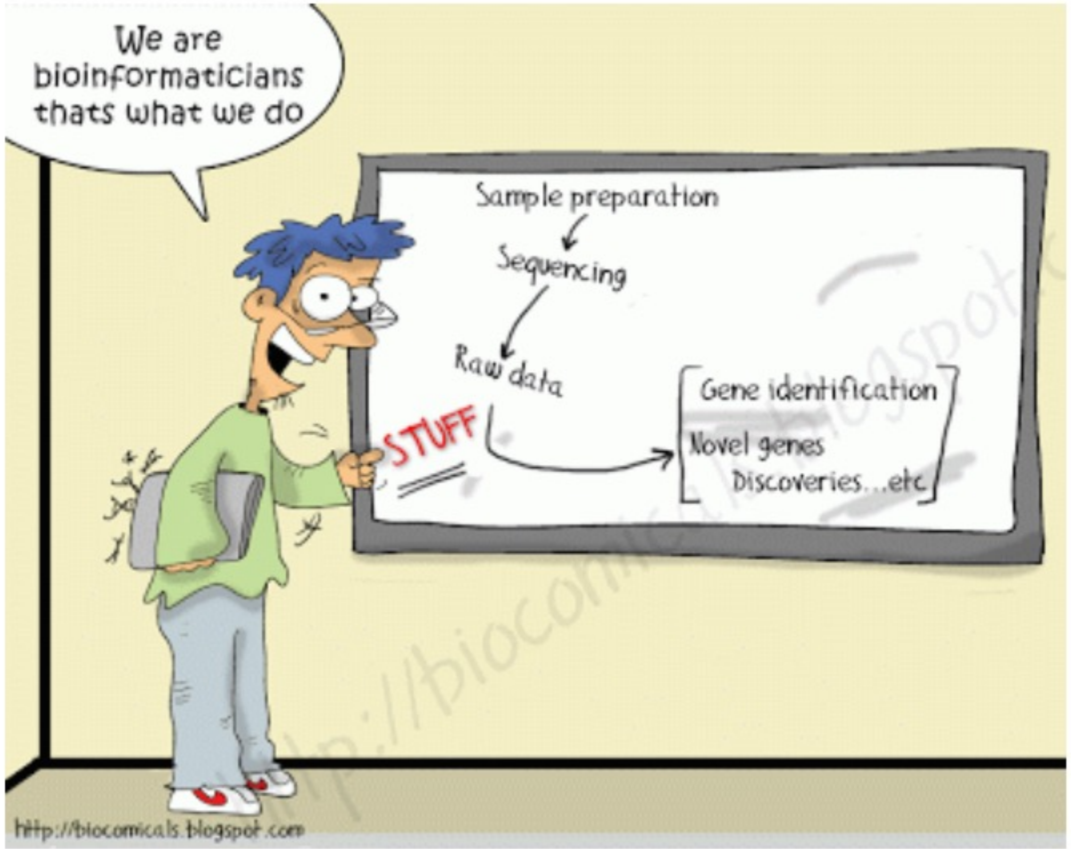
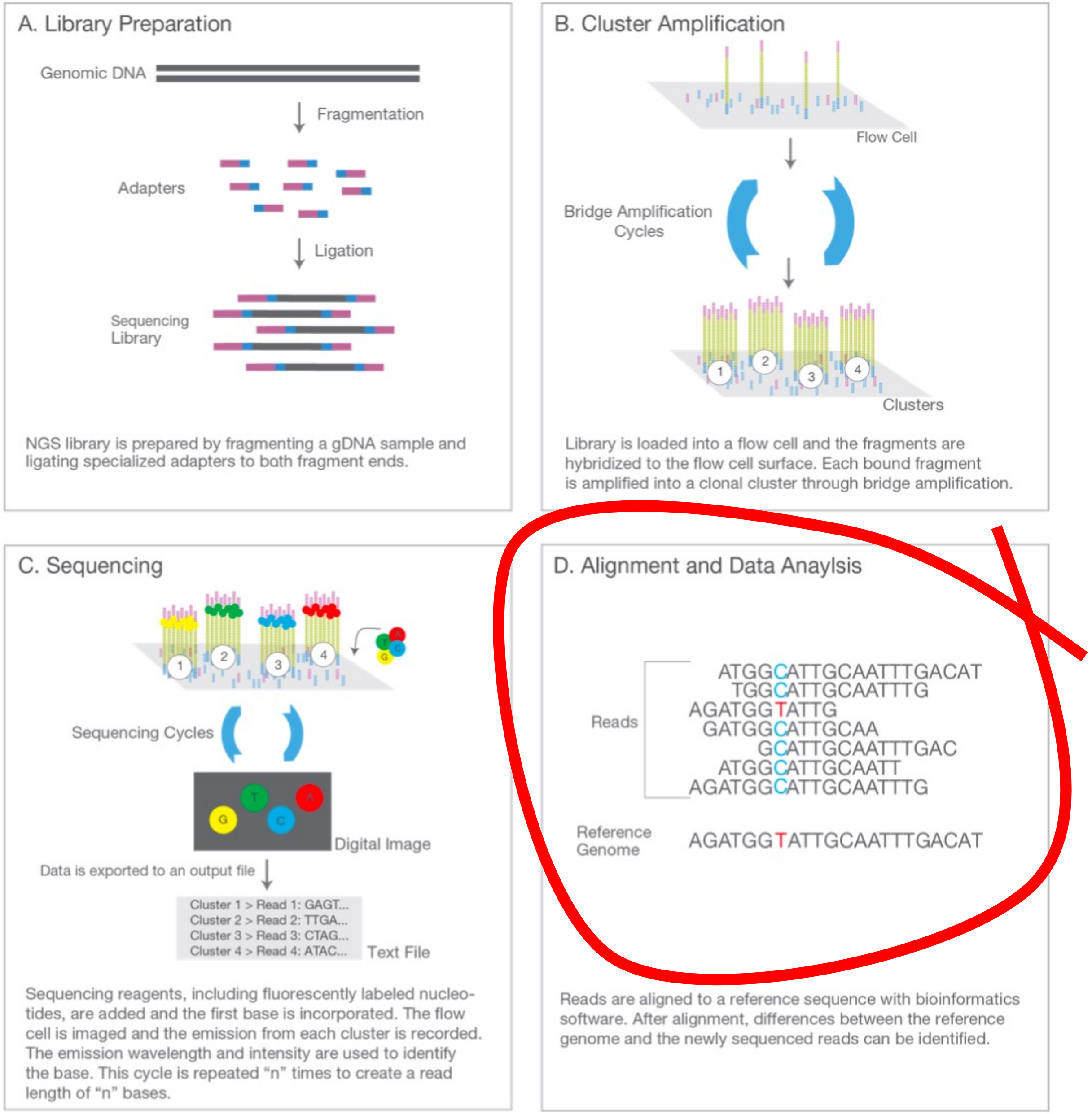
DATA ANALYSIS



Next-Generation Sequencing Overview



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

```
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
```

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

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.

EXPERIMENTAL DESIGN

Defining the technical details

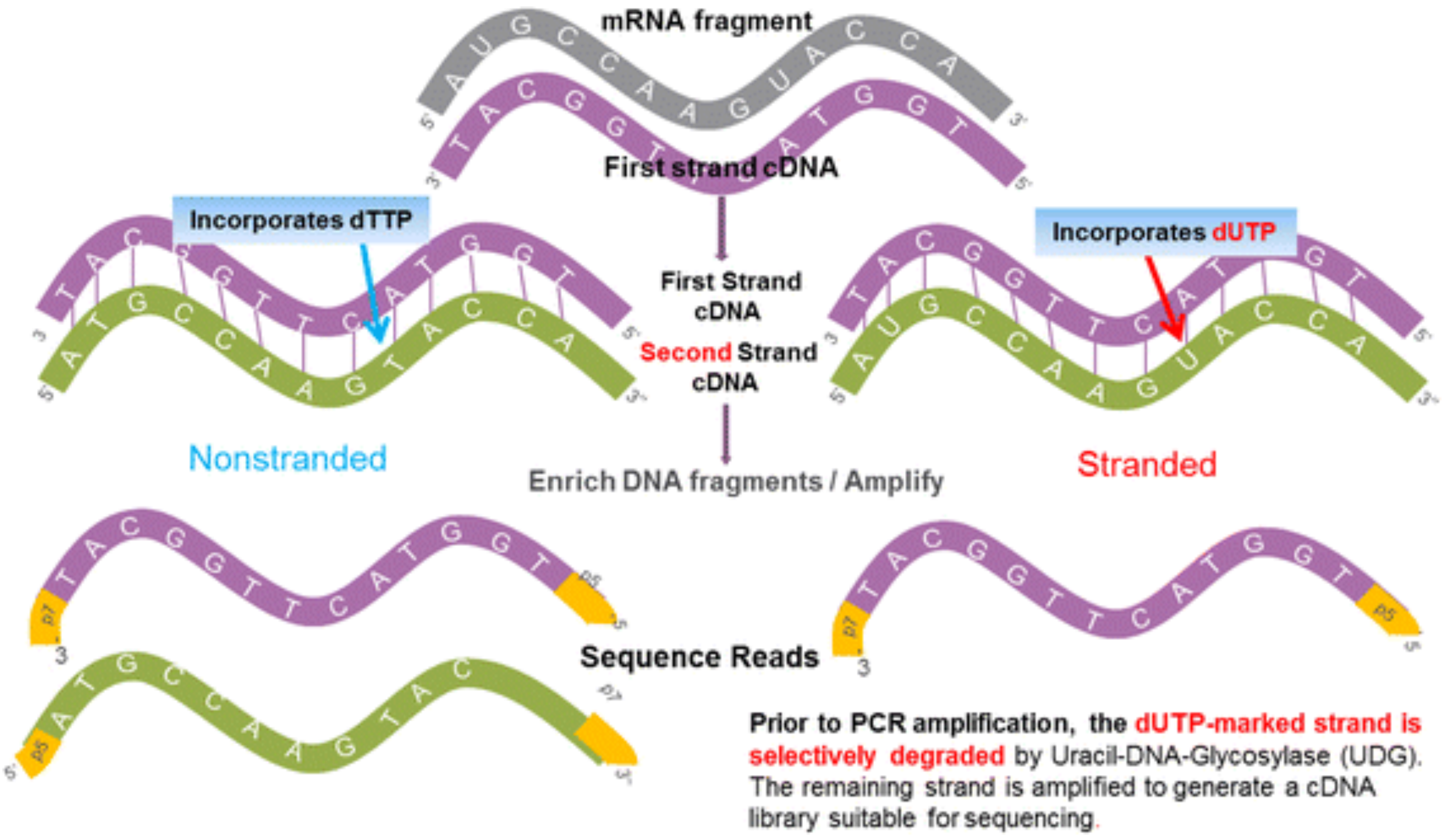
- **Choice of sequencing depth**
  If we want to measure the expression of known genes, depth can be relatively low (e.g. 20 M reads for polyA+). If we want to discover new genes and transcripts, depth must be higher (e.g. 60 M for polyA+, 120 for total RNA).

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

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

Non-stranded versus stranded RNA-seq protocol

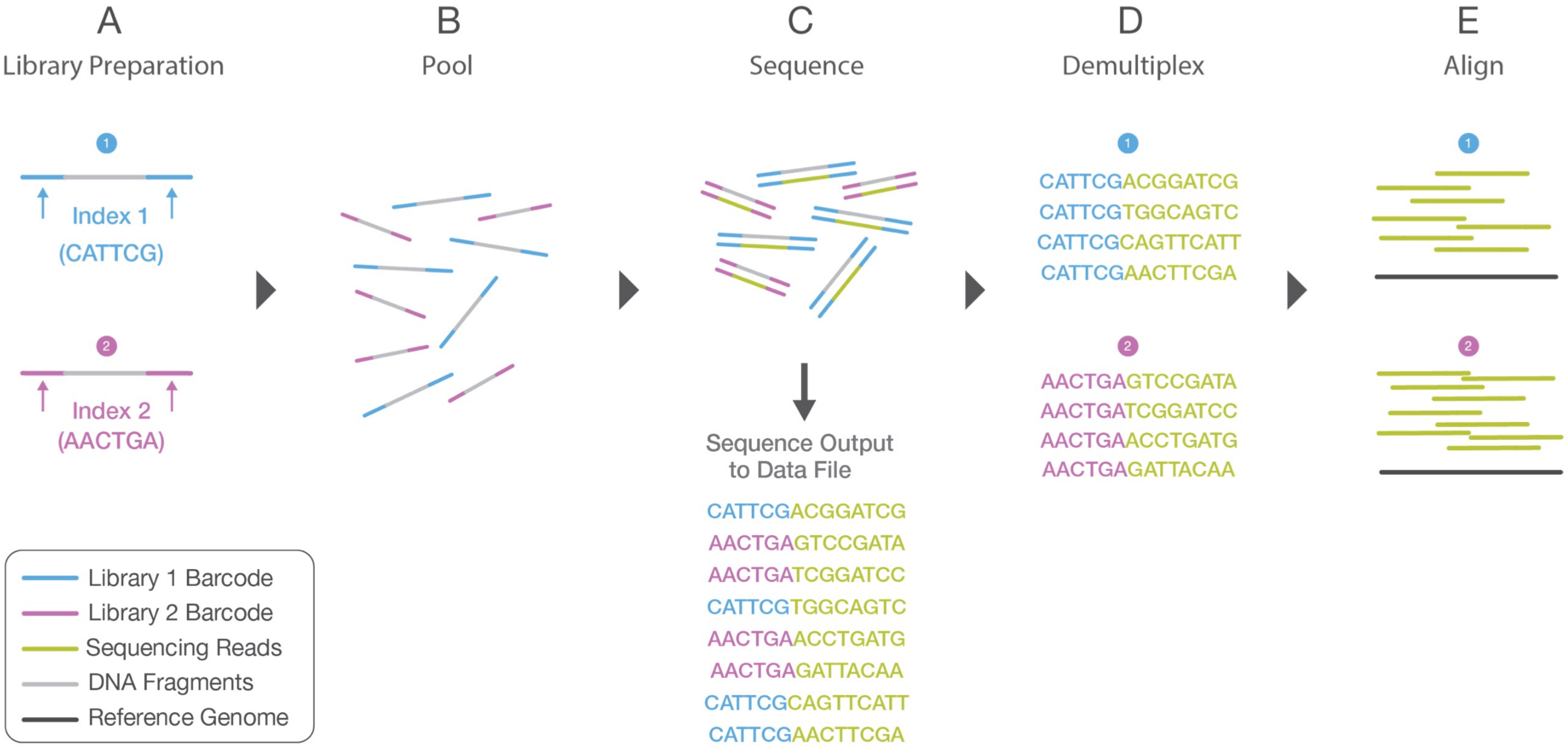
The stranded protocol differs from the non-stranded protocol in two ways. First, during cDNA synthesis, the second-strand synthesis continues as normal except the nucleotide mix includes dUTPs instead of dTTPs. Second, after library preparation, a second-strand digestion step is added. This step ensures that only the first strand survives the subsequent PCR amplification step and hence the strand information of the libraries.



Prior to PCR amplification, the dUTP-marked strand is selectively degraded by Uracil-DNA-Glycosylase (UDG). The remaining strand is amplified to generate a cDNA library suitable for sequencing.

