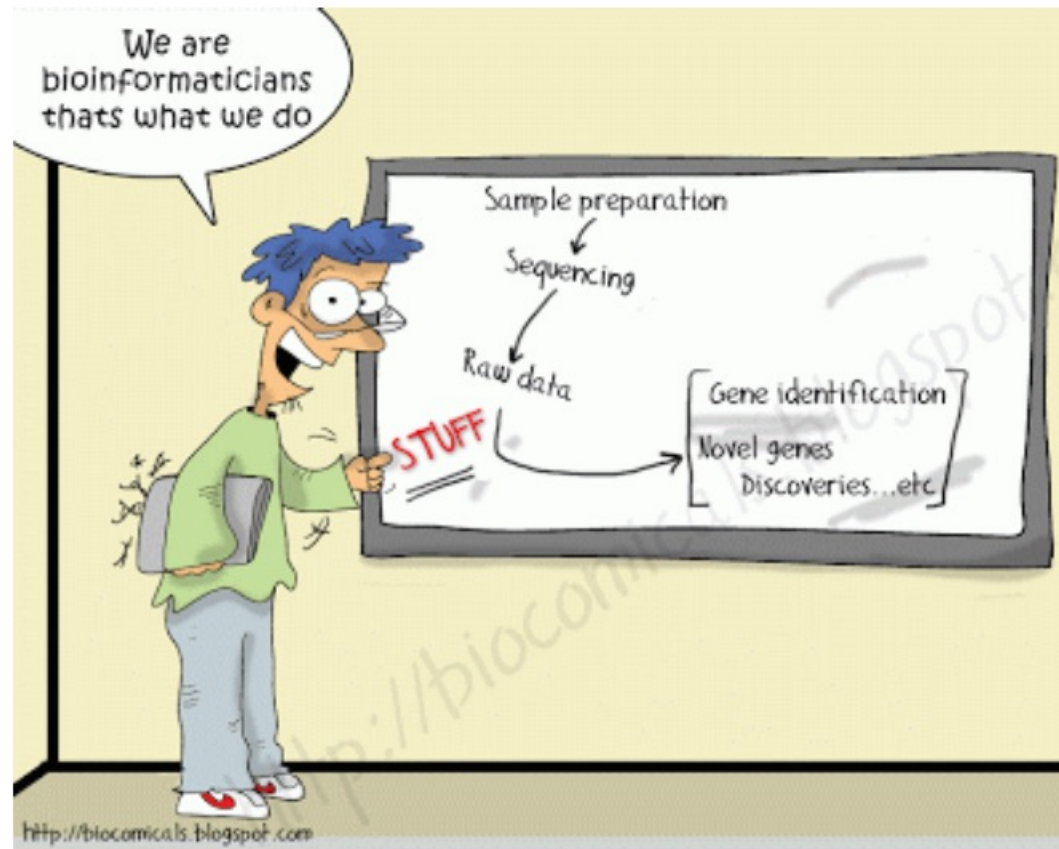
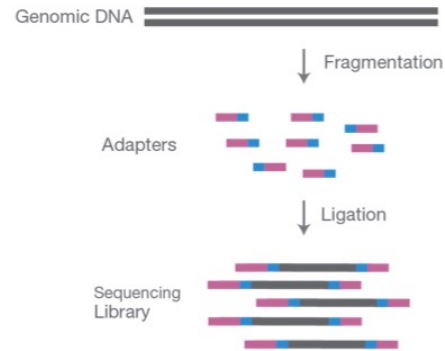


DATA ANALYSIS



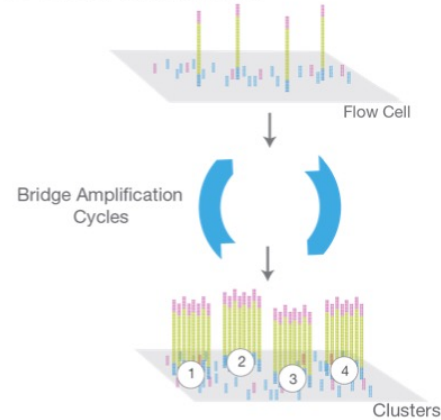
Next-Generation Sequencing Overview

A. Library Preparation



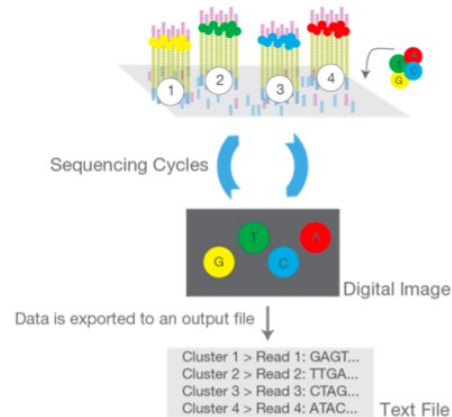
NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

B. Cluster Amplification



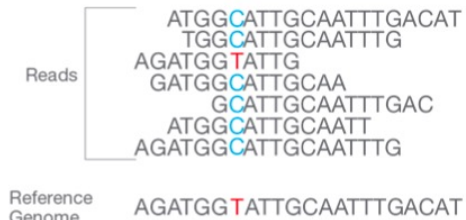
Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

C. Sequencing



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.

D. Alignment and Data Analysis



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.

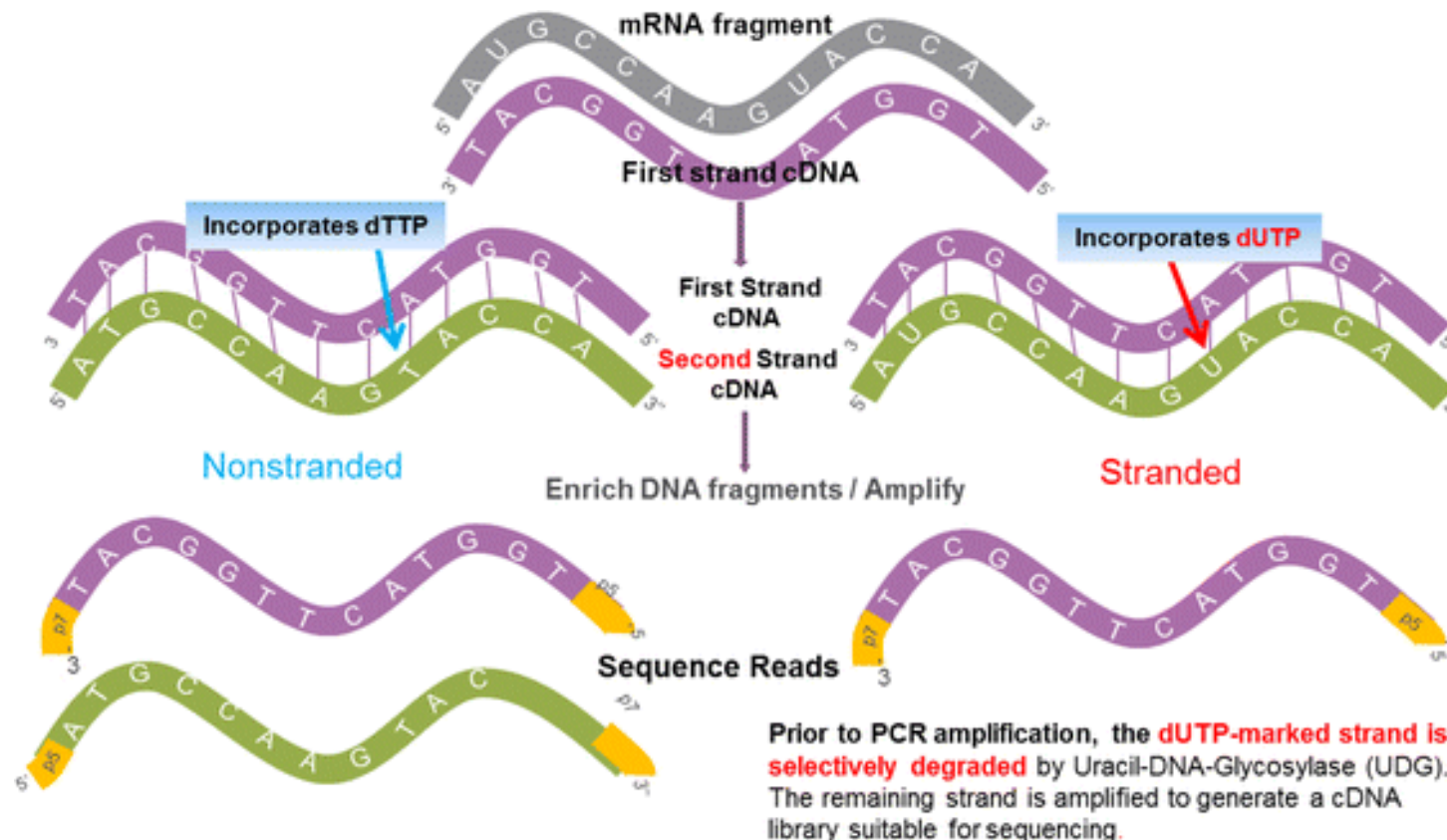


Strandedness

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

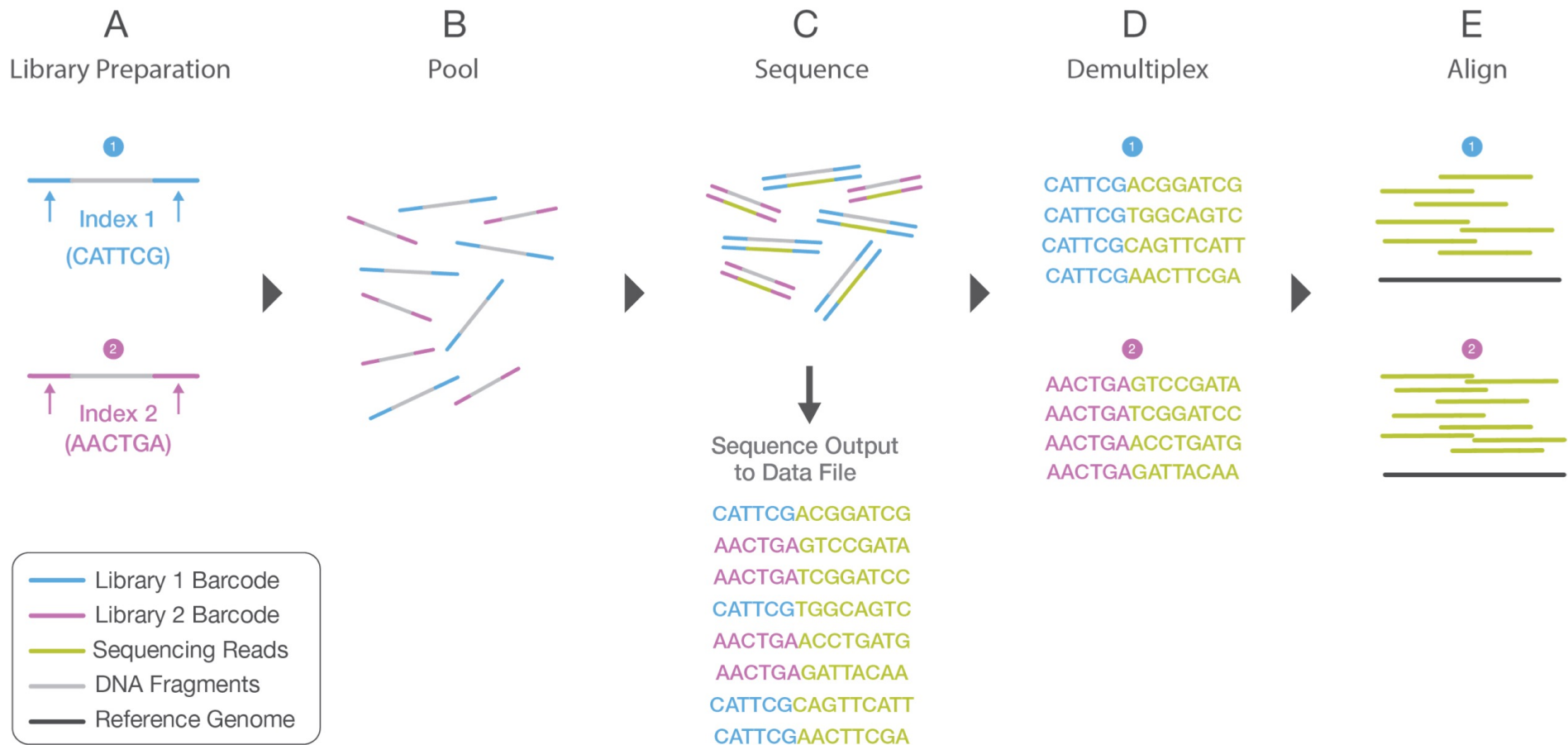
Non-stranded versus stranded RNA-seq protocol

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.



