

# EXPERIMENTAL DESIGN



**"There's a flaw in your experimental design.  
All the mice are scorpions."**

CN  
COLLECTION

## Defining the samples to be studied

### ■ Number of samples

**Biological replicates are parallel measures of biologically distinct samples,** which allow to capture random biological variations.

**Technical replicates are repeated measures of the same sample,** that represent independent measures of the random noise associated with protocols or equipment.

The greater the number of the biological replicates, the more we can trust the results, especially when testing for differential expression. With only one biological replicate, no statistical test can be performed.

## 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).

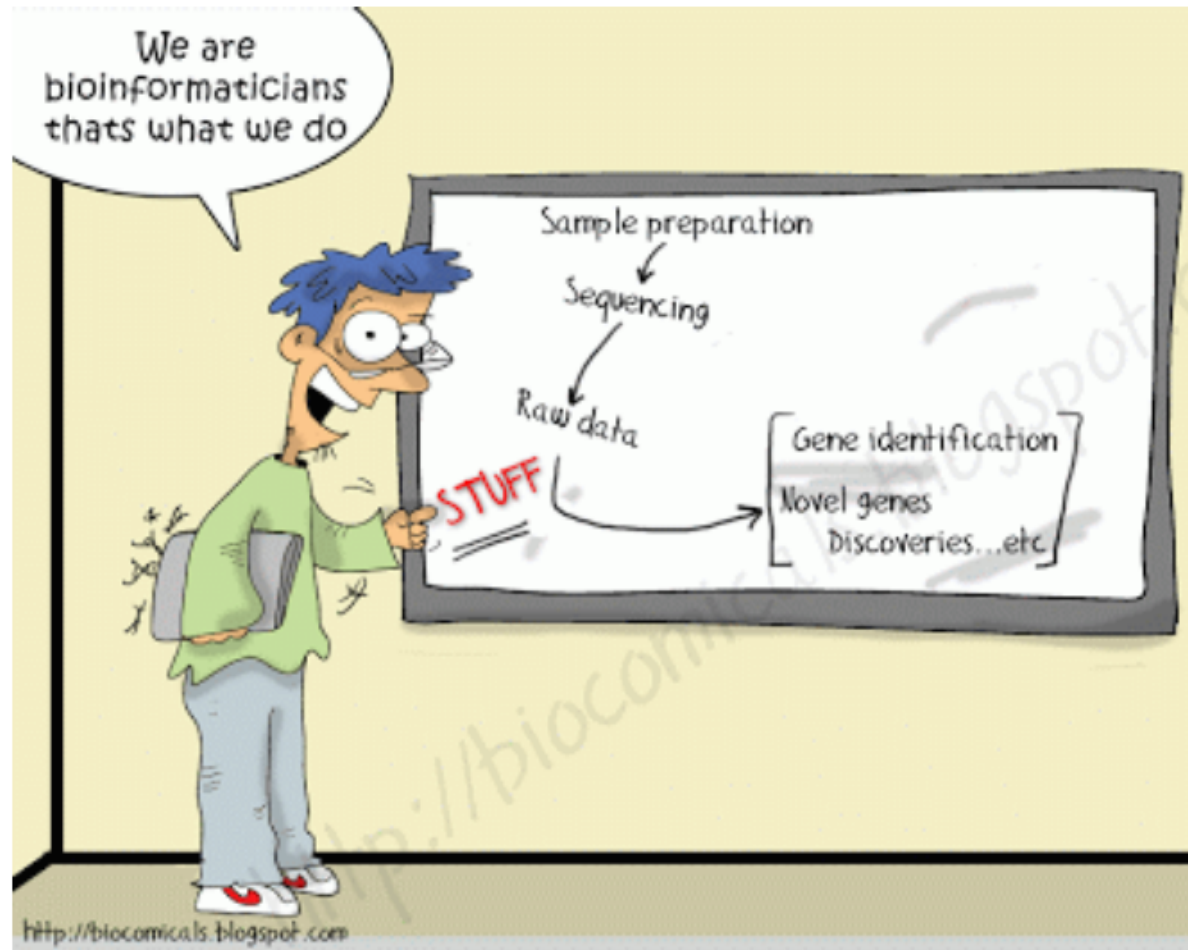
### Length and pairing of reads

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

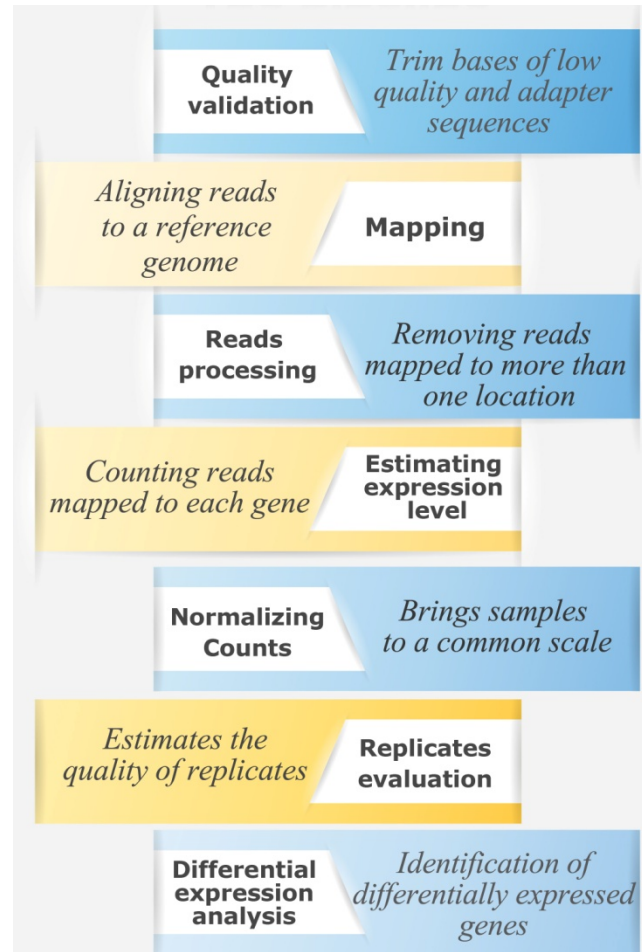
It is usually better to have a directional (stranded) sequencing: it costs slightly more, but it is able to discriminate between antisense RNAs.

