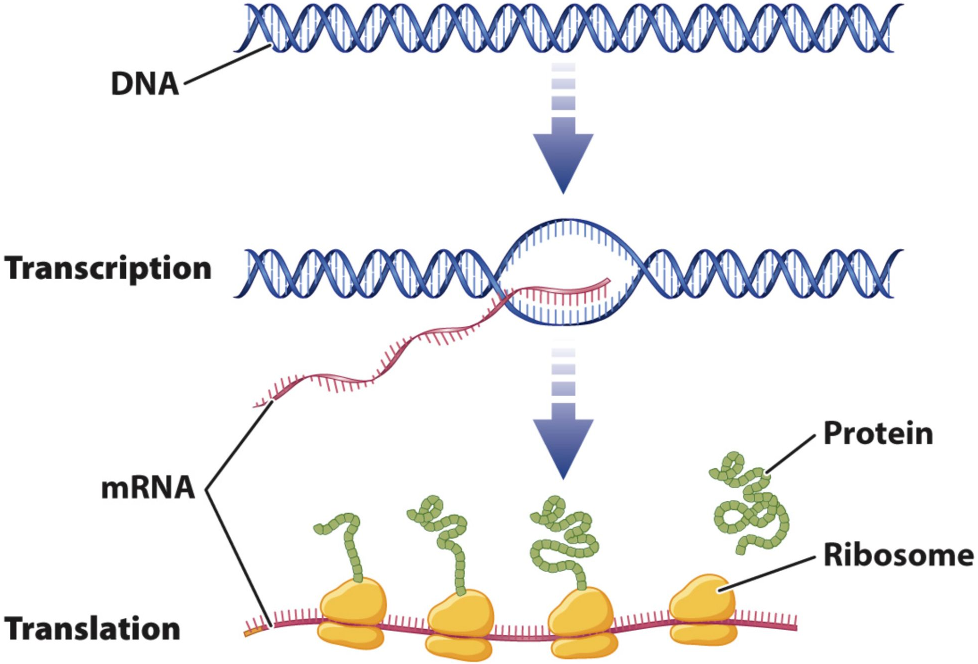
SevenBridges

Intro to RNA-seq

sevenbridges.com

Beyond DNA

Central dogma of molecular biology

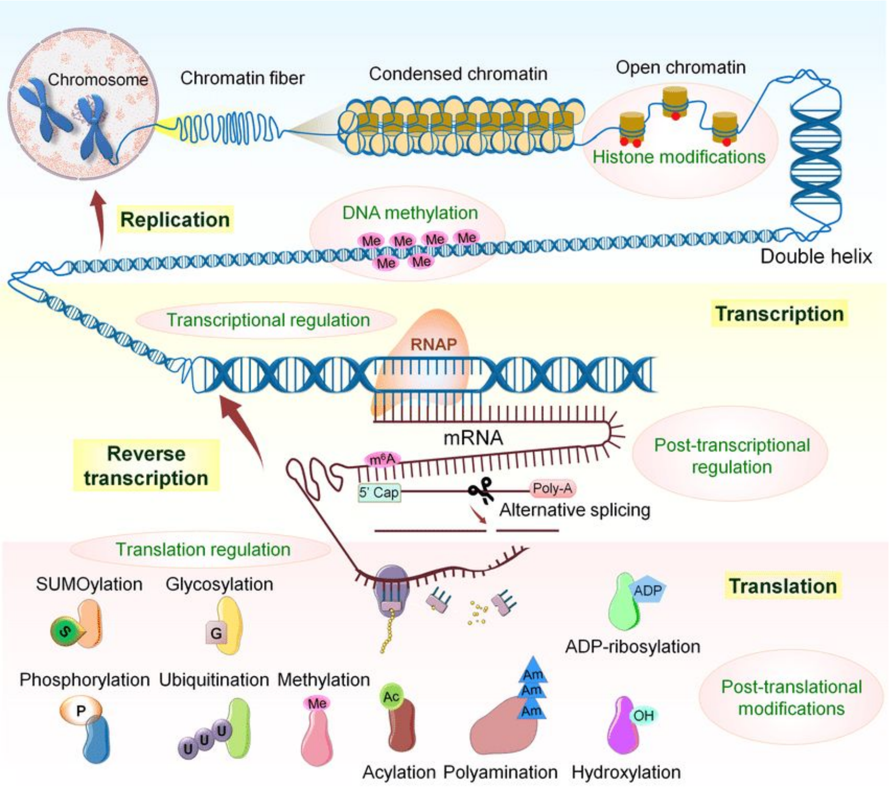


Replication - Before cell division DNA is replicated

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

Central dogma of molecular biology

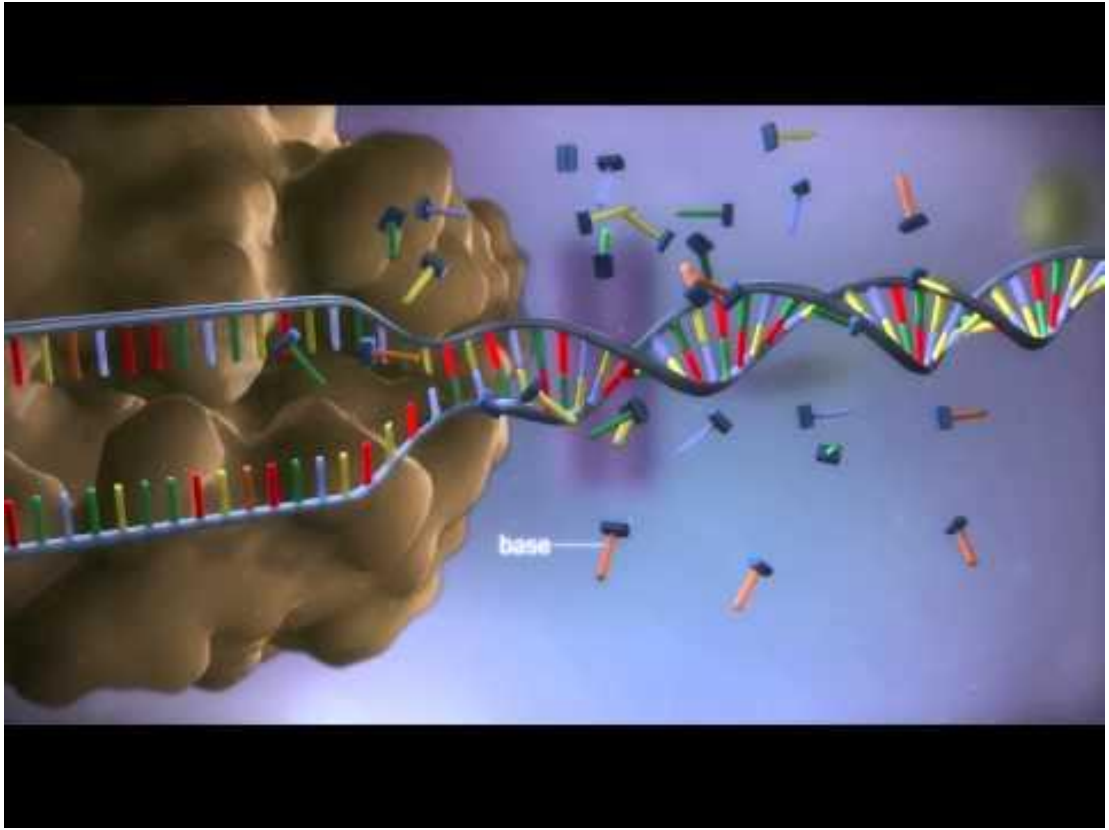


Replication - Before cell division DNA is replicated

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

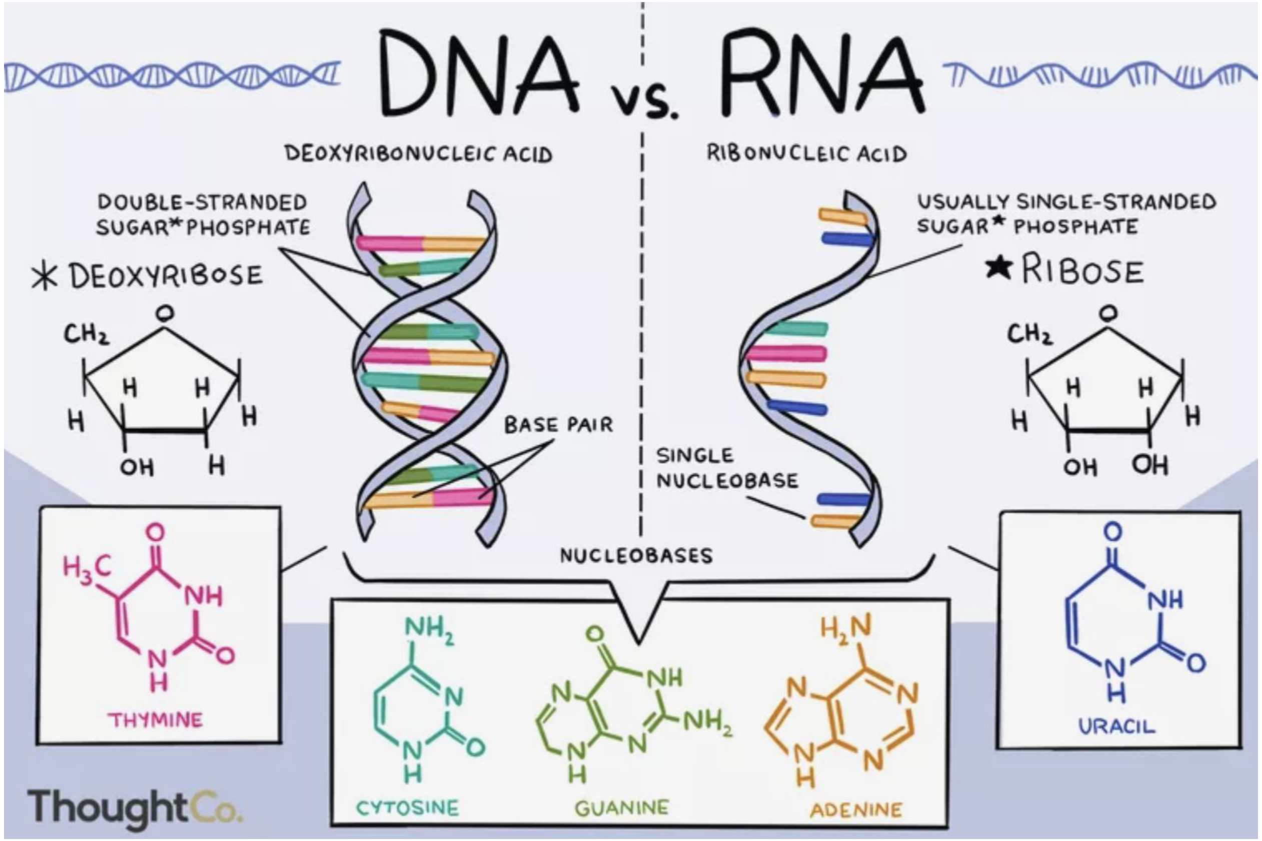
Central dogma of molecular biology - Video