Library Multiplexing Overview

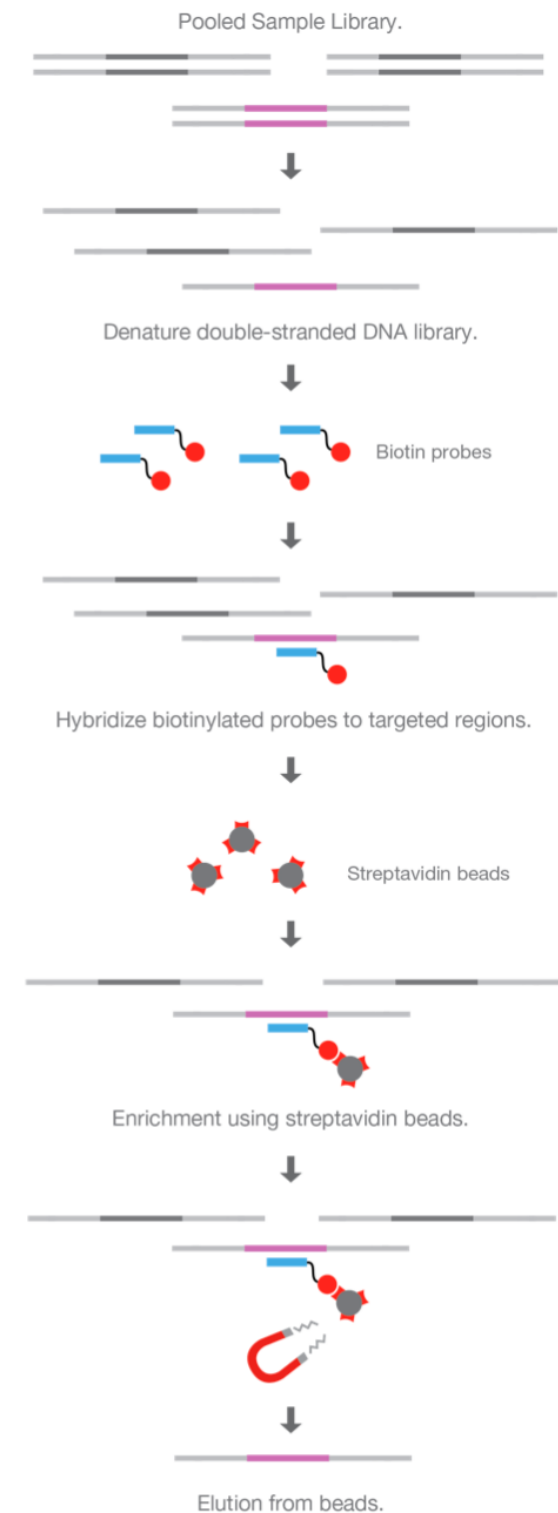
A) Unique index sequences are added to two different libraries during library preparation. (B) Libraries are pooled together and loaded into the same flow cell lane. (C) Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file. (D) A demultiplexing 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

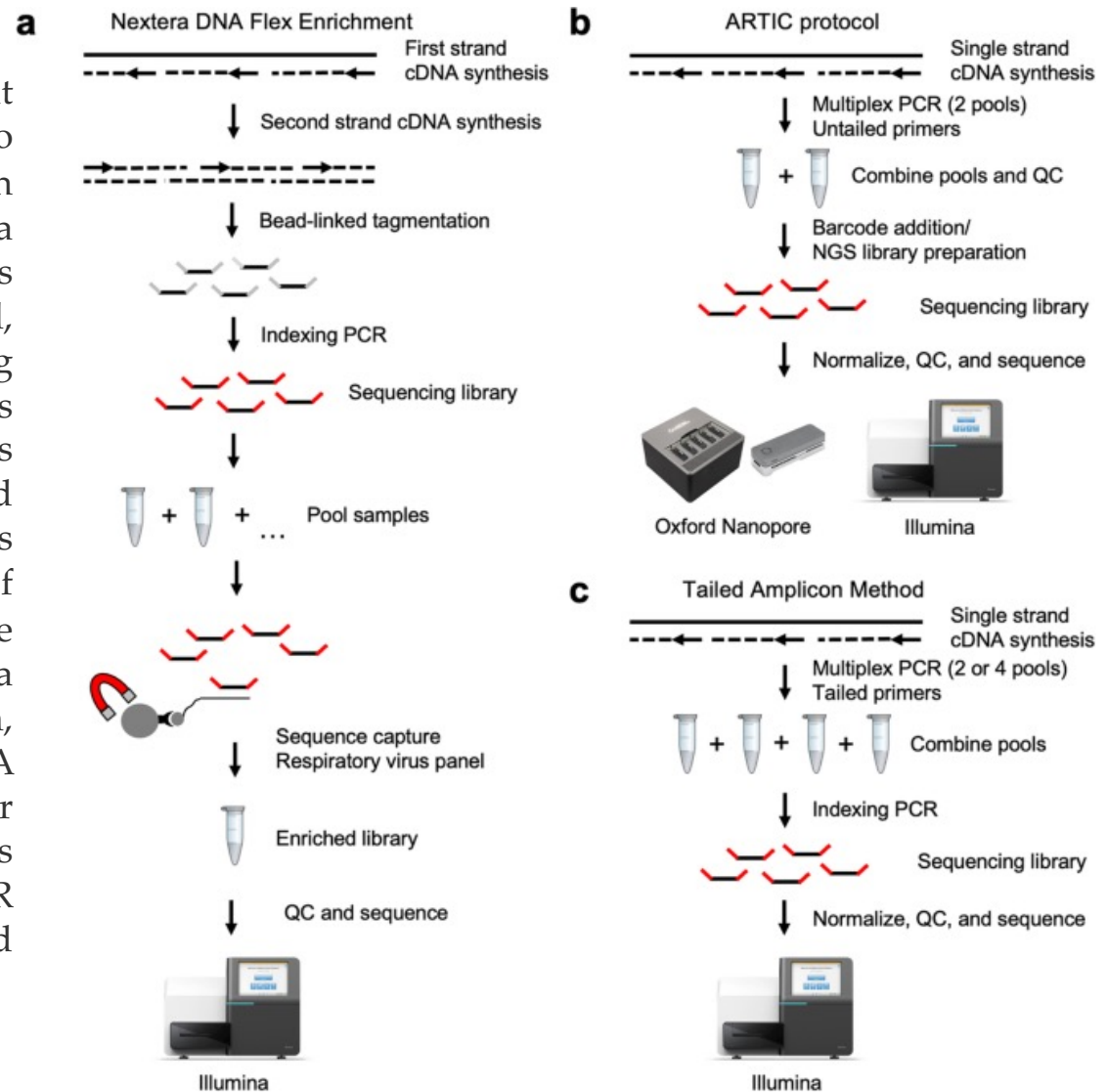
With targeted sequencing, a subset of genes or regions of the genome are isolated and sequenced. Targeted sequencing allows researchers to focus time, expenses, and data analysis on specific areas of interest and enables sequencing at much higher coverage levels. For example, a typical WGS study achieves 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.



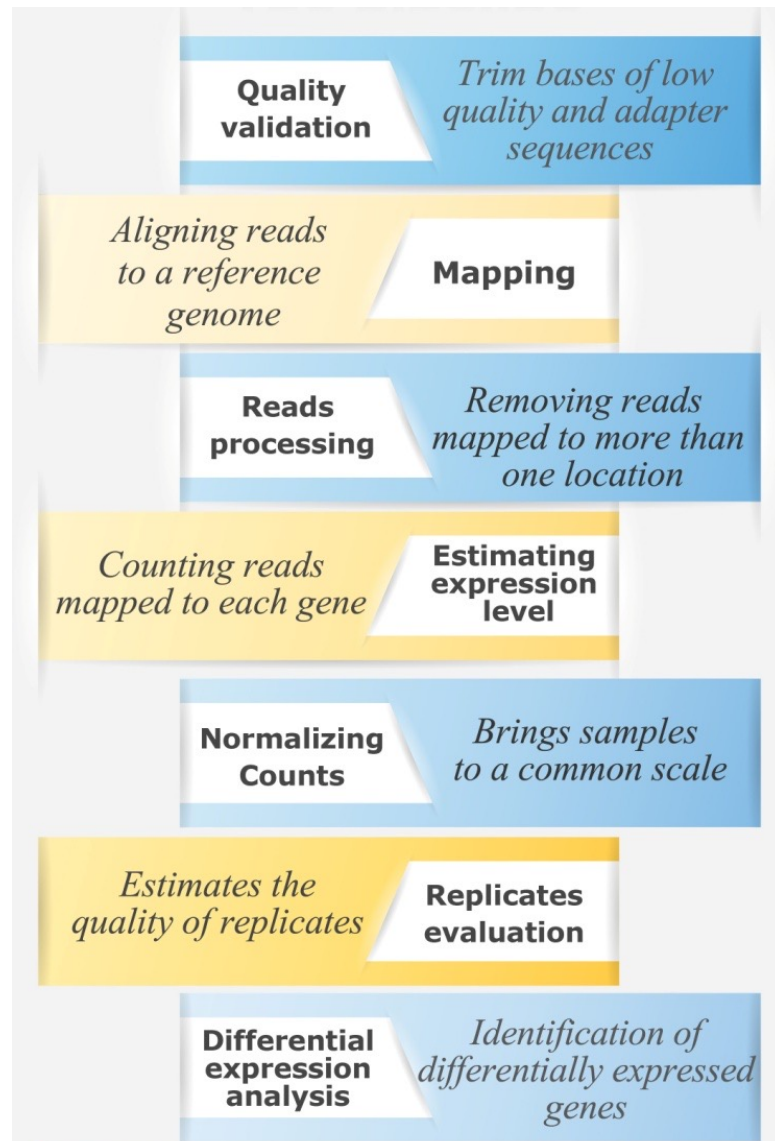
Methods for SARS-CoV-2 genome sequencing

a In Illumina's Nextera DNA Flex Enrichment protocol cDNA is fragmented and made into barcoded sequencing libraries, which are then enriched using sequence capture with a respiratory virus panel containing probes against SARS-CoV-2. **b** In the ARTIC protocol, first strand cDNA is enriched by amplifying with two pools of primers to generate amplicons tiling the SARS-CoV-2 genome. These amplicons are then subjected to either Illumina or Oxford Nanopore library preparation, using methods that either directly add adapters to the ends of the amplicons or to fragment them to enable sequencing on a wider variety of Illumina instruments. **c** The tailed amplicon approach, developed here, enriches first strand cDNA using ARTIC v3 primers containing adapter tails. This allows functional sequencing libraries to be created through a second indexing PCR reaction that adds sample-specific barcodes and flow cell adapters

