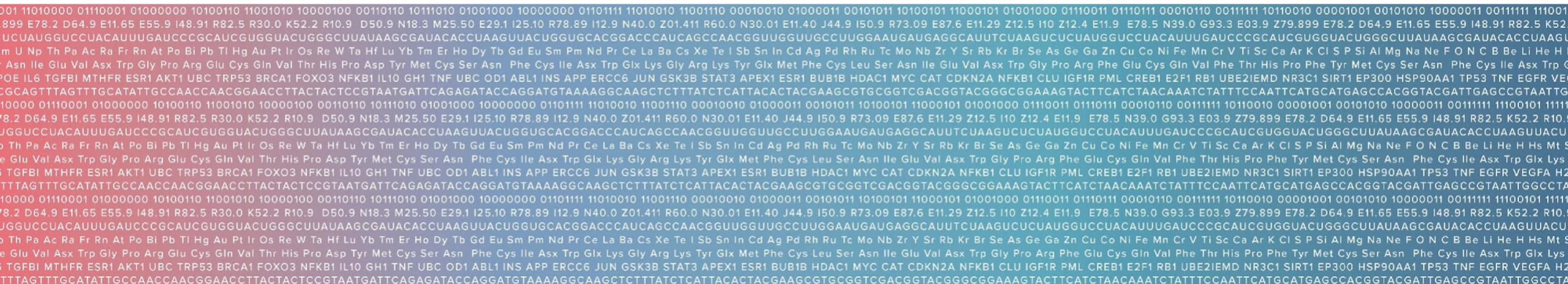
SevenBridges®

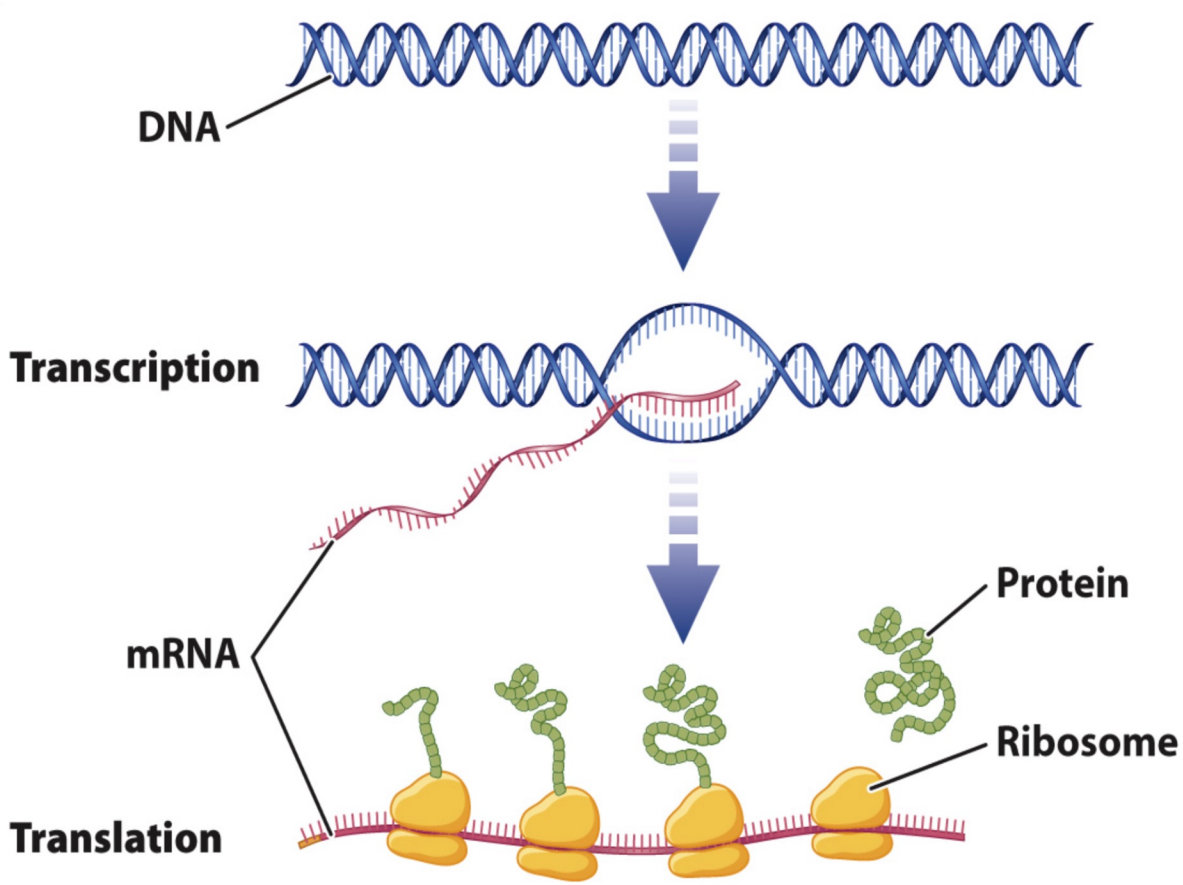
Introduction to RNA-seq

Weeks 8 and 9



Beyond DNA

Central dogma of molecular biology



Explains the flow of genetic information, from DNA to RNA, to make a protein.

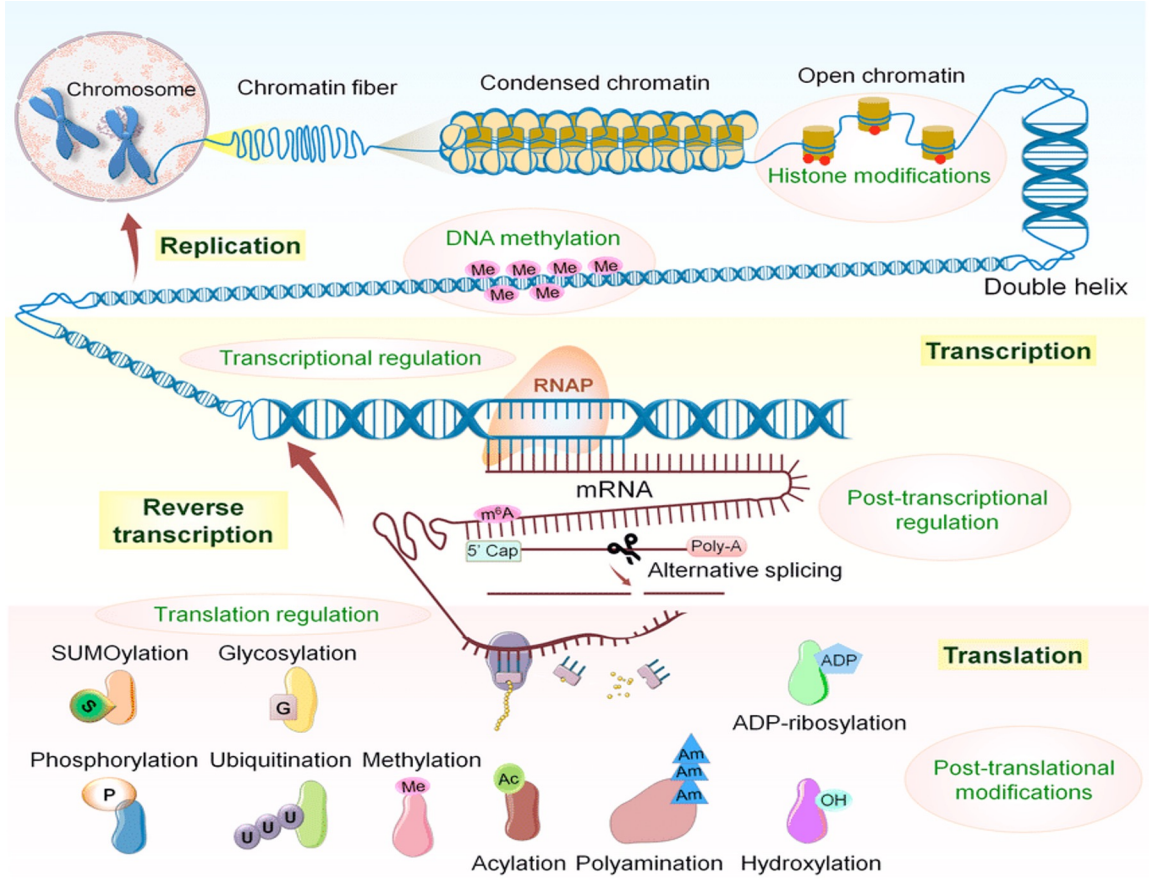
Replication - Before cell division DNA is replicated

Transcription - synthesis of an RNA molecule based on a segment of DNA

Translation - synthesis of a protein based on a sequence of an mRNA molecule

Source: 10.3390/e23060649

Central dogma of molecular biology



Explains the flow of genetic information, from DNA to RNA, to make a protein.

Replication - Before cell division DNA is replicated

Transcription - synthesis of an RNA molecule based on a segment of DNA

Translation - synthesis of a protein based on a sequence of an mRNA molecule

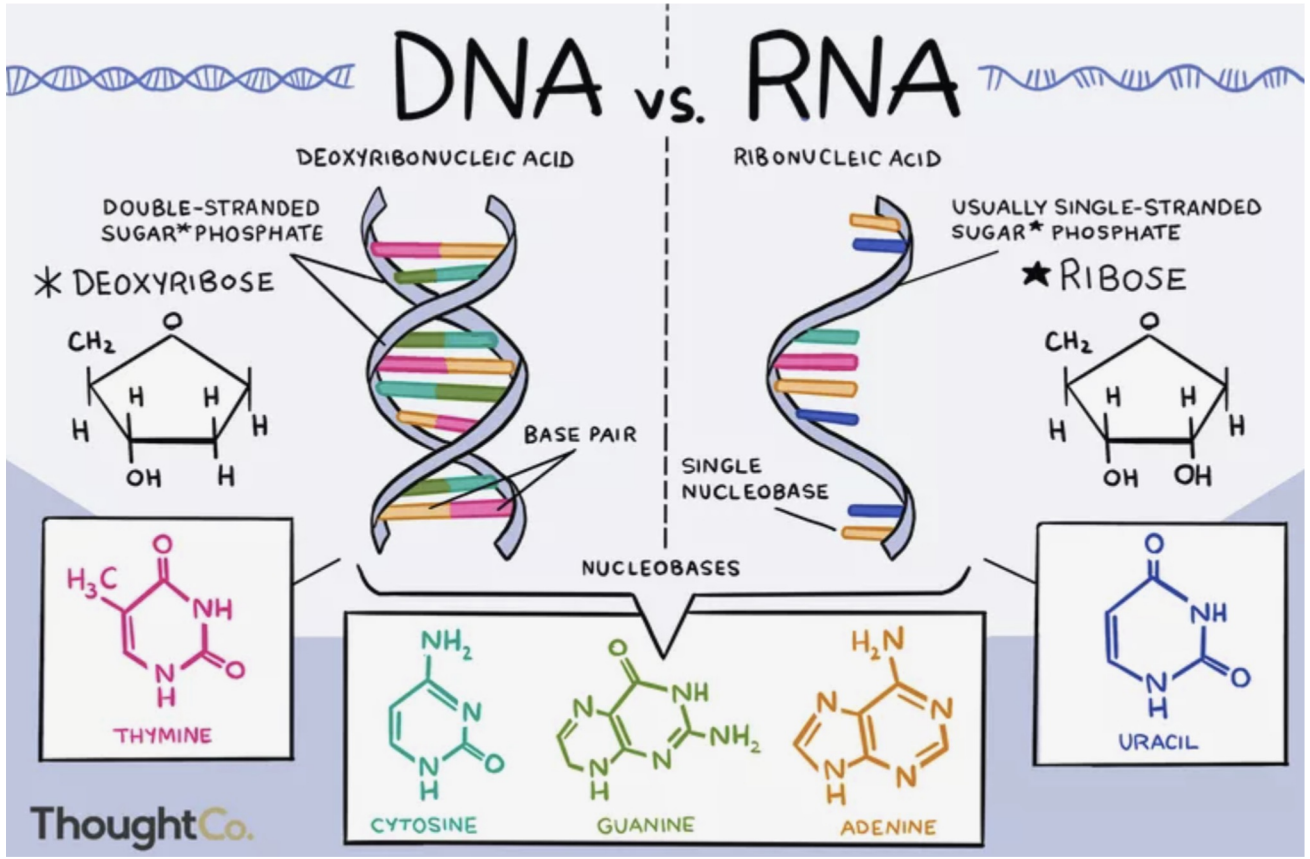
Source:
https://doi.org/10.1016/j.molp.2020.11.002

Transcriptomics

Lots of RNAs, splicing, GTF, translation

SevenBridges®

RNA vs DNA difference?

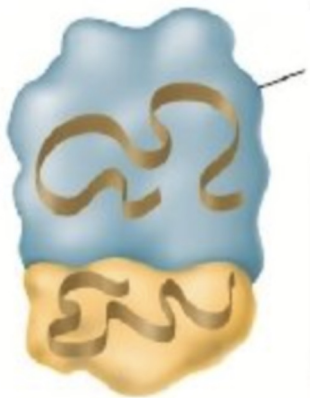


Source: [ThoughtCo.]

DNA:
- **Deoxyribonucleic acid**
- **Double strand**
- **T (thymine)**

RNA:
- **Ribonucleic acid**
- **Single strand**
- **U (uracil)**

Main types of RNA



Ribosome

Ribosomal RNA
Forms an important part of both subunits of the ribosome.

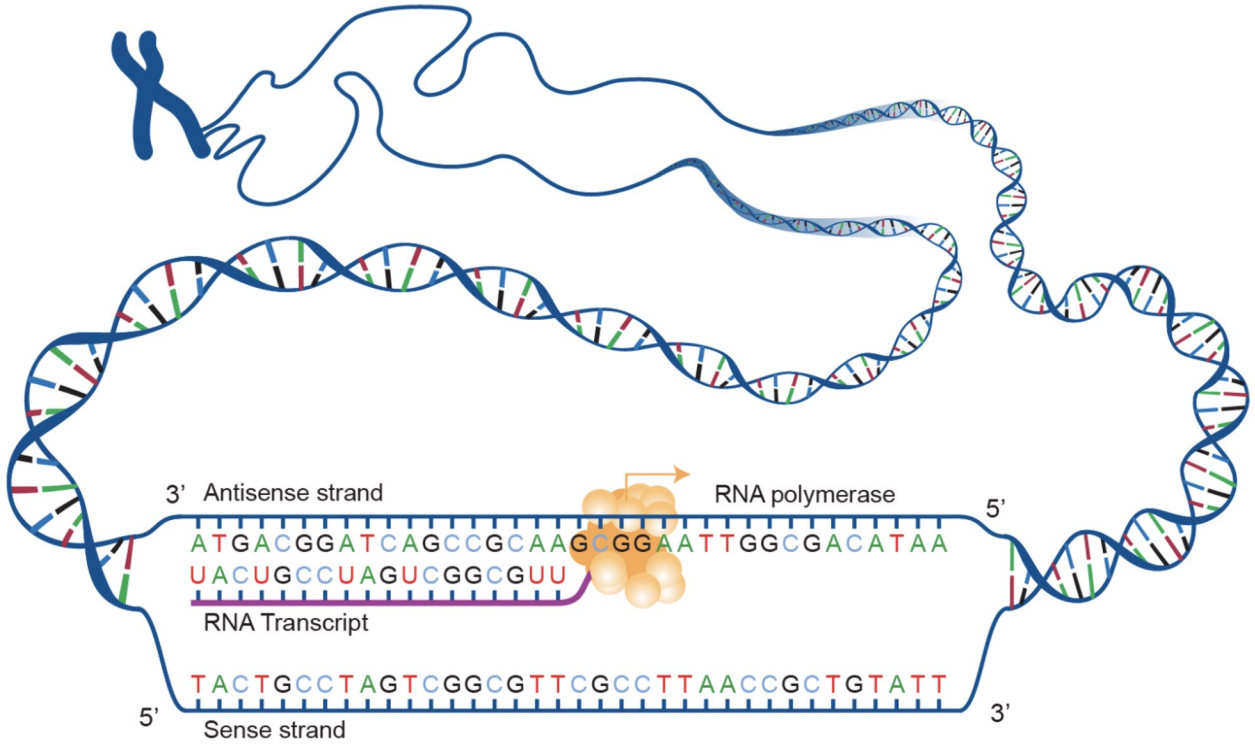
Messenger RNA
Carries instructions for polypeptide synthesis from nucleus to ribosomes in the cytoplasm.

Amino acid

Transfer RNA
Carries amino acids to the ribosome and matches them to the coded mRNA message.

Source: ThoughtCo.

Transcription



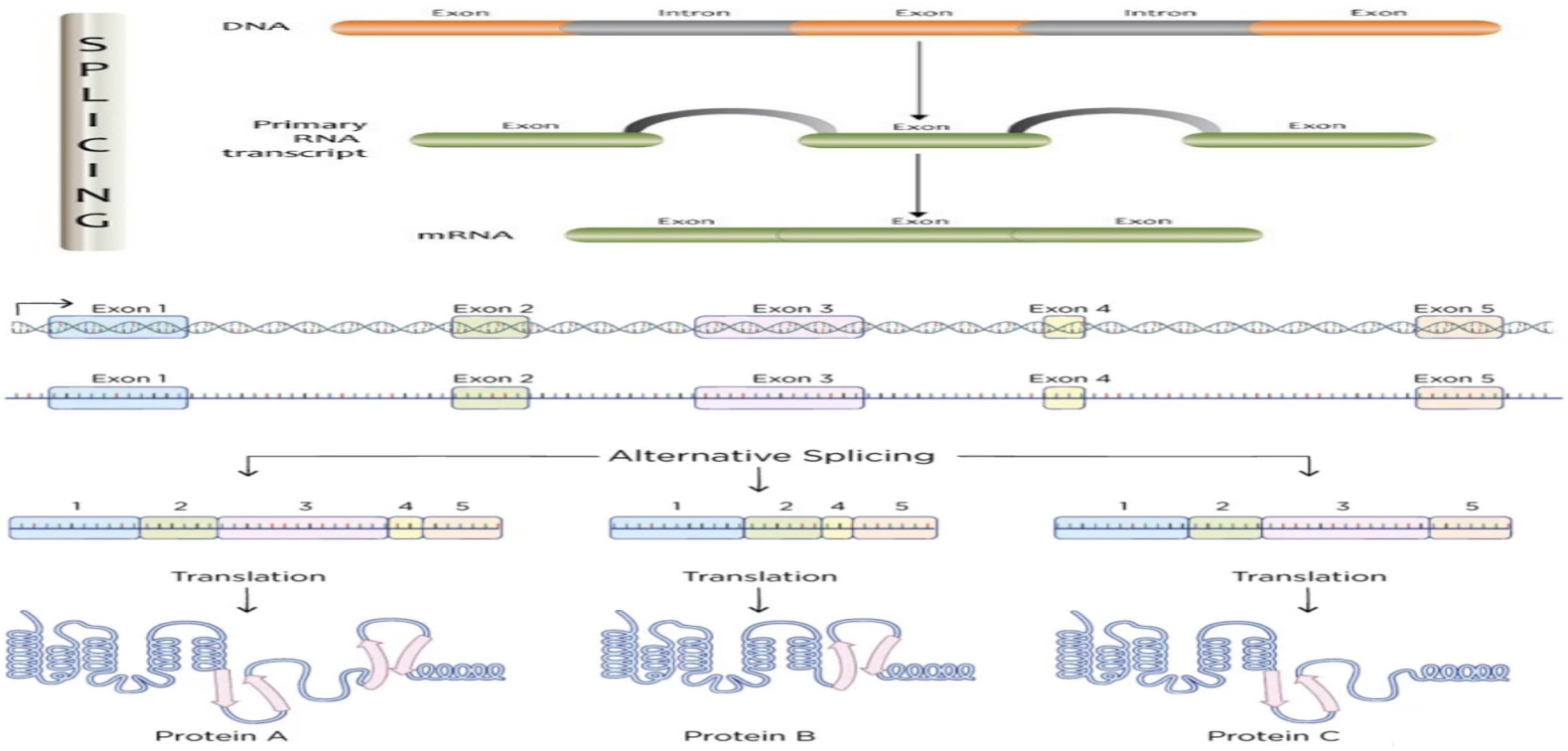
Transcription - process of making an RNA copy of a gene sequence. This copy, called a messenger RNA (mRNA) molecule, leaves the cell nucleus and enters the cytoplasm, where it directs the synthesis of the protein, which it encodes.

