SevenBridges

Intro to RNA-seq

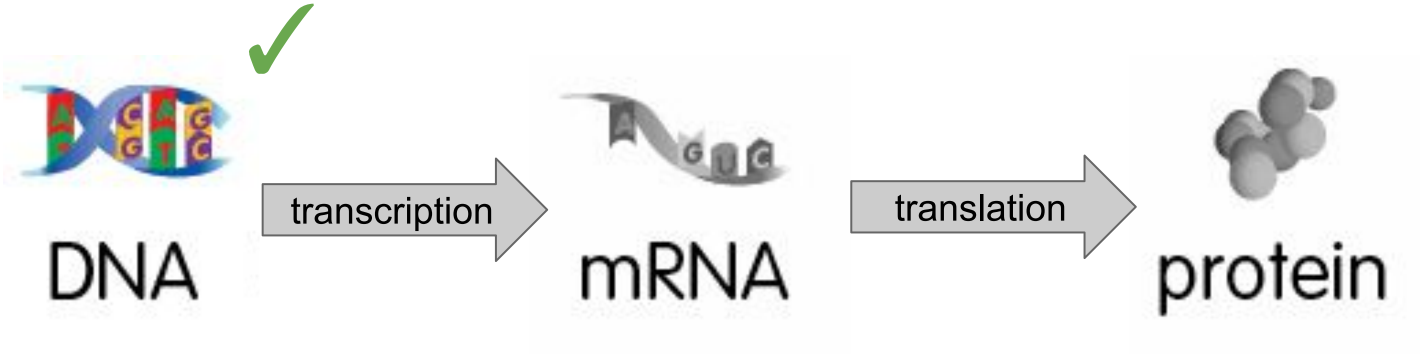
Milena Stanojevic
milena.stanojevic@sbgenomics.com

April 2021

Beyond DNA

Recap

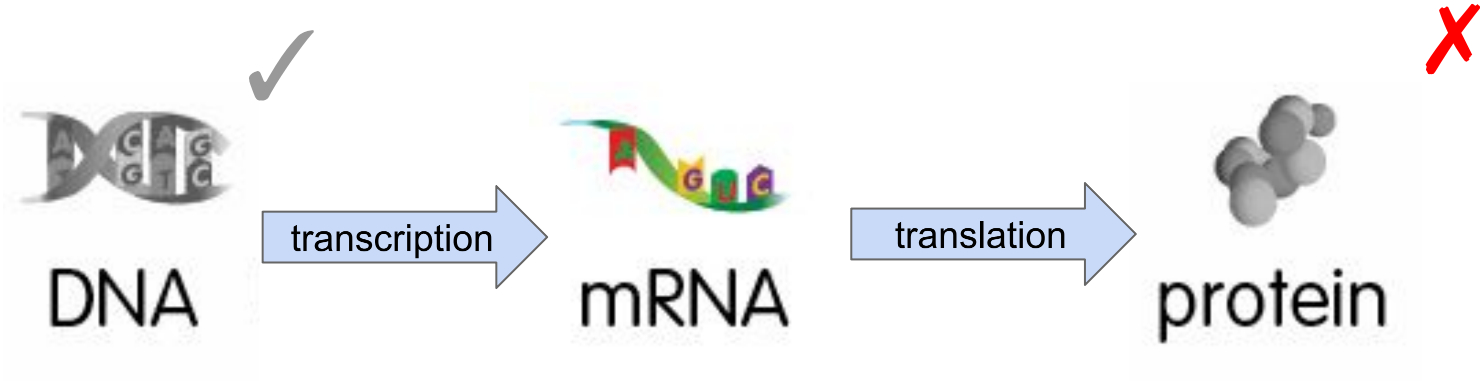
Central dogma of molecular biology



Topics covered so far:

