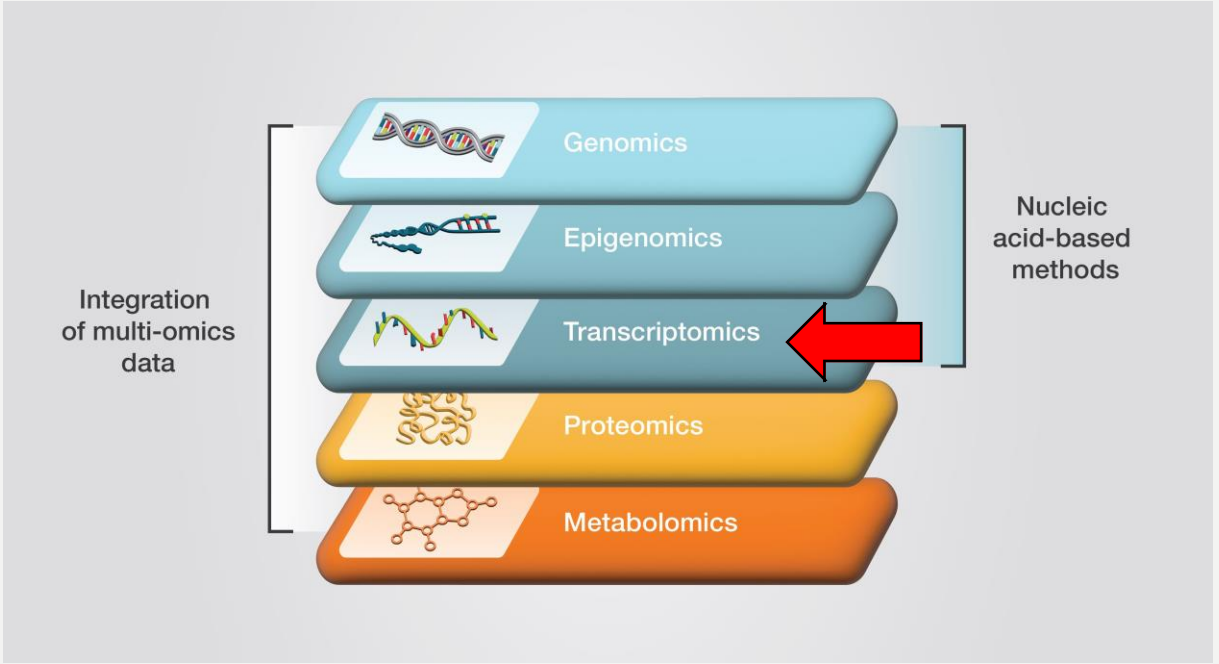
RNA-SEQUENCING

FROM RNA WITH LOVE: EXPLORING THE TRANSCRIPTOME

OMICS APPROACHES FOR THE STUDY OF BIOLOGICAL PROCESSES

With the advent of increasingly advanced technologies, such as **Next Generation Sequencing** (**NGS**), it has become possible to study biological processes in far greater depth and detail than ever before. These innovative tools have paved the way for the development of **–omics**, which enable a comprehensive and a deeper analysis of the cellular processes.

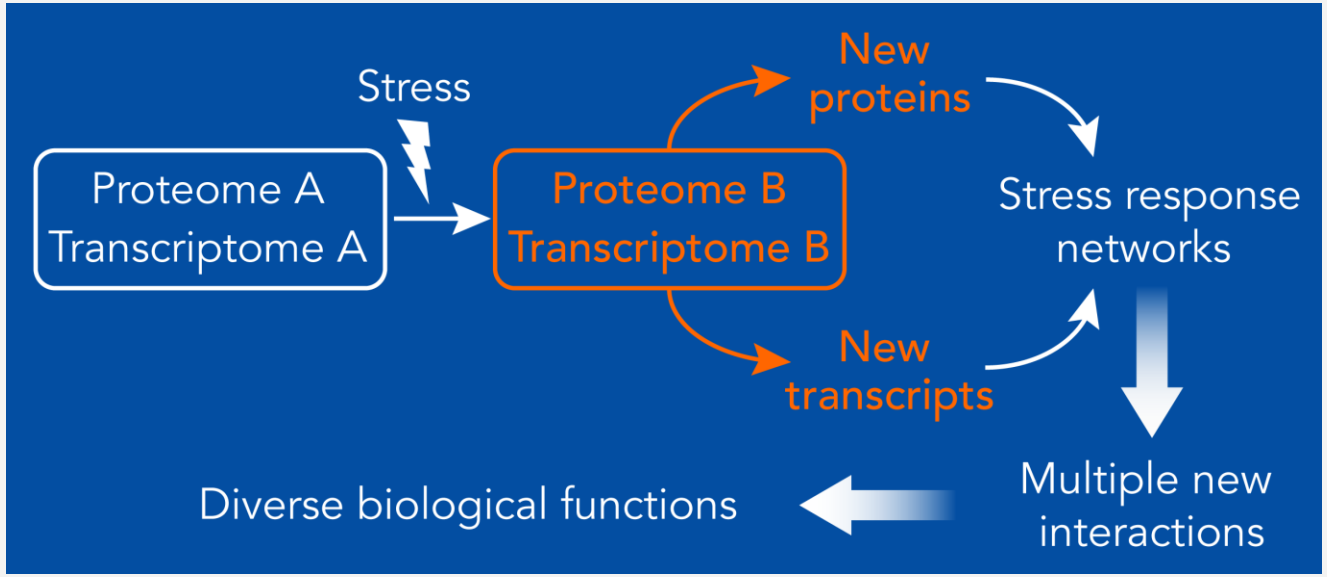


THE TRANSCRIPTOME: A SNAPSHOT OF GENE EXPRESSION

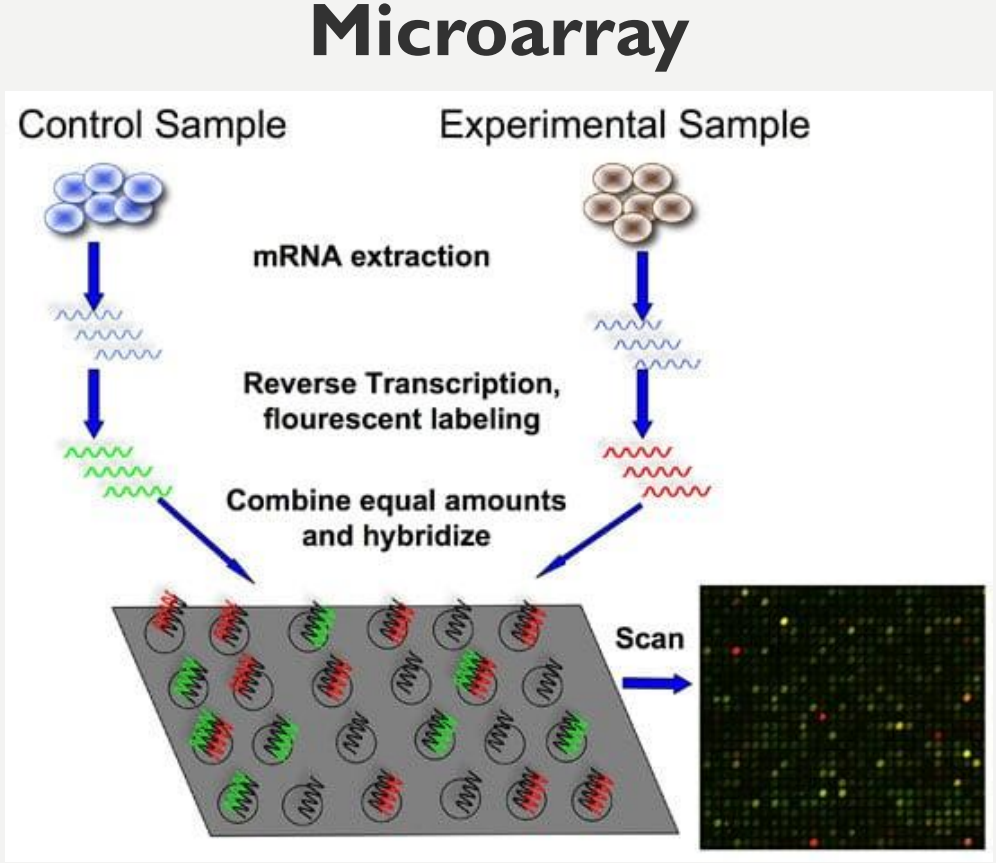
- The sum of all RNAs present in a cell is called **Transcriptome.**
- The **Transcriptome** of a cell is a **dynamic entity**: unlike the **Genome**, it constantly changes.
- **Understanding the transcriptome** is essential for interpreting **how cells communicate** or how cells **respond to external stimuli**.

THE TRANSCRIPTOME: A SNAPSHOT OF GENE EXPRESSION

- The sum of all RNAs present in a cell is called **Transcriptome.**
- The **Transcriptome** of a cell is a **dynamic entity**: unlike the **Genome**, it constantly changes.
- **Understanding the transcriptome** is essential for interpreting **how cells communicate** or how cells **respond to external stimuli**.



MICROARRAYS: THE FIRST STEP TOWARD TRANSCRIPTOMIC ANALYSIS



RNA-SEQ: HIGH-THROUGHPUT SEQUENCING FOR TRANSCRIPTOME PROFILING

- RNA-seq is essentially **massively parallel sequencing of RNA (in particular cDNA)**.
- It is based on **next-generation sequencing (NGS) platforms.**
- The introduction of RNA-seq has provided the ability to look at:

1. Transcriptional structure of genes
2. Alternatively spliced transcripts, alternative promoters and polyA sites
3. Post-transcriptional modifications
4. Identifying and studying all species of transcripts.
5. Changes in gene expression under different conditions

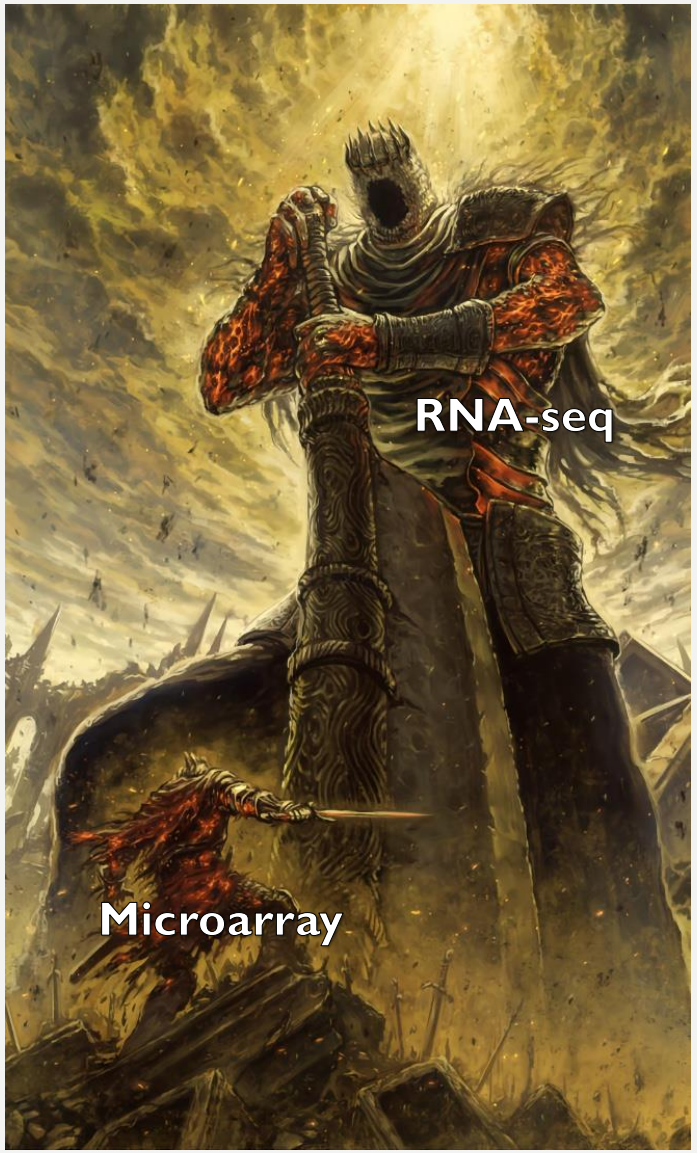
RNA-Seq VS Microarray

- **RNA-seq** has a **wider dynamic range.**
- **RNA-seq** is **more sensitive and more specific** than microarray.
- **RNA-seq** is capable of **detecting single nucleotide polymorphisms (SNPs).**
- **RNA-seq** is able to **identify and quantify novel splicing isoforms.**
- **RNA-seq** can allow the **identification of rare or low-abundance transcripts**.

RNA-SEQ: HIGH-THROUGHPUT SEQUENCING FOR TRANSCRIPTOME PROFILING

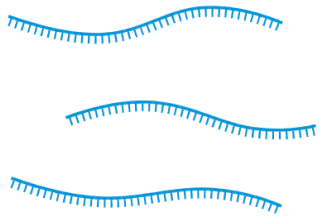
- RNA-seq is essentially **massively parallel sequencing of RNA (in particular cDNA)**.
- It is based on **next-generation sequencing (NGS) platforms.**
- The introduction of RNA-seq has provided the ability to look at:

1. Transcriptional structure of genes
2. Alternatively spliced transcripts, alternative promoters and polyA sites
3. Post-transcriptional modifications
4. Identifying and studying all species of transcripts.
5. Changes in gene expression under different conditions

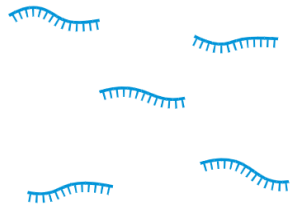


THE METHOD

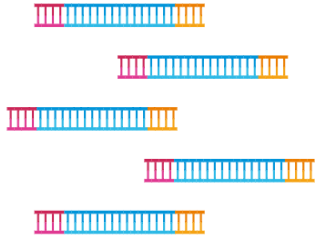
1 **Isolate RNA from samples**



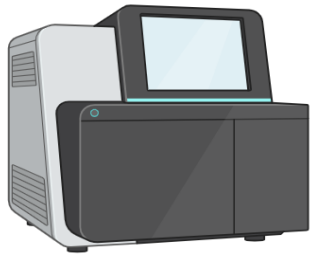
2 **Fragment RNA into short segments**



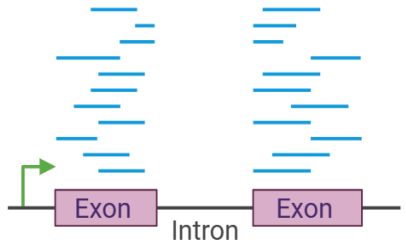
3 **Library preparation**



4 **Perform NGS sequencing**

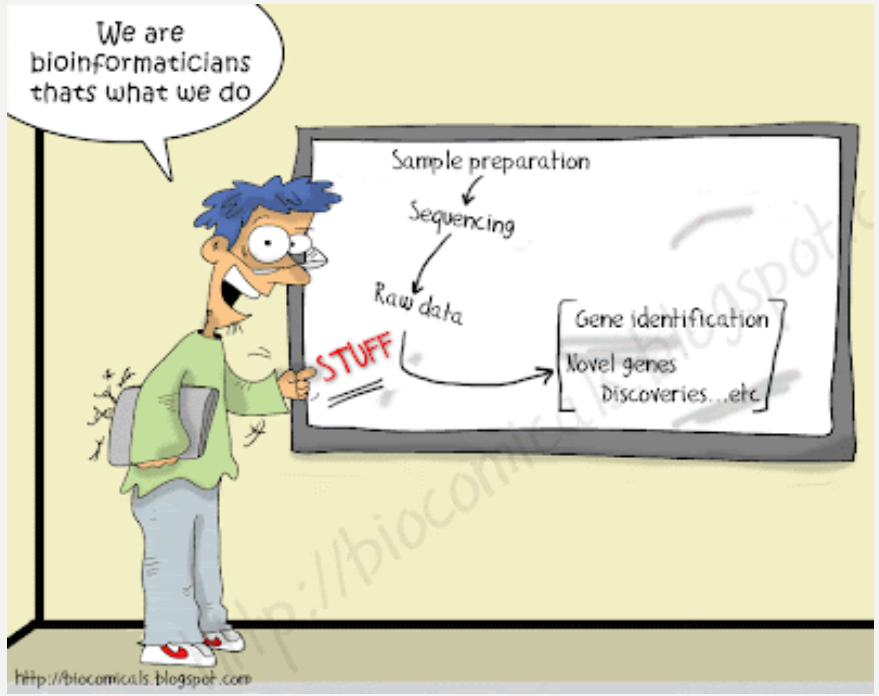


5 **Map sequencing reads to the transcriptome/genome**



A detailed knowledge of each step of RNA-seq protocol and the linked biases is essential for:

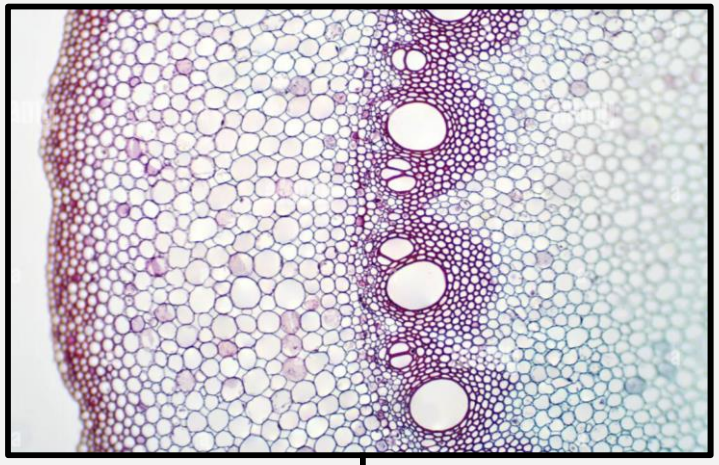
- The correct design of the experiment
- A careful interpretation of NGS data,
- Finding ways to improve library quality
- Developing bioinformatics tools to compensate for the biases