All types of RNAs are synthesized in the transcription !!!
From our perspective mRNAs are the most important

RNA polymerase which is the main transcription enzyme uses **one** of the DNA strands (the **template strand**) as a template to make a new, complementary RNA molecule.

Source: National Human Genome Research Institue (NHGRI)

Alternative splicing



Source: MCAT content

GENE (DNA): consists of introns and exons

pre-mRNA: Initial transcription product

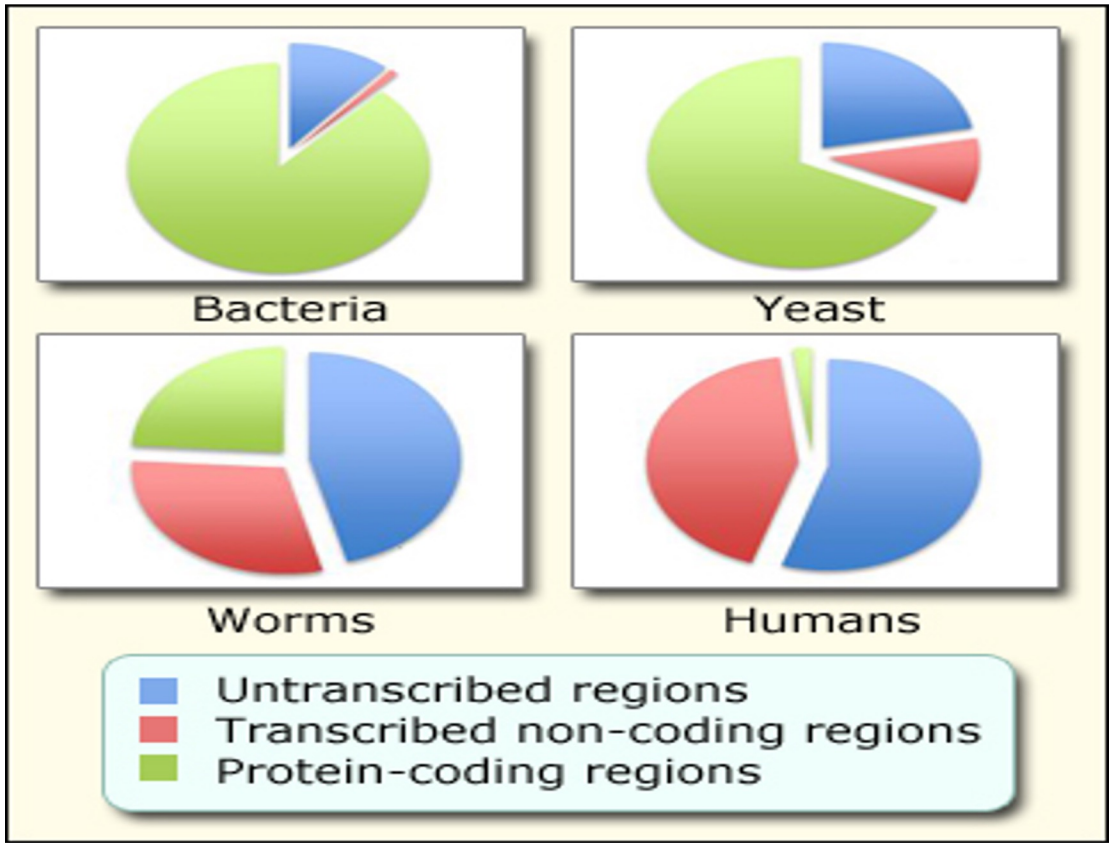
After initial transcription maturation of RNA sequence is performed in a process of splicing

Most mRNAs undergo to alternative splicing process

Proteins: One gene (usually) code multiple proteins

Alternative splicing is **the process of selecting different combinations of exons (splice sites) within a messenger RNA precursor (pre-mRNA) to produce variably spliced mRNAs**. These multiple mRNAs can encode proteins that vary in their sequence and activity, and yet arise from a single gene.

Transcription – how much DNA is transcribed



Gene - segment of DNA which is transcribed into RNA which then has a function in cell

If RNA codes for protein that RNA is called **mRNA** and the region of genome from which it is transcribed is called **protein-coding gene** (green)

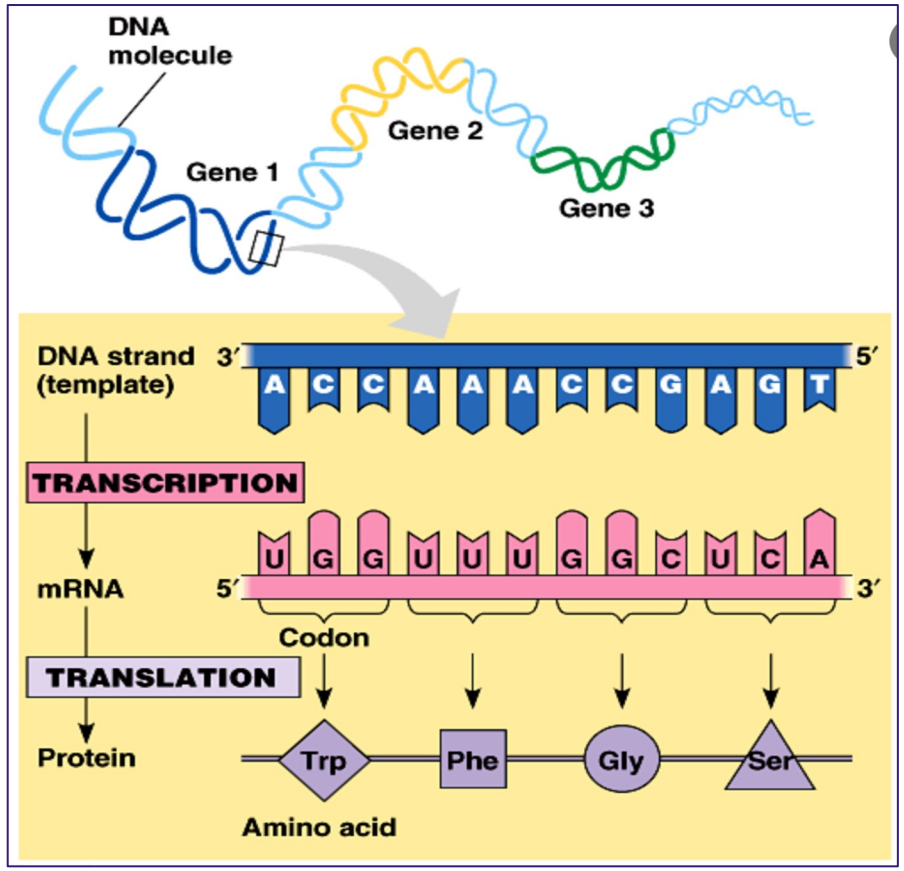
Genes which code for RNA with different functions other than protein coding - structural, regulatory, transport etc. - **non-coding genes** (red)

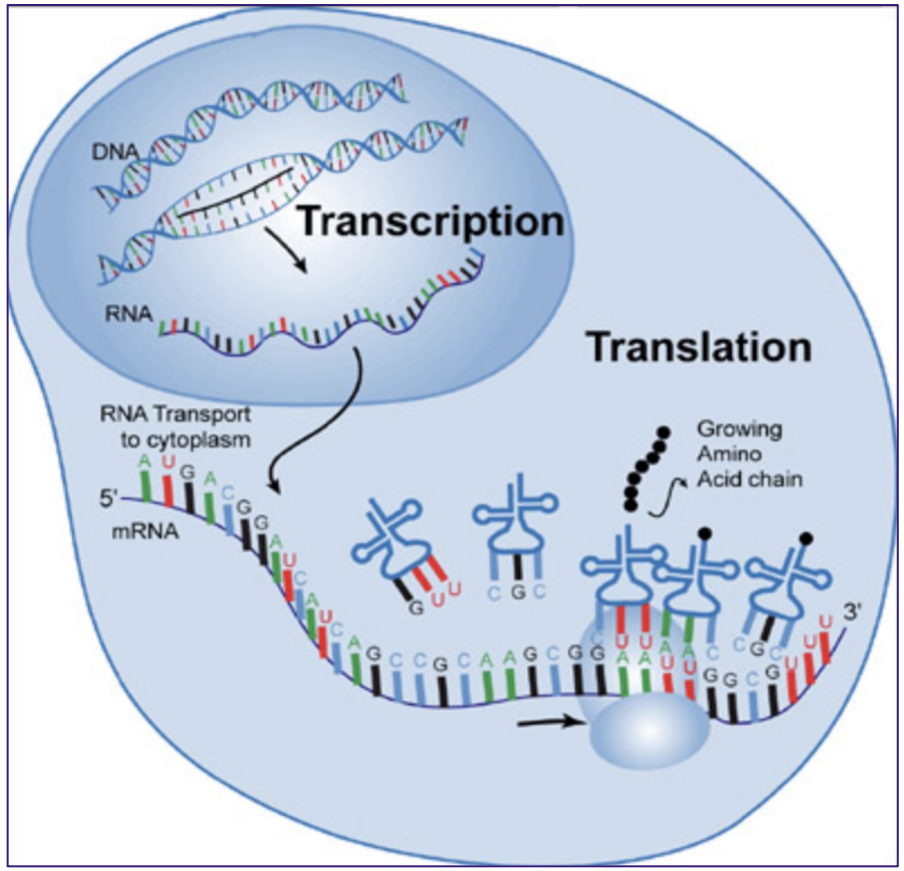
Some regions of DNA (most of it) are not transcribed at all (blue)

Although humans have the smallest percent of protein coding regions, they can synthesize a big number of proteins because of **alternative splicing**

Source: MIT course: Non-coding RNAs

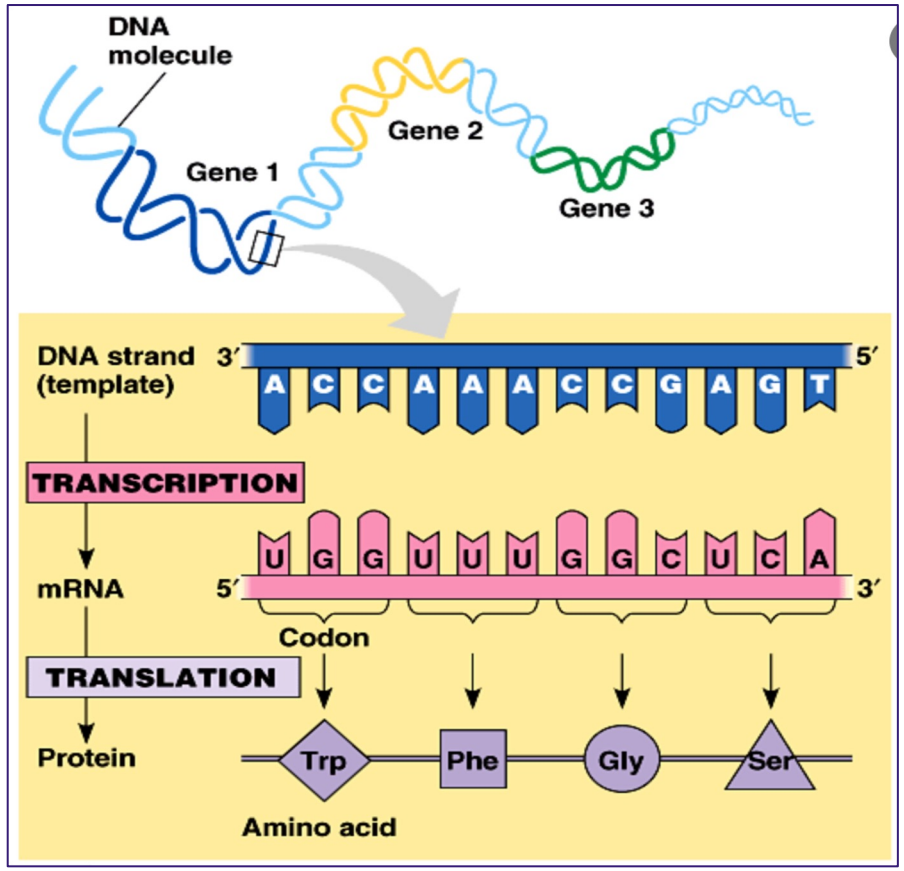
mRNAs: translation to proteins



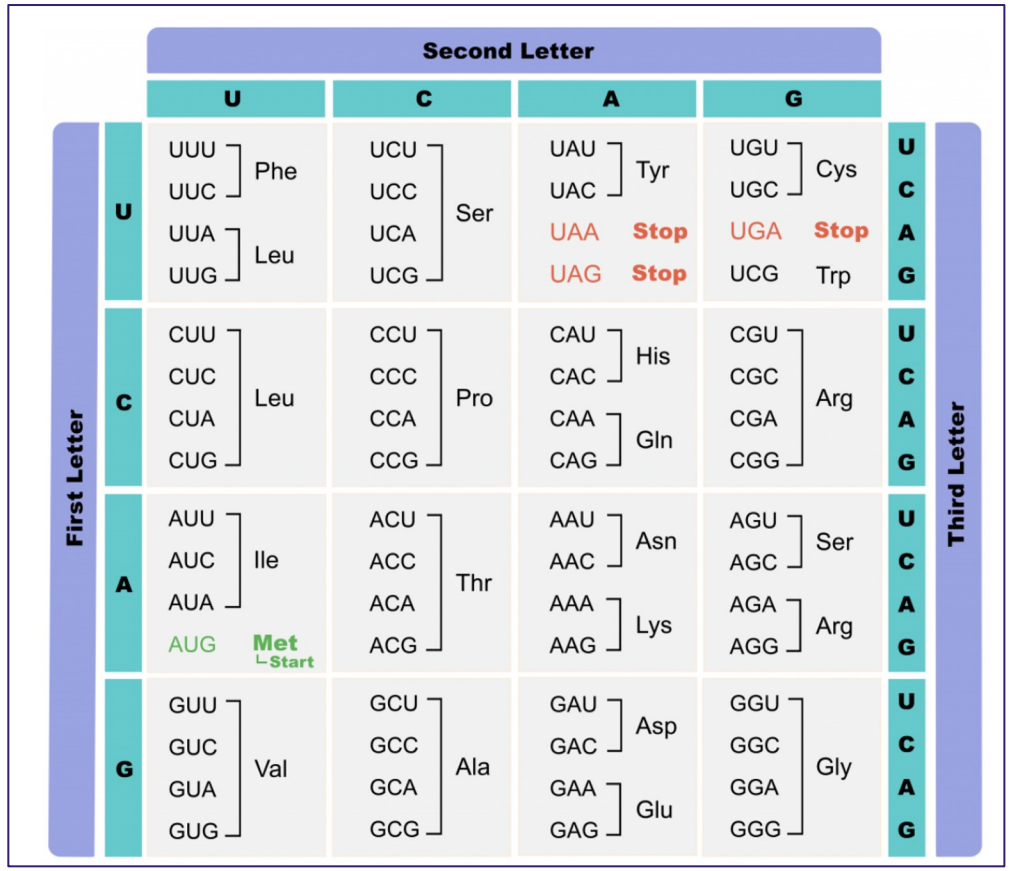


Source: modified from [10.13140/RG.2.2.24229.40166](https://doi.org/10.13140/RG.2.2.24229.40166)

mRNAs: translation to proteins



Source: modified from [10.13140/RG.2.2.24229.40166](10.13140/RG.2.2.24229.40166)



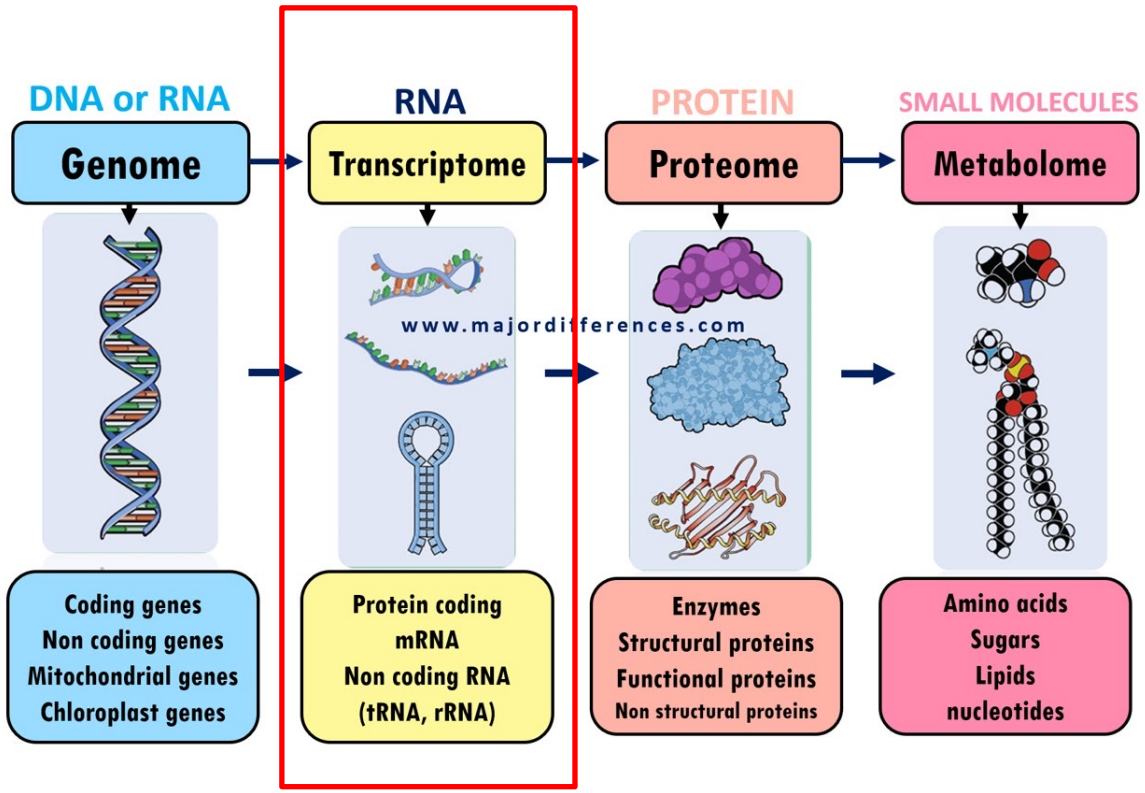
Source: [University of MIchigan]

Transcriptomics vs Genomics

Subtitle

SevenBridges®

Transcriptomics



Transcriptome - refers to all of the RNA transcripts in a cell or tissue.

