



SevenBridges

Intro to RNA-seq

Dajana Panovic

dajana.panovic@sbgenomics.com

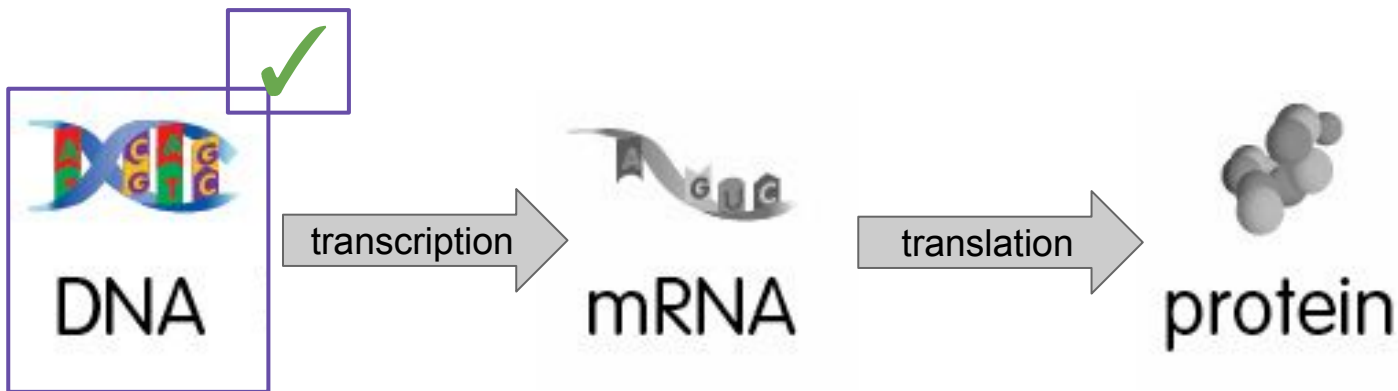
April 2022

Beyond DNA

Recap



Central dogma of molecular biology



Topics covered so far:

- Sequencing technologies
- DNA assembly
- DNA alignment
- DNA variants and variant calling

Central dogma of molecular biology



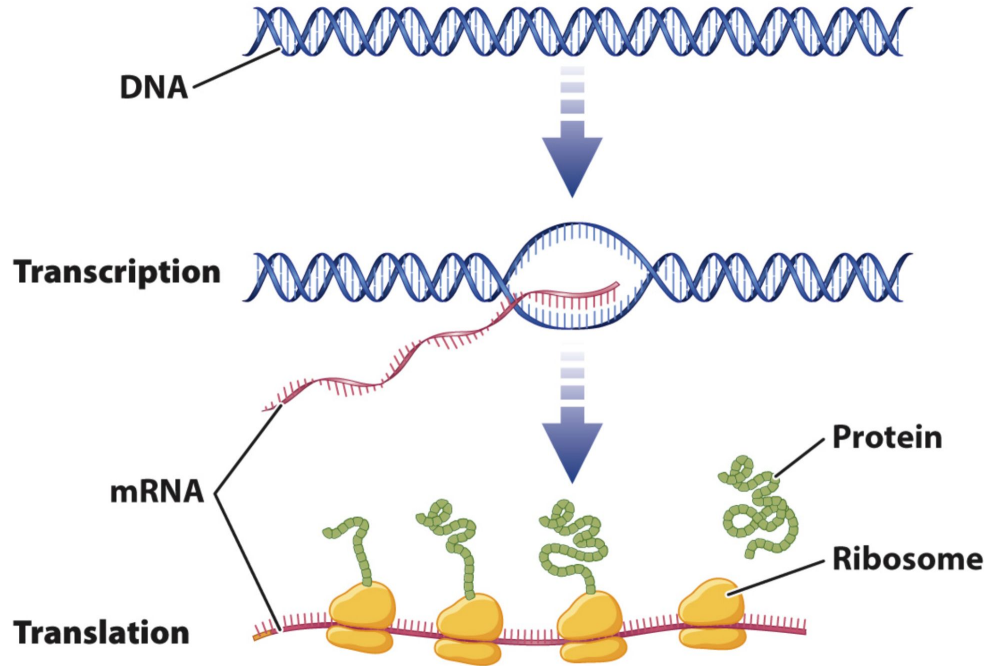
Topics covered so far:

- Sequencing technologies
- DNA assembly
- DNA alignment
- DNA variants and variant calling

Following topics:

- RNAs: transcription and translation,
- types of RNA
- mRNA: splicing, transcripts/isoforms
- RNA seq and alignment
- RNA analysis