1. ISOLATE RNA FROM SAMPLES

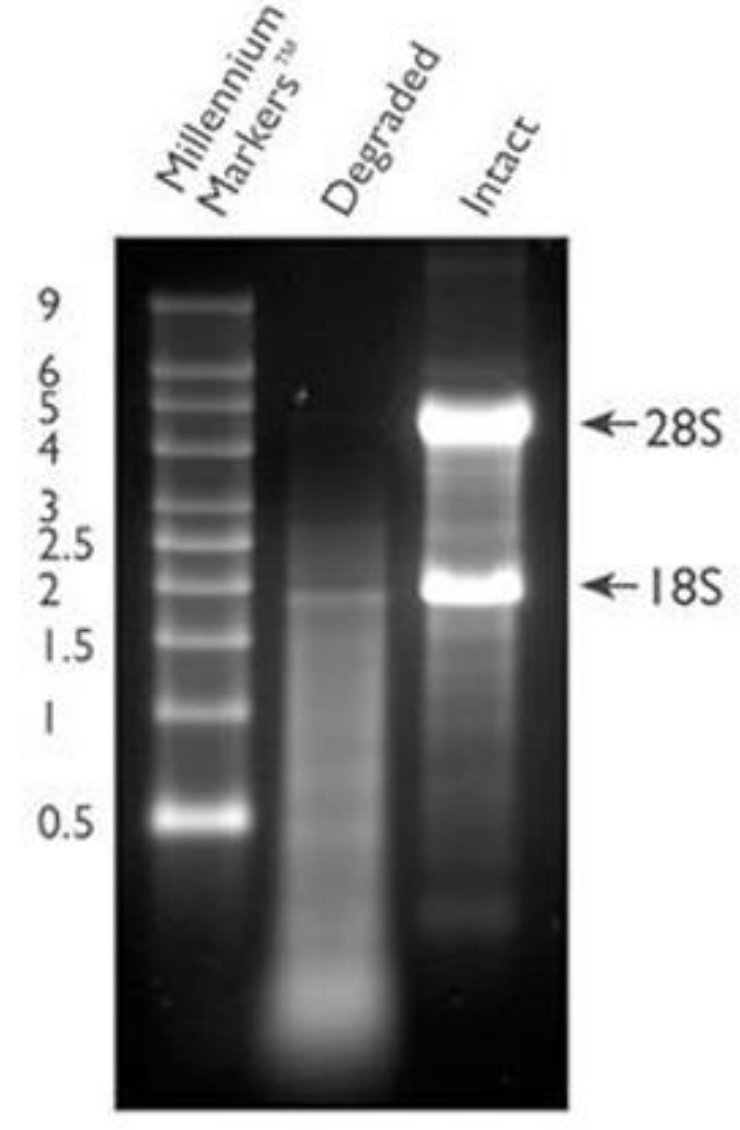
Cell Culture

Tissue



Whole Organism

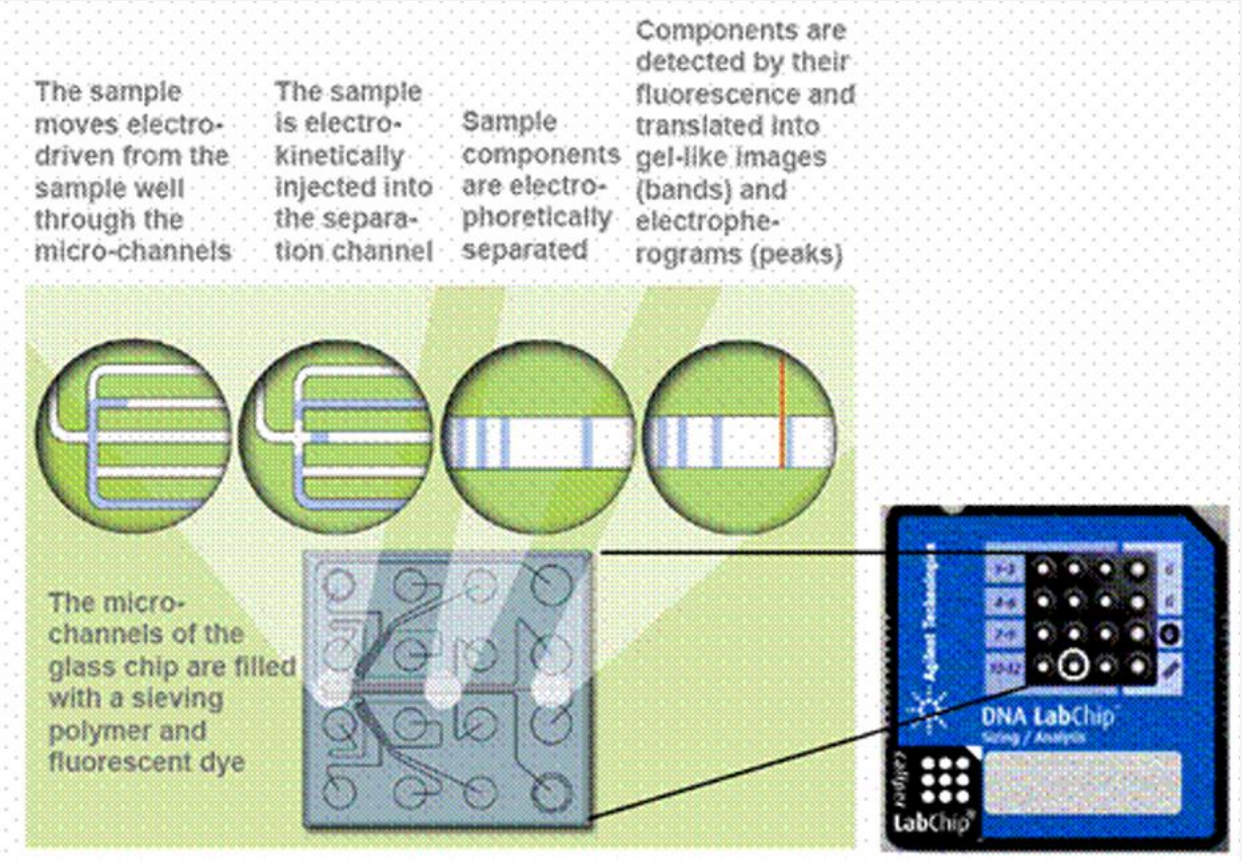
1. ISOLATE RNA FROM SAMPLES



- Intact total RNA run on a denaturing gel will have sharp, clear 28S and 18S rRNA bands
- The 28S rRNA band should be approximately twice as intense as the 18S rRNA band

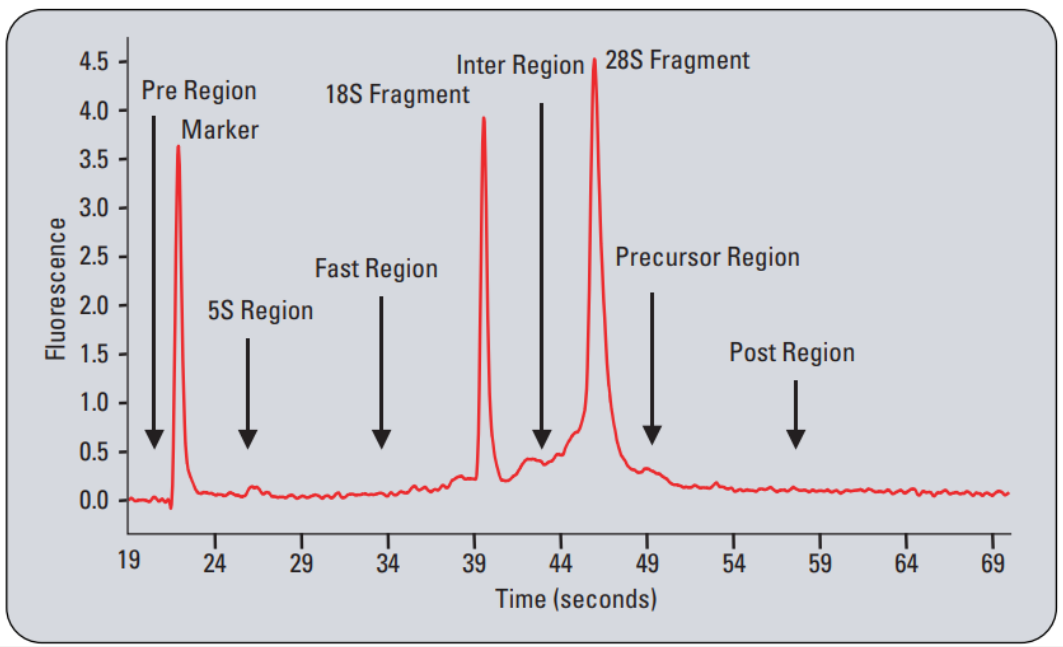
1. ISOLATE RNA FROM SAMPLES

Bioanalyzer



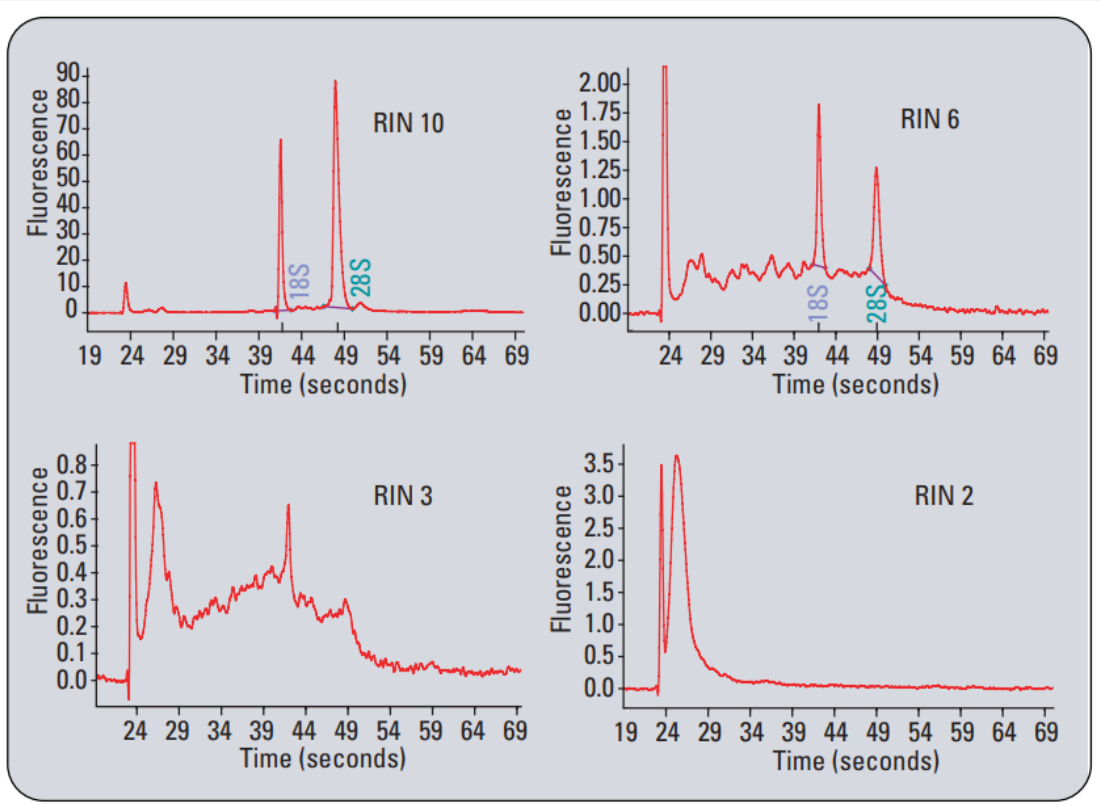
Samples are combined with a **fluorescent dye** and injected into wells in the chip. The samples move through a gel matrix in the microchannels and are separated by **electrophoresis**. The samples then are detected by fluorescence, and electropherograms and gel-like images are created by the data analysis software for sizing and quantification.

1. ISOLATE RNA FROM SAMPLES

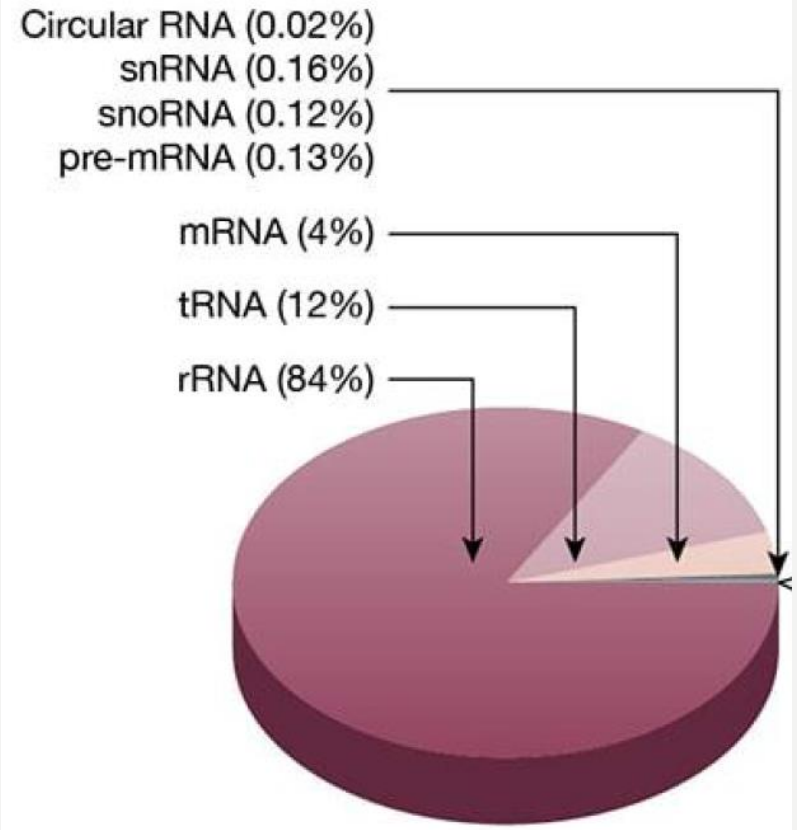


Bioanalyzer measures RNA integrity, displayed as the **RNA Integrity Number (RIN).**

To determine the RIN, the instrument software uses an algorithm that takes into account the entire electrophoretic trace of the RNA, not just the ratio of 28S and 18S rRNAs.



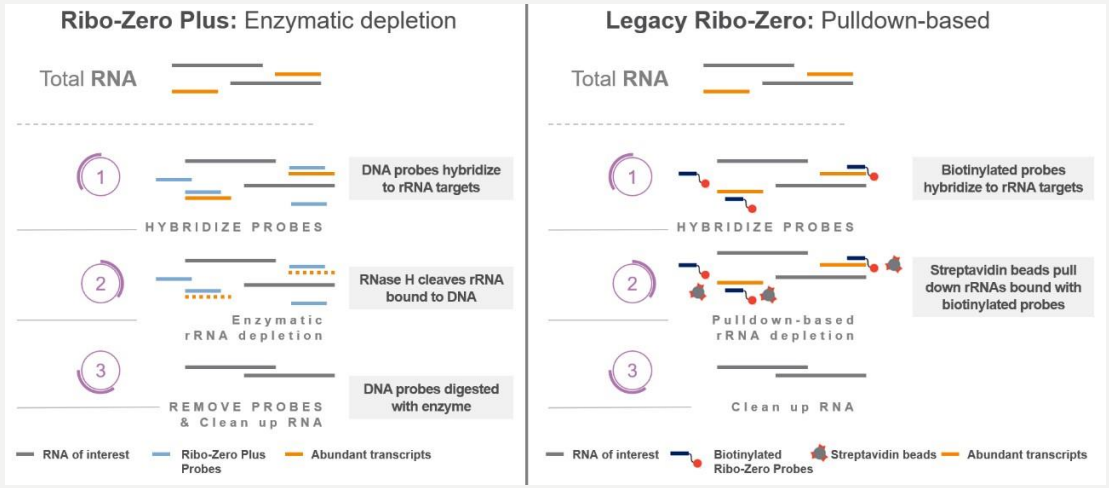
1. ISOLATE RNA FROM SAMPLES



mRNA
lncRNA
Pseudogenes
circRNA
tRNA
snoRNA
miRNA
piRNA
Etc.

1. ISOLATE RNA FROM SAMPLES

Ribo-zero

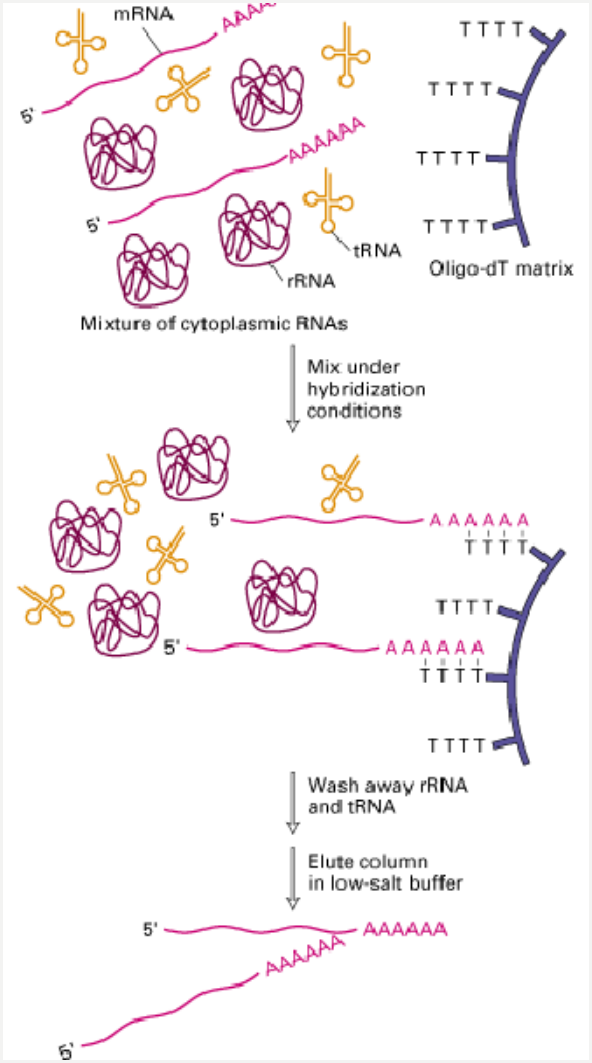


mRNA
lncRNA
Pseudogenes
circRNA
tRNA
snoRNA
miRNA
piRNA
Etc.

→ Total RNA-seq

1. ISOLATE RNA FROM SAMPLES

Poly(A) affinity selection

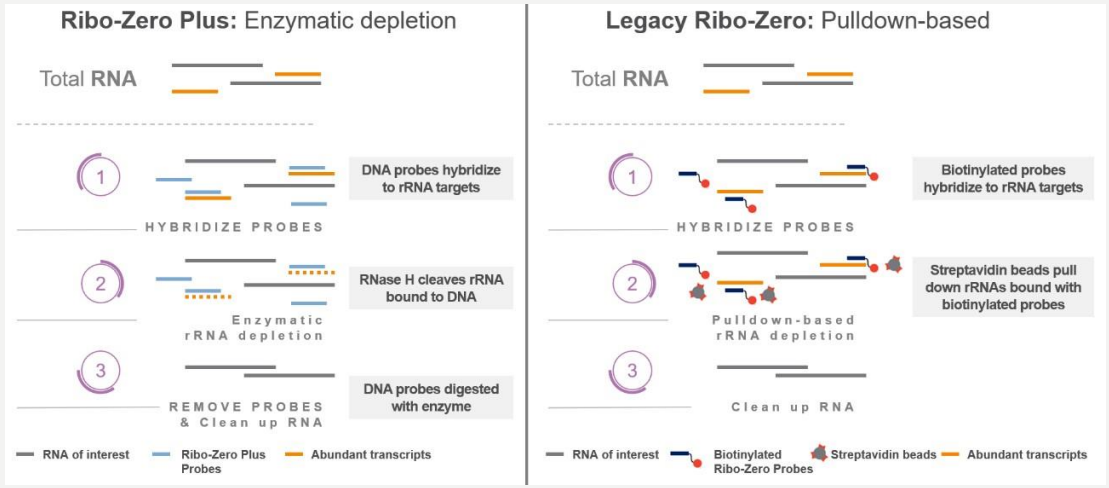


mRNA
lncRNA
Pseudogenes
circRNA
tRNA
snoRNA
miRNA
piRNA
Etc.

} **Poly(A) RNA-seq**

1. ISOLATE RNA FROM SAMPLES

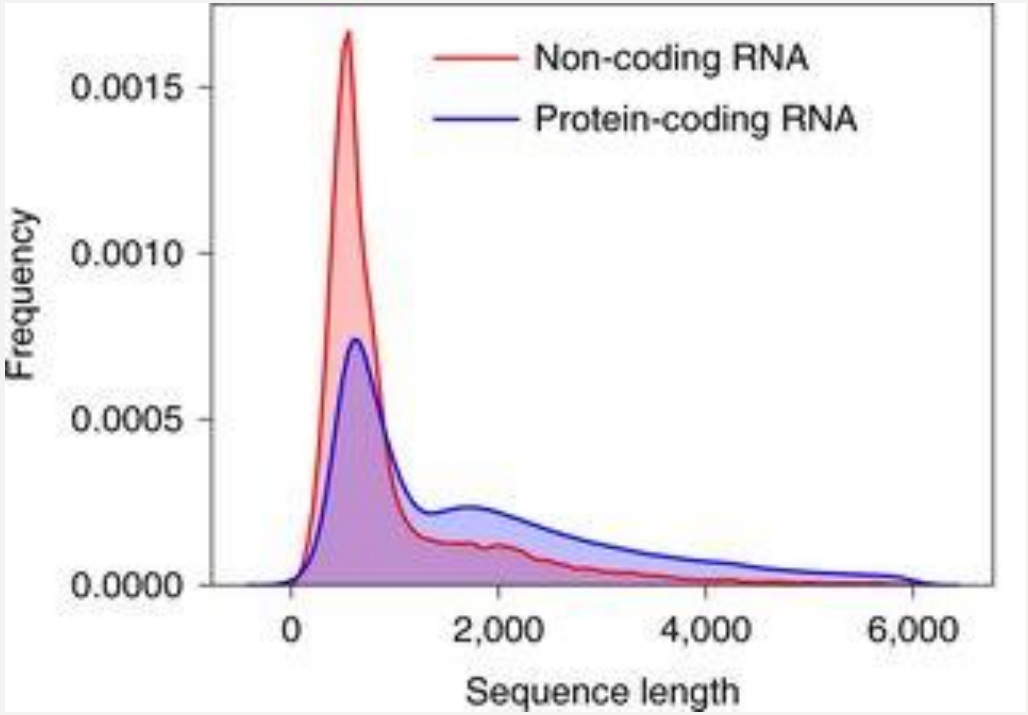
Ribo-zero



mRNA
lncRNA
Pseudogenes
circRNA
tRNA
snoRNA
miRNA
piRNA
Etc.

} **Small RNA-seq**

2. FRAGMENT RNA INTO SHORT SEGMENT



Is useful to fragment the RNAs since the **transcript length is not compatible with the Maximum Read Length of most current sequencing platforms**.