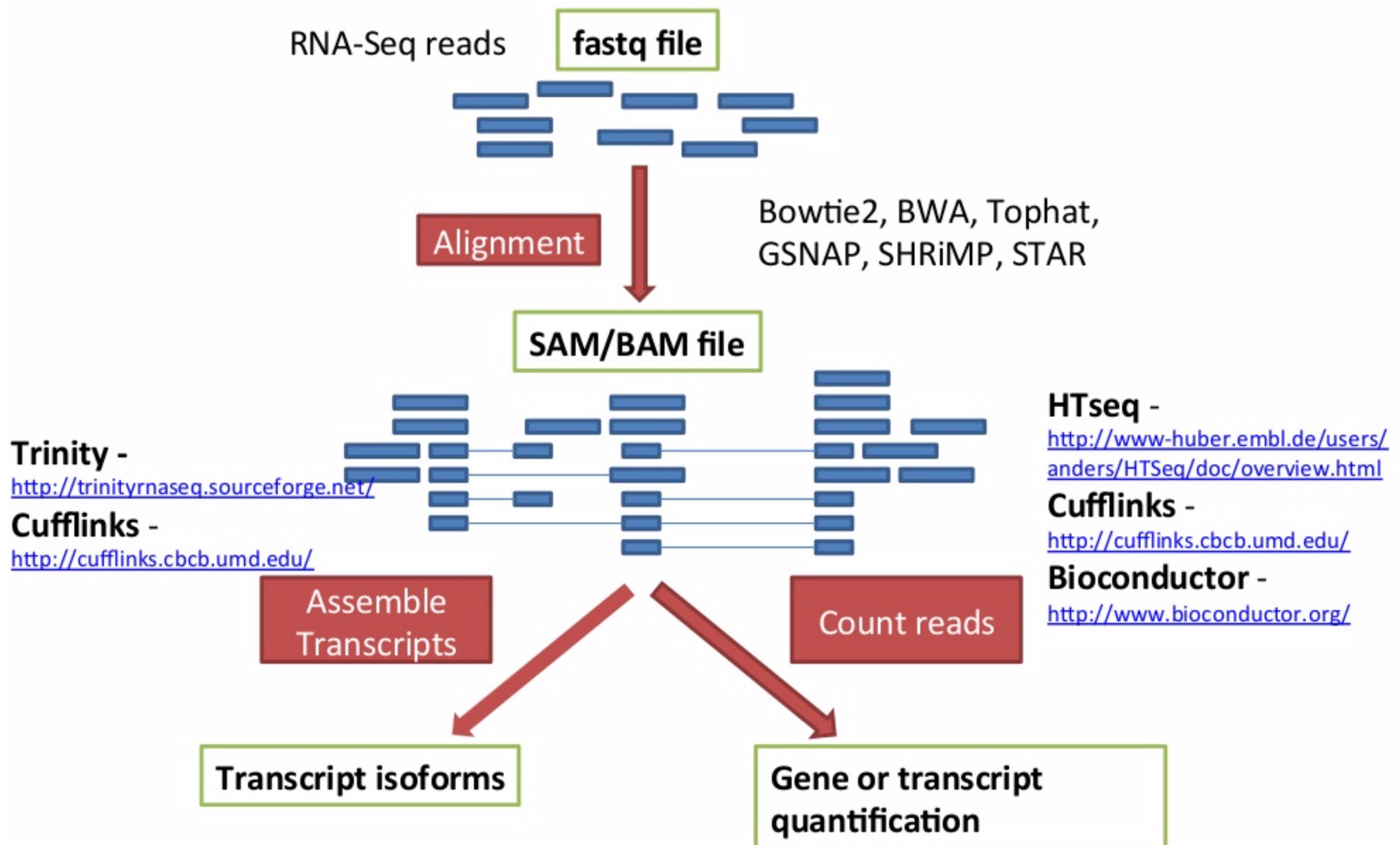


DATA ANALYSIS

■ General RNA-Seq pipeline for Differential Expression

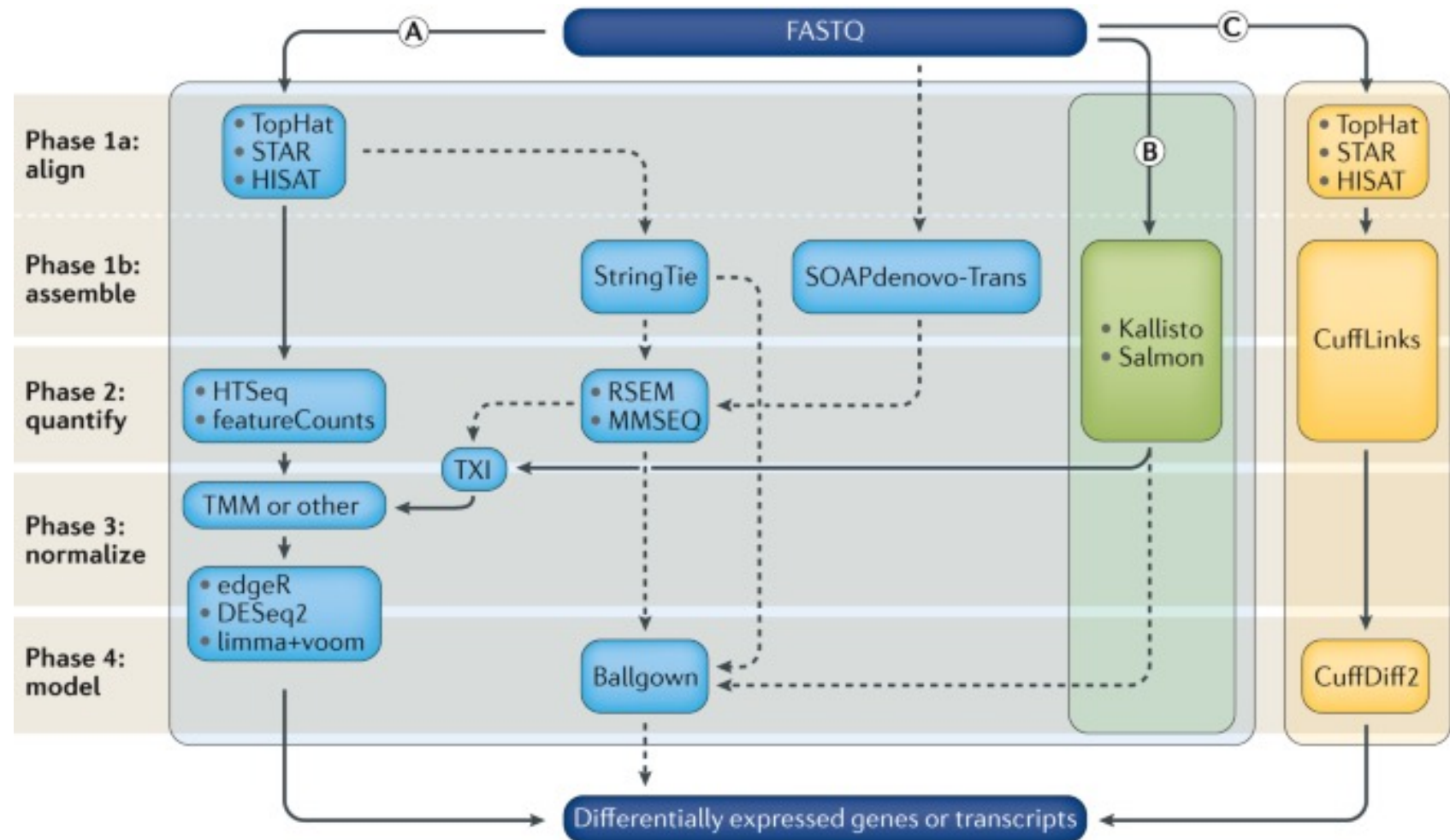


Generalized Analysis Workflow



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.



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

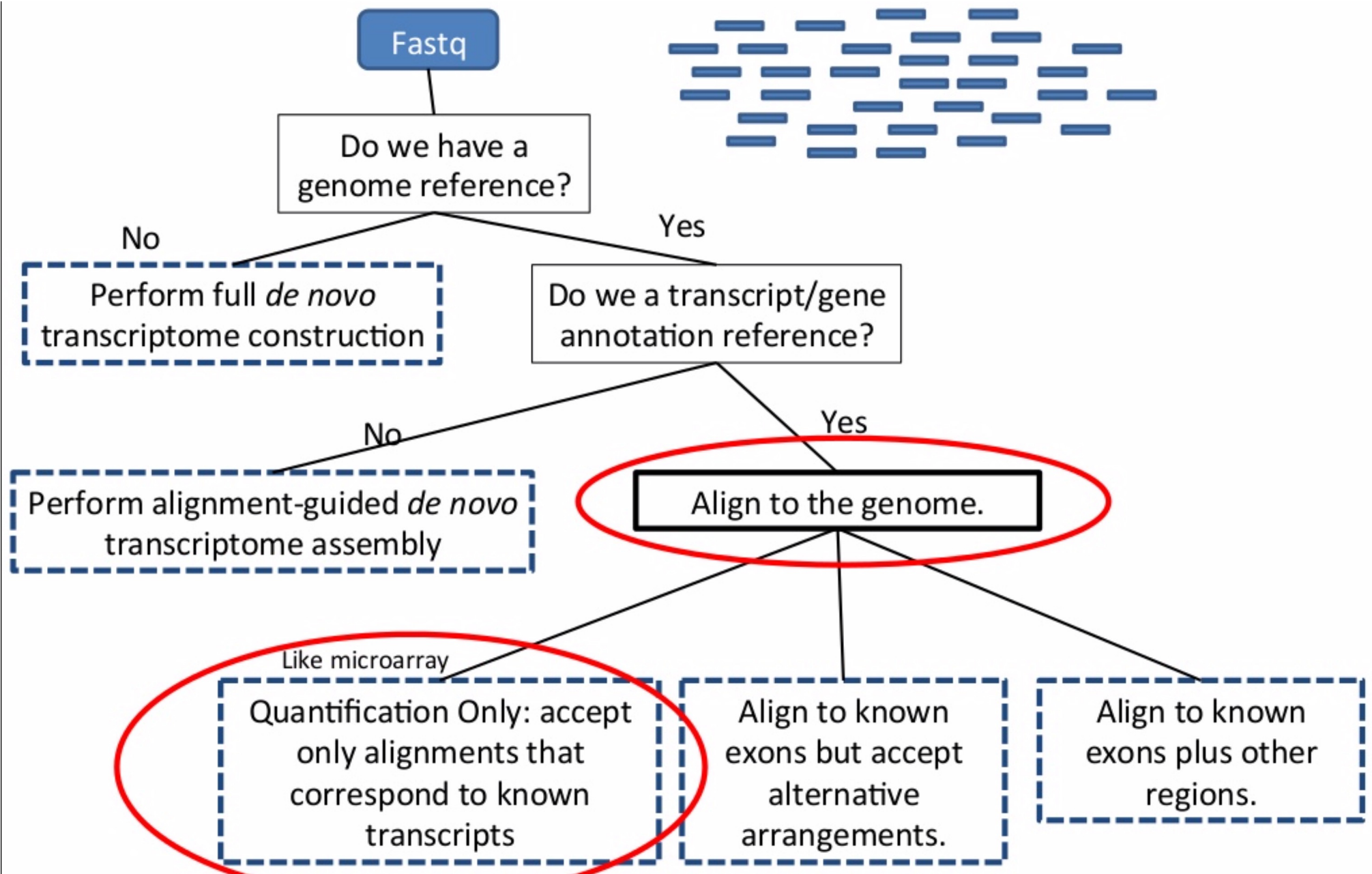
```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;7;;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;7;;;;;;;;-;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;9;7;;.7;393333
```

DATA ANALYSIS

FASTQ format

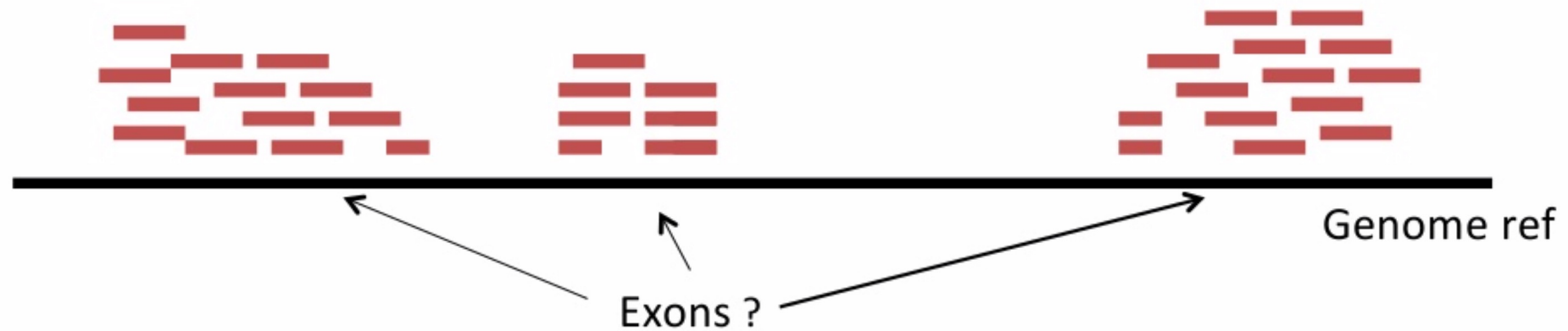
```
@SEQILMN03:128:HA5CBADXX:1:1101:1186:2059 2:N:0:GTCGTA  
NNNNNNGTTAAGATTATTGTTCATTGGCTAACTAAGCGCTACCAAGTACAAGTACAAATGC  
+  
#####0#0<BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB<<<<<<<<<<<<<<<<  
@SEQILMN03:128:HA5CBADXX:1:1101:1193:2104 2:N:0:GTCGTA  
CTATCTTCGTAACCCAAATAAATAAATAACTA.ACTCTATTTCTTGTTAGGCAGGGTATTCC  
+  
BBBFFFFFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIFB707BFFIIIIIIIIIIFFFFFFFFFFF<BBFFF  
@SEQILMN03:128:HA5CBADXX:1:1101:1227:2106 2:N:0:GTCGTA  
GGGGAGCATGACGGCCCACATCGGCGAAAACCCACTCTGGTGGGGTGAACCGGTATCCAN  
+  
BBBFFFFFFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIFFFFFBBFFFBFBFBFBFFBFF<BBFFO<BBFFBFBFBFFFB
```

DATA ANALYSIS: ALIGNMENT



What to map to?