Transcriptomics

Lots of RNAs, splicing, GTF, translation

RNA vs DNA - difference?



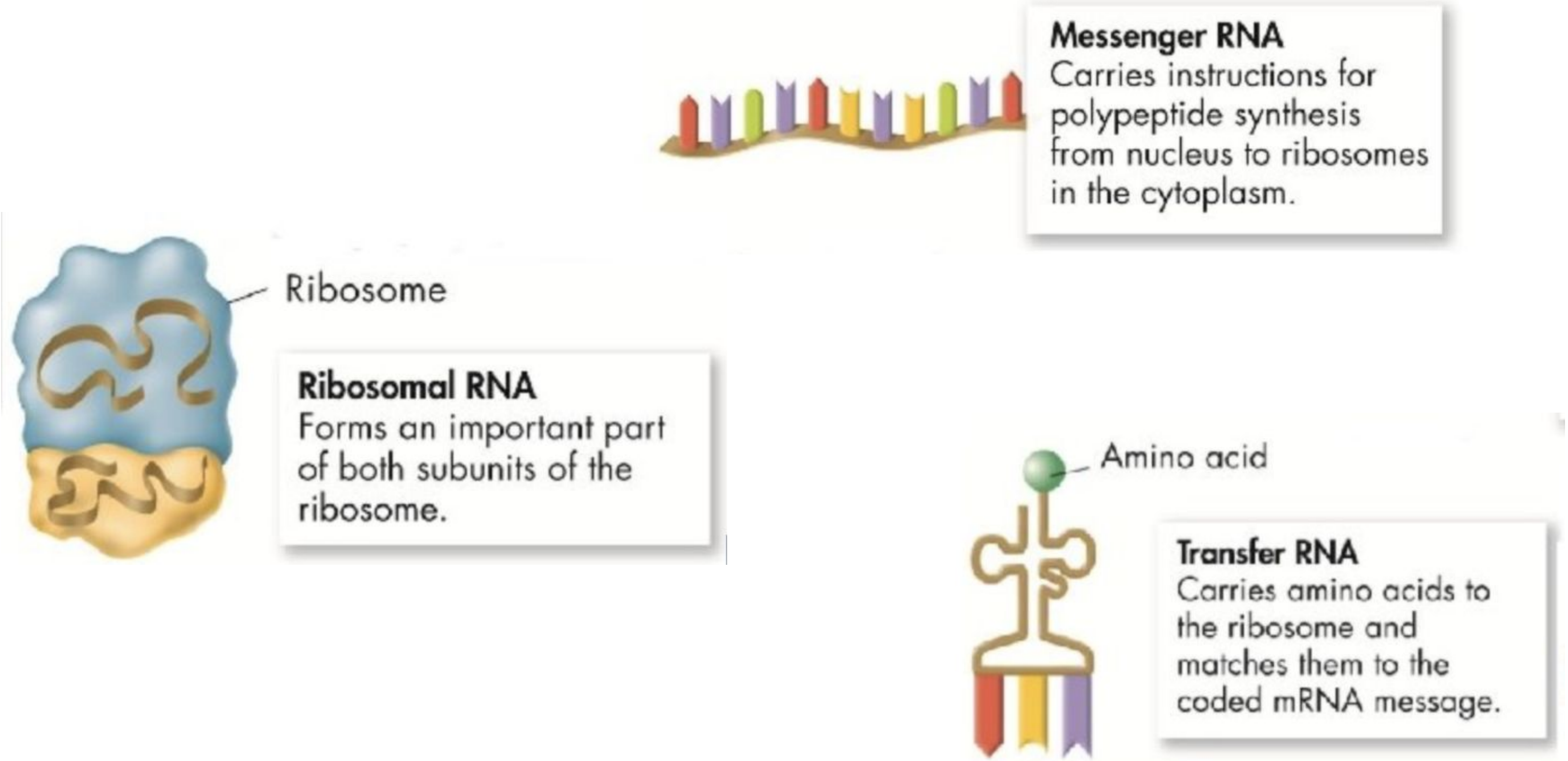
DNA:

- Deoxyribonucleic acid
- Double strand
- T (thymine)

RNA:

- Ribonucleic acid
- Single strand
- U (uracil)

Main types of RNA

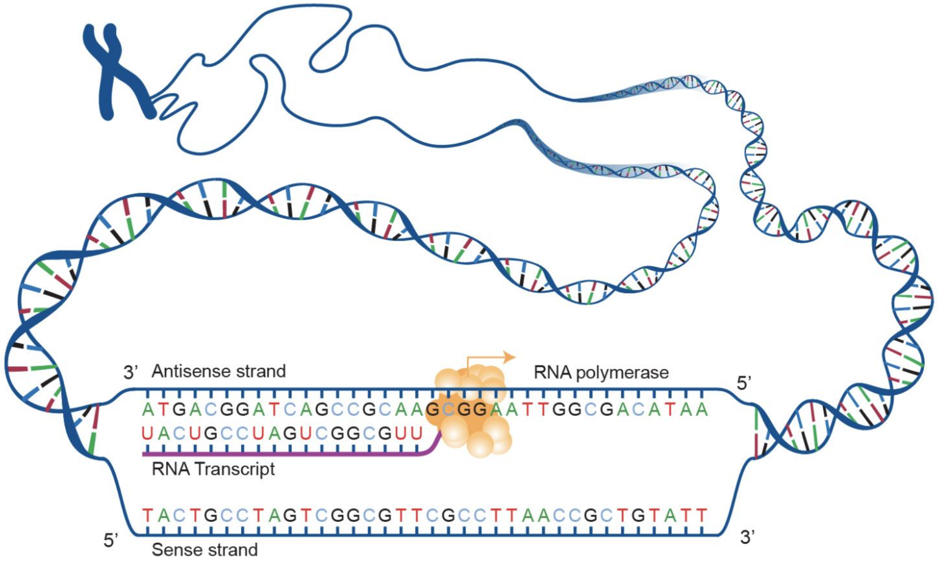


Messenger RNA
Carries instructions for polypeptide synthesis from nucleus to ribosomes in the cytoplasm.

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

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

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.