Physical Methods

- Acoustic shearing
- Sonification
- Hydrodynamic shear

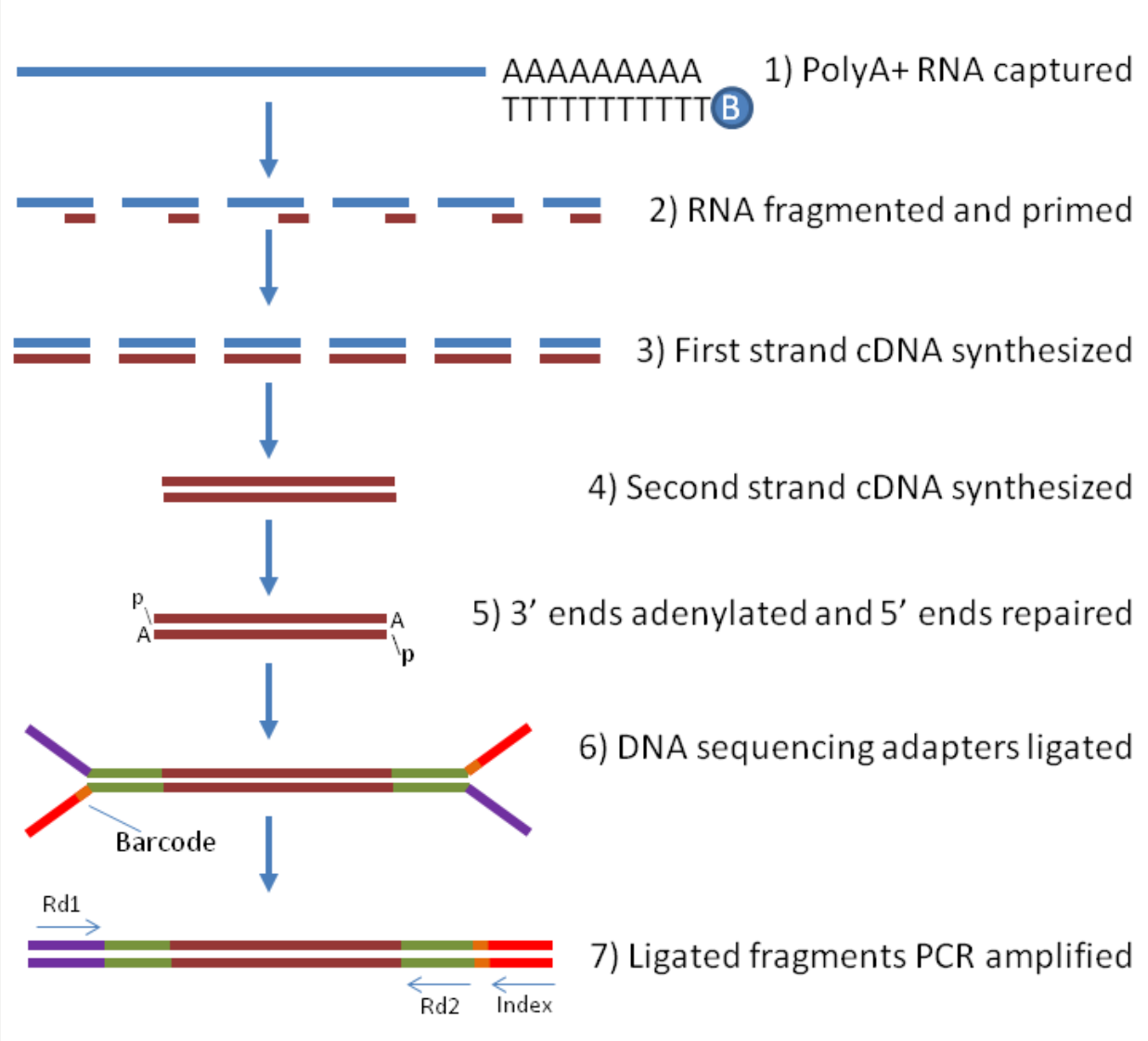
Chemical Methods

- Alkaline solutions
- Solutions with divalent cations

Enzymatic Methods

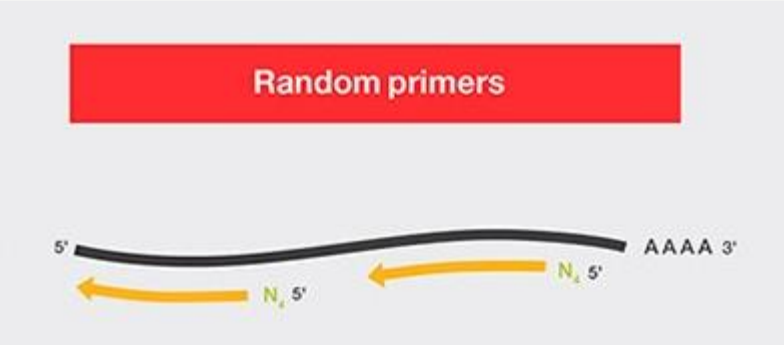
- RNAse III

3. LIBRARY PREPARATION

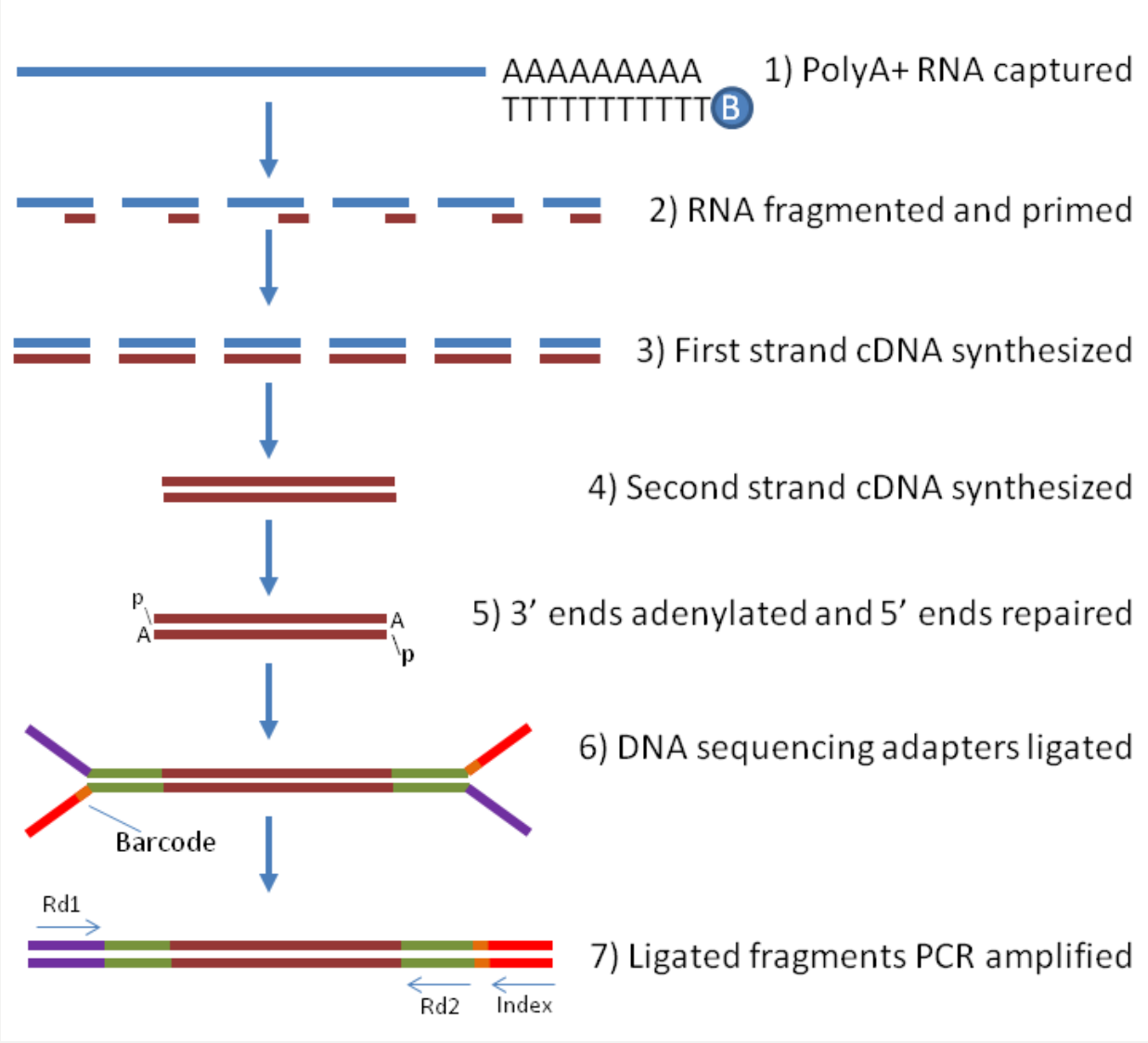


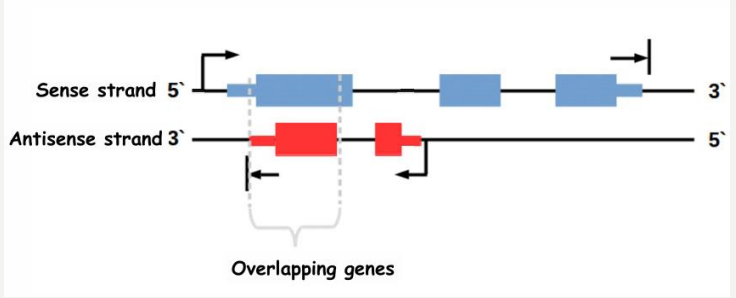
Illumina TruSeq stranded poly(A) library preparation

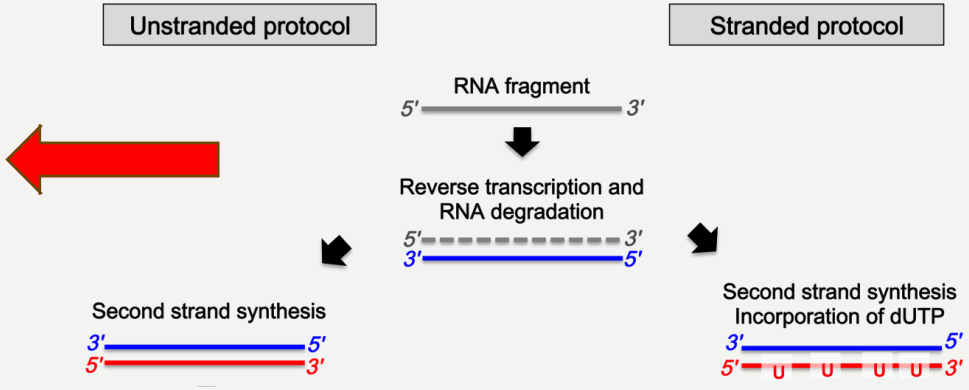
Single-stranded RNA molecules must be converted into double-stranded complementary DNAs (cDNA).