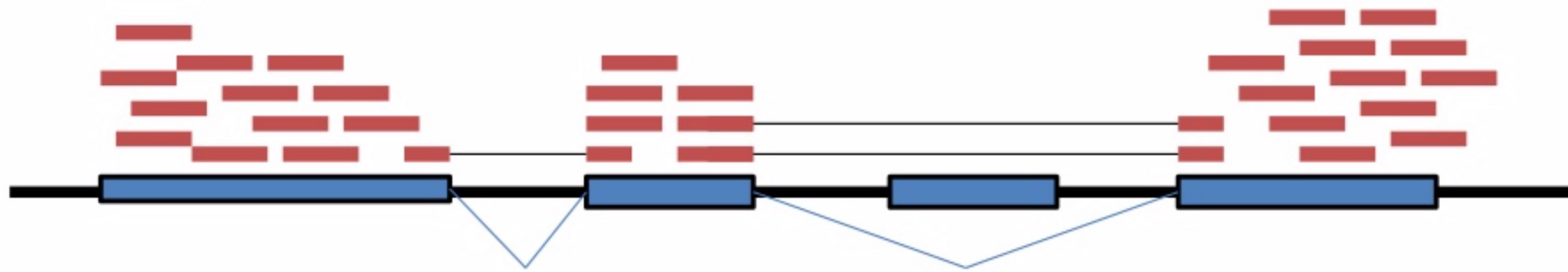
Map to a genome with no gene annotation.



- 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 effectively 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.

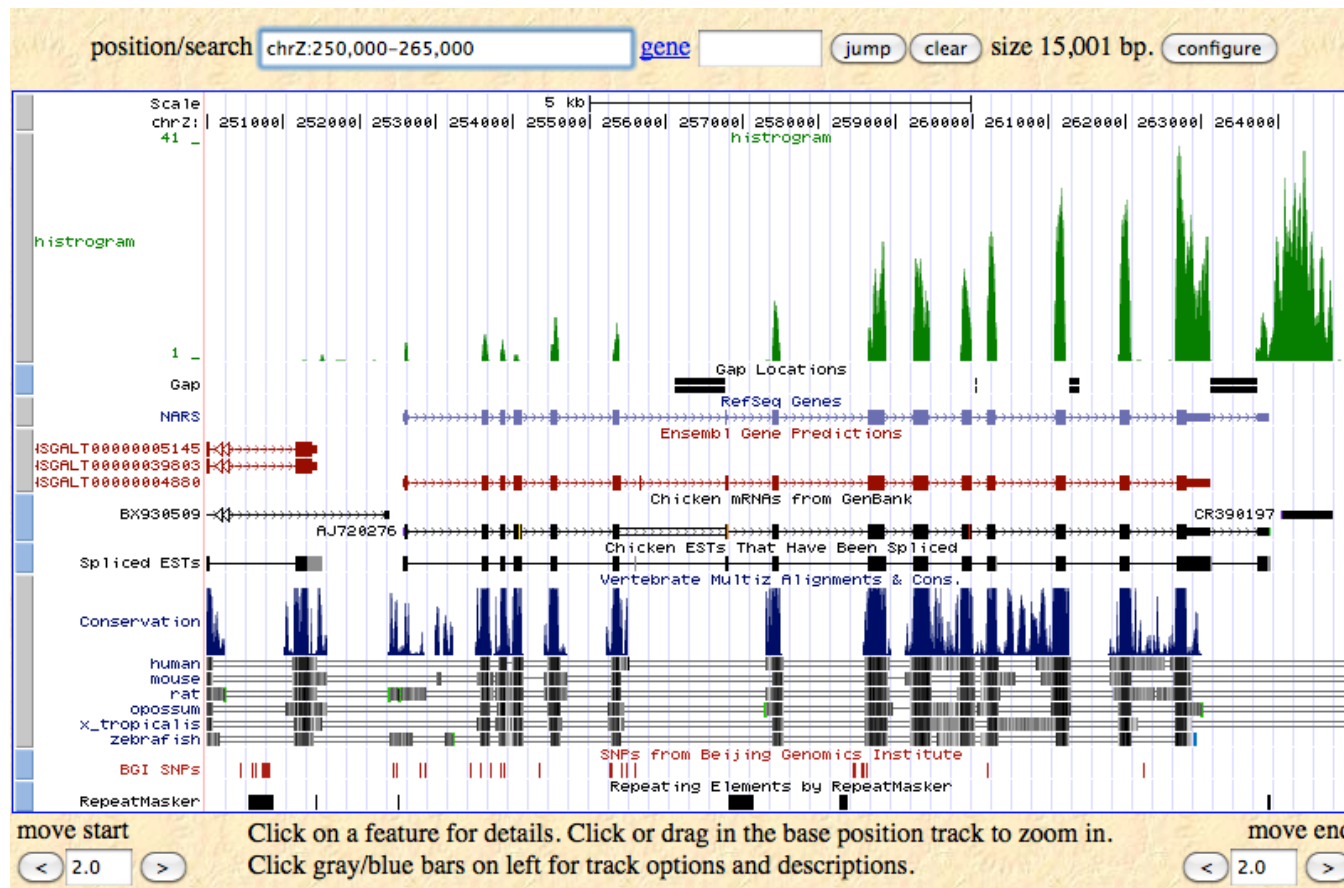
```
@HD      VN:1.0  SO:coordinate
@SQ      SN:chr20      LN:64444167
@PG      ID:TopHat      VN:2.0.14      CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-align-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr20 /data/user446/mapping_tophat/L6 18 GTGAAA L007 R1 001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714      16      chr20      190930      3      100M      *      0      0
      CCGTGTTTAAAGGTGGATGCGGTCACCTTCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCTCTCAGTCCCCCTCT
C      BBDDCCDDCCDDDDCCDDDDDCDDCCDBC?DDDDDDDDDDDDDDDCDDDDDDDDDDCCCEDDDC?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
      TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCTGACATAAGGGGCATGGACGA
G      DCDDEDDDDDDDDDDDDDDCCDDDDCCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJGJJJIJJJJJIHJJJJJHHHHHFFFFCC
      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
      GGCTTTATTGGTAAAAAGGAATAGCAGATTTAATCAGAAATCCCACCTGGCCAGCAGCACCAACCAGAAAGAAGGAAGAAGACAGGAAAAAACCA
C      DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJIHJJJIJJJJIIIGFJJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFDD@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
      0      GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACCTTGGGTCTCCGAAGCAGAACATCTCAAAATATGACCTCTCG
```

DATA ANALYSIS: ALIGNMENT

BAM file visualization



Genome Browser (UCSC)



The Sequence Read Archive (SRA)

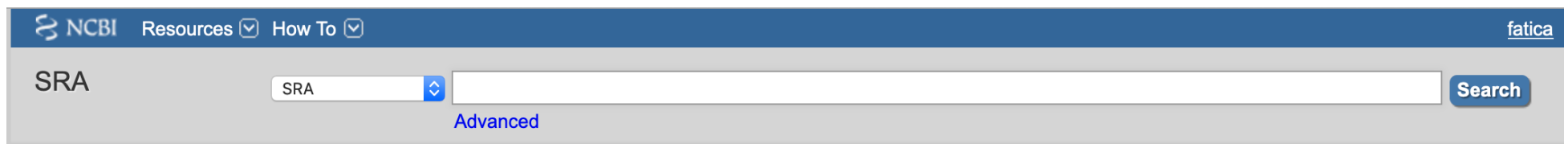
The SRA was established as a public repository for the next-generation sequence data and is operated by the International Nucleotide Sequence Database Collaboration (INSDC).

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

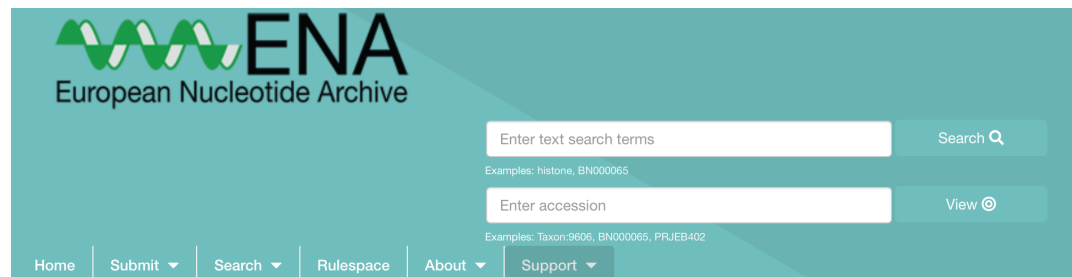
The SRA is accessible at

<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.



The screenshot shows the top navigation bar of the NCBI website with links for 'Resources' and 'How To'. Below this is the 'SRA' section, which includes a search bar with a dropdown menu set to 'SRA', a large text input field, and a 'Search' button. A link labeled 'Advanced' is positioned below the search bar.



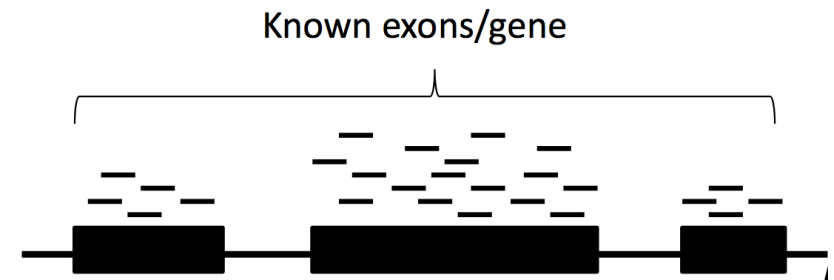
The screenshot shows the EBI ENA (European Nucleotide Archive) search interface. It features the ENA logo and a search bar with a dropdown menu. Below the search bar are two input fields: 'Enter text search terms' with a 'Search' button, and 'Enter accession' with a 'View' button. Examples of search terms and accessions are provided below each input field. A navigation bar at the bottom includes links for 'Home', 'Submit', 'Search', 'Rulespace', 'About', and '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 EXPRESSION LEVELS

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

- C : Number of mappable reads on a feature (eg. transcript, exon, etc.)
- L: Length of feature (in kb)
- N: Total number of mappable reads (in millions)

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.



RPKM = 1



RPKM = 2

FPKM = 1

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
for 4 gene samples we divide by to get the "million" scaling			
Tens of reads	3.5	4.5	10.6
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
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
for 4 gene samples we divide by 10 to get the "million" scaling			
Tens	1.5	2.025	4.51
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.

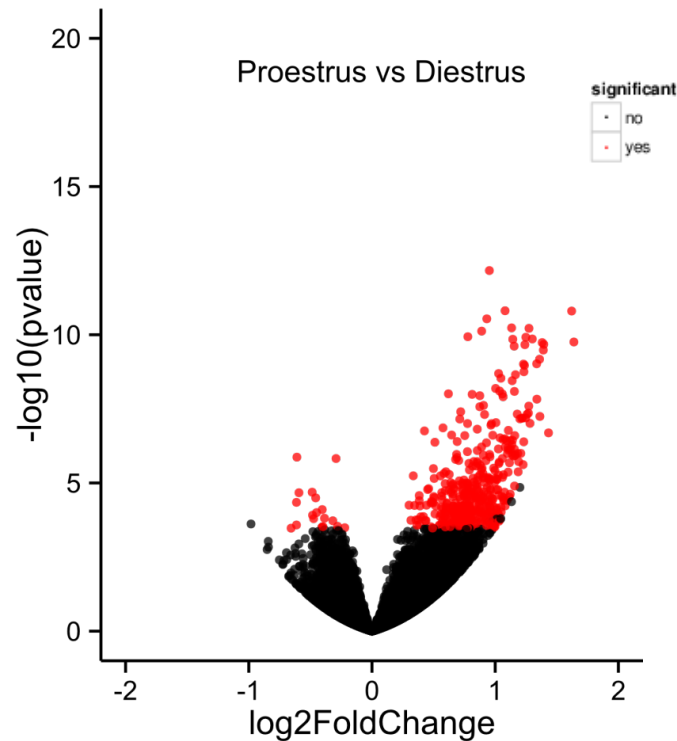
DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

