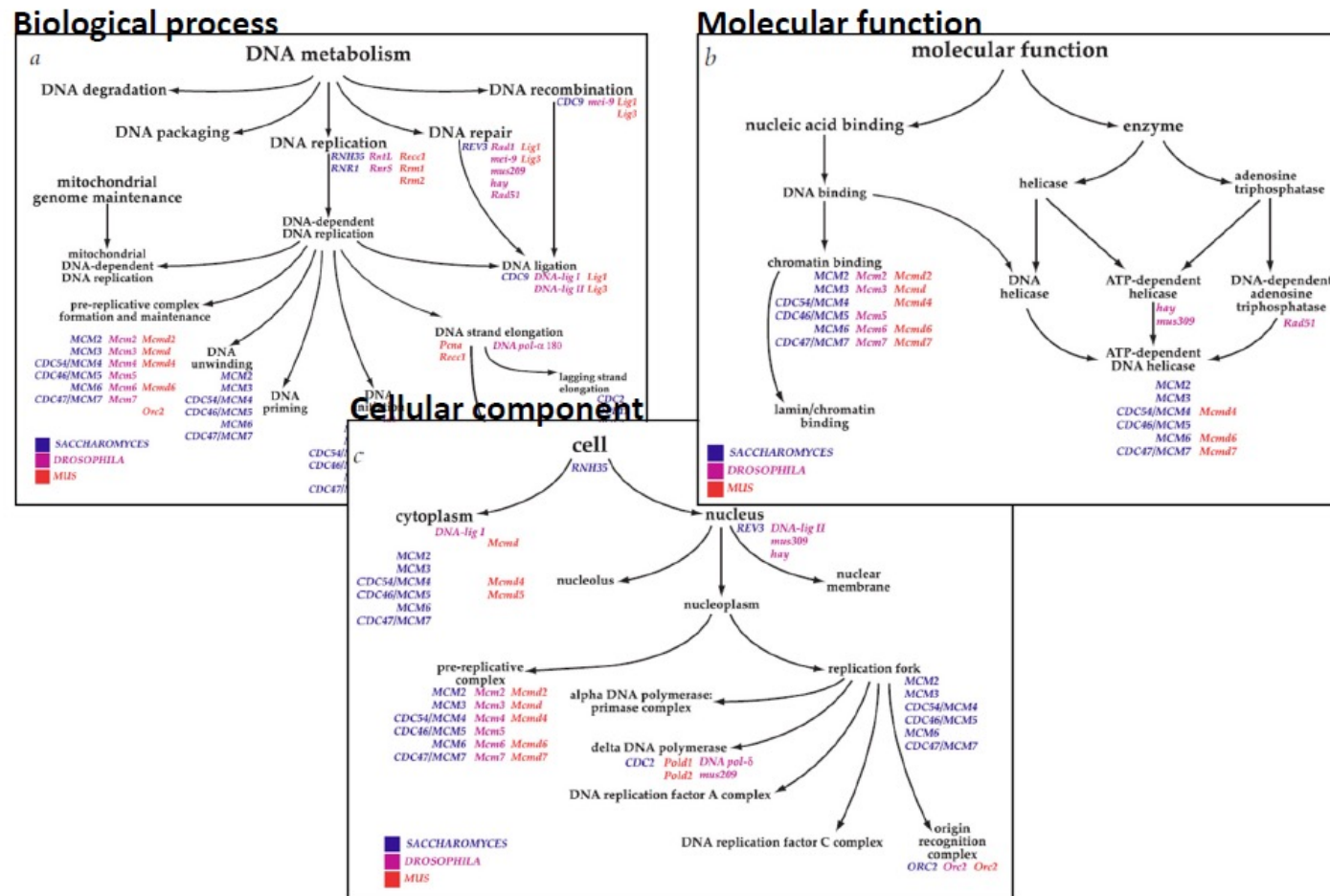
## Extracting biological meaning from DE gene lists





# DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

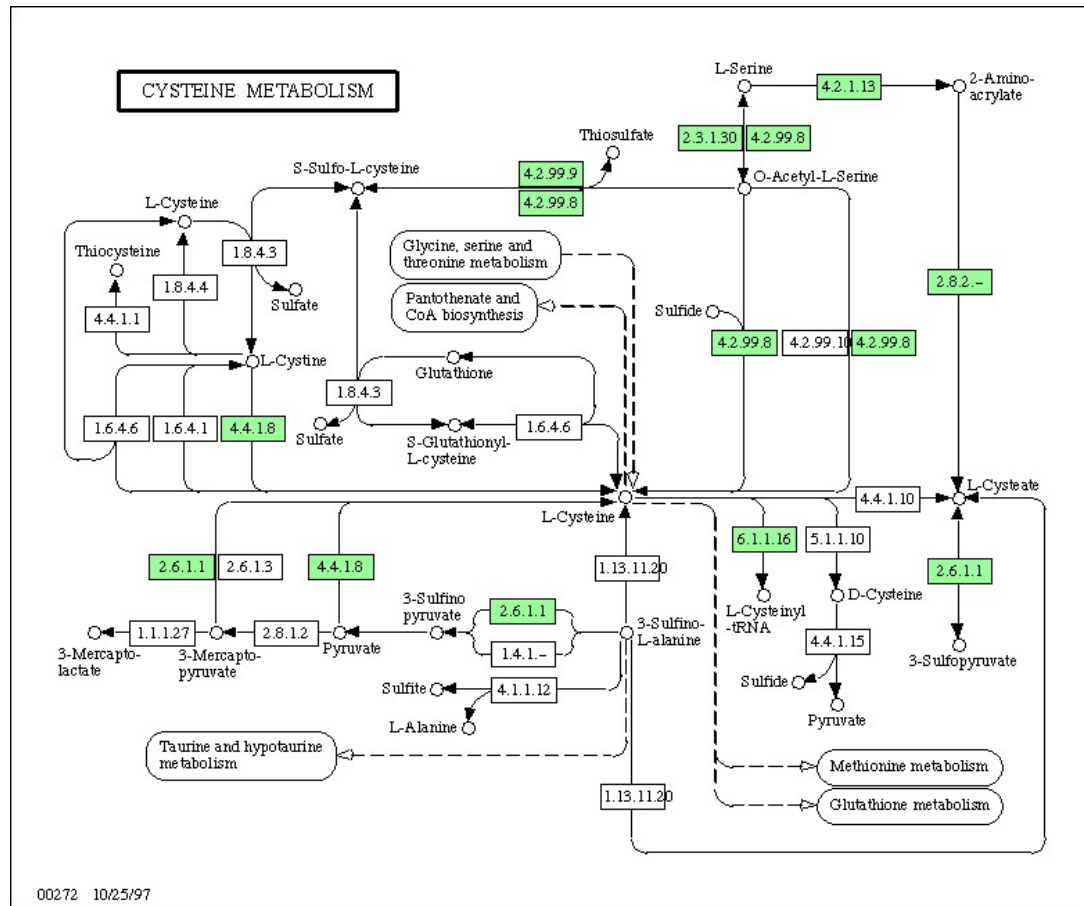
## Example of functional categories: Gene Ontology.





# DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

## Example of functional categories: Kegg pathway.



**KEGG PATHWAY** is a collection of manually drawn **pathway** maps representing our knowledge of the molecular interaction, reaction and relation networks for: 1. Metabolism

[1. Metabolism](#)

[2. Genetic Information Processing](#)

[3. Environmental Information Processing](#)

[4. Cellular Processes](#)

[5. Organismal Systems](#)

[6. Human Diseases](#)

[7. Drug Development](#)



# DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of online functional annotation tools.



**WEB-based GENE SeT AnaLysis Toolkit**

*Translating gene lists into biological insights...*

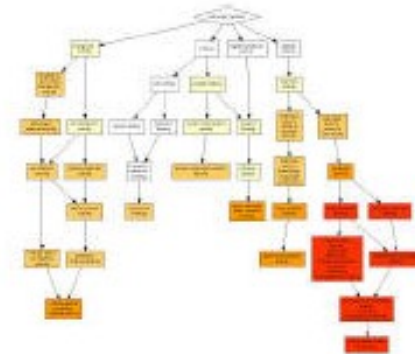


**DAVID Bioinformatics Resources 6.8**

Laboratory of Human Retrovirology and Immunoinformatics (LHRI)



**GORILLA**



*Gene Ontology enRICHment anaLysis and visuaLizAtion tool*