3. LIBRARY PREPARATION

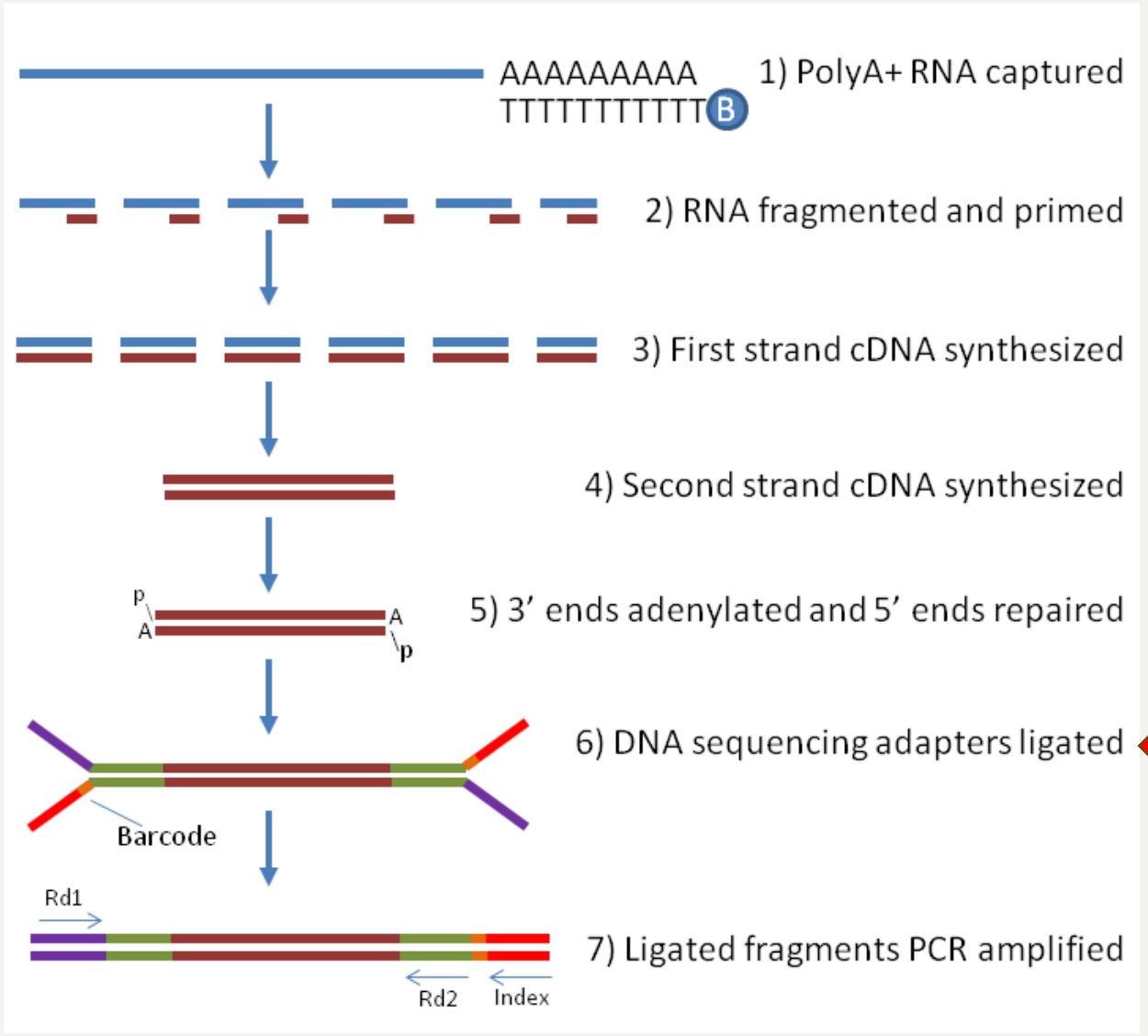


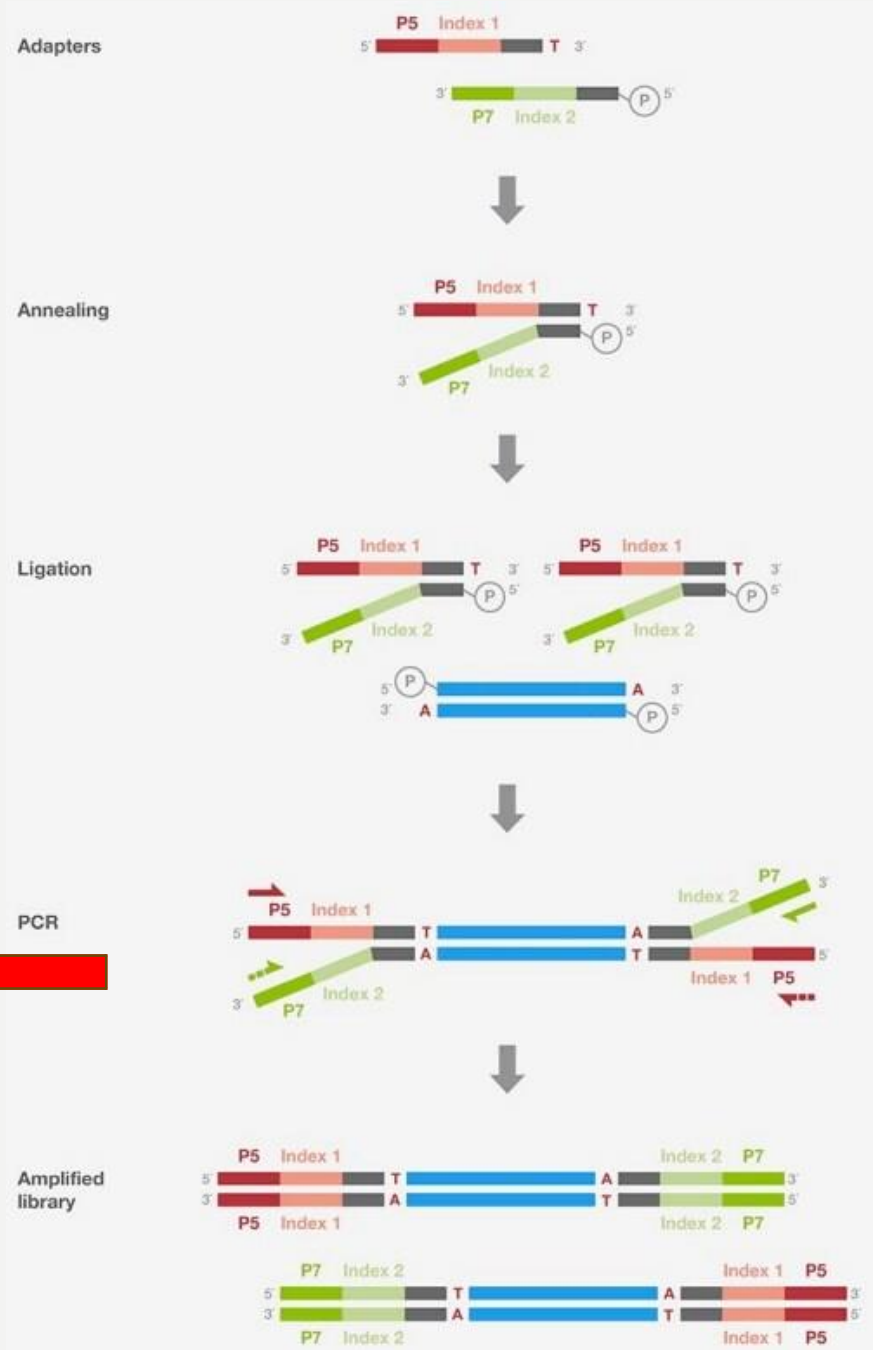




Illumina TruSeq stranded poly(A) library preparation

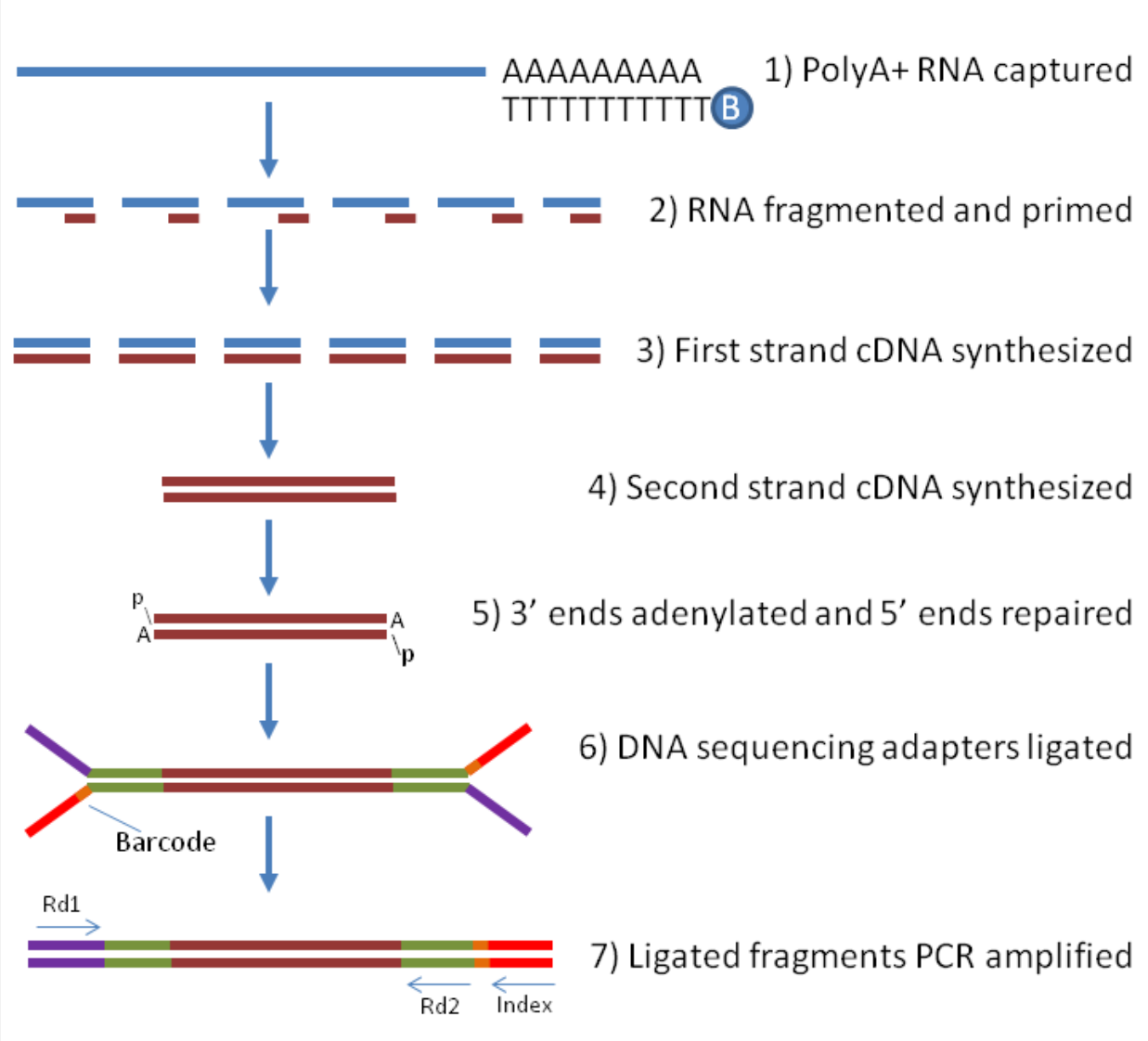
3. LIBRARY PREPARATION



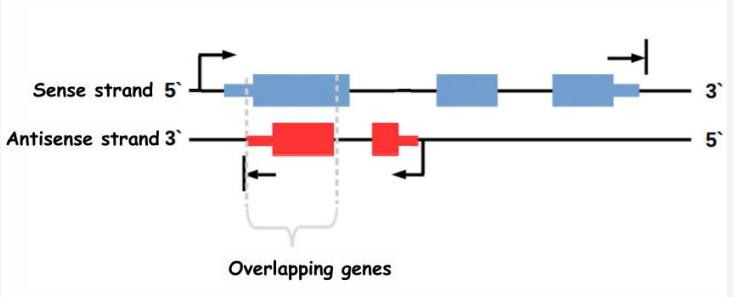


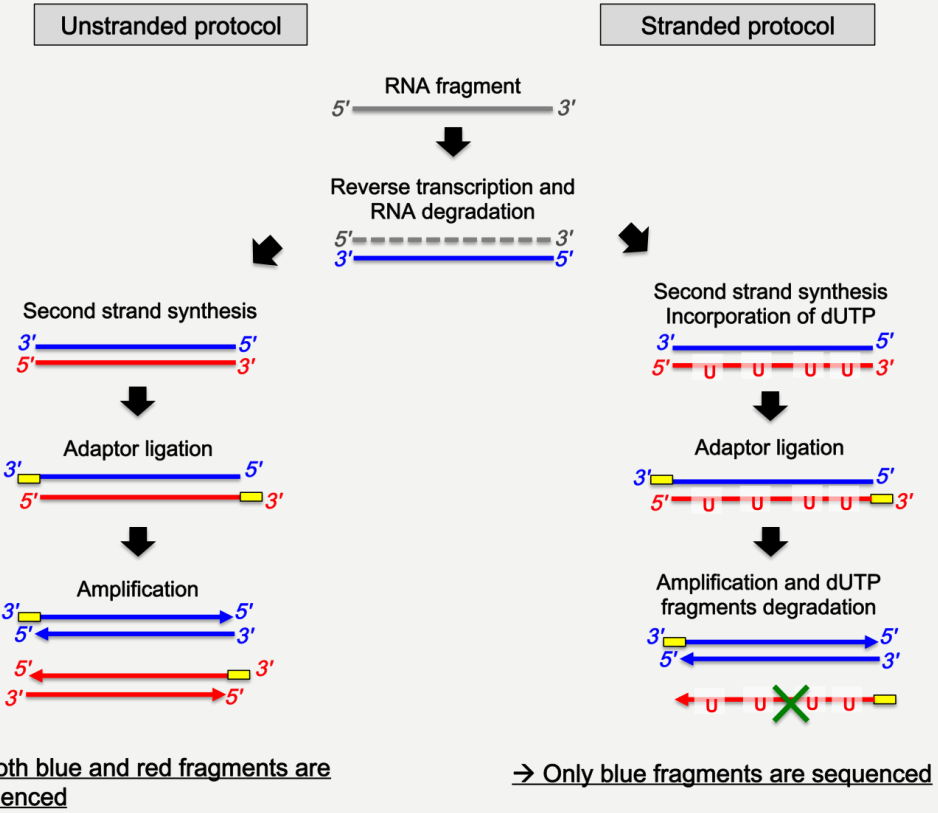
Illumina TruSeq stranded poly(A) library preparation

3. LIBRARY PREPARATION



Illumina TruSeq stranded poly(A) library preparation

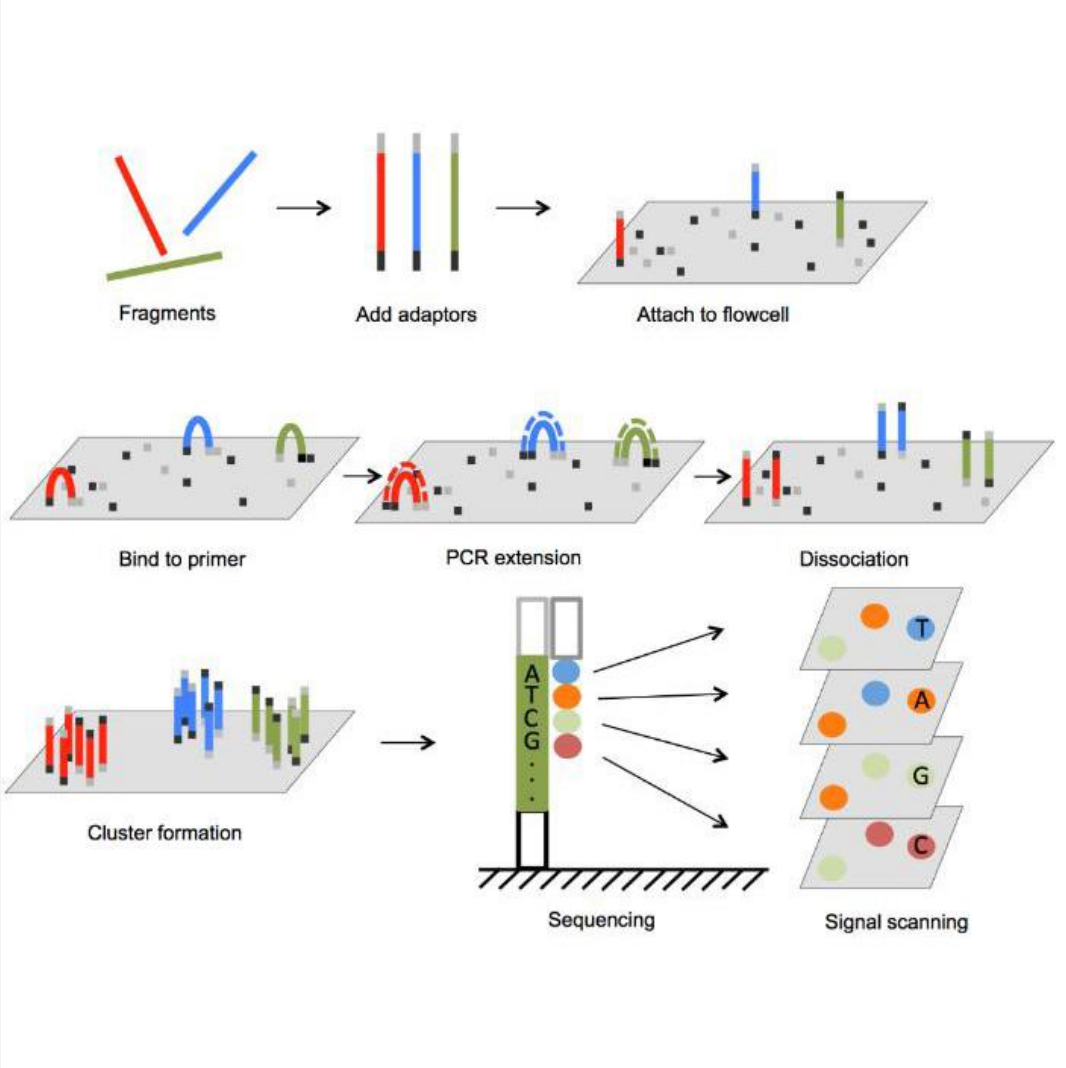


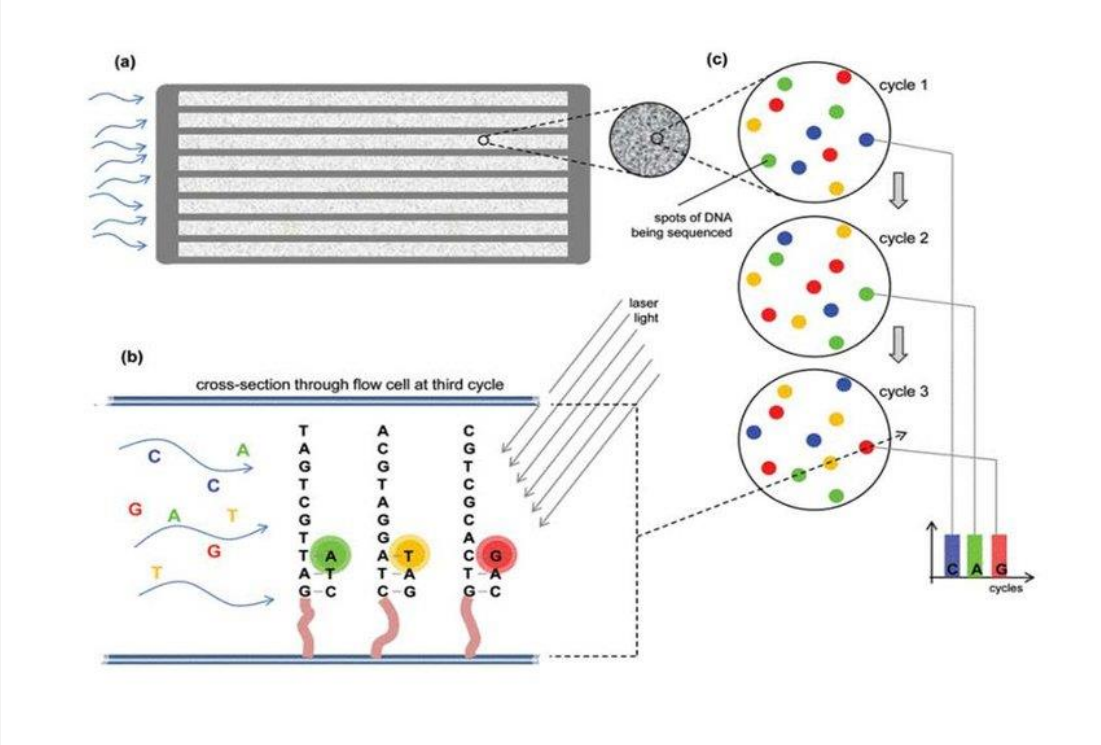


→ Both blue and red fragments are sequenced

→ Only blue fragments are sequenced

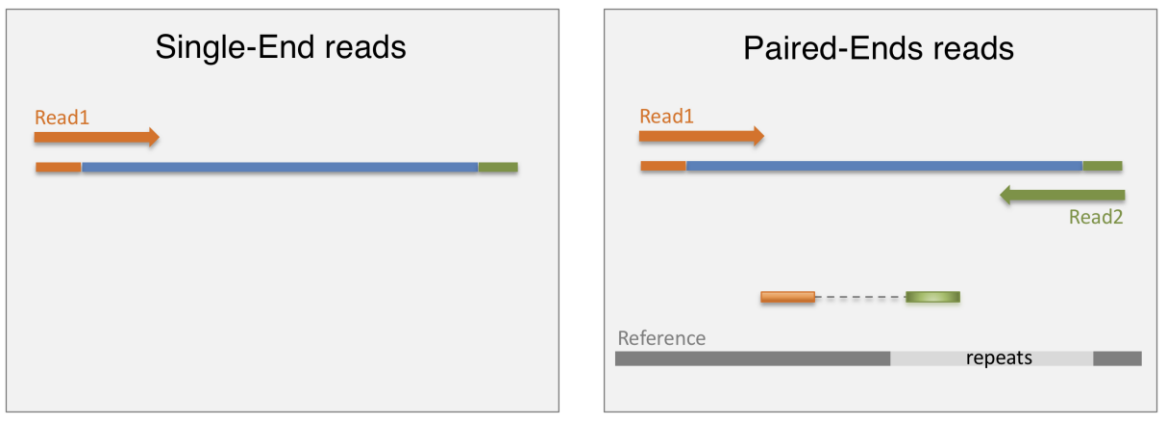
5. PERFORM NGS SEQUENCING

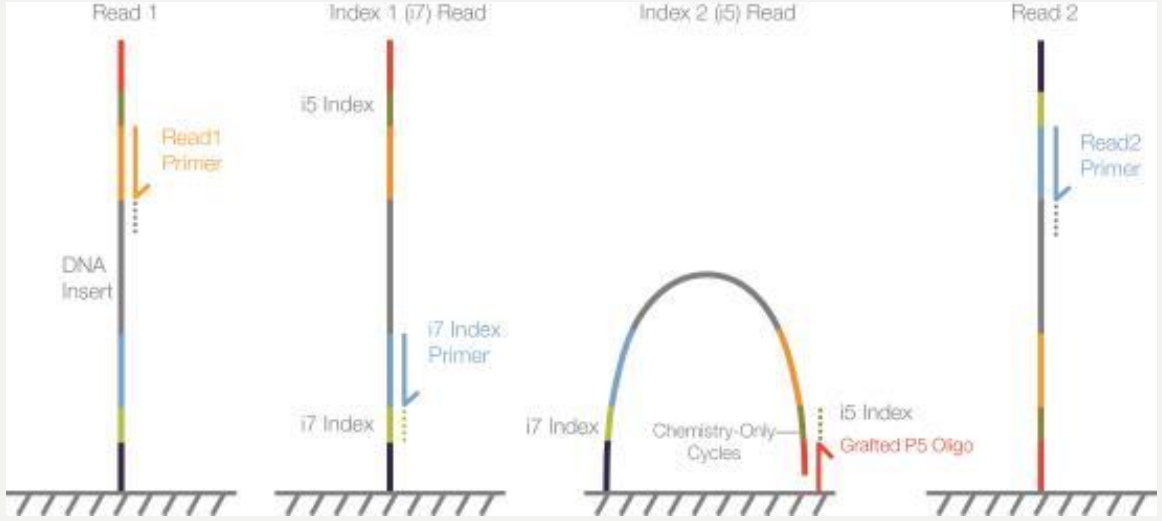




5. SINGLE-END SEQUENCING VS PAIRED END SEQUENCING

- **Single-End sequencing (SE)** : consists in sequencing the fragment from only one end
- **Paired-End sequencing (PE)** : consists in sequencing both ends of a fragment, resulting in the production of read pairs. This allows to improve the alignment, to better identify e quantify splicing variants and to detect rearrangements such as insertions, deletions and inversions



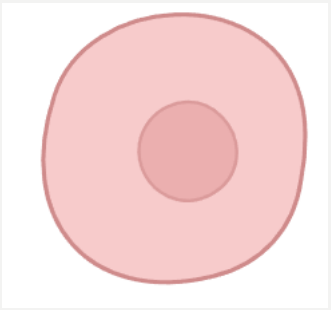


THE METHOD

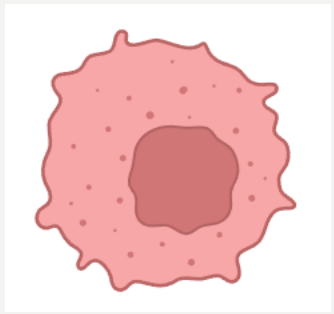
Biological Question

Mutated Gene

Cell



Cancer cell



Which genes are Upregulated or Downregulated?

Which miRNA are Upregulated or Downregulated?

Sequencing Type

mRNA
lncRNA
Pseudogenes
circRNA
tRNA
snoRNA
miRNA
piRNA
Etc.

