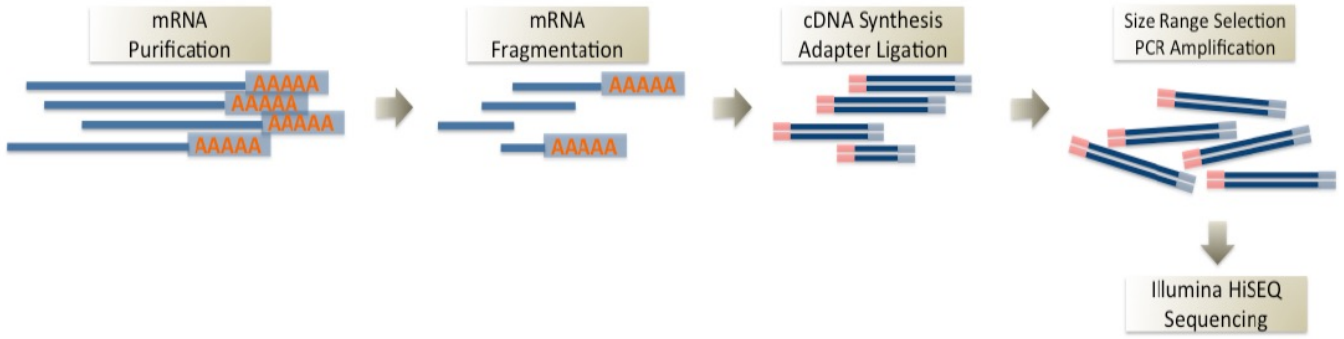
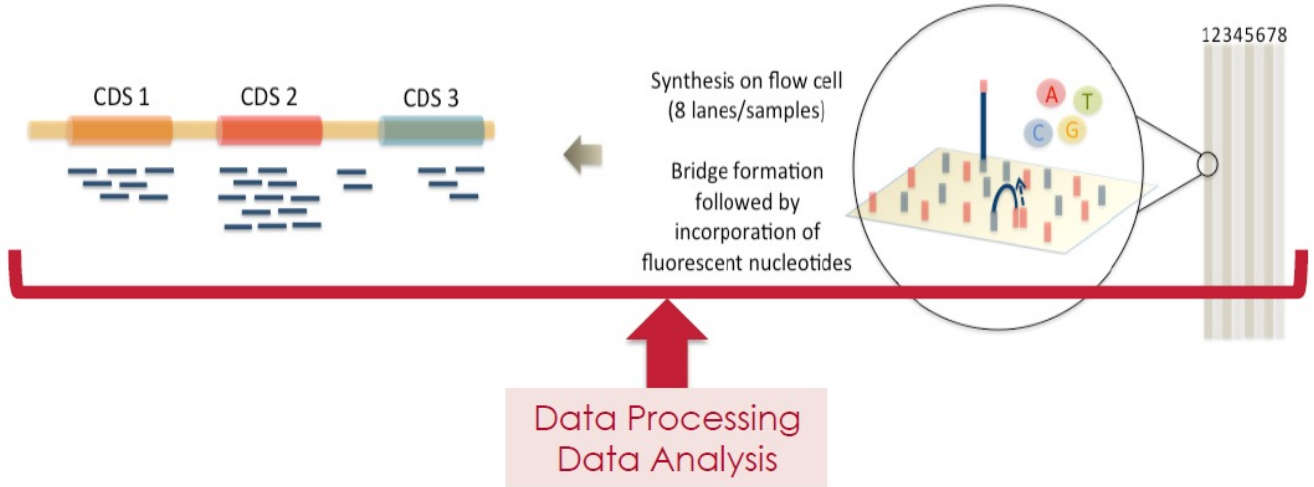
What is RNA-seq?

- 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.

The method



Library preparation



Sequencing and Data Analysis

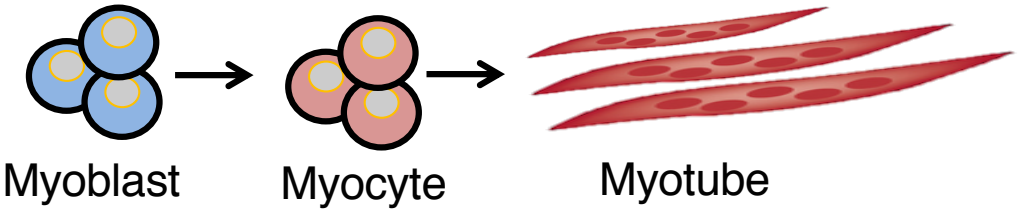
OVERVEW OF RNA-Seq EXPERIMENT

Examples of experimental design:

Treatment effect:

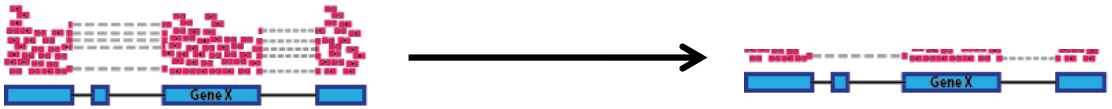
Tissue comparison:

Cell Differentiation:



Gene function:

siRNA / CRISPR-Cas9 /…



….

OVERVEW OF RNA-Seq EXPERIMENT

RNA-Seq leads you to **Identify** snd **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**
- **…**

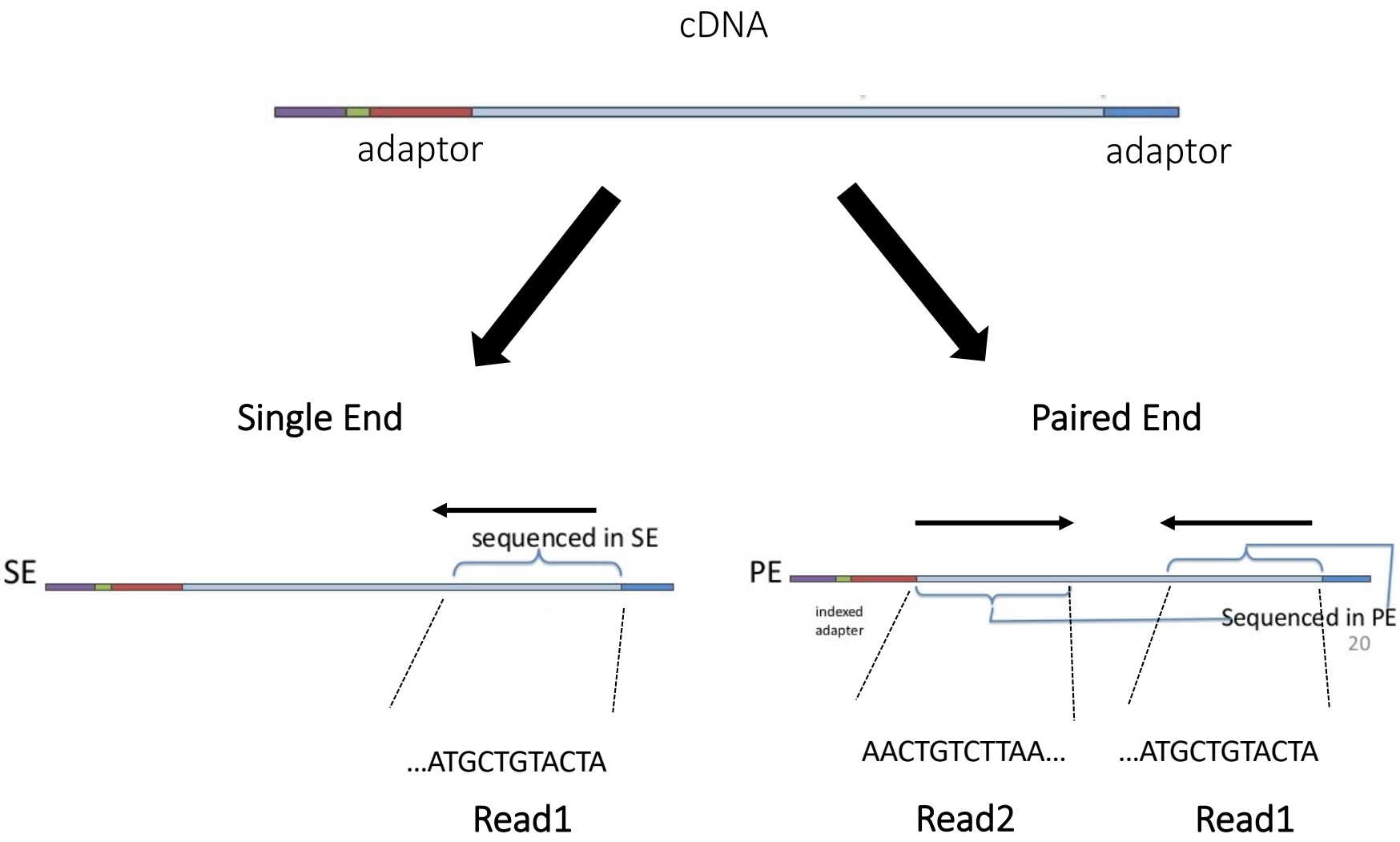
Defining the technical details

- **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).