Main transcription enzyme: RNA polymerase

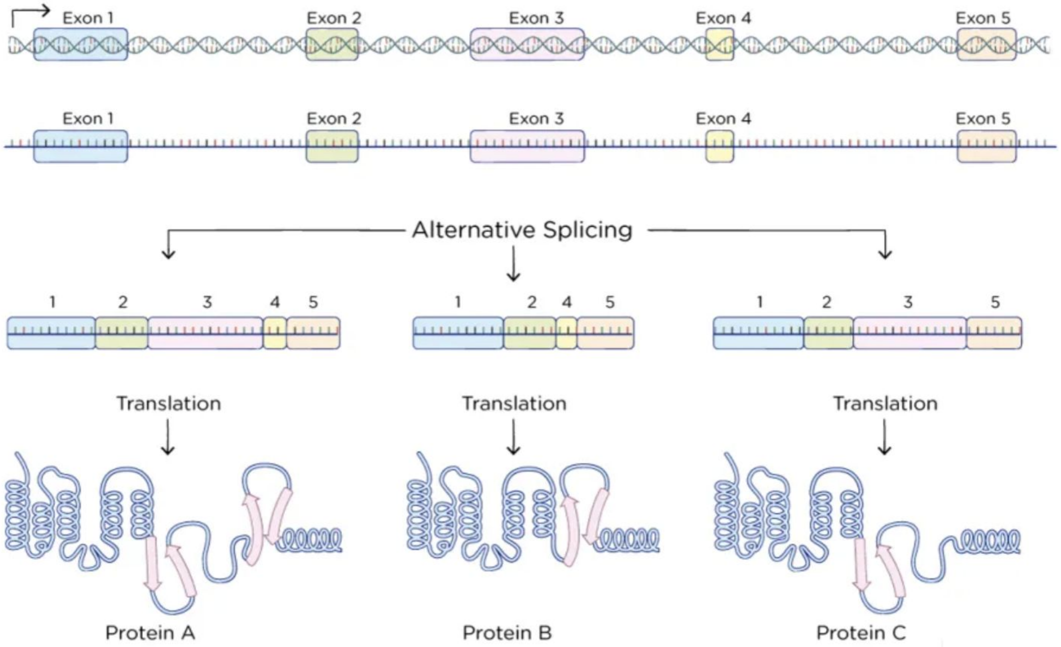
Transcription begins when RNA polymerase binds to a **promoter** sequence near the beginning of a gene (directly or through helper proteins).

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

Transcription ends in a process called **termination**.

Termination depends on sequences in the RNA, which signal that the transcript is finished.

Transcription



GENE (DNA): consists of introns and exons

pre-mRNA: Initial transcription product

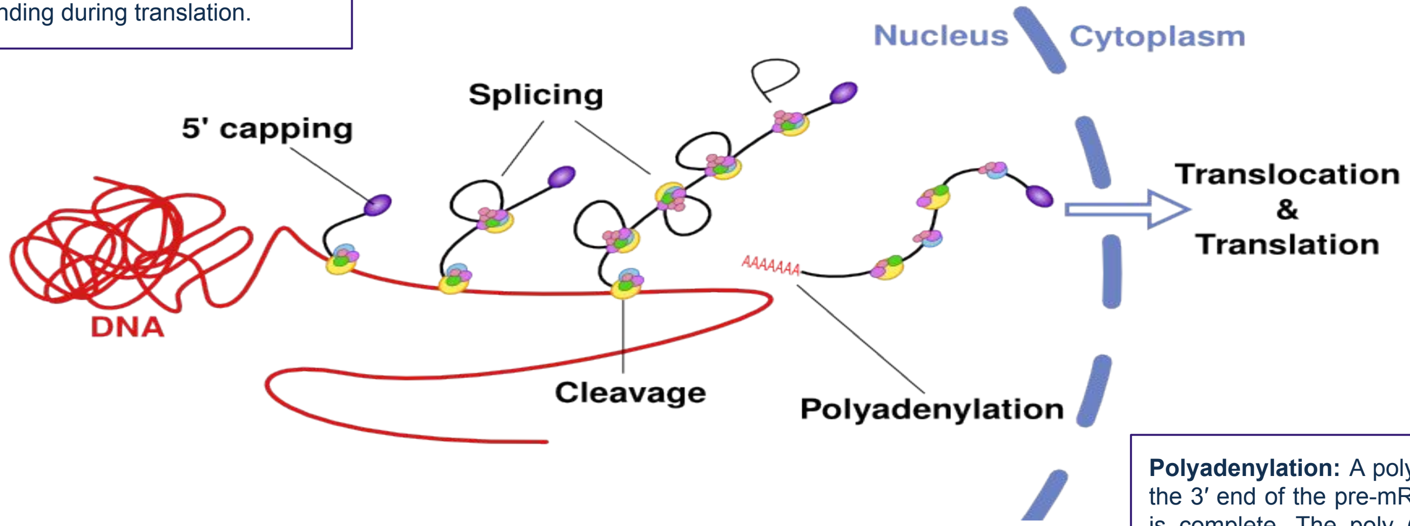
After initial transcription maturation of RNA sequence is performed in a process of alternative splicing

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.

Alternative splicing and maturation

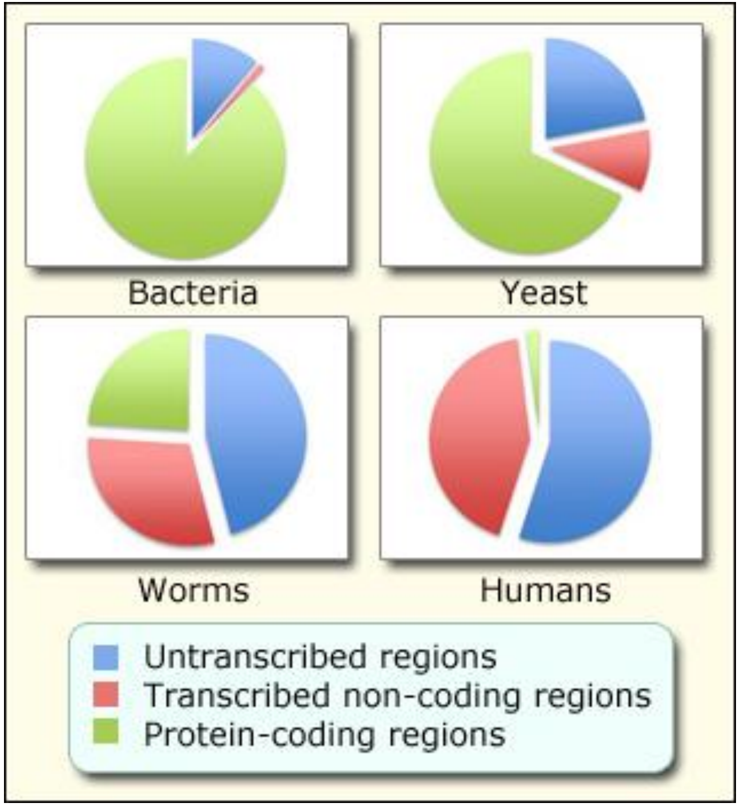
5’ capping: A cap is added to the 5′ end of the pre-mRNA while elongation is still in progress. The 5′ cap protects the nascent mRNA from degradation and assists in ribosome binding during translation.



Polyadenylation: A poly (A) tail is added to the 3′ end of the pre-mRNA once elongation is complete. The poly (A) tail protects the mRNA from degradation, aids in the export of the mature mRNA to the cytoplasm, and is involved in binding proteins involved in initiating translation.

Splicing: Introns are removed from the pre-mRNA before the mRNA is exported to the cytoplasm.

Transcription - how much of DNA is transcribed?



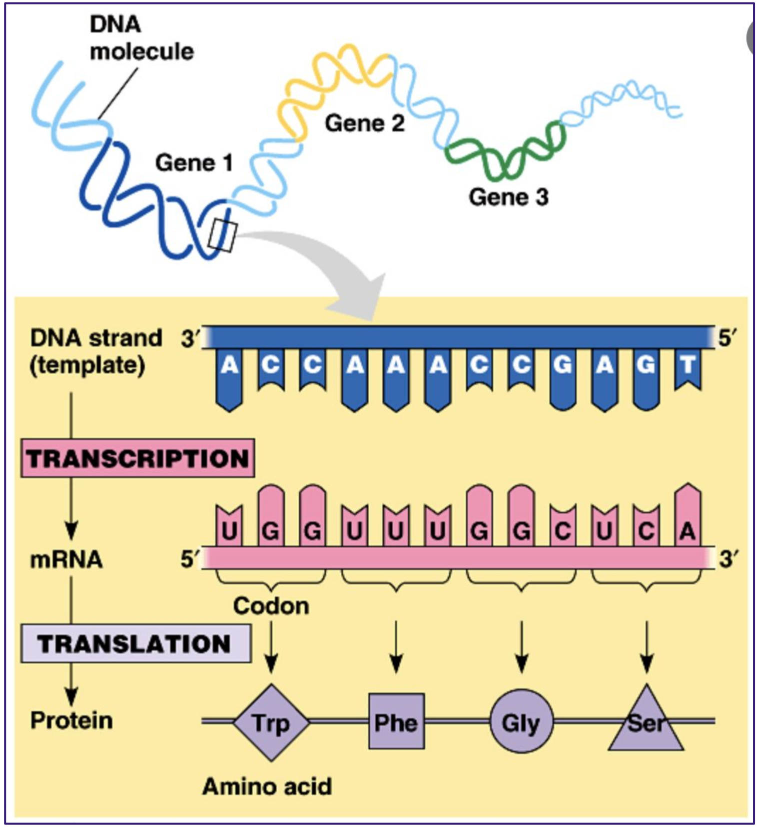
Gene - segment of DNA which is transcribed into RNA which then has a function in cell