# DATA ANALYSIS





## General RNA-Seq pipeline for Differential Expression



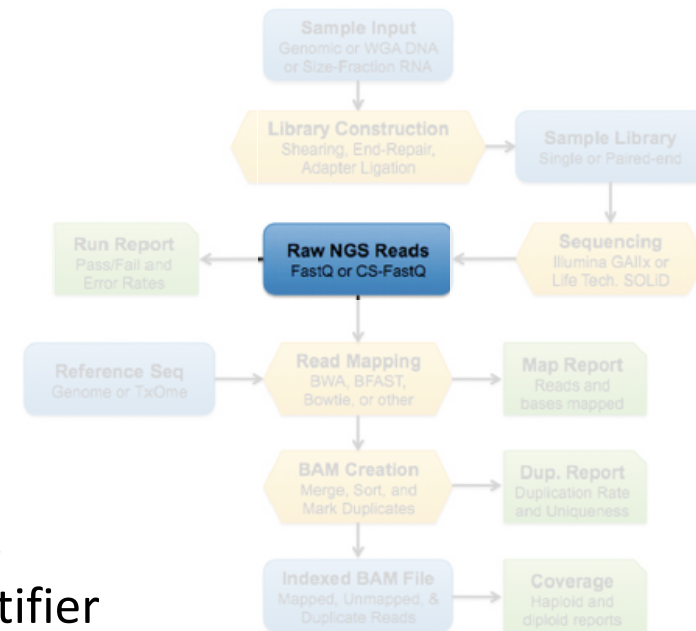
## Data format

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
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;;;9;7;;.7;393333
```



## PHRED quality score

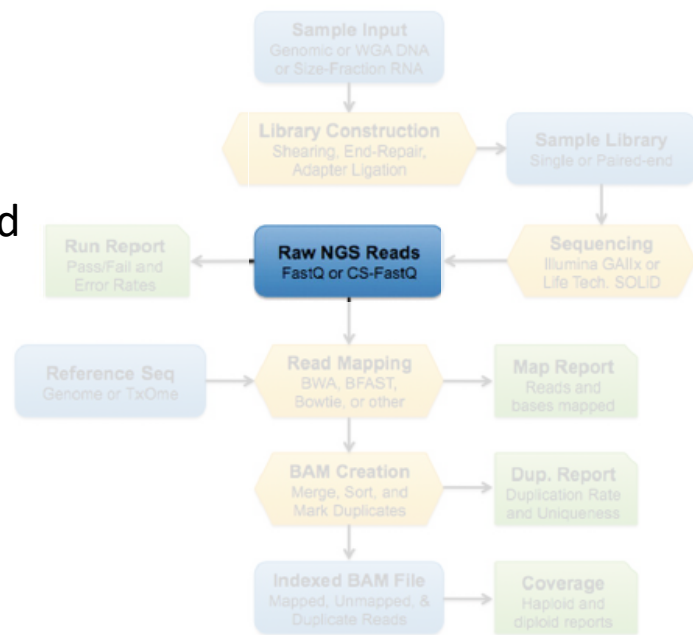
The quality score of a base, also known as a Phred or Q score, is an integer value representing the estimated probability of an error, i.e. that the base is incorrect.

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

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.

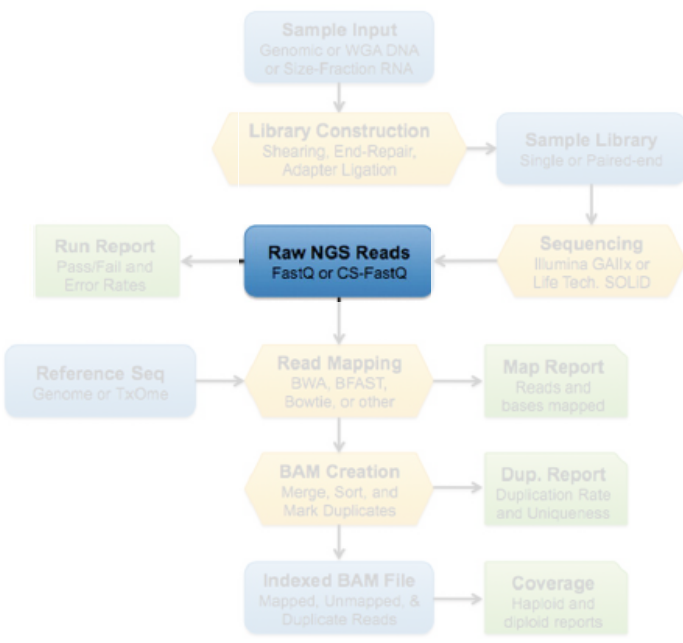
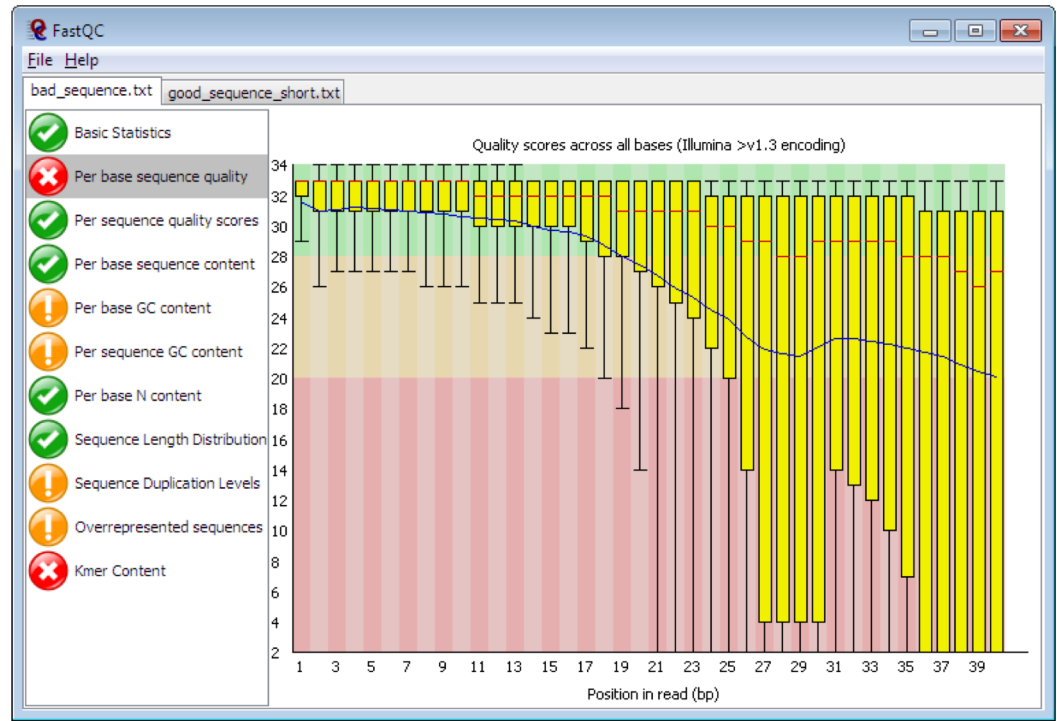
Table 1: Q-Scores and Error Probabilities

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)



## FastQC

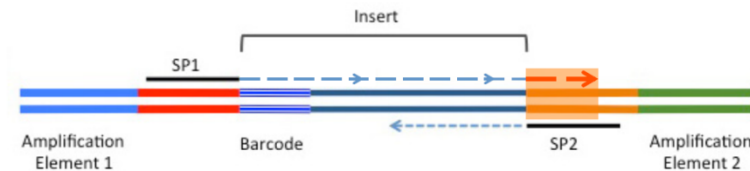
FastQC is a quality control tool for high throughput sequence data.



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

## 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.

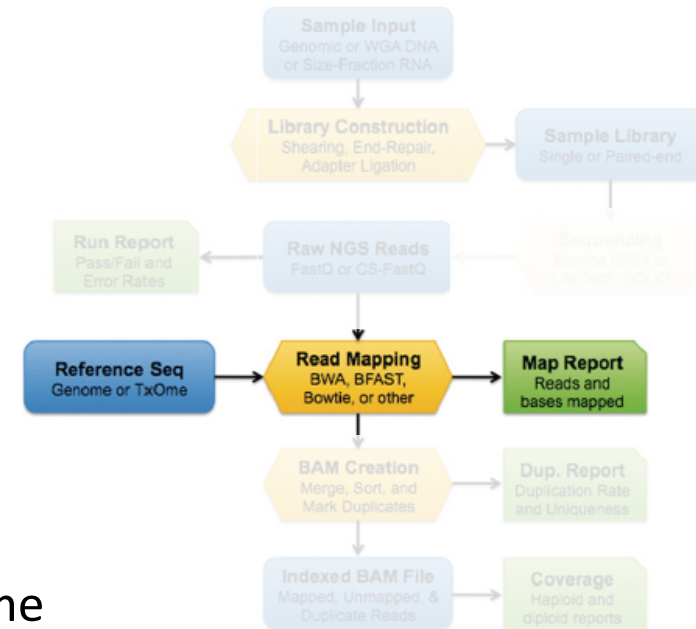
Sometimes the read terminates with ambiguous (N) bases which must be removed.

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

## Read alignment

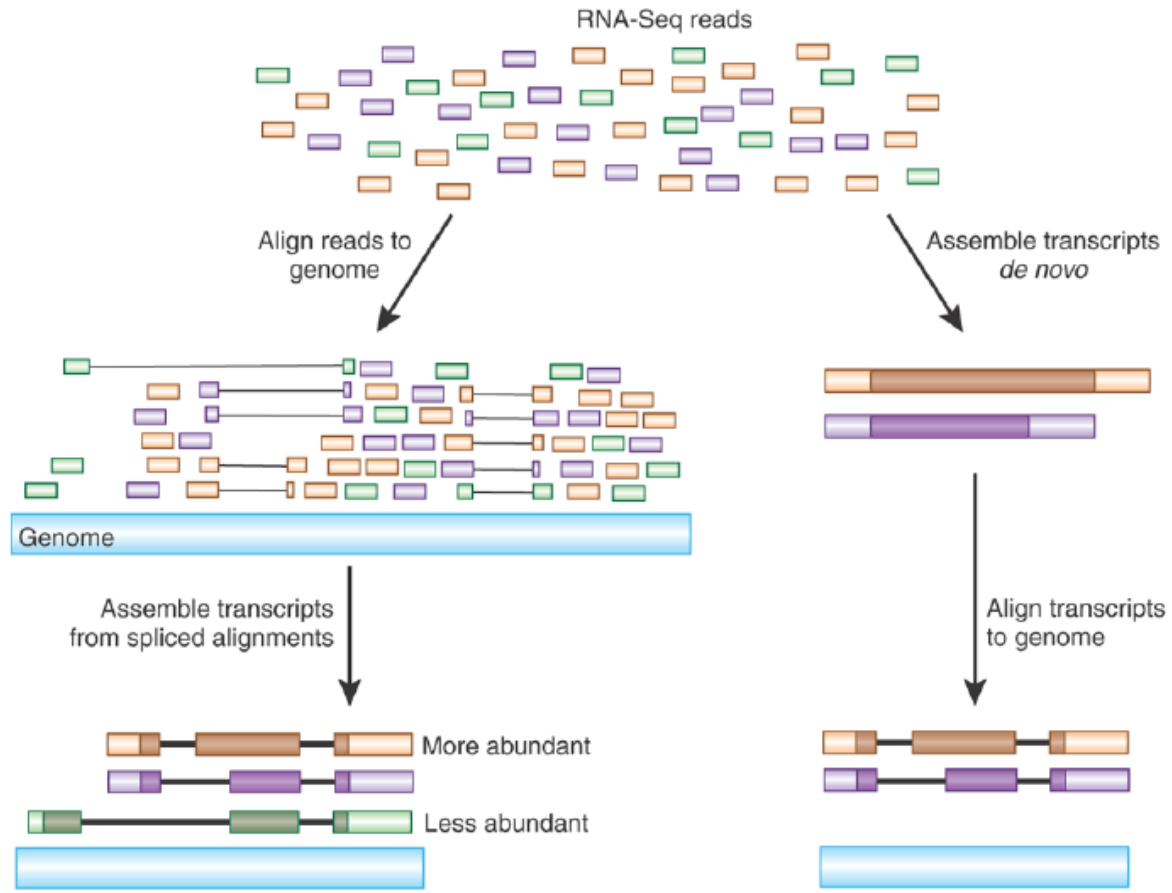
After pre-processing, we can align reads to a reference sequence.

- 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).



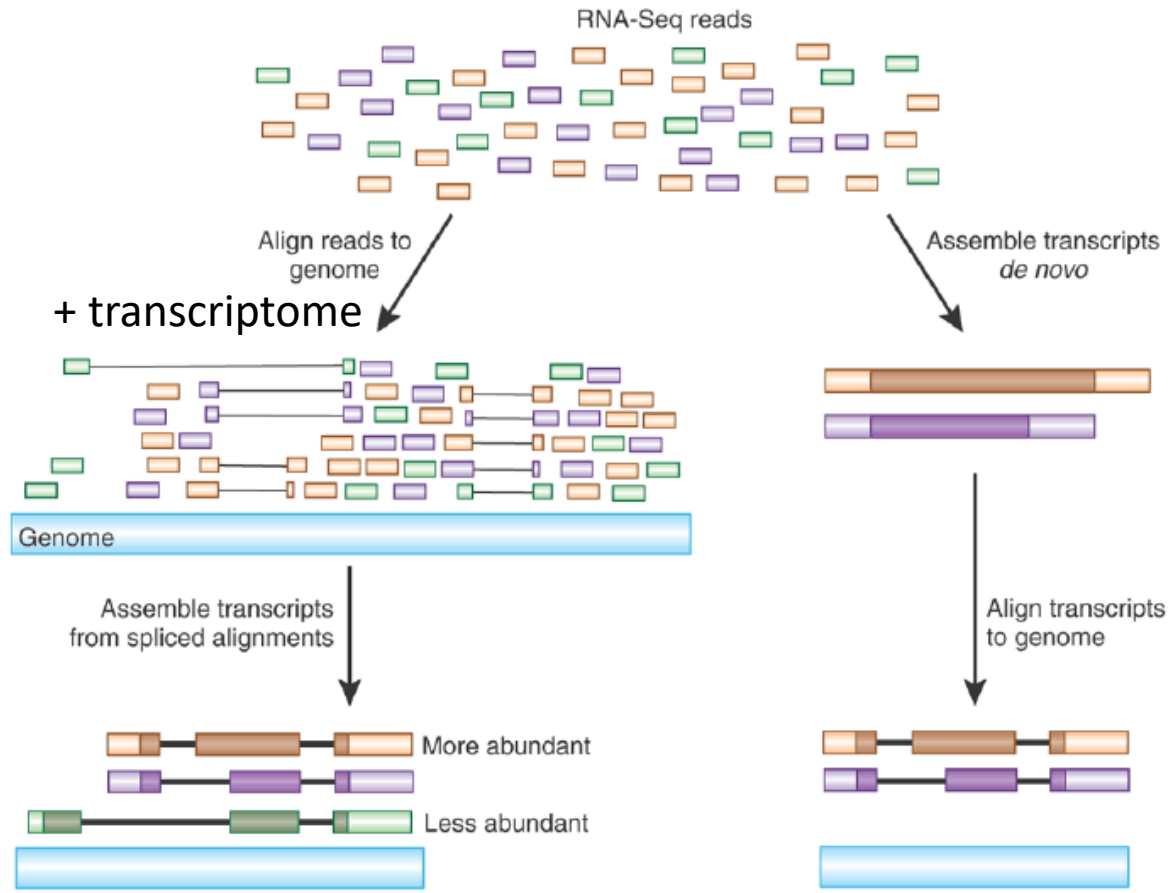