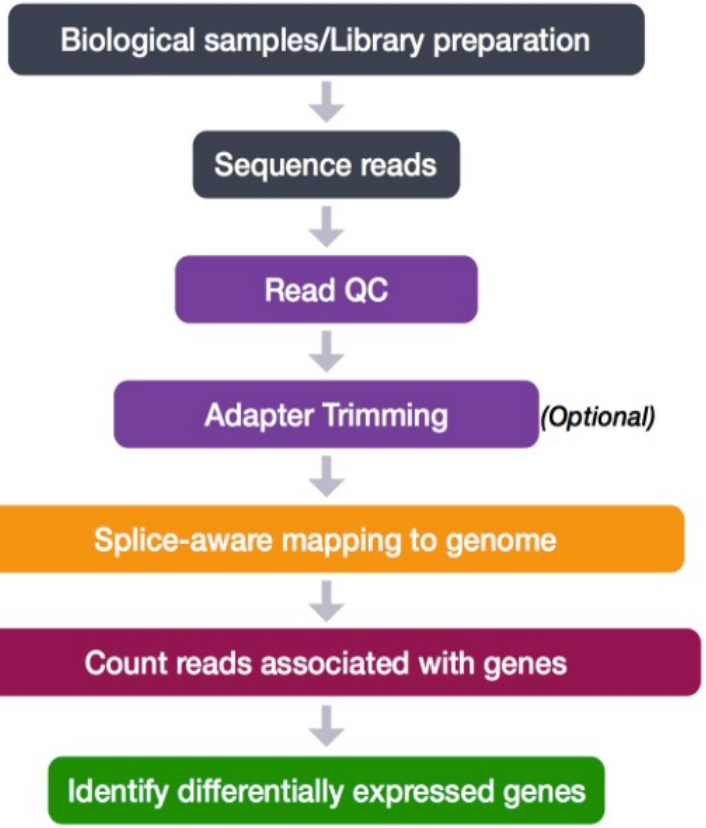
- **Length and pairing of reads**
  Theoretically speaking, read length should be > 20 bp (they usually are longer than 35 bp). PE reads are usually better (except for small RNA-Seq and Ribo-Seq), but they are more expensive.

- **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

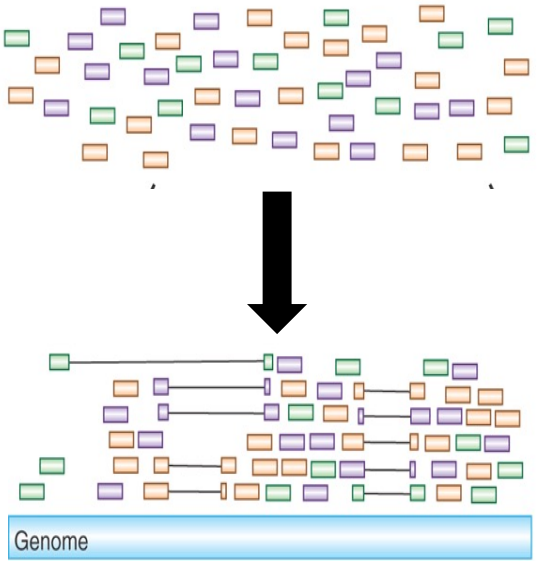


General RNA-Seq pipeline for Differential Expression



Condition A	Condition B
Sample1 Sample2 ..N	Sample1 Sample2 …N

For each sample..



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

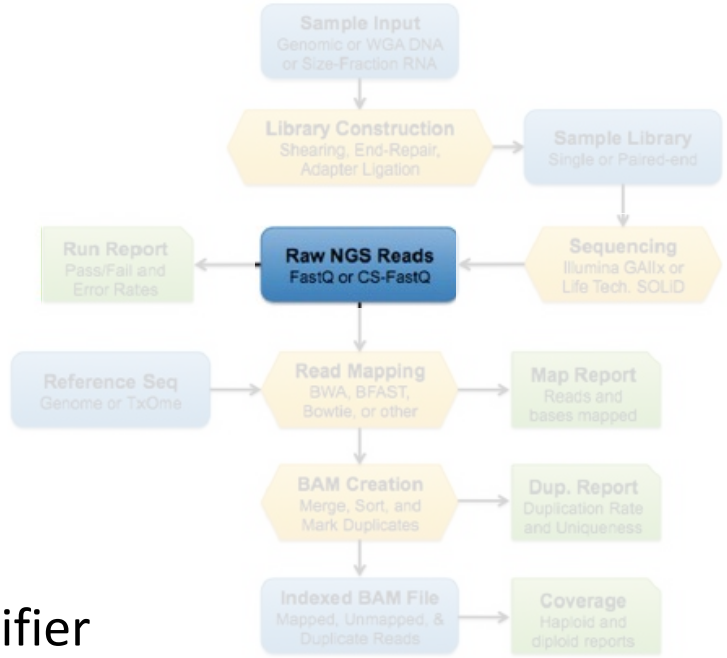
Data format

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
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