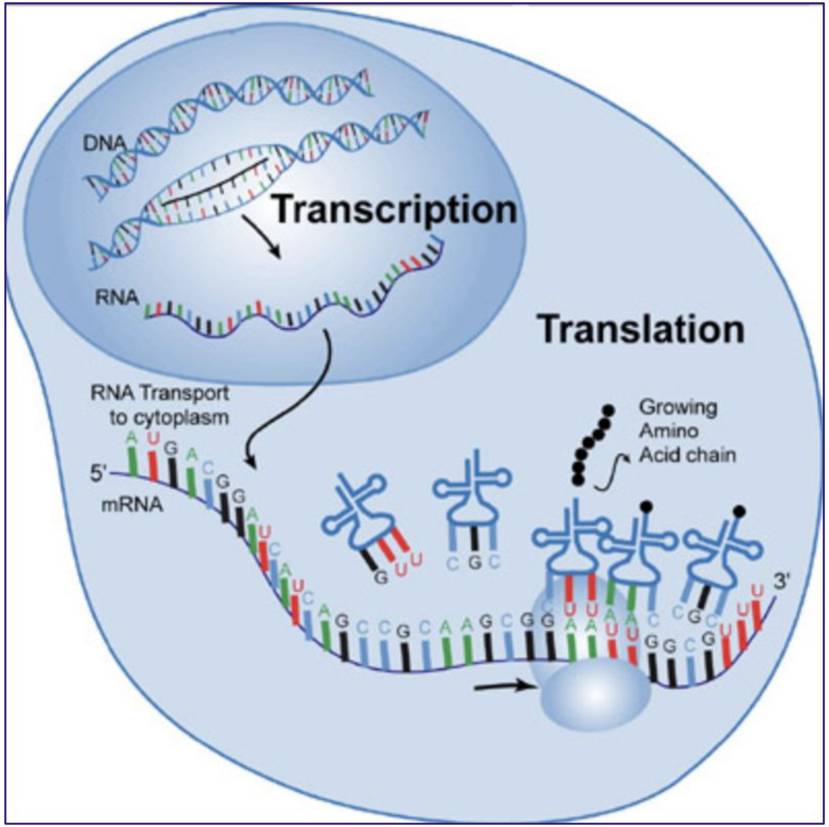
If RNA codes for protein that RNA is called **mRNA** and the region of genome from which it is transcribed is called **protein-coding gene** (green)

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

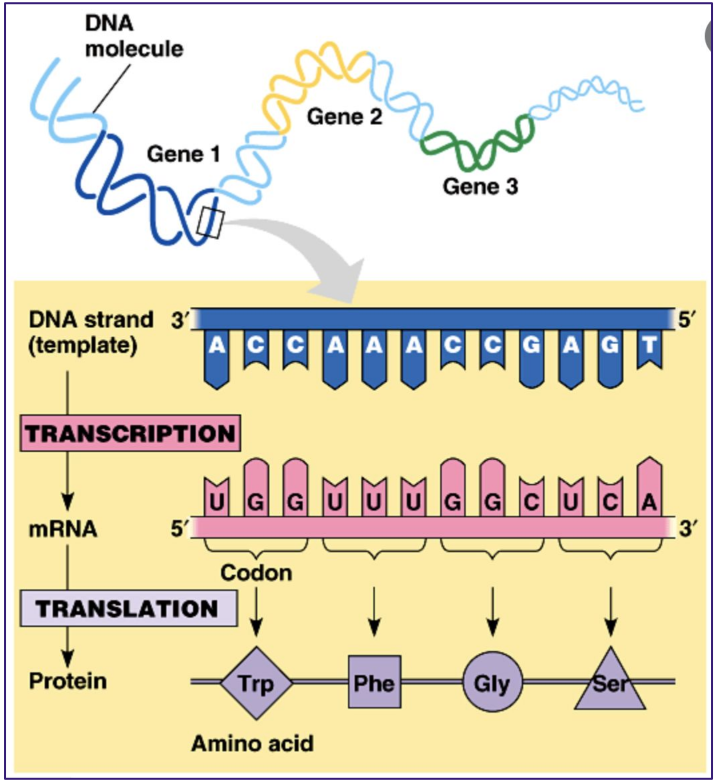
Some regions of DNA (most of it) are not transcribed at all (blue)

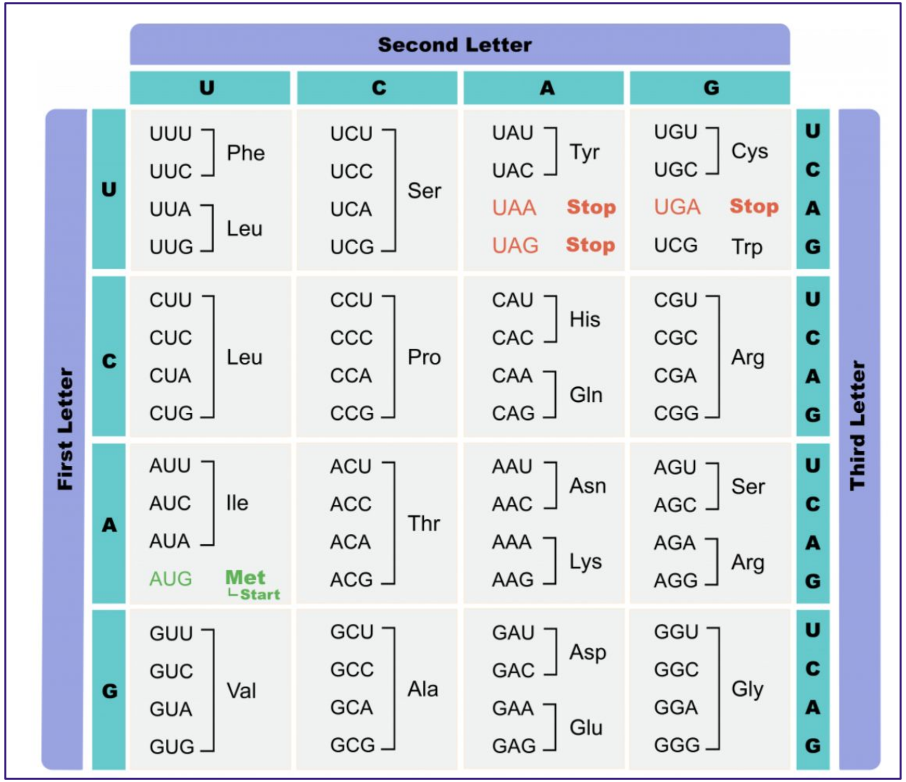
mRNAs: translation to proteins





mRNAs: translation to proteins





		Second Letter				
First Letter		U	C	A	G	Third Letter
U		UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
		UUC	UCC	UAC	UGC	C
		UUA Leu	UCA	UAA Stop	UGA Stop	A
		UUG	UCG	UAG Stop	UCG Trp	G
C		CUU Leu	CCU Pro	CAU His	CGU Arg	U
		CUC	CCC	CAC	CGC	C
		CUA	CCA	CAA Gln	CGA	A
		CUG	CCG	CAG	CGG	G
A		AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
		AUC	ACC	AAC	AGC	C
		AUA	ACA	AAA Lys	AGA Arg	A
		AUG Met Start	ACG	AAG	AGG	G
G		GUU Val	GCU Ala	GAU Asp	GGU Gly	U
		GUC	GCC	GAC	GGC	C
		GUA	GCA	GAA Glu	GGA	A
		GUG	GCG	GAG	GGG	G

RNA-seq

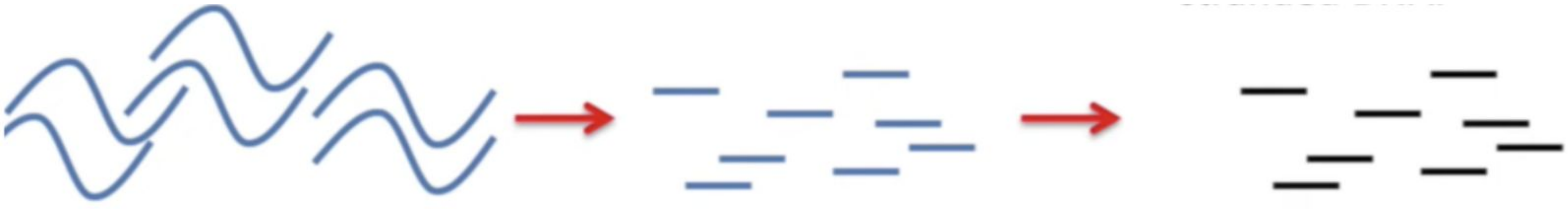
Library preparation

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

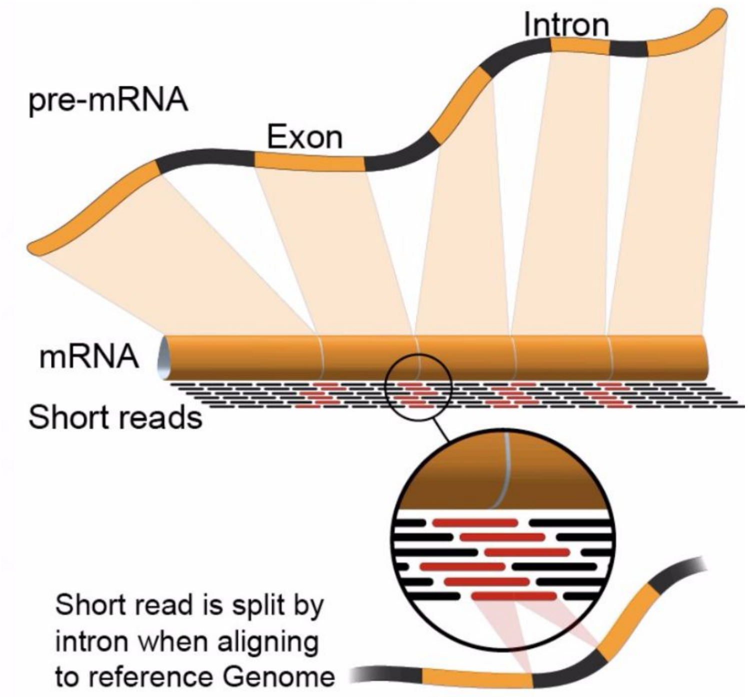
Double stranded DNA is more stable than RNA and can be easily amplified and modified. This leads us to the next step…