Central dogma of molecular biology



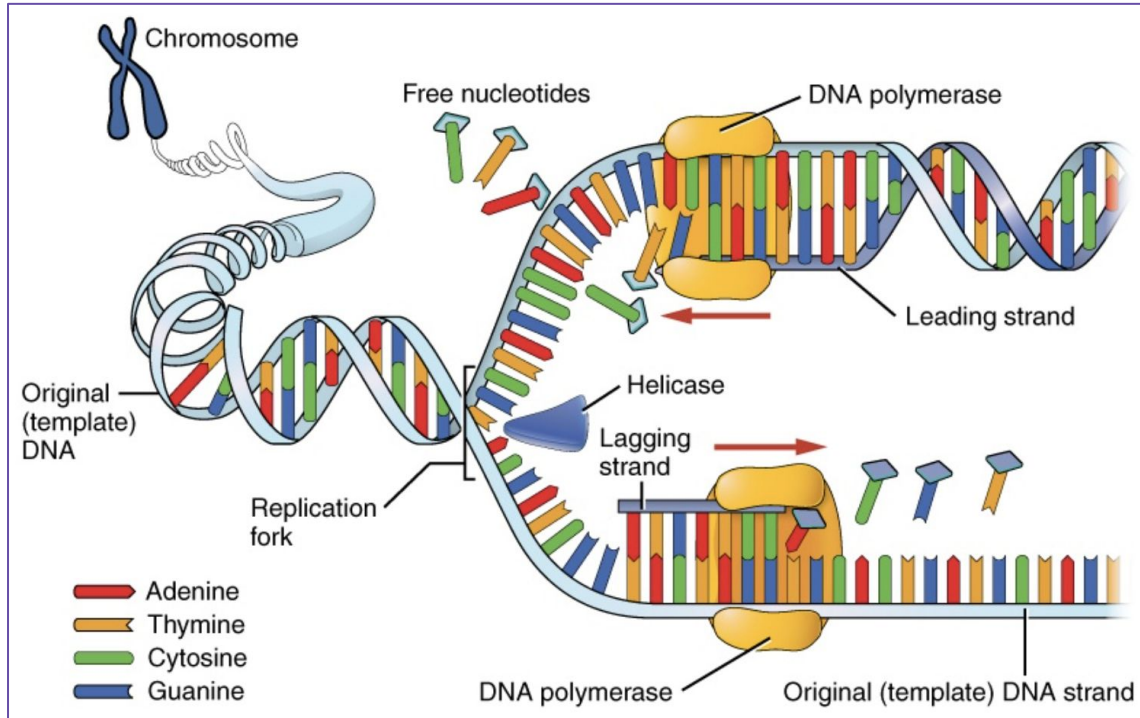
Replication - Before cell division DNA is replicated in process called replication

Transcription - synthesis of RNA molecule from small part of DNA

Translation - synthesis of protein out of RNA molecule

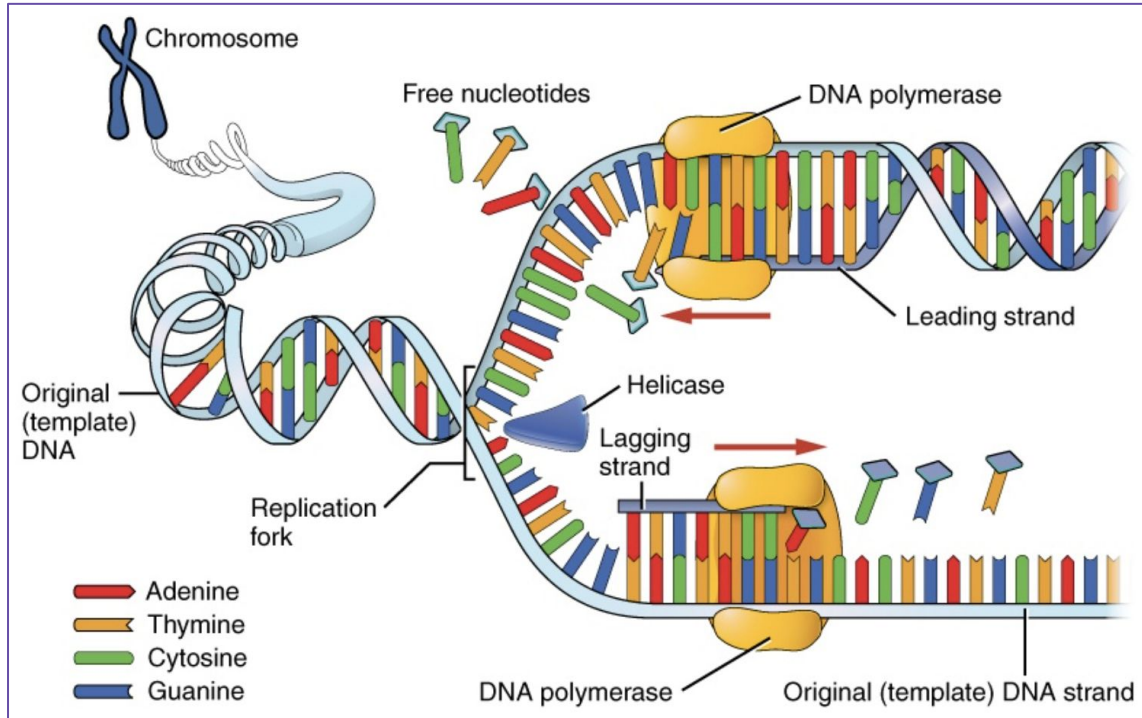
DNA replication

All cells in a body have the same DNA.