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

Central dogma of molecular biology

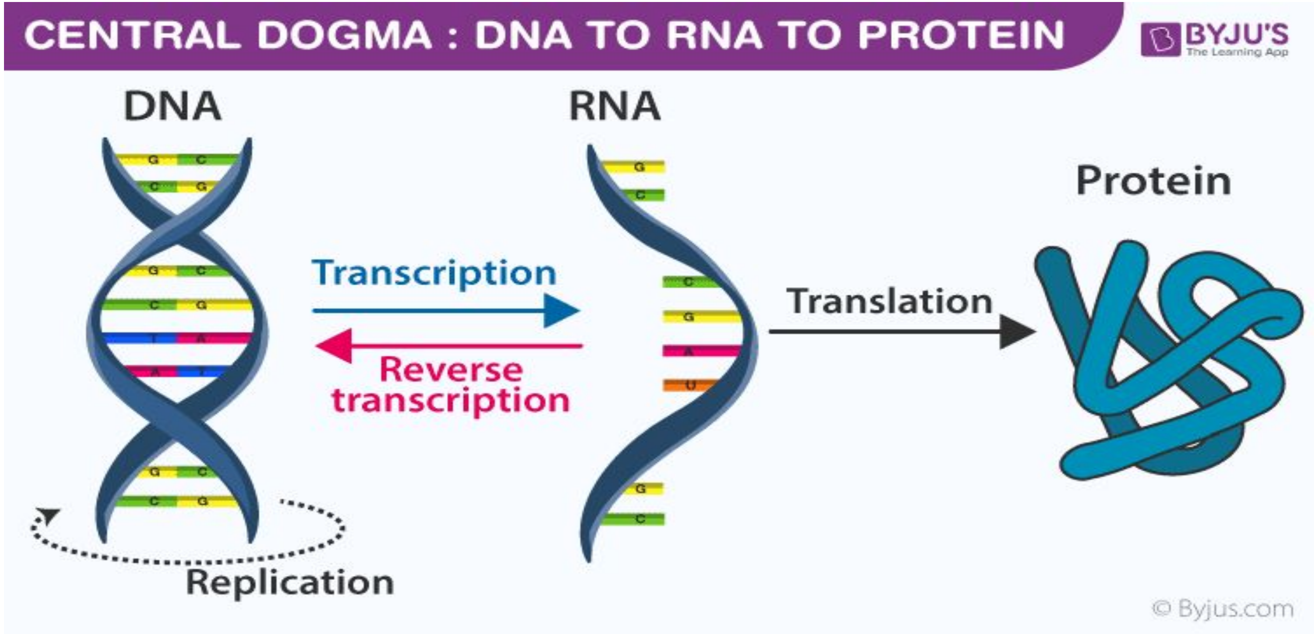


Following topics:

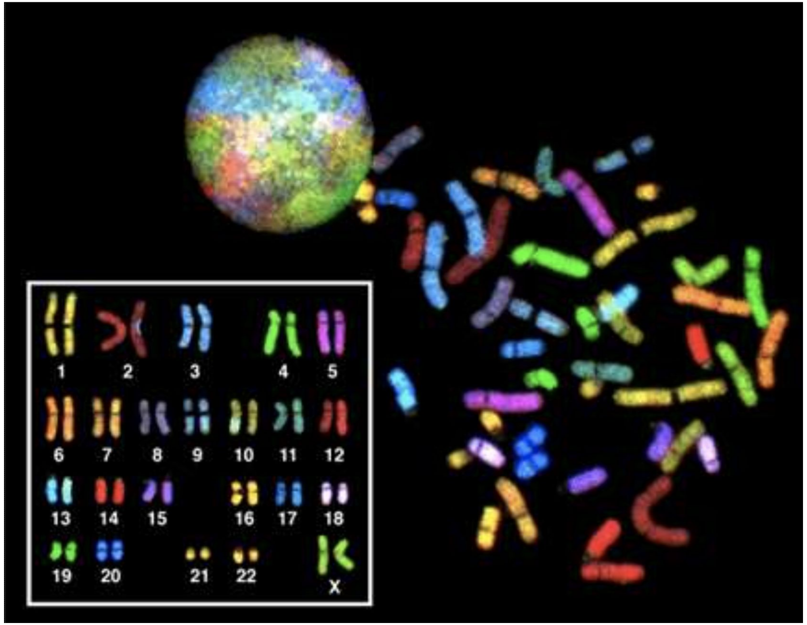
- **RNAs: transcription and translation, types of RNA**
- **mRNA: splicing, transcripts/isoforms**
- **RNA seq and alignment**
- **RNA quantification**
- **Differential expression**