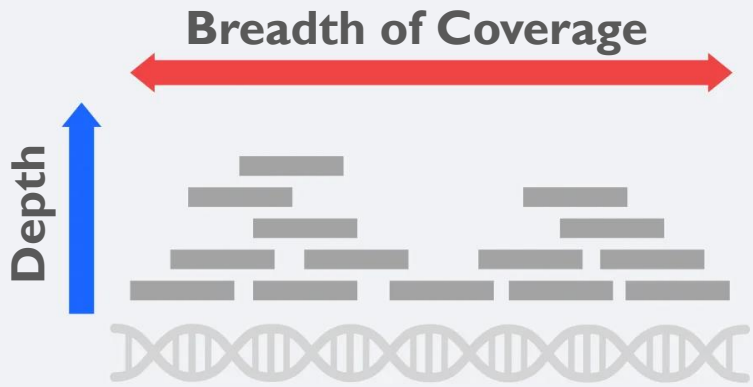
Poly(A) RNA-seq

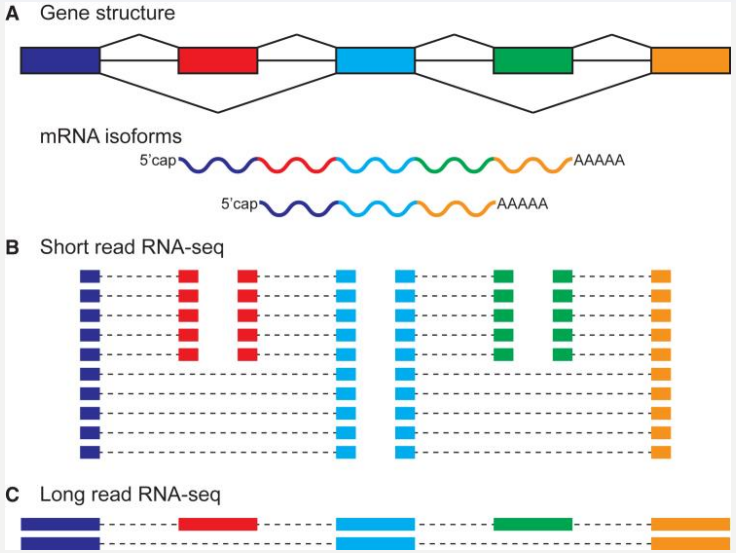
Total RNA-seq

Small RNA-seq

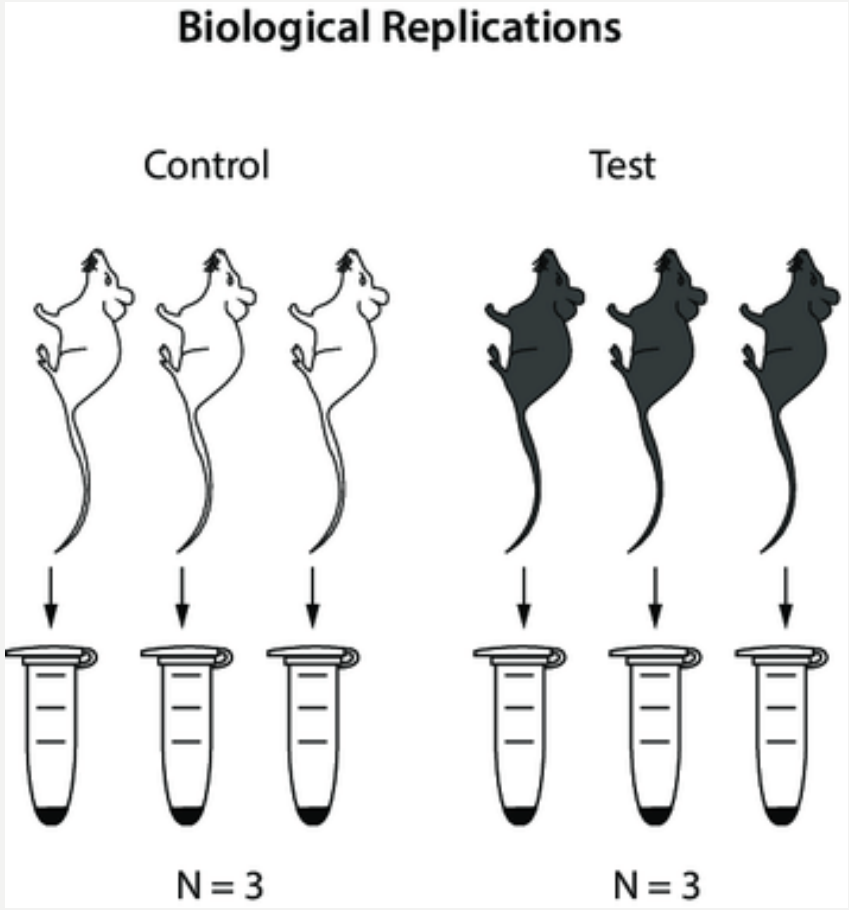
THE METHOD

Number and type of Reads



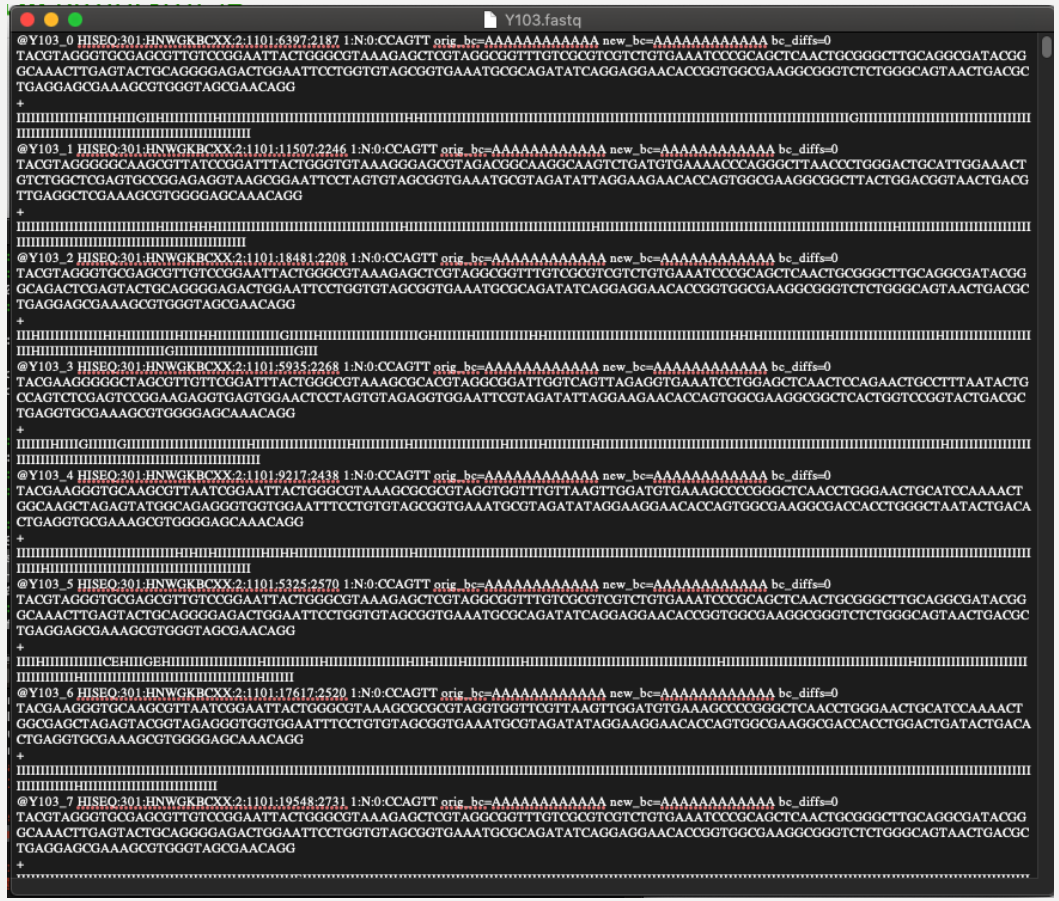


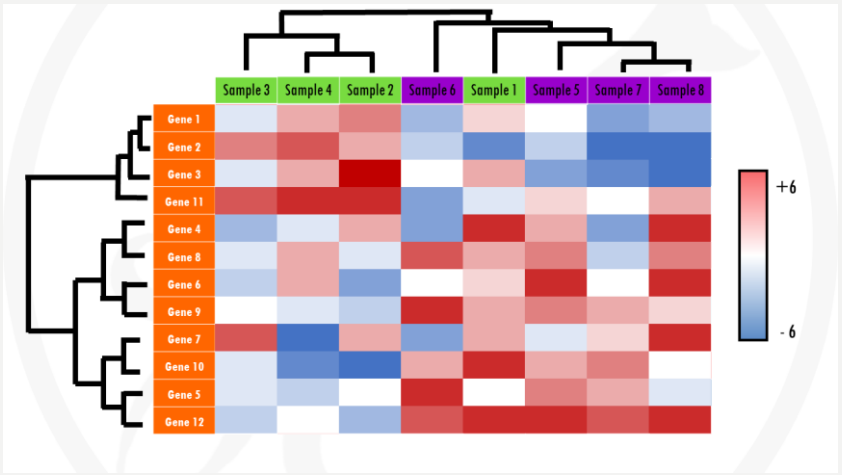
Number of replicates

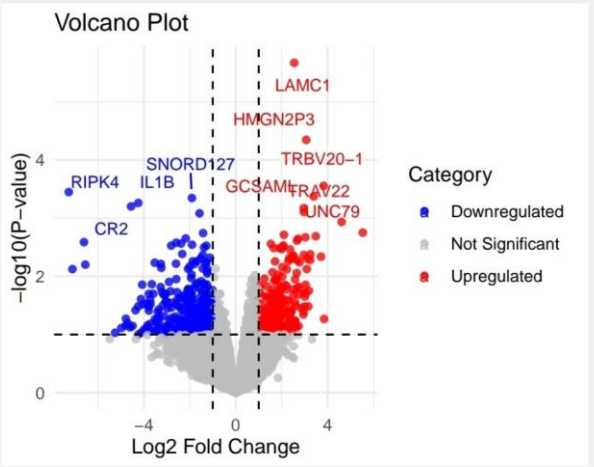


6. DATA ANALYSIS

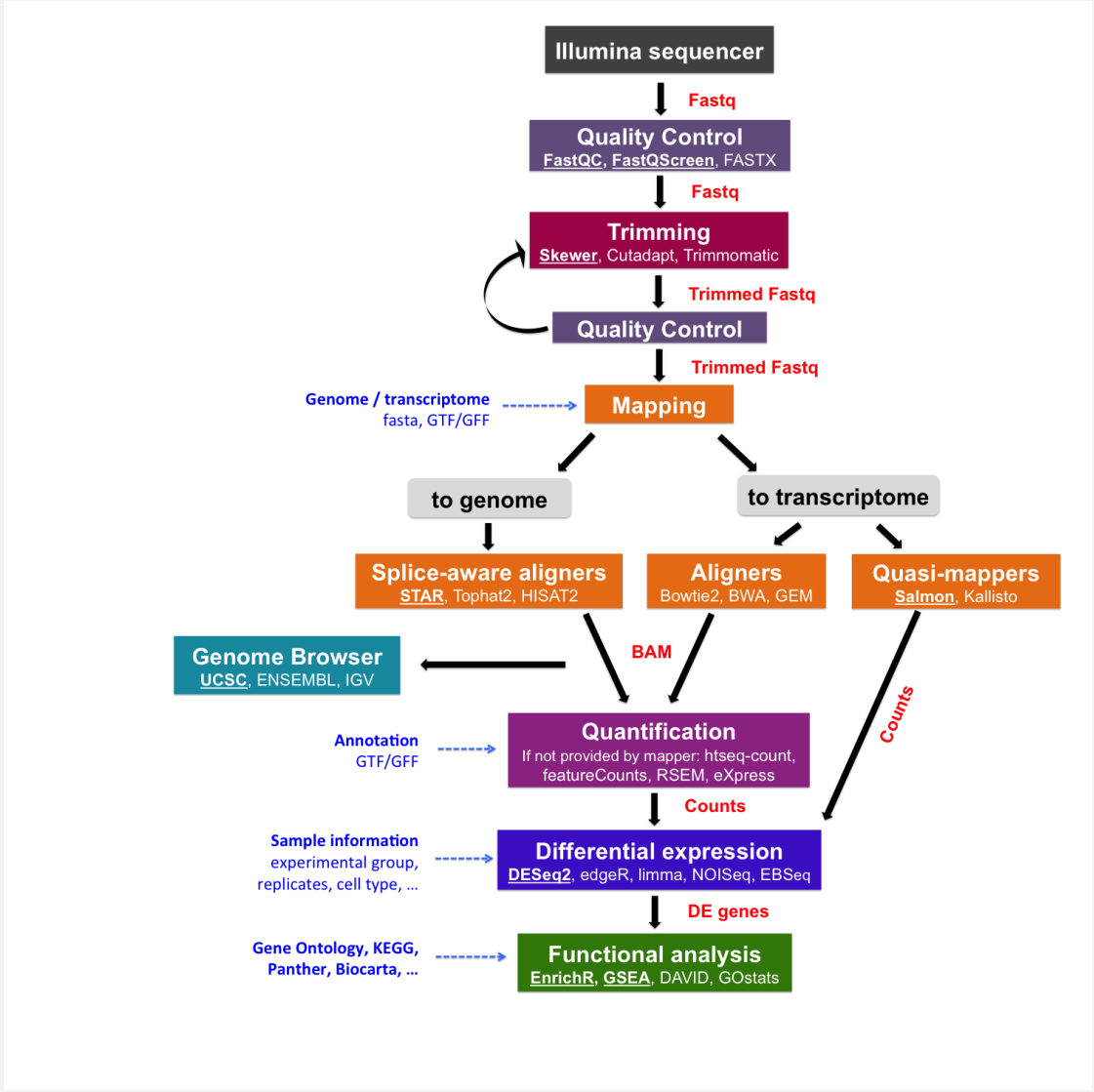
FASTQ files

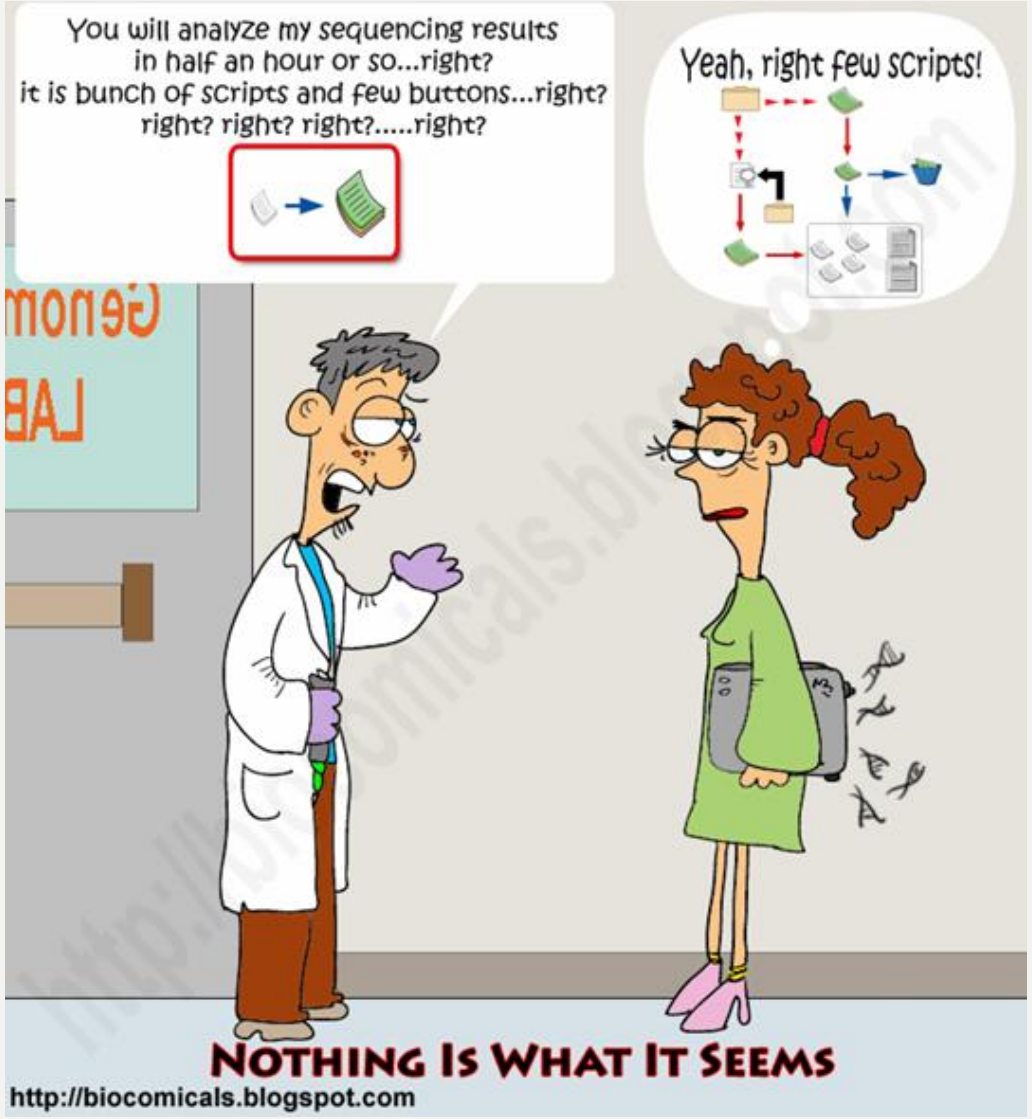






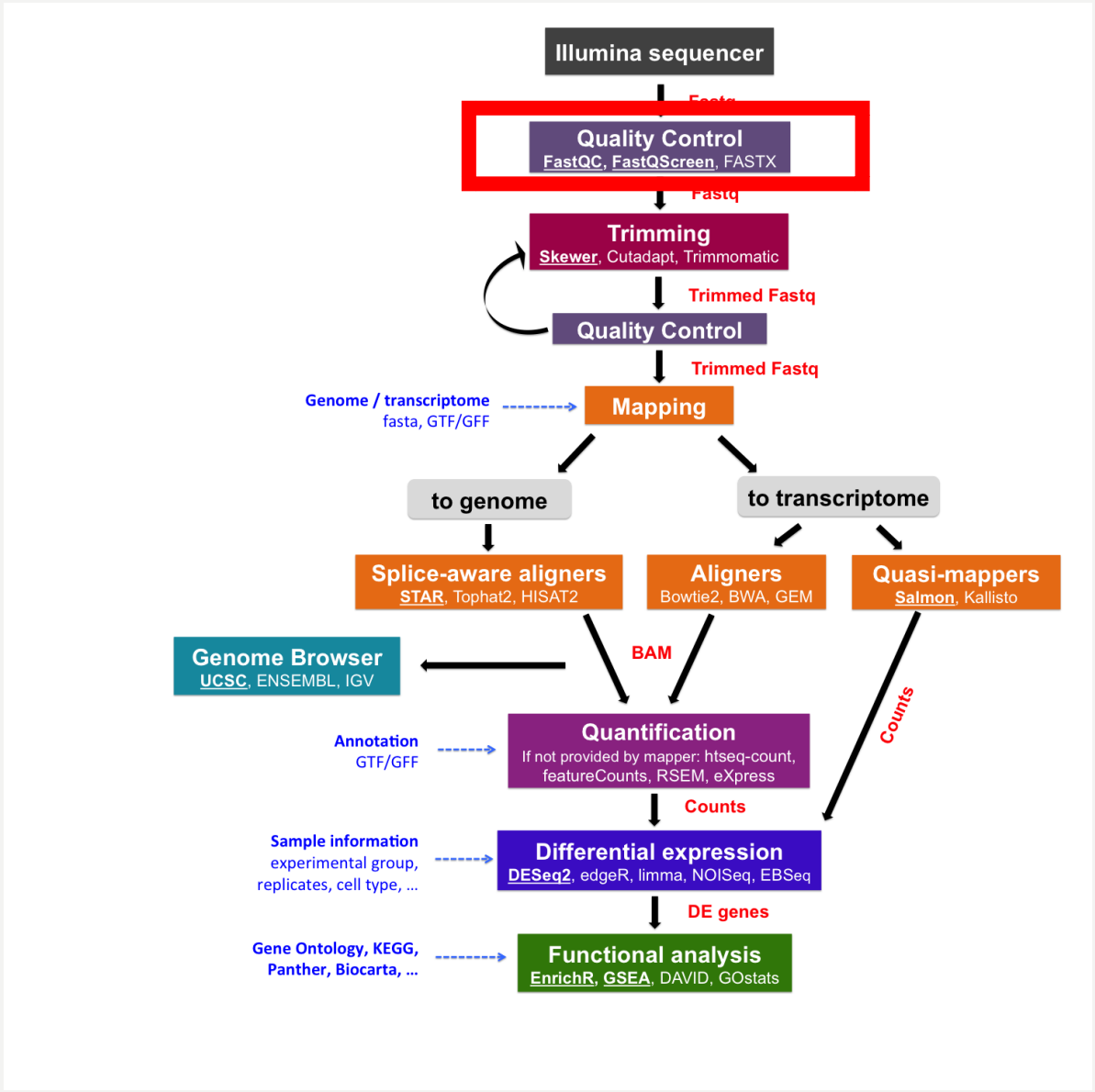
6. DATA ANALYSIS

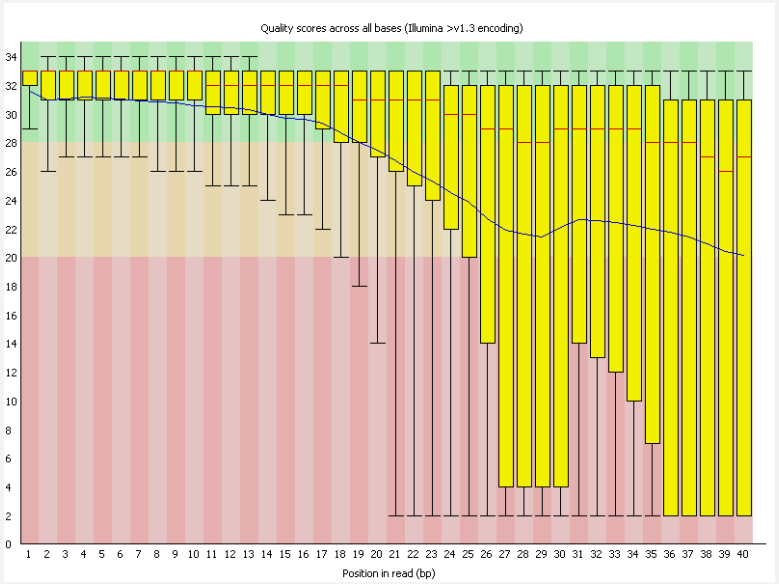


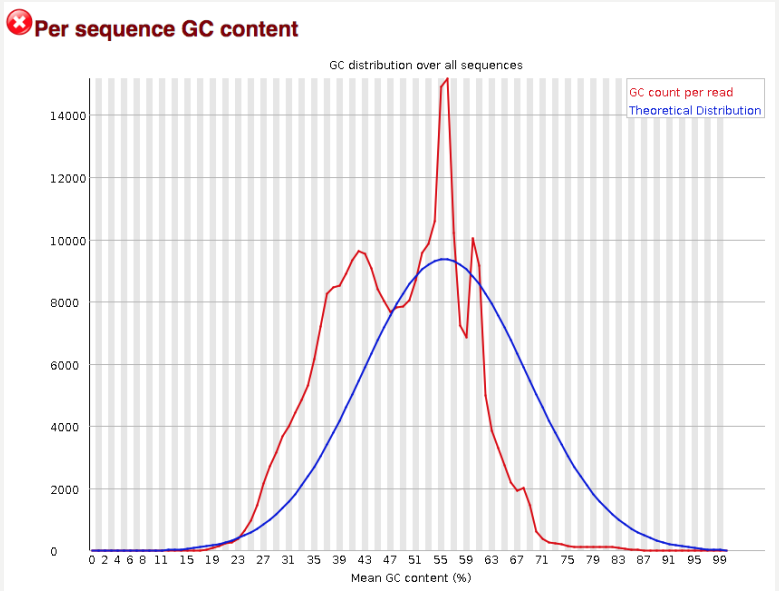