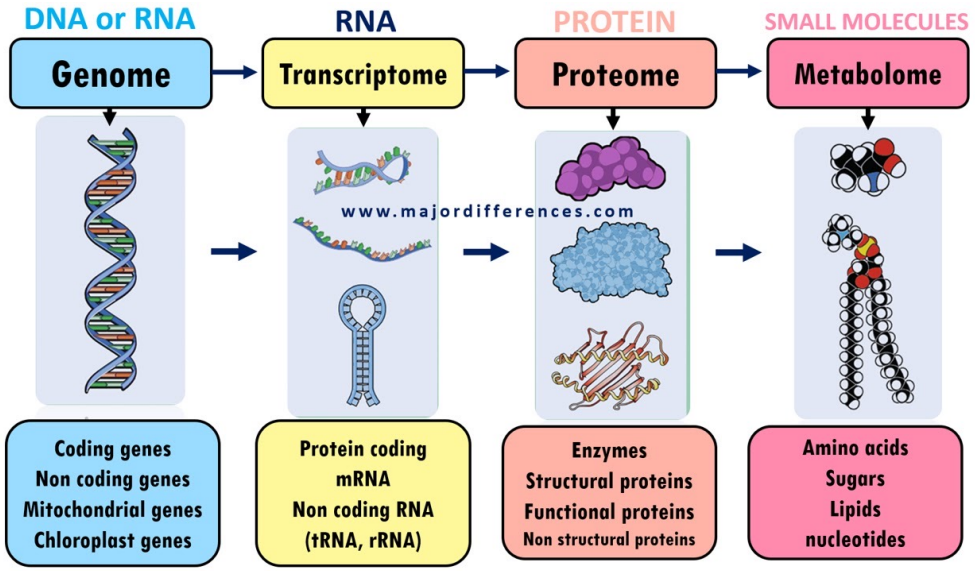
Transcriptomics – study of transcriptome which is involved in the function of cells, tissues, or organisms, across a wide range of biological conditions.

Source: modified from https://doi.org/10.1007/978-3-030-32300-4_9

Comparison

Genomics

Genomics is an interdisciplinary field of science that studies the entire genome.



Source: modified from https://doi.org/10.1007/978-3-030-32300-4_9

Transcriptomics

The transcriptomics includes the post-transcriptional era; hence, the alterations that cannot be detected at the genomics level can be revealed in the transcriptomics level.

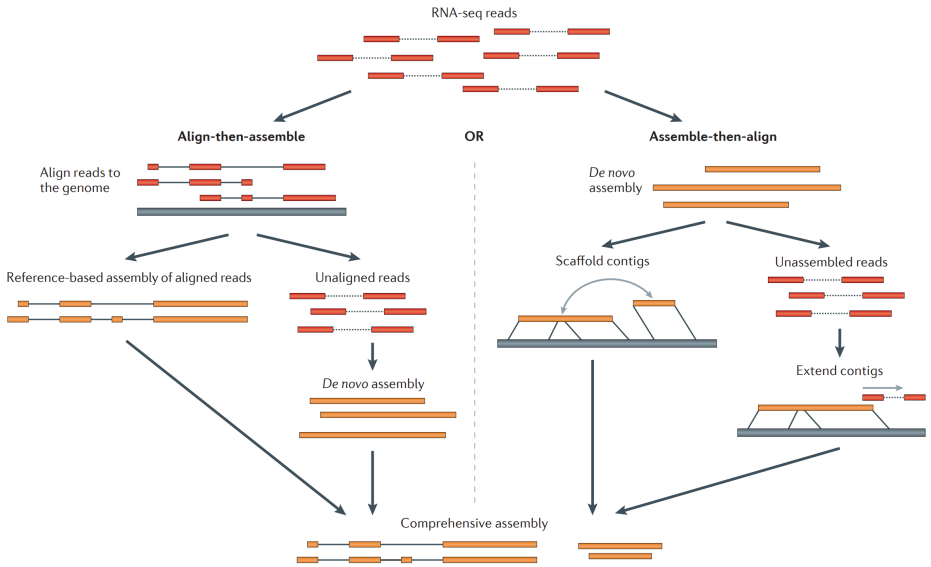
RNA-seq

Flowchart and library preparation

How to study RNA-seq

Qualitative

Transcriptome reconstruction (equvivalent to assembly)

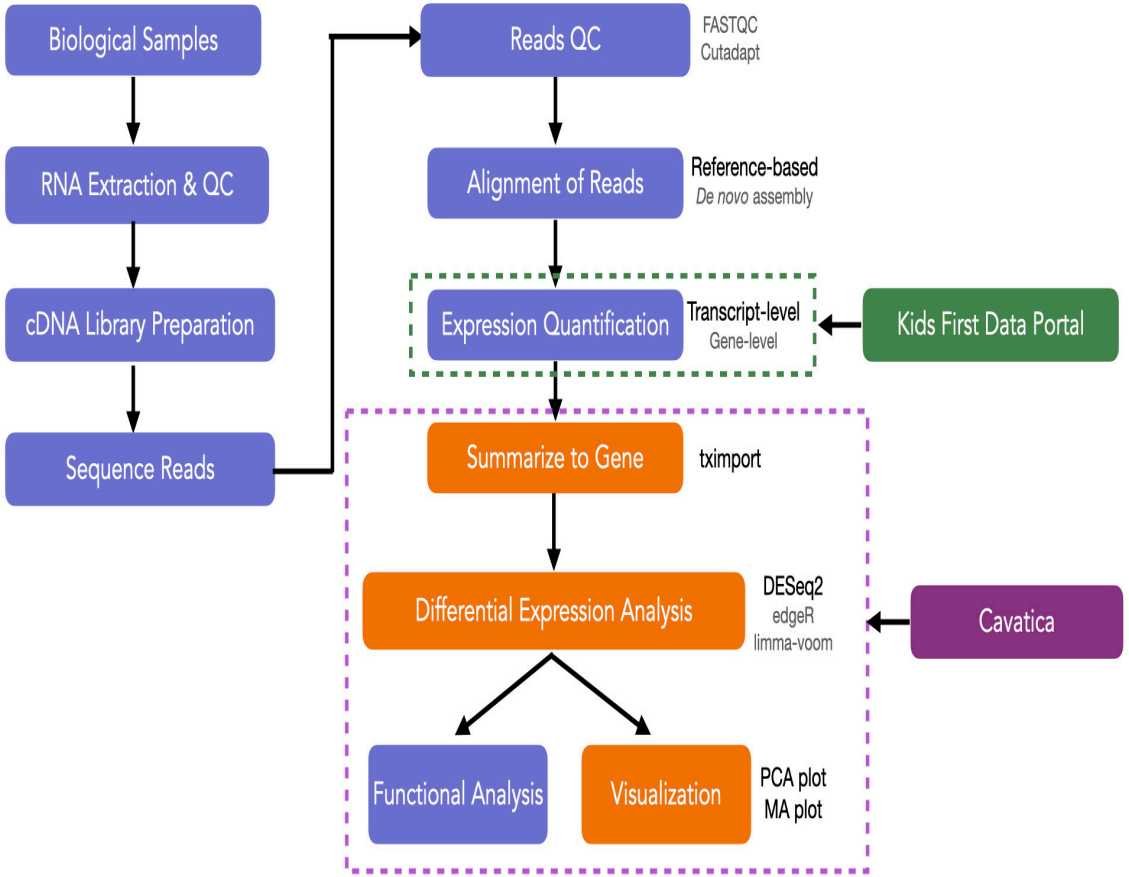


Source: https://doi.org/10.1038/nrg3068

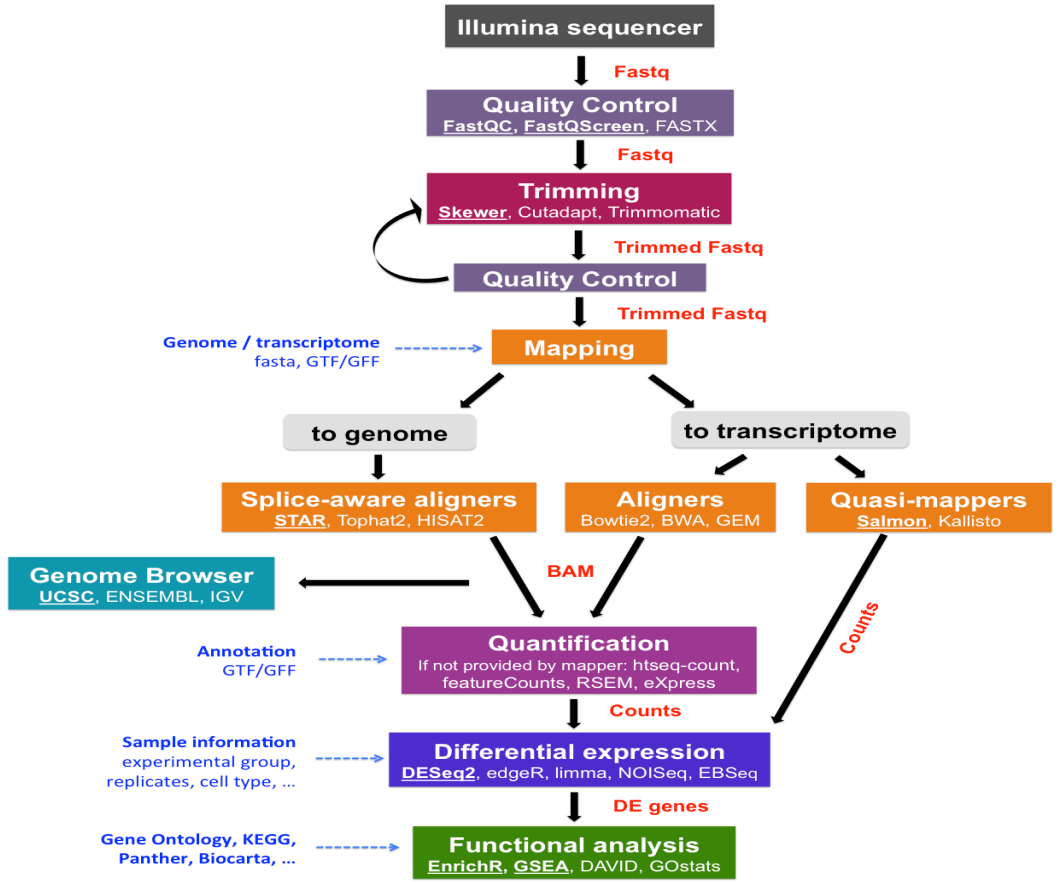
Quantitative

Evaluate differential gene expression

RNA-seq analysis flowchart



Source: Common Fund Data Ecosystem Training



Source: RNAseq_course_2019 biocerg

RNA-seq library prep

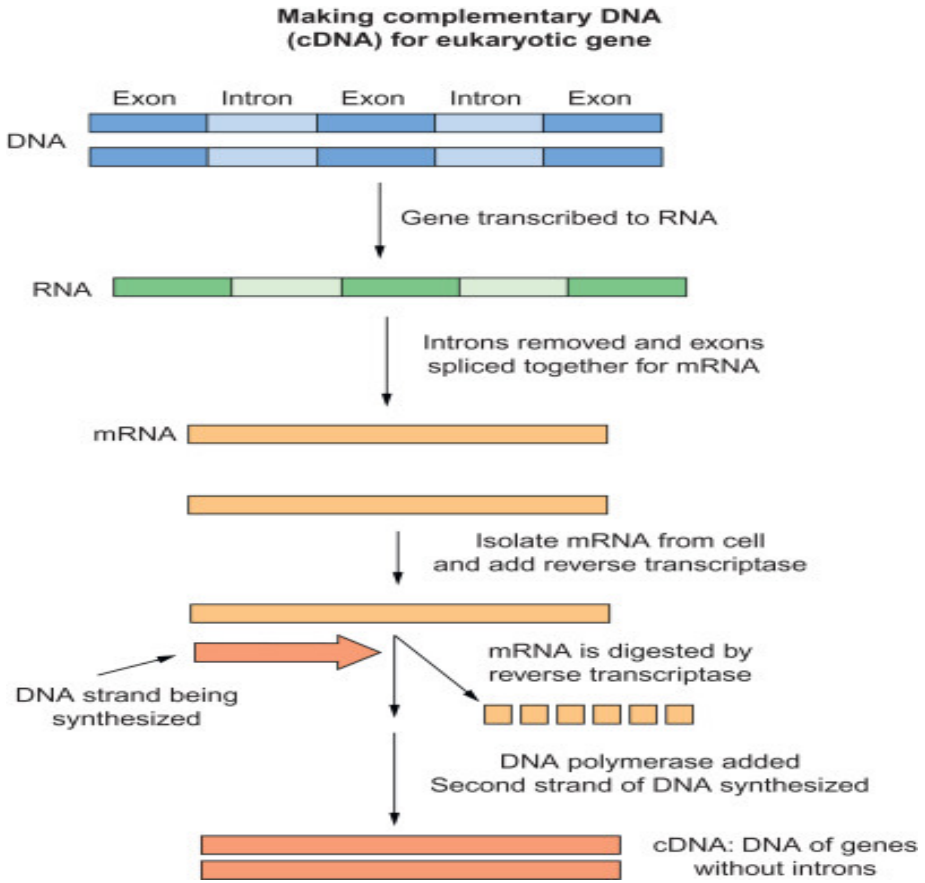
Step 1: Isolate the RNA from cells

Step 2: Break the RNA into small fragments

Step 3: Convert the RNA fragments into double stranded DNA

We do this because RNA transcripts can be thousand of bases long, but the sequencing machine can only sequence short (200-300bp) fragments

cDNA instead of mRNA



Step 3: Convert the RNA fragments into double stranded DNA

Double stranded DNA is more stable than RNA and can be easily amplified and modified. This leads us to the next step…

Source: https://doi.org/10.1016/B978-0-12-383864-3.00010-7

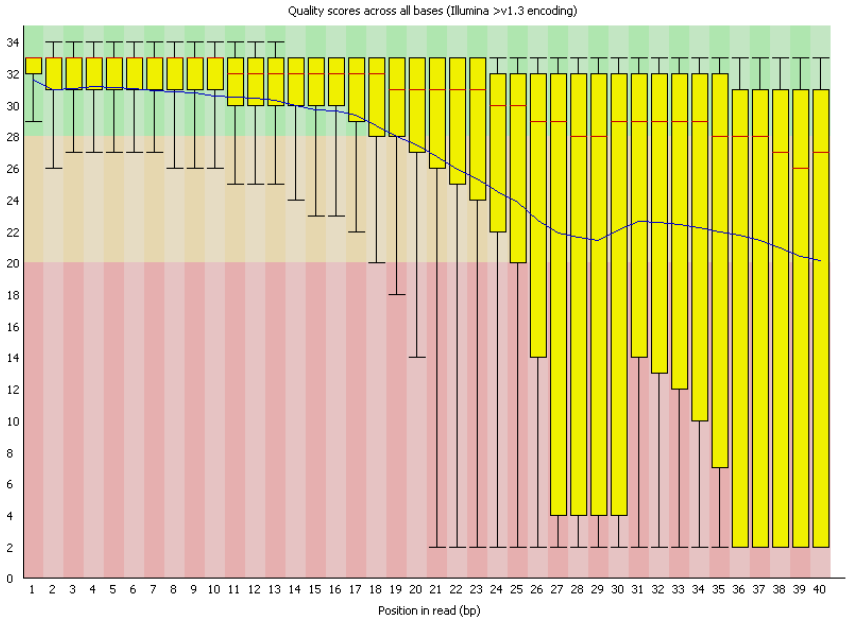
RNA-seq

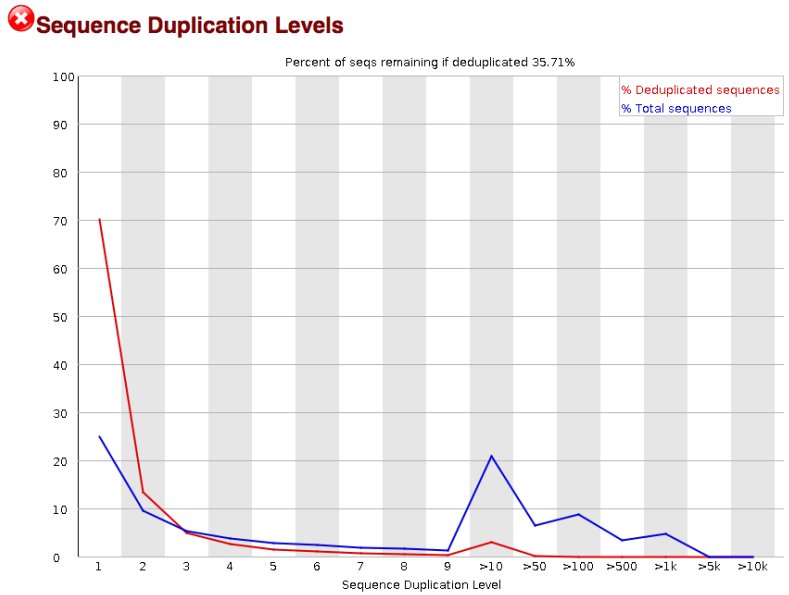
Quality control, read trimming and splice-aware alignment

