EXPERIMENTAL DESIGN

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.

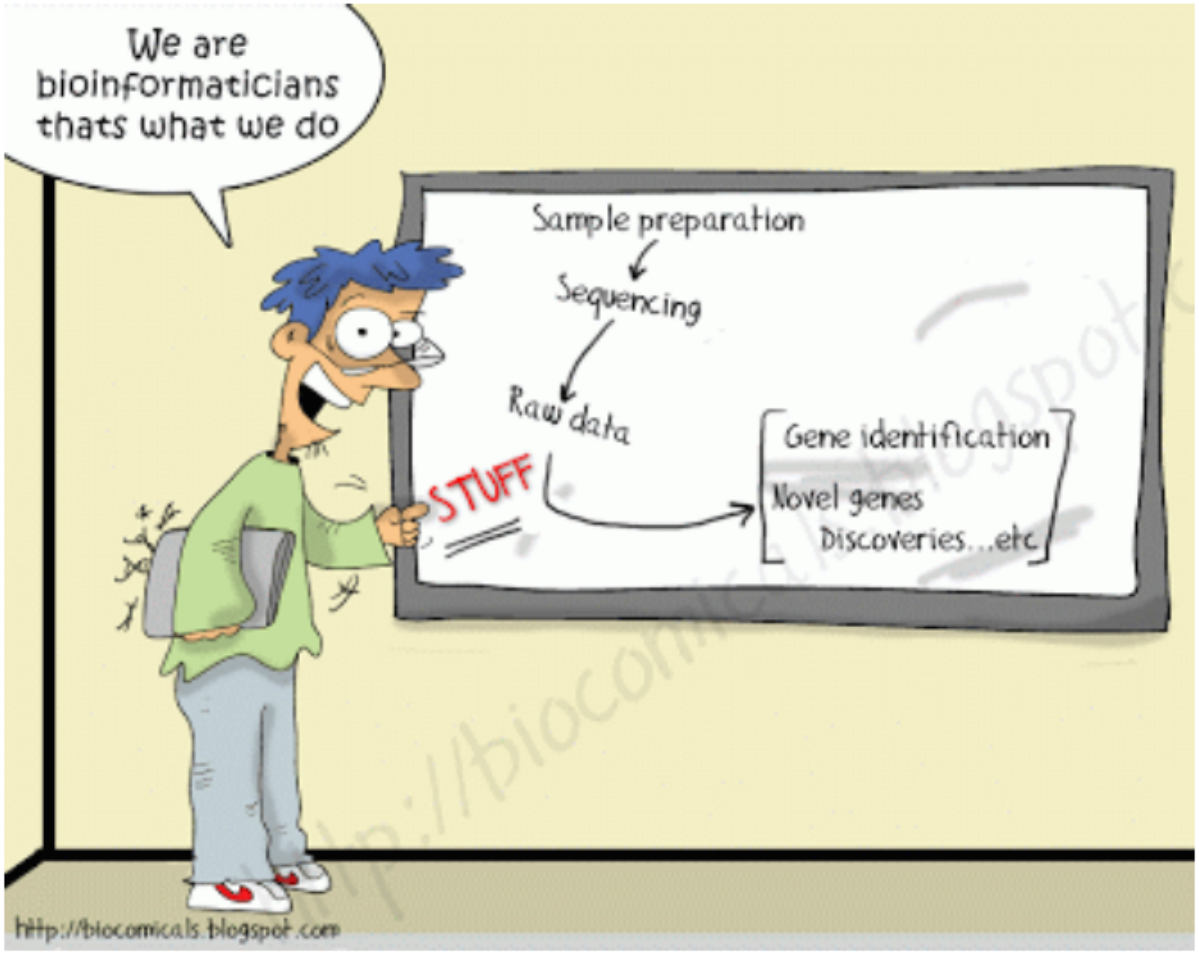
Defining the technical details

- **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).

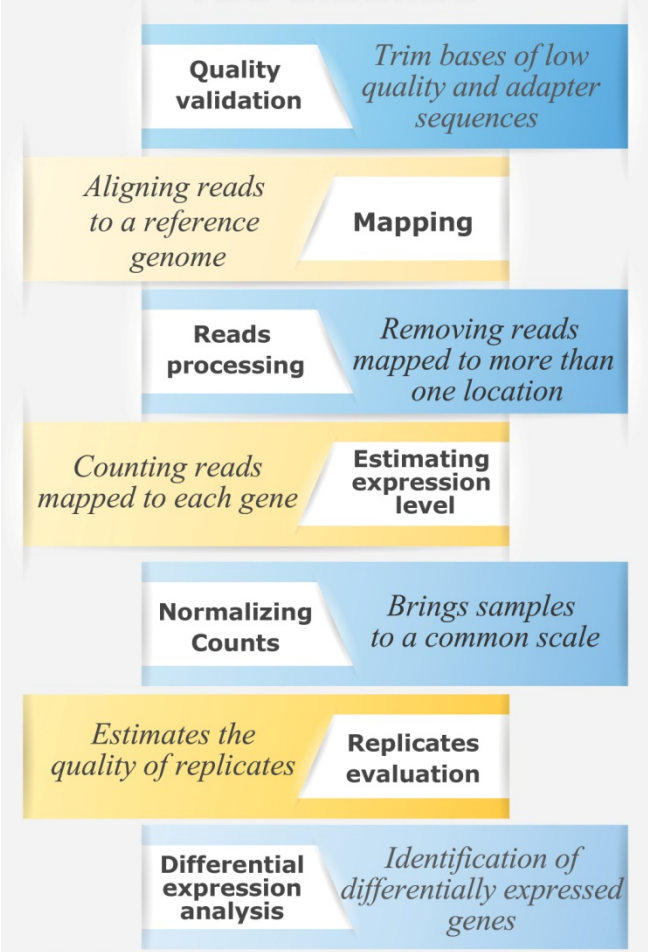
- **Length and pairing of reads**
  Theoretically speaking, read length should be > 20 bp (they usually are longer than 35 bp). PE reads are usually better (except for small RNA-Seq and Ribo-Seq), but they are more expensive.