6. DATA ANALYSIS

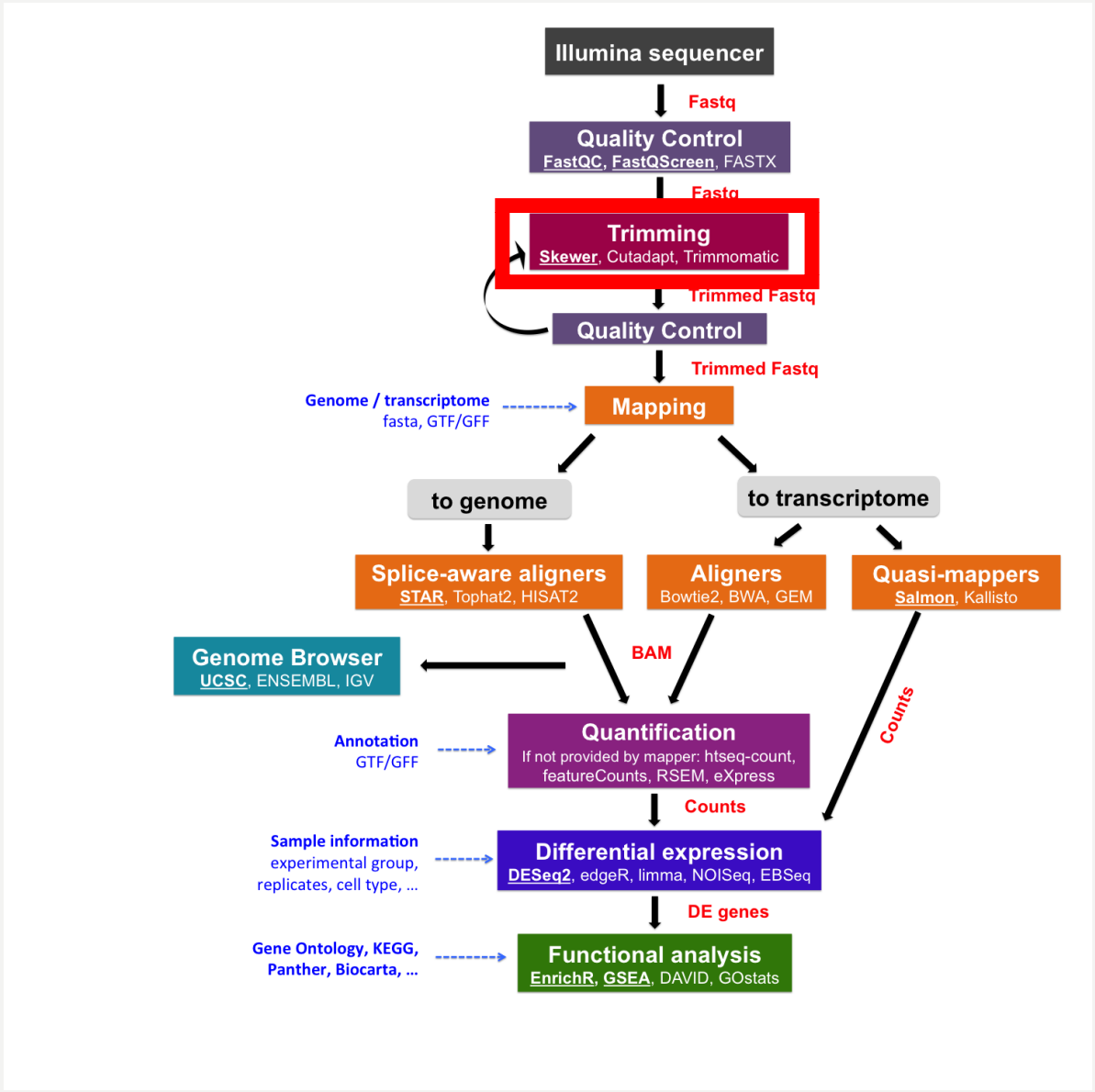
$$Q = -10 \log_{10} P$$

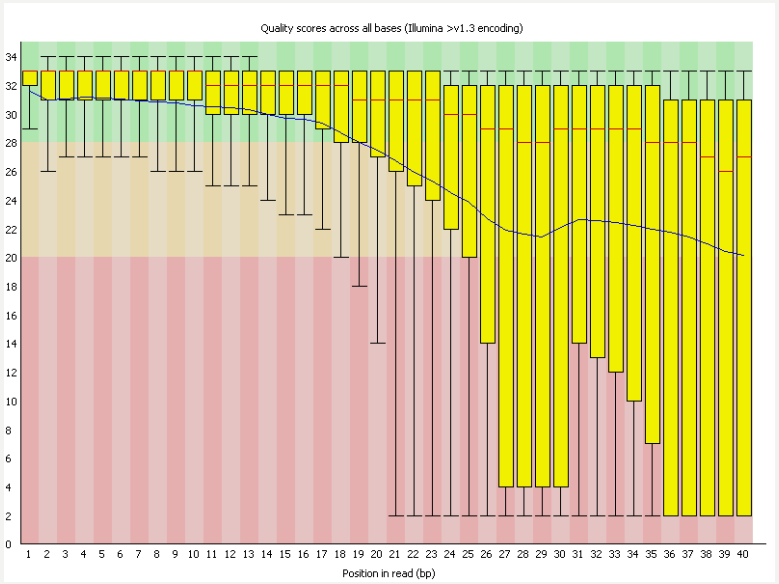


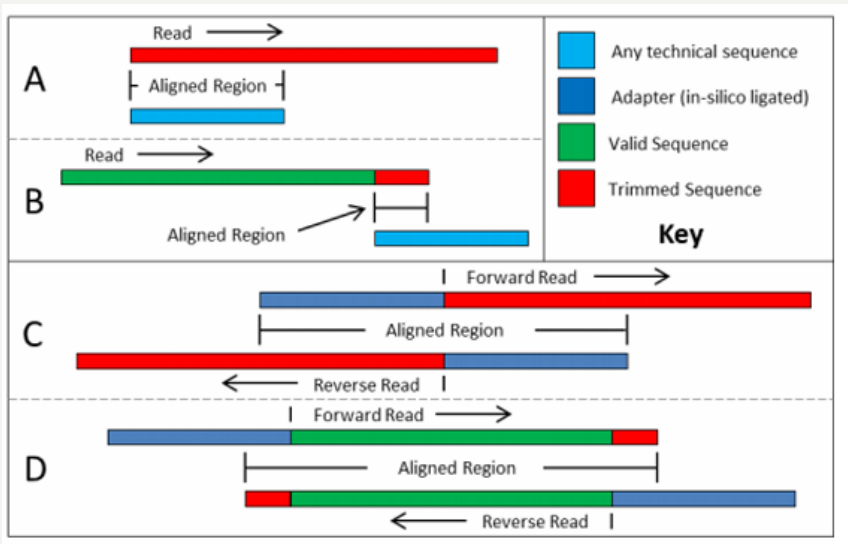




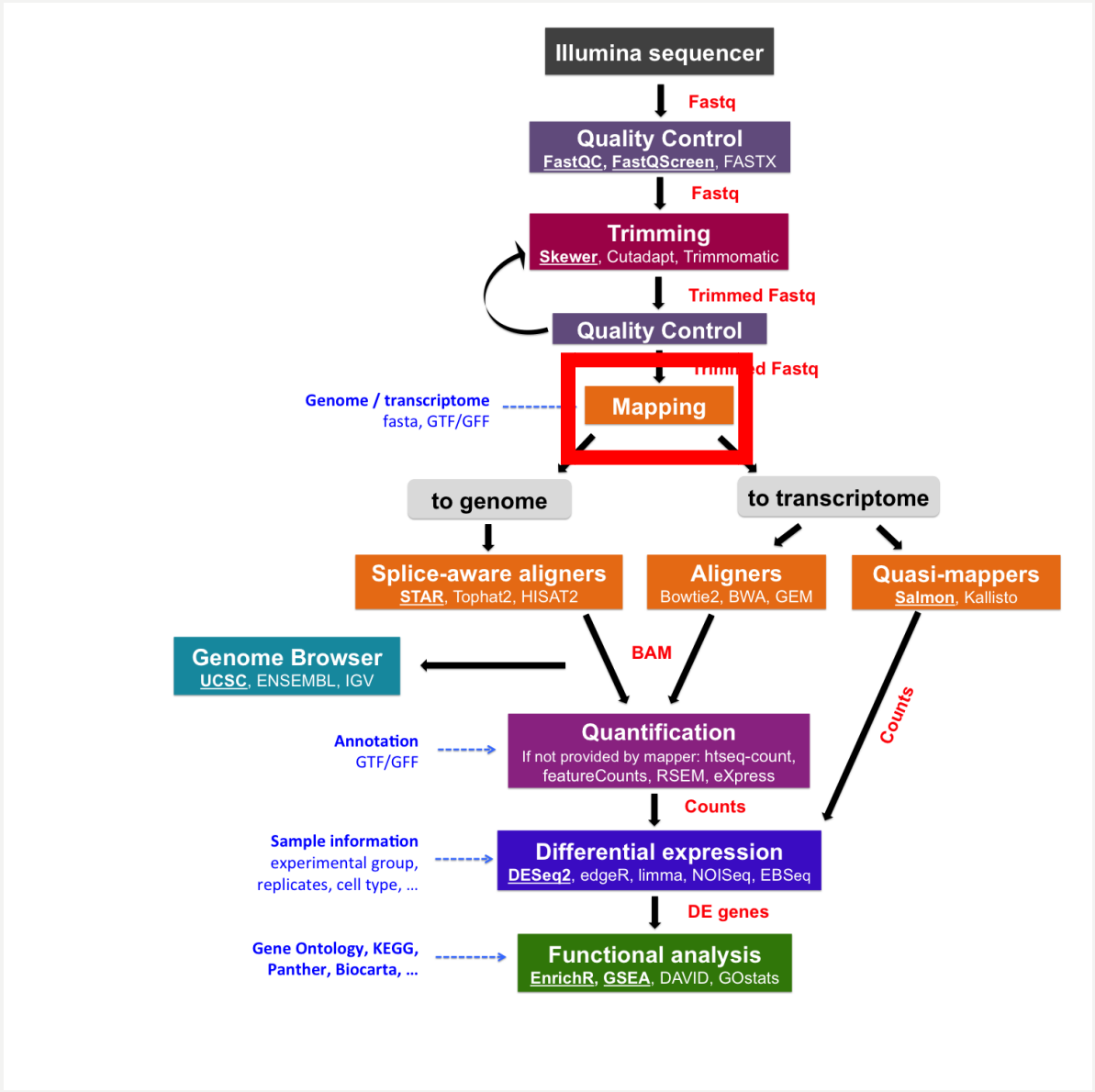
6. DATA ANALYSIS

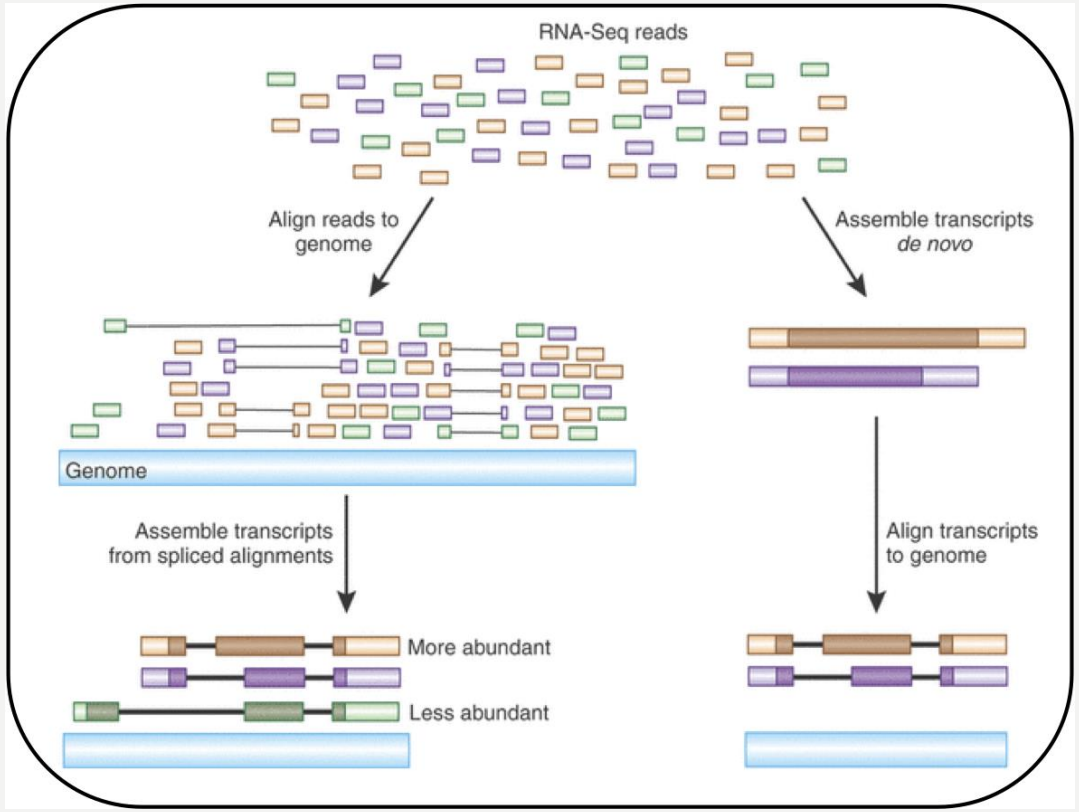






6. DATA ANALYSIS

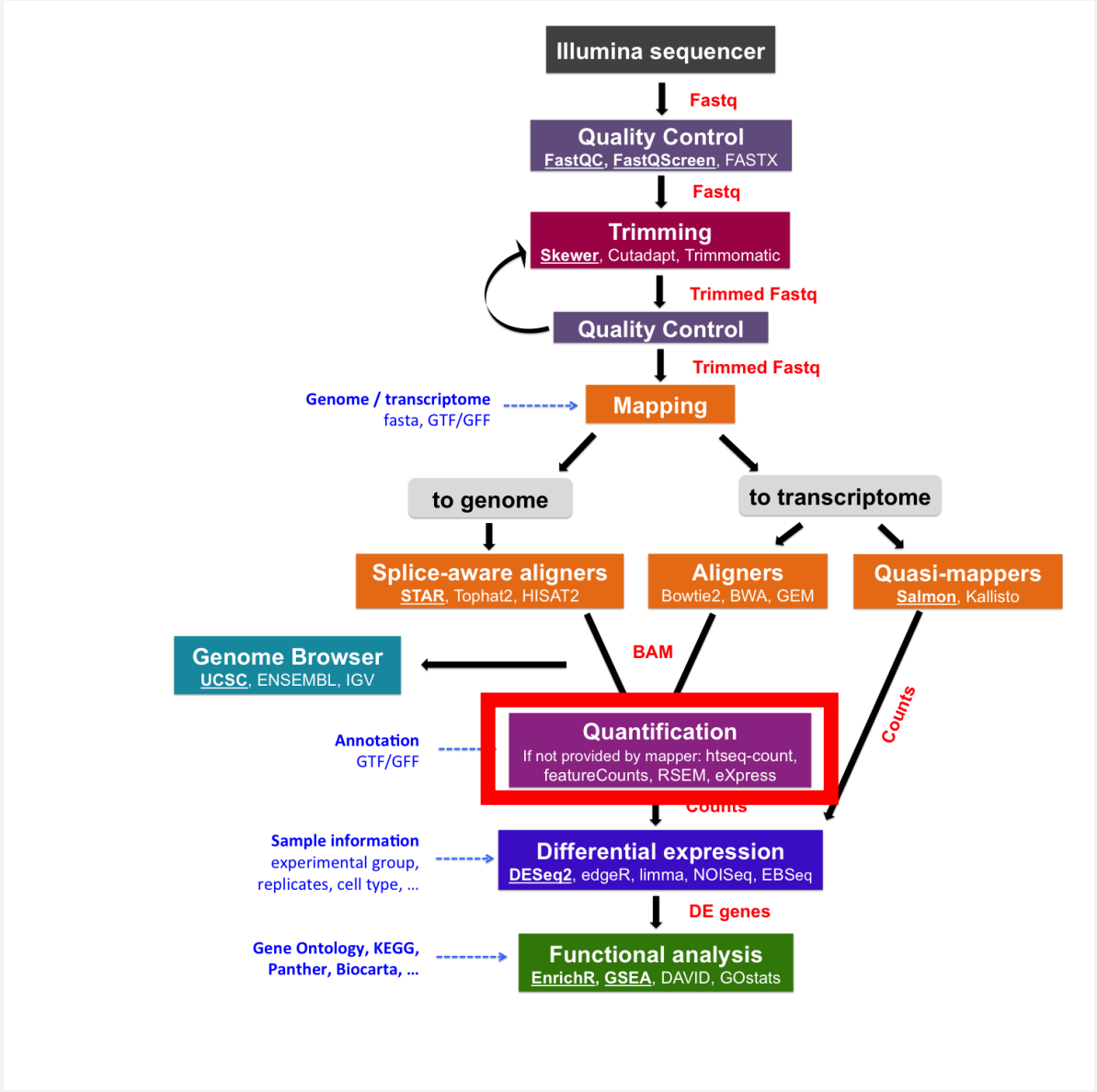


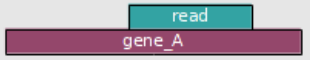
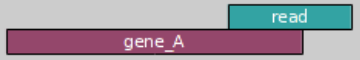
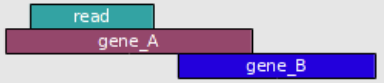
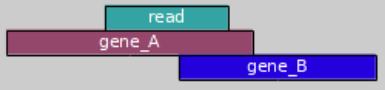
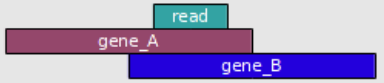
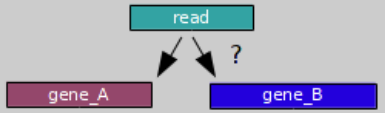