The accurate and fast alignment of millions of reads is not a simple task: many programs have been developed to address this issue.

## Main alignment strategies





## Main alignment strategies



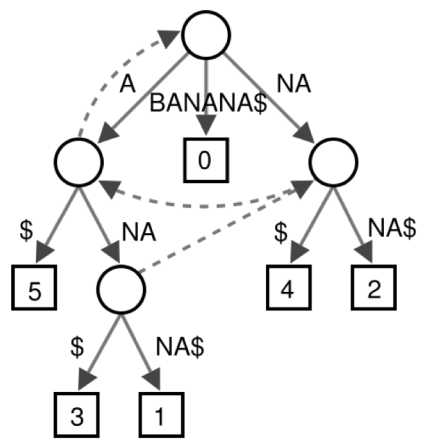
# DATA ANALYSIS: ALIGNMENT

## Alignment of NGS reads



Classical alignment techniques, such as dynamic programming, are not suitable for NGS data, due to the huge size of genomes and the high number of reads.

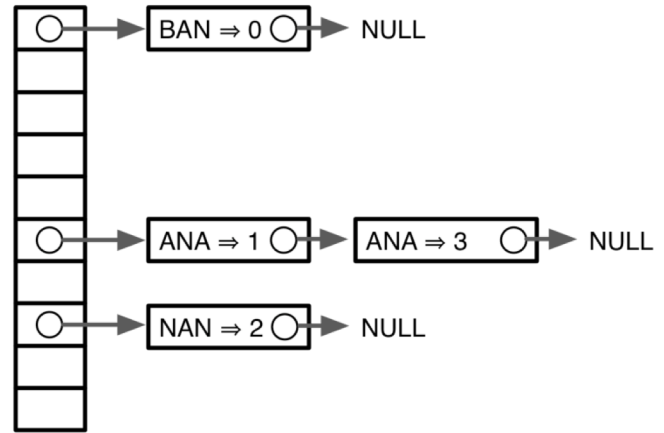
For this reason, short-read aligners are usually based on a preliminary indexing of the reference sequence. The performance of the aligner heavily depends on the way data are indexed.



Suffix tree

6	\$
5	A\$
3	ANA\$
1	ANANA\$
0	BANANA\$
4	NA\$
2	NANA\$

Suffix array



Seed hash tables

Many variants, incl. spaced seeds

## Main alignment tools

BWA, Soap2 and Bowtie are based on the Burrows-Wheeler transform, an indexing technique which allows to drastically reduce the 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

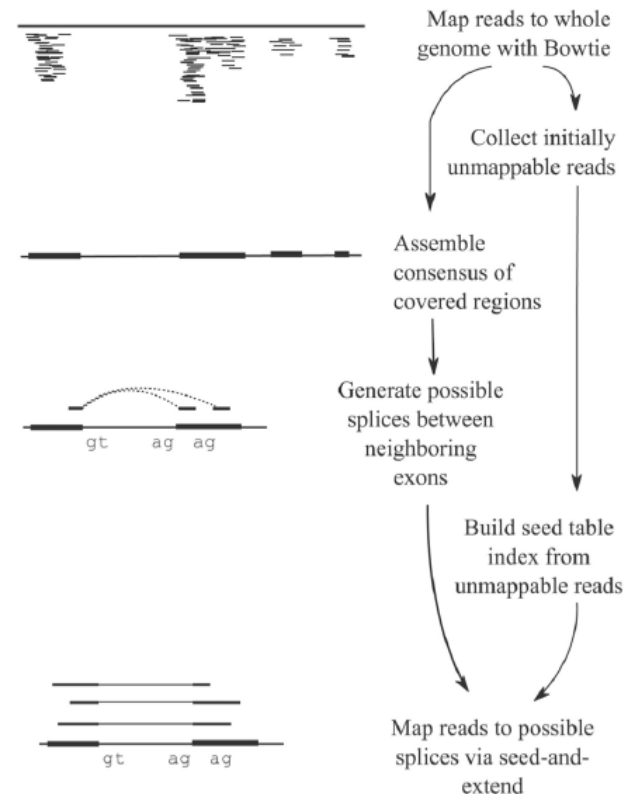
Aligner	Description	URL
Illumina platform		
ELAND	Vendor-provided aligner for Illumina data	<a href="http://www.illumina.com">http://www.illumina.com</a>
Bowtie	Ultrafast, memory-efficient short-read aligner for Illumina data	<a href="http://bowtie-bio.sourceforge.net">http://bowtie-bio.sourceforge.net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