Symbol	Exp Sample A	Exp Sample B	log2FC	p-value	padj
AC116533.1	201.723	0	-7.663366	0.00705426	0.04685518
CES1	8.326515	0.319710333	-2.8211168	0.03160709	0.1497986
TRAJ13	5.887466667	0	-2.7839734	0.53022794	0.93016935
AC245100.3	7.45177	0.257067333	-2.7491916	0.00329661	0.02539452
TRIM27	15.67494	1.771937667	-2.5887148	0.07326134	0.27982537
HIST2H2AA3	4.73542	4.33333E-06	-2.5198929	0.08293496	0.30704889
TRAJ31	4.439433333	0	-2.4434564	0.61996183	0.94844719
AC073415.1	10.40004667	1.144743333	-2.4101628	0.08903746	0.32211488
AC243913.7	3.00402	0	-2.0014492	0.5319005	0.93075919
TRAJ43	2.998836667	0	-1.9995804	0.65544674	0.94844719
TRAJ40	2.910686667	0	-1.9674219	0.72005632	0.94844719
TRAJ36	2.847366667	0	-1.9438713	0.63177075	0.94844719
TRAC	64.21976667	15.97356667	-1.9420195	1.7768E-42	7.6648E-39
TRAJ35	2.81876	0	-1.9331043	0.62529611	0.94844719
TRAJ45	2.75886	0	-1.9102952	0.67188498	0.94844719
TRAJ42	2.75886	0	-1.9102952	0.67188498	0.94844719
TRAJ3	2.676943333	0	-1.8785069	0.65544674	0.94844719
TRAJ6	2.669356667	0	-1.8755271	0.63177075	0.94844719
TRAJ4	2.62726	0	-1.8588802	0.62529611	0.94844719
NPIPA5	9.020533333	1.817733333	-1.8303523	0.57711311	0.94844719
AL022345.1	13.31823333	3.051946667	-1.8211664	0.22159401	0.59155374
AL354798.1	8.36444	1.689066667	-1.8000872	0.02610512	0.12893548
MCCC2	5.119086667	0.788869	-1.7742686	0.31473251	0.72422555
DEFB130C	2.41471	0	-1.7717631	0.27663176	0.67260404
AL033528.1	2.33771	0	-1.7388586	0.67188498	0.94844719
AC103810.6	2.312253333	0	-1.727813	0.68907179	0.94844719
AL356095.3	2.21609	0	-1.6853078	0.83918886	0.94844719
TREM2	11.3937	2.910596667	-1.6641463	4.7294E-11	4.658E-09
AP000275.2	5.66981	1.146606333	-1.635588	0.06351326	0.25265907
AC087752.2	2.910256667	0.265410667	-1.6276576	0.09783078	0.3440031
CT45A8	3.200013333	0.37426	-1.6117389	0.11279077	0.37988509
AC098934.3	2.001516667	0	-1.5856917	0.5319005	0.93075919
RHEX	32.10226667	10.09589667	-1.5769038	1.1775E-14	2.6735E-12

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.



The p-value in RNA-seq

Input Data: Normalised read counts mapped to individual genes

Null Hypothesis (H_0): Gene X shows no difference in expression between treated and control conditions

What the p-value Measures

The probability of observing a fold-change as large as (or larger than) the experimental result purely by chance, assuming H_0 is true.

$p < 0.05$: Statistically significant evidence to reject H_0 for an individual gene.

A standard p-value is valid when testing one hypothesis at a time. It breaks down when testing thousands of genes simultaneously.

Controlling False Discovery Rate (FDR)

FDR Definition: The expected proportion of false positives among all genes declared significant.

Method: P-value correction algorithms (e.g., Benjamini-Hochberg procedure) yield the Adjusted p-value (p_{adj} or q-value).

$p_{\text{adj}} < 0.05$: Means that at most 5% of your identified significant genes are expected to be false discoveries.

Benjamini–Hochberg (BH) correction

$$\text{Adjusted p-value} = \text{Raw p-value} \times \frac{\text{Total Genes (m)}}{\text{Rank (i)}}$$

Total genes tested (m) = 5

Rank / gene	Raw p	Calculation Formula	p adj	Outcome (≤ 0.05)
1 / A	0.001	$0.001 \times 5/1$	0.005	Significant
2 / B	0.008	$0.008 \times 5/2$	0.020	Significant
3 / C	0.030	$0.003 \times 5/3$	0.050	Significant
4 / D	0.045	$0.045 \times 5/4$	0.056	NOT Significant
5 / E	0.120	$0.120 \times 5/5$	0.120	NOT Significant

Gene D is discarded as a likely false positive, demonstrating why multiple testing corrections are essential in high-throughput data

Statistical vs. Biological Significance

Statistical Significance ($p_{adj} < 0.05$): "Is the observed difference real or a random coincidence?" Driven heavily by sample size and measurement precision.

Biological Significance: "Is the size of this change large enough to actually alter cellular function or phenotype?" Driven by physiological context and effect size.

Example

Gene A: Shows a 5% increase in expression (Fold Change=1.05) with a highly significant $p_{adj} = 0.0001$.

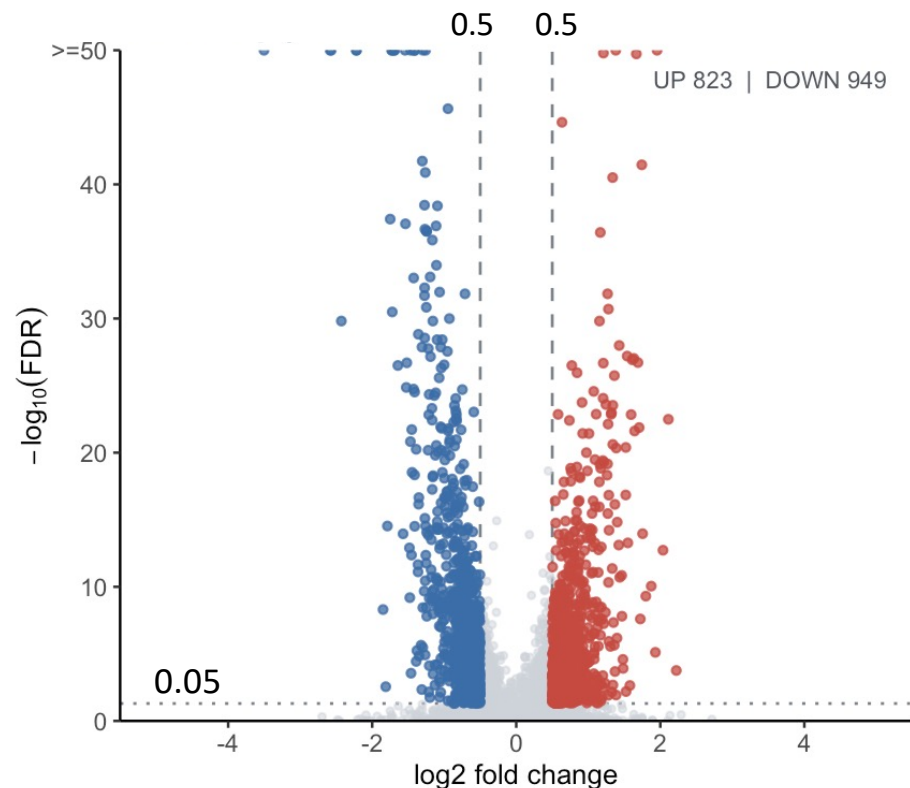
Statistically Significant: Yes ($p_{adj} < 0.05$) because high statistical power detects tiny variations with high confidence.

Biologically Significant: No. A 5% shift in mRNA abundance is typically buffered by post-transcriptional controls and has zero impact on protein levels or cell behavior.

Statistical vs. Biological Significance

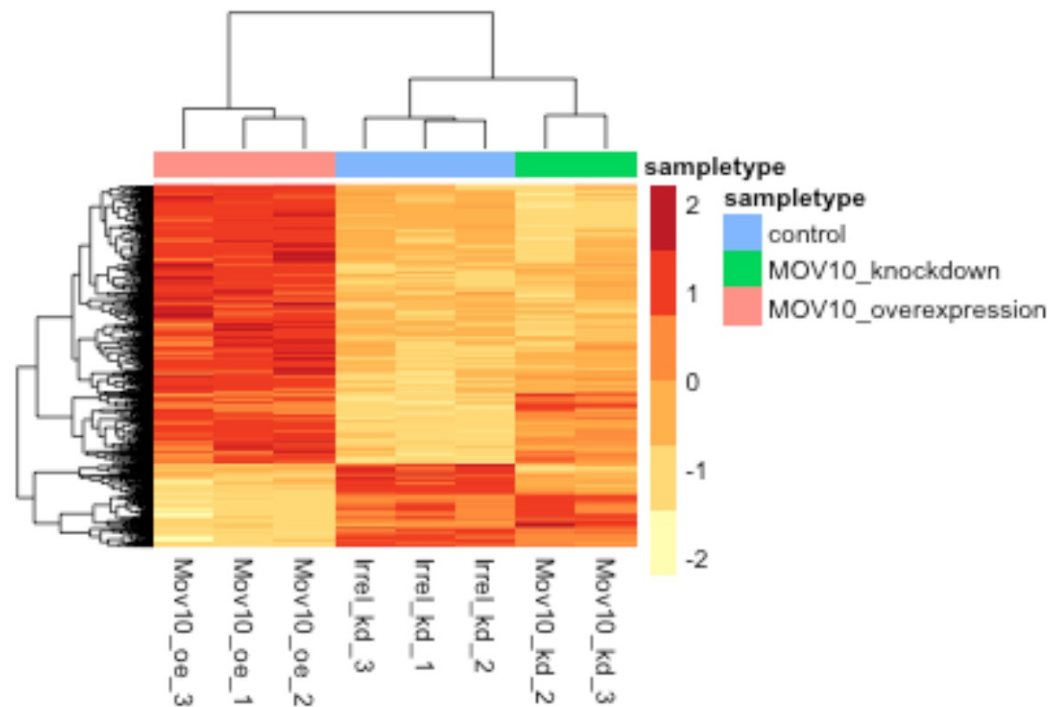
To isolate biologically meaningful target genes in RNA-seq, always filter using both metrics simultaneously:

Statistical Threshold ($\text{FDR} < 0.05$) + Effect Size Threshold ($|\text{Log}_2 \text{ Fold Change}| \geq 0.5$)