RNA-seq

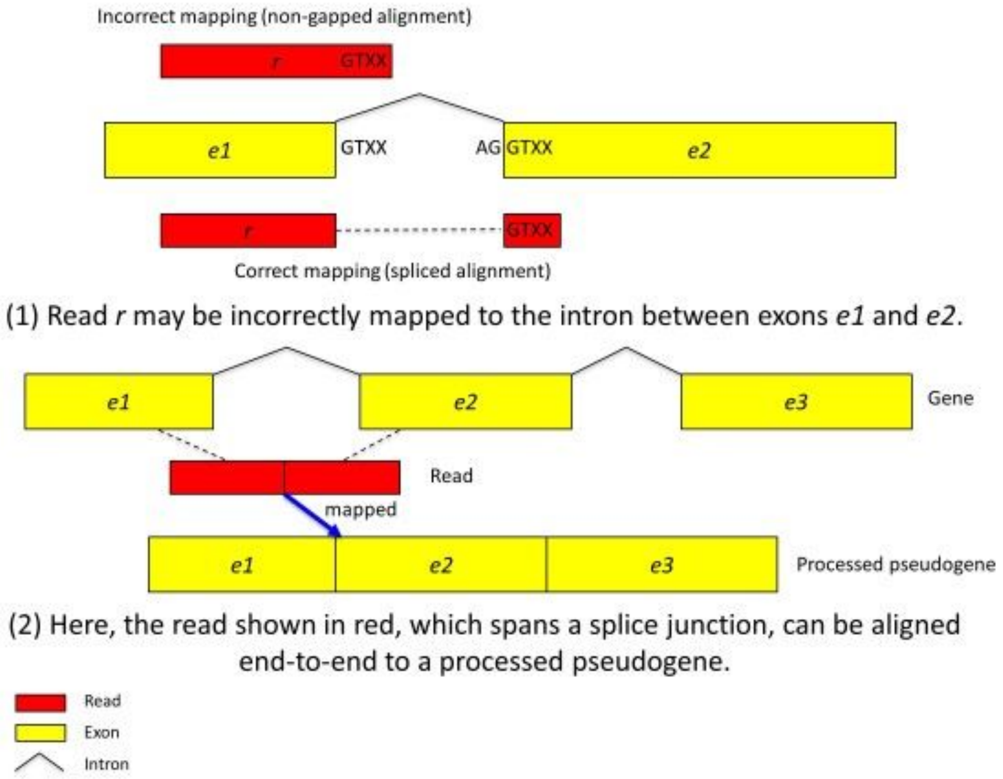
Splice-aware alignment

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


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

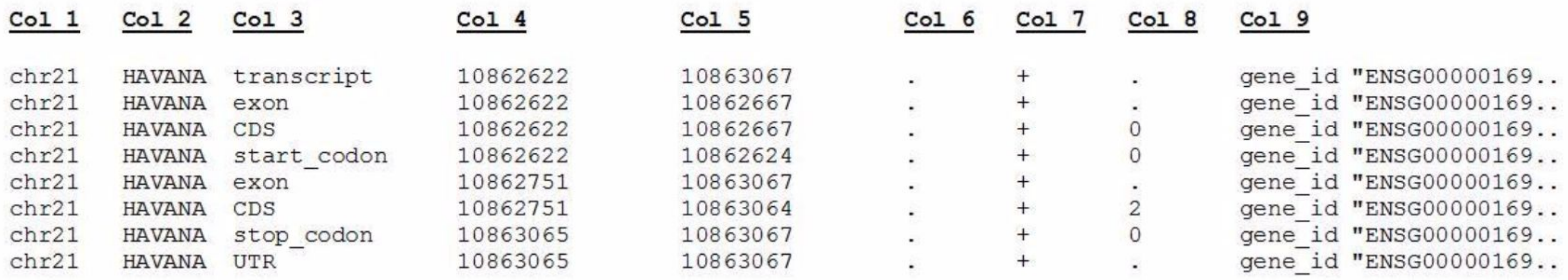
Splice-aware alignment



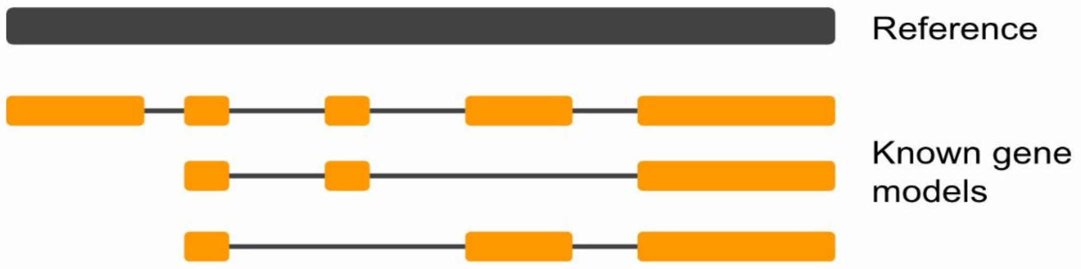
(1) Read *r* may be incorrectly mapped to the intron between exons *e1* and *e2*.

(2) Here, the read shown in red, which spans a splice junction, can be aligned end-to-end to a processed pseudogene.

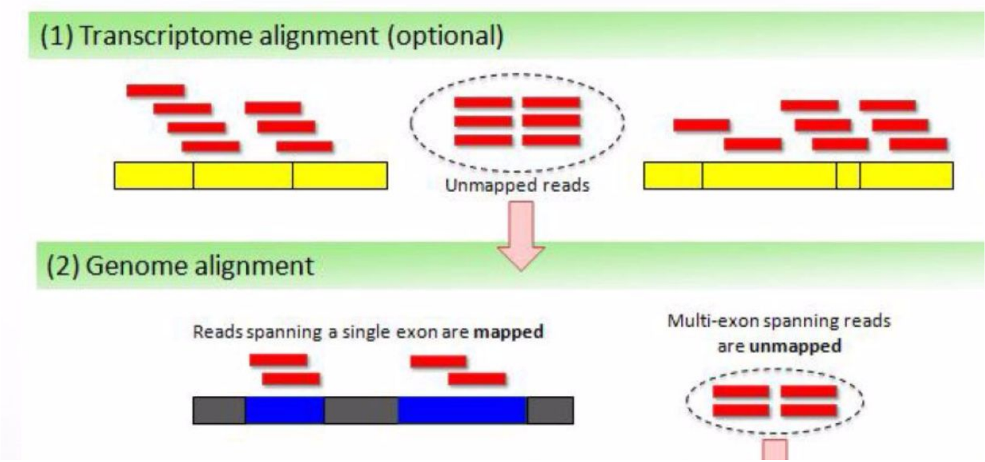
GTF (gene transfer format)



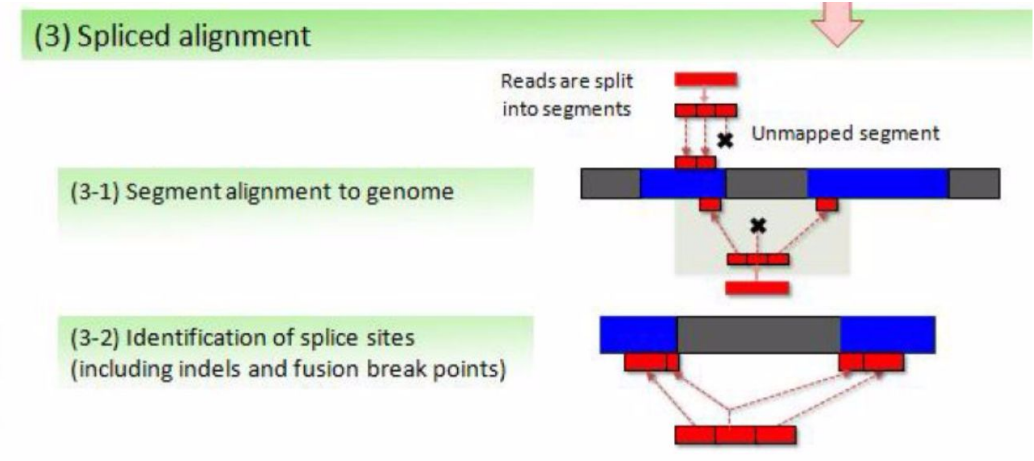
Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG0000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG0000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG0000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG0000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG0000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG0000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG0000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG0000169..



Splice-aware alignment


(1) Transcriptome alignment (optional)
Unmapped reads
(2) Genome alignment
Reads spanning a single exon are **mapped**
Multi-exon spanning reads are **unmapped**

Splice-aware alignment


(3) Spliced alignment
Reads are split into segments
Unmapped segment
(3-1) Segment alignment to genome
(3-2) Identification of splice sites (including indels and fusion break points)

Why do RNA-Seq?