## 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: TopHat, STAR, 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.

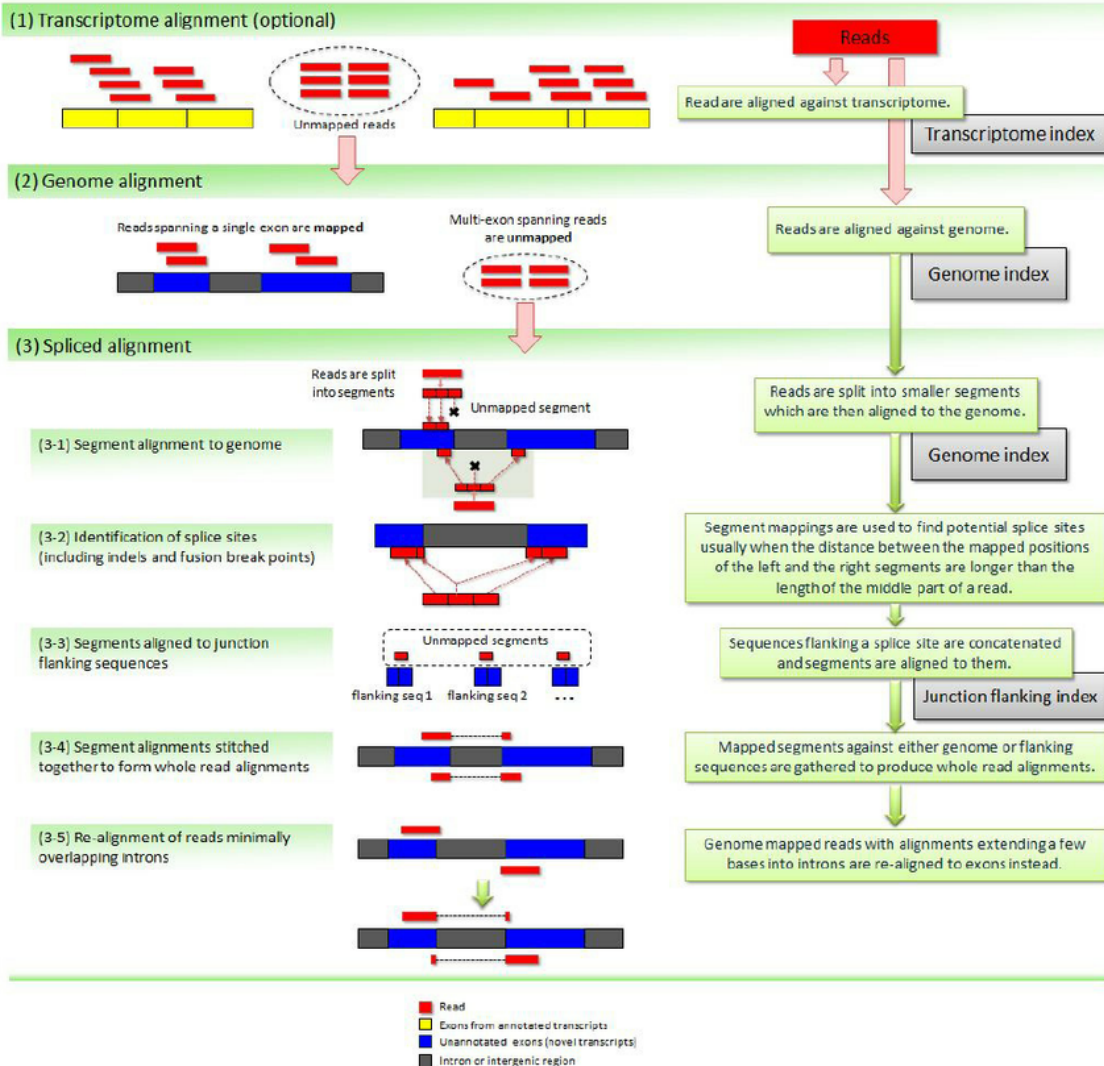
## From TopHat manual...

“TopHat generates its database of possible splice junctions from **two sources of evidence**. The first and strongest source of evidence for a splice junction is when **two segments from the same read** (for reads of at least 45bp) **are mapped at a certain distance on the same genomic sequence** or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found ab initio. **The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping.** Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million). This latter option will only report alignments across "GT-AG" introns”



# DATA ANALYSIS: ALIGNMENT

## TopHat



# DATA ANALYSIS: ALIGNMENT

## Main alignment programs

**Table 1** | Selected list of RNA-seq analysis programs

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

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

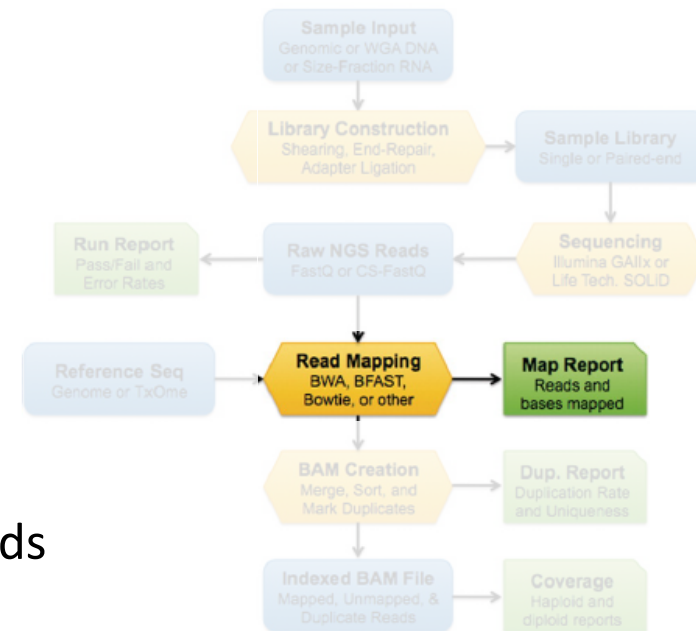
## Alignment output

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.

- **Samtools** is a suite of tools for the analysis and manipulation of SAM/BAM files (visualization, sorting, filtering, indexing etc.)



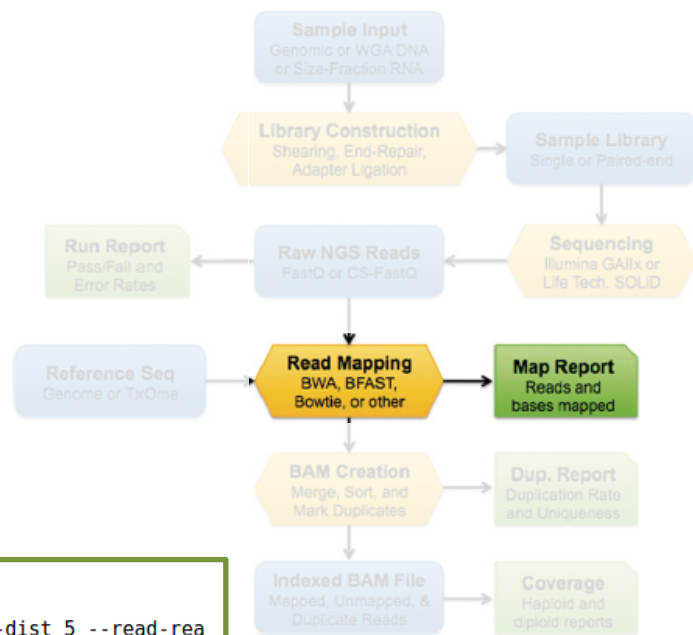


# DATA ANALYSIS: ALIGNMENT

## 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.



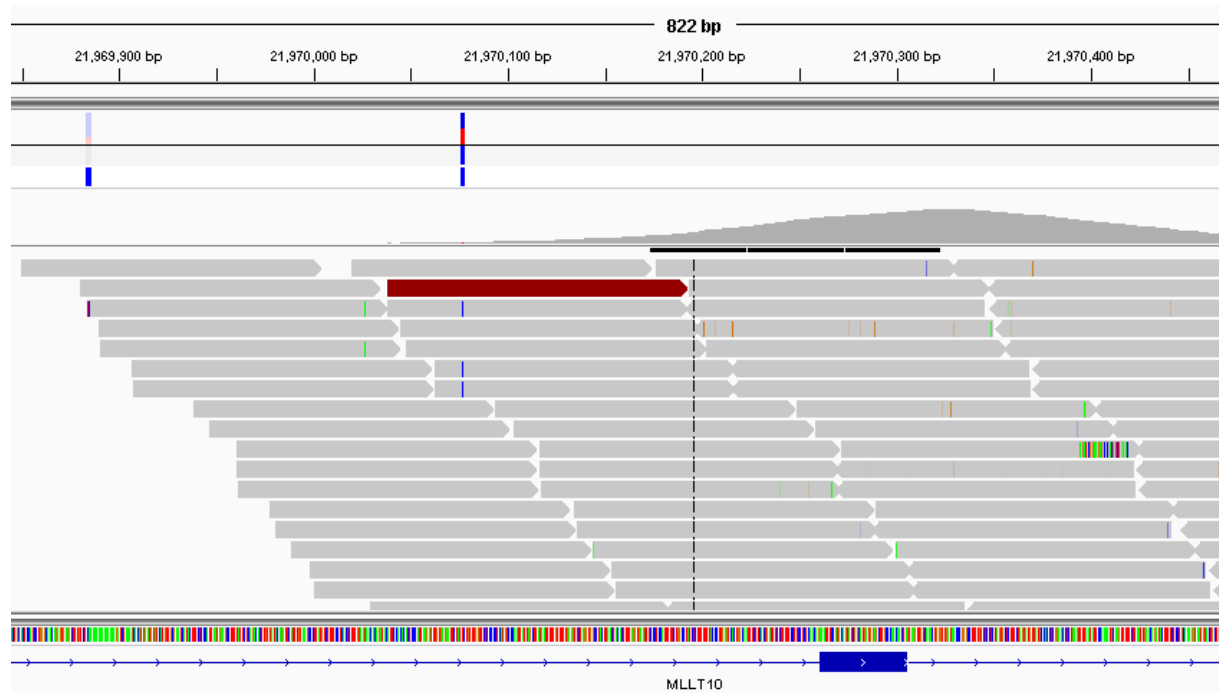
```