SevenBridges®

Quality control of raw data

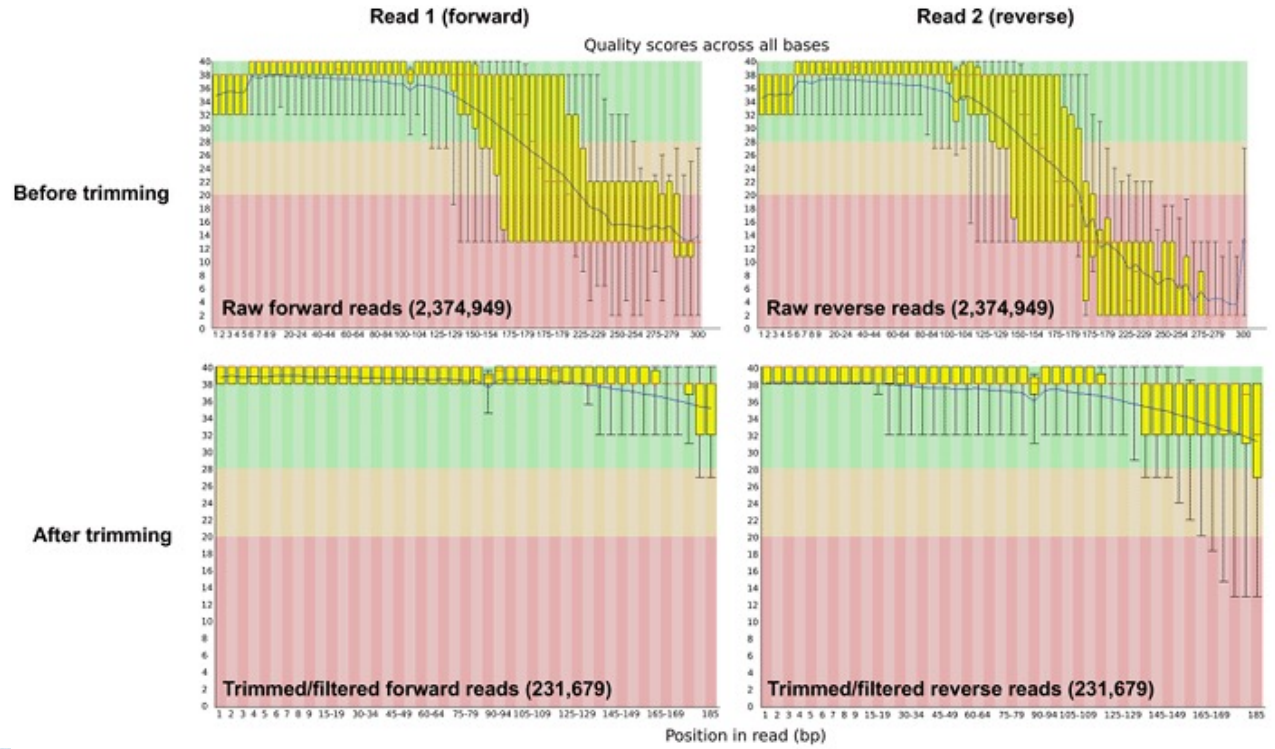
- Step which helps to quickly identify poor-quality samples in addition to flagging data issues.
- Read quality, read duplication rate, GC content…





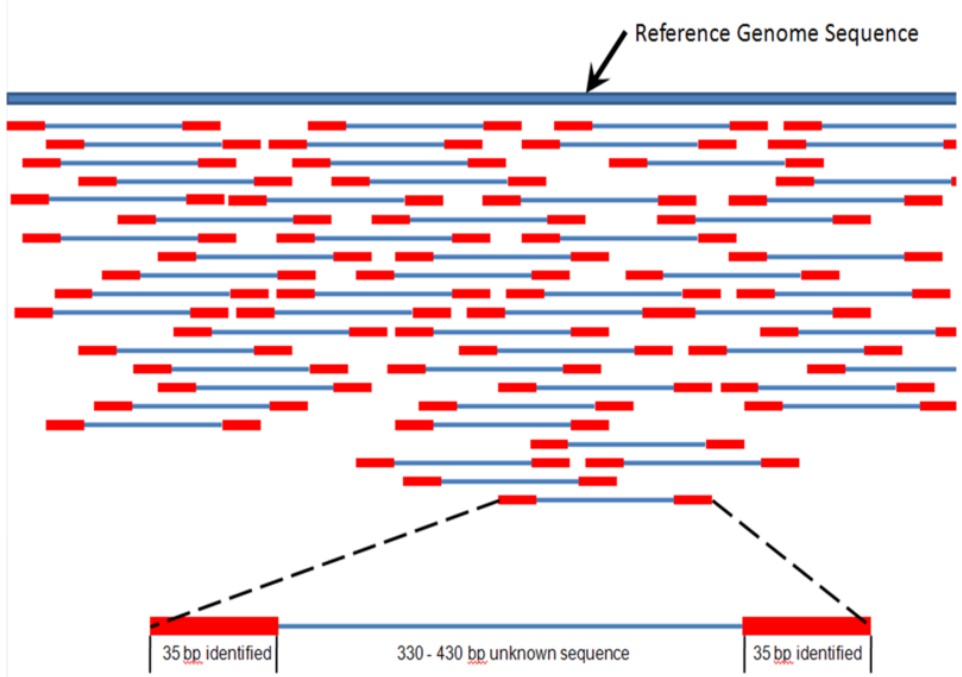
Trimming of raw data

- This step removes low-quality reads, adapter sequences and bases that fall below a certain quality threshold.
- Trimmed by sequence or by quality threshold



Reference genome/transcriptome

- A human genome reference sequence is an accepted representation of the human genome sequence that is used by researchers as a standard for comparison to DNA sequences generated in their studies.
- The Reference Transcriptomic Dataset (RTD) is an accurate and comprehensive collection of transcripts originating from a given organism.



Source: Wikipedia

GTF (gene transfer format)

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG00000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG00000169..

Building genome/transcriptome index

- Commonly used step in bioinformatics analysis
- Building index allow us to jump right to the correct place in the file and pull out just the information we need without reading much of the file
- It implies that having index save our time and improve performance

Mapping vs. Alignment

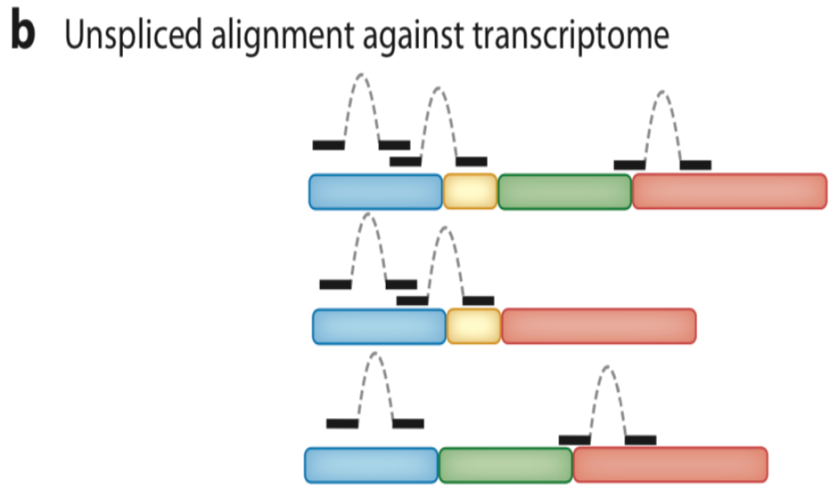
- **Mapping** - this is a term which refers to finding the approximate origin of the sequence
- **Alignment** - this is a term which refers to finding the exact difference between two sequences

Aligners used in RNA-seq analysis

- Fast (splice-unaware) aligners - align to reference transcriptome.
- Splice-aware aligners - align to reference genome
- Quasi-mappers (alignment-free mappers) - map to reference transcriptome.

Unspliced alignment

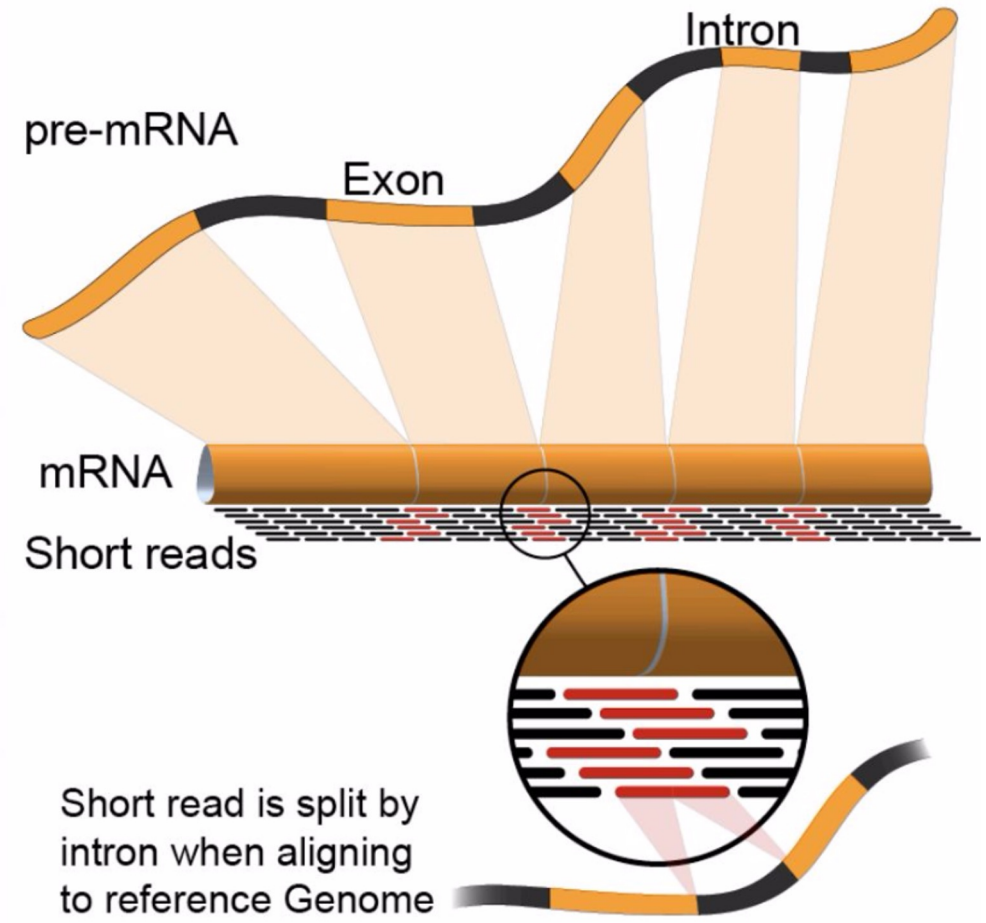
- Aligning short reads to reference transcriptome
- If aligning to a genome these tools would not map reads to splicing junctions
- Fast but less sensitive method
- Can not detect novel transcripts and novel splicing events



Source: modified from Statomics SGA sequencing intro

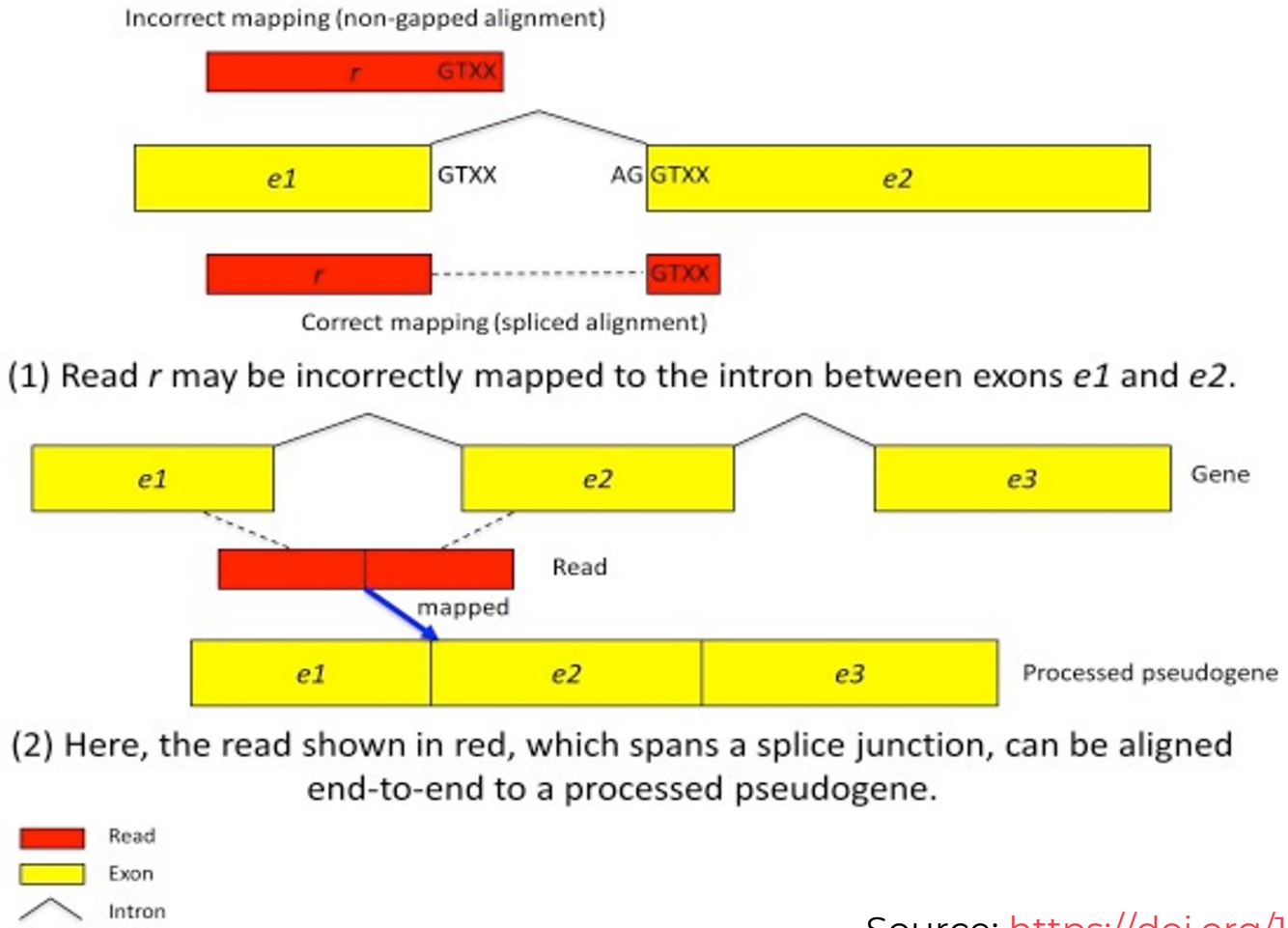
Splice-aware alignment

- Average gene size ~ 10-15 kb
- Average length of mRNA ~ 2200b
- Average exon ~ 230b
- Average number of exons ~ 9.5
- For 100b reads ~ 35% of reads would span exons
- These aligners are able to map to the splicing junctions which are present in the GTF file
- This method can detect novel transcripts and novel splicing events



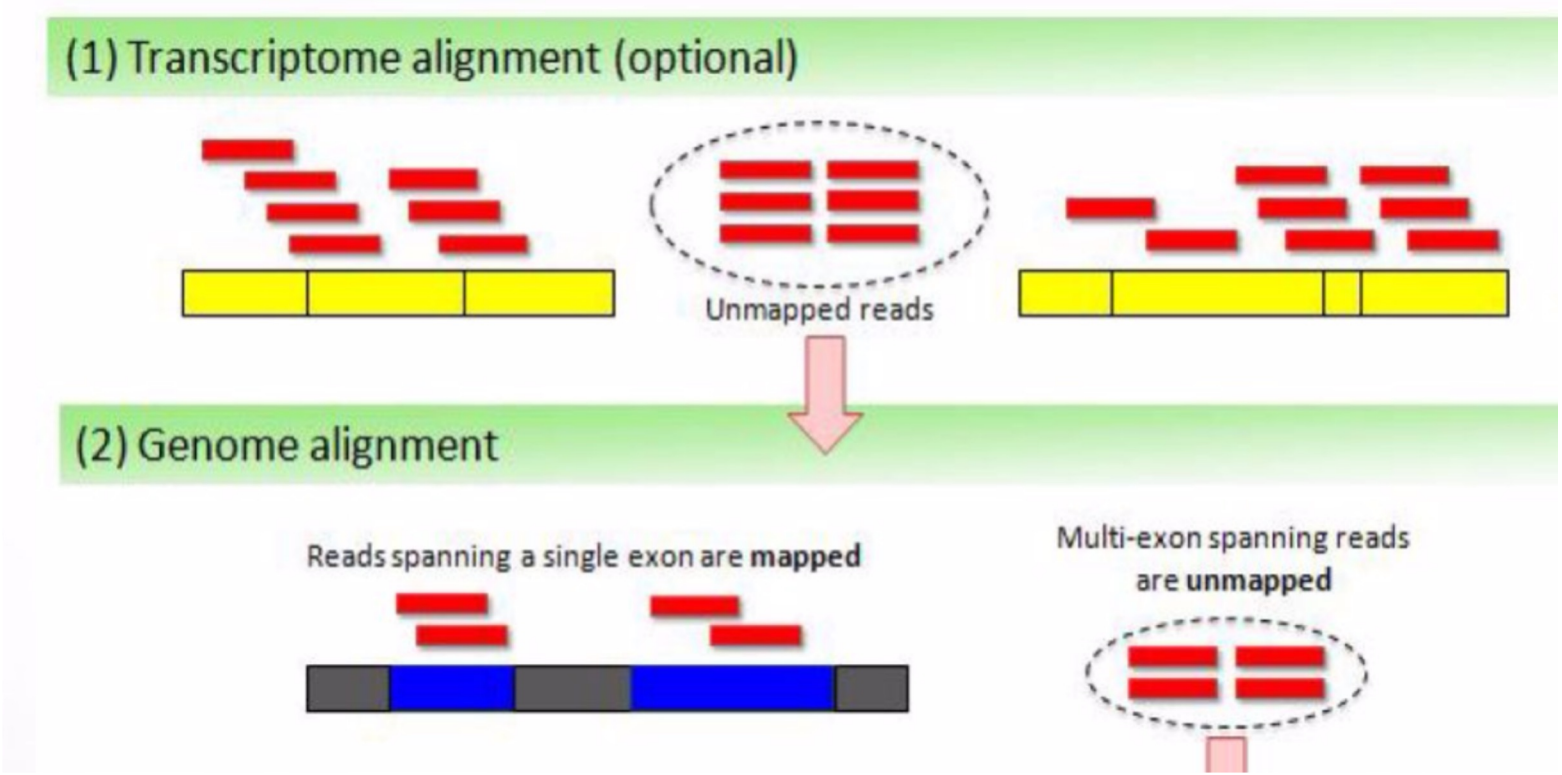
Source: https://doi.org/10.1186/gb-2013-14-4-r36