RNA-seq analysis

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

RNA-seq analysis

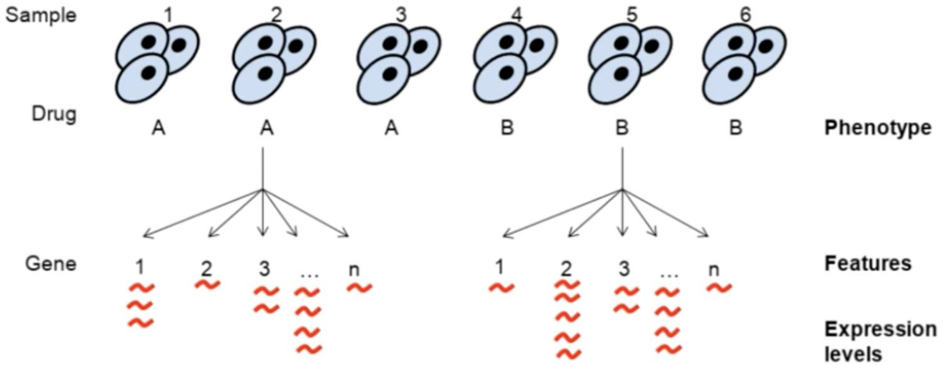
- OFTEN: **relative abundance (quantification)** of RNAs and testing for **differential expression**

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 between different cell types significantly differ



Motivation for RNA quantification

- We (usually) want to check if there is **change in transcription (expression)** between conditions (healthy/sick, treated/untreated, different tissues, etc..)

Transcriptomics

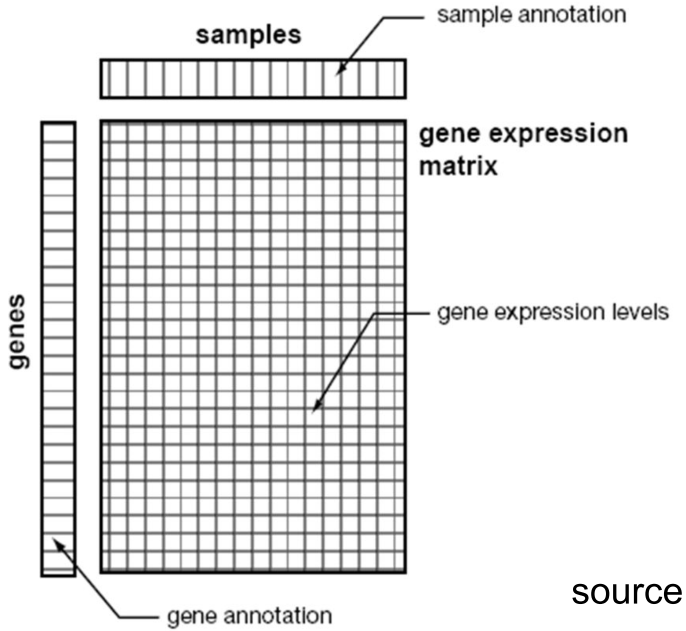
Quantification

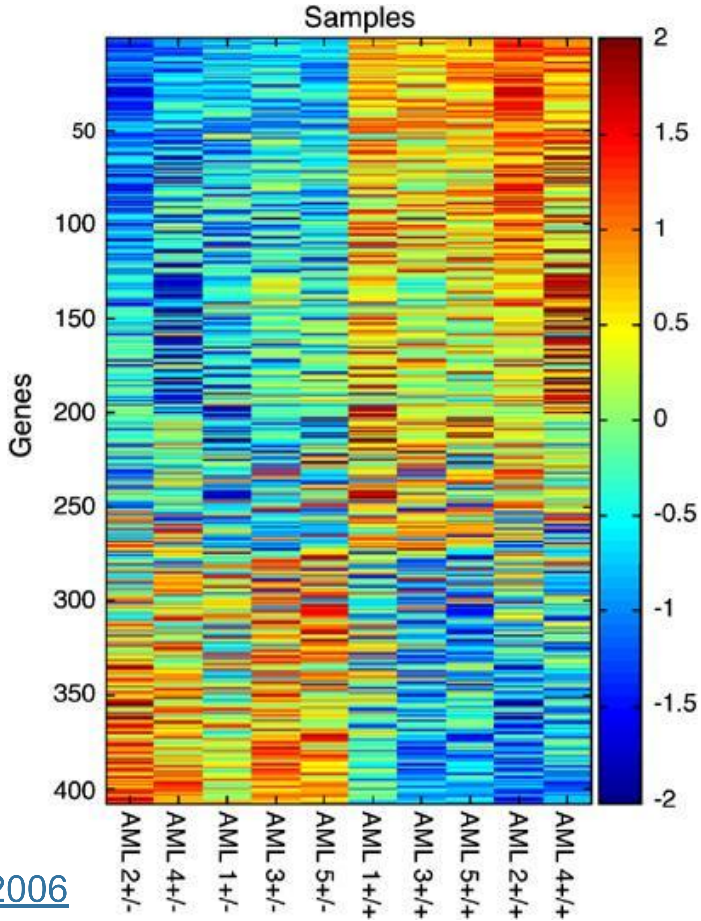
We will talk about:

- RNA quantification
- Differential expression

RNA quantification result

- Expression profiles





source: Nature Leukemia 2006

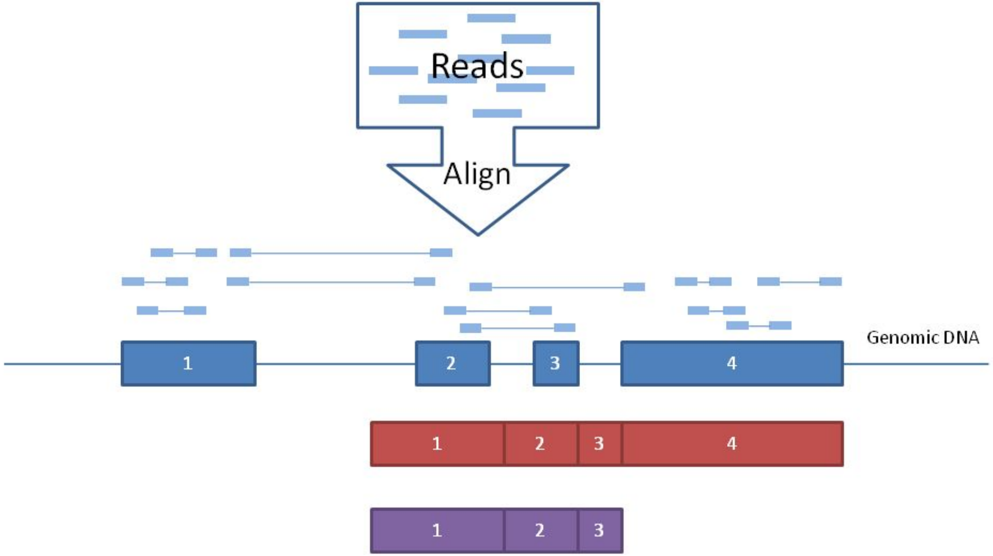
Quantification - problems

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

(1) RNA-seq: abundance estimation

Problem statement:

How to resolve alignment ambiguity?



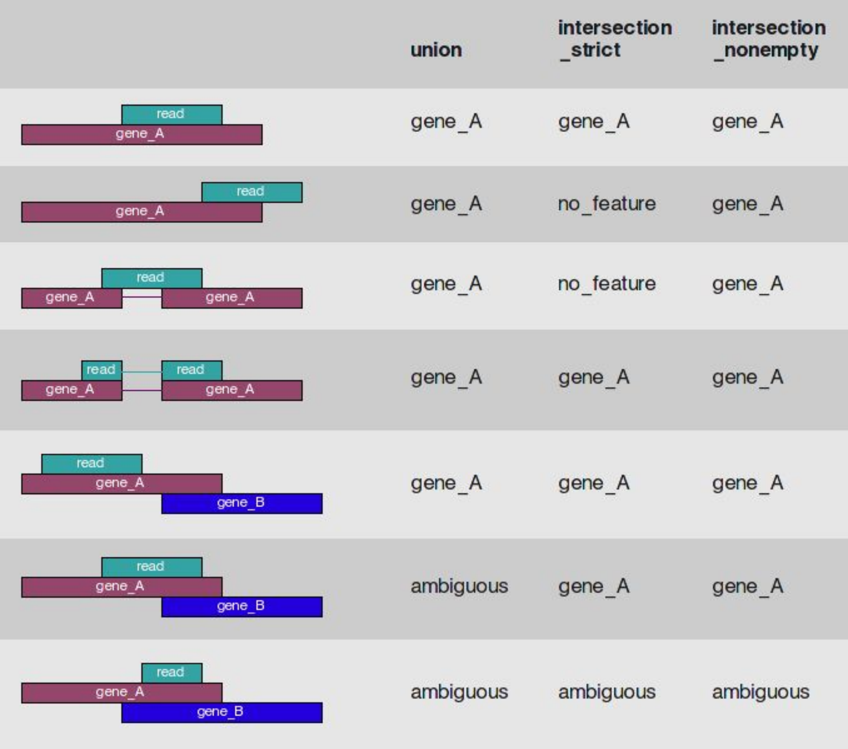
Source: http://dx.doi.org/10.13070/mm.en.3.203