FASTQ format

```
@SEQILMN03:128:HA5CBADXX:1:1101:1186:2059 2:N:0:GTCGTA
NNNNNNGTTAAGATTATTGTCATTGGCTAACTAAGCGCTACCAAGTACAAGTACAAATGC
+
############0#0<BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB<<<<<<<<<<<<<<<<<<
@SEQILMN03:128:HA5CBADXX:1:1101:1193:2104 2:N:0:GTCGTA
CTATCTTCGTAACCCAAAATAAATAAACTAACTCTATTTCTTGTGTTAGGCAGGGTATTCC
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFB707BFFIIIIIIIIIIIIFFFFFFFFFF<BBFFF
@SEQILMN03:128:HA5CBADXX:1:1101:1227:2106 2:N:0:GTCGTA
GGGGAGCATGACGGCCCACATCGGCGAAAACCCACTCTGGTGGGGTGAACCGGTATCCAN
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIFFFFFFBBFFFBBFBFFBBFF<BBFF0<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.

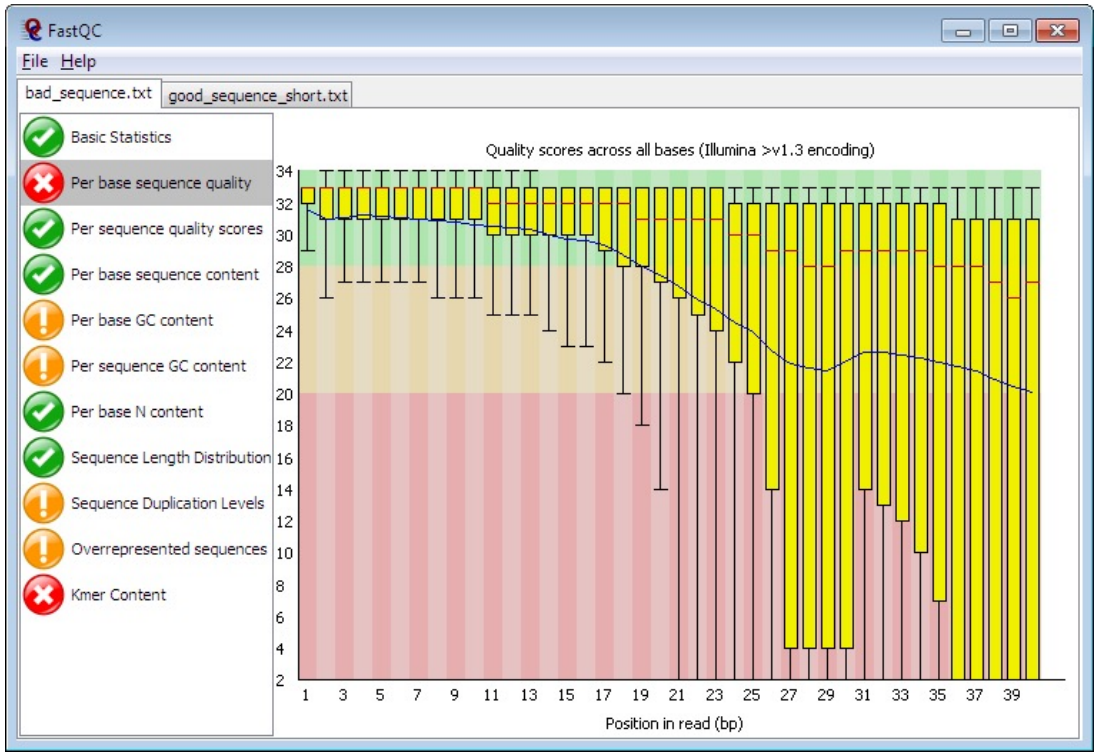
$$Q = -10 \log_{10} P$$

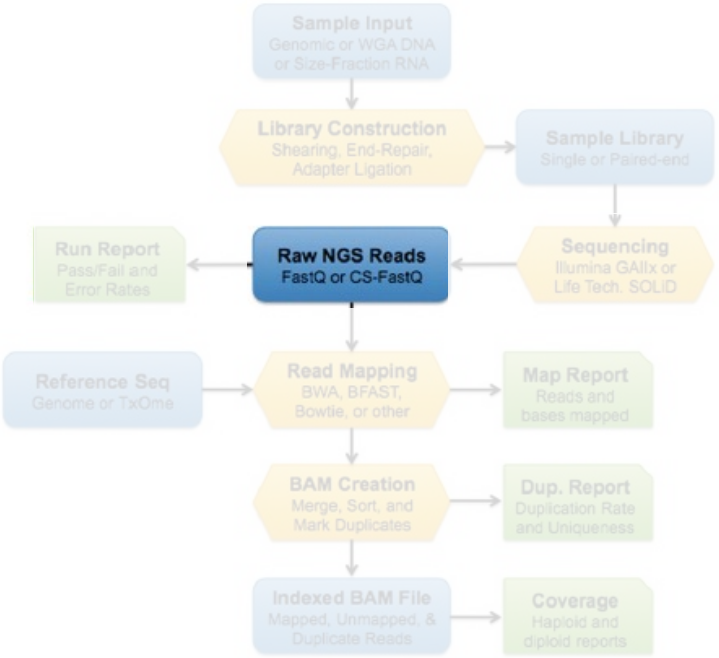
A high quality score implies that a base call is more reliable and less likely to be incorrect. For example, for base calls with a quality score of Q40, one base call in 10,000 is predicted to be incorrect. For base calls with a quality score of Q30, one base call in 1,000 is predicted to be incorrect. 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%

FastQC

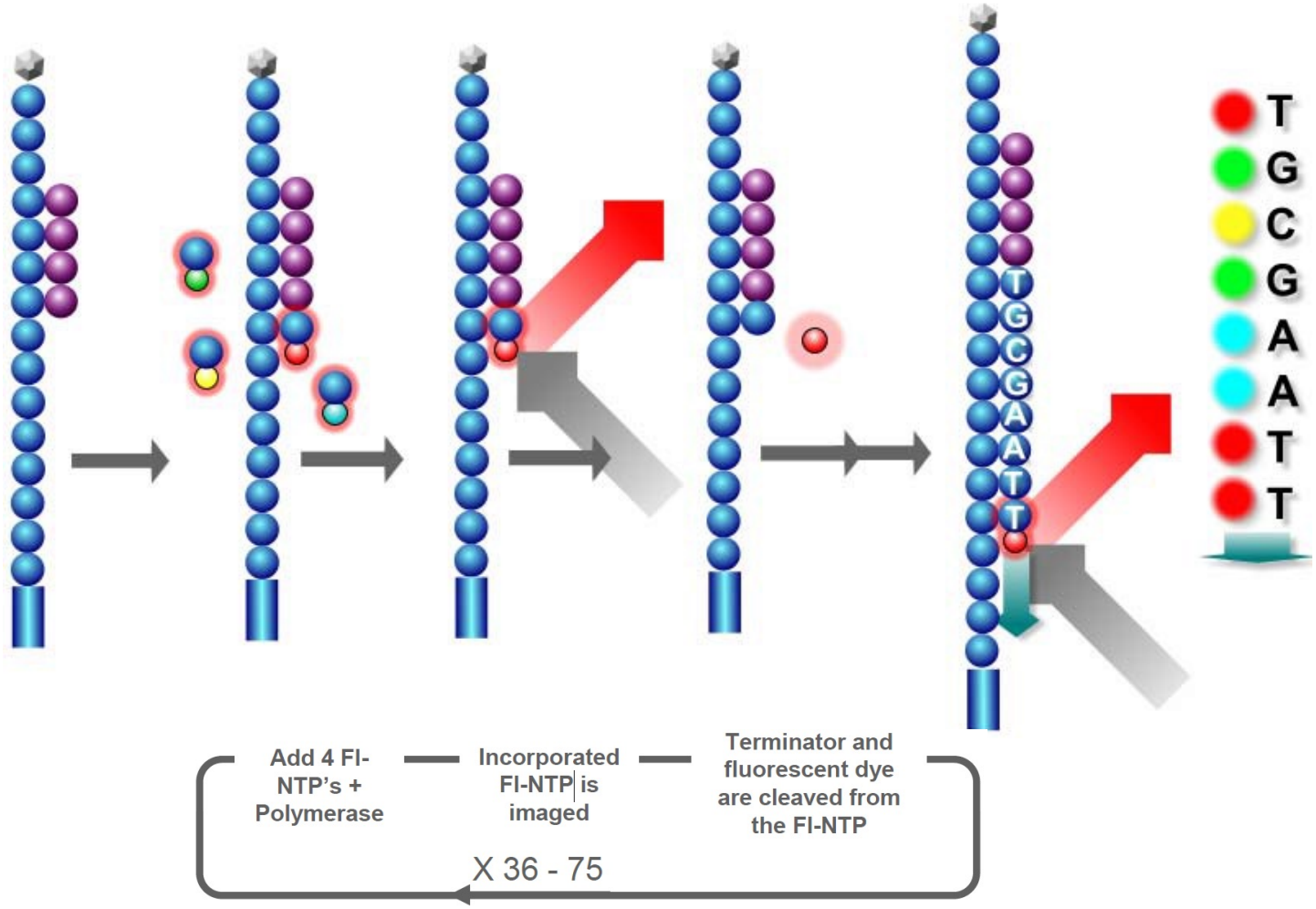
FastQC is a quality control tool for high throughput sequence data.





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

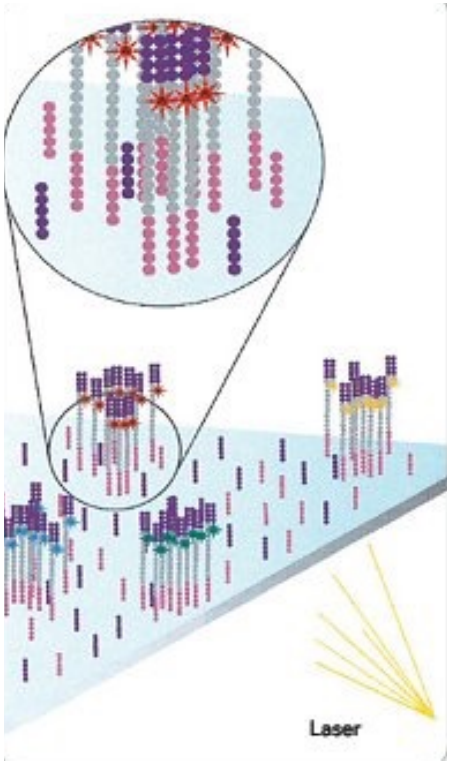
RNA-Seq: SEQUENCING REACTION



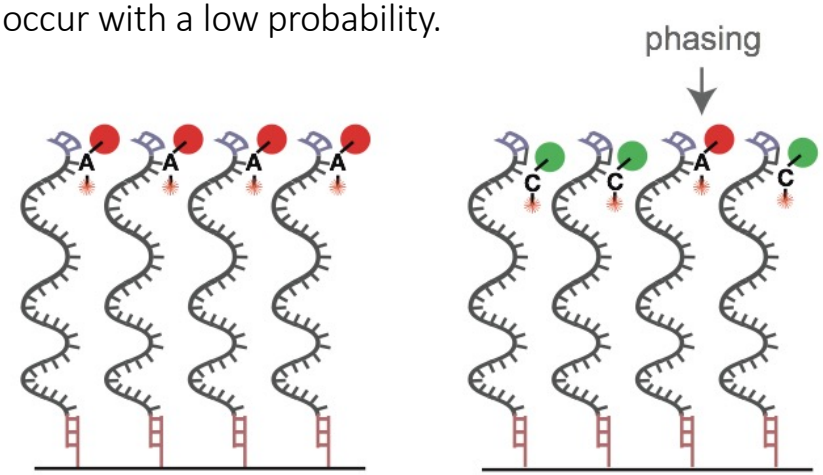
READ LENGTH = Number of reaction cycles

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.

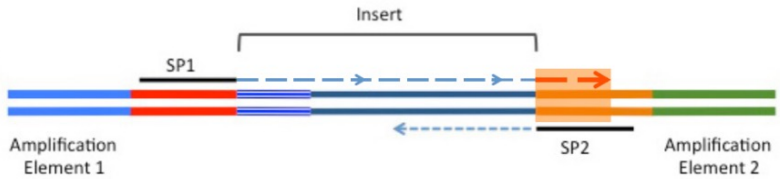