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

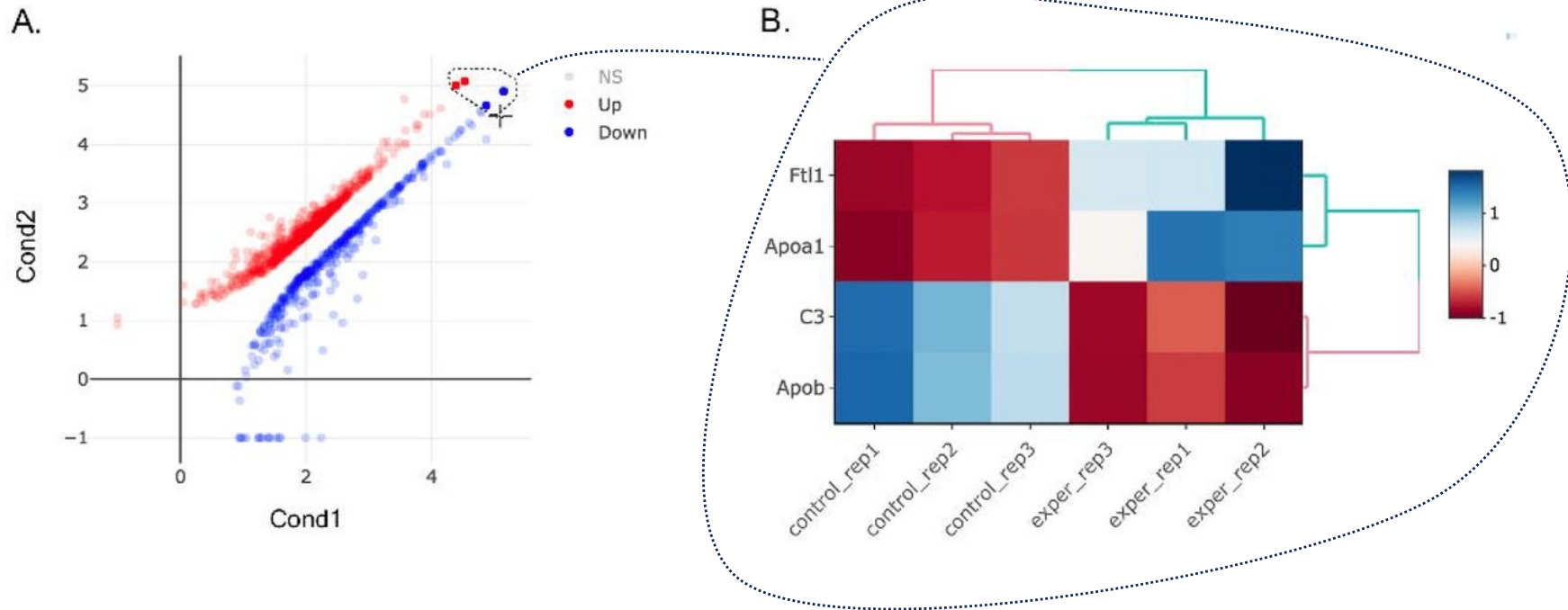
δ = SD

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.*

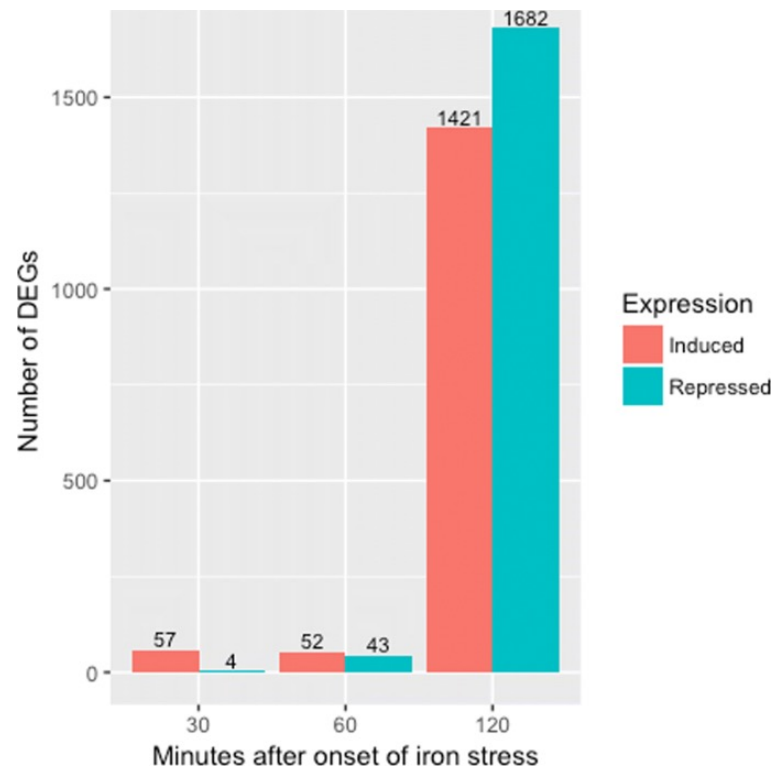
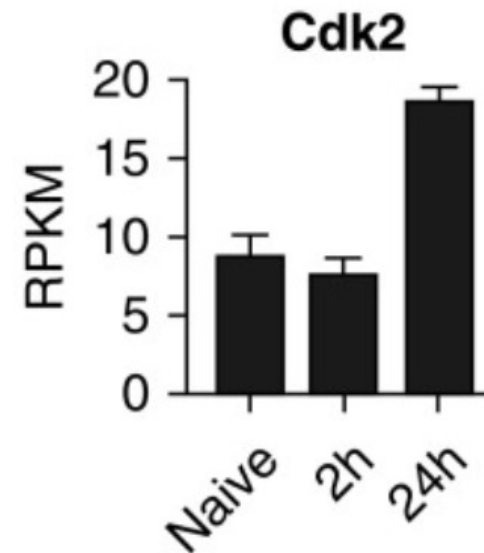
DATA ANALYSIS: QUANTIFICATION OF GENE EXPRESSION

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	negLogP	t	P.Value	adjP.Val	B
4490	RGS5	regulator of G protein signaling 5	4.9759197	14.0030864	26.5015659	4.67E-18	1.2E-13	21.293391
4309	ACD2	acyl CoA oxidase 2, branched chain	3.841977	8.50997120	26.321512	5.58E-18	2.67E-14	31.033976
7861	CHRF7	G protein-coupled receptor (metabotropic receptor type B-like)	4.1361864	6.1591491	26.1613437	6.68E-18	2.67E-14	28.884634
5514	CHRNA4	cholinergic receptor, nicotinic, alpha 4	5.349602	7.5654312	24.9512596	1.79E-17	2.58E-14	29.890214
6453	MYL7	myosin, myosin VIIA	3.5094042	7.4126031	24.3915724	2.78E-17	6.68E-14	29.543605
1203	CCNA4	collagen, type IV, alpha 3	4.4976801	7.3882025	23.6071947	5.54E-17	1.13E-13	28.890111
4696	SPPI	serpin, proteinase inhibitor 1	3.0384472	9.1039740	23.4110429	6.52E-17	1.13E-13	28.736026
6406	PLP4	PLP4, TPA, cytosolic effector	3.3482221	6.6184841	23.2043781	7.95E-17	3.38E-13	28.469643
814	CAMP2	cyclic AMP, type II, large subunit	5.788461	18.9575021	22.6149005	1.34E-16	5.78E-13	28.0110714
8871	SYND1	syndecan 1	3.3821509	6.4281317	22.4028514	4.65E-16	3.99E-13	27.897587
6261	S100B	S100 calcium binding protein B	2.5370649	6.7993111	22.3611889	1.73E-16	3.99E-13	27.8953284
29061	SYNPR	synaptic binding protein 5 (synaptic)	4.426822	6.3121287	22.0412007	3.40E-16	3.62E-13	27.22274
1368	CPE1	carbonic dehydratase II, cytoplasmic 1	3.625287	6.0601761	21.1701958	5.38E-16	4.87E-13	26.775654
5345	PDZD3	PDZ domain containing 3 (transport regulator 3)	3.5617883	6.0348763	20.7777744	7.02E-16	6.79E-13	26.336543
80702	TRAF3	TRAF domain containing 3 (ubiquitin domain, member 3)	3.1131648	6.1121822	20.4421086	1.11E-15	6.79E-13	25.079626
1410	CHAF	chaperonin, alpha 3	3.5722258	7.2798281	20.1849312	1.40E-15	1.06E-12	25.7924999
15489	C12orf43	C12orf43, human protein factor related protein 3	3.5348451	7.7396461	20.1621021	1.50E-15	1.06E-12	25.7148871
7301	TYRO3	TYRO3 protein tyrosine kinase	4.4931701	7.48015678	20.0516354	1.65E-15	1.06E-12	25.614514
7081	TRAF2	TRAF domain containing 2	2.1250426	6.1010801	19.8112021	2.03E-15	2.79E-12	25.611305
5187	PLU1	plasma membrane protein 1, ubiquitin	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
84732	SPR1	SPR1, specific, E3 ubiquitin protein ligase 2	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
5361	PLU1	plasma membrane protein 1, ubiquitin	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
8448	TRAF3	TRAF domain containing 3 (ubiquitin domain, member 3)	3.1131648	6.1121822	20.4421086	1.11E-15	6.79E-13	25.079626
202	ADAM10	ADAM10, metalloprotease, family member 10	2.7384811	6.4482299	19.4791297	4.01E-15	1.48E-12	25.0187986
4208	KUFC	kinase, protein kinase C	4.5177211	7.2762721	19.4517651	3.09E-15	1.48E-12	25.0122806
5842	MYL10	myosin, myosin X	3.4767611	6.5440051	19.3071076	4.01E-15	1.48E-12	25.0122806
4881	NPTR	neuronal protein 1	4.5177211	7.2762721	19.4517651	3.09E-15	1.48E-12	25.0122806
1310	PLA2G	phospholipase A2, group IIA	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
6497	SPR1	SPR1, specific, E3 ubiquitin protein ligase 2	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
1467	CHAF	chaperonin, alpha 3	3.5722258	7.2798281	20.1849312	1.40E-15	1.06E-12	25.7924999
5604	PLU1	plasma membrane protein 1, ubiquitin	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
852	GAB2	growth factor receptor-specific 2	2.1111291	6.0601761	19.8112021	1.14E-14	4.39E-12	23.719146
160	PLA2G	phospholipase A2, group IIA	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
5759	KUFC	kinase, protein kinase C	4.5177211	7.2762721	19.4517651	3.09E-15	1.48E-12	25.0122806
5725	SLAC	small cell carcinoma 1	3.818551	6.3161081	18.812551	3.85E-14	4.58E-12	23.5641664
1214	PLA2	phospholipase A2, group IIA	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
1314	ACV2	acyl CoA oxidase 2	3.841977	8.5099712	26.321512	5.58E-18	2.67E-14	31.033976
995	HIST1A1	histone H1A	2.862461	6.8554051	17.7114016	2.09E-14	4.58E-12	23.113554
188	DN	de novo	2.381871	6.4471441	17.7114016	2.09E-14	4.58E-12	23.113554
1071	NET1	neuronal protein 1	2.1111291	6.0601761	19.8112021	1.14E-14	4.39E-12	23.719146
8712	ADAM19	ADAM19, metalloprotease, family member 19	2.2447671	7.5183201	17.4521088	2.82E-14	4.58E-12	23.113554
1912	EDNA	endothelial domain containing 1	3.9711071	6.4741081	17.4521088	2.82E-14	4.58E-12	23.113554
302	AM1	adrenomedullin 1	3.384701	6.3741081	17.4521088	2.82E-14	4.58E-12	23.113554
1012	ELP1	elastin-like protein 1	3.384701	6.3741081	17.4521088	2.82E-14	4.58E-12	23.113554
284	GN	glycine	2.4290641	6.8109161	16.8104416	1.37E-14	4.58E-12	23.113554
6404	WDR4	WD repeat domain 4	2.618511	6.1201081	16.8104416	1.37E-14	4.58E-12	23.113554
10061	KUFC	kinase, protein kinase C	4.5177211	7.2762721	19.4517651	3.09E-15	1.48E-12	25.0122806
5489	ELC1	ELC1, type I, cytosolic	2.099241	6.1010801	19.8112021	1.14E-14	4.39E-12	23.719146
489	ND4	NADH dehydrogenase 4	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
539	PAF	phosphatidylcholine-specific phospholipase	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
18171	LYN1	LYN1, tyrosine kinase	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
8442	SH3GLB1	SH3 domain containing 1	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
722	PLA2G	phospholipase A2, group IIA	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
640	SH3GLB1	SH3 domain containing 1	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
620	SH3GLB1	SH3 domain containing 1	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
1811	PRK1	protein kinase 1 (serine/threonine)	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
871	SH3GLB1	SH3 domain containing 1	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
5391	PLU1	plasma membrane protein 1, ubiquitin	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
991	CHRF7	G protein-coupled receptor (metabotropic receptor type B-like)	4.1361864	6.1591491	26.1613437	6.68E-18	2.67E-14	28.884634
2911	MYL10	myosin, myosin X	3.4767611	6.5440051	19.3071076	4.01E-15	1.48E-12	25.0122806
2480	CHRF7	G protein-coupled receptor (metabotropic receptor type B-like)	4.1361864	6.1591491	26.1613437	6.68E-18	2.67E-14	28.884634
6442	CHRF7	G protein-coupled receptor (metabotropic receptor type B-like)	4.1361864	6.1591491	26.1613437	6.68E-18	2.67E-14	28.884634
2000	WDR1	WD repeat domain 1	2.4290641	6.8109161	16.8104416	1.37E-14	4.58E-12	23.113554
1171	SH3GLB1	SH3 domain containing 1	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
8421	SH3GLB1	SH3 domain containing 1	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
8324	PDZ	PDZ domain containing 1	2.4089143	8.4162213	19.8091779	2.13E-15	3.20E-12	25.27998
2131	SYND1	syndecan 1	3.3821509	6.4281317	22.4028514	4.65E-16	3.99E-13	27.897587
1071	NET1	neuronal protein 1	2.1111291	6.0601761	19.8112021	1.14E-14	4.39E-12	23.719146
1141	KUFC	kinase, protein kinase C	4.5177211	7.2762721	19.4517651	3.09E-15	1.48E-12	25.0122806



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) is a standardized, controlled vocabulary system to classify gene function across three structured domains:

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 to 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*.

The Gene Ontology (GO)

Input: A discrete, thresholded list of Differentially Expressed Genes (DEGs) (e.g., $p_{\text{adj}} < 0.05$ AND $|\text{Log}_2 \text{FC}| \geq 1$).

Statistical Test: Evaluates whether a GO term is enriched in your DEG list compared to the genome background using Fisher's Exact Test or Hypergeometric Test.