- **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



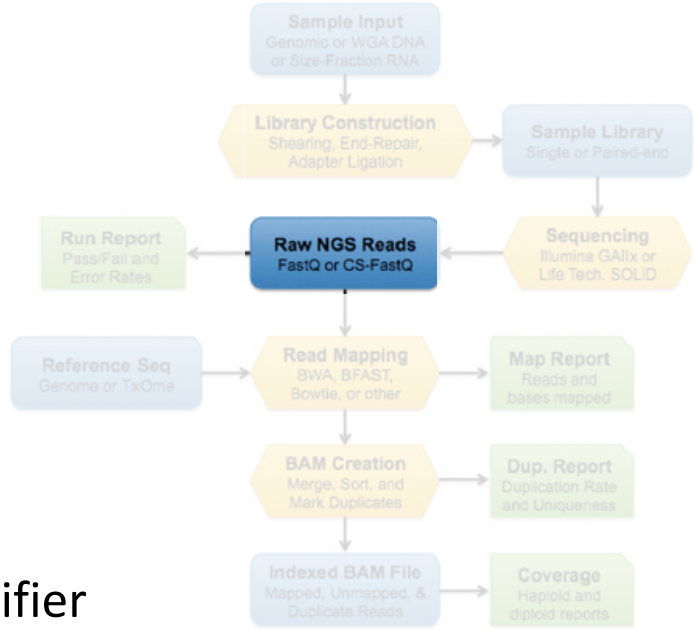
Data format

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
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<BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB<<<<<<<<<<<<<<<<<<
@SEQILMN03:128:HA5CBADXX:1:1101:1193:2104 2:N:0:GTCGTA
CTATCTTCGTAACCCAAAATAAATAAACTAACTCTATTTCTTGTGTTAGGCAGGGTATTCC
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFB707BFFIIIIIIIIIIIIFFFFFFFFFF<BBFFF
@SEQILMN03:128:HA5CBADXX:1:1101:1227:2106 2:N:0:GTCGTA
GGGGAGCATGACGGCCCACATCGGCGAAAACCCACTCTGGTGGGGTGAACCGGTATCCAN
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIFFFFFFBBFFFBBFBFFBBFF<BBFF0<BBFFBFBFFFFFB
```

PHRED quality score

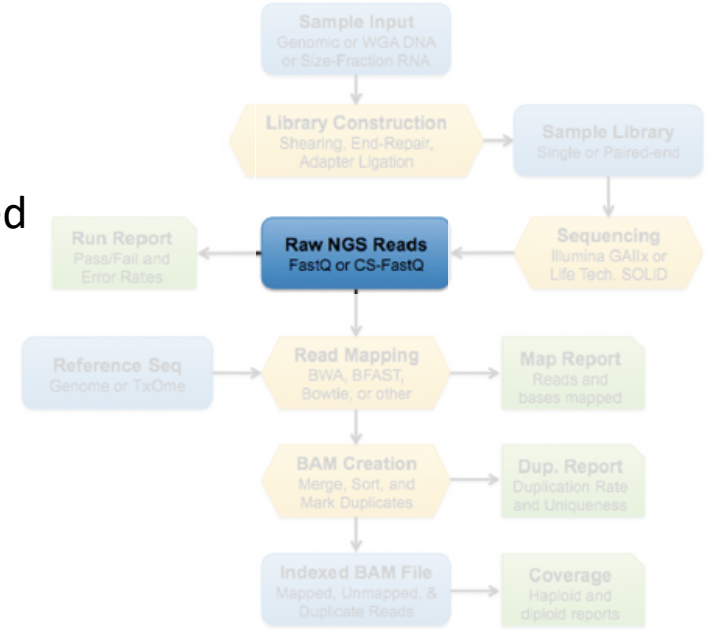
The quality score of a base, also known as a Phred or Q score, is an integer value representing the estimated probability of an error, i.e. that the base is incorrect.

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

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.

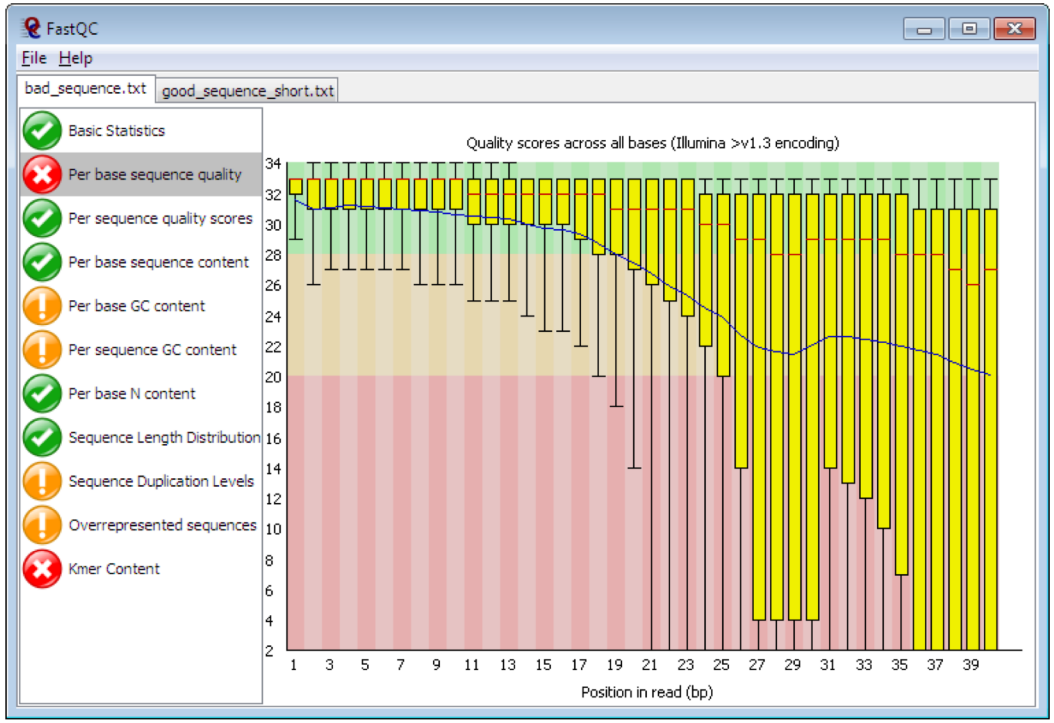
Table 1: Q-Scores and Error Probabilities

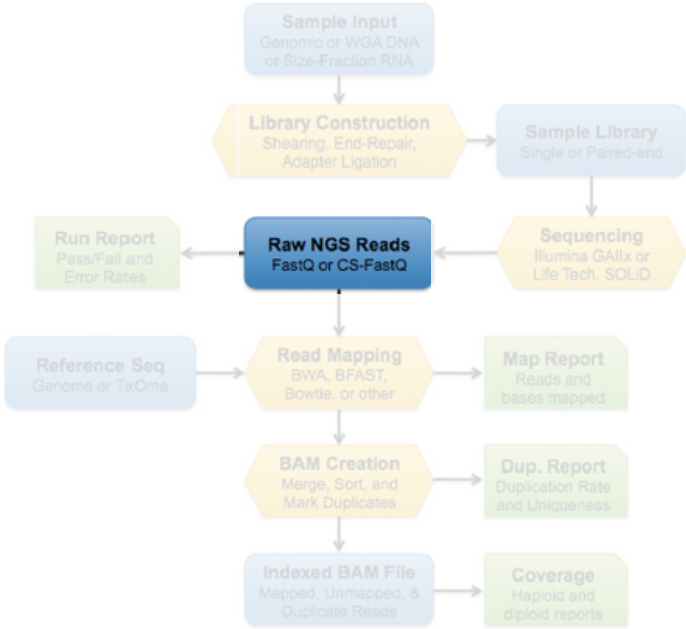
Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)



FastQC