(1) RNA-seq: abundance estimation

Raw counting

vs.

probabilistic estimation



	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	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

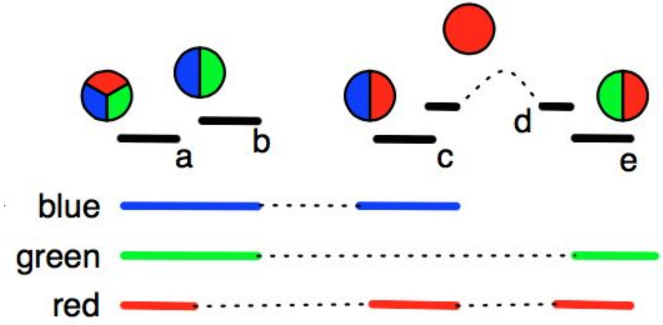
$\sum_k \rho_k = 1$, multinomial distribution

$P_i = \sum_k y_{i,k} \cdot \rho_k$, 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

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

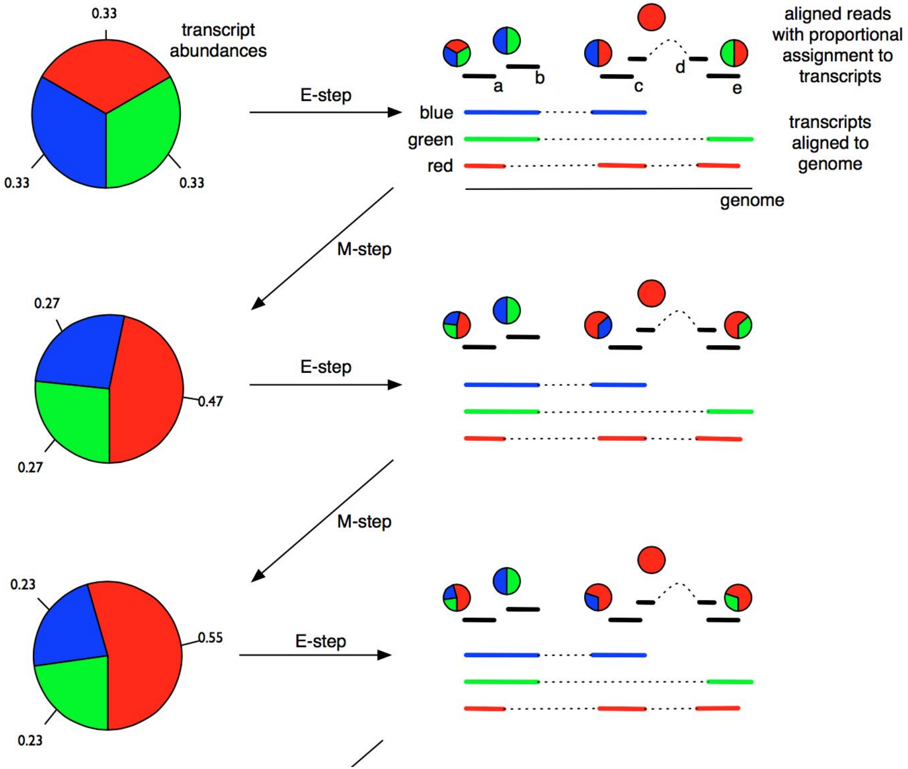
E2 step: prior $= (0.27, 0.27, 0.46)$

$$p_a = (0.27, 0.27, 0.46),\ p_b = (1/2, 1/2, 0),$$
$$p_c = (\frac{0.27}{0.46 + 0.27}, 0, \frac{0.46}{0.46 + 0.27}),\ p_d = (0, 0, 1), \ldots$$

M2 step:

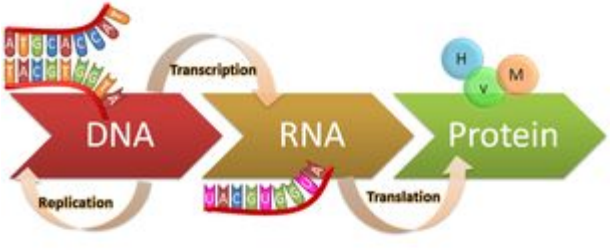
$$\rho_{blue} = (0.27 + 1/2 + 0.37 + 0 + 0)/5 = 0.23$$

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

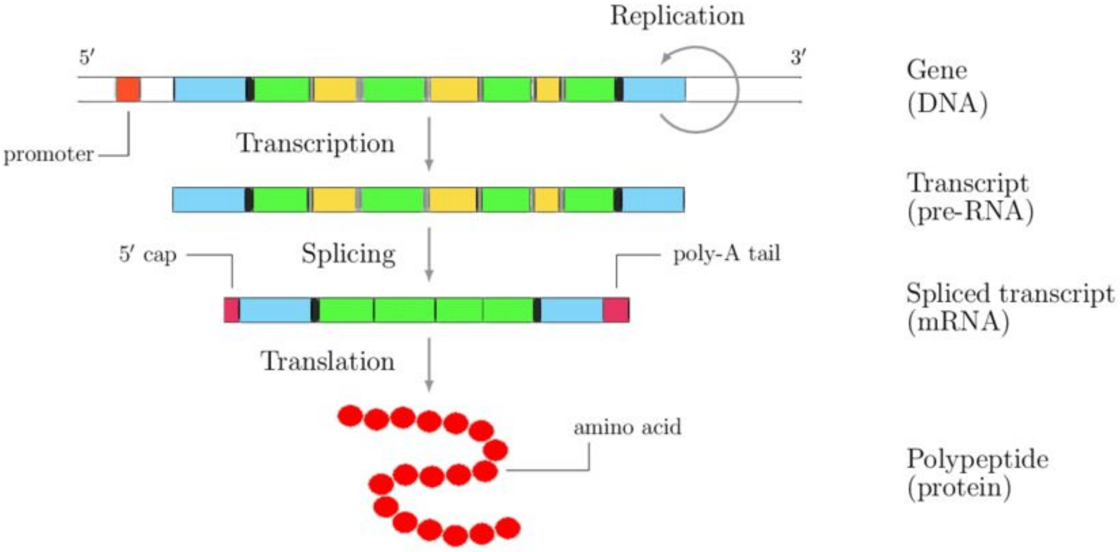


Recap

Central dogma of molecular biology



https://kaiserscience.wordpress.com/



Nadalin, Francesca. "Paired is better: local assembly algorithms for NGS paired reads and applications to RNA-Seq"

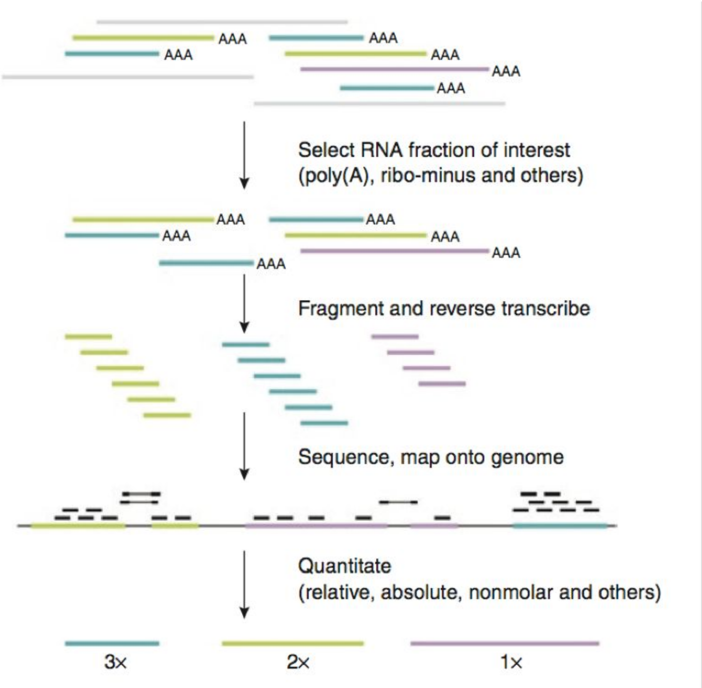
Transcriptome

A *transcriptome* the full range of messenger RNA, or mRNA, molecules expressed by an organism. The term "transcriptome" can also be used to describe the array of mRNA transcripts produced in a particular cell or tissue type.