Library Multiplexing Overview

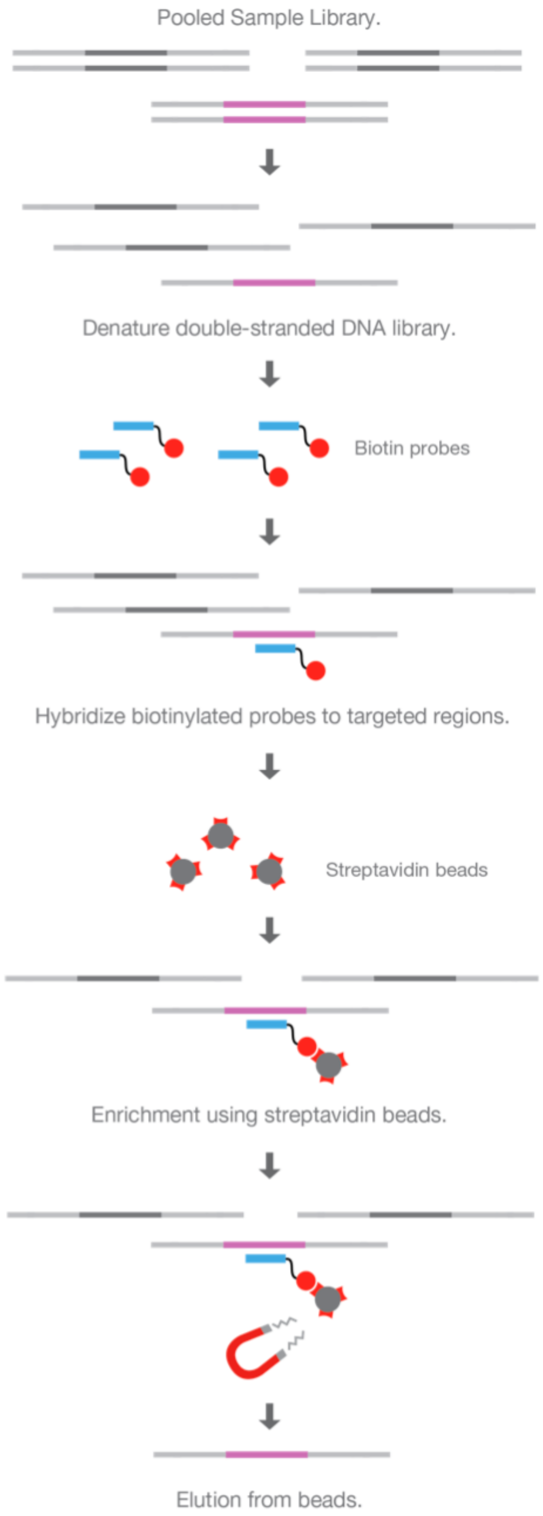
A) Unique index sequences are added to two different libraries during library preparation. (B) Libraries are pooled together and loaded into the same flow cell lane. (C) Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file. (D) A demultiplexing algorithm sorts the reads into different files according to their indexes. (E) Each set of reads is aligned to the appropriate reference sequence



Target Enrichment Workflow

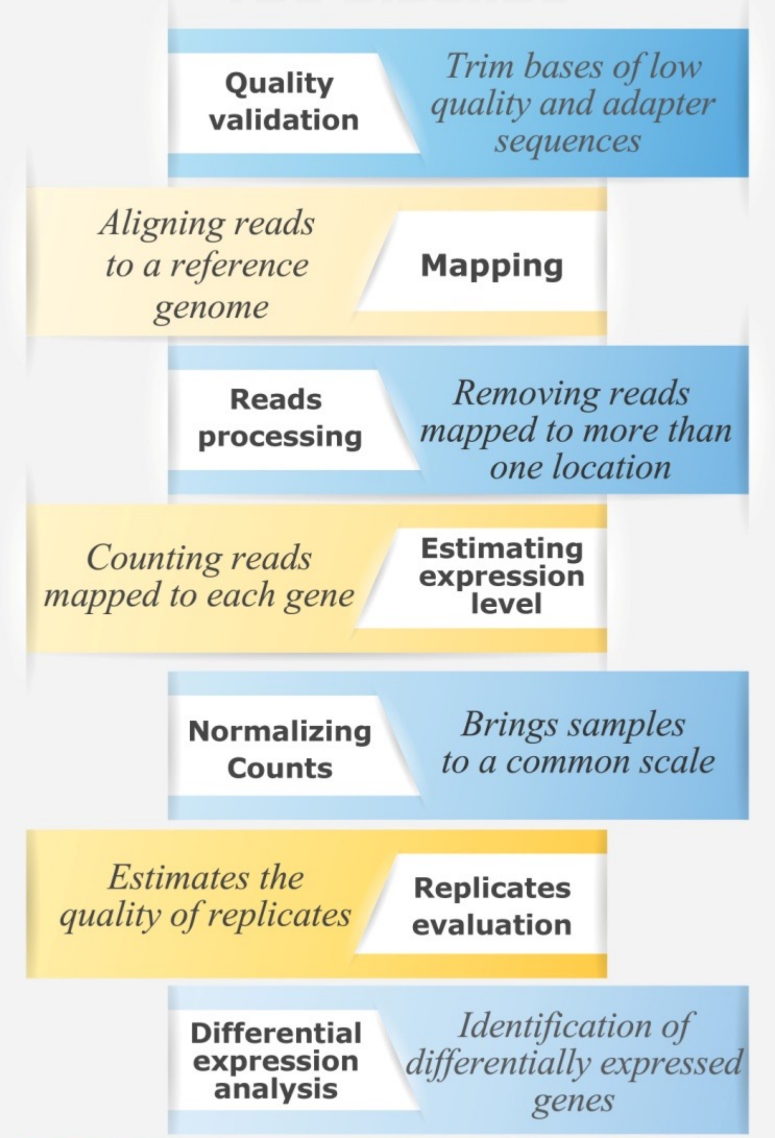
With targeted sequencing, a subset of genes or regions of the genome are isolated and sequenced. Targeted sequencing allows researchers to focus time, expenses, and data analysis on specific areas of interest and enables sequencing at much higher coverage levels. For example, a typical WGS study achieves coverage levels of 30–50× per genome, while a targeted resequencing project can easily cover the target region at 500–1000× or higher. This higher coverage allows researchers to identify **rare variants**, variants that would be too rare and too expensive to identify with WGS or CE-based sequencing.

Targeted sequencing panels can be purchased with fixed, preselected content or can be custom designed. A wide variety of targeted sequencing library prep kits are available, including kits with probe sets focused on specific areas of interest such as cancer, cardiomyopathy, or epidemiology.


Pooled Sample Library.
Denature double-stranded DNA library.
Biotin probes
Hybridize biotinylated probes to targeted regions.
Streptavidin beads
Enrichment using streptavidin beads.
Elution from beads.

DATA ANALYSIS

- **General RNA-Seq pipeline for Differential Expression**



Quality validation
Trim bases of low quality and adapter sequences

Aligning reads to a reference genome
Mapping

Reads processing
Removing reads mapped to more than one location

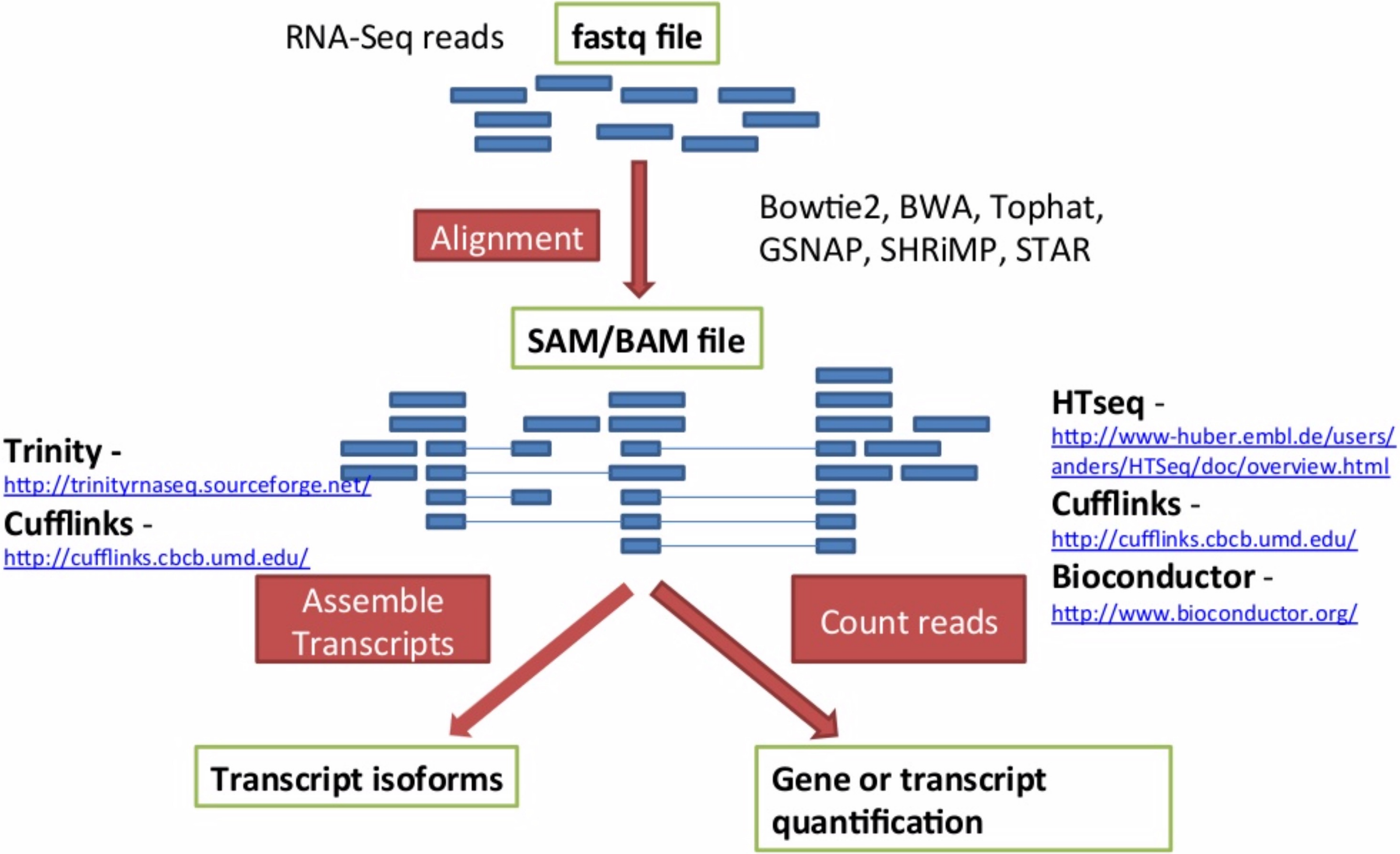
Counting reads mapped to each gene
Estimating expression level

Normalizing Counts
Brings samples to a common scale

Estimates the quality of replicates
Replicates evaluation

Differential expression analysis
Identification of differentially expressed genes

Generalized Analysis Workflow



RNA-Seq reads

fastq file

Alignment

Bowtie2, BWA, Tophat, GSNAP, SHRiMP, STAR

SAM/BAM file

Trinity - http://trinityrnaseq.sourceforge.net/

Cufflinks - http://cufflinks.cbcb.umd.edu/

HTseq - http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html

Cufflinks - http://cufflinks.cbcb.umd.edu/

Bioconductor - http://www.bioconductor.org/

Assemble Transcripts

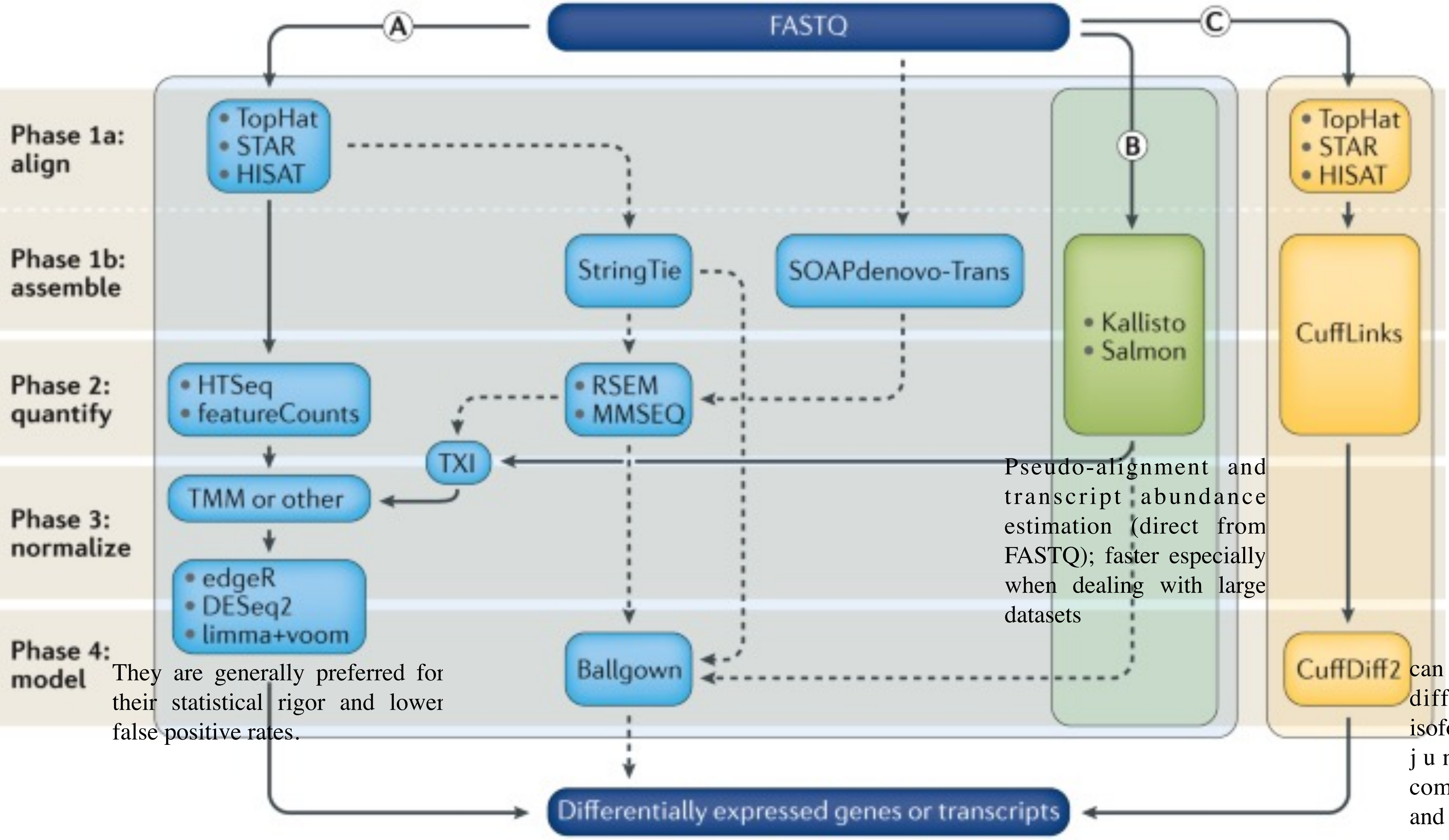
Count reads

Transcript isoforms

Gene or transcript quantification

Workflow for differential gene expression

Computational analysis for differential gene expression (DGE) begins with raw RNA sequencing (RNA-seq) reads in FASTQ format and can follow a number of paths. Three popular workflows (A, B and C, represented by the solid lines) are given as examples, and some of the more common alternative tools (represented by the dashed lines) are indicated.



They are generally preferred for their statistical rigor and lower false positive rates.

Pseudo-alignment and transcript abundance estimation (direct from FASTQ); faster especially when dealing with large datasets

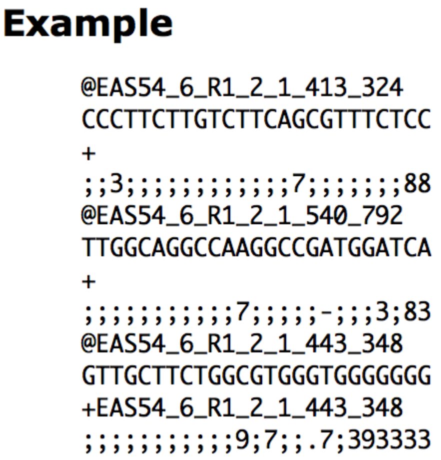
can be used to identify differentially expressed isoforms and analyze splice junction usage. computationally intensive and high false positive rate

DATA ANALYSIS

Data format

- Usually, the format of the file containing the sequence of the reads is FASTQ.
  It is composed of four-lines blocks:
  - **the first line** begins with @ and contains the ID of the read and optional information.
  - **the second line** is the sequence
  - the third line begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again
  - **the fourth line** encodes the quality values for the sequence in Line 2.