But over time (with increasing read length) they add up and pollute the light signal more and more. The signal gets more and more asynchronous. And since the light signal is used to calculate quality scores the asynchronous signal results in a decreasing sequence quality score.

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.

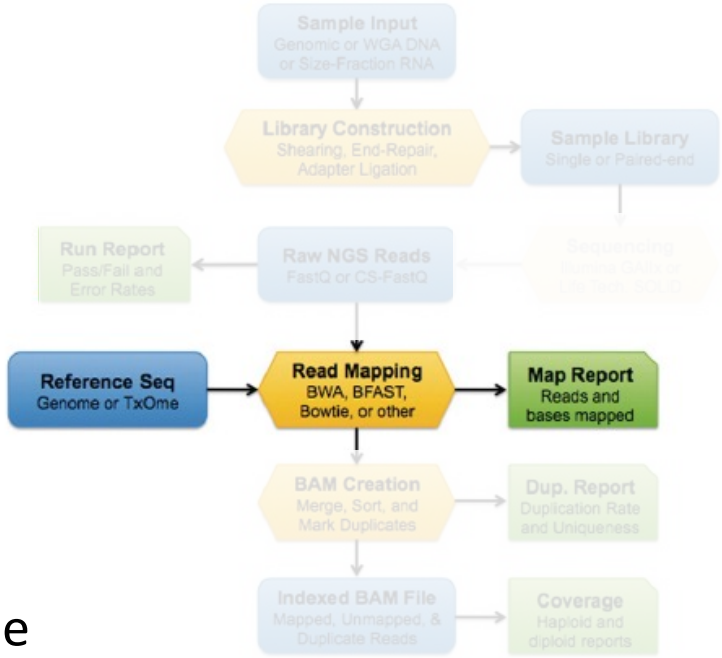


- If the overall quality of the read is low, it must be removed. A trimming is useful if quality decreases too much towards the end of the read.
- Sometimes the read terminates with ambiguous (N) bases which must be removed.
- Some of the most common preprocessing tool are FASTX-Toolkit, Cutadapt, Trimmomatic, 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).

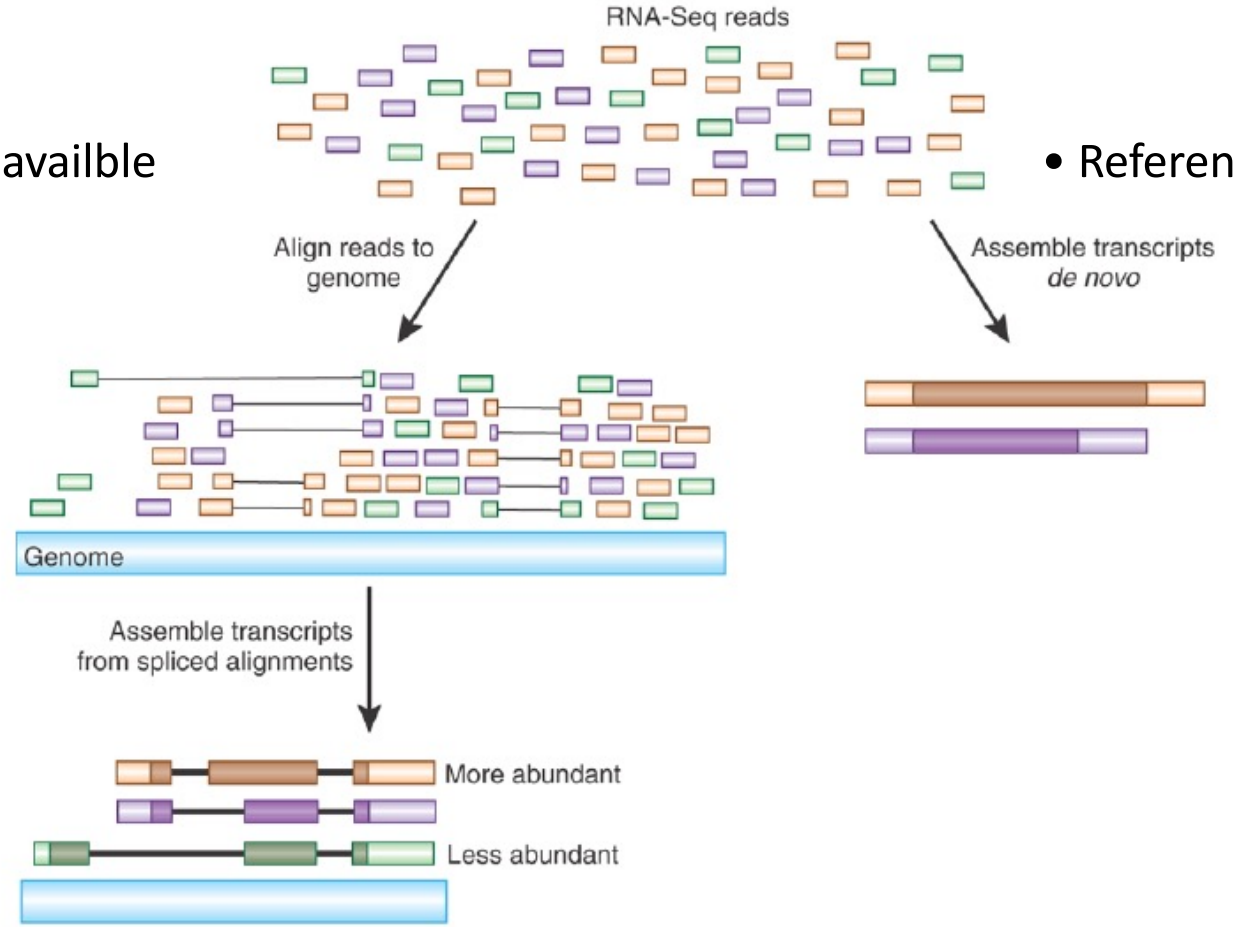
The accurate and fast alignment of millions of reads is not a simple task: many programs have been developed to address this issue.

Read alignment

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

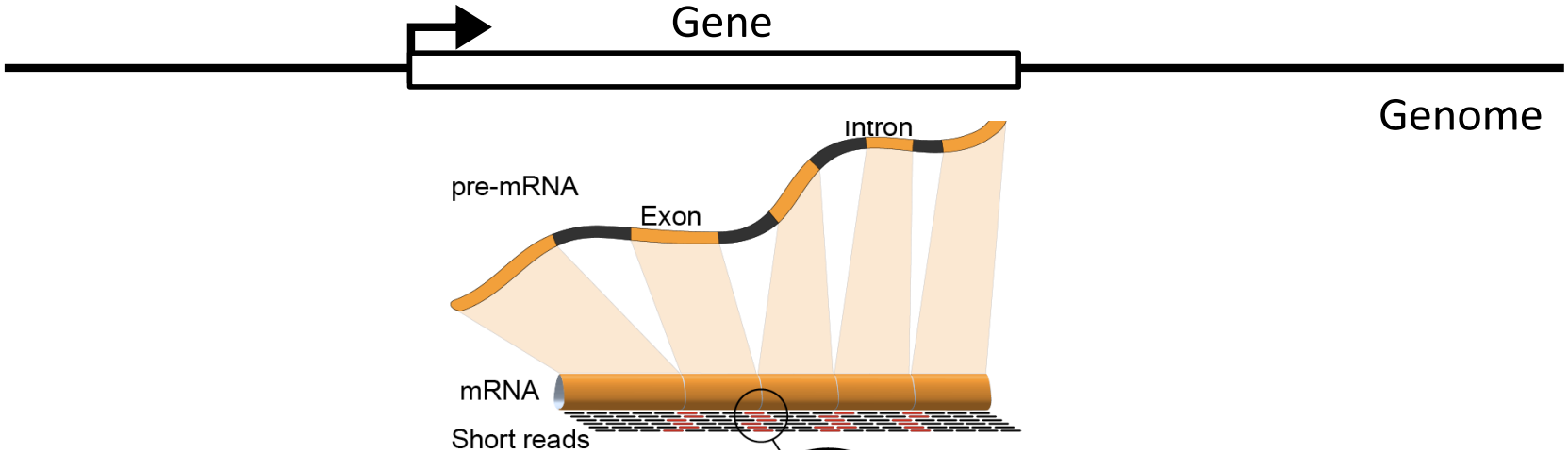
Main alignment strategies

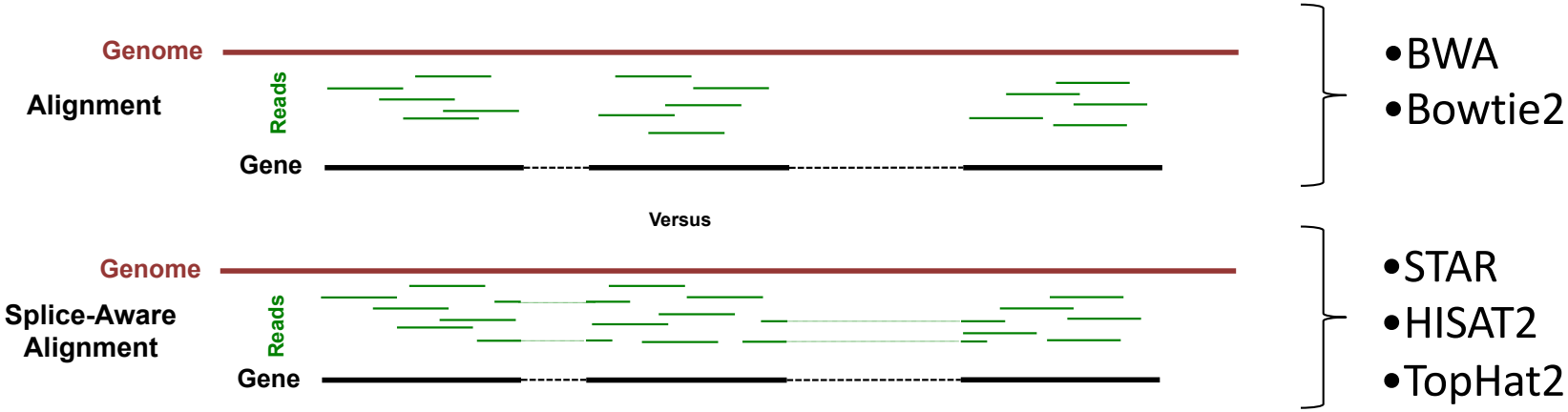
- Reference availble
- Reference not availble



- transcriptome

DATA ANALYSIS: ALIGNMENT





Splice-aware 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

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.

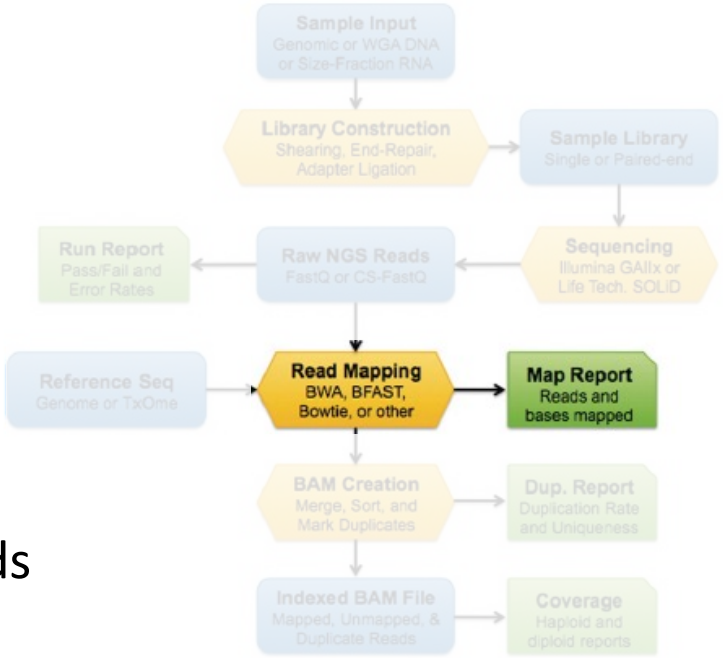
Main alignment programs

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

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.

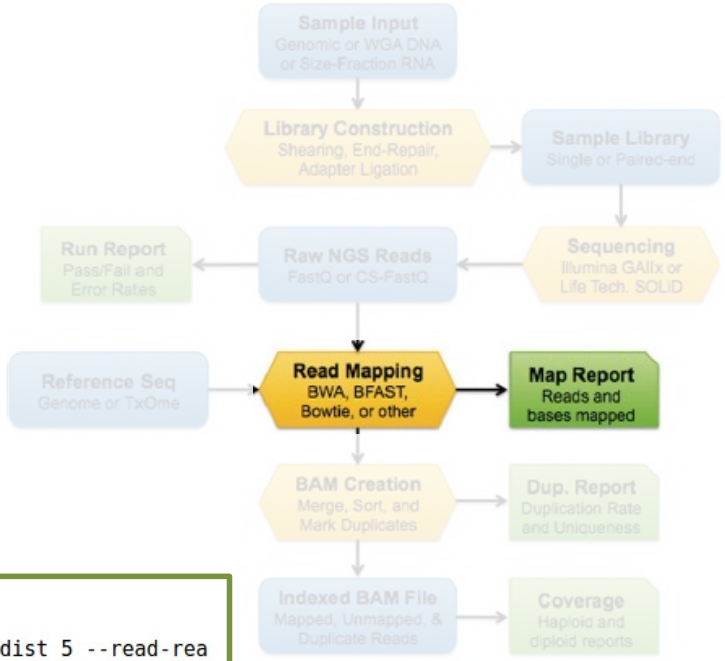
DATA ANALYSIS: ALIGNMENT

SAM file structure

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

- **header** reports **general information**.

- **body reports information about reads**. Each line describes a read (aligned or not): alignment position, sequence, quality etc.

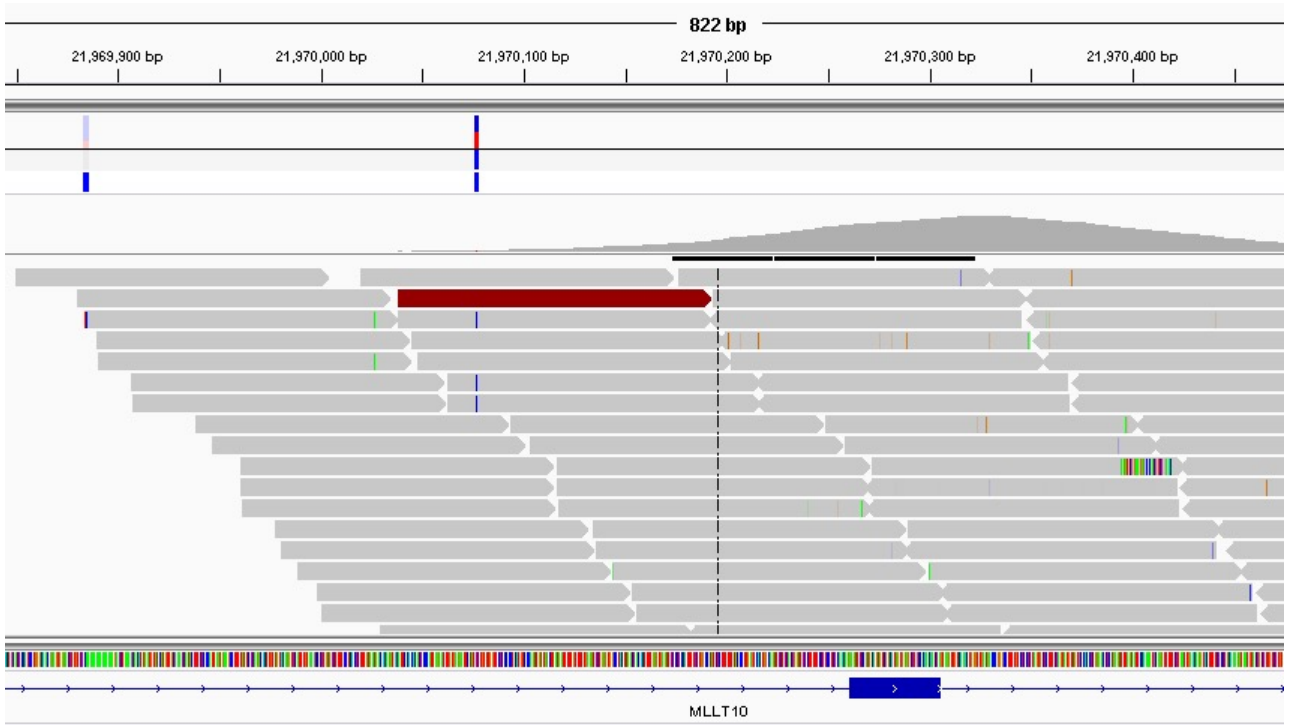