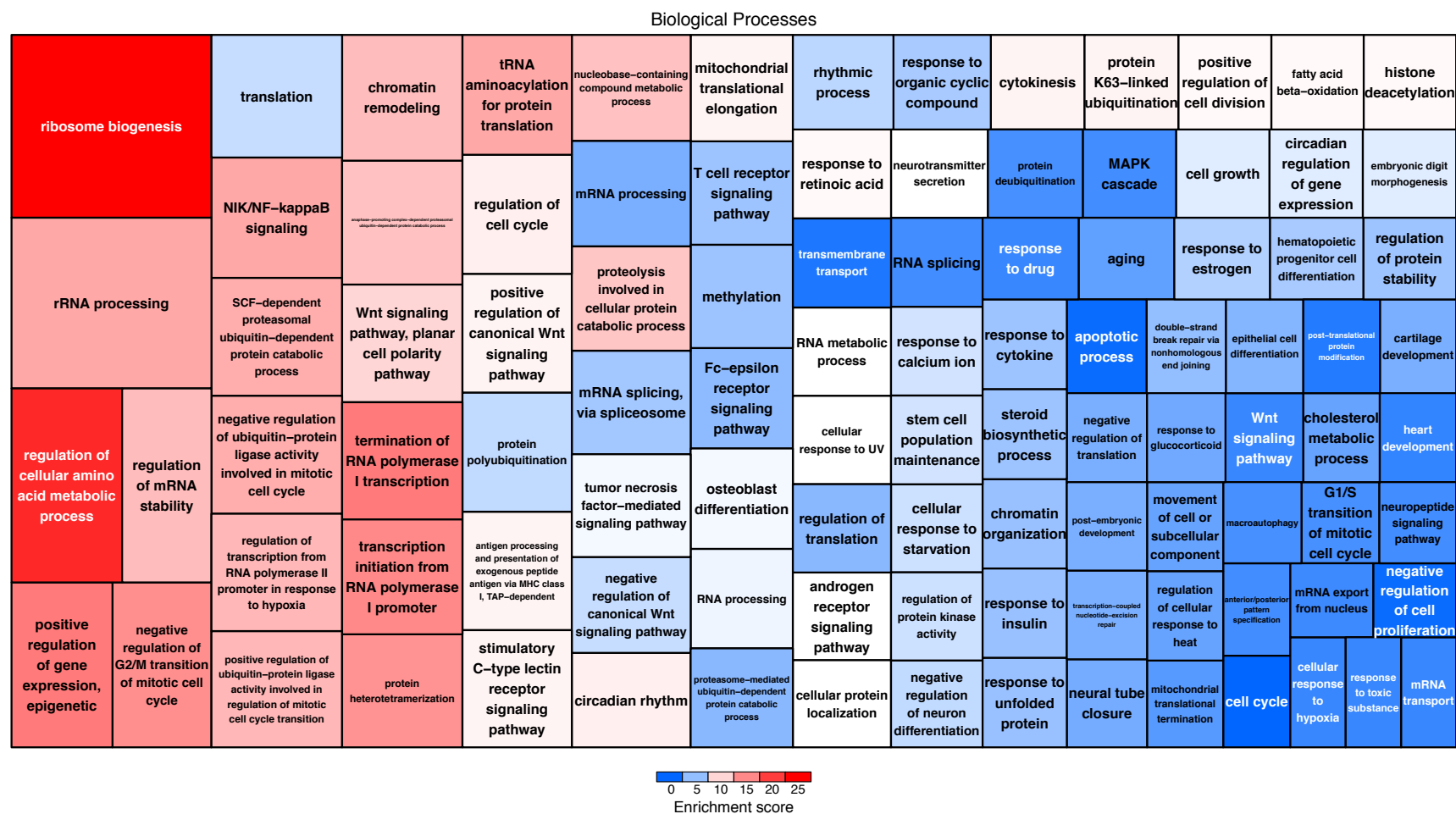
Key Limitations of GO

Cut-off Dependent: Arbitrary thresholds (p_{adj} , Fold Change) exclude genes with small but biologically meaningful changes.

Ignores Directionality & Ranking: Treats all genes in the input list equally, ignoring their magnitude of change or rank order.

DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of functional categories: Gene Ontology.



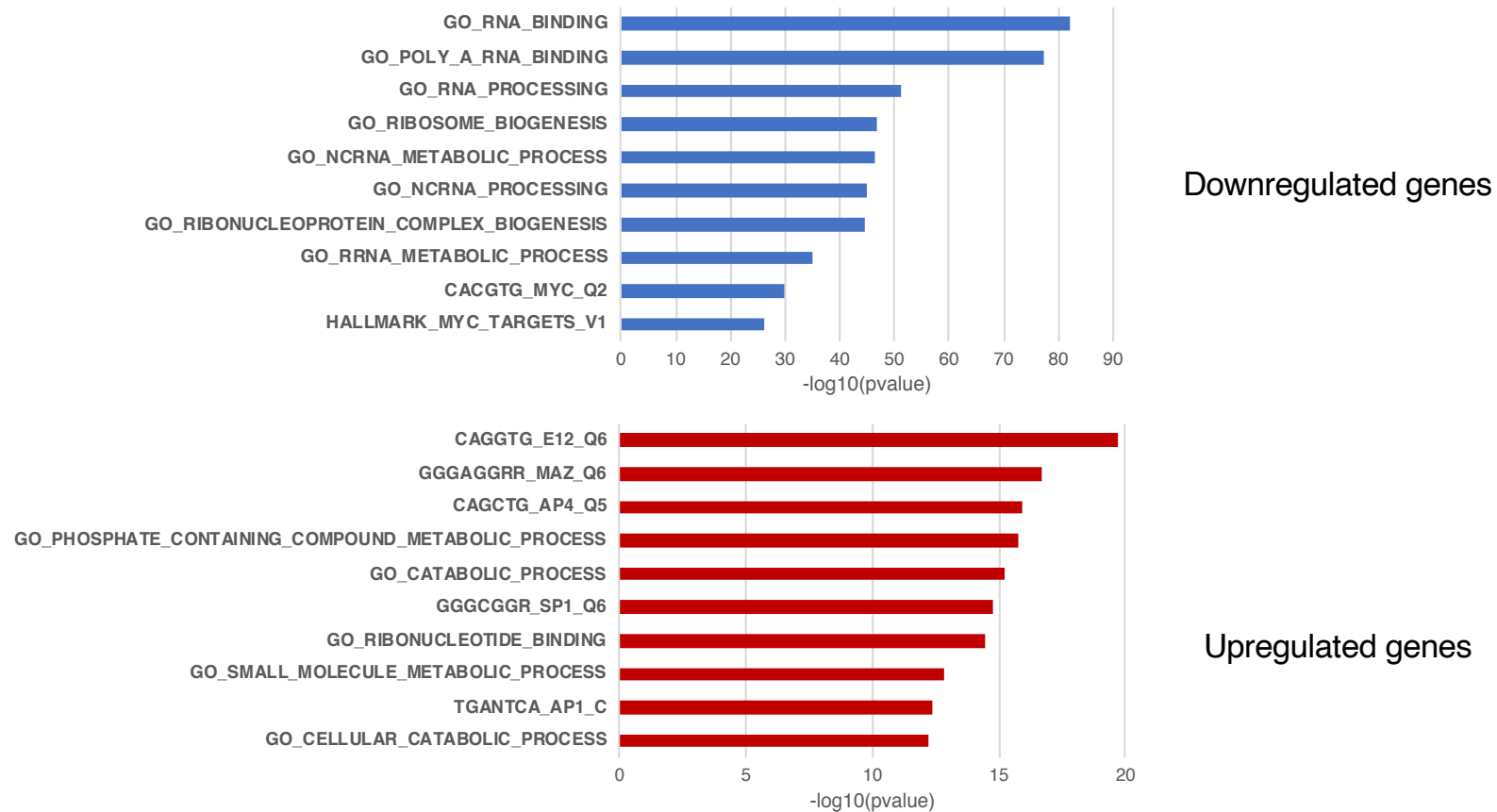
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

Example of functional categories: Gene Ontology.



DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

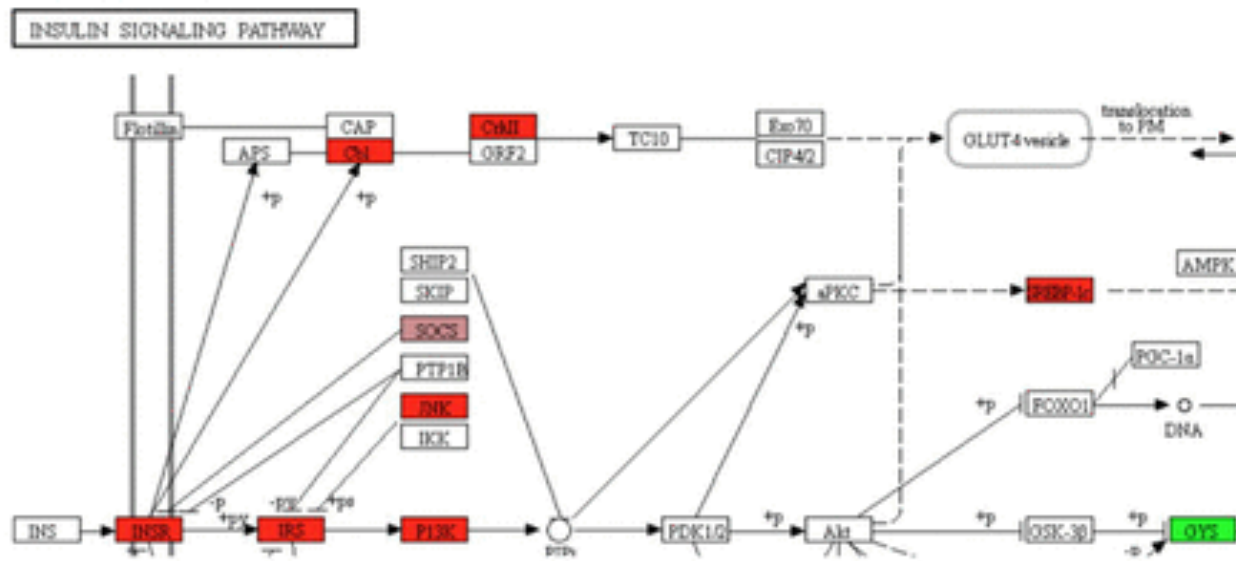
Example of functional categories: Gene Ontology.