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

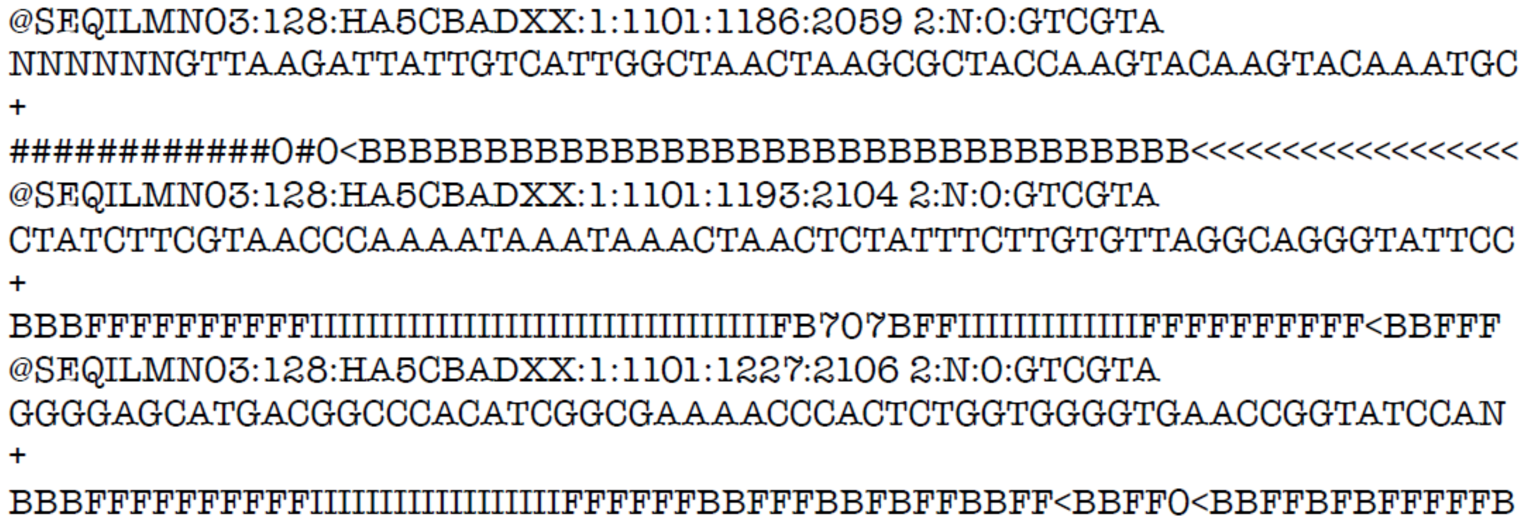


Example

```
@EAS54_6_R1_2_1_413_324
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
```

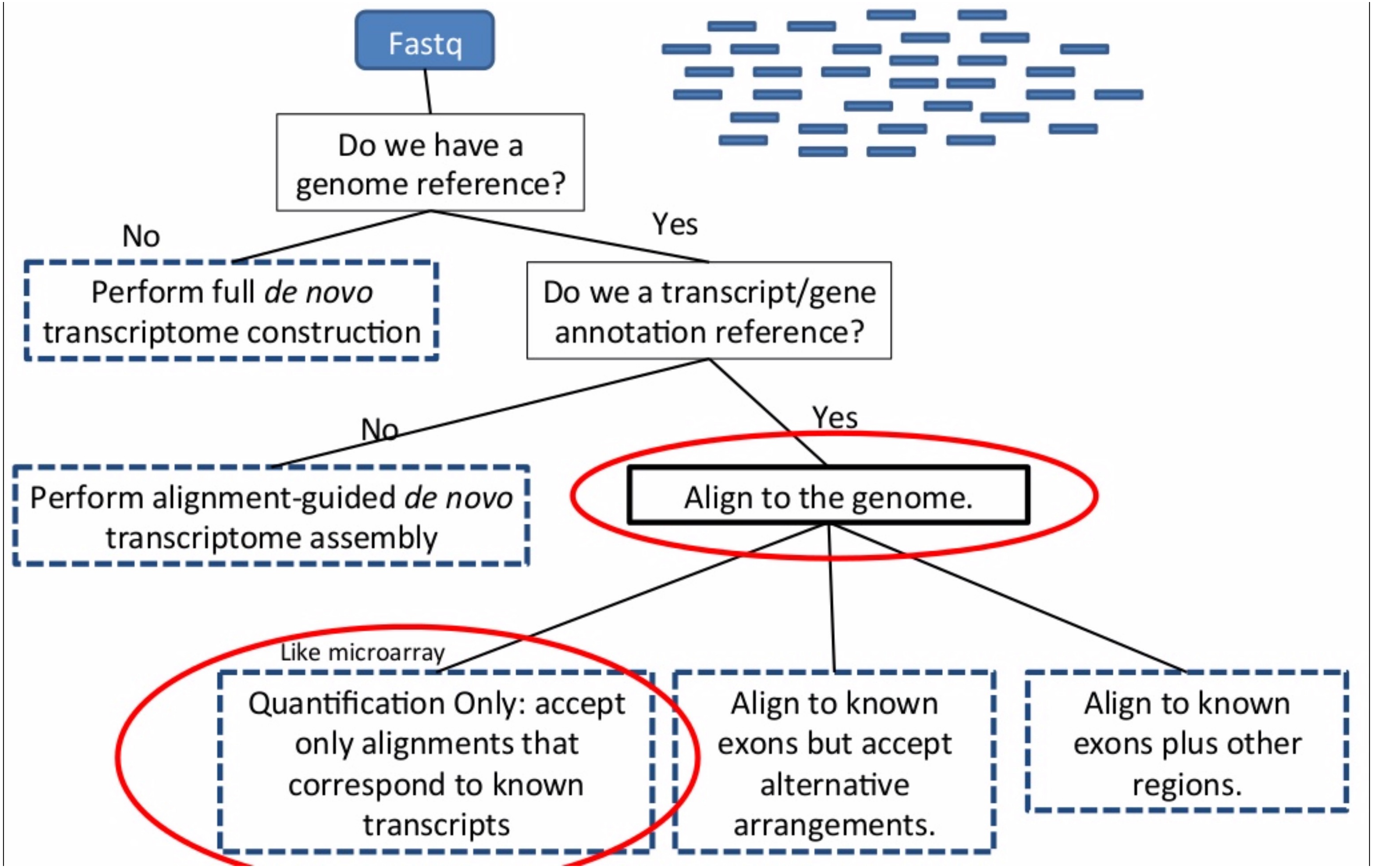
DATA ANALYSIS

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
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFB707BFFIIIIIIIIIIIIIIIFFFFFFFFFF<BBFFF
@SEQILMN03:128:HA5CBADXX:1:1101:1227:2106 2:N:0:GTCGTA
GGGGAGCATGACGGCCCACATCGGCGAAAACCCACTCTGGTGGGGTGAACCGGTATCCAN
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIFFFFFFBBFFFBBFBFFBBFF<BBFF0<BBFFBFBFFFFFB
```

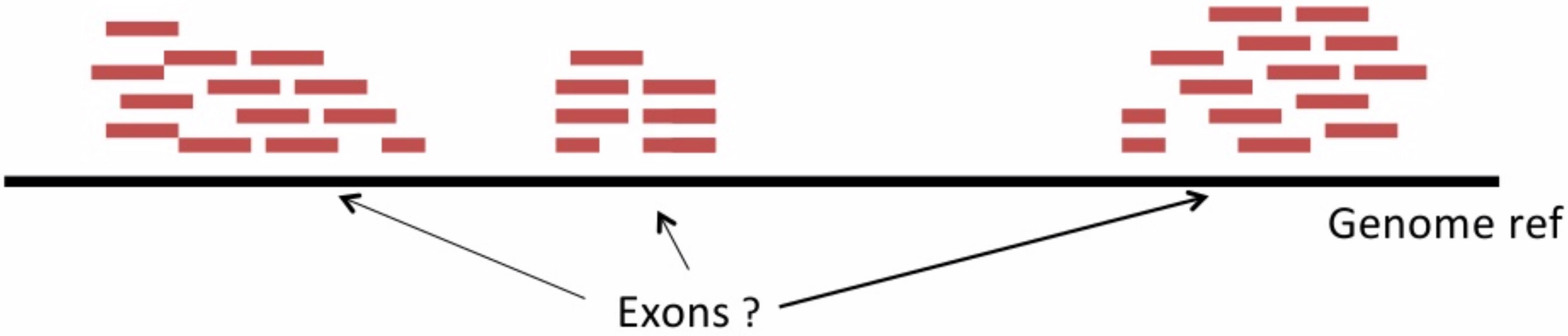
DATA ANALYSIS: ALIGNMENT



Fastq
Do we have a genome reference?
No
Yes
Perform full *de novo* transcriptome construction
Do we a transcript/gene annotation reference?
No
Yes
Perform alignment-guided *de novo* transcriptome assembly
Align to the genome.
Like microarray
Quantification Only: accept only alignments that correspond to known transcripts
Align to known exons but accept alternative arrangements.
Align to known exons plus other regions.

What to map to?

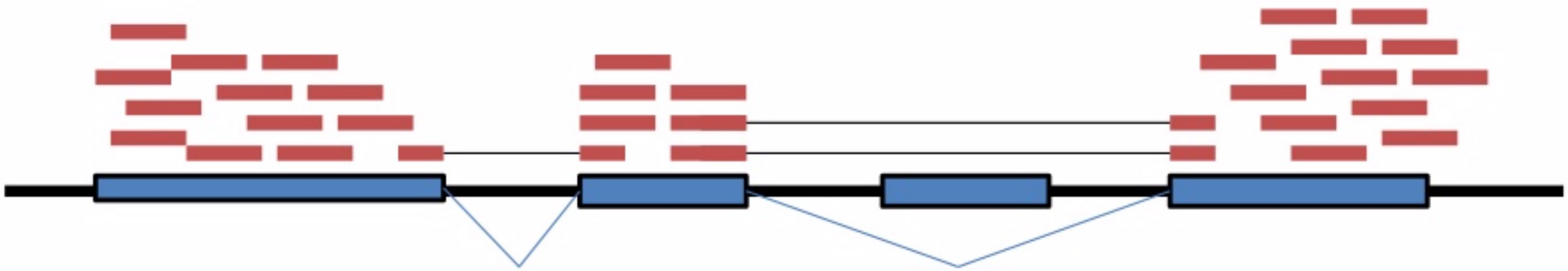
Map to a genome with no gene annotation.



- Assembling transcripts from exon regions is difficult and requires complex statistical algorithms.
- Identifying alternative transcript isoforms is unreliable.
- Usually this is best for a novel or unannotated genomes.

What to map to?

Map to the genome, with knowledge of transcript annotations



- •Well annotated genome reference is required.
- •To effecively map to exon junctions, you need a mapping algorithm that can divide the sequencing reads and map portions independently.
- •Identifying alternative transcript isoforms involves complex algorithms.

Which sequence mappers to use?

- RNASeq Alignment algorithm must be
  - Fast
  - Able to handle SNPs, indels, and sequencing errors
  - Maintain accurate quantification
  - Allow for introns for reference genome alignment(spliced alignment detection)
- Burrows Wheeler Transform(BWT) mappers
  - Fast
  - Limited mismatches allowed (<3)
  - Limited indel detection ability
  - Examples: Bowtie2, BWA, Tophat
  - Use cases: large and conserved genome and transcriptomes
- Hash Table mappers
  - Require large amount of RAM for indexing
  - More mismatches allowed
  - Indel detection
  - Examples: GSNAP, SHRiMP, STAR
  - Use case: highly variable or smaller genomes, transcriptomes

DATA ANALYSIS: ALIGNMENT

Alignment output

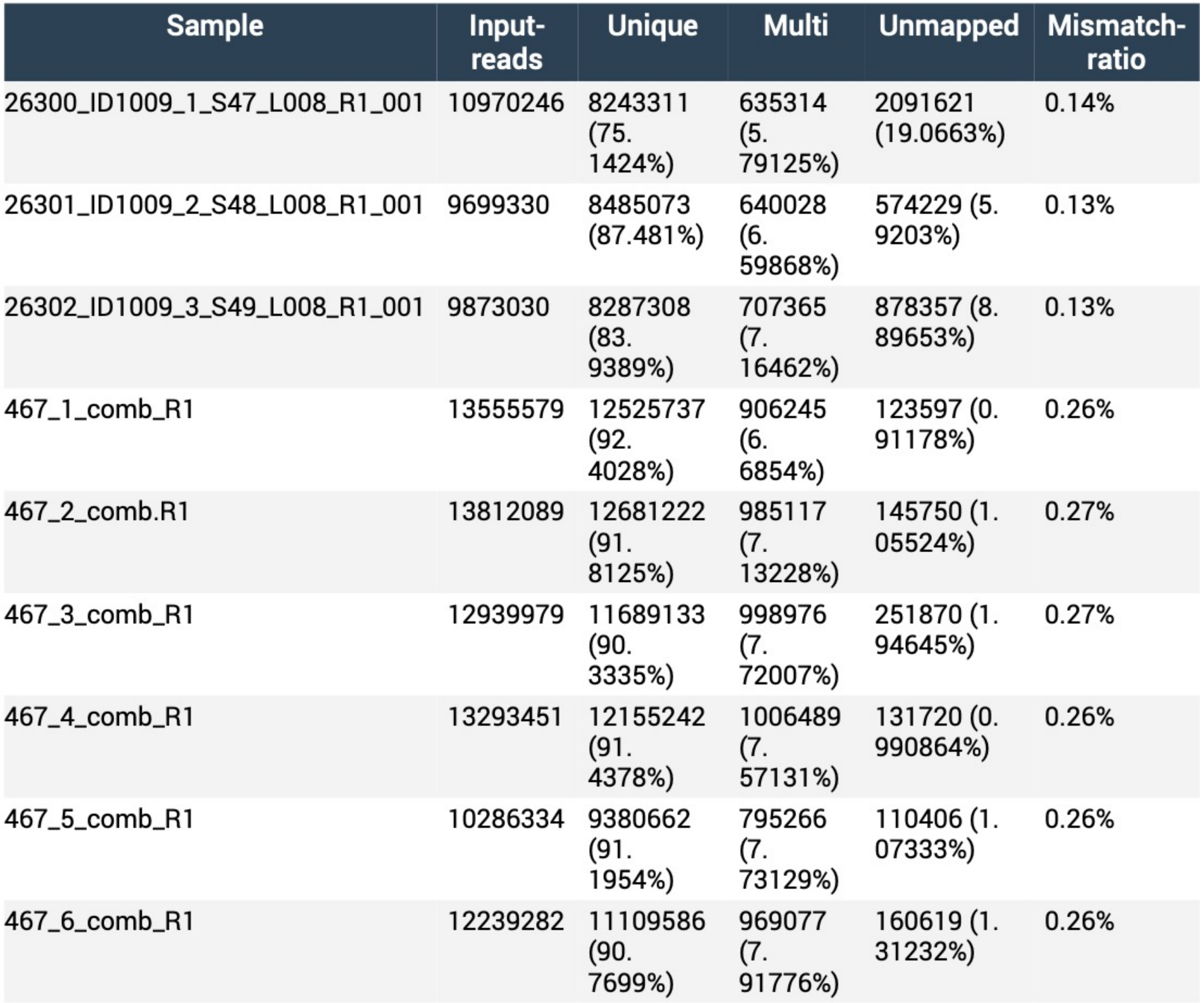
- After alignment, mapped and unmapped reads are usually exported in SAM/BAM format.

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

  - **BAM** is the binary equivalent of SAM.

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

DATA ANALYSIS: ALIGNMENT



Sample	Input-reads	Unique	Multi	Unmapped	Mismatch-ratio
26300_ID1009_1_S47_L008_R1_001	10970246	8243311 (75.1424%)	635314 (5.79125%)	2091621 (19.0663%)	0.14%
26301_ID1009_2_S48_L008_R1_001	9699330	8485073 (87.481%)	640028 (6.59868%)	574229 (5.9203%)	0.13%
26302_ID1009_3_S49_L008_R1_001	9873030	8287308 (83.9389%)	707365 (7.16462%)	878357 (8.89653%)	0.13%
467_1_comb_R1	13555579	12525737 (92.4028%)	906245 (6.6854%)	123597 (0.91178%)	0.26%
467_2_comb.R1	13812089	12681222 (91.8125%)	985117 (7.13228%)	145750 (1.05524%)	0.27%
467_3_comb_R1	12939979	11689133 (90.3335%)	998976 (7.72007%)	251870 (1.94645%)	0.27%
467_4_comb_R1	13293451	12155242 (91.4378%)	1006489 (7.57131%)	131720 (0.990864%)	0.26%
467_5_comb_R1	10286334	9380662 (91.1954%)	795266 (7.73129%)	110406 (1.07333%)	0.26%
467_6_comb_R1	12239282	11109586 (90.7699%)	969077 (7.91776%)	160619 (1.31232%)	0.26%

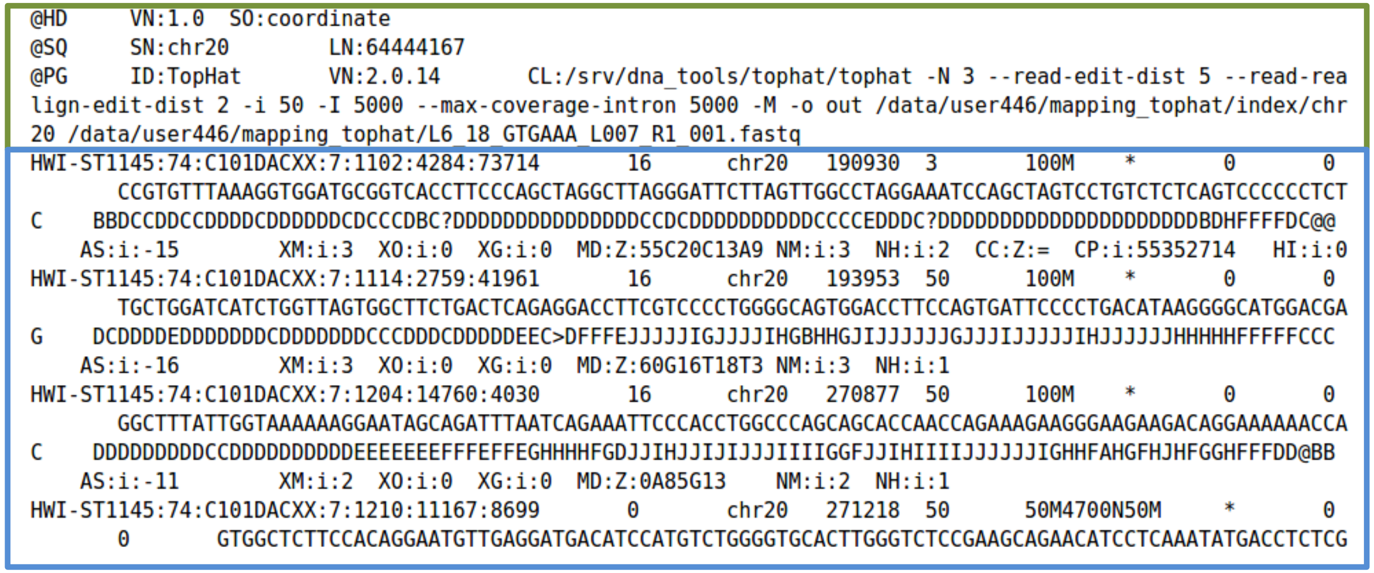
DATA ANALYSIS: ALIGNMENT

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.

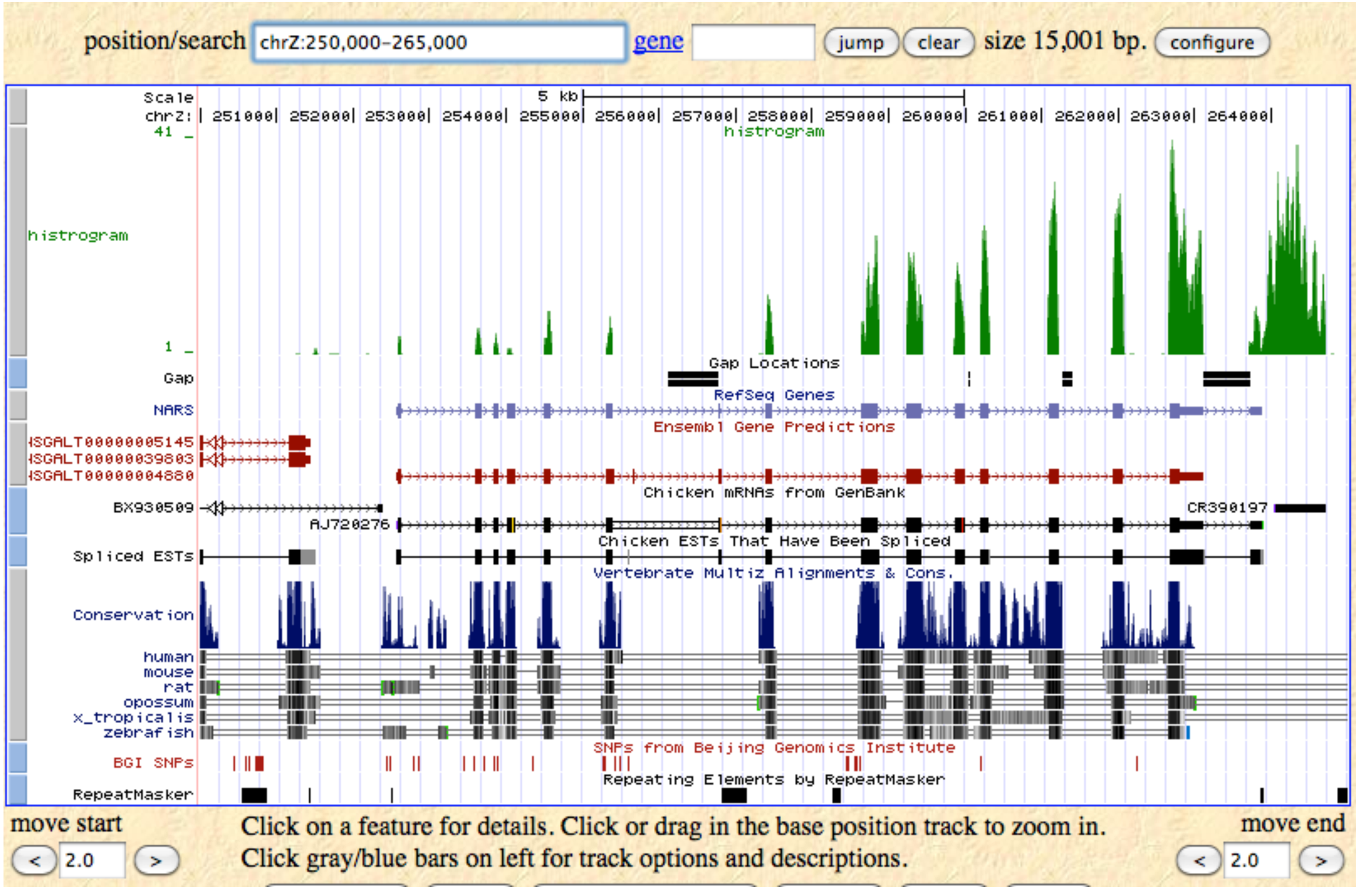