Splice-unaware aligners

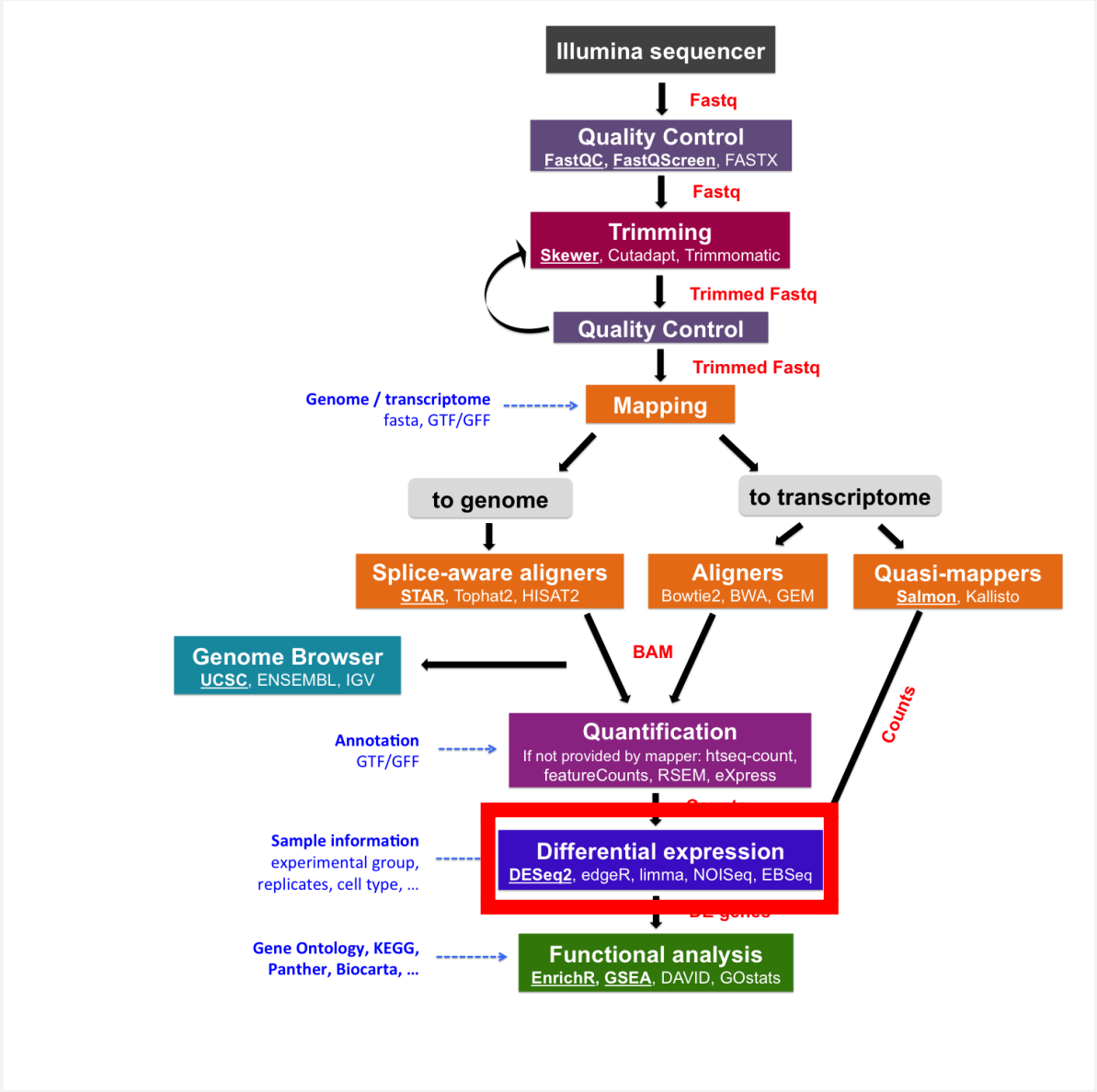
Splice-aware aligners

6. DATA ANALYSIS



	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

6. DATA ANALYSIS

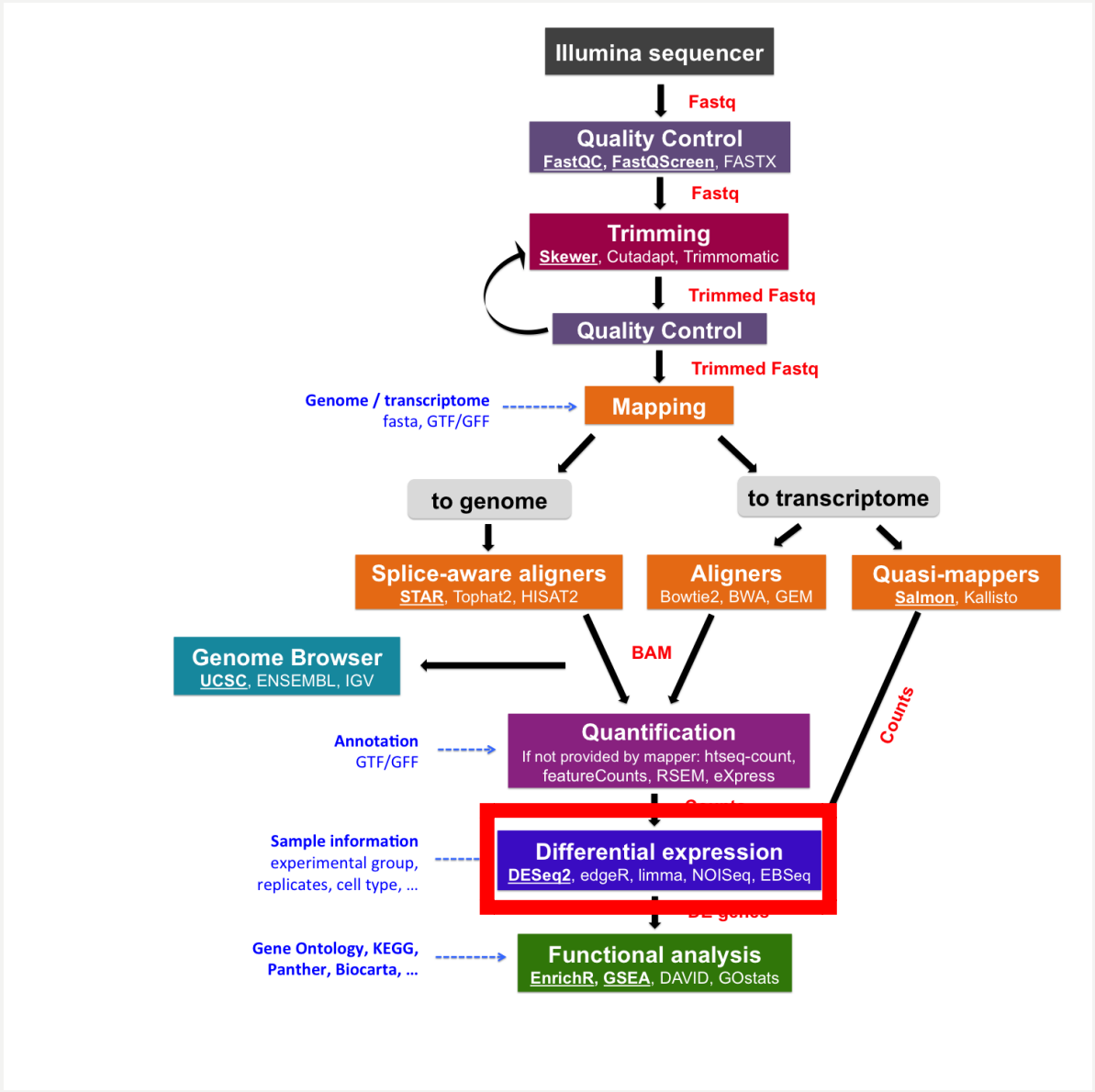


Each column is a sample

Each row is a gene

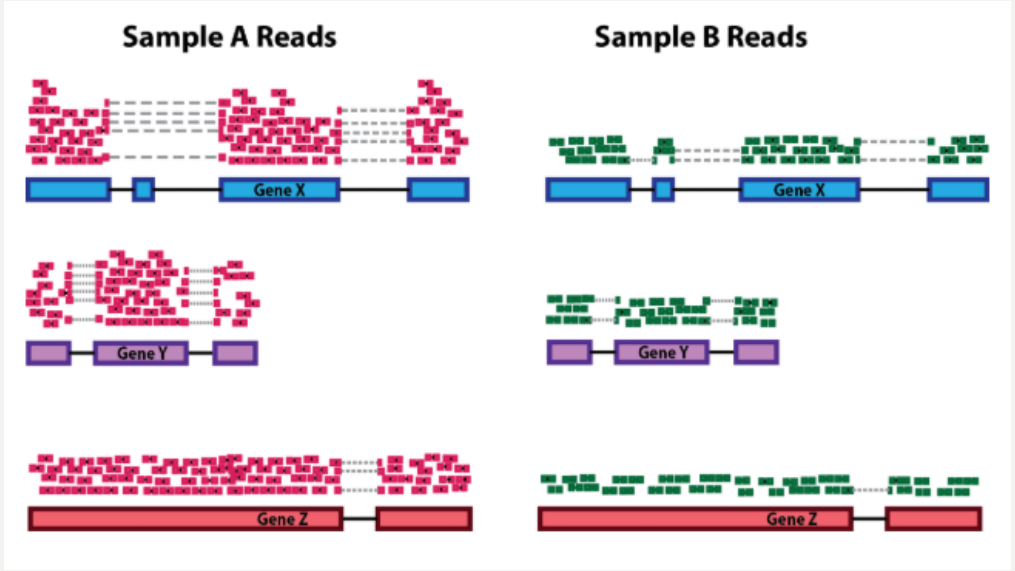
GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

6. DATA ANALYSIS

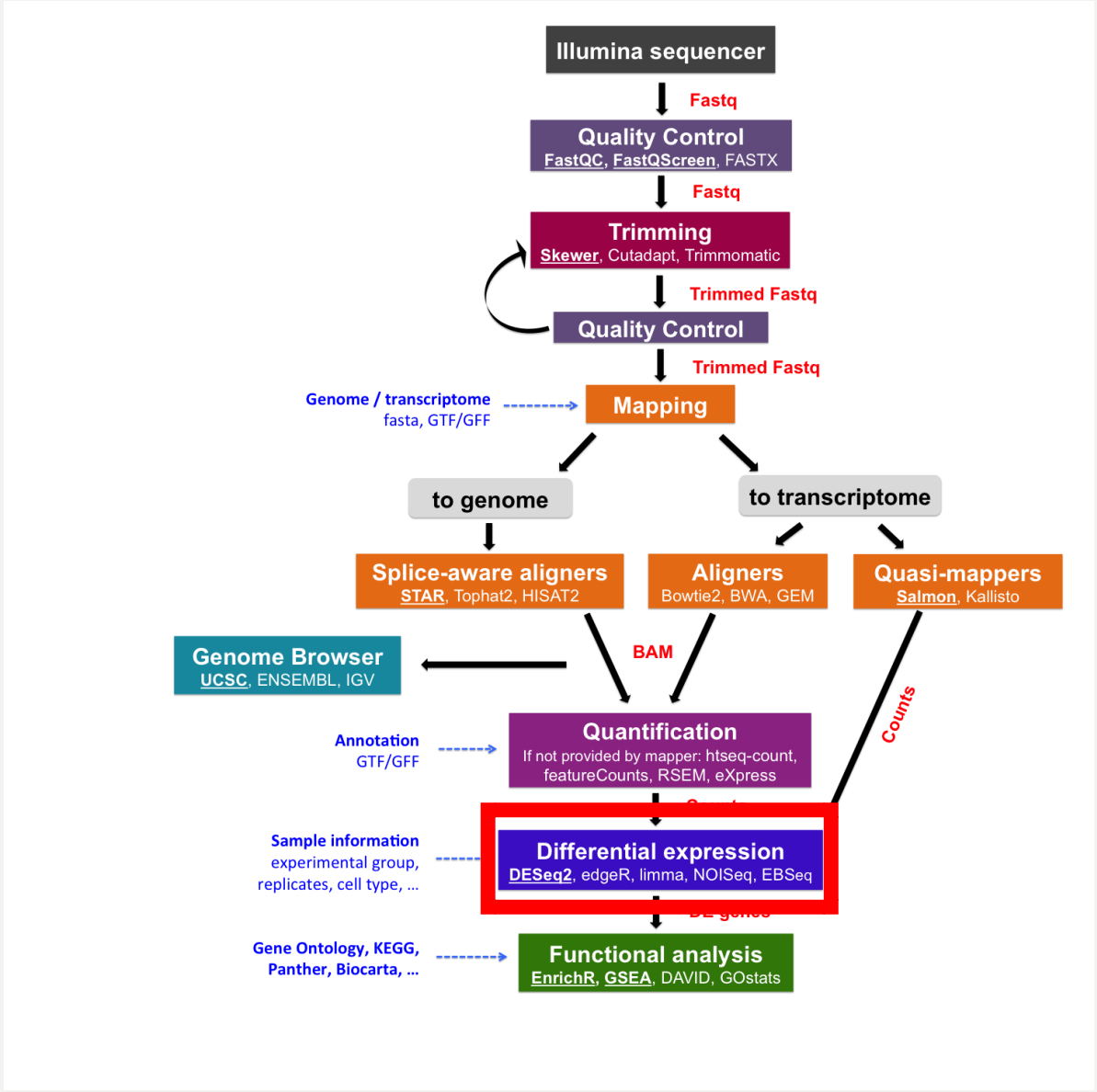


Count Normalization

Sequencing Depth

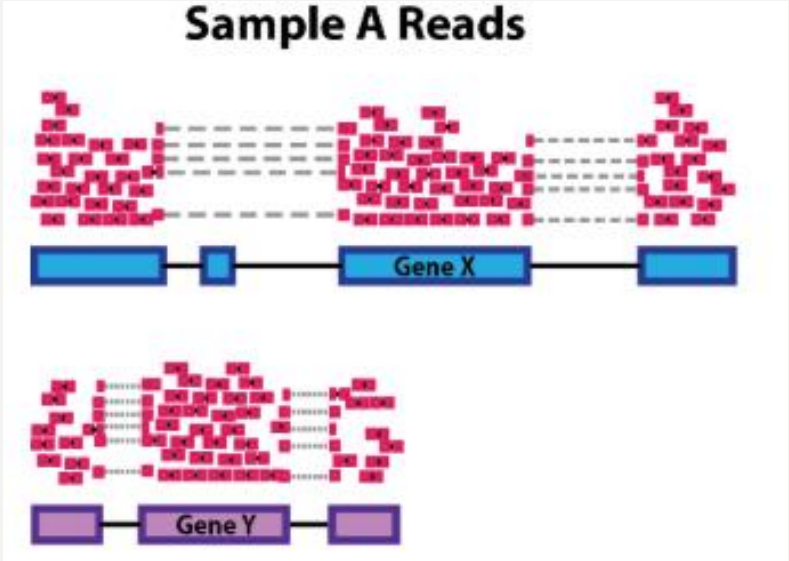


6. DATA ANALYSIS

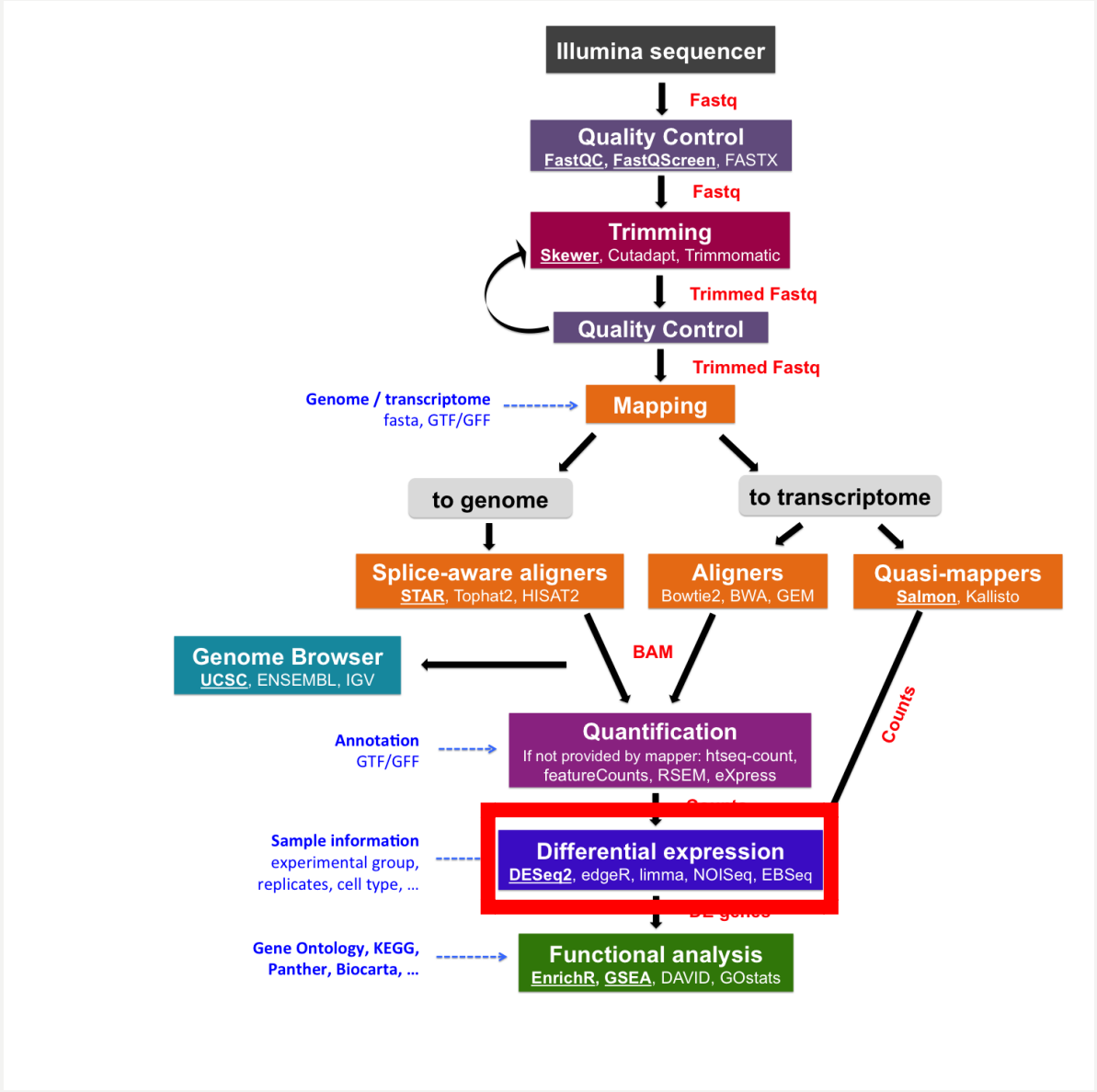


Count Normalization

Gene Length



6. DATA ANALYSIS



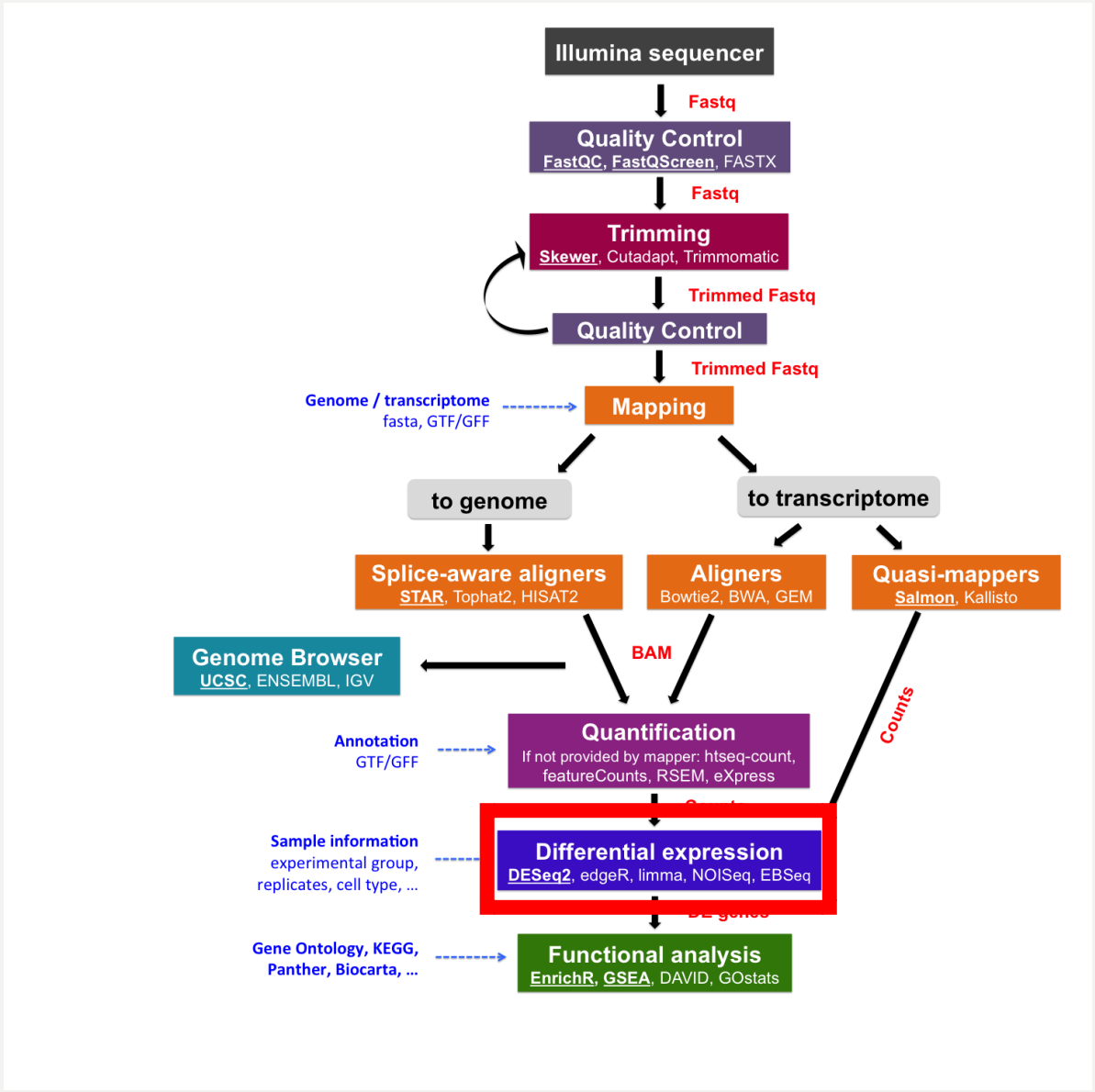
Count Normalization

RPKM/FPKM(Reads/Fragment per kilobase per million mappable reads= $\frac{C}{LN}$

TPM(Transcripts per milion)= $\frac{C/L}{\sum_x \frac{C_x}{L_x}} \times 10^6$

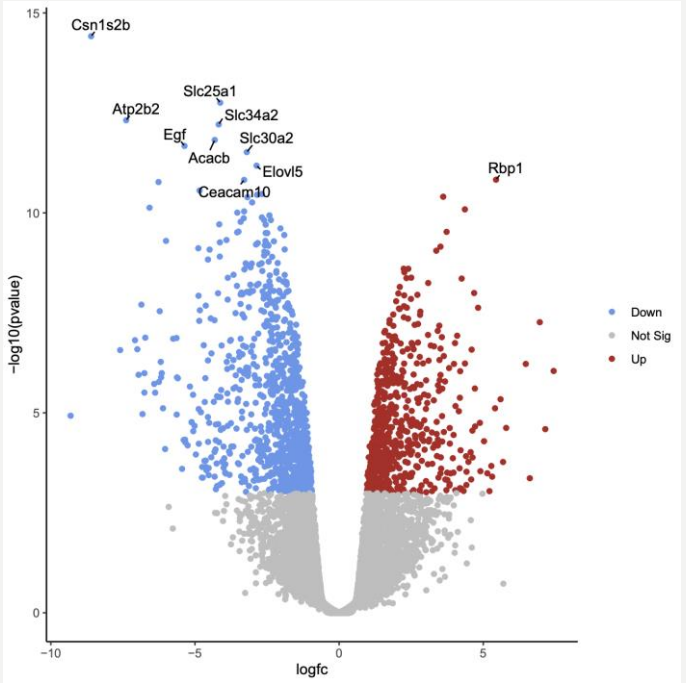
C= Number of mappable reads on a feature (e.g. transcript, exon etc.)
L= Length of feature
N= Total number of mappable reads (in millions)