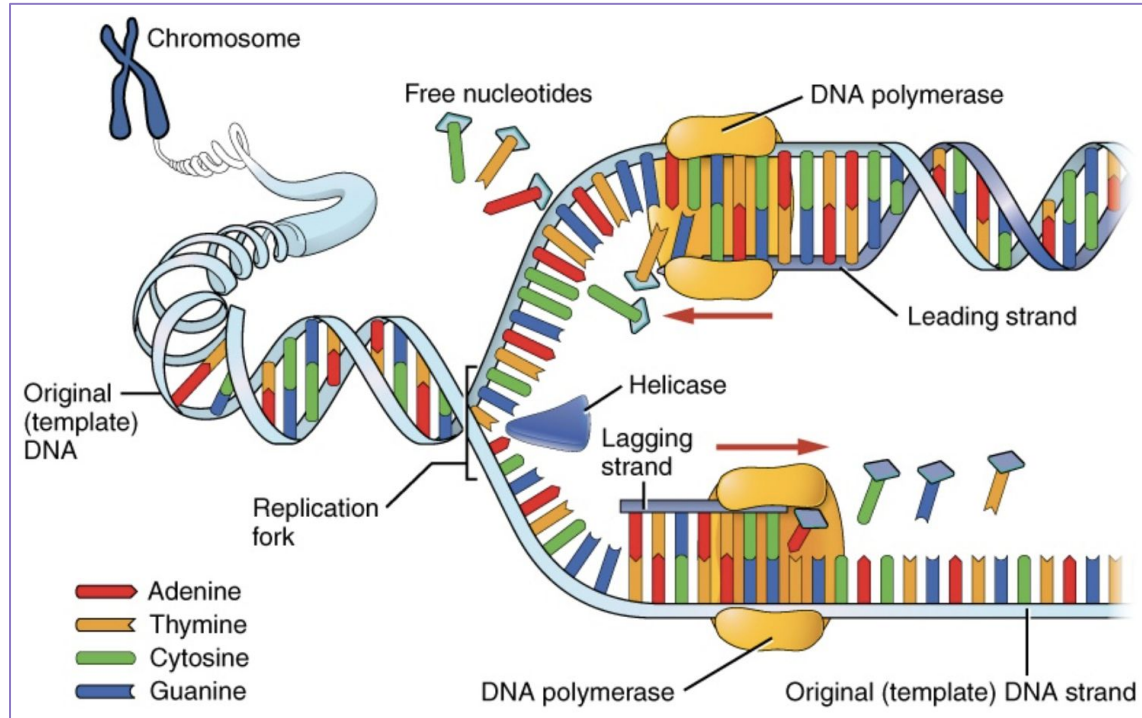
DNA replication

All cells in a body have the same DNA. **What differs is RNA.**



DNA replication

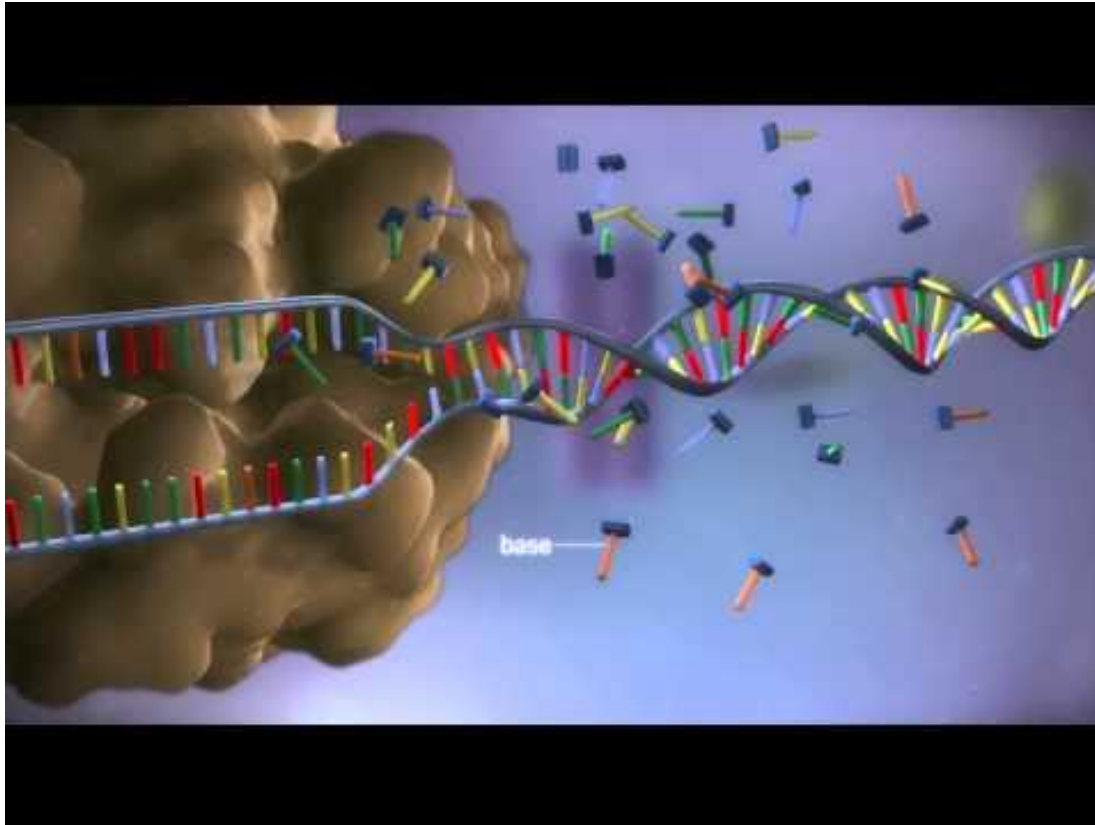
All cells in a body have the same DNA. **What differs is RNA.**



When cell is about to divide the DNA material inside the cell is replicated. That results in two cells with the very same genomic material after division.