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



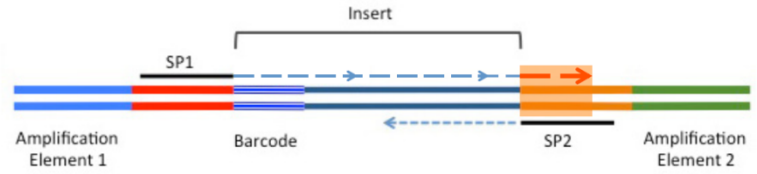


http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help

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.

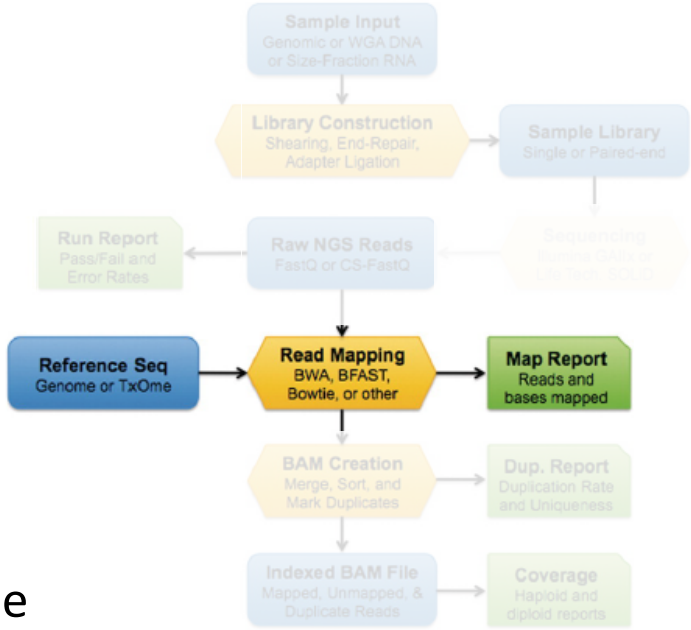


- If the overall quality of the read is low, it must be removed. A trimming is useful if quality decreases too much towards the end of the read.
- Sometimes the read terminates with ambiguous (N) bases which must be removed.
- Some of the most common preprocessing tool are FASTX-Toolkit, Cutadapt, Trimmomatic,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

DATA ANALYSIS: ALIGNMENT

Read alignment

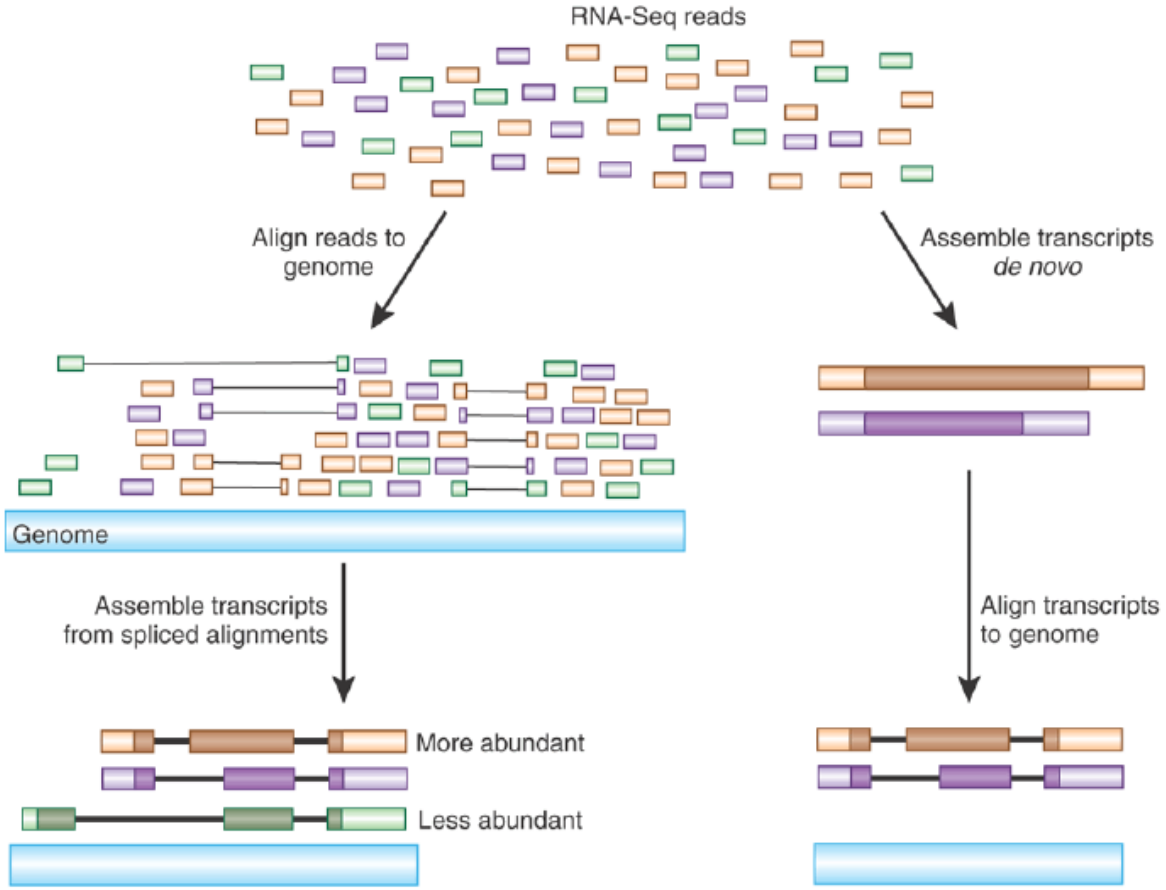
After pre-processing, we can align reads to a reference sequence.



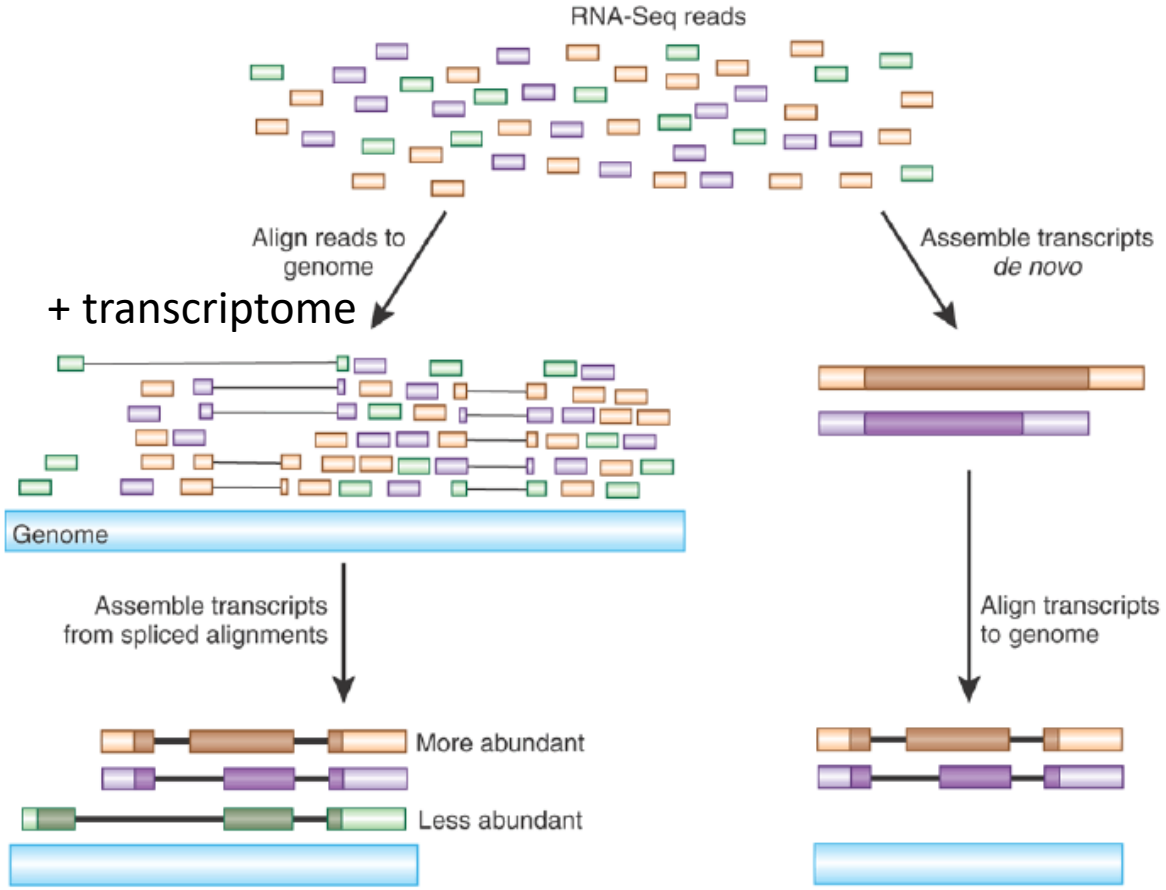
- to align a read means finding the region of the genome to which it belongs.
- if the genome sequence of the organism is known, reads can be aligned to it.
- other approaches have to be used if the genome sequence is not known (de novo transcriptome assembly).

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

Main alignment strategies



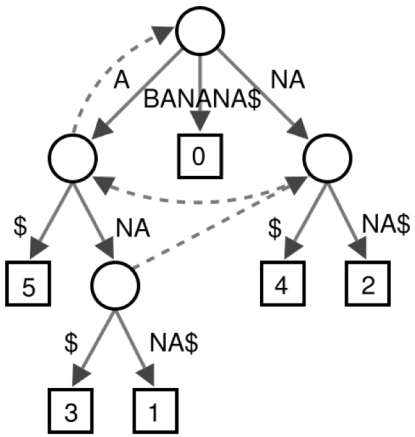
Main alignment strategies



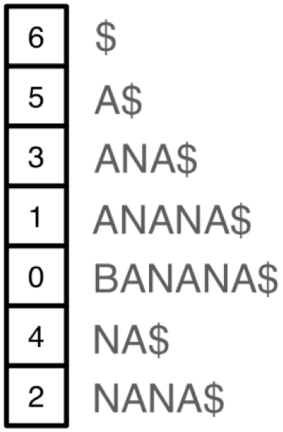