DATA ANALYSIS: Pathway analysis

Example of functional categories: KEGG pathway

KEGG Pathway

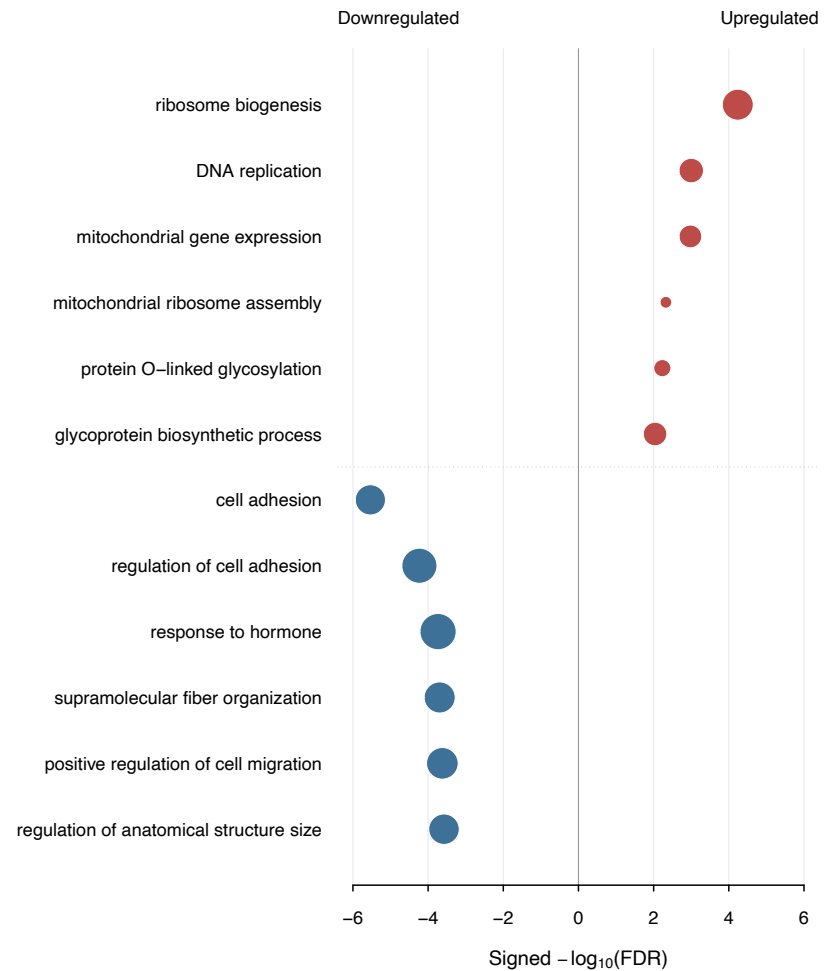


Upregulated

Downregulated

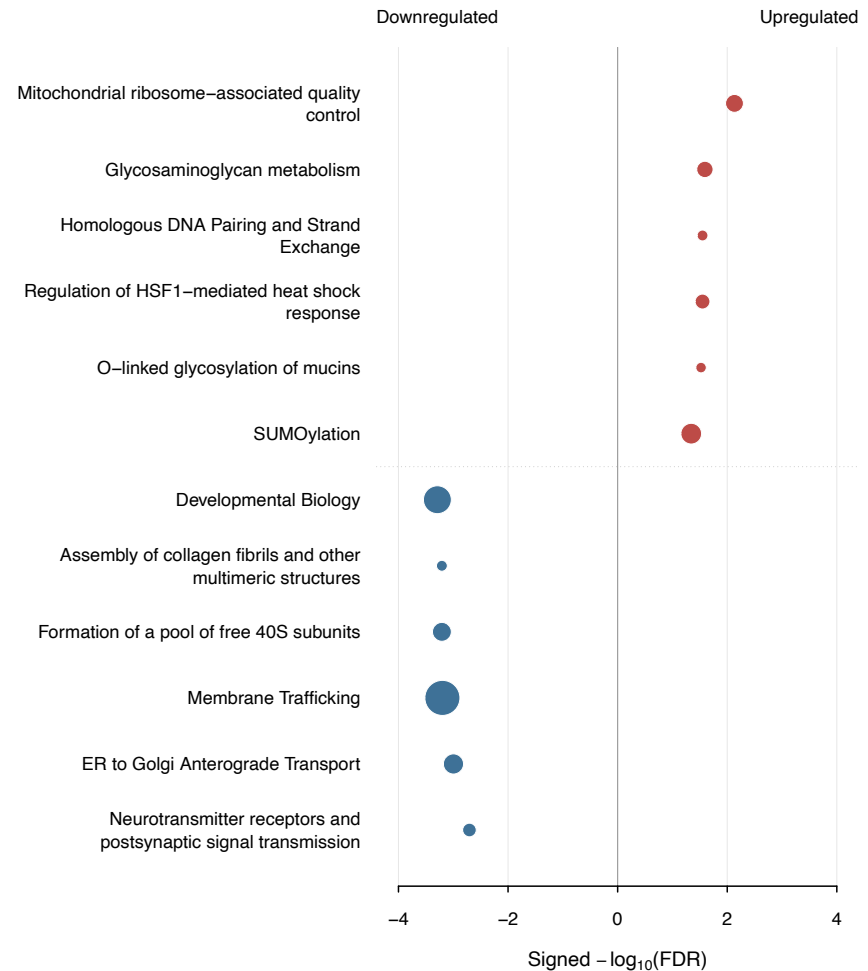
DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS

GO Biological processes



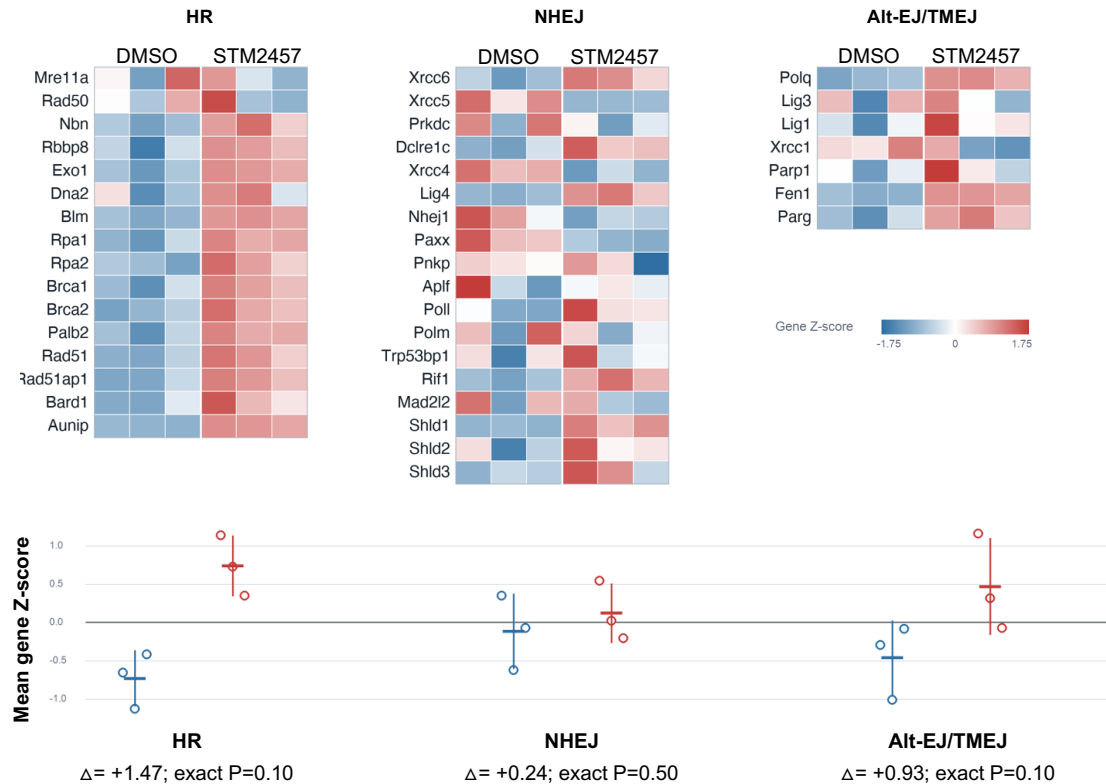
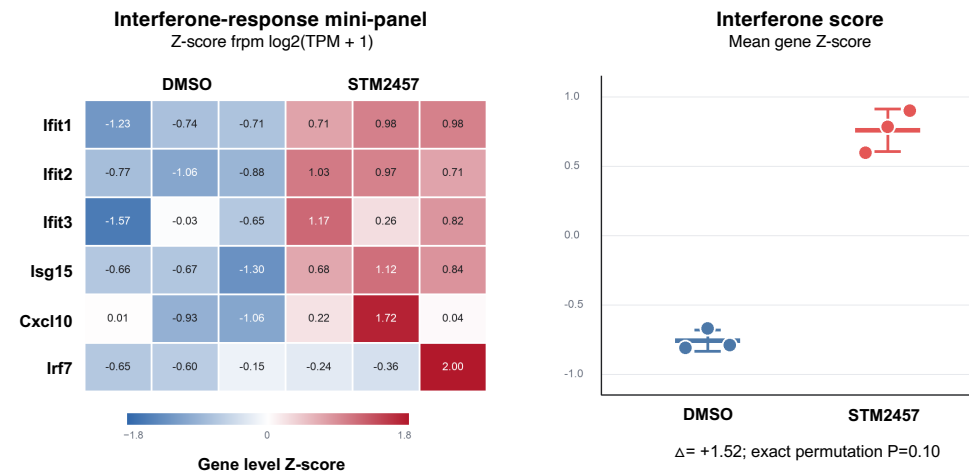
Overlapping genes ● 20 ● 60 ● 120

Reactome pathways



Overlapping genes ● 20 ● 60 ● 120

DATA ANALYSIS: FUNCTIONAL ENRICHMENT ANALYSIS



Gene Set Enrichment Analysis (GSEA)

Threshold-Free Approach: Uses all detected genes from the RNA-seq experiment - no arbitrary cut-offs.

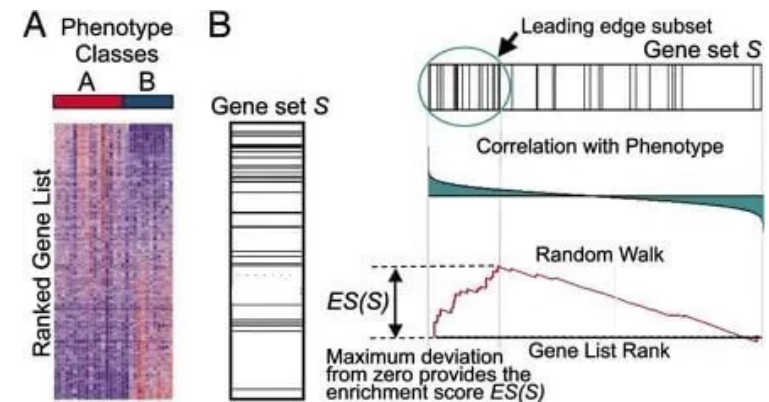
Detects Coordinated Shifts: Identifies pathways where individual genes show subtle, modest changes that together drive a massive biological response (e.g., metabolic or inflammatory cascades).

The GSEA Workflow

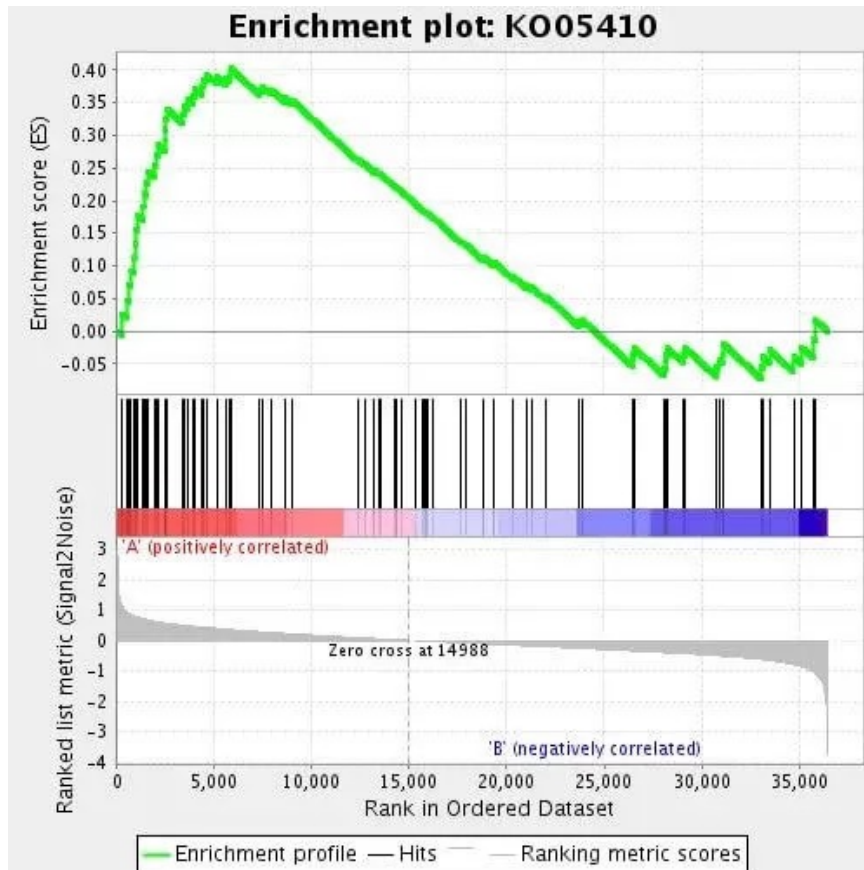
Rank All Genes: Order the entire gene list by a metric reflecting differential expression (e.g., $\text{sign}(\text{Log2 FC}) \times -\log_{10}(\text{padj})$).

Calculate Enrichment Score (ES): Walk down the ranked list, increasing the score when a gene belongs to the pathway set and decreasing it when it does not.

Statistical Evaluation: Permute labels to derive significance (p-value) and normalize scores across different gene set sizes.



Gene Set Enrichment Analysis (GSEA)



Enrichment Score (ES): The green curve reaches a positive peak ($ES \approx 0.40$) on the left, demonstrating significant pathway enrichment.

Gene Set Hits: Vertical black lines are densely clustered at the top of the ranking, showing coordinated up-regulation in Phenotype A.

Ranked List Metric: All $\sim 36,000$ genes are ranked by Signal-to-Noise ratio, transitioning from Phenotype A (red) to Phenotype B (blue).

In summary, pathway KO05410 is strongly activated/up-regulated in Phenotype A compared to Phenotype B

GO *vs* GSEA

Feature	GO	GSEA
Input Data	List of top DEGs (e.g., 200 genes passing threshold)	Entire ranked dataset (e.g., 18,000 genes)
Cut-off dependency	High (sensitive to threshold selection)	None (threshold-free)
Biological Sensitivity	Detects strong single-gene drivers	Detects small, coordinated pathway shifts

Both GO and GSEA can utilize the exact same functional databases (GO_BP, KEGG, Reactome etc.); they differ solely in their statistical approach to testing pathway enrichment.