```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

BAM file visualization

IGV

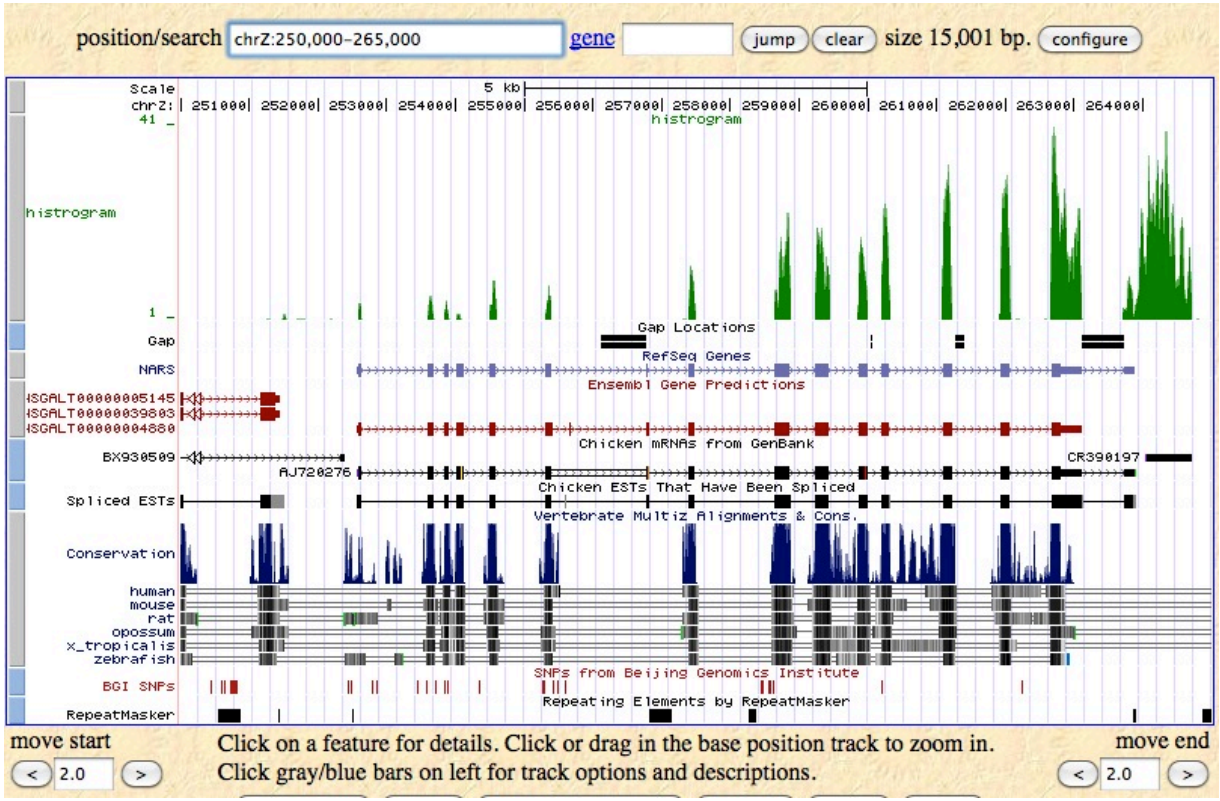
IGV is a standalone program which allows a highly interactive visualization of BAM files (and other genomic annotation formats).