Alignment of NGS reads

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.

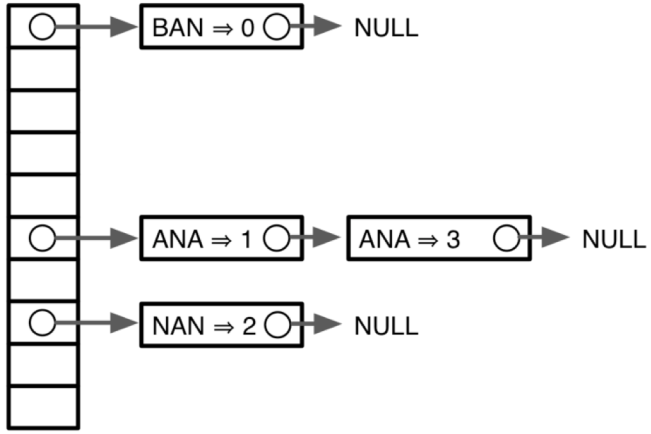
For this reason, short-read aligners are usually based on a preliminary indexing of the reference sequence. The performance of the aligner heavily depends on the way data are indexed.



Suffix tree



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	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

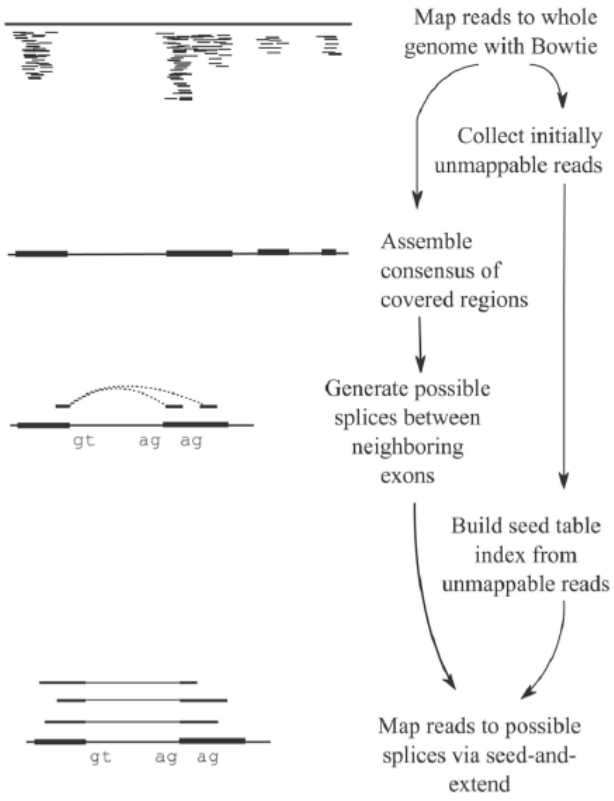
- The algorithms discussed so far are not able to align reads on splicing junctions, unless we use the transcriptome sequence as a reference.

- There are several programs that are able to perform spliced alignments: 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.

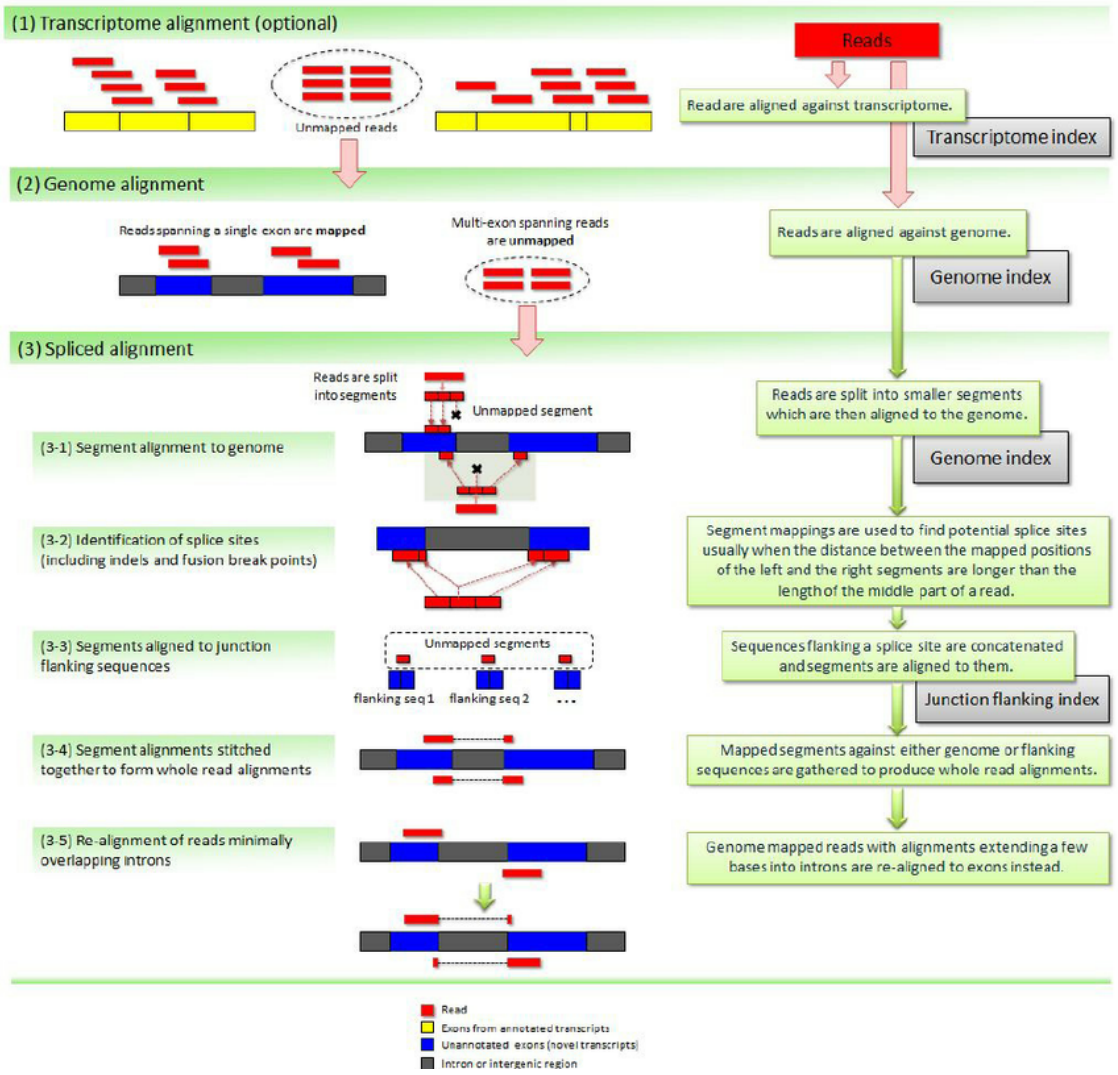
From TopHat manual…

"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



Main alignment programs

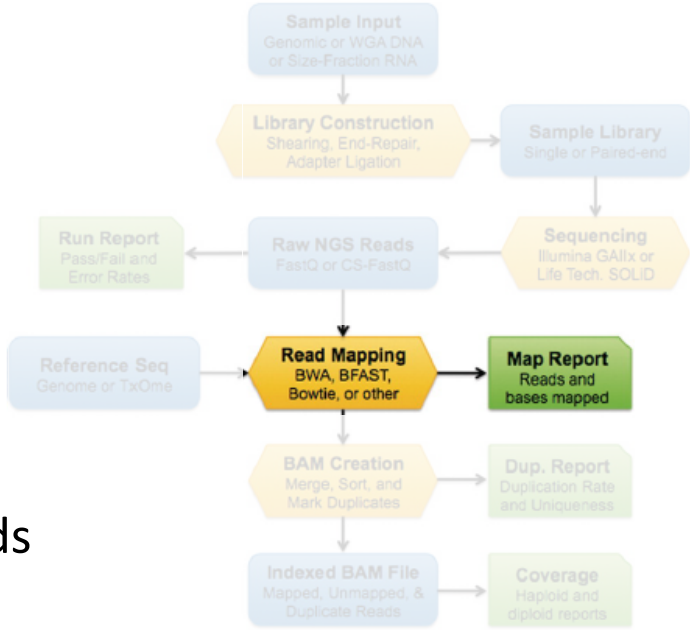
Table 1 | Selected list of RNA-seq analysis programs

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.

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

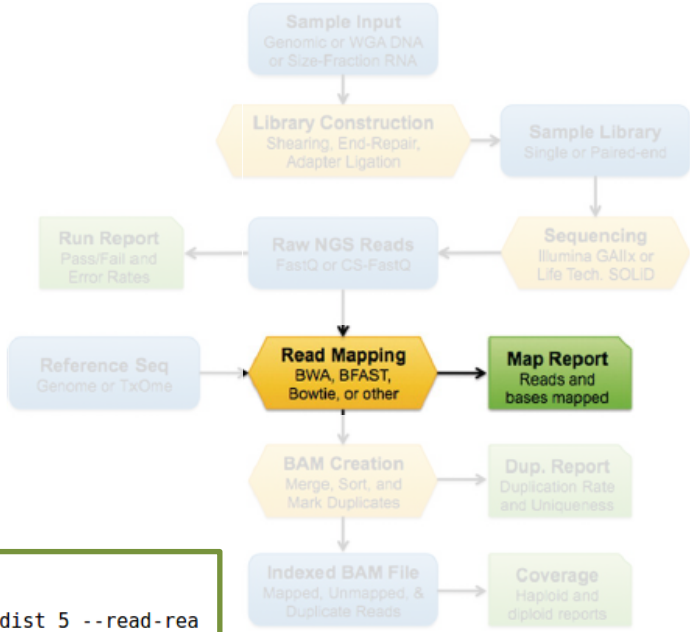
DATA ANALYSIS: ALIGNMENT

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.

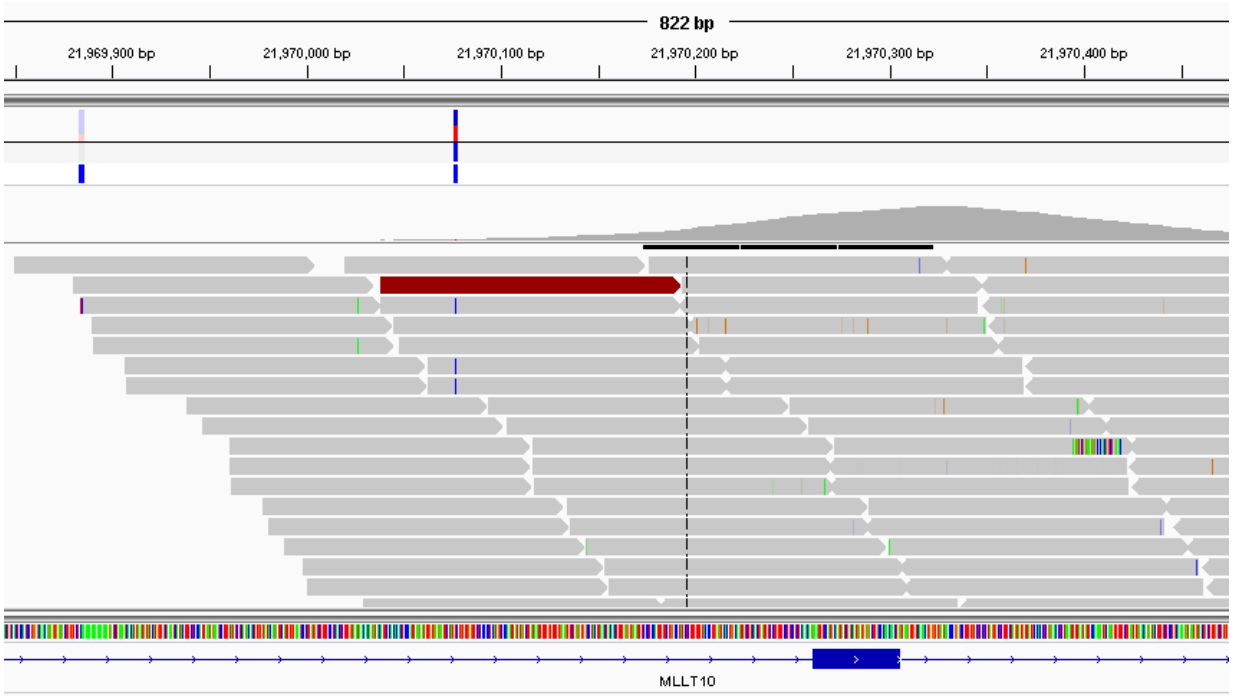