Splice-aware alignment



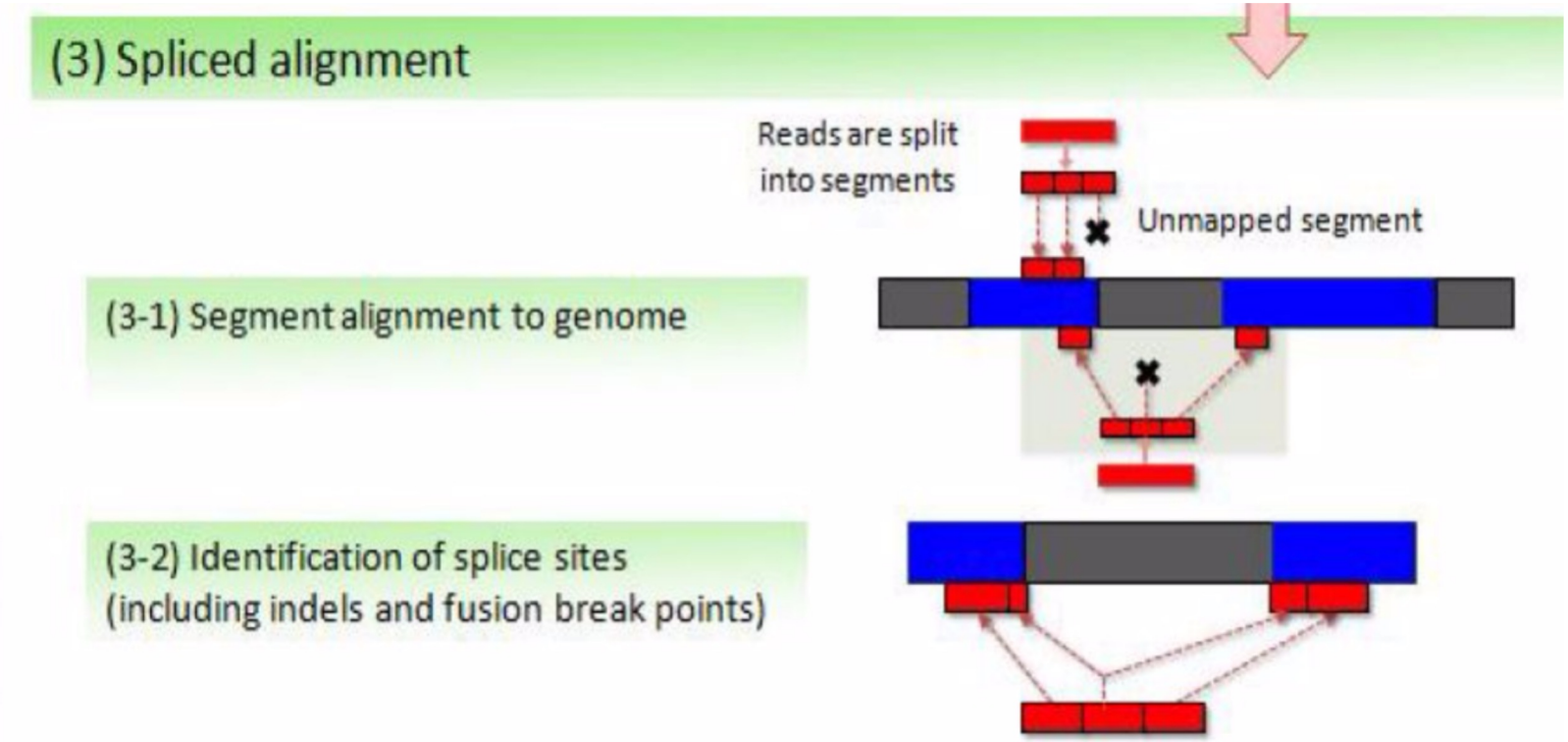
Source: https://doi.org/10.1186/gb-2013-14-4-r36

Splice-aware alignment



Source: https://doi.org/10.1186/gb-2013-14-4-r36

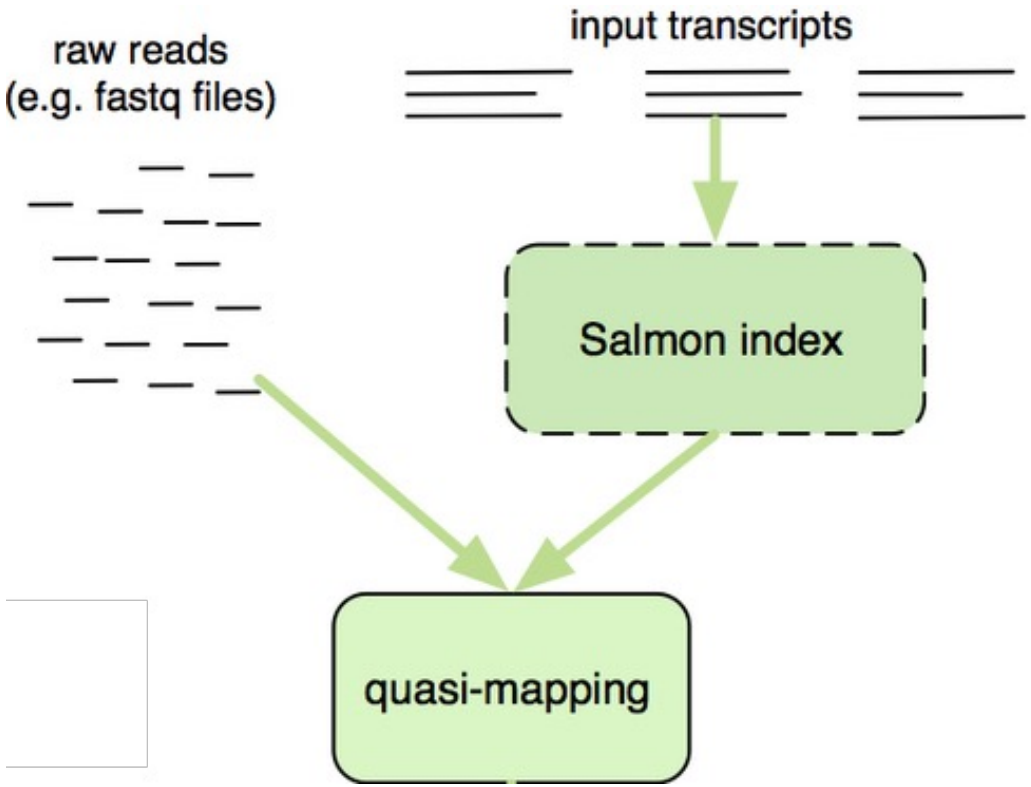
Splice-aware alignment



Source: https://doi.org/10.1186/gb-2013-14-4-r36

Quasi-mappers (alignment-free mappers)

- These tools are way faster than the previous ones because they don't need to report the resulting alignments (BAM/SAM files) but only associate a read to a given transcript for quantification.
- They don't discover novel transcript variants (or splicing events) or detect variations



Source: Introduction to RNA-Seq - hbctraining

Why do RNA-seq?

RNA-seq analysis

- RARELY: (splice-aware) alignment -> variant calling
- EVEN MORE RARELY: transcriptome assembly

RNA-seq analysis

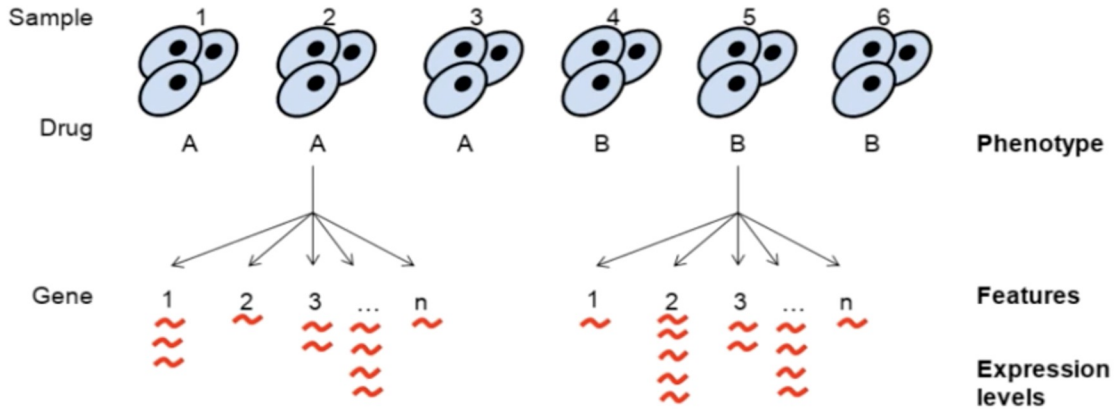
- OFTEN: **relative abundance (quantification)** of RNAs and testing for **differential expression**

New term:

- When gene products are created (through transcription and translation) we say that gene is **expressed**

Why we analyze RNA

- All cells in the body have the same DNA
- However, set of RNA molecules and thus proteins between different cell types significantly differ
- Cells also differ by the number of transcripts that are produced for one gene



Source: Differential expression analysis - datacamp

Motivation for RNA quantification

- **Quantification** is a method which estimate the relative abundance of transcripts based on the provided data
- We (usually) want to check if there is **change in transcription (expression)** between conditions (healthy/sick, treated/untreated, different tissues, etc..)

Transcriptomics

Quantification

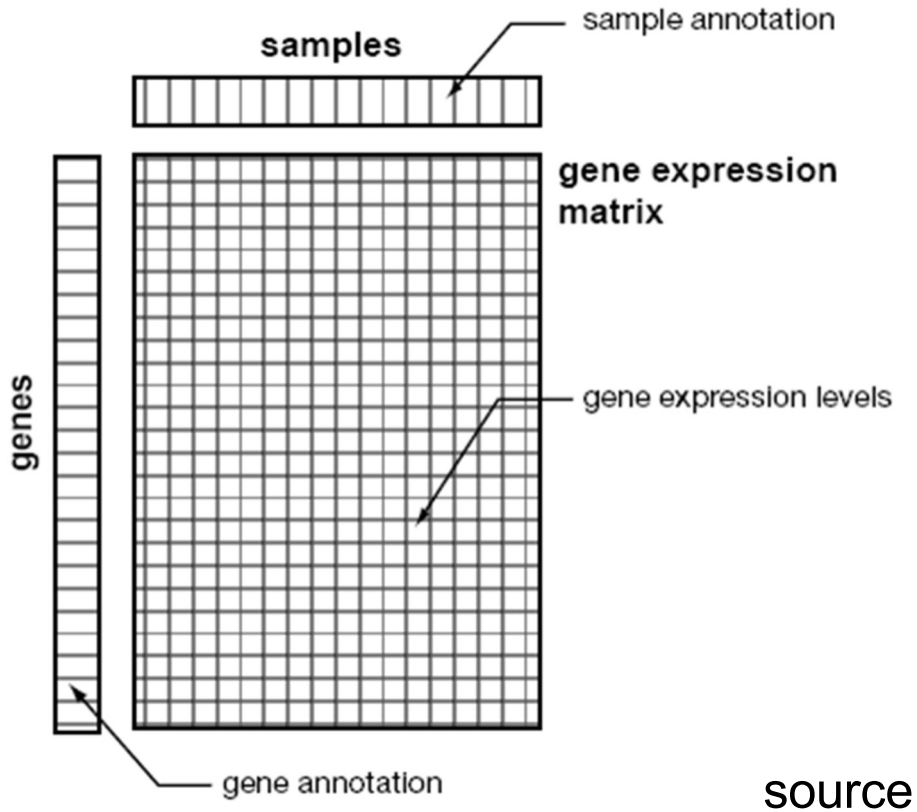
SevenBridges®

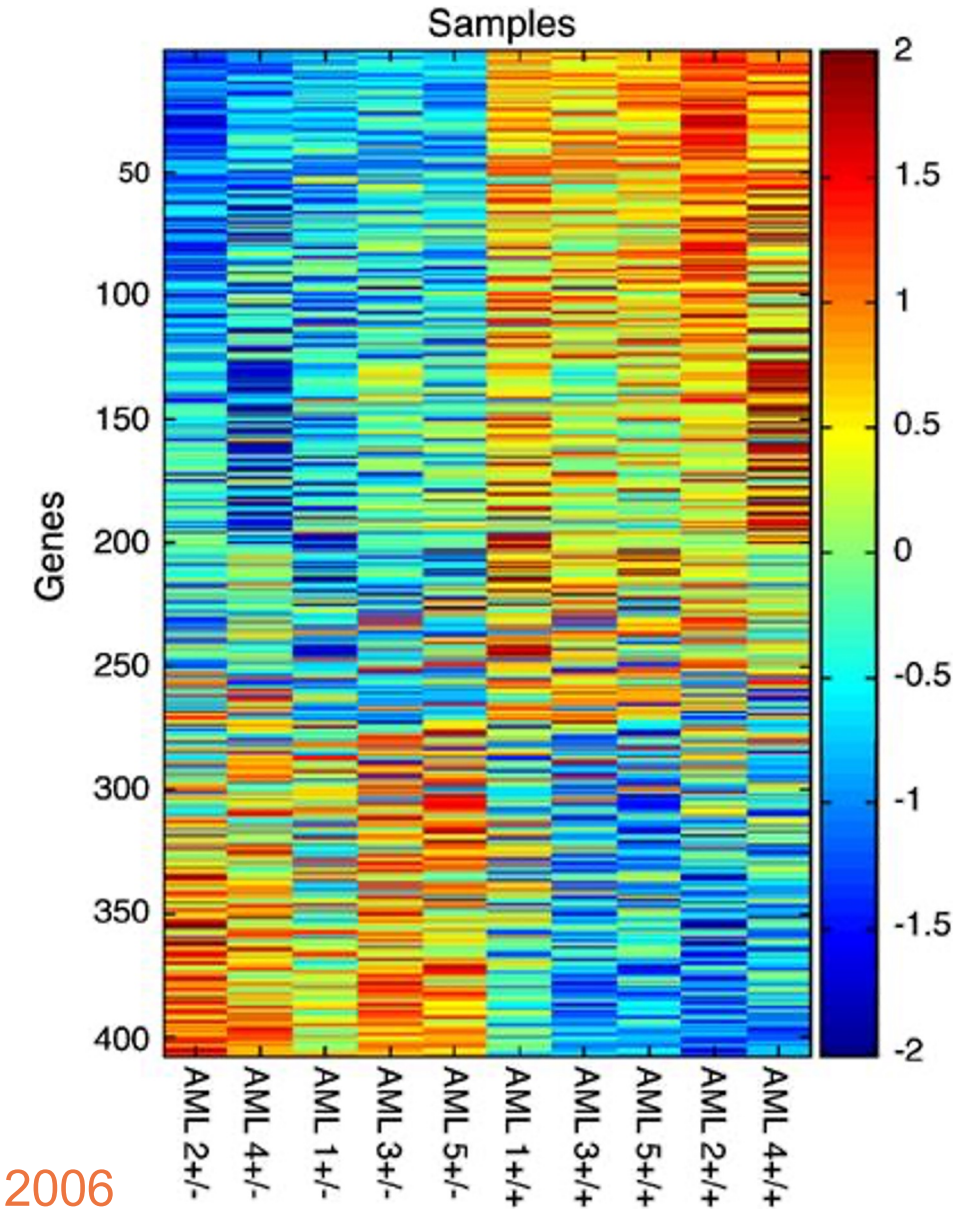
We will talk about:

- RNA quantification
- Differential expression

RNA quantification result

- Expression profiles





source: Nature Leukemia 2006

RNA-seq glossary

- Alignment-based quantification
- Alignment-free quantification
- Map to reference genome
- Map to reference transcriptome
- Gene-level analysis
- Transcript-level analysis

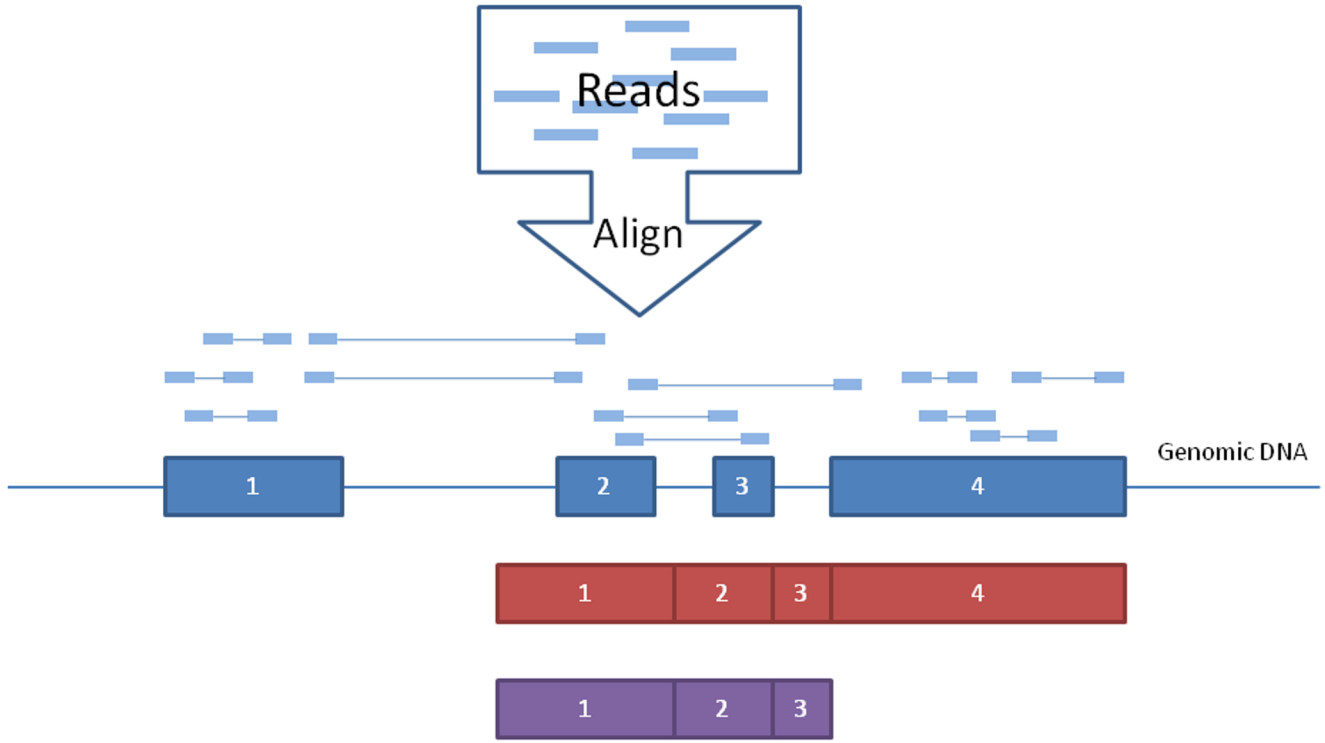
Quantification - problems

- Quantification = Counting reads?
- We can be interested in gene expression quantification (**gene counts**), but also in transcript quantification (**transcript counts**)

(1) RNA-seq: abundance estimation

Problem statement:

How to resolve alignment ambiguity?