Processes inside the different cells are regulated by various biological mechanisms and depending on the cell type and its purpose RNA material will differ.

Central dogma of molecular biology - Video

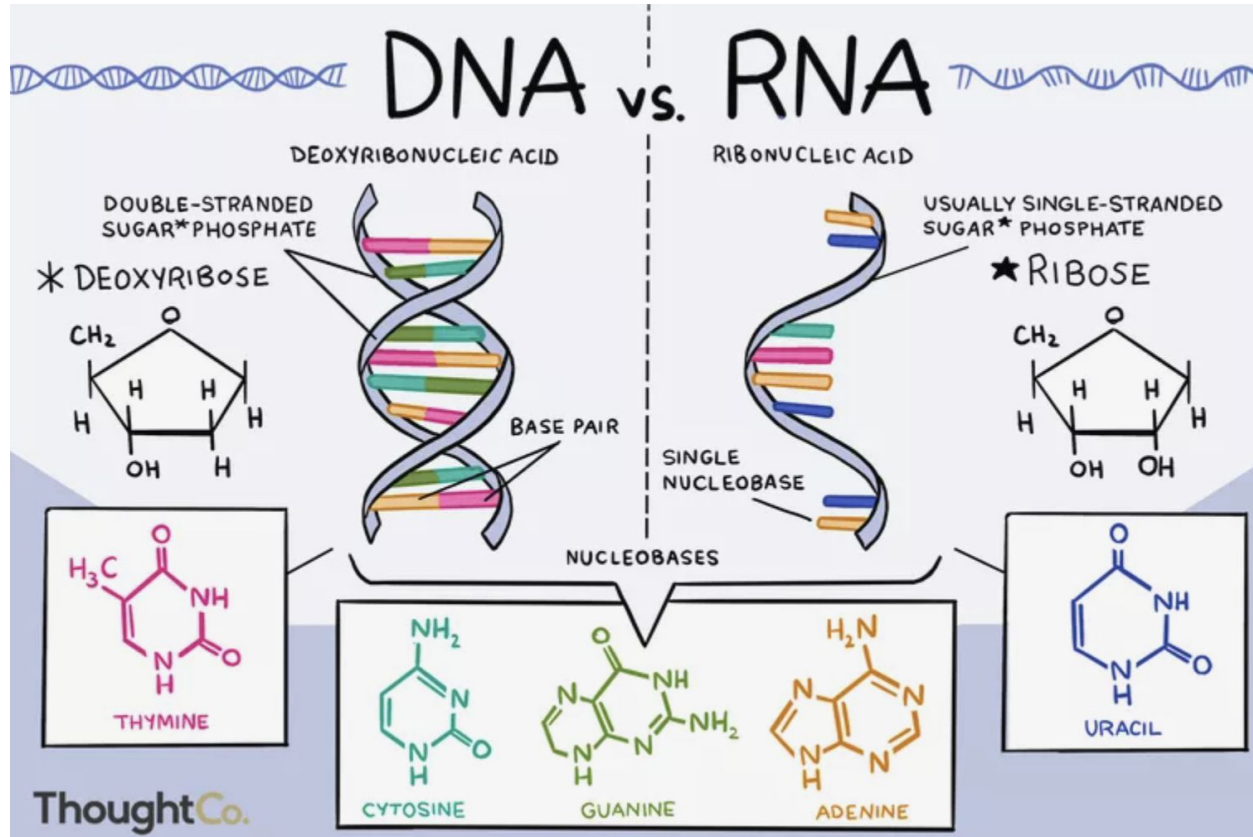


Transcriptomics

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.



Ribosome

Ribosomal RNA

Forms an important part of both subunits of the ribosome.