BAM file visualization

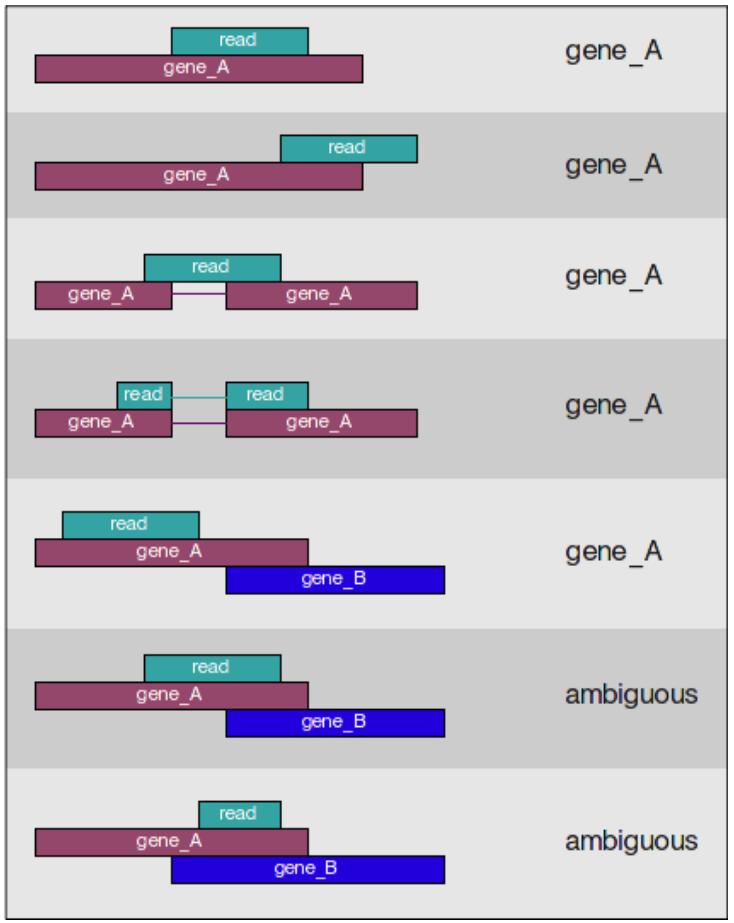
Genome Browser (UCSC)

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:

Count matrix

Each column is a sample

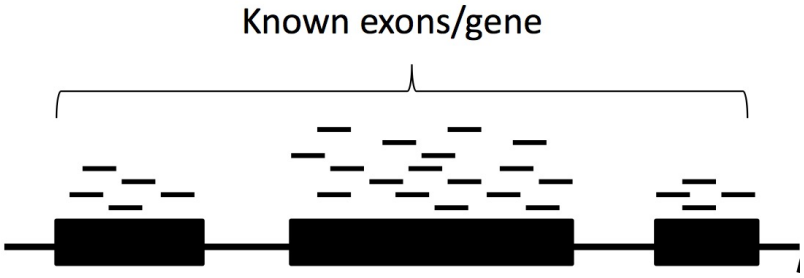
Each row is a gene

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

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)



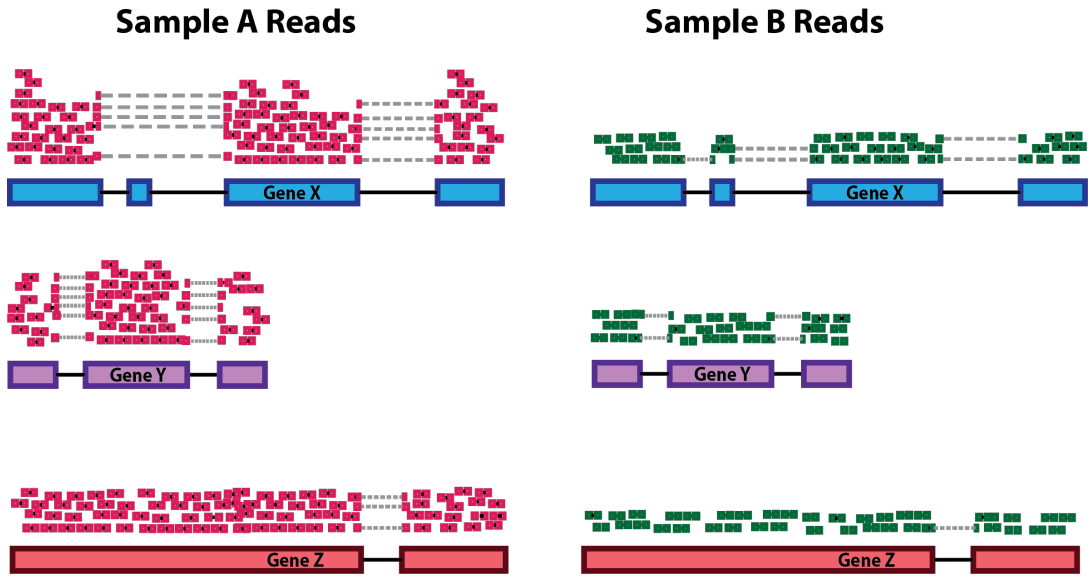
- The raw number of reads mapping on a gene (**read count**) requires a normalization.

Why?

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Why normalization is required before DE analyis?

- **Sequencing Depth**



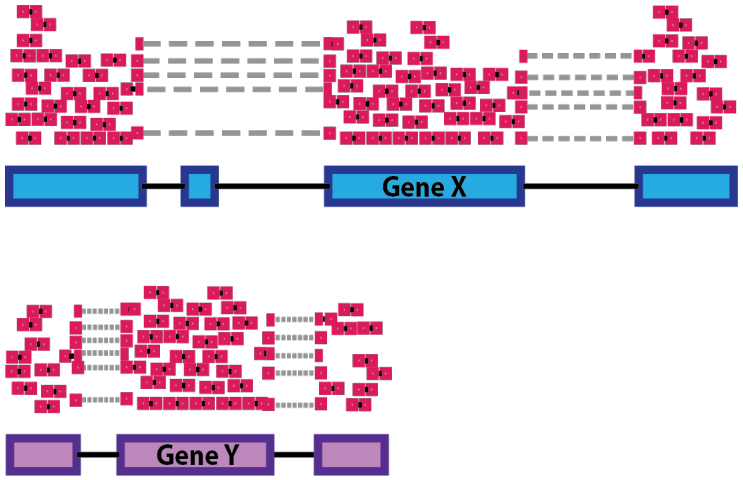
the number of reads mapped on a gene depends on sequencing depth: to normalize for the total number of mapped reads is important to compare the expression levels of the same gene obtained from two different sequencing experiments.

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Why normalization is required before DE analyis?

- **Gene length**

Sample A Reads



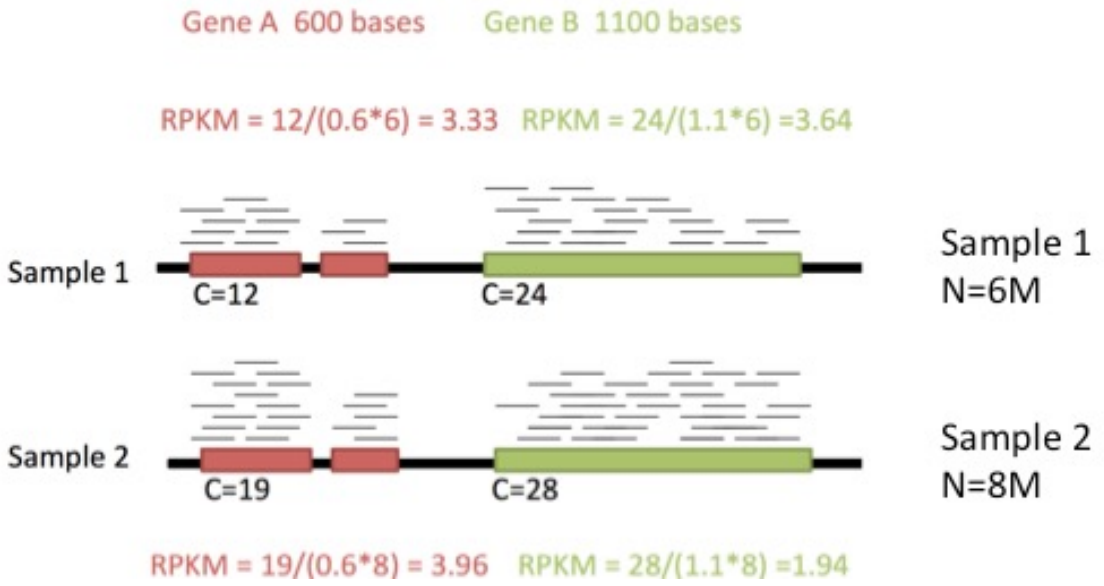
longer genes will have a greater number of reads mapped on them compared to equally expressed shorter genes: to normalize for gene length is important to compare the expression of distinct genes.