```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq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

DATA ANALYSIS: ALIGNMENT

BAM file visualization

- Genome Browser (UCSC)



position/search chrZ:250,000-265,000

size 15,001 bp.

Click on a feature for details. Click or drag in the base position track to zoom in.
Click gray/blue bars on left for track options and descriptions.

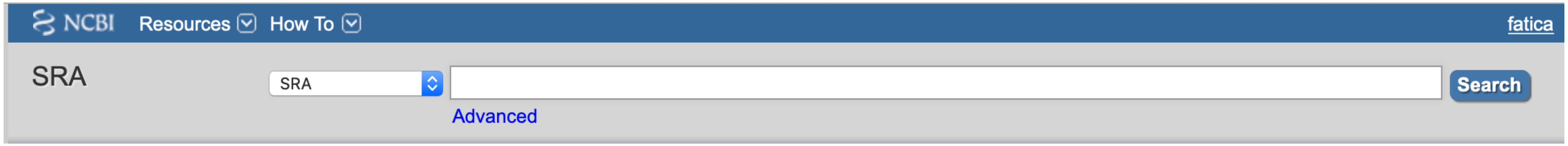
The Sequence Read Archive (SRA)

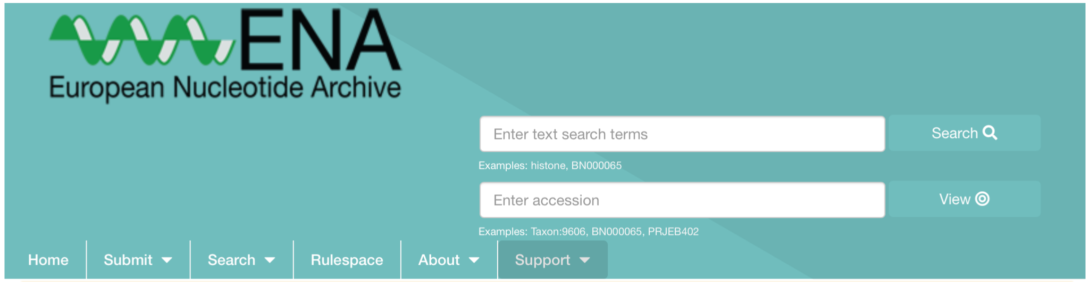
The SRA was established as a public repository for the next-generation sequence data and is operated by the International Nucleotide Sequence Database Collaboration (INSDC).

INSDC partners include the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ).

The SRA is accessible at http://www.ncbi.nlm.nih.gov/Traces/sra from NCBI, at http://www.ebi.ac.uk/ena from EBI and at http://trace.ddbj.nig.ac.jp from DDBJ.

total amount of data in 2019 was more than 14 petabytes (1 petabyte = 1 million gigabytes). For reference, one petabyte is equivalent to more than 4,000 digital photos per day for a lifetime.

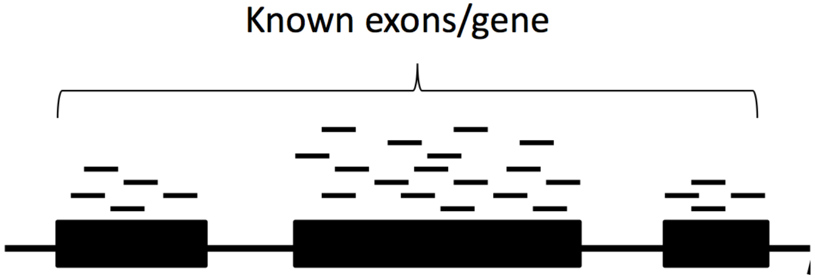



ENA
European Nucleotide Archive
Enter text search terms
Search
Examples: histone, BN000065
Enter accession
View
Examples: Taxon:9606, BN000065, PRJEB402
Home
Submit
Search
Rulespace
About
Support

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.

DATA ANALYSIS: ESTIMATING EXPESSION LEVELS

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 → 2 kb transcript	500 reads
gene B → 600 bp transcript	250 reads

The number of fragments sequenced are proportional to the **abundance** and **length** of the transcript.

↓

Normalize by transcript exon model **length** and **sequence depths** of the different samples.

RPKM (Reads per kilobase and million mappable reads):

Given 10 million mappable reads

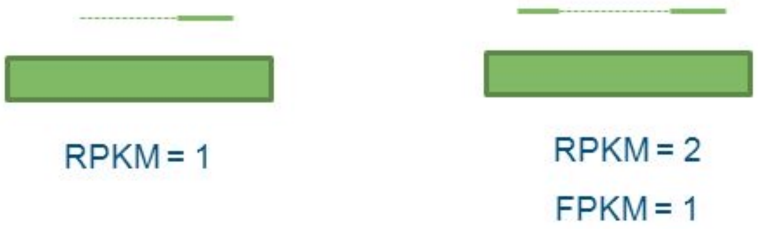
RPKM, Gene A: 500 reads / 2 / 10 = 25 RPKM
RPKM, Gene B: 250 reads / 0,6 / 10 = 42 RPKM

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Measures of gene expression: FPKM (paired-end) and RPKM (single-end)

- - 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

Measures of gene expression: TPM (transcripts per milion)

TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here's how you calculate TPM:

1. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
2. Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
3. Divide the RPK values by the "per million" scaling factor. This gives you TPM.

When calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.

When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.

Here's an example. If the TPM for gene A in Sample 1 is 3.33 and the TPM in sample B is 3.33, then I know that the exact same proportion of total reads mapped to gene A in both samples. This is because the sum of the TPMs in both samples always add up to the same number (so the denominator required to calculate the proportions is the same, regardless of what sample you are looking at.)

With RPKM or FPKM, the sum of normalized reads in each sample can be different. Thus, if the RPKM for gene A in Sample 1 is 3.33 and the RPKM in Sample 2 is 3.33, I would not know if the same proportion of reads in Sample 1 mapped to gene A as in Sample 2. This is because the denominator required to calculate the proportion could be different for the two samples.

RPKM vs TPM

RPKM

Gene	Counts_rep1	Counts_rep2	Counts_rep3
A 2kb	10	12	30
B 4kb	20	25	60
C 1kb	5	8	15
D 10kb	0	0	1

Total reads	35	45	106	
Tens of reads	3.5	4.5	10.6	for 4 gene samples we divide by to get the "million" scaling

Gene	RPM_1	RPM_2	RPM_3
A 2kb	2.86	2.67	2.83
B 4kb	5.71	5.56	5.66
C 1kb	1.43	1.78	1.42
D 10kb	0.00	0.00	0.09

Gene	RPKM_1	RPKM_2	RPKM_3
A 2kb	1.43	1.33	1.42
B 4kb	1.43	1.39	1.42
C 1kb	1.43	1.78	1.42
D 10kb	0.00	0.00	0.01

1. Count up the total reads in a sample and divide that number by 1,000,000 – this is our "per million" scaling factor.
2. Divide the read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
3. Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

TPM

Gene	Counts_rep1	Counts_rep2	Counts_rep3
A 2kb	10	12	30
B 4kb	20	25	60
C 1kb	5	8	15
D 10kb	0	0	1

Gene	RPK_rep1	RPK_rep2	RPK_rep3	RPK=Counts/kb
A 2kb	5	6	15	
B 4kb	5	6.25	15	
C 1kb	5	8	15	
D 10kb	0	0	0.1	

total RPKM	15	20.25	45.1	
Tens	1.5	2.025	4.51	for 4 gene samples we divide by 10 to get the "milion" scaling

Gene	TPM_1	TPM_2	TPM_3
A 2kb	3.33	2.96	3.33
B 4kb	3.33	3.09	3.33
C 1kb	3.33	3.95	3.33
D 10kb	0.00	0.00	0.02

1. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
2. Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
3. Divide the RPK values by the "per million" scaling factor. This gives you TPM.

DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

What is differential expression (DE) analysis?

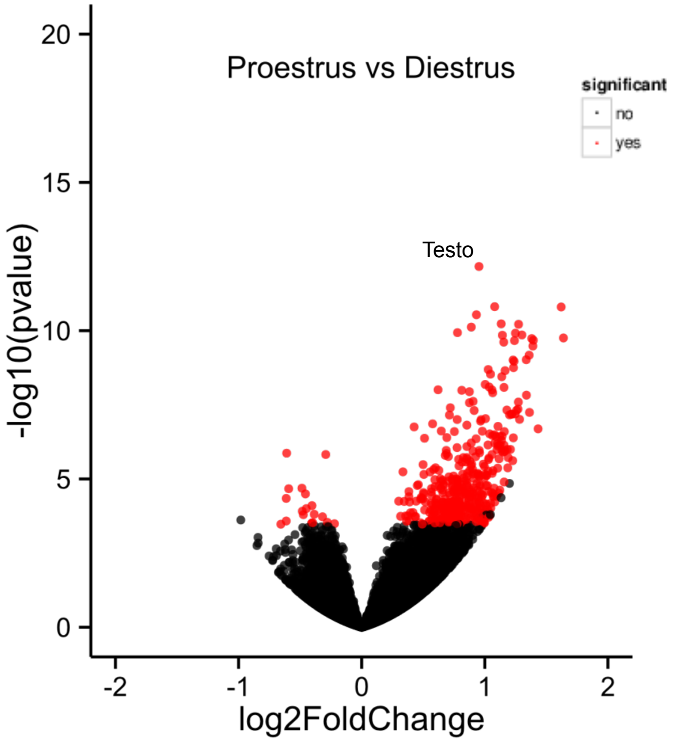
- DE analysis allows to find **genes** (or other genomic features like transcripts and exons) **that are expressed at significantly different levels between two groups of samples** (conditions): patients treated with drugs VS controls, healthy VS sick individuals , different tissues and different differentiation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

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

DATA ANALYSIS: DIFFERENTIAL EXPRESSION ANALYSIS

Fold Change (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.



STATISTICS

When carrying out a statistical significance test, one initially assumes the so-called "null hypothesis," according to which there is no difference between the groups regarding the parameter under consideration. According to the null hypothesis, the groups are equal to each other, and the observed difference is attributable to chance.

Obviously, the null hypothesis can be either true or false. Now you must decide: do you accept or reject the null hypothesis?

To decide, you must analyze your data with a statistical test. If the test "advises" you to reject the null hypothesis, then the observed difference is declared statistically significant. If, however, the test "advises" you to accept the null hypothesis, then the difference is statistically non-significant.

The significance level of a test can be chosen arbitrarily by the experimenter. However, a probability level of 0.05 (5%) or 0.01 (1%) is usually chosen. This probability (called the P-value) represents a quantitative estimate of the probability that the observed differences are due to chance.

More precisely, the P-value is "the probability of obtaining a result as extreme or more extreme than the one observed if the difference is entirely due to sampling variability alone, thereby assuming that the initial null hypothesis is true" (Signorelli).