```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       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>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    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

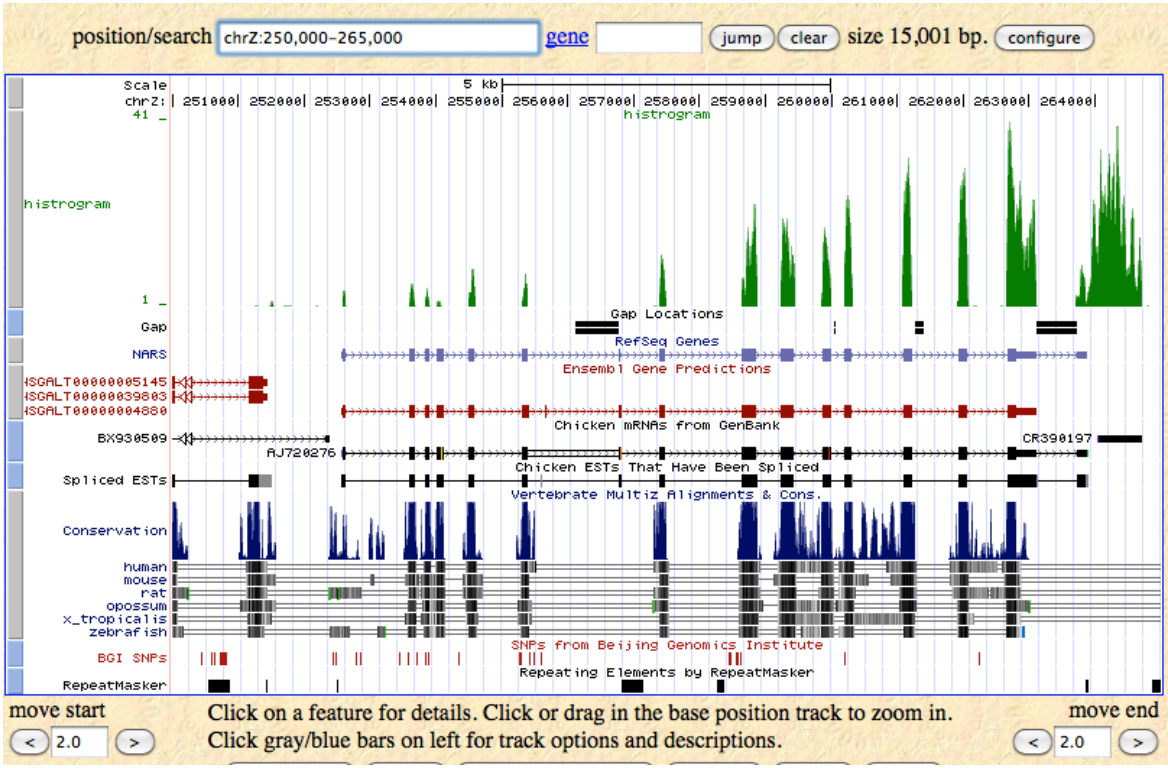
IGV is a standalone program which allows a highly interactive visualization of BAM files (and other genomic annotation formats).



BAM file visualization

Genome Browser (UCSC)

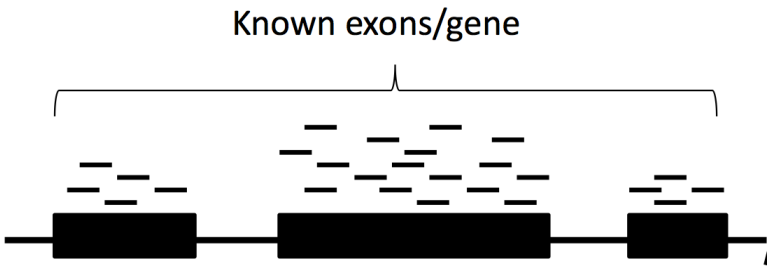
Visualization is less interactive, but many supplementary tracks are available.



DATA ANALYSIS: 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?
  - **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.

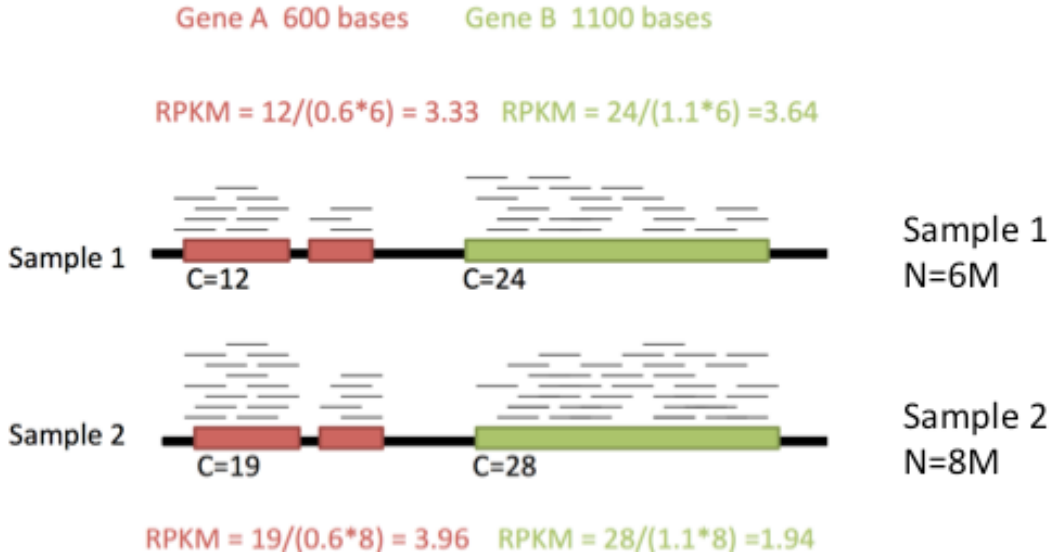
- **RPKM** and **FPKM** are two normalized measures of gene expression.

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

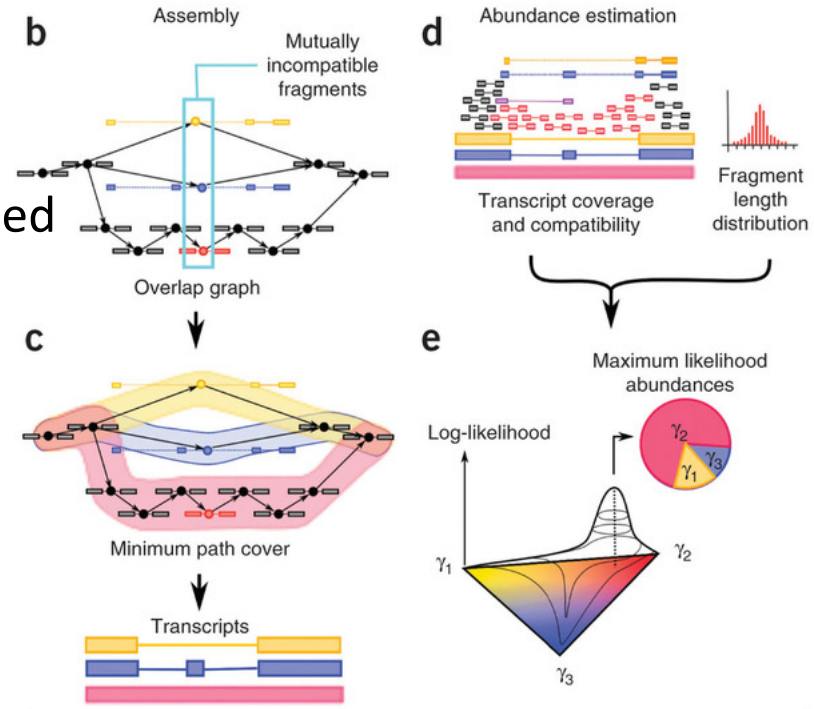
Cufflinks

Cufflinks is a software which is able to:

- Assemble transcripts from reads mapped to the genome,

- estimate the abundance of these transcripts (FPKM),

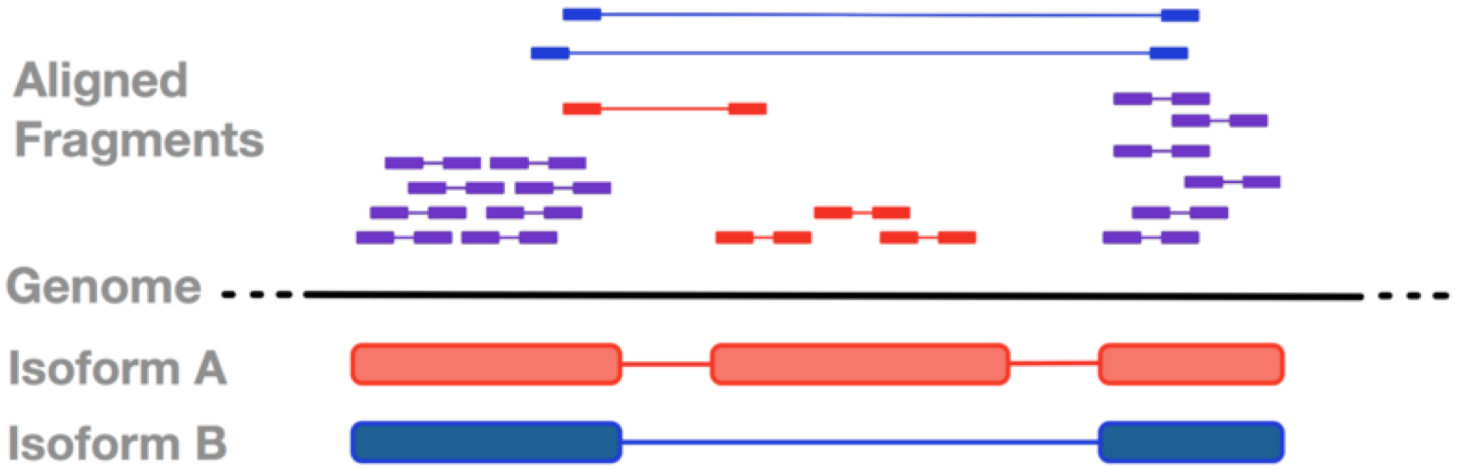
- test for differential expression (DE).



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-ambiguosly 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.

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.

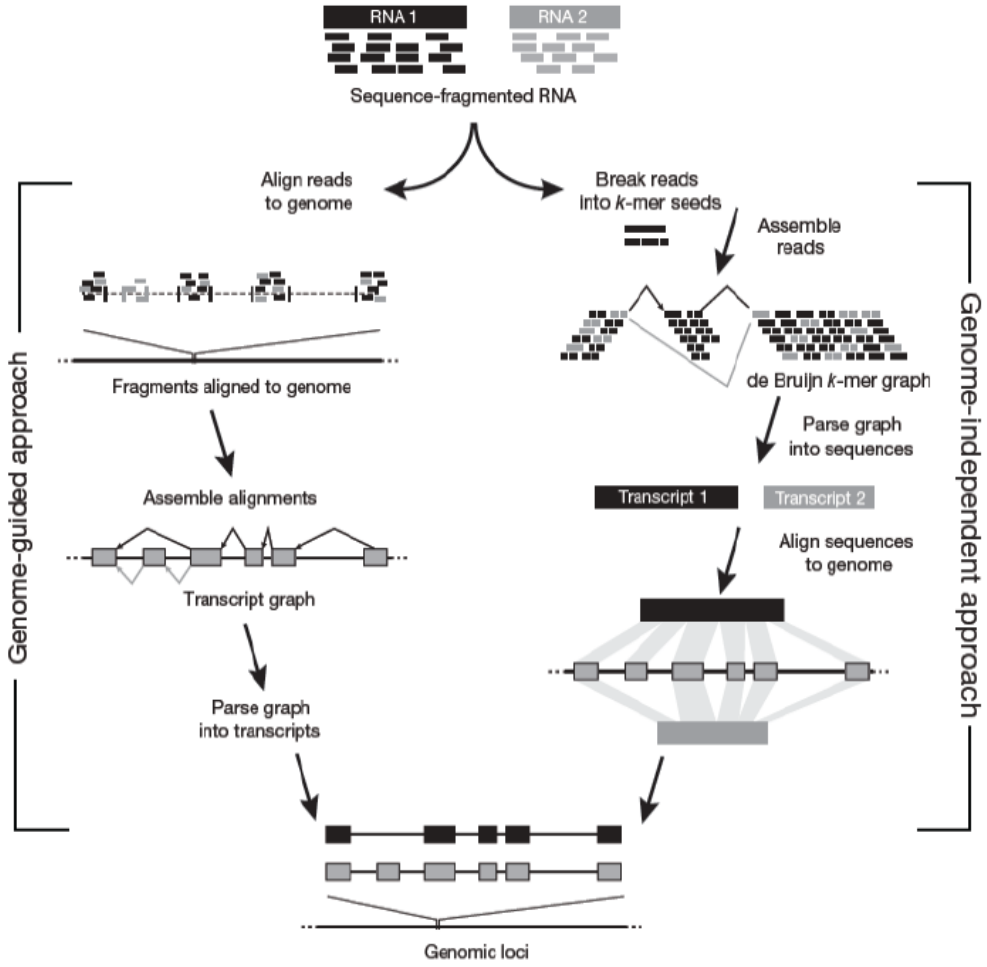
- **genome-independent** programs use the overlap between reads to assemble transcripts; alignment to the genome is not required. They are thus useful in the absence of a reference genome, but also to find transcripts coming from genes which underwent structural variations (indels, fusions etc.). These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

Tools for de novo discovery of transcripts



DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

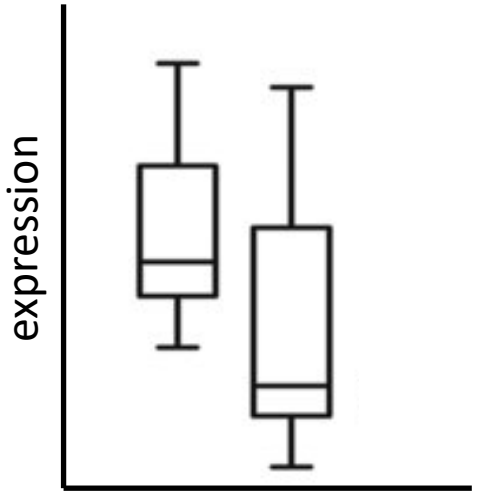
What is differential expression (DE) analysis?

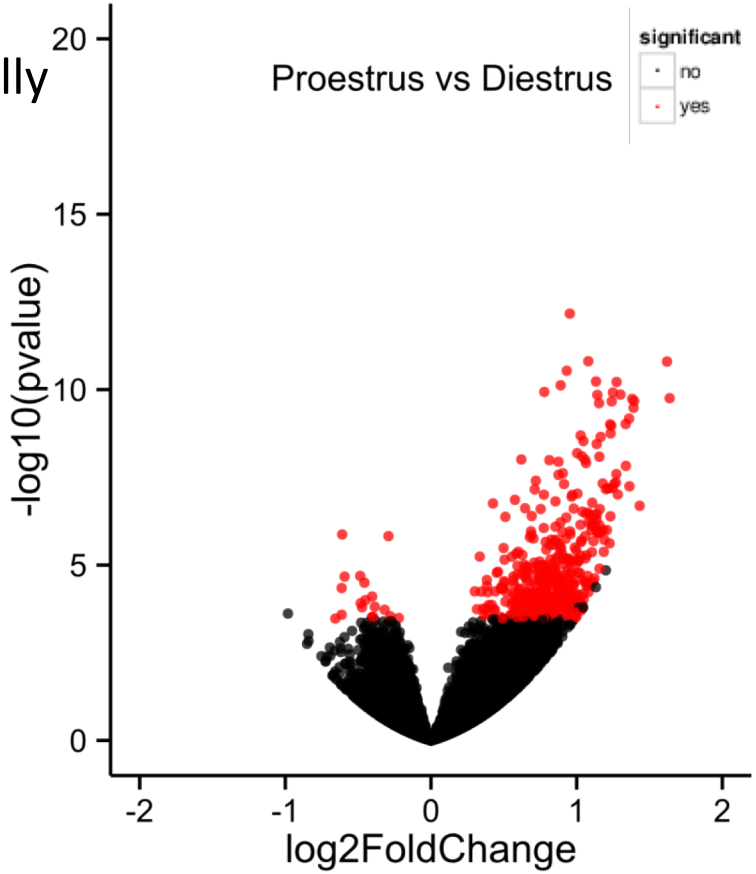
DE analysis allows to find **genes** (or other genomic features like transcripts and exons) **that are expressed at significantly different levels between two groups of samples** (conditions): patients treated with drugs VS controls, healthy VS sick individuals , different tissues and different differentiation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