Measures of gene expression: RPKM

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

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

- 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)



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.

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

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**	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)**	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**

DATA ANALYSIS: DE NOVO DISCOVERY

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 Genome-guided assembly	G.Mor.Se Scripture[28] Cufflinks[29]	Assembles exons Reports all isoforms 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[61] TransABySS[56] Trinity	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads

DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

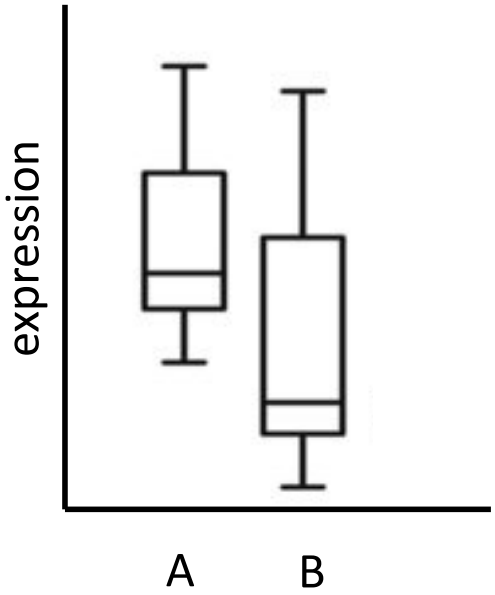
What is differential expression (DE) analysis?

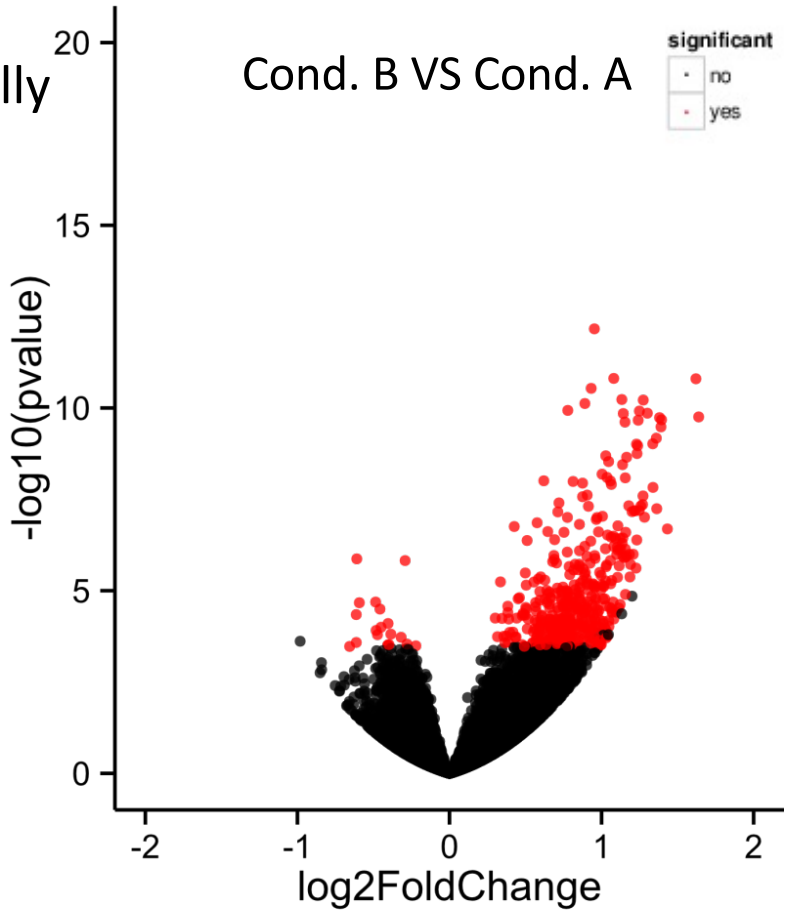
- DE analysis allows to find **genes** (or other genomic features like transcripts and exons) **that are expressed at significantly different levels between two groups of samples** (conditions): patients treated with drugs VS controls, healthy VS sick individuals , different tissues and different differentiation 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.

FC vs p-value

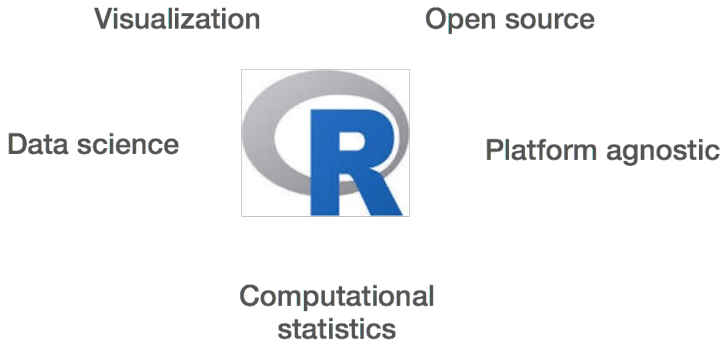
High absolute FC values are not necessarily associated with significant P-values, especially when the expression of the gene is highly variable.





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

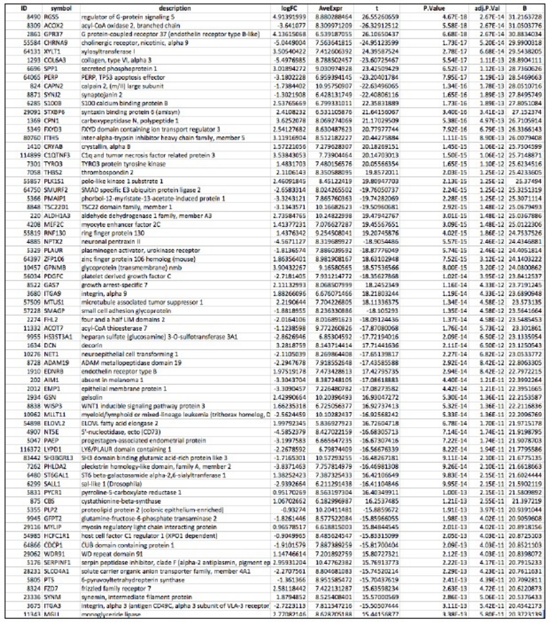
- Benchmark of DE genes analysis tools:

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.
NBPSeq	- 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 / 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.

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:

Group of genes ⟷ **Biological processes** ⟷ **Phenotype**



And now what ?

DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Extracting biological meaning from DE gene lists

- What do we need to perform a functional enrichment analysis?

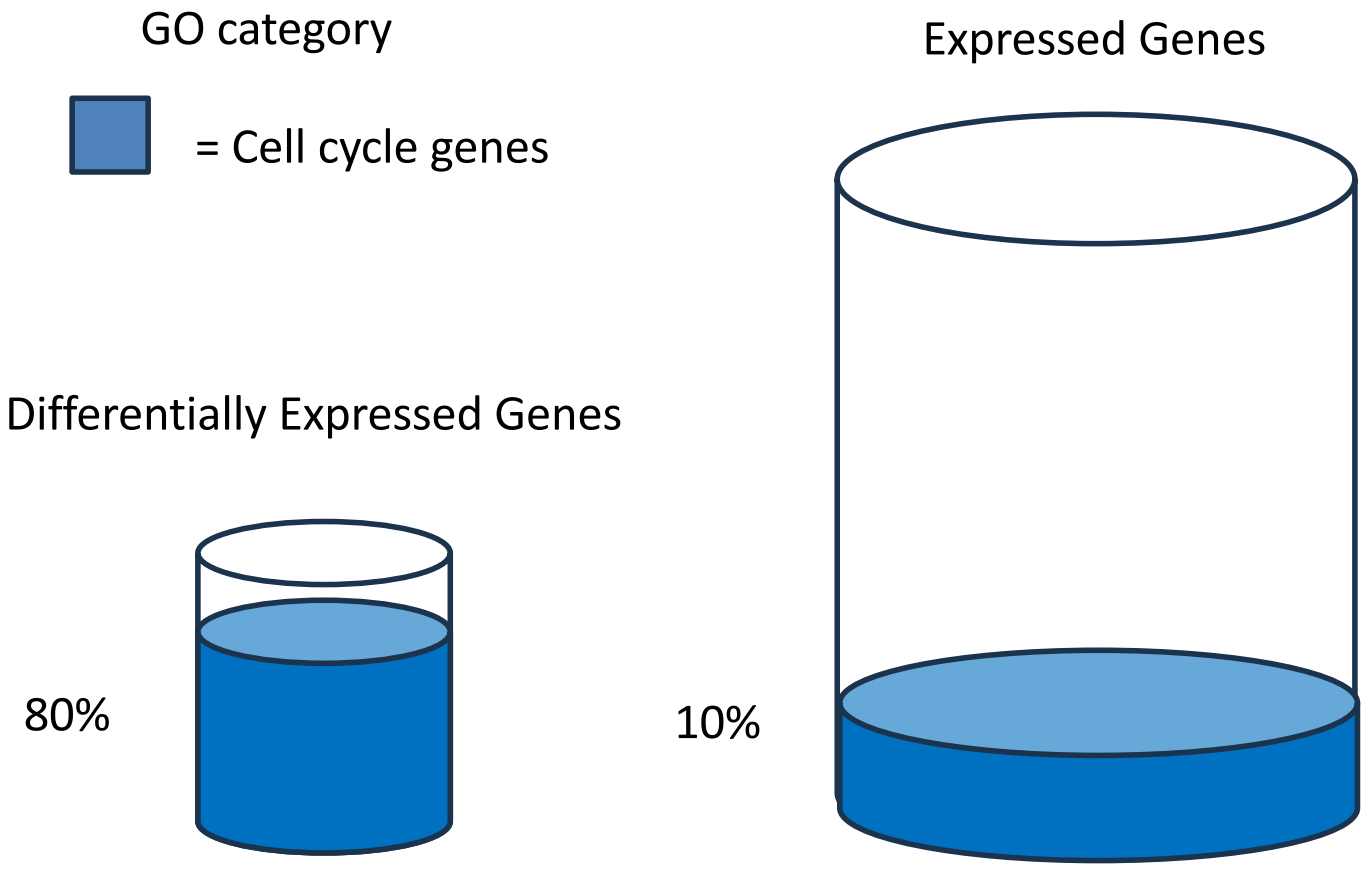
- A list of “interesting” genes.