The RNA-Seq pipeline

Quantitative/qualitative analysis goals:

- Transcriptome reconstruction
- Estimation of gene/transcript expression levels


AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)
3×
2×
1×

Pepke S., Wold B., Mortazavi A., Computation for ChIP-seq and RNA-seq studies. Nature Methods 6:11, Nov 2009

Normalization

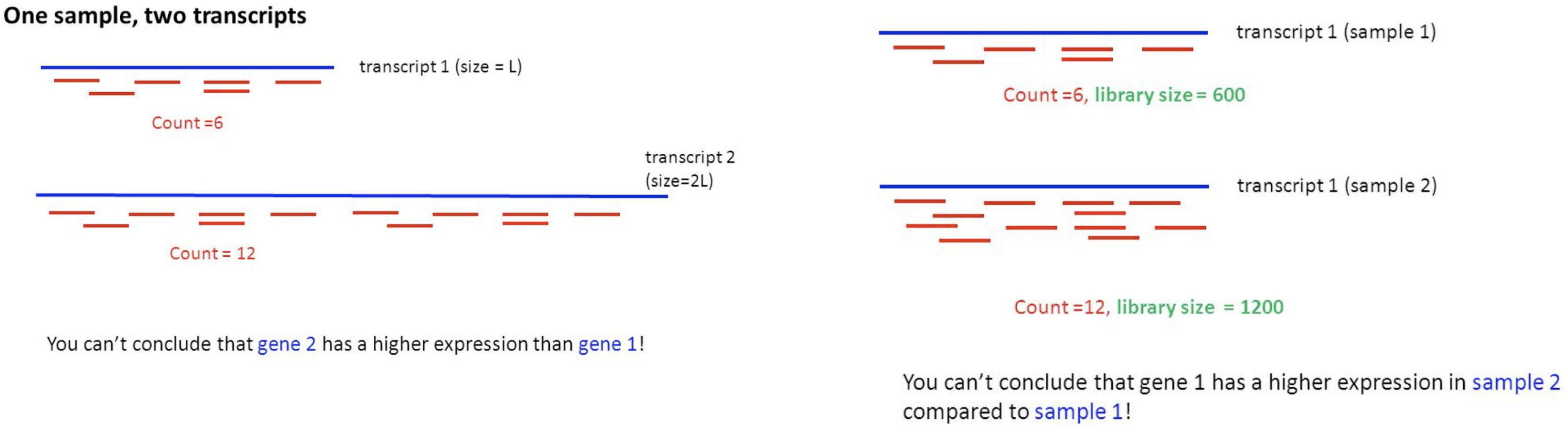
RNA-seq: data normalization

Problem statement:

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

if we observe reads from sampled transcripts?

RNA-seq: data normalization



One sample, two transcripts

transcript 1 (size = L)

Count =6

transcript 2 (size=2L)

Count = 12

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

transcript 1 (sample 1)

Count =6, library size = 600

transcript 1 (sample 2)

Count =12, library size = 1200

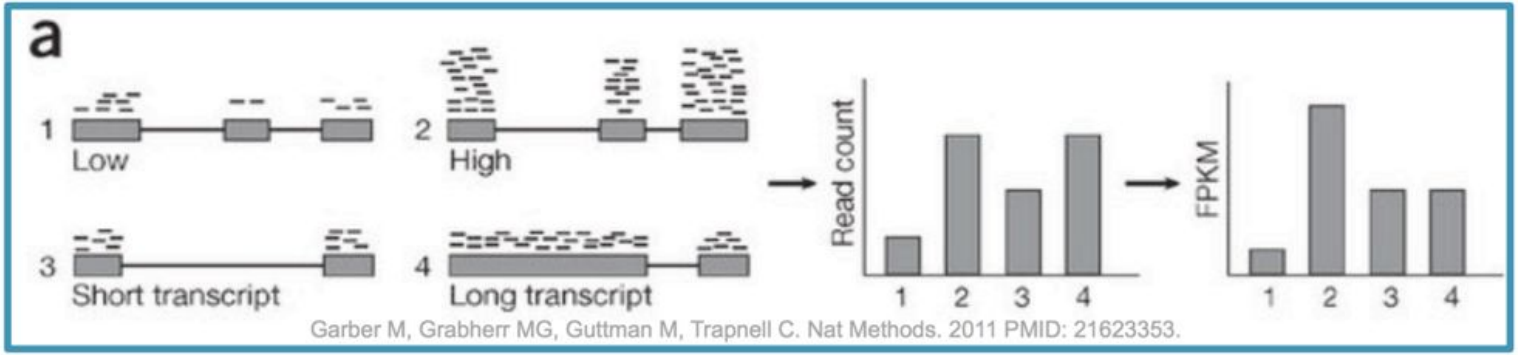
You can’t conclude that gene 1 has a higher expression in sample 2 compared to sample 1!

- We need to account for gene length and library size

RNA-seq: data normalization

Let X_i be number of reads aligned to ith transcript

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



Garber M, Grabherr MG, Guttman M, Trapnell C. Nat Methods. 2011 PMID: 21623353.

(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)

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

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

X_i - number of reads aligned to transcript ‘i’

N - total number of reads

l_i - read length

Example

```
##              Rep1 Counts     Rep2 Counts     Rep3 Counts
## Gene Name
## A(5kb)       100000          120000          300000
## B(10kb)      200000          250000          600000
## C(1kb)       100000          80000           150000
## D(20kb)      0               0               10000
```

FPKM

We first divide by the total number of reads...

```
##              Rep1 Counts     Rep2 Counts     Rep3 Counts
## Gene Name
## A(5kb)       250000.0        266666.67       283018.87
## B(10kb)      500000.0        555555.56       566037.74
## C(1kb)       250000.0        177777.78       141509.43
## D(20kb)      0.0             0.00            9433.96
```

FPKM

…and then by gene length.

```
##              Rep1 Counts     Rep2 Counts     Rep3 Counts
## Gene Name
## A(5kb)       50000.0         53333.33        56603.77
## B(10kb)      50000.0         55555.56        56603.77
## C(1kb)       250000.0        177777.78       141509.43
## D(20kb)      0.0             0.00            471.70
```

TPM

We first divide by the gene length...

```
##              Rep1 Counts     Rep2 Counts     Rep3 Counts
## Gene Name
## A(5kb)       50000.0         60000.0         150000.0
## B(10kb)      50000.0         62500.0         150000.0
## C(1kb)       100000.0        80000.0         150000.0
## D(20kb)      0.0             0.0             1000.0
```

TPM

…and then we perform the library size normalization, using abundances already normalized for gene length:

```
##              Rep1 Counts     Rep2 Counts     Rep3 Counts
## Gene Name
## A(5kb)       250000.0        296296.296      332594.235
## B(10kb)      250000.0        308641.975      332594.235
## C(1kb)       500000.0        395061.728      332594.235
## D(20kb)      0.0             0.000           2217.295
```

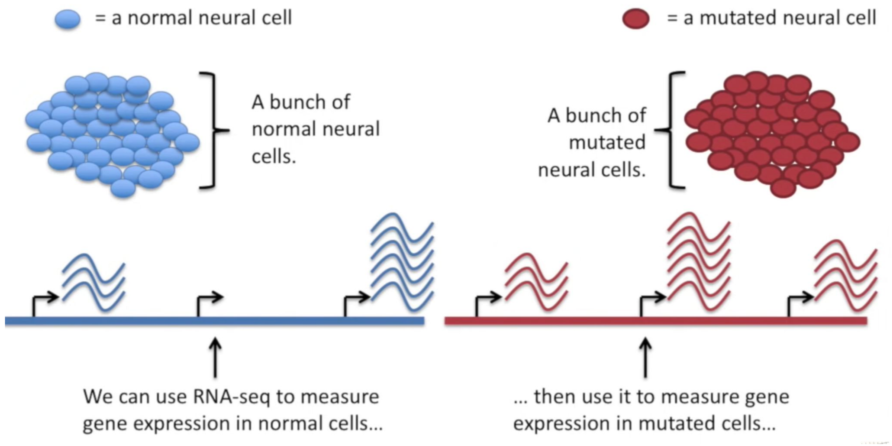
Transcriptomics

Differential expression

Differential expression:

Problem statement:

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



● = a normal neural cell

● = a mutated neural cell

A bunch of normal neural cells.

A bunch of mutated neural cells.

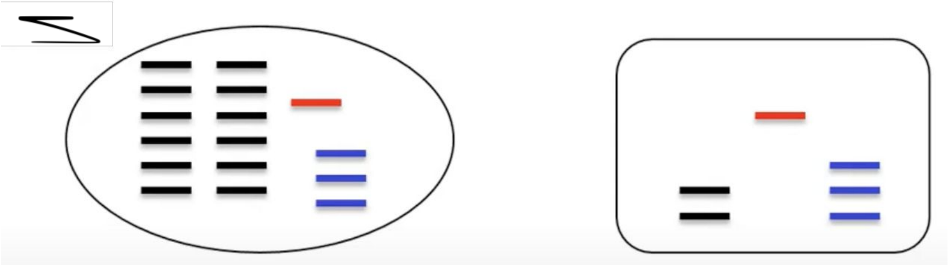
We can use RNA-seq to measure gene expression in normal cells...

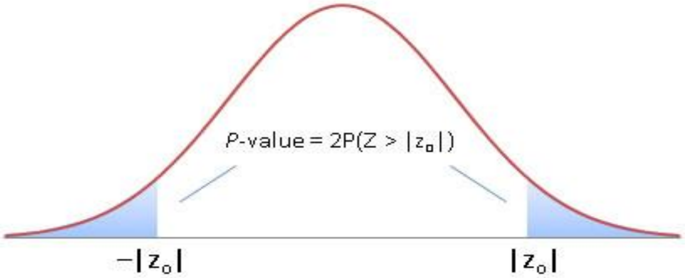
... then use it to measure gene expression in mutated cells...

(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.





(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

(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