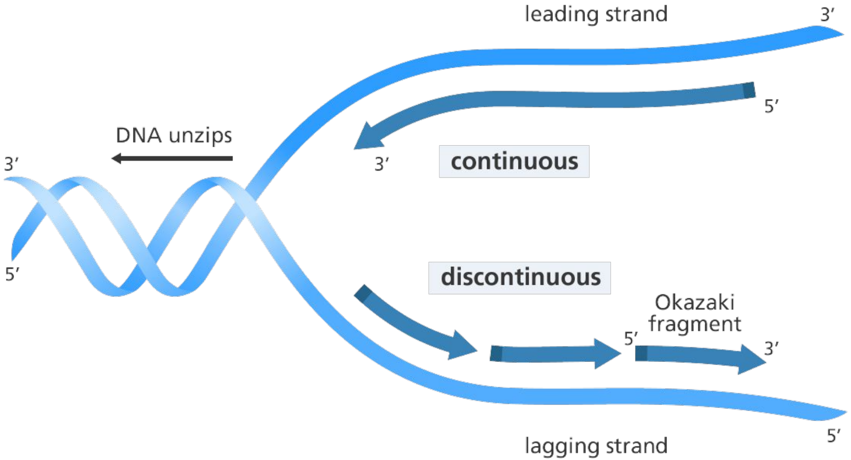
+ Cancer bioinformatics

Central dogma of molecular biology



DNA replication





Results in **all** cells in a body having same DNA

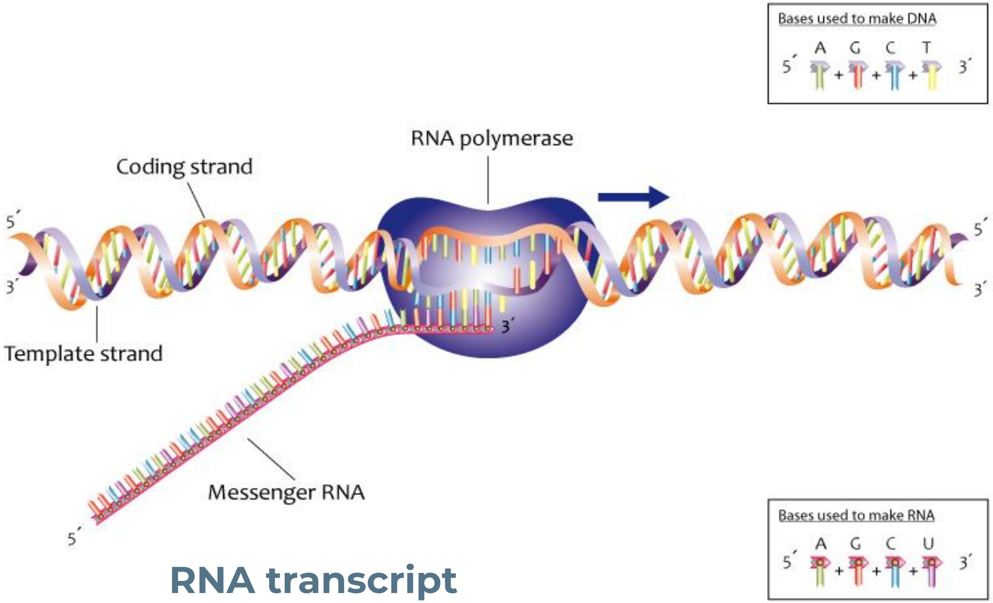
How is that we have different cells in a body?

Transcriptomics

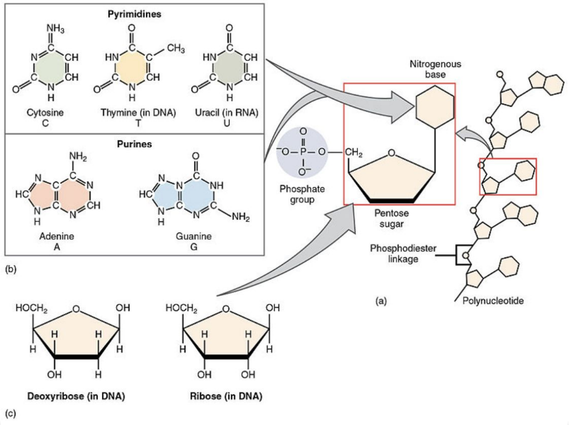
Lots of RNAs, splicing, GTF, translation