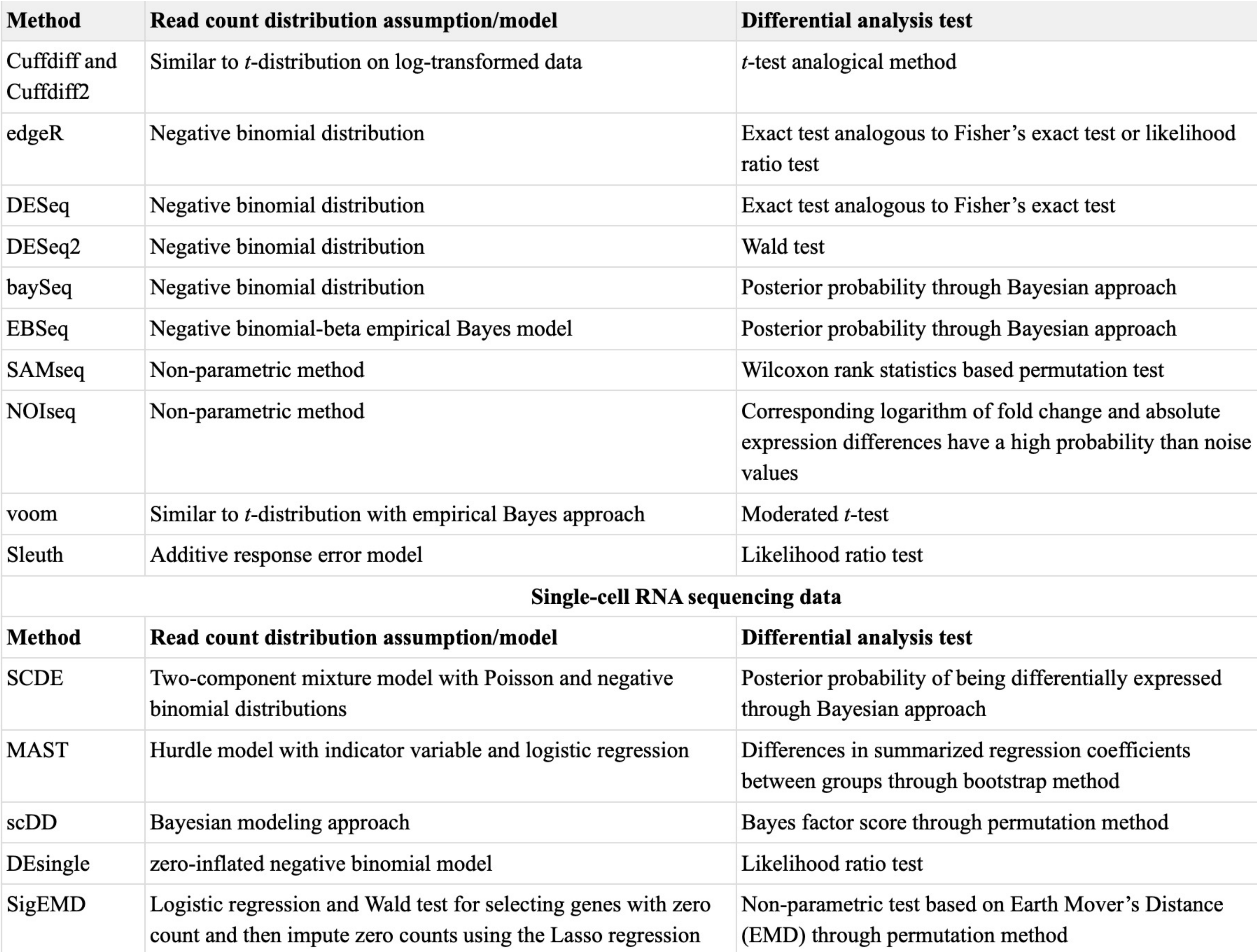
Note that P is a probability and can therefore only assume values between 0 and 1. A P-value that approaches 0 indicates a low probability that the observed difference can be attributed to chance.

Example: In a hypothetical experiment, a drug was shown to have an anti-hypertensive effect: in the treated subjects, systolic pressure decreased, on average, by 2 mm of Hg compared to untreated subjects, and this difference was found to be "statistically significant."

This does not automatically imply that the drug is a good anti-hypertensive; in fact, it is likely to be practically useless in therapy, as such a limited reduction (2 mm Hg) has no clinical interest.

This example highlights the important distinction between statistical significance (the difference is unlikely due to chance) and clinical significance (the difference is large enough to matter in real-world patient care).

STATISTICAL METHODS FOR RNA SEQUENCING DIFFERENTIAL ANALYSIS



Method	Read count distribution assumption/model	Differential analysis test
Cuffdiff and Cuffdiff2	Similar to *t*-distribution on log-transformed data	*t*-test analogical method
edgeR	Negative binomial distribution	Exact test analogous to Fisher's exact test or likelihood ratio test
DESeq	Negative binomial distribution	Exact test analogous to Fisher's exact test
DESeq2	Negative binomial distribution	Wald test
baySeq	Negative binomial distribution	Posterior probability through Bayesian approach
EBSeq	Negative binomial-beta empirical Bayes model	Posterior probability through Bayesian approach
SAMseq	Non-parametric method	Wilcoxon rank statistics based permutation test
NOIseq	Non-parametric method	Corresponding logarithm of fold change and absolute expression differences have a high probability than noise values
voom	Similar to *t*-distribution with empirical Bayes approach	Moderated *t*-test
Sleuth	Additive response error model	Likelihood ratio test

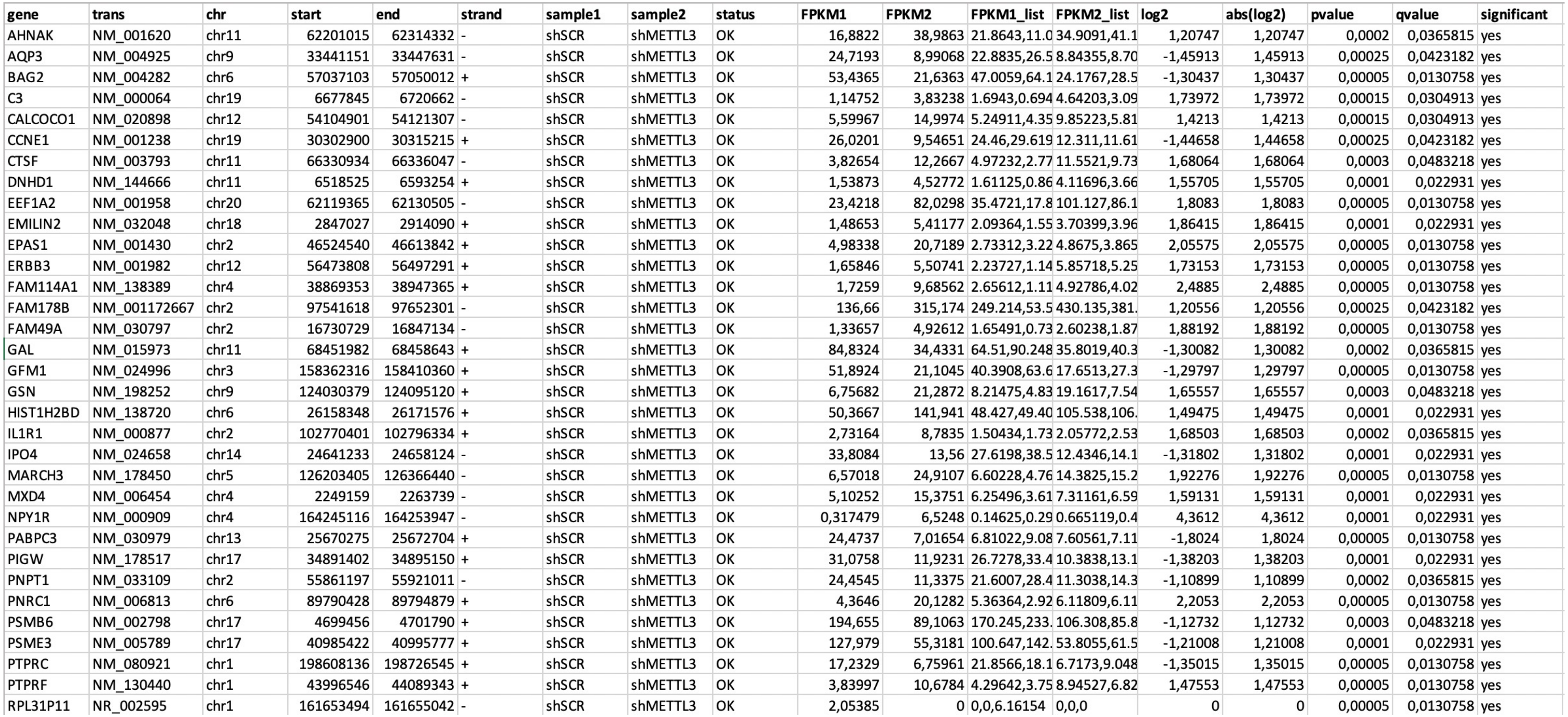
Single-cell RNA sequencing data

Method	Read count distribution assumption/model	Differential analysis test
SCDE	Two-component mixture model with Poisson and negative binomial distributions	Posterior probability of being differentially expressed through Bayesian approach
MAST	Hurdle model with indicator variable and logistic regression	Differences in summarized regression coefficients between groups through bootstrap method
scDD	Bayesian modeling approach	Bayes factor score through permutation method
DEsingle	zero-inflated negative binomial model	Likelihood ratio test
SigEMD	Logistic regression and Wald test for selecting genes with zero count and then impute zero counts using the Lasso regression	Non-parametric test based on Earth Mover's Distance (EMD) through permutation method

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

By transcripts

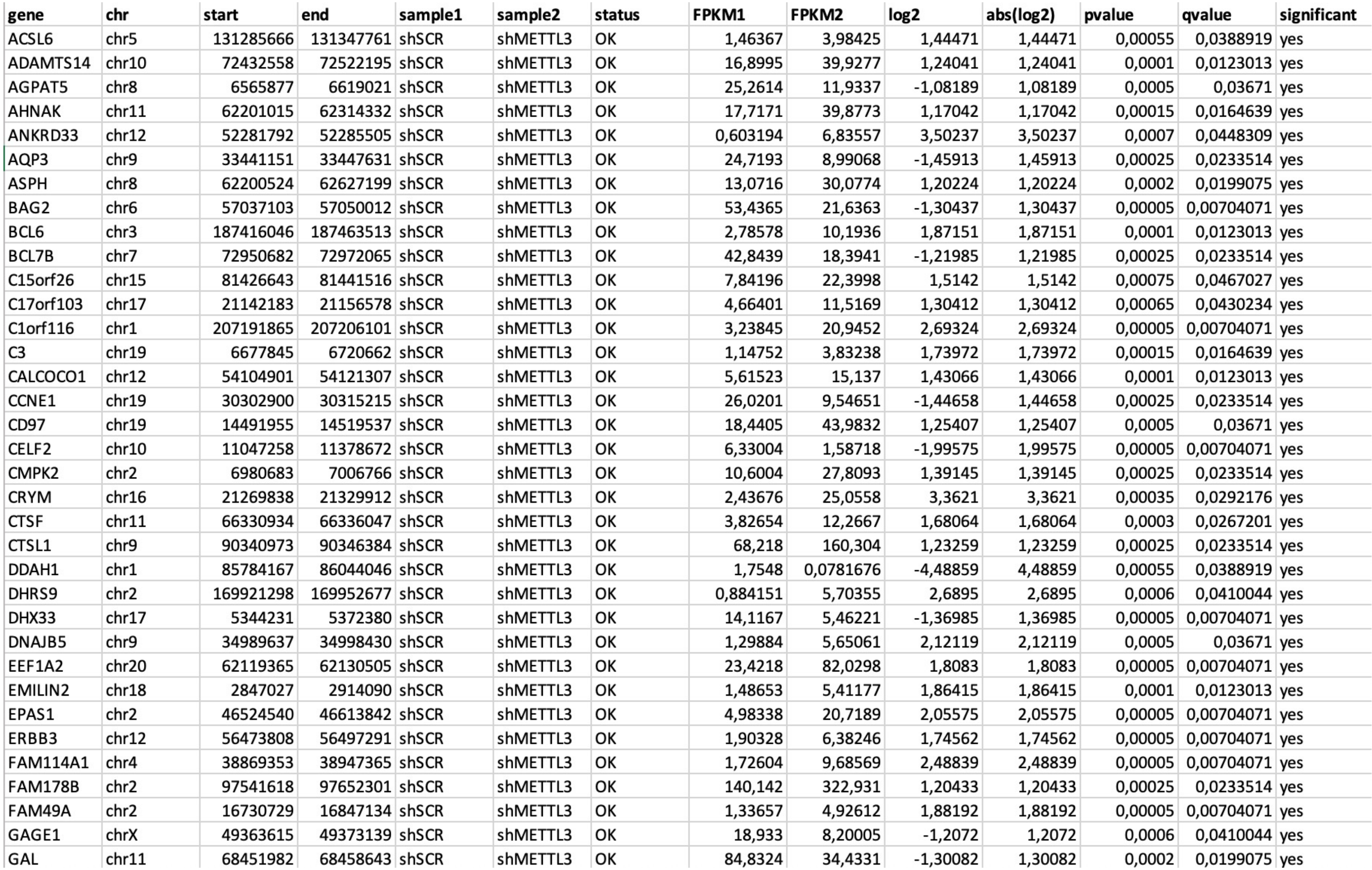
The q-value is an adjusted p-value, taking in to account the false discovery rate (FDR)



gene	trans	chr	start	end	strand	sample1	sample2	status	FPKM1	FPKM2	FPKM1_list	FPKM2_list	log2	abs(log2)	pvalue	qvalue	significant
AHNAK	NM_001620	chr11	62201015	62314332	-	shSCR	shMETTL3	OK	16,8822	38,9863	21.8643,11.0	34.9091,41.1	1,20747	1,20747	0,0002	0,0365815	yes
AQP3	NM_004925	chr9	33441151	33447631	-	shSCR	shMETTL3	OK	24,7193	8,99068	22.8835,26.5	8.84355,8.70	-1,45913	1,45913	0,00025	0,0423182	yes
BAG2	NM_004282	chr6	57037103	57050012	+	shSCR	shMETTL3	OK	53,4365	21,6363	47.0059,64.1	24.1767,28.5	-1,30437	1,30437	0,00005	0,0130758	yes
C3	NM_000064	chr19	6677845	6720662	-	shSCR	shMETTL3	OK	1,14752	3,83238	1.6943,0.694	4.64203,3.09	1,73972	1,73972	0,00015	0,0304913	yes
CALCOCO1	NM_020898	chr12	54104901	54121307	-	shSCR	shMETTL3	OK	5,59967	14,9974	5.24911,4.35	9.85223,5.81	1,4213	1,4213	0,00015	0,0304913	yes
CCNE1	NM_001238	chr19	30302900	30315215	+	shSCR	shMETTL3	OK	26,0201	9,54651	24.46,29.619	12.311,11.61	-1,44658	1,44658	0,00025	0,0423182	yes
CTSF	NM_003793	chr11	66330934	66336047	-	shSCR	shMETTL3	OK	3,82654	12,2667	4.97232,2.77	11.5521,9.73	1,68064	1,68064	0,0003	0,0483218	yes
DNHD1	NM_144666	chr11	6518525	6593254	+	shSCR	shMETTL3	OK	1,53873	4,52772	1.61125,0.86	4.11696,3.66	1,55705	1,55705	0,0001	0,022931	yes
EEF1A2	NM_001958	chr20	62119365	62130505	-	shSCR	shMETTL3	OK	23,4218	82,0298	35.4721,17.8	101.127,86.1	1,8083	1,8083	0,00005	0,0130758	yes
EMILIN2	NM_032048	chr18	2847027	2914090	+	shSCR	shMETTL3	OK	1,48653	5,41177	2.09364,1.55	3.70399,3.96	1,86415	1,86415	0,0001	0,022931	yes
EPAS1	NM_001430	chr2	46524540	46613842	+	shSCR	shMETTL3	OK	4,98338	20,7189	2.73312,3.22	4.8675,3.865	2,05575	2,05575	0,00005	0,0130758	yes
ERBB3	NM_001982	chr12	56473808	56497291	+	shSCR	shMETTL3	OK	1,65846	5,50741	2.23727,1.14	5.85718,5.25	1,73153	1,73153	0,00005	0,0130758	yes
FAM114A1	NM_138389	chr4	38869353	38947365	+	shSCR	shMETTL3	OK	1,7259	9,68562	2.65612,1.11	4.92786,4.02	2,4885	2,4885	0,00005	0,0130758	yes
FAM178B	NM_001172667	chr2	97541618	97652301	-	shSCR	shMETTL3	OK	136,66	315,174	249.214,53.5	430.135,381.	1,20556	1,20556	0,00025	0,0423182	yes
FAM49A	NM_030797	chr2	16730729	16847134	-	shSCR	shMETTL3	OK	1,33657	4,92612	1.65491,0.73	2.60238,1.87	1,88192	1,88192	0,00005	0,0130758	yes
GAL	NM_015973	chr11	68451982	68458643	+	shSCR	shMETTL3	OK	84,8324	34,4331	64.51,90.248	35.8019,40.3	-1,30082	1,30082	0,0002	0,0365815	yes
GFM1	NM_024996	chr3	158362316	158410360	+	shSCR	shMETTL3	OK	51,8924	21,1045	40.3908,63.6	17.6513,27.3	-1,29797	1,29797	0,00005	0,0130758	yes
GSN	NM_198252	chr9	124030379	124095120	+	shSCR	shMETTL3	OK	6,75682	21,2872	8.21475,4.83	19.1617,7.54	1,65557	1,65557	0,0003	0,0483218	yes
HIST1H2BD	NM_138720	chr6	26158348	26171576	+	shSCR	shMETTL3	OK	50,3667	141,941	48.427,49.40	105.538,106.	1,49475	1,49475	0,0001	0,022931	yes
IL1R1	NM_000877	chr2	102770401	102796334	+	shSCR	shMETTL3	OK	2,73164	8,7835	1.50434,1.73	2.05772,2.53	1,68503	1,68503	0,0002	0,0365815	yes
IPO4	NM_024658	chr14	24641233	24658124	-	shSCR	shMETTL3	OK	33,8084	13,56	27.6198,38.5	12.4346,14.1	-1,31802	1,31802	0,0001	0,022931	yes
MARCH3	NM_178450	chr5	126203405	126366440	-	shSCR	shMETTL3	OK	6,57018	24,9107	6.60228,4.76	14.3825,15.2	1,92276	1,92276	0,00005	0,0130758	yes
MXD4	NM_006454	chr4	2249159	2263739	-	shSCR	shMETTL3	OK	5,10252	15,3751	6.25496,3.61	7.31161,6.59	1,59131	1,59131	0,0001	0,022931	yes
NPY1R	NM_000909	chr4	164245116	164253947	-	shSCR	shMETTL3	OK	0,317479	6,5248	0.14625,0.29	0.665119,0.4	4,3612	4,3612	0,0001	0,022931	yes
PABPC3	NM_030979	chr13	25670275	25672704	+	shSCR	shMETTL3	OK	24,4737	7,01654	6.81022,9.08	7.60561,7.11	-1,8024	1,8024	0,00005	0,0130758	yes
PIGW	NM_178517	chr17	34891402	34895150	+	shSCR	shMETTL3	OK	31,0758	11,9231	26.7278,33.4	10.3838,13.1	-1,38203	1,38203	0,0001	0,022931	yes
PNPT1	NM_033109	chr2	55861197	55921011	-	shSCR	shMETTL3	OK	24,4545	11,3375	21.6007,28.4	11.3038,14.3	-1,10899	1,10899	0,0002	0,0365815	yes
PNRC1	NM_006813	chr6	89790428	89794879	+	shSCR	shMETTL3	OK	4,3646	20,1282	5.36364,2.92	6.11809,6.11	2,2053	2,2053	0,00005	0,0130758	yes
PSMB6	NM_002798	chr17	4699456	4701790	+	shSCR	shMETTL3	OK	194,655	89,1063	170.245,233.	106.308,85.8	-1,12732	1,12732	0,0003	0,0483218	yes
PSME3	NM_005789	chr17	40985422	40995777	+	shSCR	shMETTL3	OK	127,979	55,3181	100.647,142.	53.8055,61.5	-1,21008	1,21008	0,0001	0,022931	yes
PTPRC	NM_080921	chr1	198608136	198726545	+	shSCR	shMETTL3	OK	17,2329	6,75961	21.8566,18.1	6.7173,9.048	-1,35015	1,35015	0,00005	0,0130758	yes
PTPRF	NM_130440	chr1	43996546	44089343	+	shSCR	shMETTL3	OK	3,83997	10,6784	4.29642,3.75	8.94527,6.82	1,47553	1,47553	0,00005	0,0130758	yes
RPL31P11	NR_002595	chr1	161653494	161655042	-	shSCR	shMETTL3	OK	2,05385	0	0,0,6.16154	0,0,0	0	0	0,00005	0,0130758	yes