@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
CCGTGTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTAGTTGGCCTAGGAAATCCAGCTAGTCTGTCTCTCAGTCCCCCTCT
C BBDDCCDDCCDDDDCCDDDDDDCCDDBC?DDDDDDDDDDDDDDCCDDDDDDDDDDCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBHFFFDCC@
AS:i:-15 XM:i:3 X0: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
TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCTGGGGCAGTGGACCTTCCAGTGATTCCTGACATAAGGGCATGGACGA
G DCDDEDDDDDDDDDDDDDDDDDDDDCCDDDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJJJJJJJJJJJJJJJJJJJJJJHHHHHHFFFC
AS:i:-16 XM:i:3 X0: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
GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCACCTGGCCAGCAGCACCAACCAGAAAGAGGGAAGAAGACAGGAAAAACCA
C DDDDDDDDDCCDDDDDDDDDEEEEEFFFEFFEGHHHHFGDJJIHJJJJJJIIIGGFJJJIHIIIIJJJJJJJJJJJJJJJJJJJJJJHGHFAHGFHJHFGGHHFFDD@BB
AS:i:-11 XM:i:2 X0: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 GTGGCTCTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACCTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
  
```

## BAM file visualization

### IGV

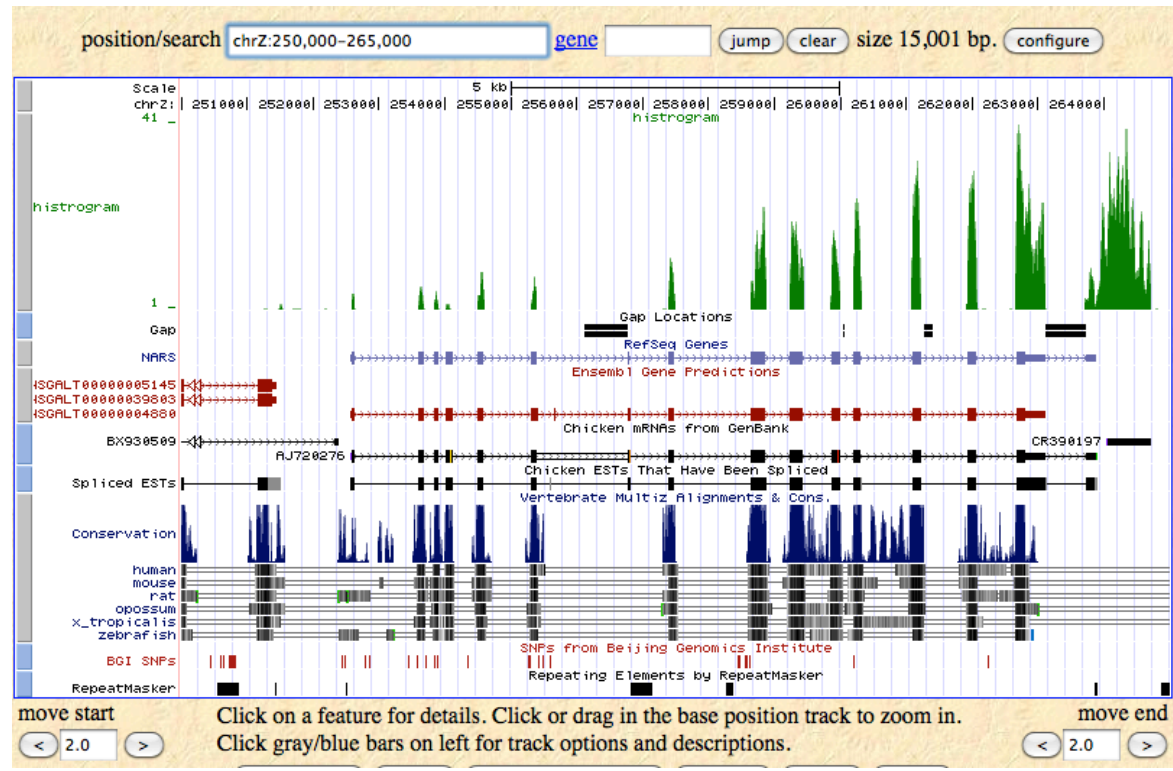
IGV is a standalone program which allows a highly interactive visualization of BAM files (and other genomic annotation formats).



## BAM file visualization

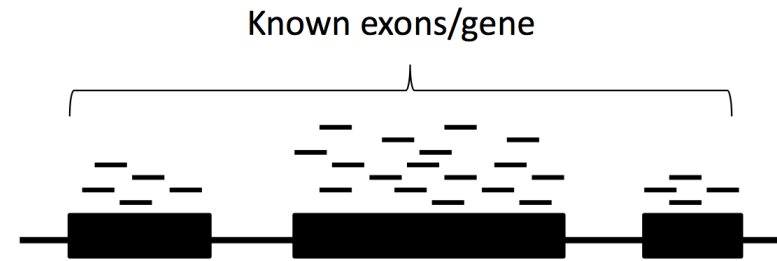
### Genome Browser (UCSC)

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



## 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?

- **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.

- **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.

- **RPKM** and **FPKM** are two normalized measures of gene expression.

# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

## Measures of gene expression: RPKM

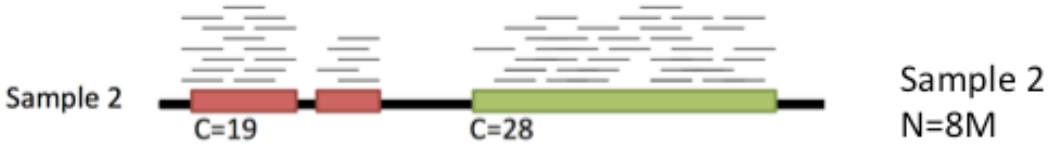
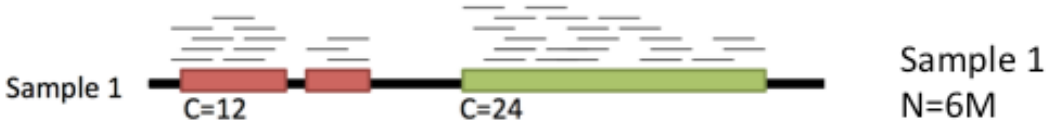
- RPKM stands for “Reads per Kilobase of exon per Million mapped reads”

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

- 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)

Gene A 600 bases    Gene B 1100 bases

$RPKM = 12 / (0.6 * 6) = 3.33$      $RPKM = 24 / (1.1 * 6) = 3.64$



$RPKM = 19 / (0.6 * 8) = 3.96$      $RPKM = 28 / (1.1 * 8) = 1.94$

## Measures of gene expression: FPKM

- FPKM stands for “Fragments per Kilobase of exon per Million mapped fragments”

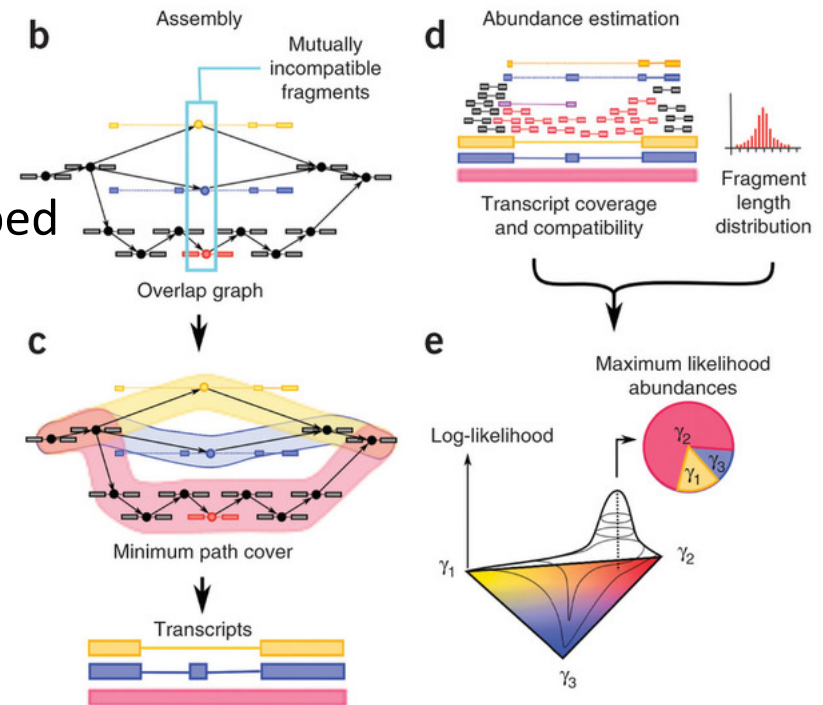
-The unit used for quantification is no longer the single read, but the fragment. In single-end sequencing, each read represents a fragment, so FPKM = RPKM. In paired-end sequencing, each fragment is represented by a read pair: this way, each read pair is not counted twice.



## Cufflinks

Cufflinks is a software which is able to:

- Assemble transcripts from reads mapped to the genome,
- estimate the abundance of these transcripts (FPKM),
- test for differential expression (DE).

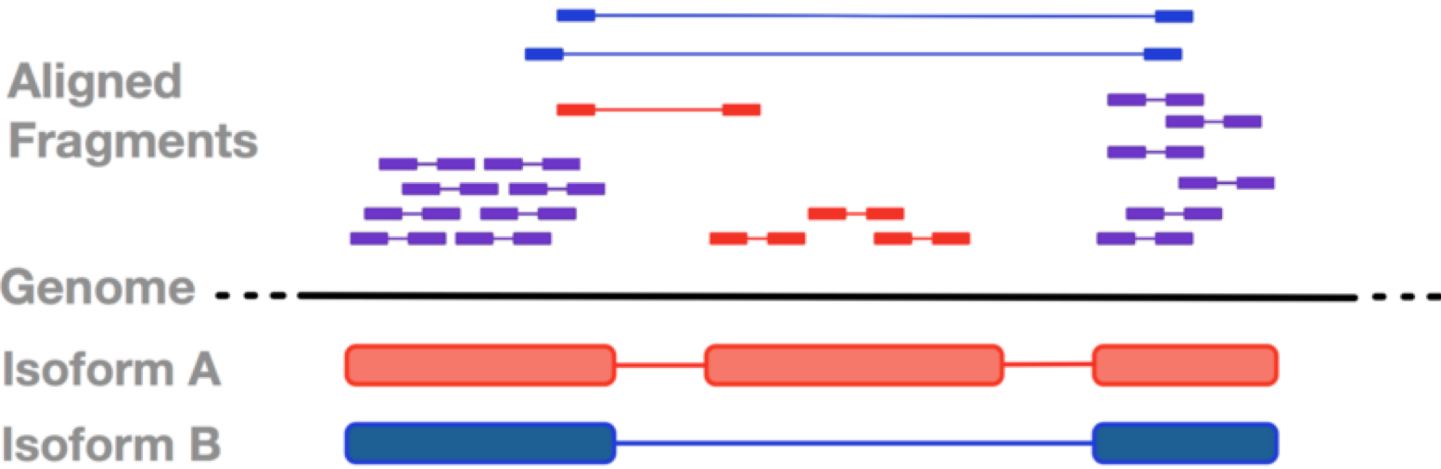


# DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

## Cufflinks: quantification



Expression values are expressed as FPKM  
Distinct transcripts belonging to the same gene may share some exons. How does Cufflinks assign reads to the correct isoforms? It uses non-ambiguously mapping reads to estimate the probability of each ambiguous reads coming from a certain isoform.



The expression of a gene is equal to the sum of the FPKMs of its isoforms.



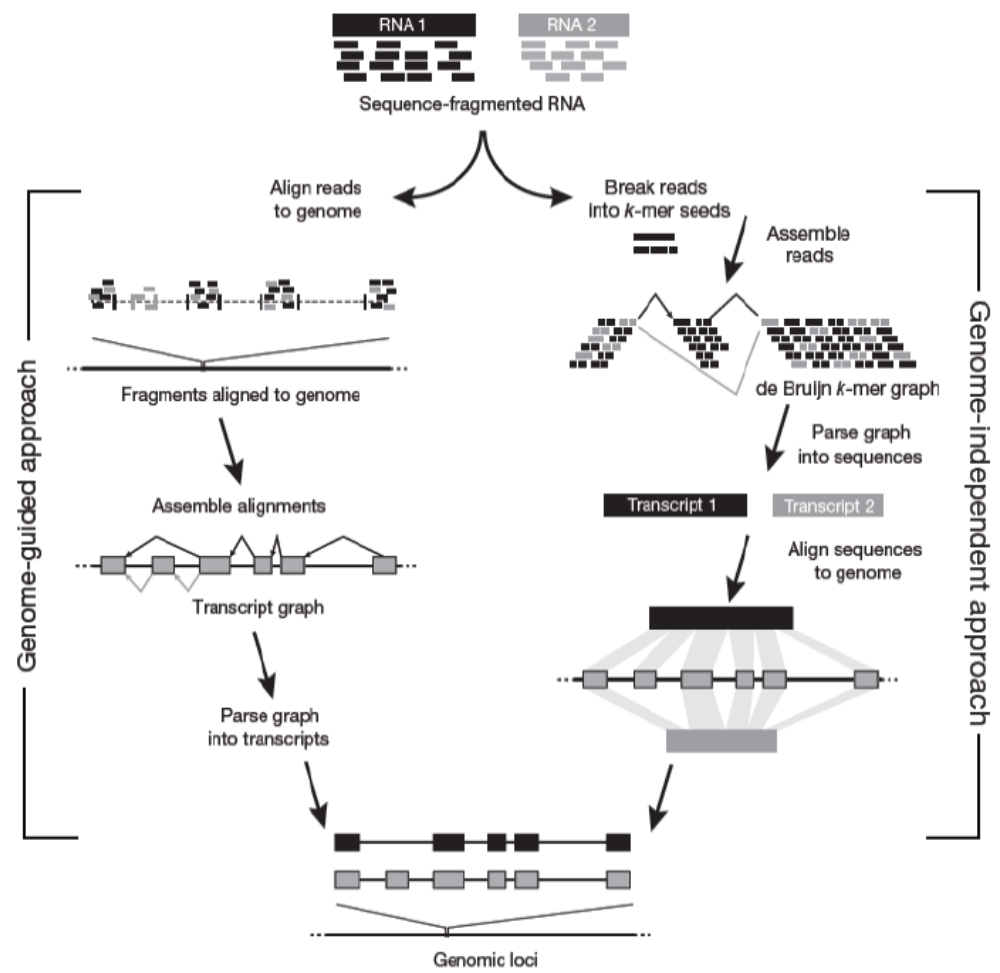
## Tools for de novo discovery of transcripts

- **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 <sup>28</sup> Cufflinks <sup>29</sup>	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 <sup>61</sup> TransABySS <sup>56</sup> Trinity	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads

## Tools for de novo discovery of transcripts



## 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).

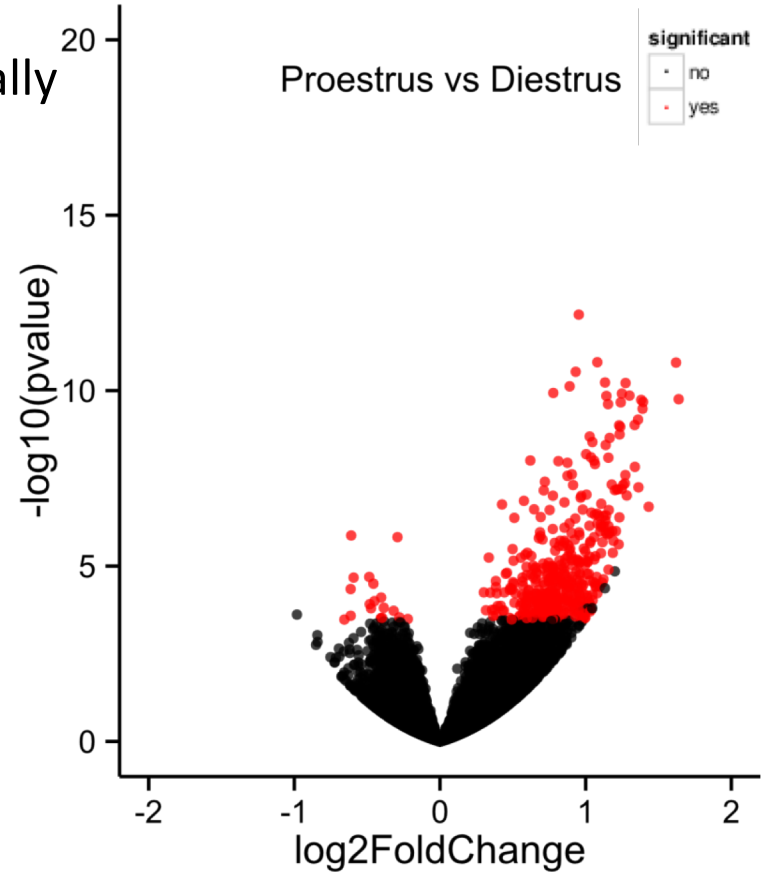
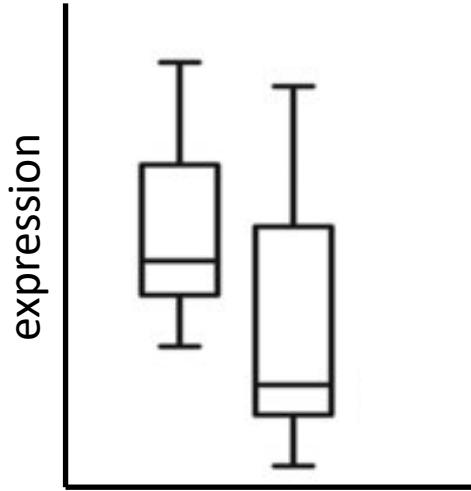
For each analyzed gene, the result will be:

- **Fold Change (FC)**: the ratio of the average expression of gene in condition A to the average expression in condition B.  $\log_2$  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

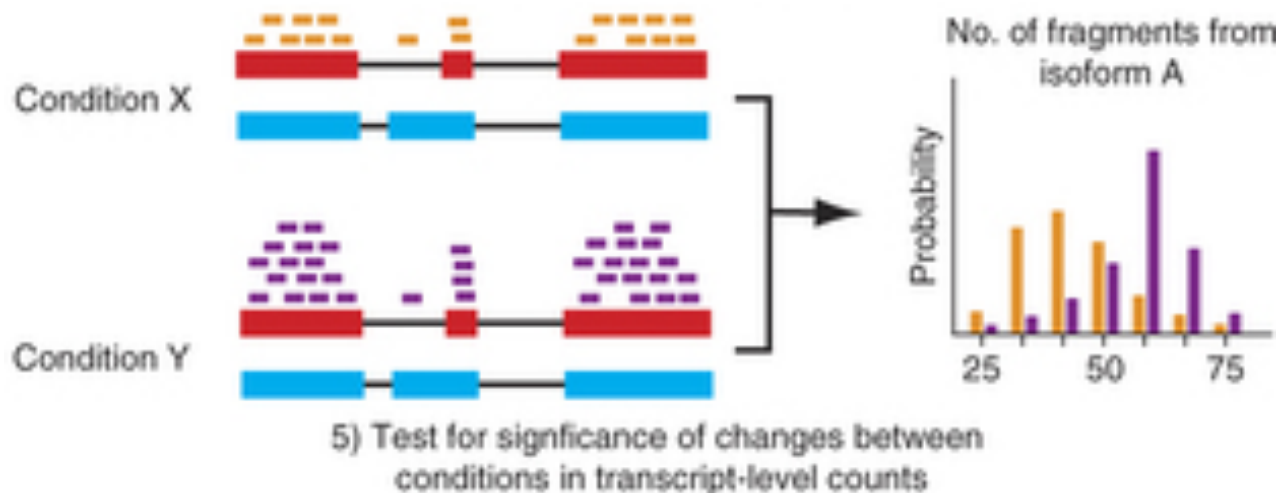
## FC vs p-value

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



## Cuffdiff: differential expression test

Cuffdiff is able to perform a differential expression test, both at isoform and gene level.



# 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	AveLogP	t	P-value	adj.P.Val	B
6490	RYR2	regulator of G protein signaling 2	4.3729791	0.0046694	18.2510193	4.81E-15	2.7E-14	23.02101
8300	ACAD2	acyl-CoA oxidase 2, branched-chain	3.8452777	0.0097103	26.2512512	5.58E-18	2.47E-14	31.0537676
6100	CHRNA2	alpha 2 nicotinic acetylcholine receptor (nonbrain receptor type II-like)	4.1212488	0.0127761	21.2610437	1.68E-15	2.07E-14	28.88148
5164	CHRNA3	cholinergic acetylcholine receptor, alpha 3	5.1616104	0.0044155	24.4511259	7.92E-17	2.49E-14	29.9901828
6411	MYL7	myosin heavy chain 7	3.3504422	0.0126612	24.2619224	3.78E-17	6.08E-14	24.5436305
230	CD44	cell surface receptor type 1, hyaluronate	4.4974885	0.0082617	26.8627544	5.54E-17	1.11E-13	28.8906111
6006	SPPI	serpin peptidase inhibitor 1	3.0284272	0.0091429	23.4210429	6.33E-17	1.32E-13	28.7359241
6005	SPPI2	serpin peptidase inhibitor 2	3.1892228	0.0091429	23.4210429	7.49E-17	1.32E-13	28.5606663
614	CANP2	caprin 2, 60 kDa subunit	5.7384402	0.0175087	22.6540605	1.34E-16	1.29E-13	28.0537176
8871	FNX2	fibronectin type 2	3.1802106	0.0031139	22.4506616	1.49E-14	1.89E-13	27.8499164
6285	UCDML	urocortin-like domain-containing protein 3	2.1767649	0.0091811	22.2613889	1.73E-16	1.49E-13	27.8095284
29292	STRBP	strapped domain-binding protein 3	4.4248242	0.0212269	19.4461007	2.40E-15	1.42E-13	27.221474
1349	OPN1	osteopontin, alpha 1, polypeptide 1	3.4252768	0.0047819	21.7719509	5.38E-16	6.97E-13	26.7399164
3349	FOXP1	forkhead domain-containing fork transcription regulator 1	2.8417882	0.0018713	20.7977744	7.02E-14	6.79E-13	26.3364343
8070	TRIS	trans alpha-tropomyosin kinase chain family, member 3	1.1319104	0.0113827	20.4427384	1.13E-15	8.90E-13	26.1007949
4142	SRFBP1	serpin family B member 1	1.0122259	0.0082617	20.3181261	1.40E-15	1.06E-12	25.7302099
14488	CD137A3	CD137 tumor necrosis factor related protein 3	3.1044813	0.0160466	19.1471913	1.50E-15	1.06E-12	25.7164871
2501	TRPC1	TRPC1 transient receptor channel	3.4827701	0.0016076	20.0050504	1.60E-15	1.05E-12	25.4214614
7058	TRSD1	transmembrane domain-containing protein 2	2.1106143	0.0030895	19.8271911	2.03E-15	1.25E-12	25.4133605
52851	TRSD2	transmembrane domain-containing protein 2	2.4028141	0.0112189	19.8001793	2.13E-15	1.25E-12	25.3119464
64752	SAHHA2	SAH domain-containing ubiquitin protein ligase 2	2.0584214	0.0212652	18.7650727	2.24E-15	1.25E-12	25.3251232
1586	NRAMP1	natural resistance-associated macrophage protein 1	3.1343211	0.0076763	19.7433760	1.38E-15	1.25E-12	25.3077114
8848	TGCT2B1	TGCT domain family, member 1	-1.1343114	0.0116823	19.3039081	2.92E-15	1.48E-12	25.0784401
122	ACTA13B	actin filament-capping protein 1 family, member A3	2.7158161	0.0143288	19.2416767	3.01E-15	1.48E-12	25.0715754
4238	MYL2C	myosin light chain 2C	1.4237212	0.0162787	18.8251951	3.09E-15	1.48E-12	25.0222806
5810	MYL3C	myosin light chain 3C	1.4784241	0.0162787	18.8251951	4.05E-15	1.48E-12	25.0175154