- 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.

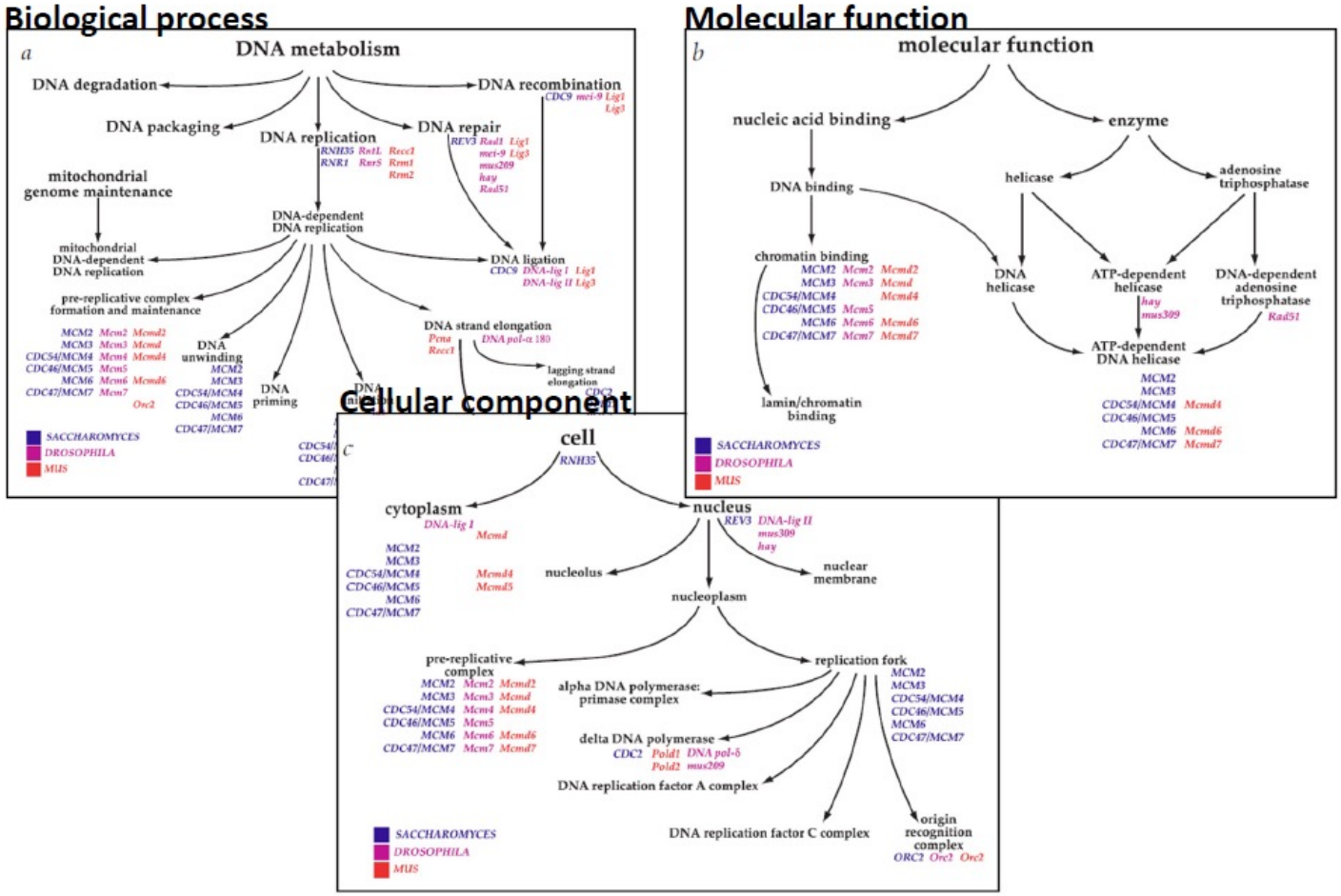
- A test which is able to tell what categories are significantly over or under-represented in our list compared to background.

Extracting biological meaning from DE gene lists



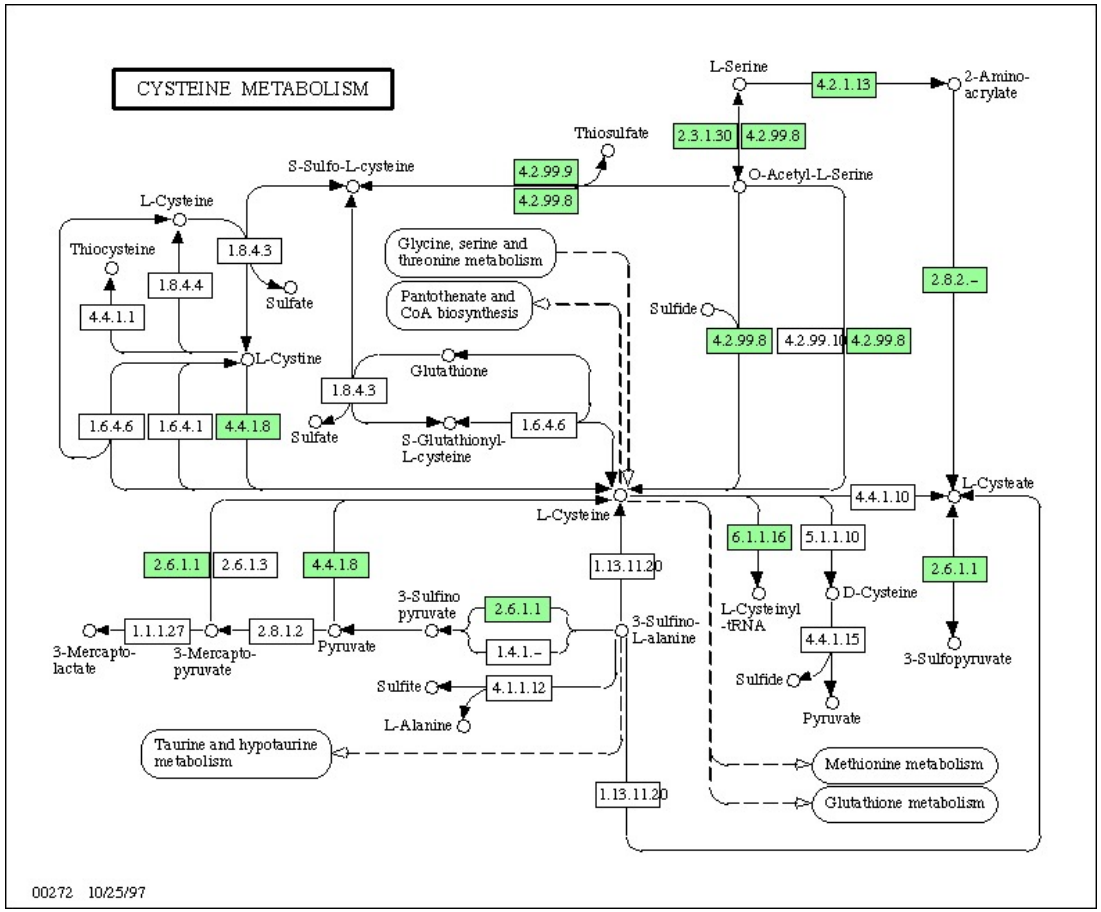
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

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.

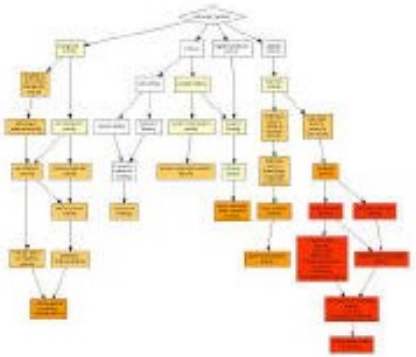
WebGestalt

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