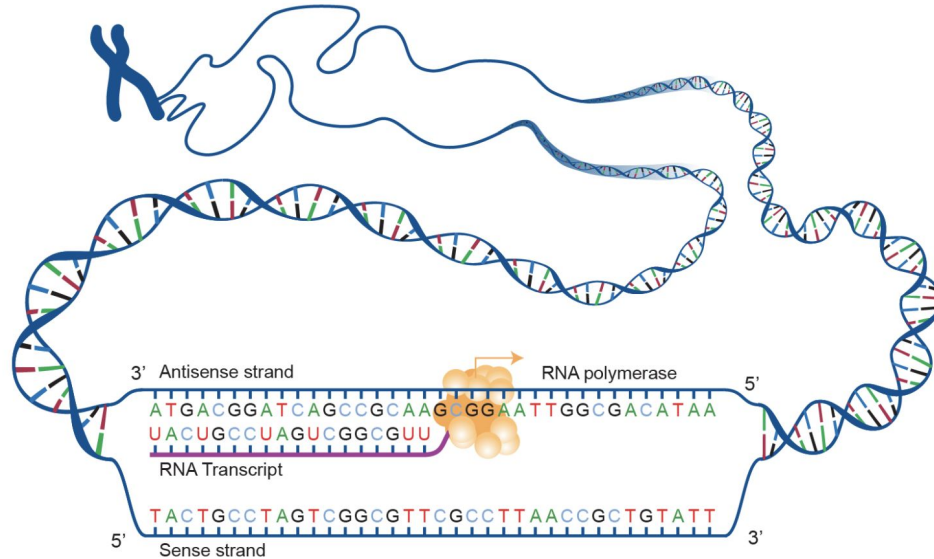


Amino acid

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

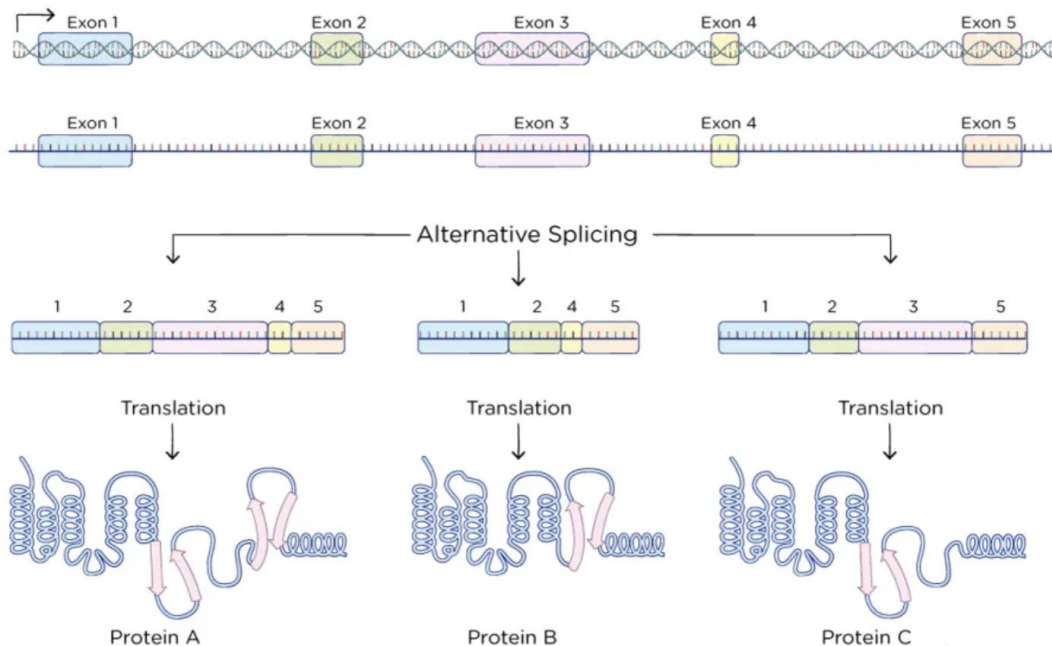
Transcription begins when RNA polymerase binds to a **promoter** sequence near the beginning of a gene (directly or through helper proteins).

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

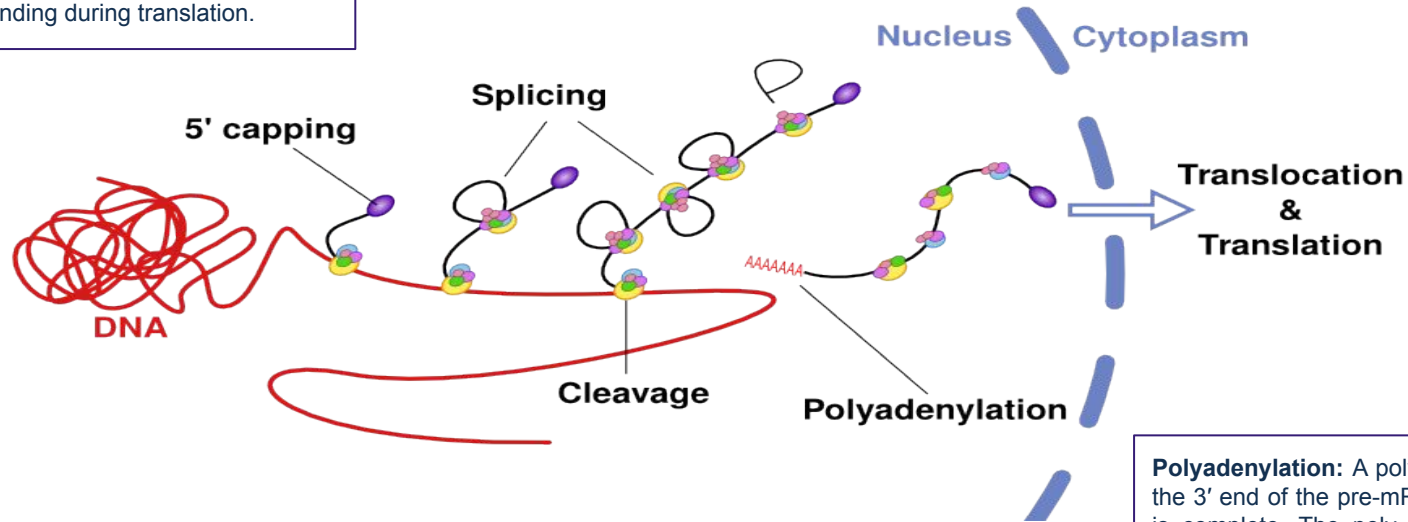
After initial transcription maturation of RNA sequence is performed in a process of **alternative splicing**