Transcription

Synthesis of RNAs from some relatively small genomic regions of DNA

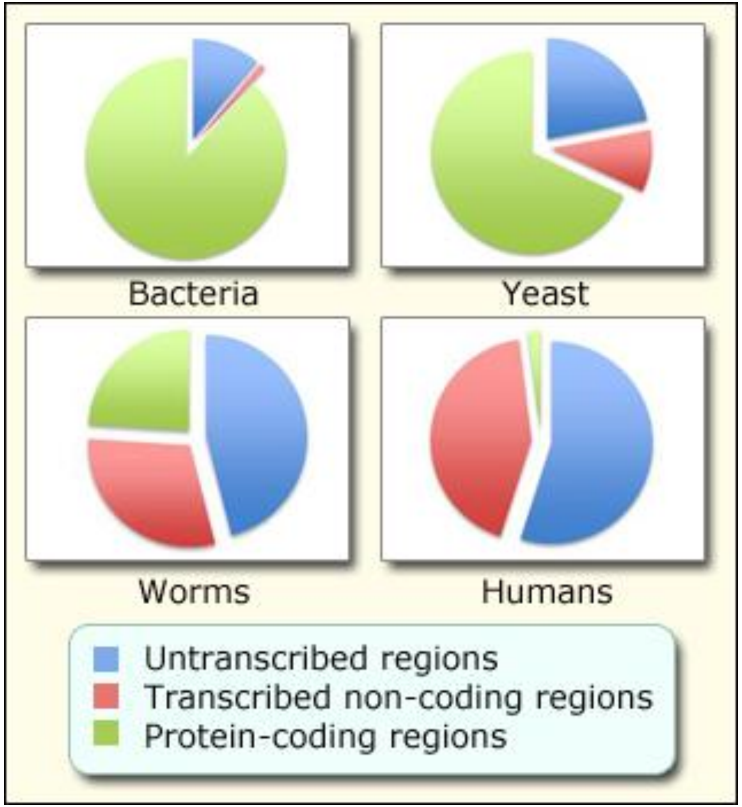


RNA transcript
single-stranded nucleic acid



RNA vs DNA - difference?

Transcription - how much of DNA is transcribed?



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

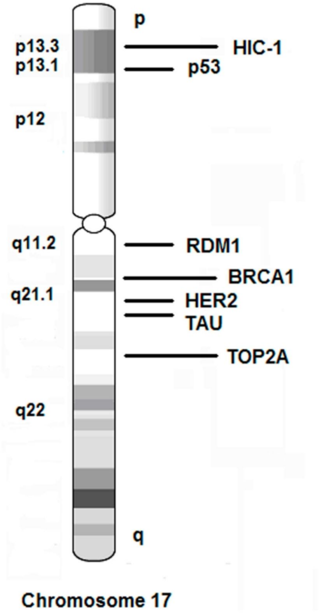
If RNA codes for protein that RNA is called **mRNA** and the region of genome from which it is transcribed is called **protein-coding gene** (green)

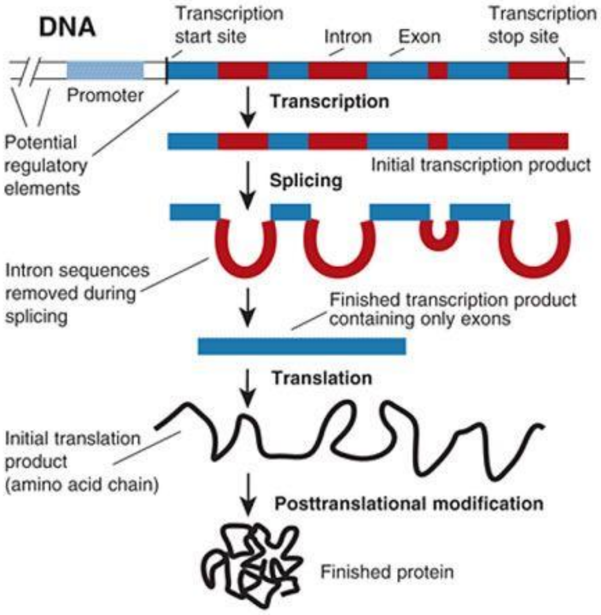
Genes which code for RNA with different functions other than protein coding - structural, regulatory, transport etc. - **non-coding genes** (red)