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

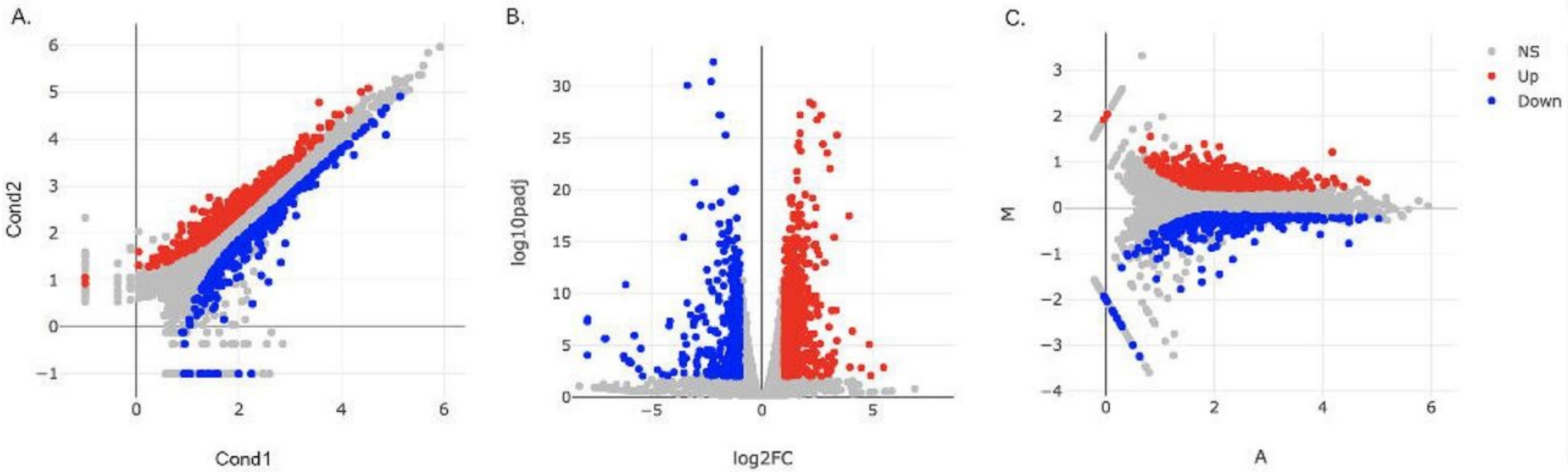
By genes



gene	chr	start	end	sample1	sample2	status	FPKM1	FPKM2	log2	abs(log2)	pvalue	qvalue	significant
ACSL6	chr5	131285666	131347761	shSCR	shMETTL3	OK	1,46367	3,98425	1,44471	1,44471	0,00055	0,0388919	yes
ADAMTS14	chr10	72432558	72522195	shSCR	shMETTL3	OK	16,8995	39,9277	1,24041	1,24041	0,0001	0,0123013	yes
AGPAT5	chr8	6565877	6619021	shSCR	shMETTL3	OK	25,2614	11,9337	-1,08189	1,08189	0,0005	0,03671	yes
AHNAK	chr11	62201015	62314332	shSCR	shMETTL3	OK	17,7171	39,8773	1,17042	1,17042	0,00015	0,0164639	yes
ANKRD33	chr12	52281792	52285505	shSCR	shMETTL3	OK	0,603194	6,83557	3,50237	3,50237	0,0007	0,0448309	yes
AQP3	chr9	33441151	33447631	shSCR	shMETTL3	OK	24,7193	8,99068	-1,45913	1,45913	0,00025	0,0233514	yes
ASPH	chr8	62200524	62627199	shSCR	shMETTL3	OK	13,0716	30,0774	1,20224	1,20224	0,0002	0,0199075	yes
BAG2	chr6	57037103	57050012	shSCR	shMETTL3	OK	53,4365	21,6363	-1,30437	1,30437	0,00005	0,00704071	yes
BCL6	chr3	187416046	187463513	shSCR	shMETTL3	OK	2,78578	10,1936	1,87151	1,87151	0,0001	0,0123013	yes
BCL7B	chr7	72950682	72972065	shSCR	shMETTL3	OK	42,8439	18,3941	-1,21985	1,21985	0,00025	0,0233514	yes
C15orf26	chr15	81426643	81441516	shSCR	shMETTL3	OK	7,84196	22,3998	1,5142	1,5142	0,00075	0,0467027	yes
C17orf103	chr17	21142183	21156578	shSCR	shMETTL3	OK	4,66401	11,5169	1,30412	1,30412	0,00065	0,0430234	yes
C1orf116	chr1	207191865	207206101	shSCR	shMETTL3	OK	3,23845	20,9452	2,69324	2,69324	0,00005	0,00704071	yes
C3	chr19	6677845	6720662	shSCR	shMETTL3	OK	1,14752	3,83238	1,73972	1,73972	0,00015	0,0164639	yes
CALCOCO1	chr12	54104901	54121307	shSCR	shMETTL3	OK	5,61523	15,137	1,43066	1,43066	0,0001	0,0123013	yes
CCNE1	chr19	30302900	30315215	shSCR	shMETTL3	OK	26,0201	9,54651	-1,44658	1,44658	0,00025	0,0233514	yes
CD97	chr19	14491955	14519537	shSCR	shMETTL3	OK	18,4405	43,9832	1,25407	1,25407	0,0005	0,03671	yes
CELF2	chr10	11047258	11378672	shSCR	shMETTL3	OK	6,33004	1,58718	-1,99575	1,99575	0,00005	0,00704071	yes
CMPK2	chr2	6980683	7006766	shSCR	shMETTL3	OK	10,6004	27,8093	1,39145	1,39145	0,00025	0,0233514	yes
CRYM	chr16	21269838	21329912	shSCR	shMETTL3	OK	2,43676	25,0558	3,3621	3,3621	0,00035	0,0292176	yes
CTSF	chr11	66330934	66336047	shSCR	shMETTL3	OK	3,82654	12,2667	1,68064	1,68064	0,0003	0,0267201	yes
CTSL1	chr9	90340973	90346384	shSCR	shMETTL3	OK	68,218	160,304	1,23259	1,23259	0,00025	0,0233514	yes
DDAH1	chr1	85784167	86044046	shSCR	shMETTL3	OK	1,7548	0,0781676	-4,48859	4,48859	0,00055	0,0388919	yes
DHRS9	chr2	169921298	169952677	shSCR	shMETTL3	OK	0,884151	5,70355	2,6895	2,6895	0,0006	0,0410044	yes
DHX33	chr17	5344231	5372380	shSCR	shMETTL3	OK	14,1167	5,46221	-1,36985	1,36985	0,00005	0,00704071	yes
DNAJB5	chr9	34989637	34998430	shSCR	shMETTL3	OK	1,29884	5,65061	2,12119	2,12119	0,0005	0,03671	yes
EEF1A2	chr20	62119365	62130505	shSCR	shMETTL3	OK	23,4218	82,0298	1,8083	1,8083	0,00005	0,00704071	yes
EMILIN2	chr18	2847027	2914090	shSCR	shMETTL3	OK	1,48653	5,41177	1,86415	1,86415	0,0001	0,0123013	yes
EPAS1	chr2	46524540	46613842	shSCR	shMETTL3	OK	4,98338	20,7189	2,05575	2,05575	0,00005	0,00704071	yes
ERBB3	chr12	56473808	56497291	shSCR	shMETTL3	OK	1,90328	6,38246	1,74562	1,74562	0,00005	0,00704071	yes
FAM114A1	chr4	38869353	38947365	shSCR	shMETTL3	OK	1,72604	9,68569	2,48839	2,48839	0,00005	0,00704071	yes
FAM178B	chr2	97541618	97652301	shSCR	shMETTL3	OK	140,142	322,931	1,20433	1,20433	0,00025	0,0233514	yes
FAM49A	chr2	16730729	16847134	shSCR	shMETTL3	OK	1,33657	4,92612	1,88192	1,88192	0,00005	0,00704071	yes
GAGE1	chrX	49363615	49373139	shSCR	shMETTL3	OK	18,933	8,20005	-1,2072	1,2072	0,0006	0,0410044	yes
GAL	chr11	68451982	68458643	shSCR	shMETTL3	OK	84,8324	34,4331	-1,30082	1,30082	0,0002	0,0199075	yes

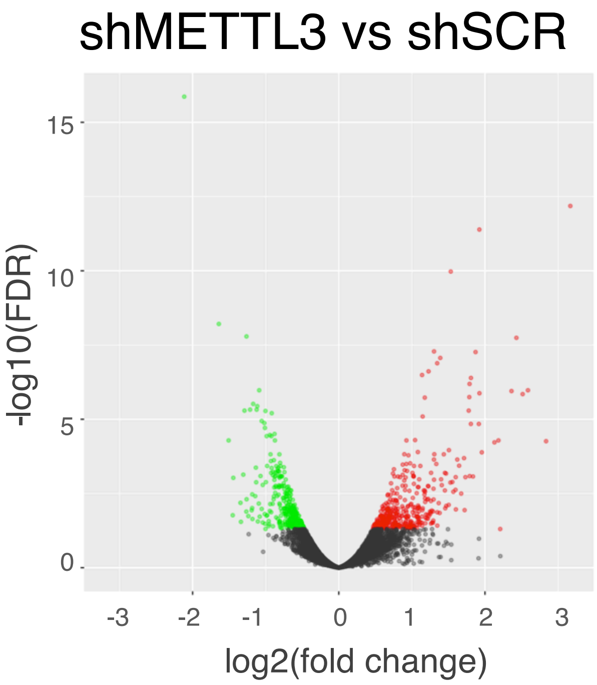
DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

A scatterplot is an effective visualization tool that plots read count distributions across all genes and samples. Specifically, it represents each row (gene) as a point in each scatterplot.



a Scatter plot. **b** Volcano plot. **c** MA plot DE genes are located in each plot while padj < 0.01 and |log2foldChange|>1.

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION



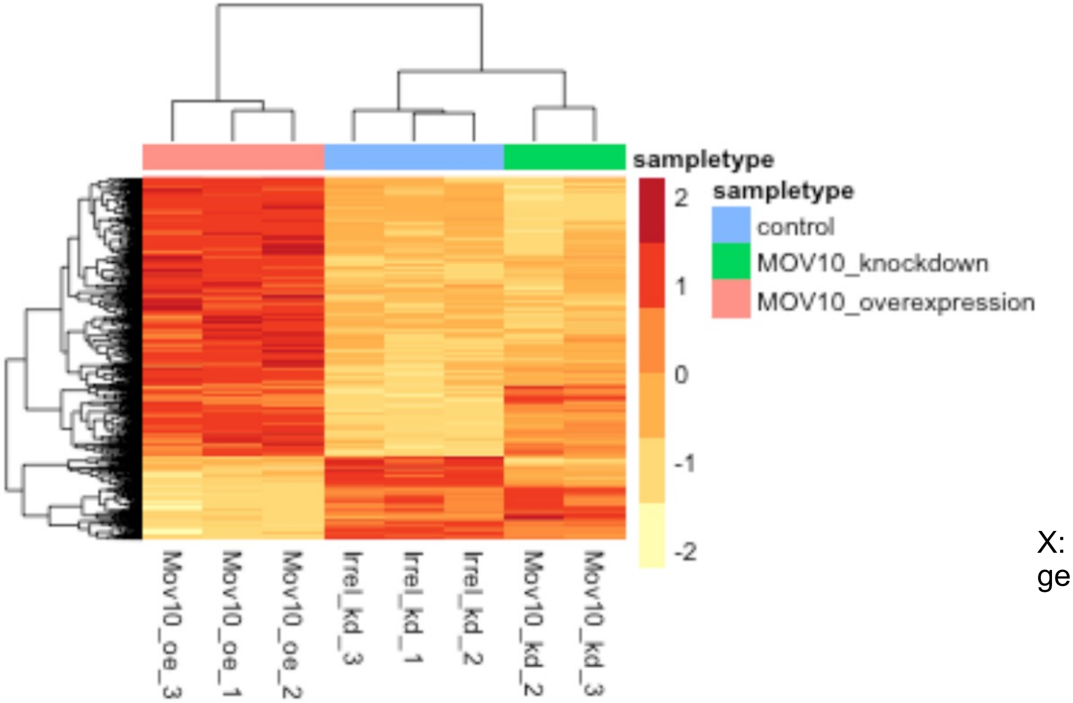
A commonly used plot is a **volcano plot**; in which you have the log of FDR are plotted on the y-axis and log2 fold change values on the x-axis.

The **false discovery rate (FDR)** is a statistical approach used in multiple hypothesis testing to correct for multiple comparisons. It is typically used in high-throughput experiments in order to correct for random events that falsely appear significant.

The *q*-value can be interpreted as the FDR: the proportion of false positives among all positive results. Just as the **p-value** gives the expected false positive rate obtained by rejecting the null hypothesis for any result with an equal or smaller p-value, the **q-value** gives the expected FDR obtained by rejecting the null hypothesis for any result with an equal or smaller q-value.

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

we could also extract the normalized values of *all* the significant genes and plot a **heatmap** of their expression



$$Z = \frac{X - \mu}{\delta}$$

X = score

μ = mean

δ = SD

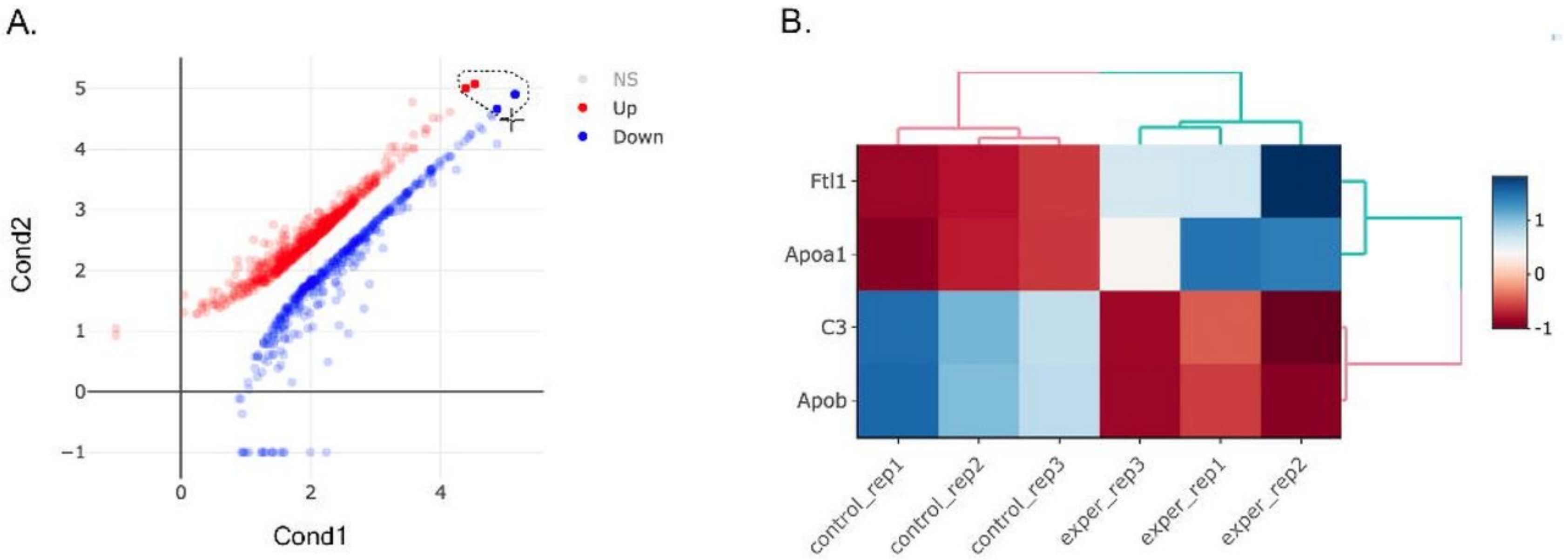
X: normalized read count for a specific gene within an individual sample.