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

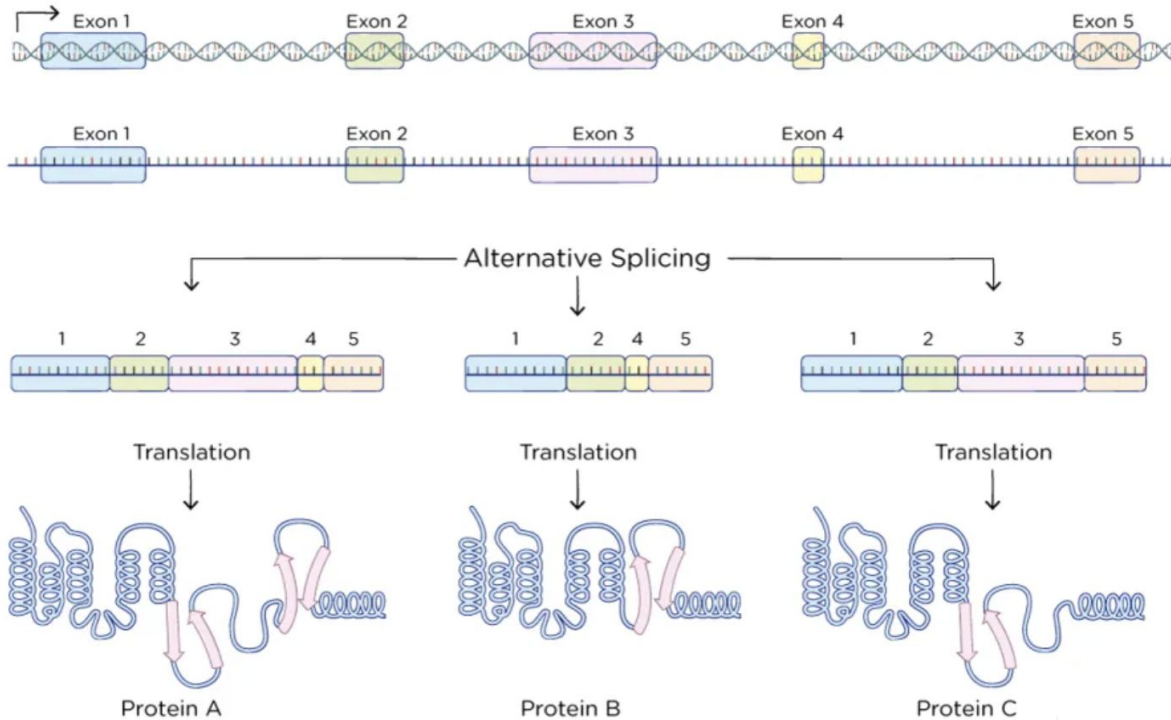
5' capping: A cap is added to the 5' end of the pre-mRNA while elongation is still in progress. The 5' cap protects the nascent mRNA from degradation and assists in ribosome binding during translation.



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

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.

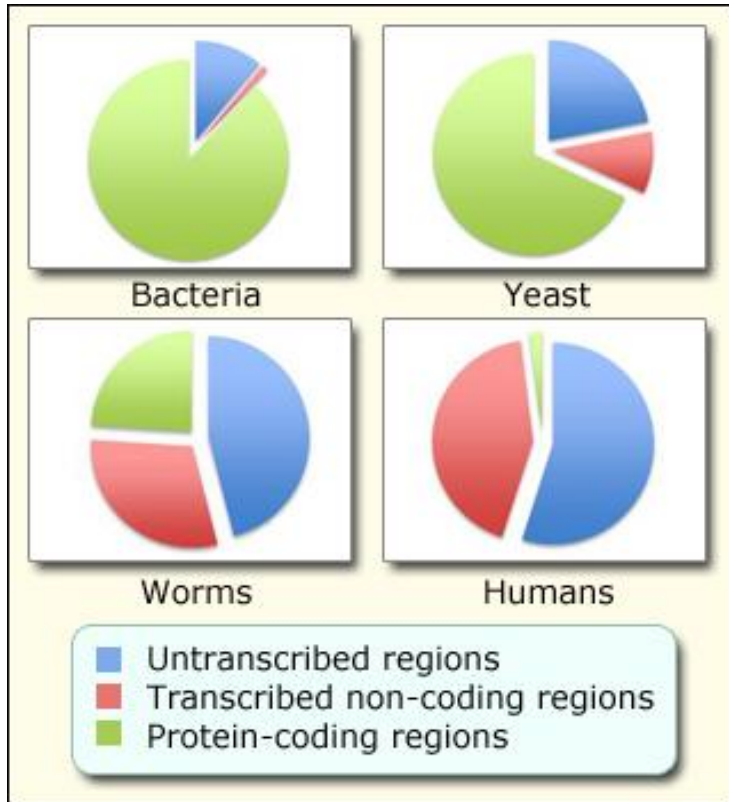
mRNAs: transcripts and isoforms



Molecules of mRNA that are transcribed from protein-coding genes are called **transcripts**.

All transcripts that originate from the same gene, created by the process of alternative splicing, are called **isoforms**.

Transcription - how much of DNA is transcribed?