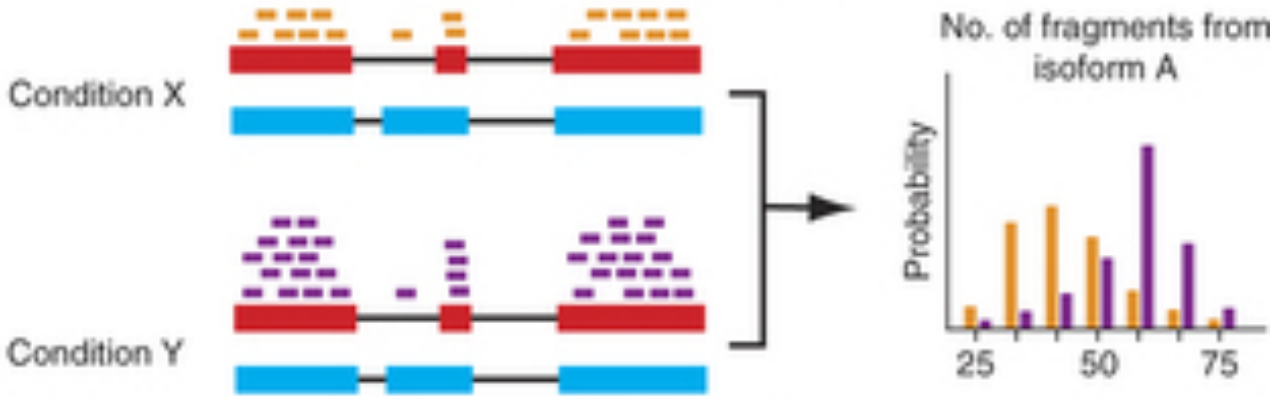
High absolute FC values are not necessarily associated with significant P-values, especially when the expression of the gene is highly variable.





Cuffdiff: differential expression test

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

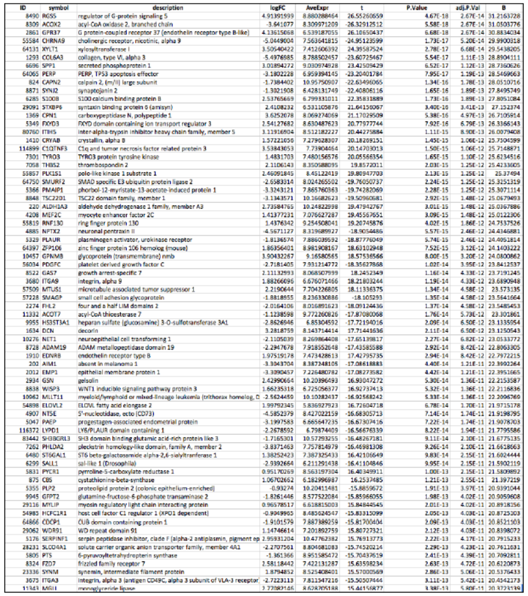


5) Test for signficance of changes between conditions in transcript-level counts

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 ?

Extracting biological meaning from DE gene lists

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

- A list of “interesting” genes.

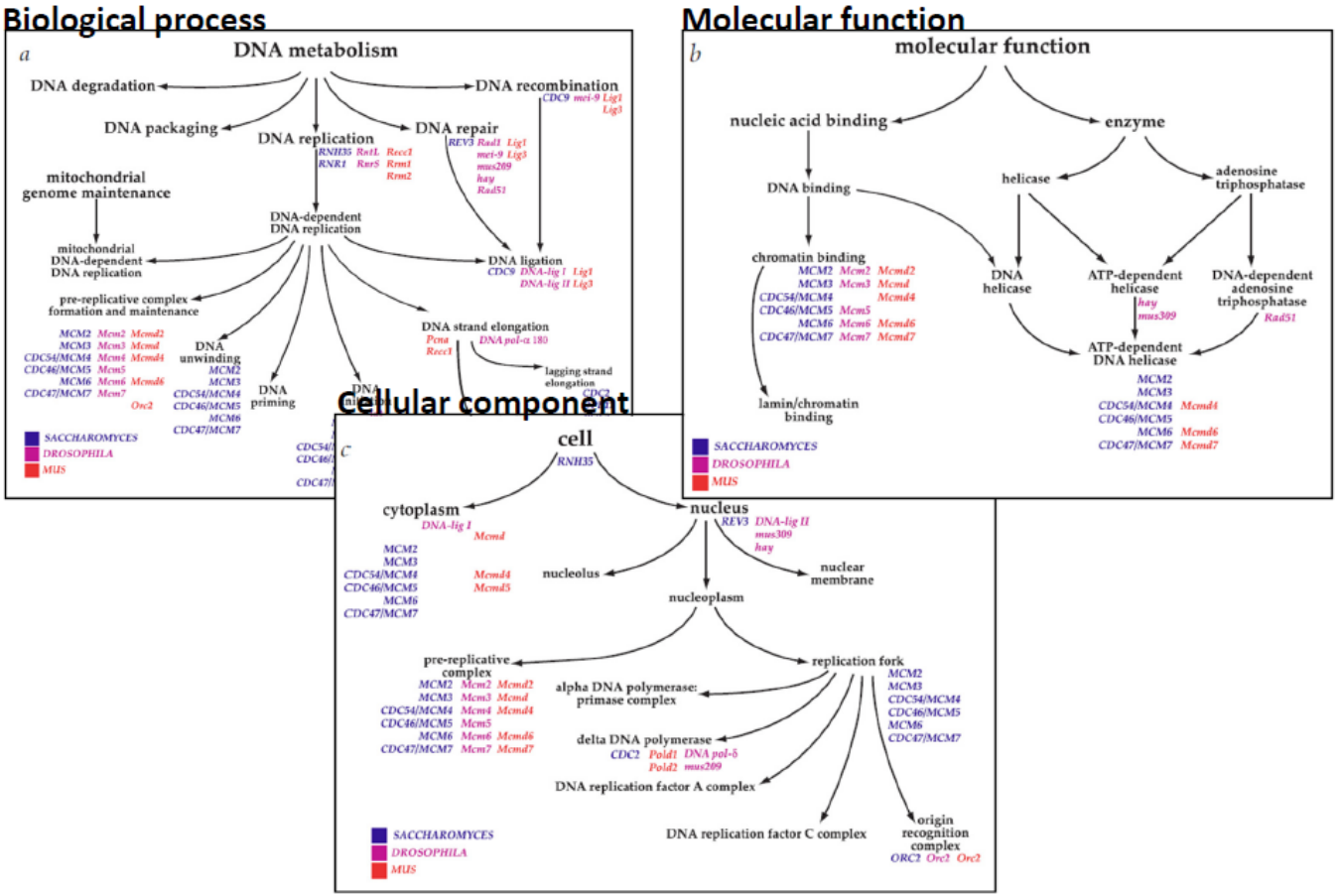
- A background gene list, representing the “universe” of possible genes that could be called as significantly regulated in the experiment. This list should contain only genes that are “called” as expressed (to avoid biological bias) in the experiment.

- 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.

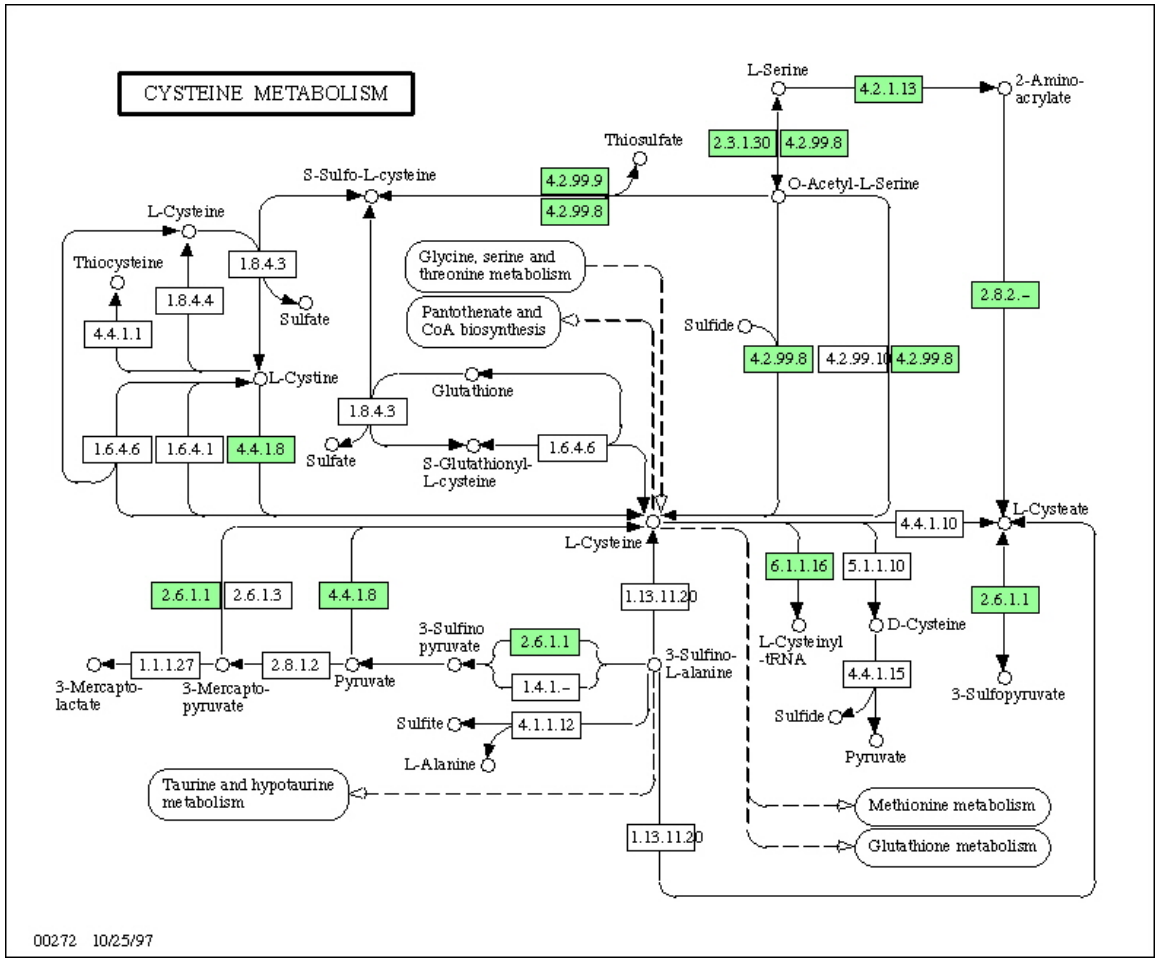
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of functional categories: Gene Ontology.