In this heatmap Z-scores are calculated for each row (each gene) and these are plotted instead of the normalized expression values; this ensures that the expression patterns/trends that we want to visualize are not overwhelmed by the expression values.

Z-scores are computed on a gene-by-gene basis by subtracting the mean and then dividing by the standard deviation. The Z-scores are computed ***after the clustering****, so that it only affects the graphical aesthetics and the colour visualization is improved.*

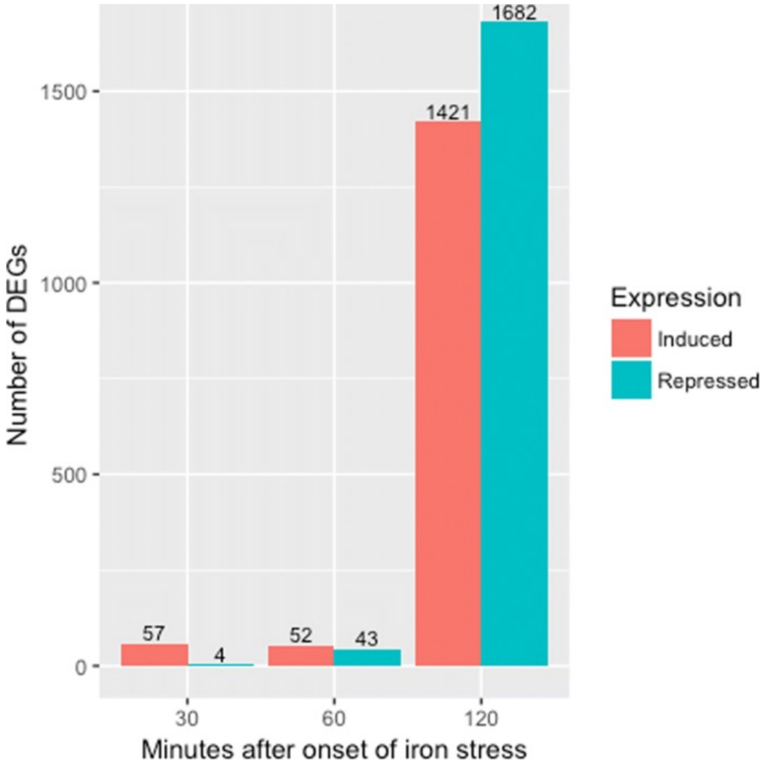
DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

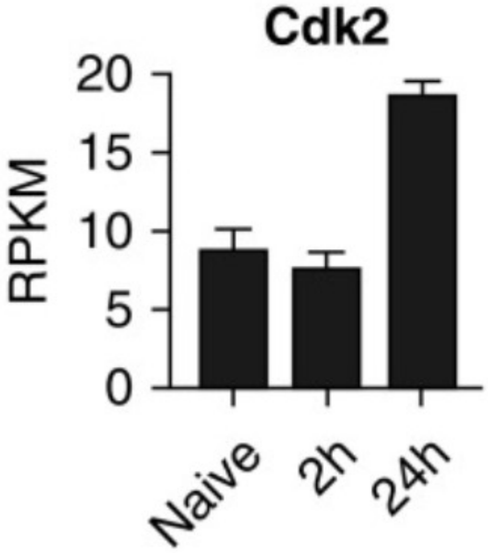
To explore the underlying data for any set of regions in a plot, we can draw heatmaps for any selected region from any main plot. Figure shows the heatmap for the selected genes. Conversely, in any heatmap the users can select a subset of regions (such as based on similar expression pattern) for downstream analysis such as gene ontology, disease and pathway analysis.



DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Histograms are useful to visualize gene expression responses across time points or to visualize individual gene analysis

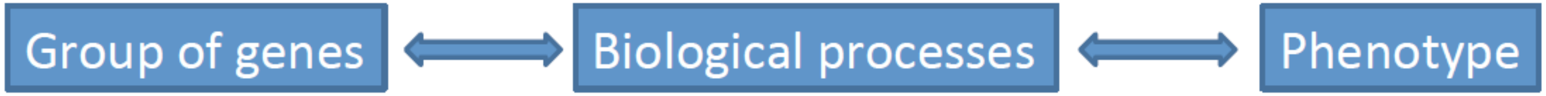




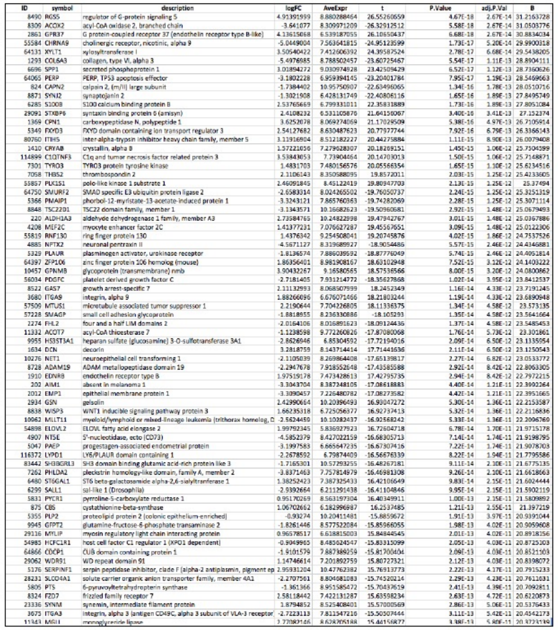
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

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:



Group of genes ↔ Biological processes ↔ Phenotype



ID	symbol	description	logFC	AveExpr	t	P.Value	adj.P.Val	B
8490	RGS5	regulator of G-protein signaling 5	4.91391999	8.880288464	26.55260659	4.67E-18	2.67E-14	31.2163728
8309	ACOX2	acyl-CoA oxidase 2, branched chain	-3.641077	8.309971209	-26.32912512	5.58E-18	2.67E-14	31.0503776
2861	GPR37	G protein-coupled receptor 37 (endothelin receptor type B-like)	4.13615068	6.539187055	26.10650437	6.68E-18	2.67E-14	30.8834034
55584	CHRNA9	cholinergic receptor, nicotinic, alpha 9	-5.0449004	7.563641815	-24.95123599	1.73E-17	5.20E-14	29.9900318
64131	XYLT1	xylosyltransferase I	3.50540422	7.412606392	24.39587524	2.78E-17	6.68E-14	29.5438205
1293	COL6A3	collagen, type VI, alpha 3	-5.4976985	8.788502457	-23.60725467	5.54E-17	1.11E-13	28.8904111
6696	SPP1	secreted phosphoprotein 1	3.01894272	9.030974928	23.42509429	6.52E-17	1.12E-13	28.7360626
64065	PERP	PERP, TP53 apoptosis effector	-3.1802228	6.959394145	-23.20401784	7.95E-17	1.19E-13	28.5469663
824	CAPN2	calpain 2, (m/II) large subunit	-1.7384402	10.95750907	-22.63496065	1.34E-16	1.78E-13	28.0510716
8871	SYNJ2	synaptojanin 2	-1.3021908	6.428131749	-22.40806116	1.65E-16	1.89E-13	27.8495749
6285	S100B	S100 calcium binding protein B	2.53765669	6.799331011	22.35831889	1.73E-16	1.89E-13	27.8051084
29091	STXBP6	syntaxin binding protein 6 (amisyn)	2.4108232	6.531105876	21.64156067	3.40E-16	3.41E-13	27.152374
1369	CPN1	carboxypeptidase N, polypeptide 1	3.6252078	8.069274069	21.17029509	5.38E-16	4.97E-13	26.7105914
5349	FXYD3	FXYD domain containing ion transport regulator 3	2.54127682	8.630487623	20.77977744	7.92E-16	6.79E-13	26.3366143
80760	ITIH5	inter-alpha-trypsin inhibitor heavy chain family, member 5	3.11916904	8.512182227	20.44275884	1.11E-15	8.90E-13	26.0079408
1410	CRYAB	crystallin, alpha B	1.57221056	7.279628307	20.18269151	1.45E-15	1.06E-12	25.7504399
114899	C1QTNF3	C1q and tumor necrosis factor related protein 3	3.53843053	7.73904464	20.14703013	1.50E-15	1.06E-12	25.7148871
7301	TYRO3	TYRO3 protein tyrosine kinase	1.4831703	7.480156576	20.05566354	1.65E-15	1.10E-12	25.6234516
7058	THBS2	thrombospondin 2	2.1106143	8.350588095	19.8572011	2.03E-15	1.25E-12	25.4233605
55857	PLK1S1	polo-like kinase 1 substrate 1	2.46091845	8.45122419	19.80947703	2.13E-15	1.25E-12	25.37494
64750	SMURF2	SMAD specific E3 ubiquitin protein ligase 2	-2.6583314	8.024265502	-19.76050737	2.24E-15	1.25E-12	25.3251319
5366	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1	-3.3243121	7.865760363	-19.74282069	2.28E-15	1.25E-12	25.3071114
8848	TSC22D1	TSC22 domain family, member 1	-3.1343571	10.16682623	-19.50960681	2.92E-15	1.48E-12	25.0679493
220	ALDH1A3	aldehyde dehydrogenase 1 family, member A3	2.73584765	10.24822998	19.47942767	3.01E-15	1.48E-12	25.0367886
4208	MEF2C	myocyte enhancer factor 2C	1.41377231	7.076627287	19.45567651	3.09E-15	1.48E-12	25.0122306
55819	RNF130	ring finger protein 130	1.4376342	9.254508041	19.20745876	4.02E-15	1.86E-12	24.7537526
4885	NPTX2	neuronal pentraxin II	-4.5671127	8.319689927	-18.9054486	5.57E-15	2.46E-12	24.4346881
5329	PLAUR	plasminogen activator, urokinase receptor	-1.8136574	7.886039592	-18.87776049	5.74E-15	2.46E-12	24.4051814
64397	ZFP106	zinc finger protein 106 homolog (mouse)	1.86356401	8.981908167	18.63102948	7.52E-15	3.12E-12	24.1403222
10457	GPNMB	glycoprotein (transmembrane) nmb	3.90432267	9.16580565	18.57536566	8.00E-15	3.20E-12	24.0800862
56034	PDGFC	platelet derived growth factor C	-2.7181405	7.931214772	-18.35627868	1.02E-14	3.95E-12	23.8412537
8522	GAS7	growth arrest-specific 7	2.11132993	8.068507999	18.2452349	1.16E-14	4.33E-12	23.7191245
3680	ITGA9	integrin, alpha 9	1.88266096	6.676071466	18.21803244	1.19E-14	4.33E-12	23.6890948
57509	MTUS1	microtubule associated tumor suppressor 1	2.2190644	7.704226805	18.11336375	1.34E-14	4.58E-12	23.573135
57228	SMAGP	small cell adhesion glycoprotein	-1.8818955	8.236330886	-18.105293	1.35E-14	4.58E-12	23.5641664
2274	FHL2	four and a half LIM domains 2	-2.0164106	8.016891623	-18.09124436	1.37E-14	4.58E-12	23.5485453
11332	ACOT7	acyl-CoA thioesterase 7	-1.1238598	9.772260826	-17.87080068	1.76E-14	5.73E-12	23.301861
9955	HS3ST3A1	heparan sulfate (glucosamine) 3-O-sulfotransferase 3A1	-2.8626946	6.85304592	-17.72194016	2.09E-14	6.50E-12	23.1335954
1634	DCN	decorin	3.2818759	8.143714414	17.71441636	2.11E-14	6.50E-12	23.1250543
10276	NET1	neuroepithelial cell transforming 1	-2.1105039	8.269864408	-17.65139817	2.27E-14	6.82E-12	23.0533772
8728	ADAM19	ADAM metallopeptidase domain 19	-2.2947678	7.918552648	-17.43585588	2.92E-14	8.42E-12	22.8063305
1910	EDNRB	endothelin receptor type B	1.97519178	7.473428613	17.42795735	2.94E-14	8.42E-12	22.7972215
202	AIM1	absent in melanoma 1	-3.3043704	8.387248105	-17.08618883	4.40E-14	1.21E-11	22.3992264
2012	EMP1	epithelial membrane protein 1	-3.3090457	7.226480782	-17.08273582	4.42E-14	1.21E-11	22.3951665
2934	GSN	gelsolin	2.42990664	10.20396493	16.93047272	5.30E-14	1.36E-11	22.2153587
8838	WISP3	WNT1 inducible signaling pathway protein 3	1.66235318	6.725056377	16.92737413	5.32E-14	1.36E-11	22.2116836
10962	MLLT11	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, D	-2.5624450	10.10282437	-16.92568242	5.33E-14	1.36E-11	22.2096760
54898	ELOVL2	ELOVL fatty acid elongase 2	1.99792345	5.836927923	16.72604718	6.78E-14	1.70E-11	21.9715178
4907	NT5E	5'-nucleotidase, ecto (CD73)	-4.5852379	8.427022159	-16.68305713	7.14E-14	1.74E-11	21.9198795
5047	PAEP	progestagen-associated endometrial protein	-3.1997583	6.665647235	-16.67307416	7.22E-14	1.74E-11	21.9078703
116372	LYPD1	LY6/PLAUR domain containing 1	-2.2678592	6.79874409	-16.56676339	8.22E-14	1.94E-11	21.7795586
83442	SH3BGRL3	SH3 domain binding glutamic acid-rich protein like 3	-1.7165301	10.57293255	-16.48267181	9.11E-14	2.10E-11	21.6775135
7262	PHLDA2	pleckstrin homology-like domain, family A, member 2	-3.8371463	7.757814979	-16.46981308	9.26E-14	2.10E-11	21.6618663
6480	ST6GAL1	ST6 beta-galactosamide alpha-2,6-sialyltranferase 1	1.38252423	7.387325433	16.42106649	9.83E-14	2.15E-11	21.6024444
6299	SALL1	sal-like 1 (Drosophila)	-2.9392664	6.211291438	-16.41104846	9.95E-14	2.15E-11	21.5902119
5831	PYCR1	pyrroline-5-carboxylate reductase 1	0.95170269	8.563197304	16.40349911	1.00E-13	2.15E-11	21.5809892
875	CBS	cystathionine-beta-synthase	1.06702662	6.182996987	16.2537485	1.21E-13	2.55E-11	21.397219
5355	PLP2	proteolipid protein 2 (colonic epithelium-enriched)	-0.93274	10.20411481	-15.8859672	1.91E-13	3.97E-11	20.9391044
9945	GFPT2	glutamine-fructose-6-phosphate transaminase 2	-1.8261446	8.577522084	-15.85966055	1.98E-13	4.02E-11	20.9059608
29116	MYLIP	myosin regulatory light chain interacting protein	0.96578517	6.618815003	15.84844545	2.01E-13	4.02E-11	20.8918156
54985	HCFC1R1	host cell factor C1 regulator 1 (XPO1 dependent)	-0.9049965	8.485624547	-15.83315099	2.05E-13	4.03E-11	20.8725103
64866	CDCP1	CUB domain containing protein 1	-1.9101579	7.887389259	-15.81700404	2.09E-13	4.03E-11	20.8521103
29062	WDR91	WD repeat domain 91	1.14746614	7.201892759	15.80727321	2.12E-13	4.03E-11	20.8398072
5176	SERPINF1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment ep	2.95931204	10.47762382	15.76913773	2.22E-13	4.17E-11	20.7915233
28231	SLCO4A1	solute carrier organic anion transporter family, member 4A1	-2.2707561	8.804081083	-15.74520214	2.29E-13	4.23E-11	20.7611631
5805	PTS	6-pyruvoyltetrahydropterin synthase	-1.361366	8.951585472	-15.70437619	2.41E-13	4.39E-11	20.7092811
8324	FZD7	frizzled family receptor 7	2.58118442	7.422131287	15.63598234	2.63E-13	4.72E-11	20.6220873
23336	SYNM	synemin, intermediate filament protein	1.8794852	8.525408401	15.57000569	2.86E-13	5.06E-11	20.5376433
3675	ITGA3	integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	-2.7223113	7.811547216	-15.50507444	3.11E-13	5.42E-11	20.4542173
11343	MGLL	monoglyceride lipase	2.77082146	8.628705188	15.44156877	3.38E-13	5.80E-11	20.3723139

And now what ?