4885	AP2A2	adaptor protein 2 complex, alpha 2 subunit	4.4571117	0.0110897	18.5054465	5.57E-15	1.48E-12	24.4344081
1122	NRXN1	neuronal cell adhesion molecule 1	2.4216141	0.0069582	18.2877169	5.76E-15	1.48E-12	24.2652241
6437	PRPLD	proteoglycan core protein 2-like domain-containing protein 1	1.8956411	0.0080817	18.5511294	7.52E-15	1.32E-12	24.1603222
3047	PRNAB	glycoprotein (transmembrane) emb	3.0701217	0.0026251	18.7116048	6.00E-15	1.28E-12	24.0839661
5054	RCOR1	retinoblastoma corepressor 1	2.7181405	0.0113472	18.3547958	1.07E-14	1.05E-12	24.0411327
8512	CAF1	growth arrest specific 7	2.1113293	0.0064079	18.2412491	1.16E-14	1.43E-12	23.7191245
2680	TSPAN	transmembrane protein 1	1.8826095	0.0071246	18.2212024	1.38E-14	1.03E-12	23.6800648
5700	MTUS1	microtubule-associated tumor suppressor 1	2.102644	0.0092085	18.1116375	1.34E-14	1.48E-12	23.5731135
5718	SNAP25	small cell neuroendocrine protein 2	1.8814651	0.0130081	18.0141051	1.35E-14	1.48E-12	23.5641664
1214	MYL2	myosin light chain 2, beta and a half-like domain 2	2.034208	0.0093012	18.0114436	1.37E-14	1.48E-12	23.5480463
12132	ACOT2	acyl-CoA oxidase 2	2.128488	0.0072004	17.8189066	1.76E-14	1.28E-12	23.5405664
5995	HSST1A1	heparan sulfate (glycosaminoglycan) 3-O-sulfotransferase 1A1	2.8202466	0.0034082	17.7211015	2.09E-14	6.50E-12	23.1335954
6184	CDNA1	cdna1	1.1814791	0.0417448	17.7444168	1.15E-14	6.50E-12	23.1305643
8070	MYL7	myosin heavy chain 7	2.1070709	0.0094408	17.6513817	2.27E-14	6.50E-12	23.0537176
4728	ADAM19	ADAM metallopeptidase domain 19	1.2867878	0.0025258	17.6431886	2.82E-14	6.50E-12	23.0695825
3210	EDH8	endothelin receptor type 8	1.9751918	0.0142813	17.4271925	2.04E-14	6.50E-12	22.7922215
302	ASB1	absent in melanoma 1	1.1047061	0.0074825	17.0818883	4.40E-14	1.21E-12	23.005284
2012	EMPS1	epithelial membrane protein 1	3.1096417	0.0080729	17.0671982	4.42E-14	1.21E-12	23.0952885
8244	SDN1	SDN1	2.4296461	0.0219663	16.9252722	5.36E-14	1.21E-12	23.0258971
8818	WDR5	WD repeat domain 5	1.6282818	0.0126617	16.9271913	5.32E-14	1.36E-12	23.2114848
8242	SDN2	SDN2	1.1524465	0.0126617	16.9271913	5.32E-14	1.36E-12	23.2067965
5488	SDN3	SDN3	1.0972345	0.0192723	16.7620474	6.79E-14	1.26E-12	23.1711738
4907	MYL4	myosin light chain 4, beta and a half-like domain 2	2.1822379	0.0046572	16.6650713	7.22E-14	1.26E-12	23.1967926
4908	MYL5	myosin light chain 5, beta and a half-like domain 2	3.1997568	0.0064715	16.6710726	7.22E-14	1.26E-12	23.1967926
38473	MYL6	myosin light chain 6, beta and a half-like domain 2	2.1822379	0.0046572	16.6650713	7.22E-14	1.26E-12	23.1967926
8442	MYO10A1	myosin class I0 heavy chain non-protein like 1	-1.1165101	0.0177015	16.4481781	9.11E-14	1.26E-12	21.4771135
1121	MYO10B1	myosin class I0 heavy chain non-protein like 1	0.8374841	0.0177015	16.4481781	9.11E-14	1.26E-12	21.4771135
6440	MYO11A1	myosin class I1b heavy chain alpha-2A-subunit-like 1	1.3825243	0.0018519	16.4210549	9.48E-14	1.26E-12	21.6024444
6295	MYO11B1	myosin class I1b heavy chain alpha-2B-subunit-like 1	2.1938841	0.0018519	16.4210549	9.48E-14	1.26E-12	21.6024444
5811	PKOX1	pyruvate carboxylase oxidase 1	0.9112619	0.0119104	16.4044911	1.00E-13	2.35E-12	21.5608901
915	ME	myosin essential light chain	1.0070041	0.0029687	16.2174849	1.21E-13	2.35E-12	21.9397219
3151	MYR2	myosin regulator 2 (non-muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
5895	MYR1	myosin regulator 1 (non-muscle/embryonic)	1.8262446	0.0075228	15.8556055	1.26E-13	1.42E-12	20.9956666
25118	MYL9	myosin light chain 9, embryonic	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
5486	MYR2B1	myosin regulator 2 (muscle/embryonic)	0.8099893	0.0062417	15.8313099	2.00E-13	1.42E-12	20.8721203
5486	MYR2B2	myosin regulator 2 (muscle/embryonic)	0.8099893	0.0062417	15.8313099	2.00E-13	1.42E-12	20.8721203