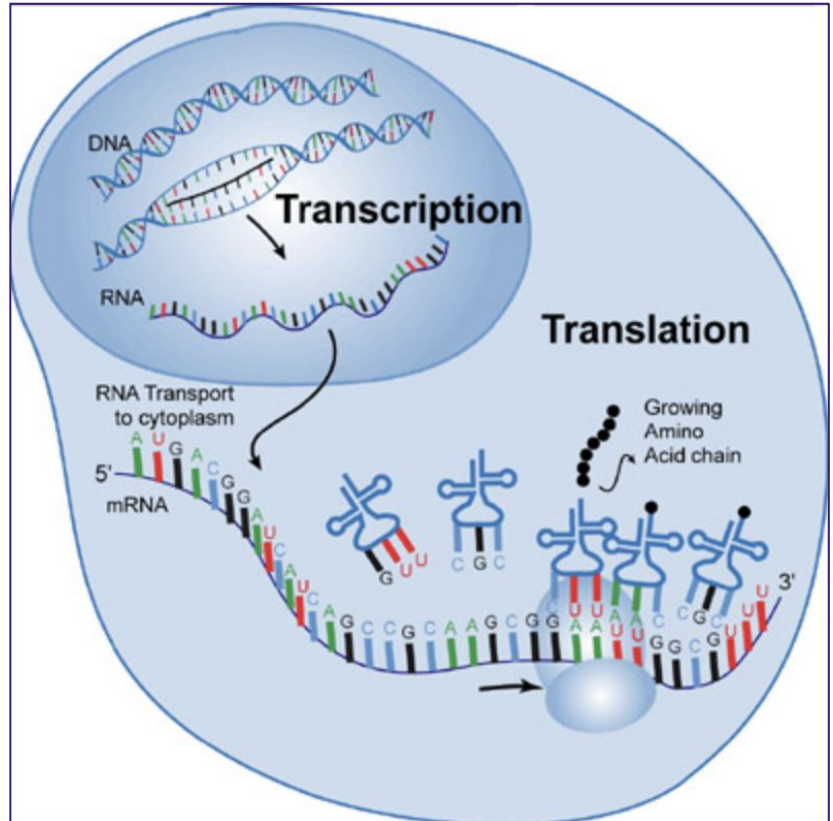
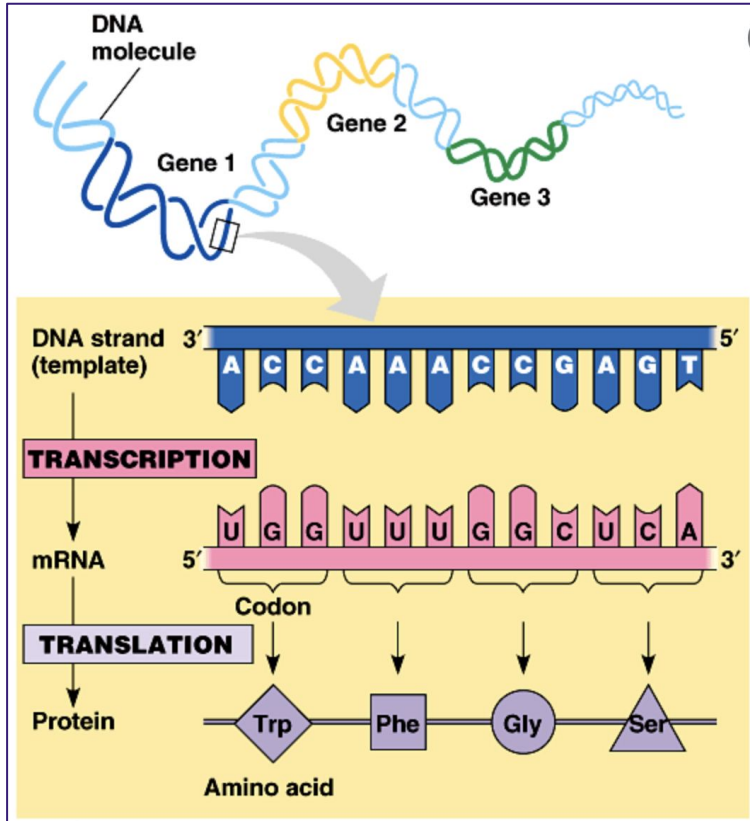
Gene - segment of DNA which is transcribed into RNA which then has a function in cell

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)

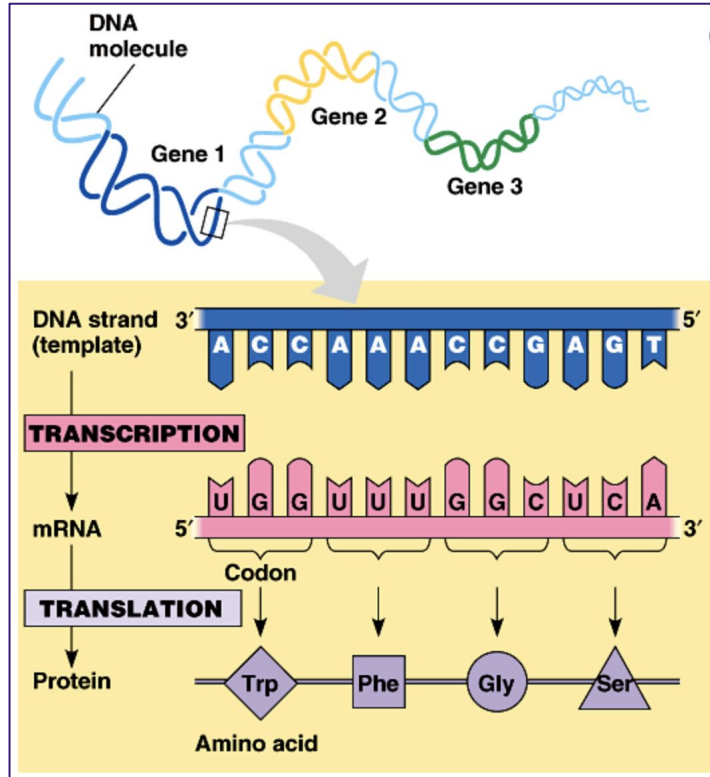
Genes which code for RNA with different functions other than protein coding - structural, regulatory, transport etc. - **non-coding genes** (red)

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

mRNAs: translation to proteins



mRNAs: translation to proteins



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

GTF (gene transfer format)

<u>Col 1</u>	<u>Col 2</u>	<u>Col 3</u>	<u>Col 4</u>	<u>Col 5</u>	<u>Col 6</u>	<u>Col 7</u>	<u>Col 8</u>	<u>Col 9</u>
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..