6. DATA ANALYSIS

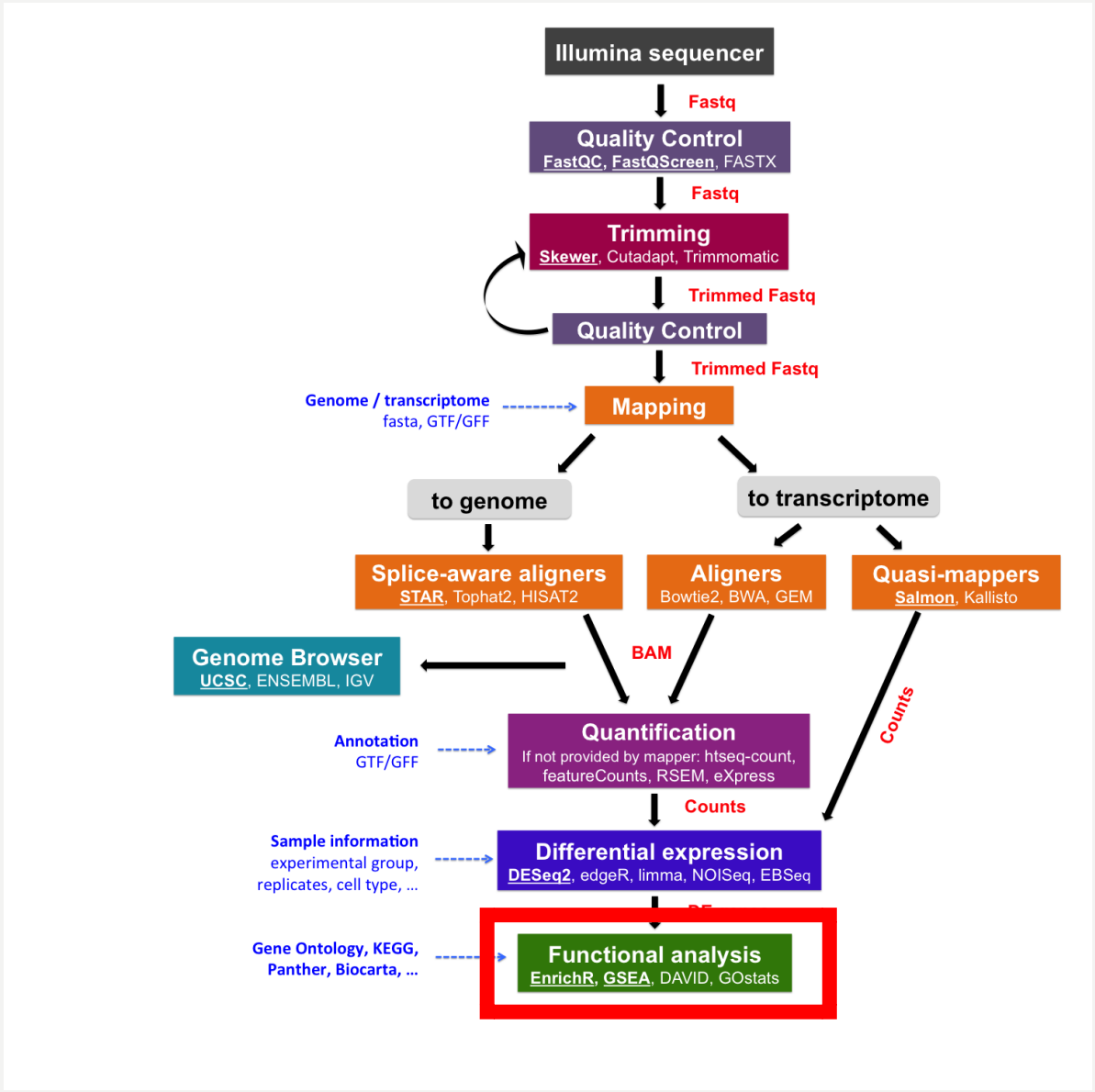


DE analysis allows to find **genes** (or other genomic features like transcripts and exons) **that are expressed at significantly different levels between two groups of samples** (conditions)

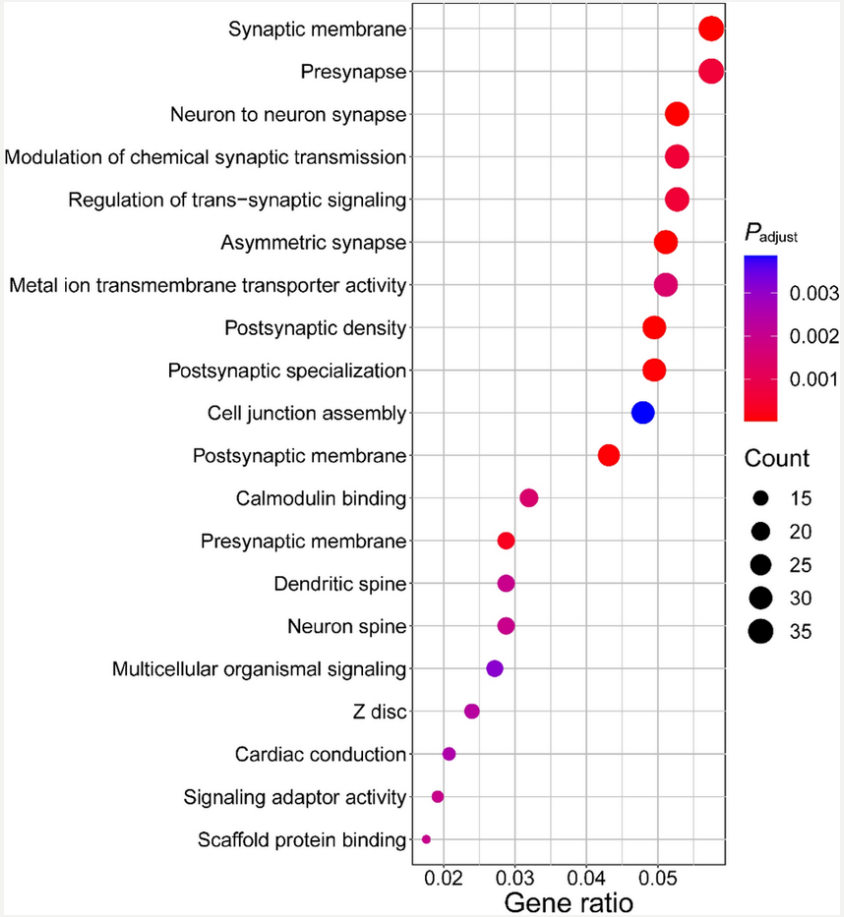
- $Fold\ Change = \frac{Mean\ FPKMCondition_A}{Mean\ FPKMCondition_B}$
- $Pvalue$



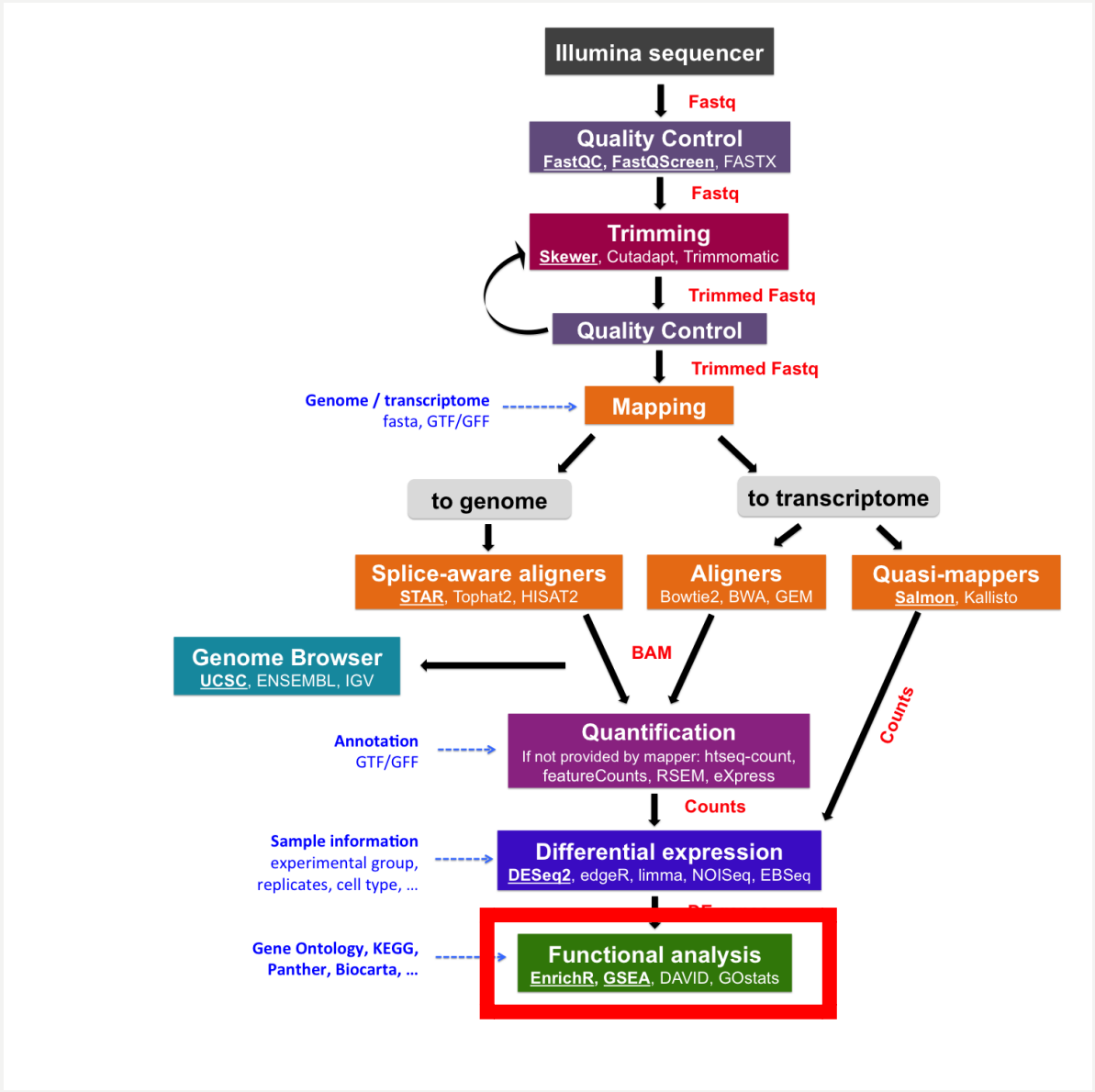
6. DATA ANALYSIS



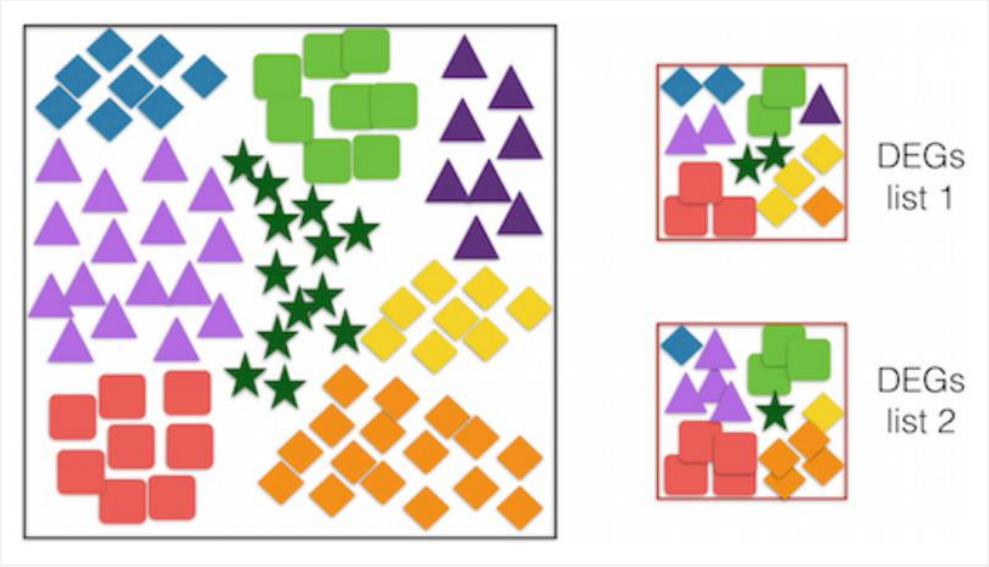
An enrichment analysis will **find which GO terms are over-represented** (or under-represented) using annotations for that gene set.



6. DATA ANALYSIS



An enrichment analysis will **find which GO terms are over-represented** (or under-represented) using annotations for that gene set.

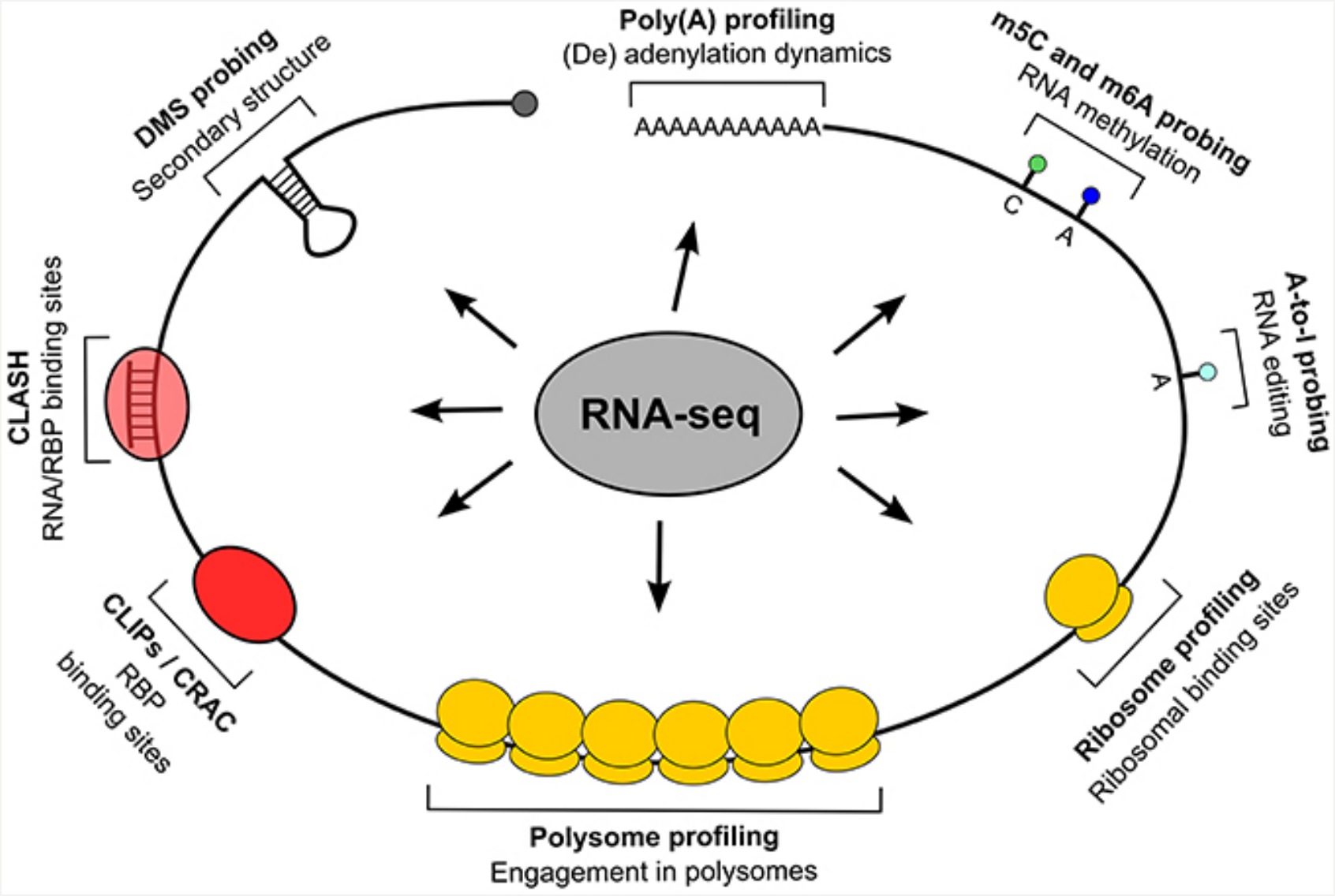


Molecular Function

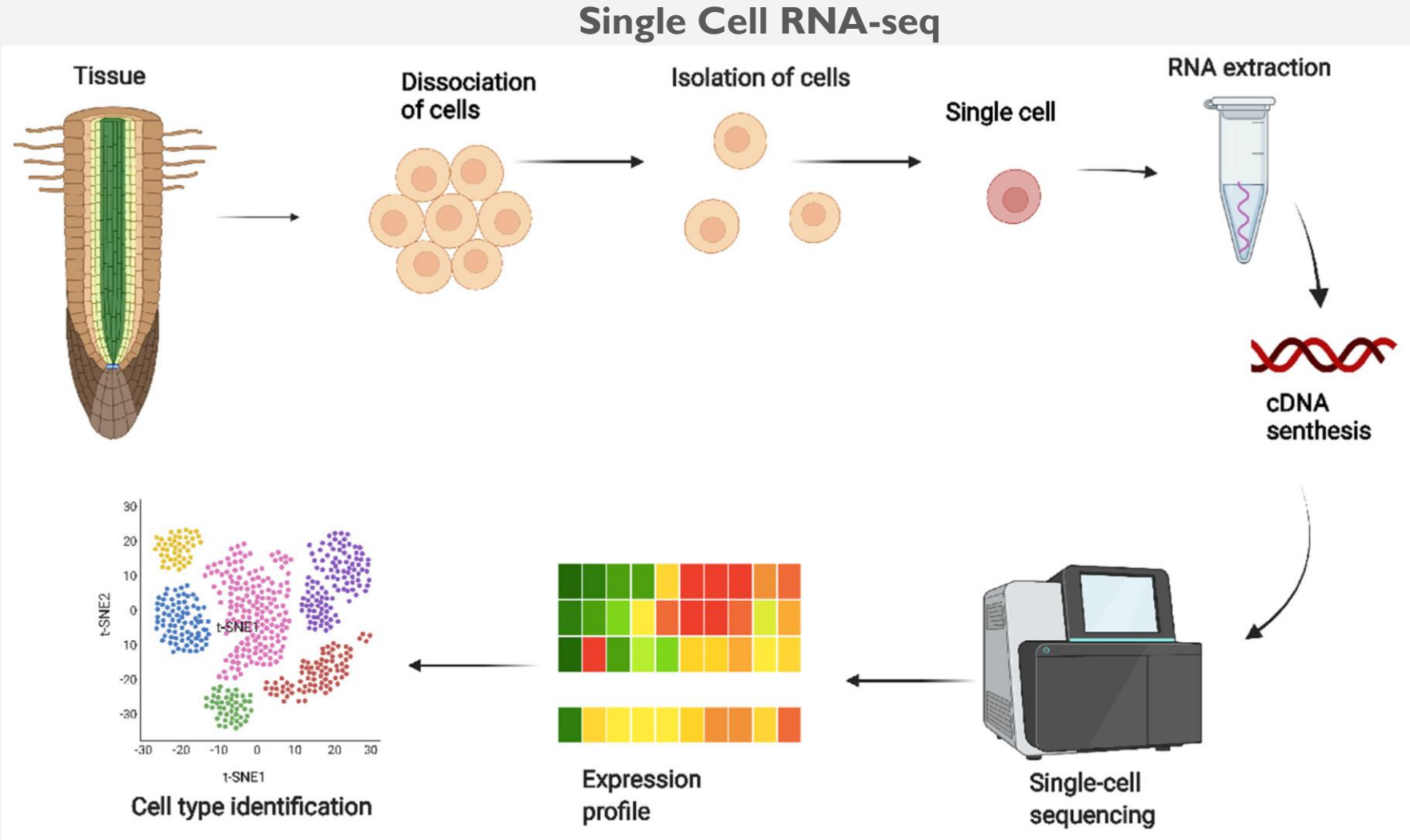
Biological Process

Cellular Component

APPLICATION



APPLICATION



APPLICATION

Spatial Transcriptomics