2902	MYR1B1	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B2	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B3	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B4	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B5	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B6	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B7	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B8	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B9	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B10	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B11	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B12	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B13	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B14	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B15	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B16	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B17	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B18	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B19	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B20	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B21	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B22	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B23	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B24	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B25	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B26	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B27	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.21E-13	1.42E-12	20.9892044
2902	MYR1B28	myosin regulator 1 (muscle/embryonic)	0.810214	0.0241481	15.8450272	1.		

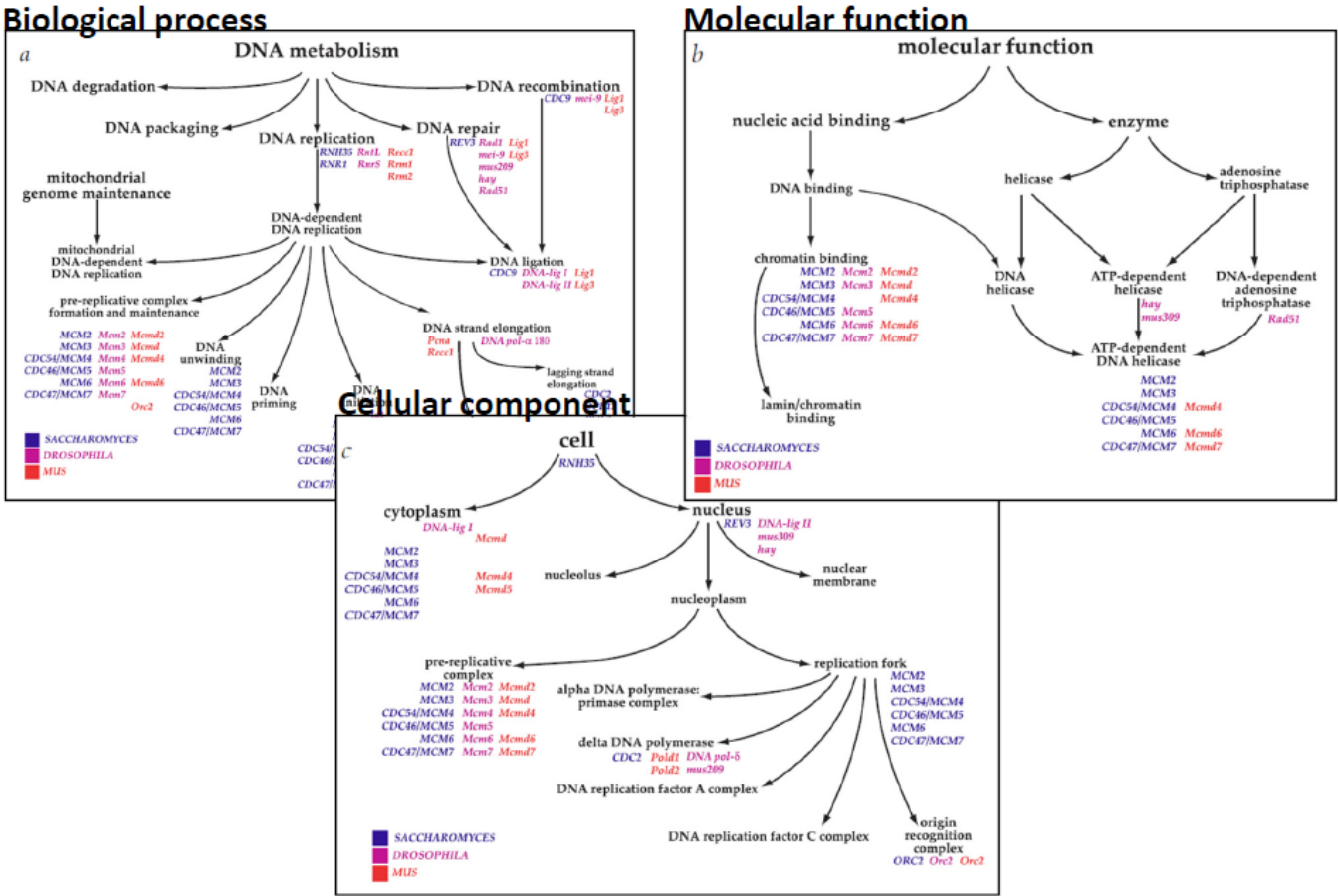
## Extracting biological meaning from DE gene lists

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

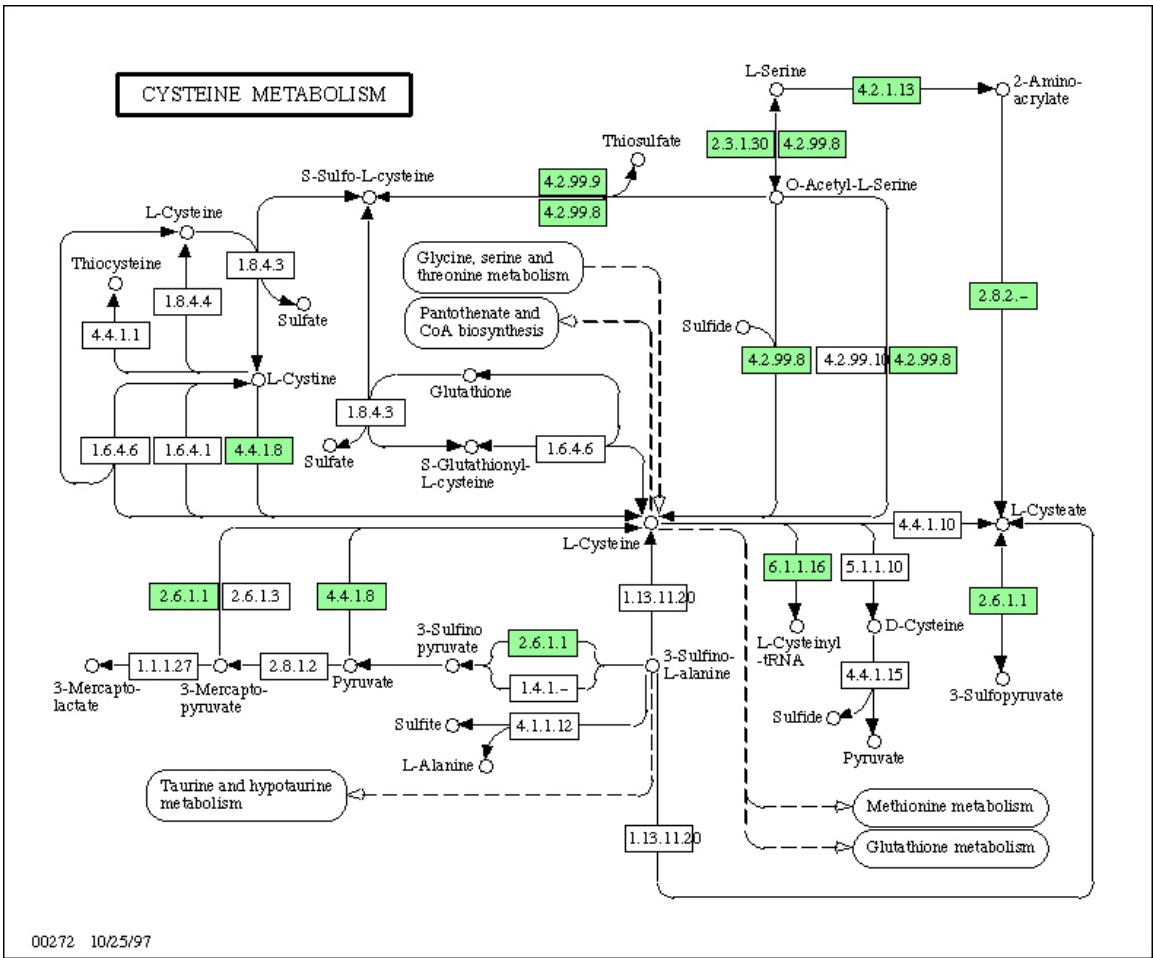
## Example of functional categories: Gene Ontology.





# DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

## Example of functional categories: Kegg pathway.

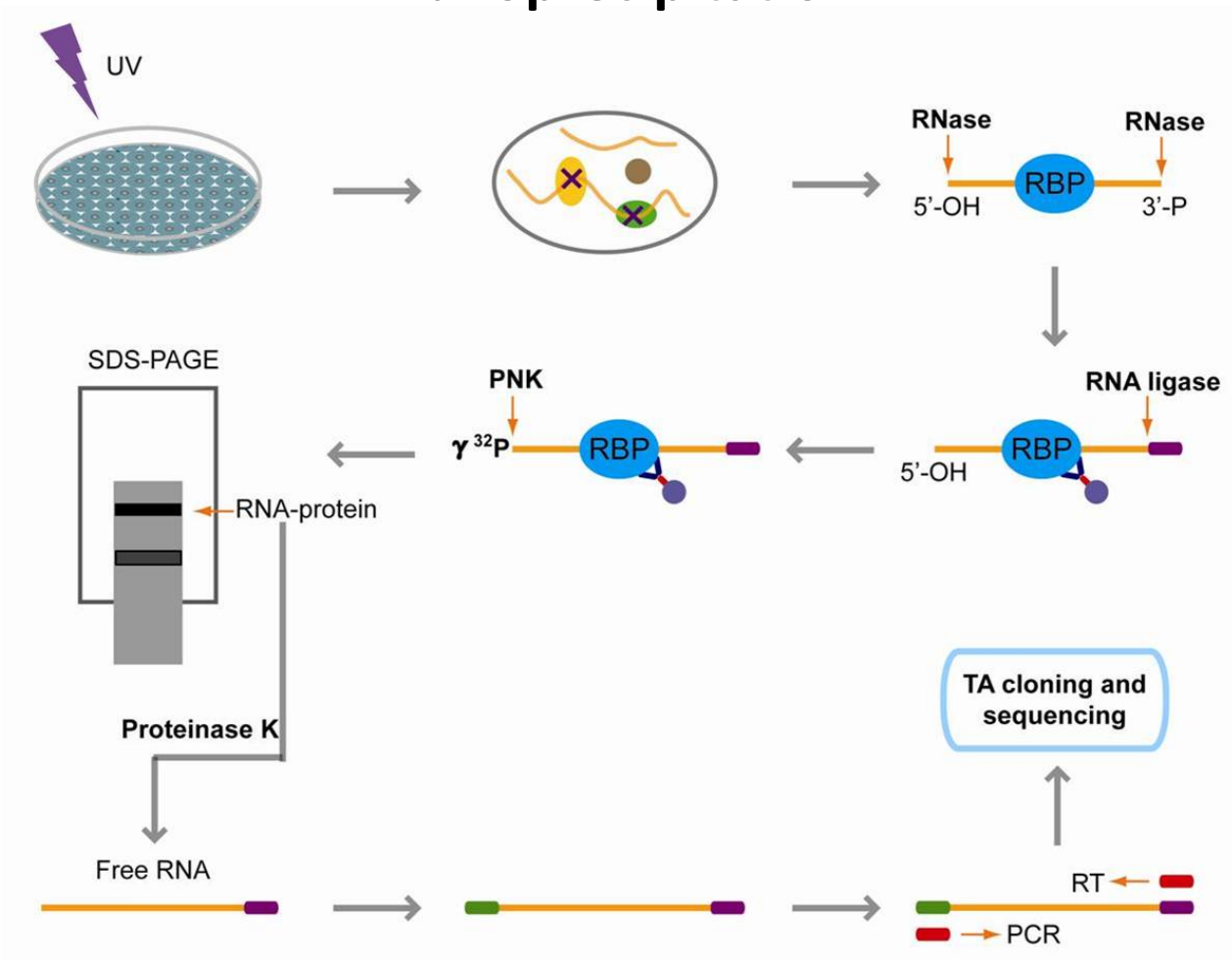


# DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

## Example of online functional annotation tool: DAVID.

The screenshot shows the DAVID Bioinformatics Resources 6.8 website. The header includes the DAVID logo and the text "DAVID Bioinformatics Resources 6.8 National Institute of Allergy and Infectious Diseases (NIAID), NIH". A navigation bar contains links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About Us. A red banner welcomes users to DAVID 6.8 with updated knowledgebase and provides a link to DAVID 6.7. A "Shortcut to DAVID Tools" sidebar lists categories like Functional Annotation, Gene Functional Classification, Gene ID Conversion, and Gene Name Batch Viewer. The main content area features a "Welcome to DAVID 6.8" section with a search bar and a list of key features, including identifying enriched biological themes, discovering enriched functional-related gene groups, and visualizing genes on BioCarta & KEGG pathway maps. A "What's Important in DAVID?" section lists updates such as new requirements to cite DAVID and support for Affy Exon and Gene arrays. A "Statistics of DAVID" section includes a bar chart titled "DAVID Bioinformatic Resources Citations" showing an increasing trend from 2004 to 2015, and a list of statistics: > 21,000 Citations, Average Daily Usage: ~2,600 gene lists/sublists from ~800 unique researchers, and Average Annual Usage: ~1,000,000 gene lists/sublists from >5,000 research institutes world-wide.

## high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation



# CLIP-Seq: DATA ANALYSIS