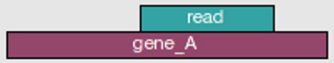
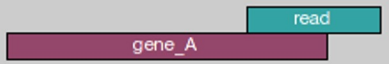
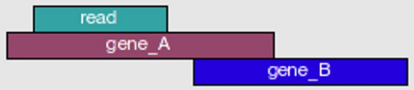
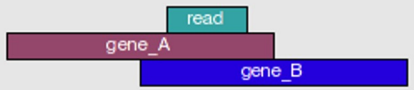
Source: http://dx.doi.org/10.13070/mm.en.3.203

(1) RNA-seq: abundance estimation

Raw counting

vs.

probabilistic estimation

	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	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

HTSeq counting model

(2) RNA-seq: abundance estimation

- For transcript quantification we usually use different probabilistic methods

- E.g. Expectation Maximization algorithm (EML or EM), Maximum Likelihood estimation

(2) RNA-seq: abundance estimation

Maximum likelihood example

$i = 5$ single-end, equal-length reads (a,b,c,d,e)

$k = 3$ transcripts (blue, green, red)

$\rho = (\rho_{blue}, \rho_{green}, \rho_{red})$ relative abundances of transcripts

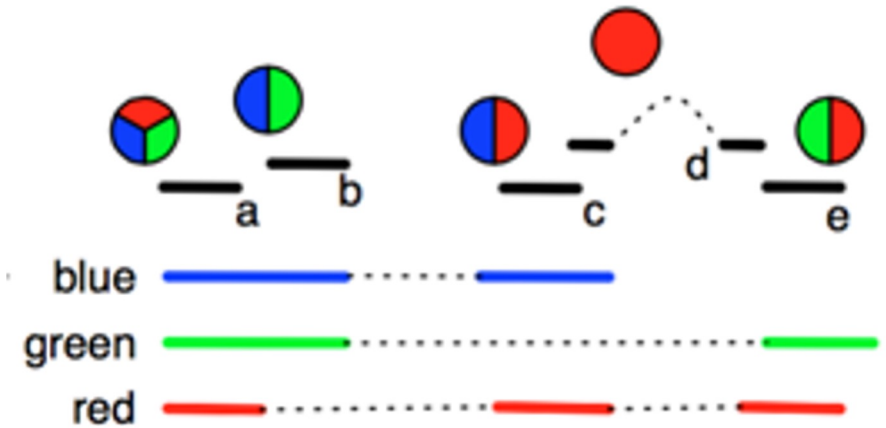
$\sum_k \rho_k = 1$, multinomial distribution

$P_i = \sum_k y_{i,k} \cdot \rho_k$, probability of detecting i-th read

where $y_{i,k} = 1$ if i-th read aligns to k-th transcript, otherwise 0

$$\boxed{L(\rho) = \prod_i \sum_k y_{i,k} \cdot \rho_k}$$

Analytical solution $\rho = (0.18, 0.18, 0.64)$



Adapted from: Lior Pachter 2011, arxiv: 1104.3889v2

(2) RNA-seq: abundance estimation

EM example

$$(\rho_{blue}, \rho_{green}, \rho_{red}) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), \text{ uniform prior}$$

E1 step: Proportional assignment

$p_a = (1/3, 1/3, 1/3),\ p_b = (1/2, 1/2, 0),$

$p_c = (1/2, 0, 1/2),\ p_d = (0, 0, 1),\ p_e = (0, 1/2, 1/2)$

M1 step: recalculate abundances

$\rho_{blue} = (1/3 + 1/2 + 1/2 + 0 + 0)/5 = 0.27$

E2 step: prior $= (0.27, 0.27, 0.46)$

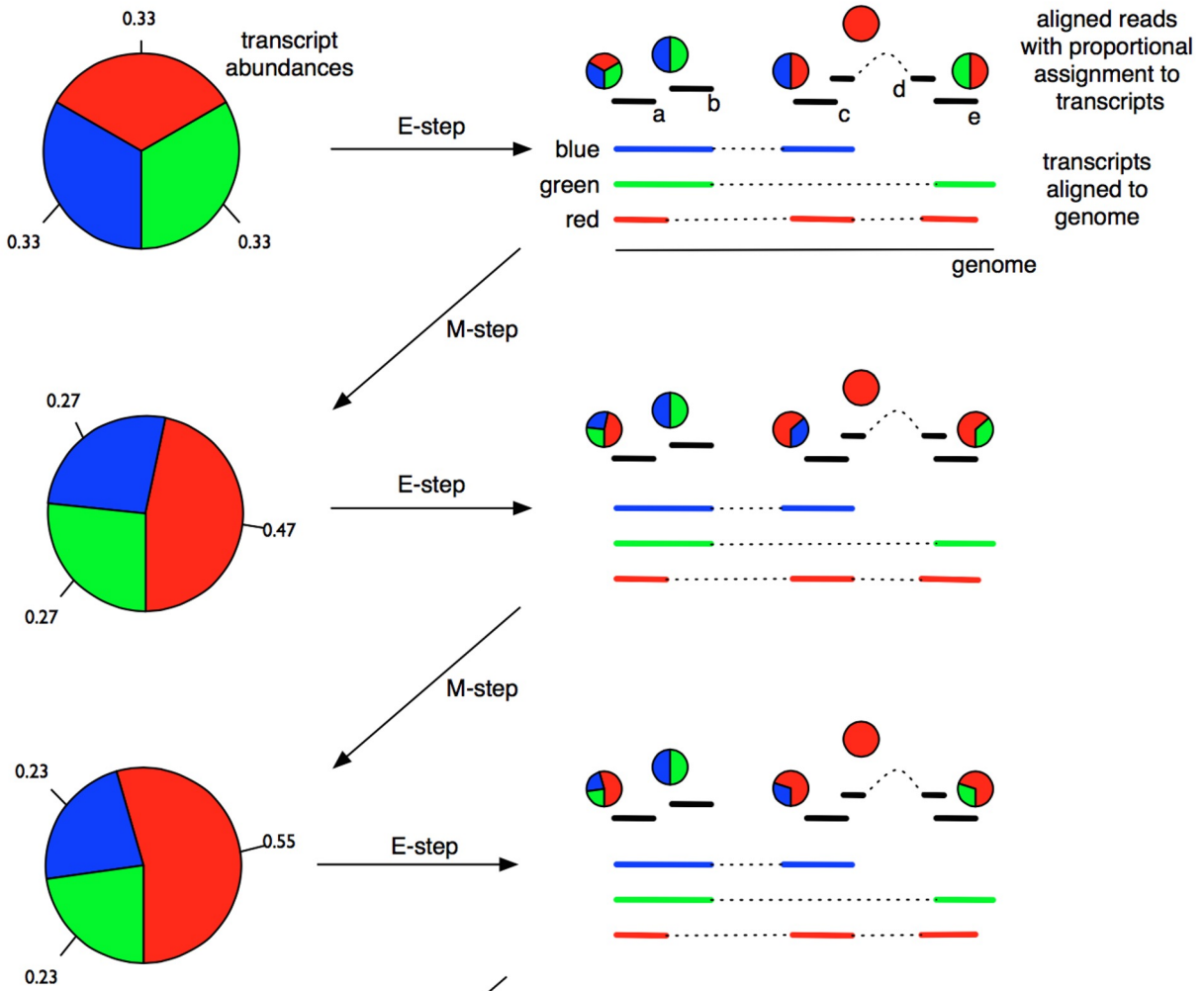
$p_a = (0.27, 0.27, 0.46),\ p_b = (1/2, 1/2, 0),$

$p_c = (\frac{0.27}{0.46 + 0.27}, 0, \frac{0.46}{0.46 + 0.27}),\ p_d = (0, 0, 1), \ldots$

M2 step:

$\rho_{blue} = (0.27 + 1/2 + 0.37 + 0 + 0)/5 = 0.23$

Iterative convergance $\rho_{blue} = 0.33, 0.27, 0.23, \ldots, 0.18$



Normalization

RNA-seq: data normalization

Problem statement:

Can we compare expression of genes (within and between samples)

if we observe reads from sampled transcripts?

Within-sample normalization

Normalisation that is performed to make gene or transript counts comparabale WITHIN ONE sample

- RPKM
- FPKM
- TPM

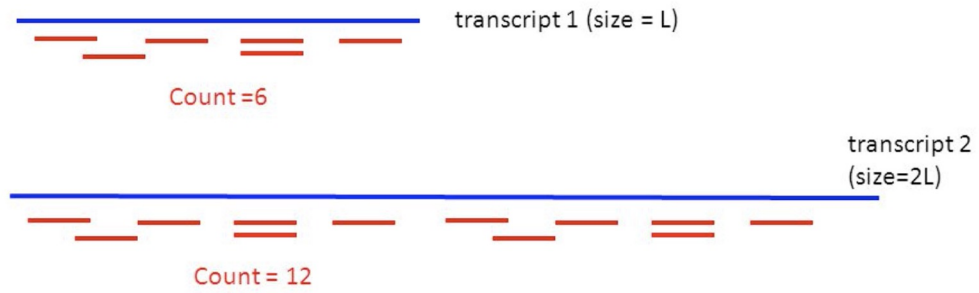
Between-sample normalization

Normalisation that is performed to make gene or transript counts comparabale ACROSS samples

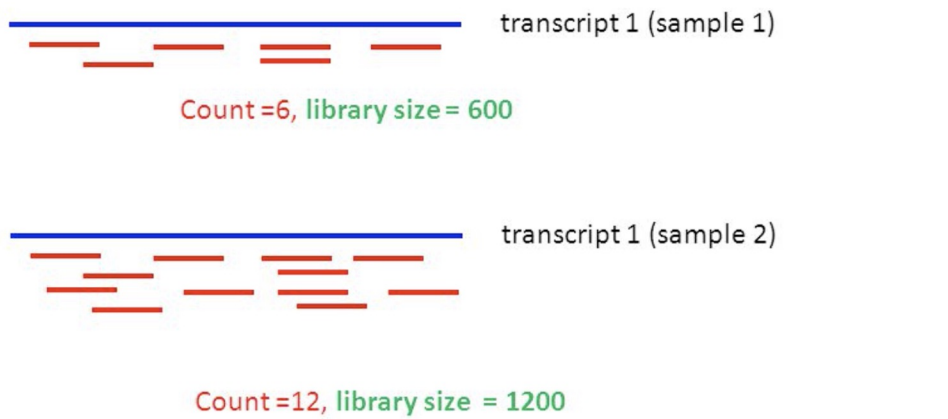
- TMM
- DESeq

RNA-seq: data normalization

One sample, two transcripts



You can't conclude that gene 2 has a higher expression than gene 1!



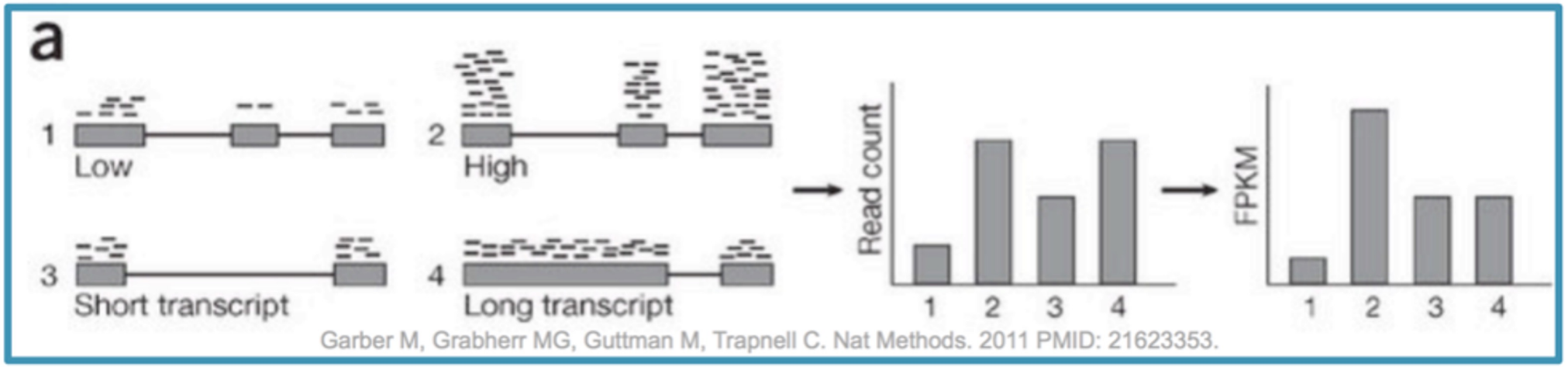
You can't conclude that gene 1 has a higher expression in sample 2 compared to sample 1!

- We need to account for **gene length** and **library size**

RNA-seq: data normalization within-sample

Let X_i be number of reads aligned to ith transcript

$$\sum_i X_i \neq \text{expression of a gene}$$



Source: 10.1038/nmeth.1613

RPKM, FPKM and TPM

- RPKM(Reads per kilobase million) - Firstly normalizes by sequencing depth and then by gene length. Used for single-end sequencing data
- FPKM(Fragments per kilobase million) - Similar to the first method but used for paired-end sequencing data
- TPM(Transcripts per million) - Firstly normalizes by gene length and the by sequencing depth

(2) RNA-seq: data normalization

Relative units (adjust for transcript length and sequencing depth):

- Transcripts per million (TPM)
- Fragments per kilobase of exon per million reads (FPKM)

$$FPKM_i = \frac{X_i}{\left(\frac{\tilde{l}_i}{10^3}\left(\frac{N}{10^6}\right)\right)} = \frac{X_i}{\tilde{l}_i N} \cdot 10^9$$

$$TPM_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\Sigma_j \frac{X_j}{\tilde{l}_j}}\right) \cdot 10^6$$

X_i - number of reads aligned to transcript ‘i’

N - total number of reads

l_i - read length

RPKM

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

RPKM

Normalize for read depth

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
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

RPKM

Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
A (2kb)	2.86	2.67	2.83
B (4kb)	5.71	5.56	5.66
C (1kb)	1.43	1.78	1.43
D (10kb)	0	0	0.09

RPKM

Normalize for gene length

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPMK
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	0	0.009

TPM

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

TPM

Normalize for gene length

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

TPM

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1
Total RPK	15	20.25	45.1
Tens of RPK	1.5	2.025	4.51

TPM

Normalize for sequencing depth

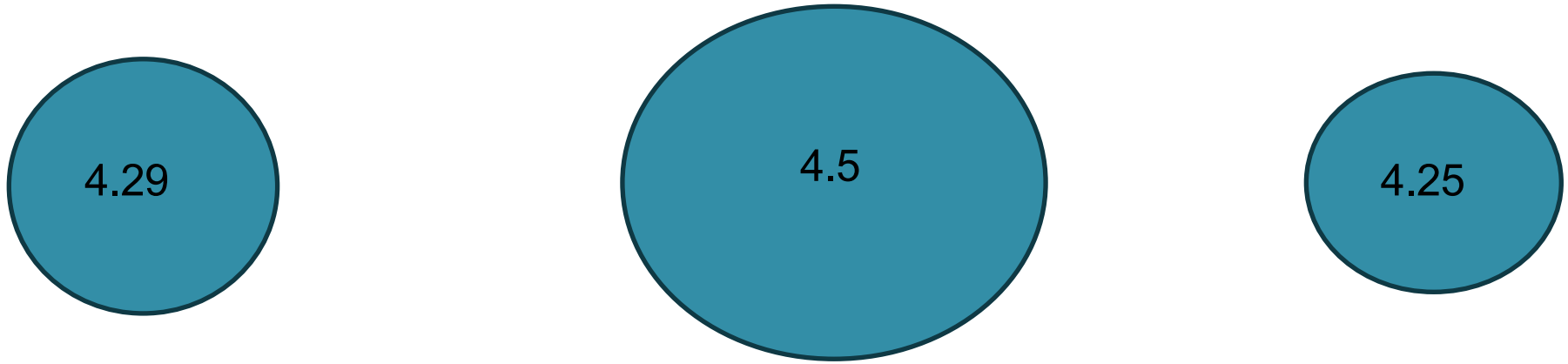
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

TPM vs RPKM (FPKM)

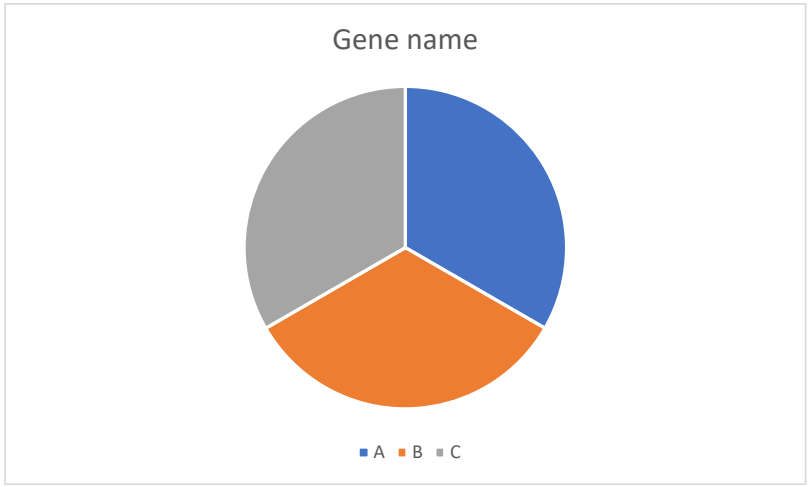
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total	10	10	10

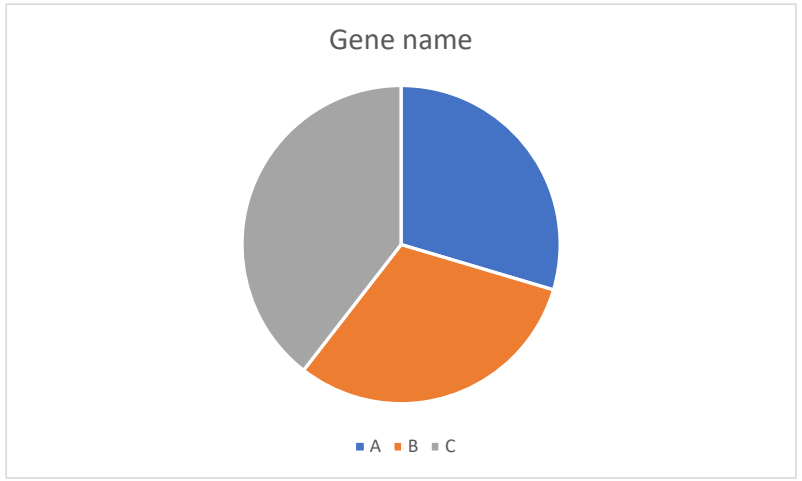
Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPMK
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total	4.29	4.5	4.25