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

Transcription

Process: Synthesis of RNAs from some relatively small genomic regions



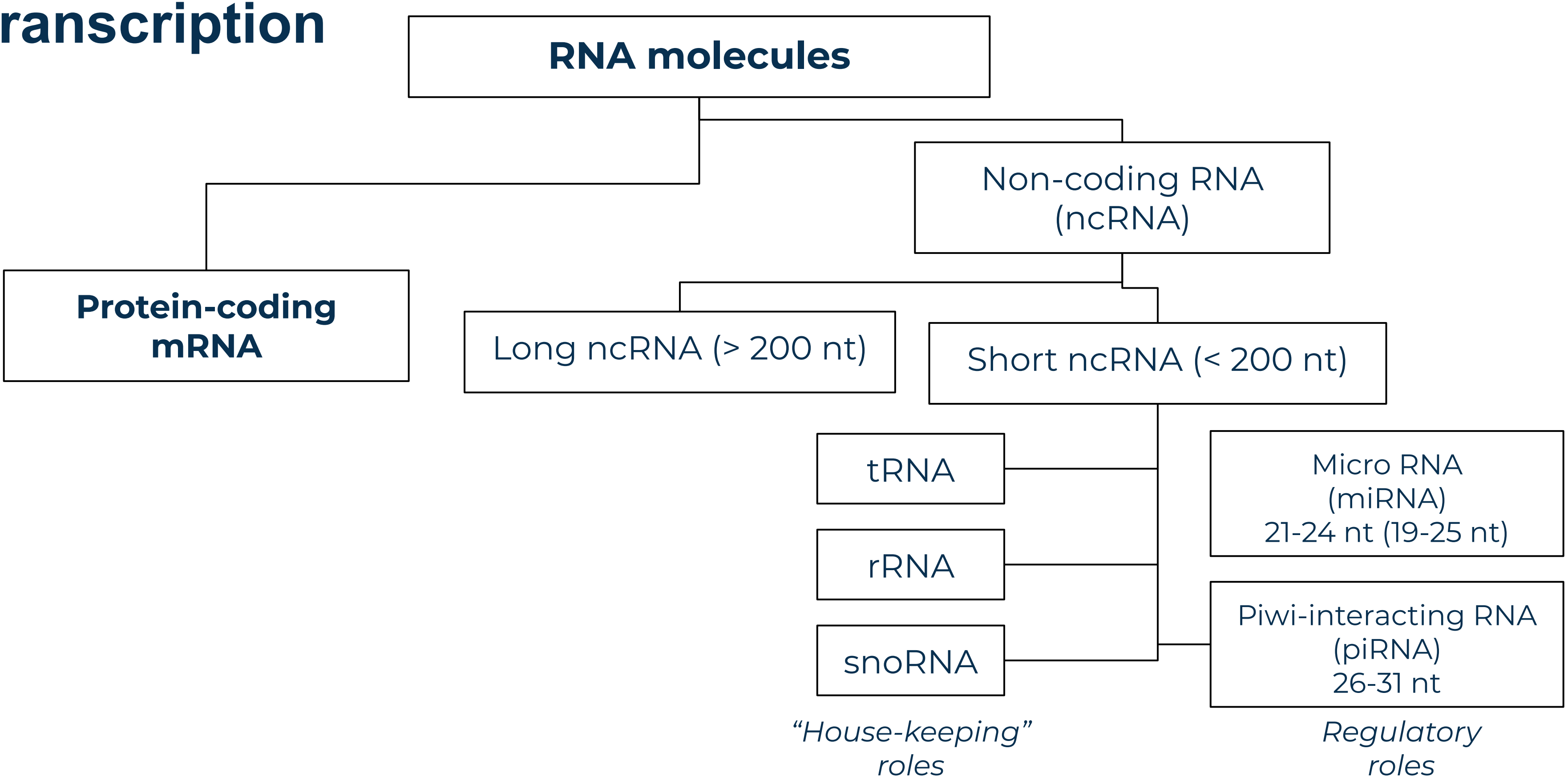

DNA
Transcription start site
Intron
Exon
Transcription stop site
Promoter
Potential regulatory elements
Transcription
Initial transcription product
Splicing
Intron sequences removed during splicing
Finished transcription product containing only exons
Translation
Initial translation product (amino acid chain)
Posttranslational modification
Finished protein