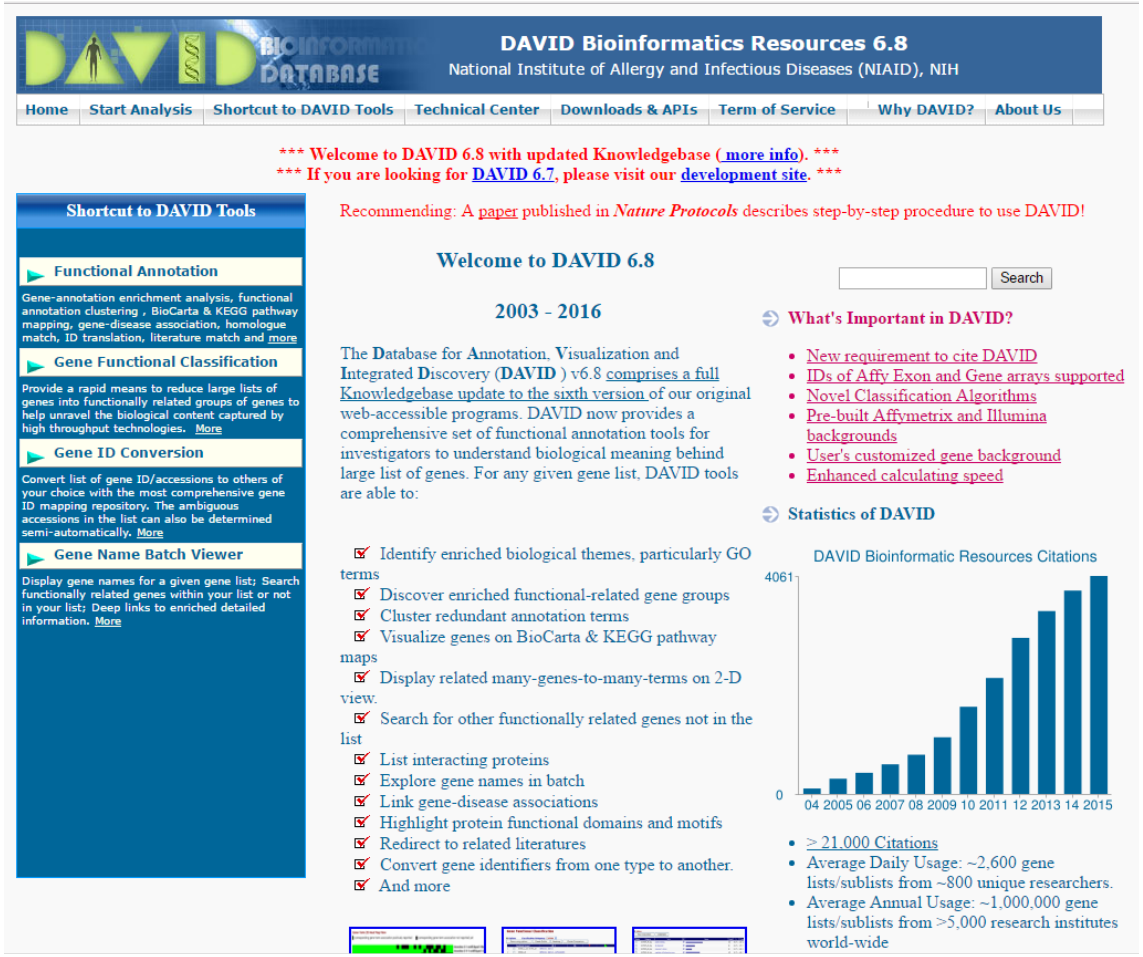
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of functional categories: Kegg pathway.



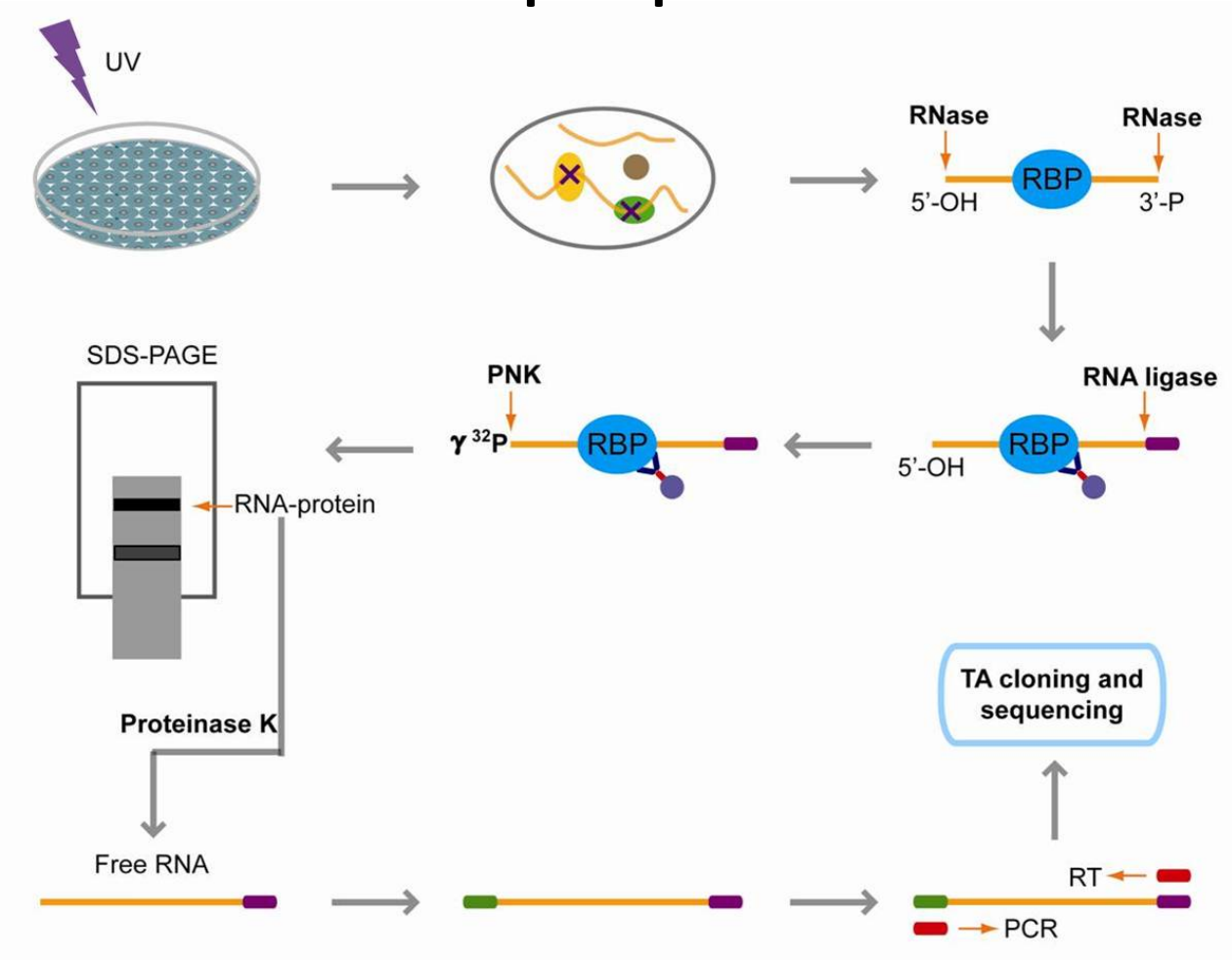
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of online functional annotation tool: DAVID.



CLIP-Seq

high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation



CLIP-Seq: DATA ANALYSIS