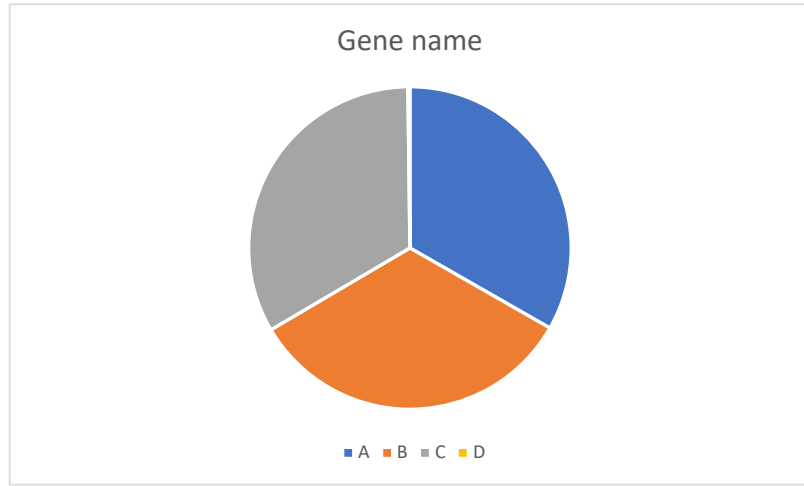
RPKM



TPM







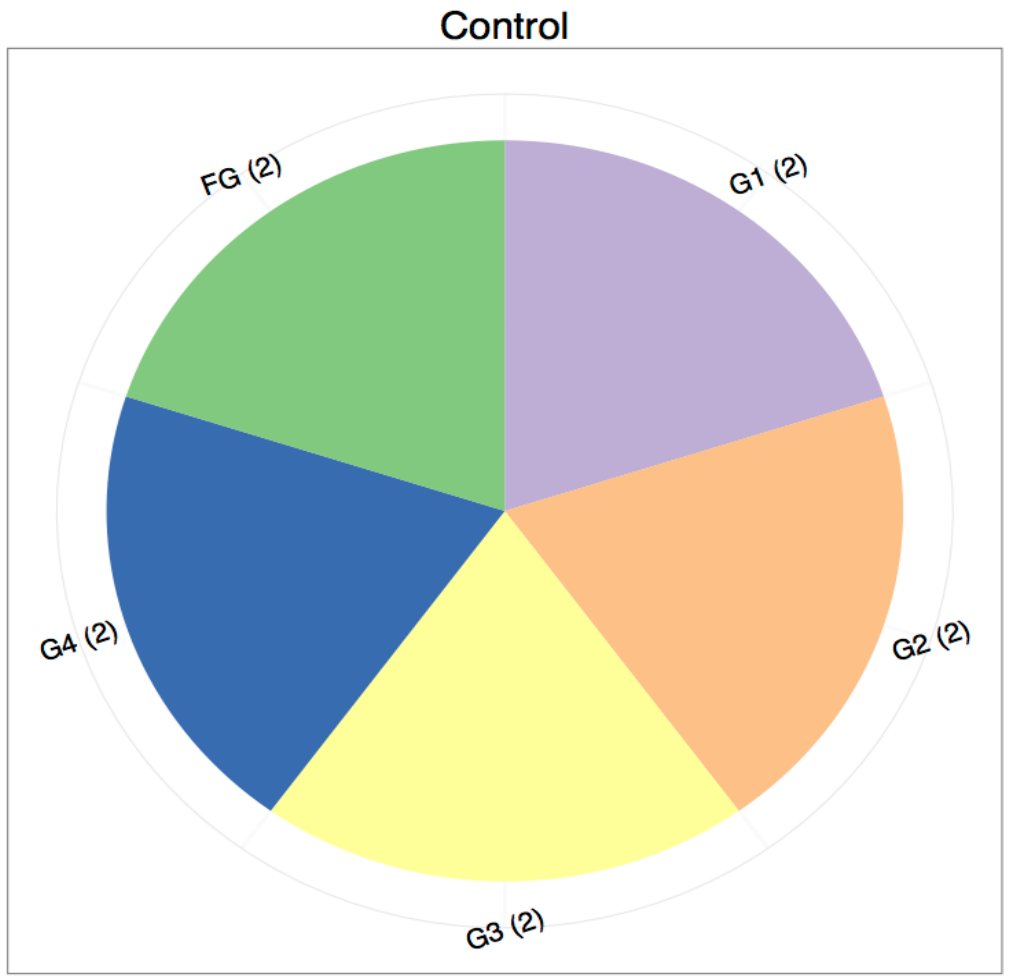
RNA-seq: data normalization between-samples

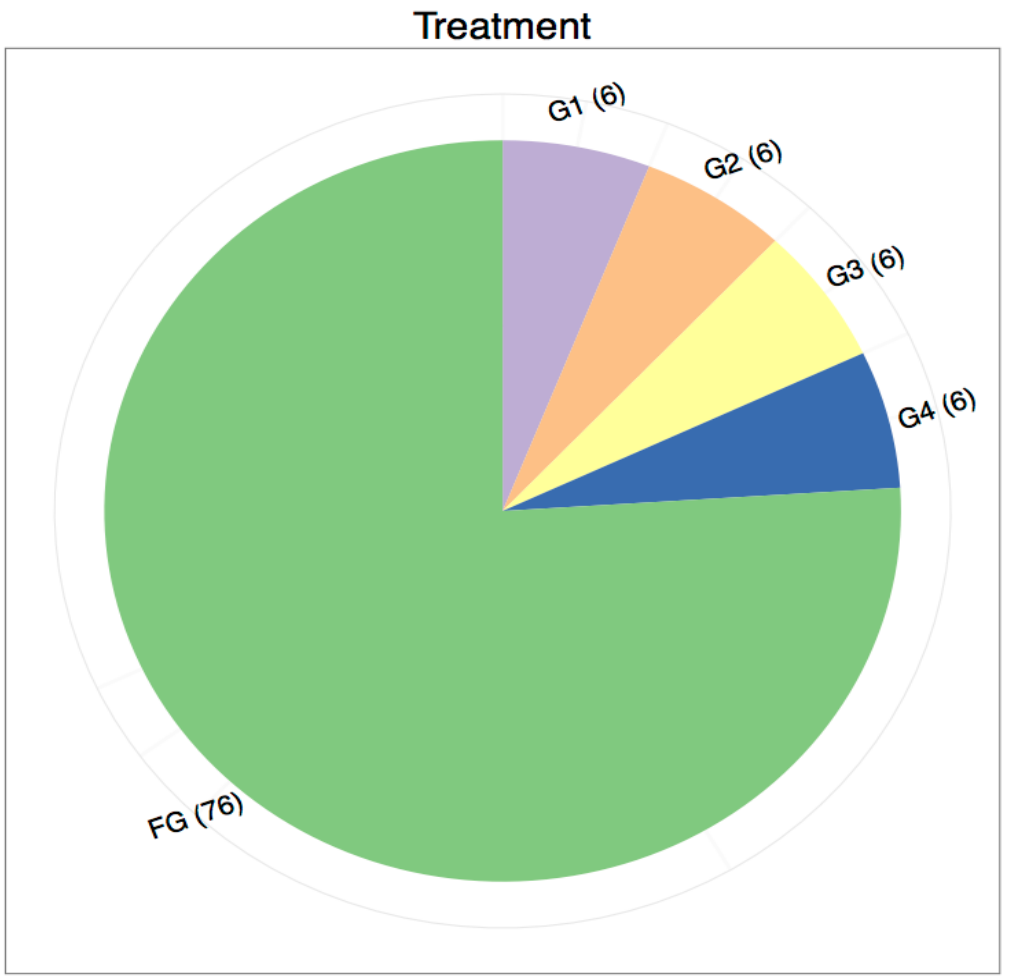
- This method adresess two issues:
  - Variable sequencing depth (the total number of reads sequenced) between experiments.
  - Finding a “control set” of expression features which should have relatively similar expression patterns across experiments (e.g. genes that are not differentially expressed) to serve as a baseline.

Between-normalization procedures/algorithms

- TMM
- DESeq – for the most of tools

Example





Example

- Here we divide number of counts for certain gene with total number of counts

Gene	Control Counts	Treatment Counts	Control Normalized	Treatment Normalized
G1	2.00	6.00	0.20	0.06
G2	2.00	6.00	0.20	0.06
G3	2.00	6.00	0.20	0.06
G4	2.00	6.00	0.20	0.06
FG	2.00	76.00	0.20	0.76

Example

Gene	Control Counts	Treatment Counts	Control Normalized	Treatment Normalized
G1	2.00	6.00	0.25	0.25
G2	2.00	6.00	0.25	0.25
G3	2.00	6.00	0.25	0.25
G4	2.00	6.00	0.25	0.25
FG	2.00	76.00	0.25	3.17

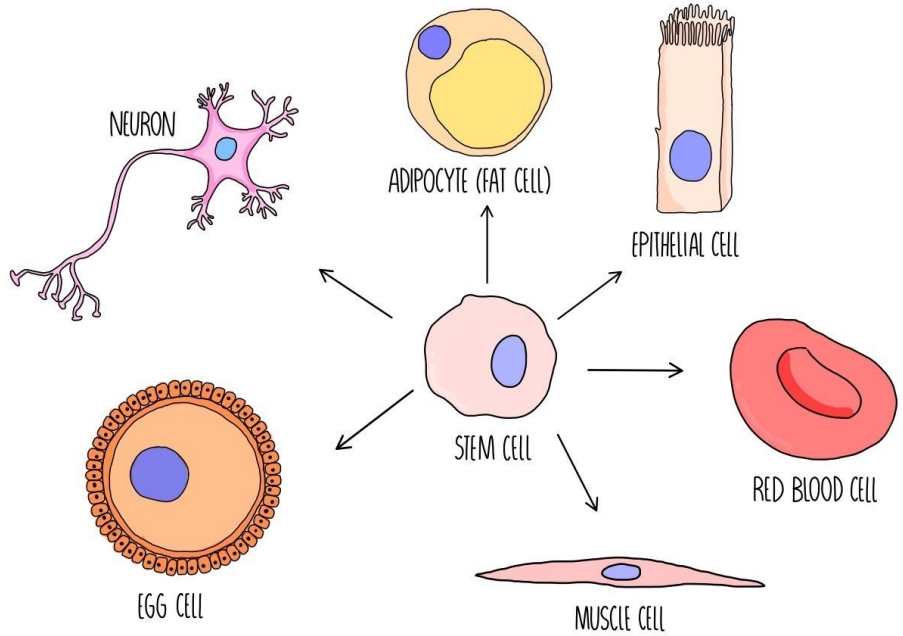
Transcriptomics

Differential expression

SevenBridges®

Differential expression

Differential gene expression is the process where different genes are activated in a cell, giving that cell a specific purpose that defines its function. When a stem cell has differentiated, it becomes a somatic cell with a specific phenotype.

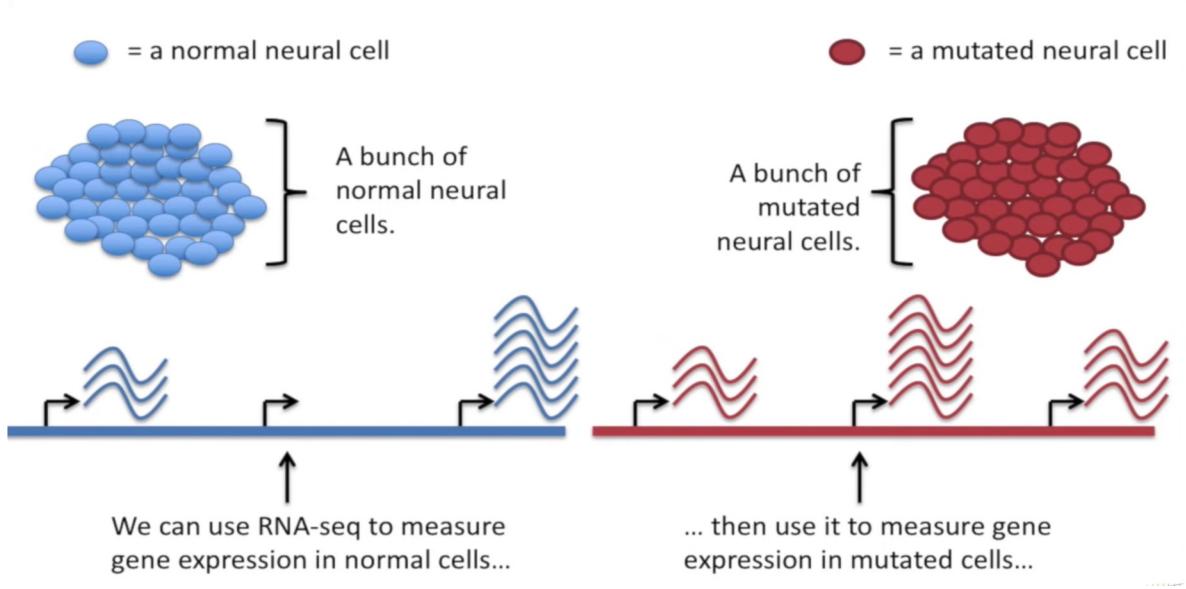


Source: The science hive

Differential expression:

Problem statement:

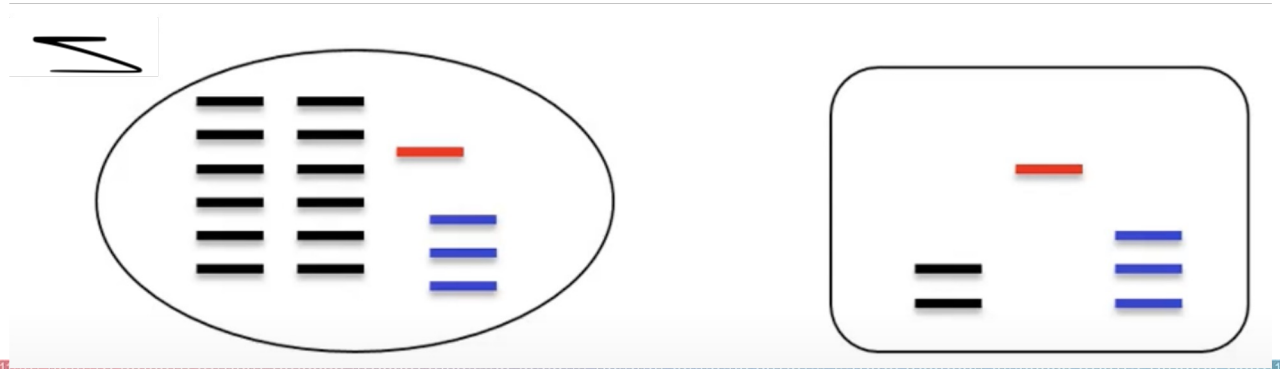
From thousands of genes, how do we know which ones are really differentially expressed and not observed changed by coincidence?

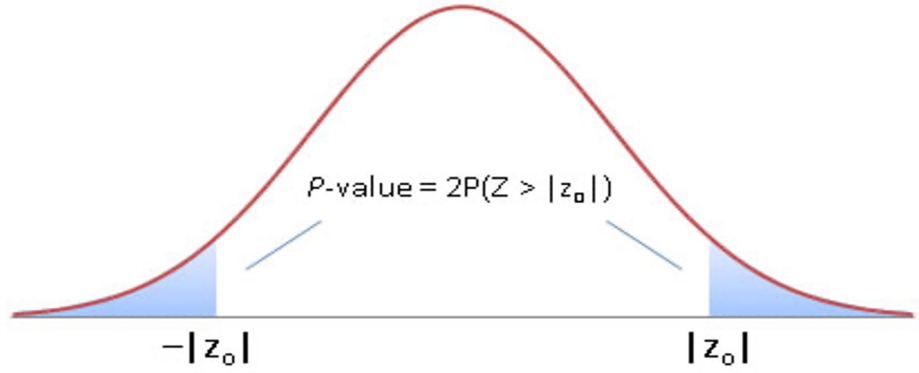


(3) RNA-seq: multiple testing

Measure of statistical significance

- **Null hypothesis**: there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.
- The **p-value** is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.
- The **alternative hypothesis** is considered true if the statistic observed would be an unlikely realization of the null hypothesis according to the p-value





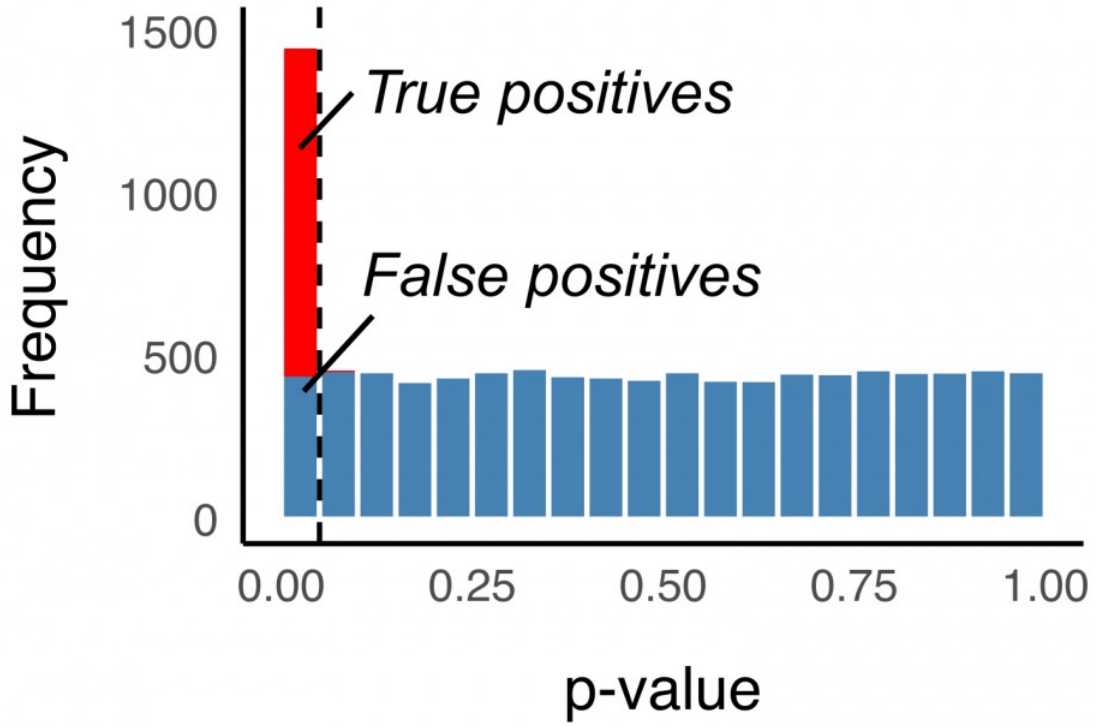
Source: RPubs

RNA-seq differential expression: choice of test

- The number of samples or replicates are usually small in a typical RNA-seq experiment
- This is the reason why parametric tests are used for estimating statistical significance for differential expression

(3) RNA-seq: multiple testing

- In genomic studies you don't usually fit just one regression model or calculate just one p-value. You calculate many p-values.
- *human_hg19_genes_2015.gtf* has about 26,000 genes and 54,000 transcripts.
- Suppose 1200 out of 20,000 genes are found significant at 0.05 level.
  - No correction: you should expect 0.05 * 20,000 = 1000 false positives
  - Solution: Multiple testing correction