We can perform GO or Pathway analysis directly on the results of differential expression analysis or on a subset of selected genes from any of the plots described above.

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.

The Gene Ontology (GO)

The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects:

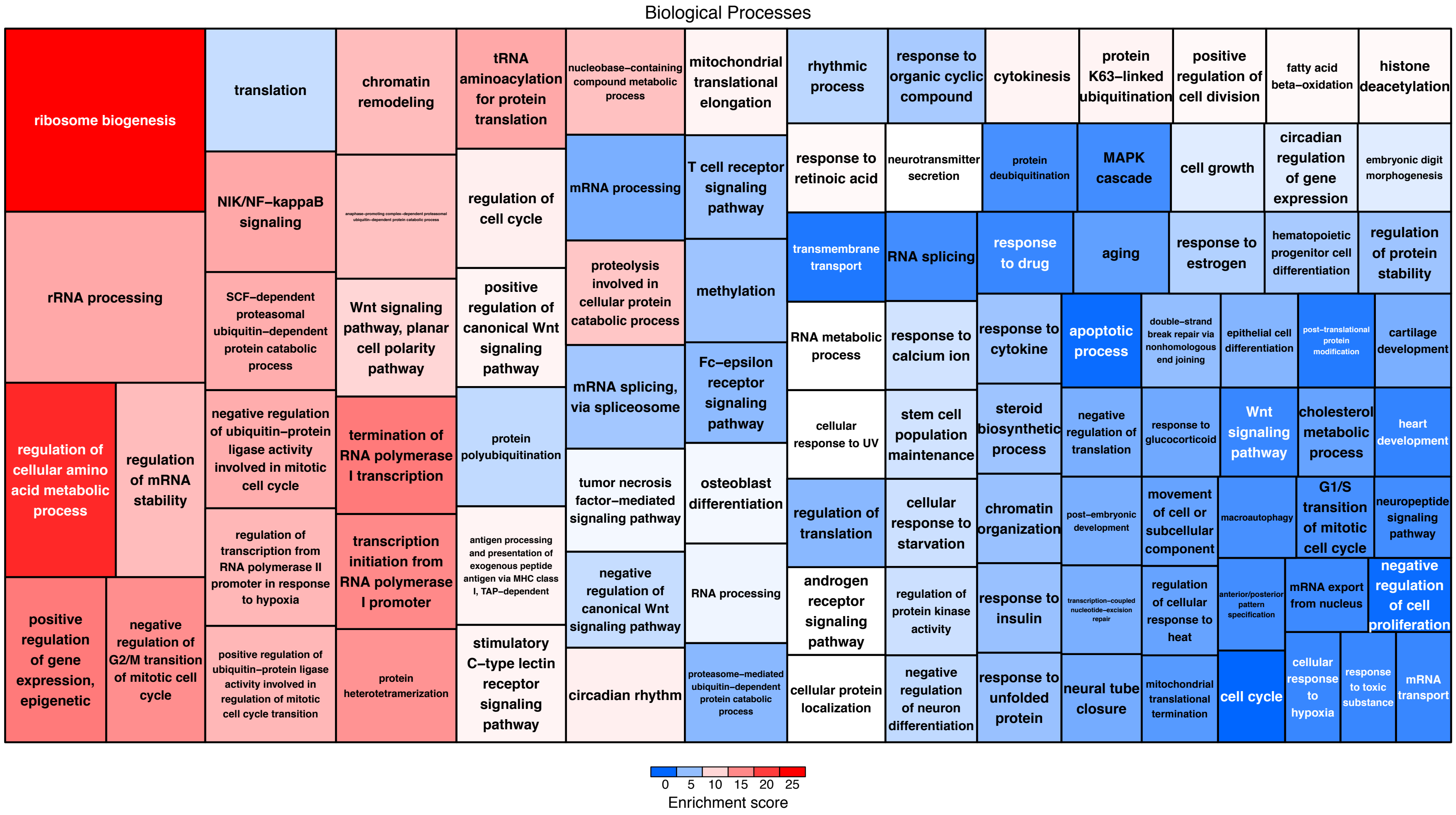
Molecular Function: molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as “catalysis” or “transport”. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place.

Cellular Component: The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (*e.g.*, *mitochondrion*), or stable macromolecular complexes of which they are parts (*e.g.*, the *ribosome*). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.

Biological Process: The larger processes, or ‘biological programs’ accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*.

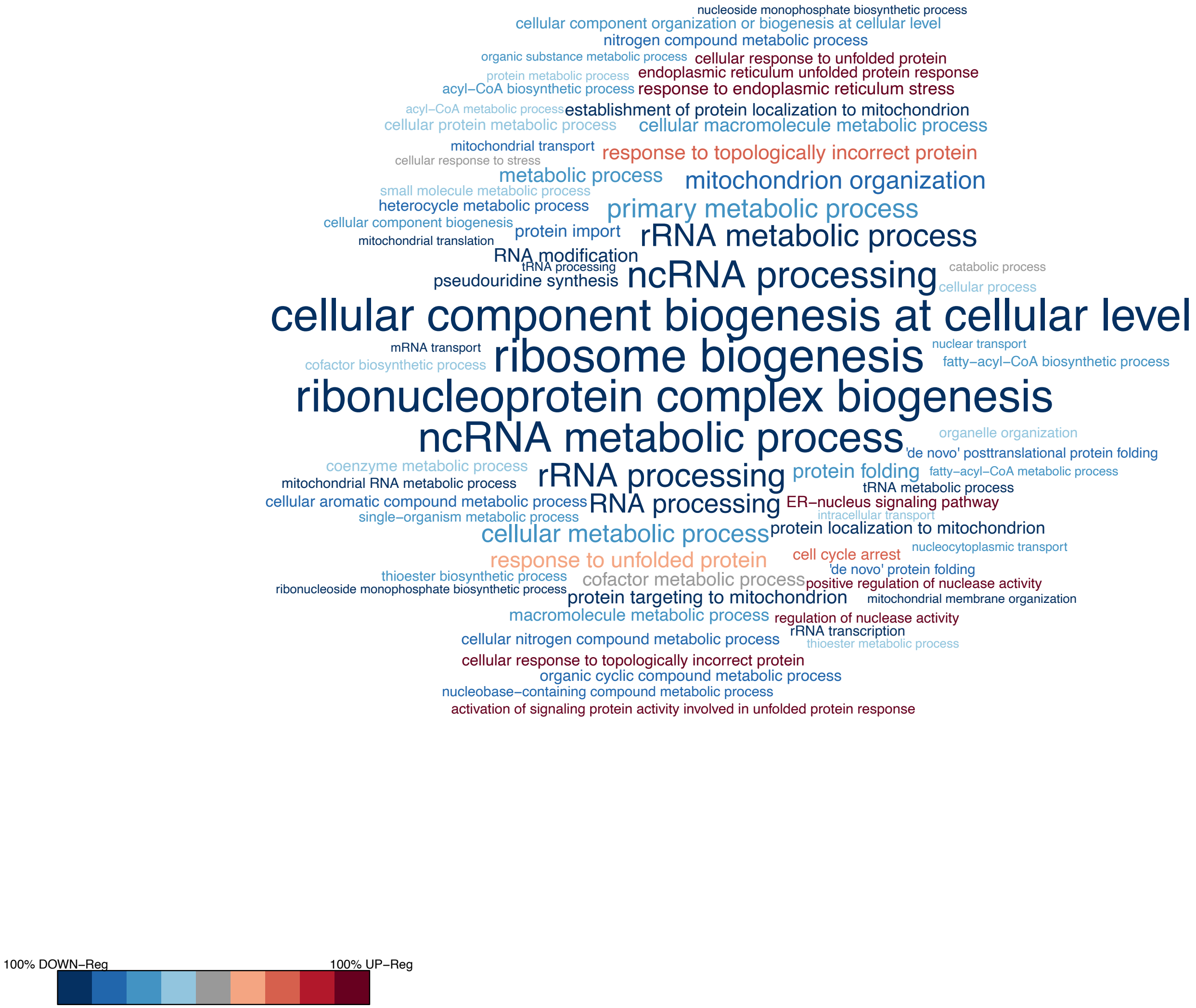
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of functional categories: Gene Ontology.



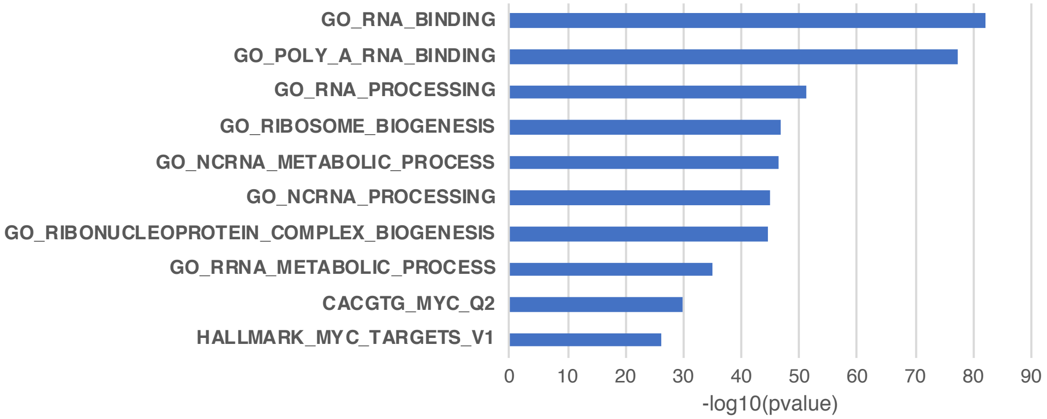
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of functional categories: Gene Ontology.

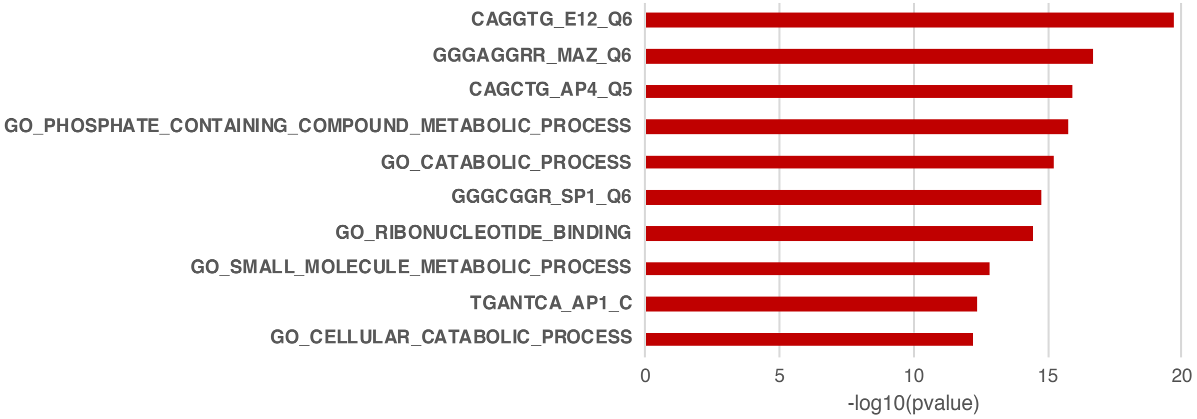


DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of functional categories: Gene Ontology.



Downregulated genes



Upregulated genes

DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Measures of gene expression: TPM (transcripts per milion)

TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here's how you calculate TPM:

1. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
2. Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
3. Divide the RPK values by the "per million" scaling factor. This gives you TPM.

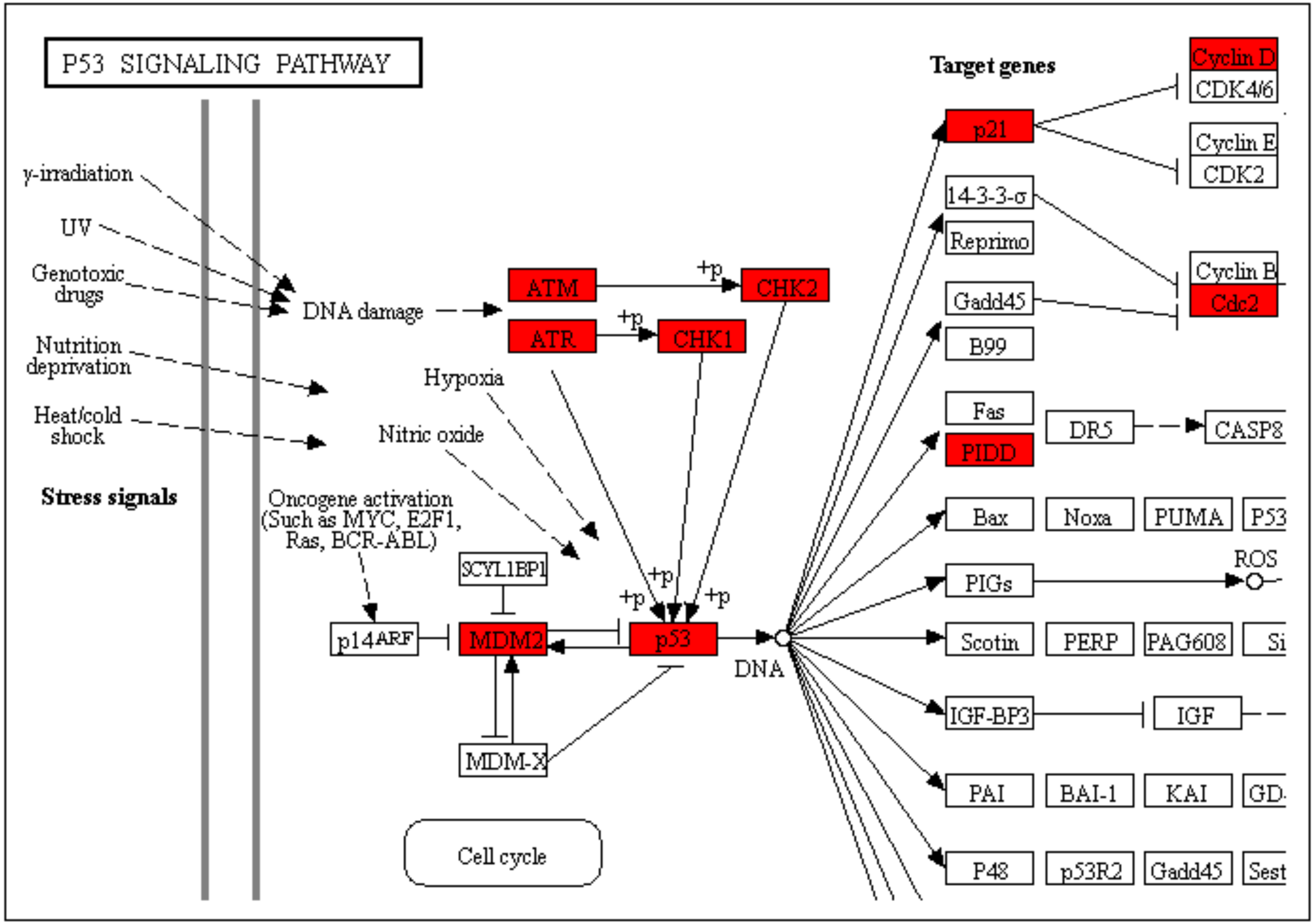
When calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.
When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.

Here's an example. If the TPM for gene A in Sample 1 is 3.33 and the TPM in sample B is 3.33, then I know that the exact same proportion of total reads mapped to gene A in both samples. This is because the sum of the TPMs in both samples always add up to the same number (so the denominator required to calculate the proportions is the same, regardless of what sample you are looking at.)

With RPKM or FPKM, the sum of normalized reads in each sample can be different. Thus, if the RPKM for gene A in Sample 1 is 3.33 and the RPKM in Sample 2 is 3.33, I would not know if the same proportion of reads in Sample 1 mapped to gene A as in Sample 2. This is because the denominator required to calculate the proportion could be different for the two samples.

DATA ANALYSIS: Pathway analysis

Example of functional categories: KEGG pathway


P53 SIGNALING PATHWAY
Stress signals
γ-irradiation
UV
Genotoxic drugs
Nutrition deprivation
Heat/cold shock
DNA damage
Hypoxia
Nitric oxide
Oncogene activation (Such as MYC, E2F1, Ras, BCR-ABL)
ATM
ATR
CHK2
CHK1
+p
SCYL1BP1
p14ARF
MDM2
p53
MDM-X
DNA
Cell cycle
Target genes
p21
14-3-3-σ
Reprimo
Gadd45
B99
Fas
PIDD
Bax
PIGs
Scotin
IGF-BP3
PAI
P48
Cyclin D
CDK4/6
Cyclin E
CDK2
Cyclin B
Cdc2
DR5
CASP8
Noxa
PUMA
P53
ROS
PERP
PAG608
Si
IGF
BAI-1
KAI
GD
p53R2
Gadd45
Sest