Chr 17 - protein coding genes example

Eucariotic gene structure

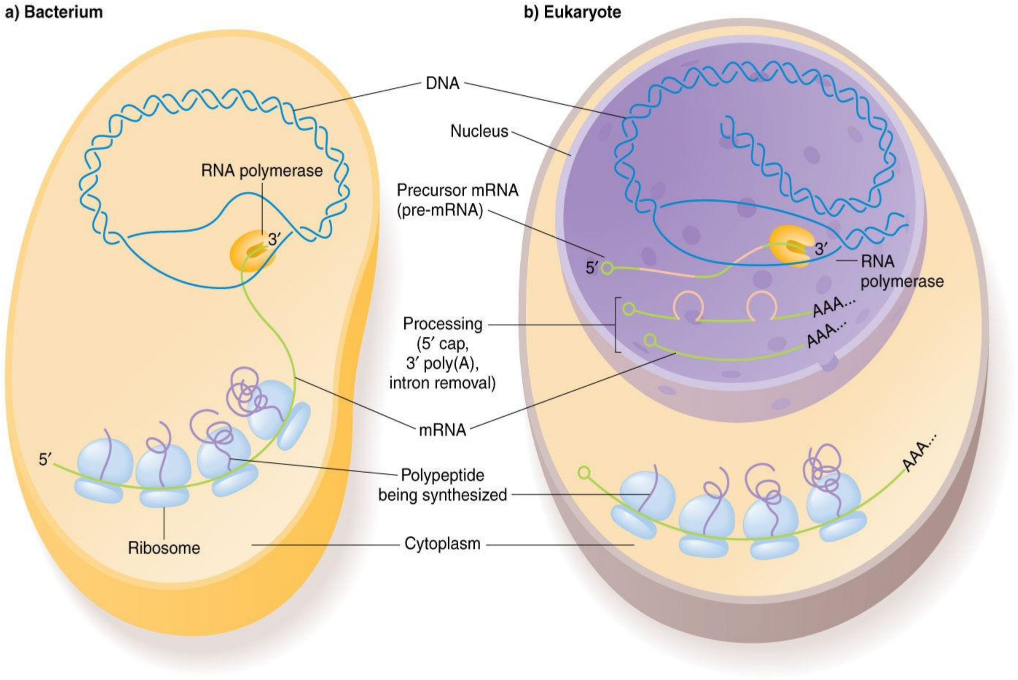
Human cells use splicing and other processes to make multiple proteins from the instructions encoded in a single gene