Source: GeneVia Technologies

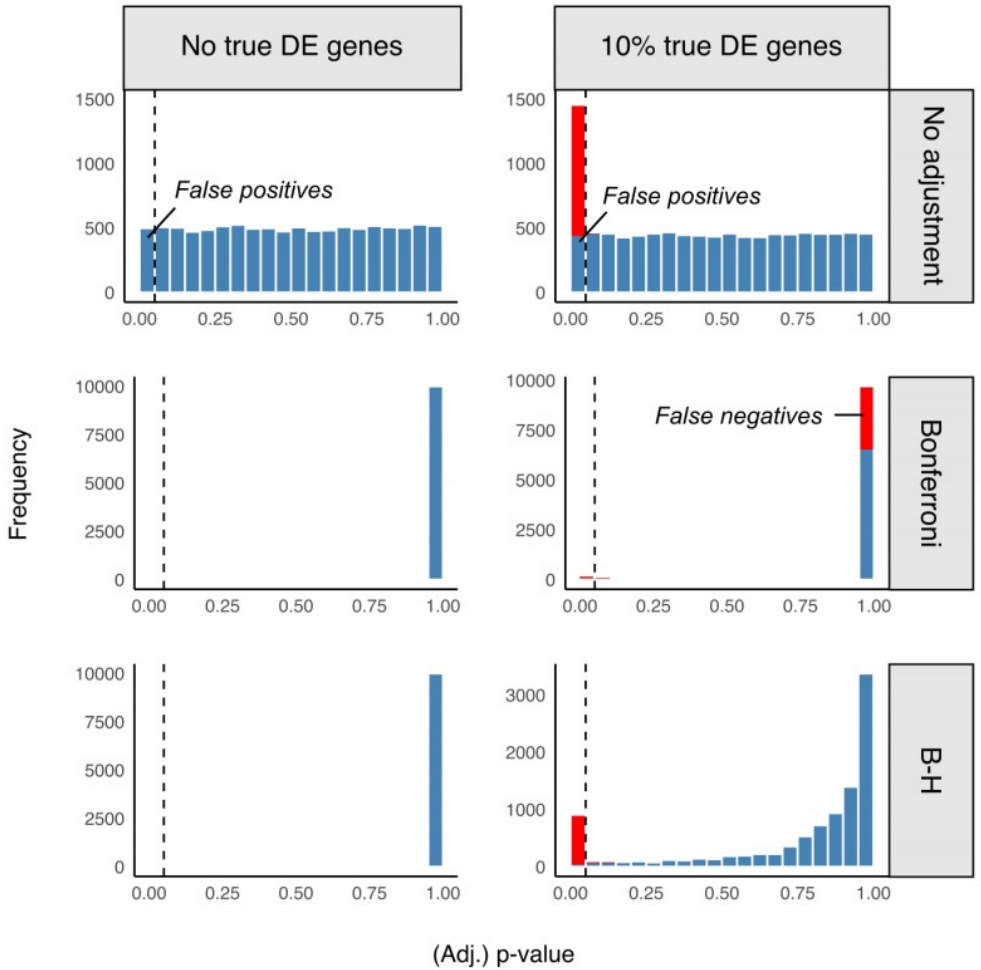
(3) RNA-seq: multiple testing

Multiple testing correction procedures:

- Bonferroni correction
  - p_value * total_number_of_tests_performed

For more info see also:

- BH (Benjamini-Hochberg) procedure
- BY (Benjamini–Yekutieli) procedure



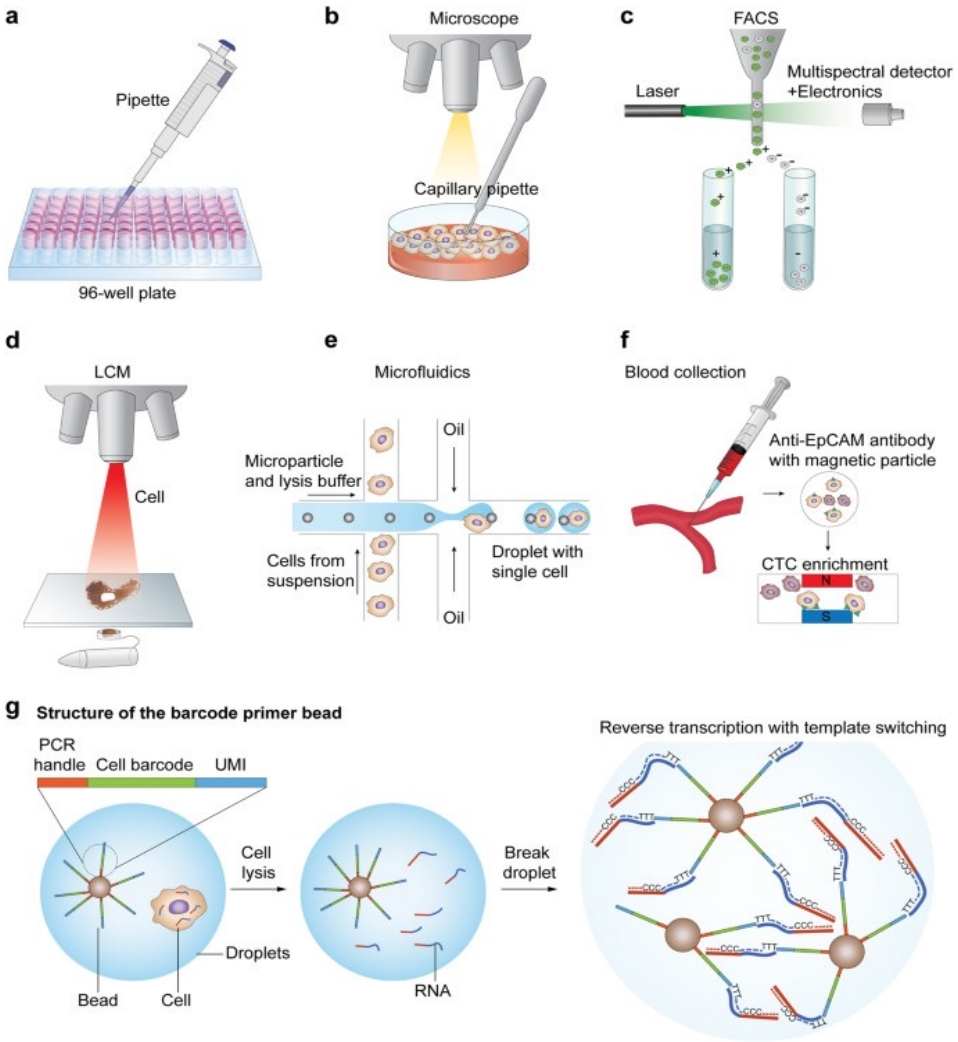
Source: GeneVia Technologies

Single-cell sequencing

SevenBridges®

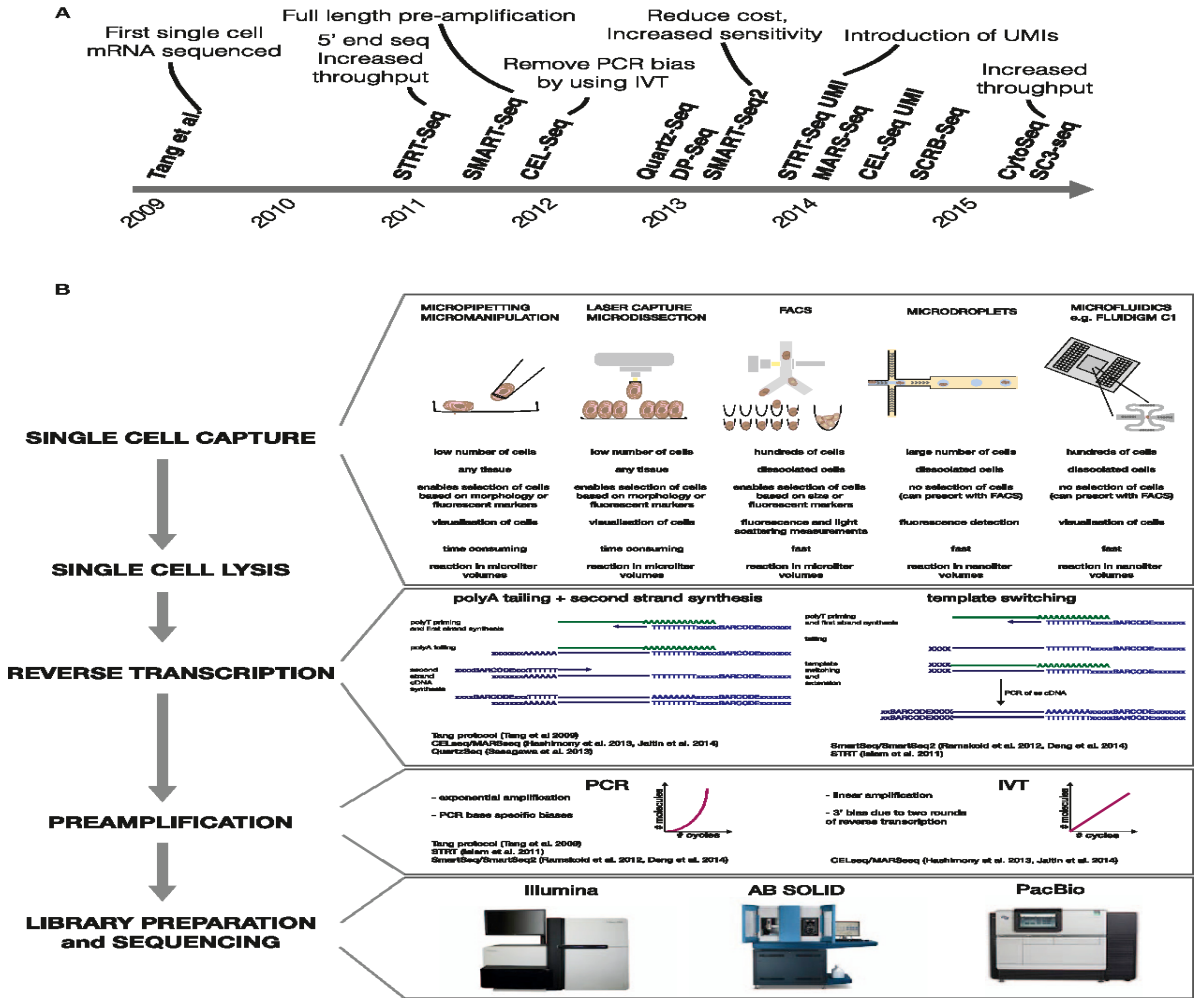
Single-cell sequencing

- Studying cells at the single-cell level offers unique opportunities to dissect the interplay between intrinsic cellular processes and extrinsic stimuli such as the local environment or neighboring cells in cell fate determination
- To date, most transcriptome studies are conducted on a ‘population level’ usually averaging the transcriptomes of millions of cells. However, in some cases such as stem cells, circulating tumor cells (CTCs) and other rare populations, sufficient material cannot be obtained for analysis on such a scale. In addition, bulk approaches fail to detect the subtle but potentially biologically meaningful differences between seemingly identical cells.



Source: [doi: https://doi.org/10.1038/s12276-018-0071-8](https://doi.org/10.1038/s12276-018-0071-8)

The flowchart of Single-cell sequencing



Source: [DOI: 10.1016/j.molcel.2015.04.005](https://doi.org/10.1016/j.molcel.2015.04.005)

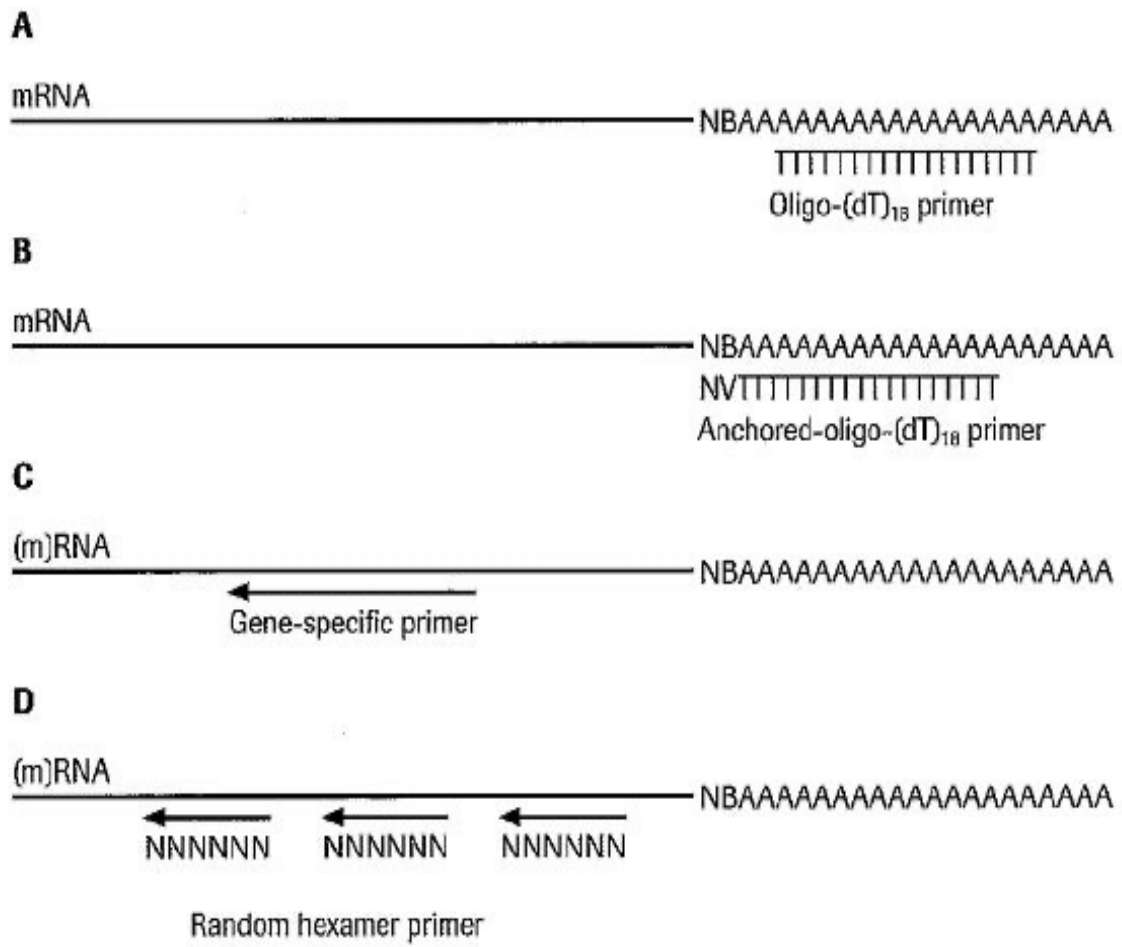
Novel era of sequencing

Direct RNA sequencing

SevenBridges®

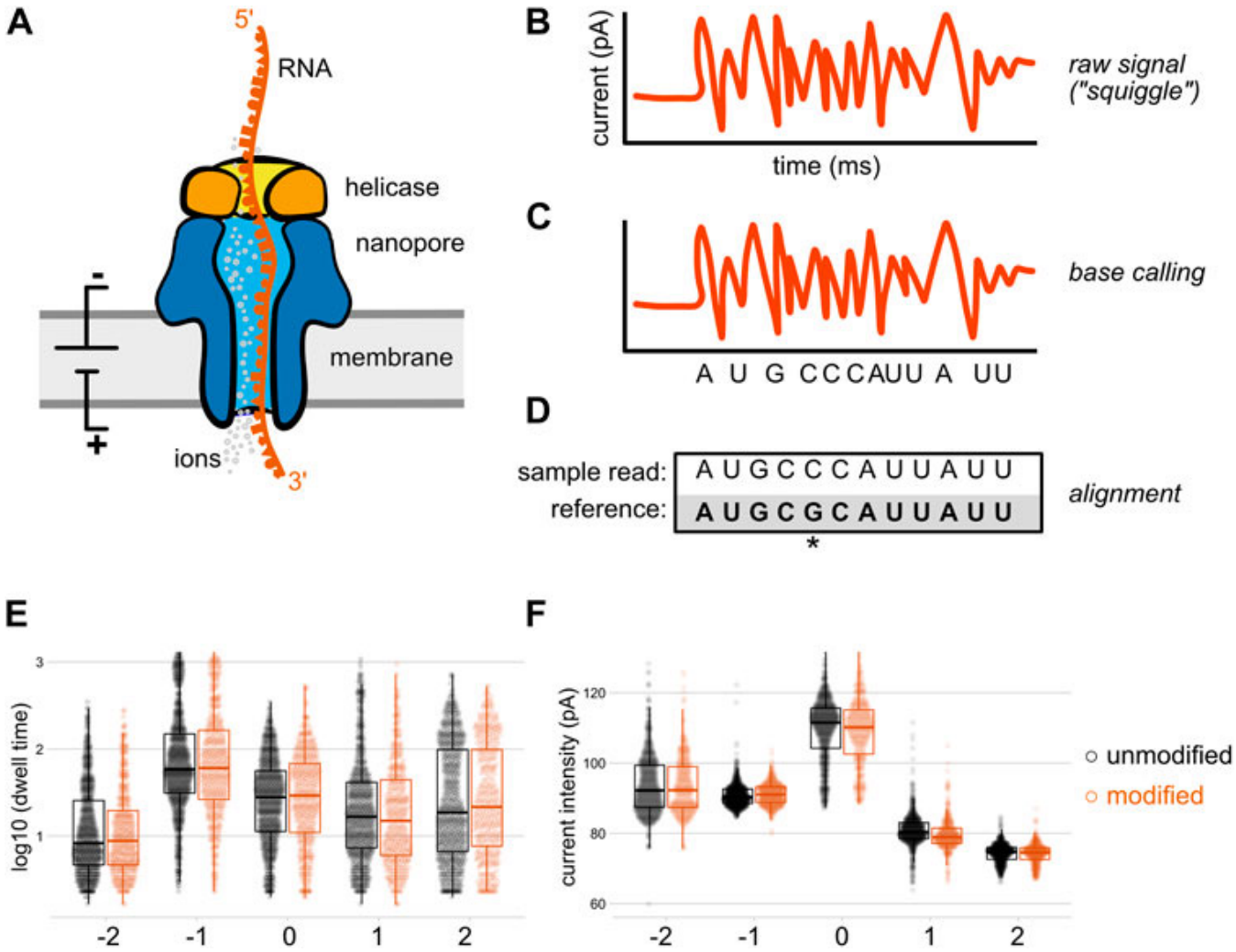
Short read RNA-seq and its limitations

- Converting RNA into cDNA (reverse transcription):
  - Not-so-random nature of random priming and negative influence of certain secondary structures on the processivity of reverse transcription
  - Above phenomena results in uneven coverage along the length of transcripts and the under- and overrepresentation in libraries of certain sequences.
- Inaccurate quantification of gene expression at the transcript level for genes with a complex array of alternatively spliced isoforms
- Average eukaryote mRNA is 2–2.5 kb while the average fragment size of RNA-seq libraries is no longer than 250–300 bp



Nanopore based direct long-read RNA-seq

- Relies on the measurement of alterations in ionic current when a nucleic acid transits a protein nanopore embedded in the membrane of a flowcell
- Good method for detecting RNA-modifications



Source:
10.3389/fgene.2022.1037134

Thank You!