Reference



Known gene
models

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 library prep

Step 4: Add sequencing adaptors



Step 5: PCR amplify



Step 6: QC

- 1) Verify library concentration
- 2) Verify library fragment lengths

The adaptors do two things:

- 1) Allow the sequencing machine to recognize the fragments
- 2) Allow sequencing of different samples in the same time, since different samples can use different adapters

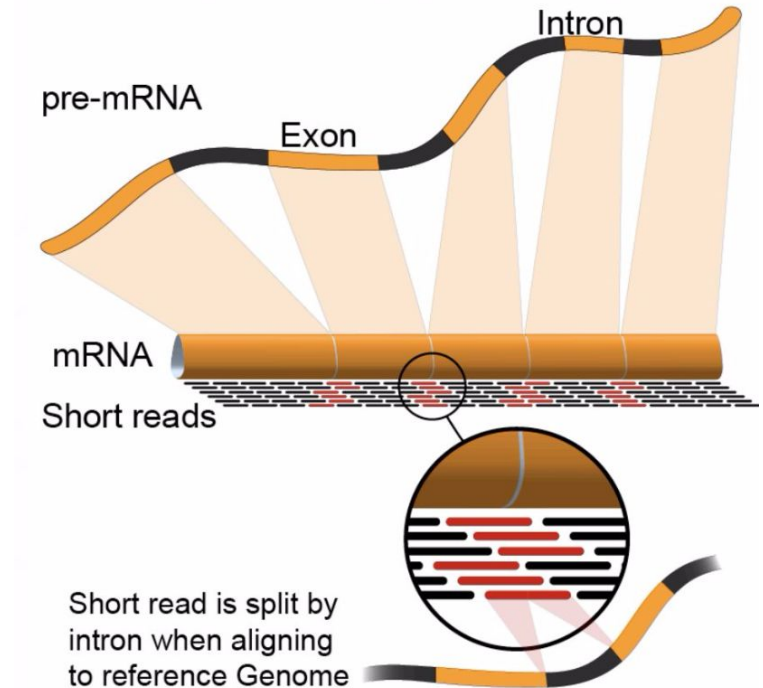
RNA-seq

Splice-aware alignment

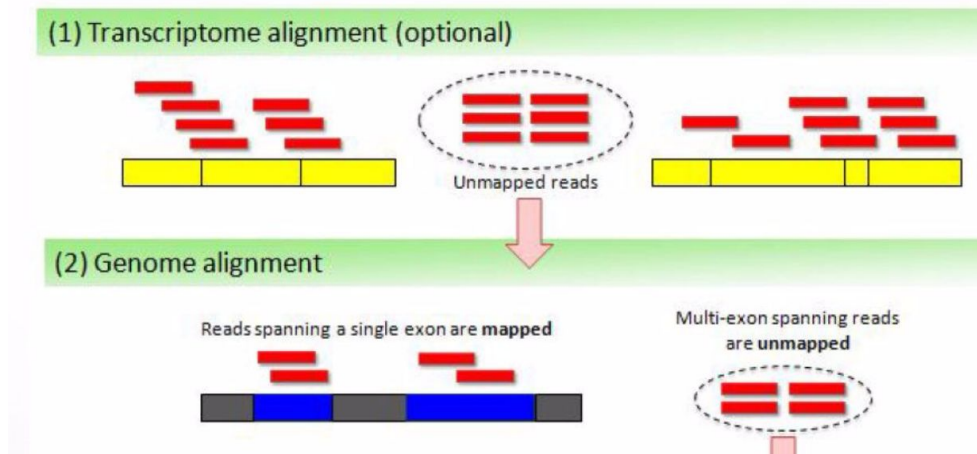


Splice-aware alignment

- Average gene size ~ 10-15 kbp
- Average length of mRNA ~ 2200bp
- Average exon ~ 230bp
- Average number of exons ~ 9.5
- For 100bp reads ~ 35% of reads would span exons



Splice-aware alignment



Splice-aware alignment

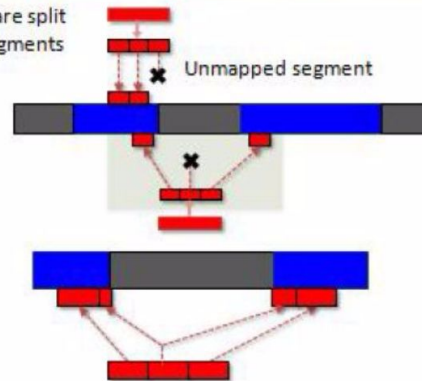
(3) Spliced alignment

(3-1) Segment alignment to genome

(3-2) Identification of splice sites (including indels and fusion break points)

Reads are split
into segments

Unmapped segment



Questions?