Transcription

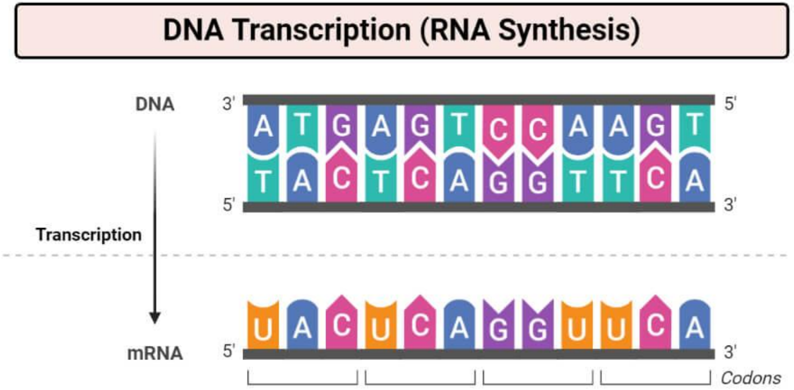


mRNAs

Synthesis and maturation in nucleus

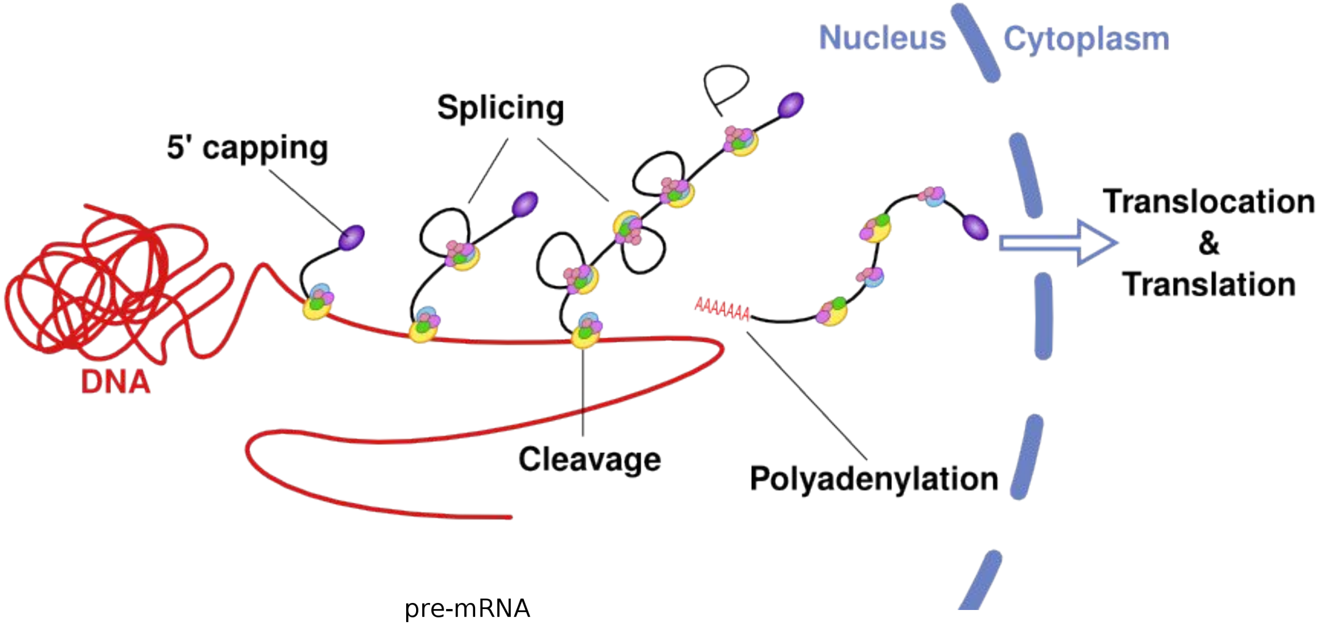


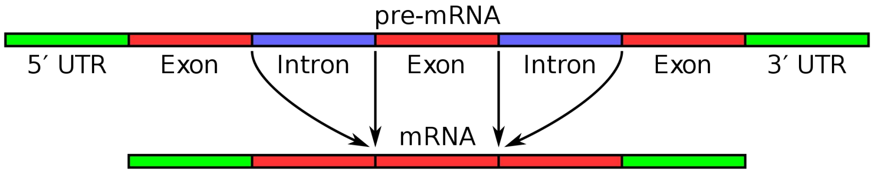
© 2010 Pearson Education, Inc.



mRNAs: maturation

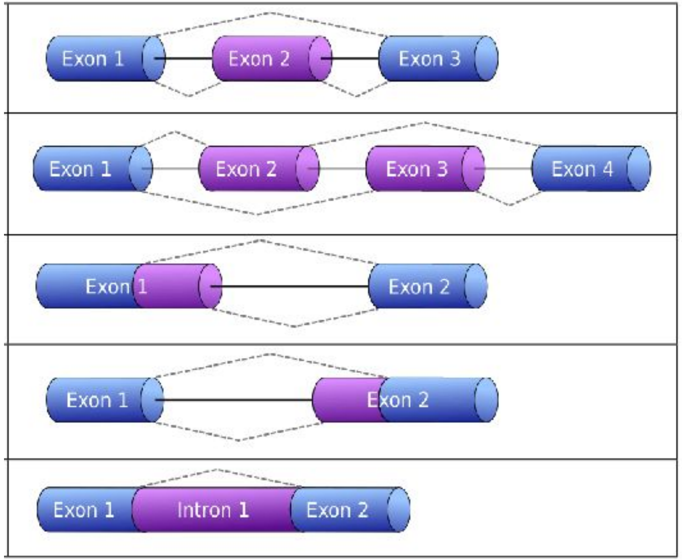
mature mRNA = spliced transcript + poly-A tail

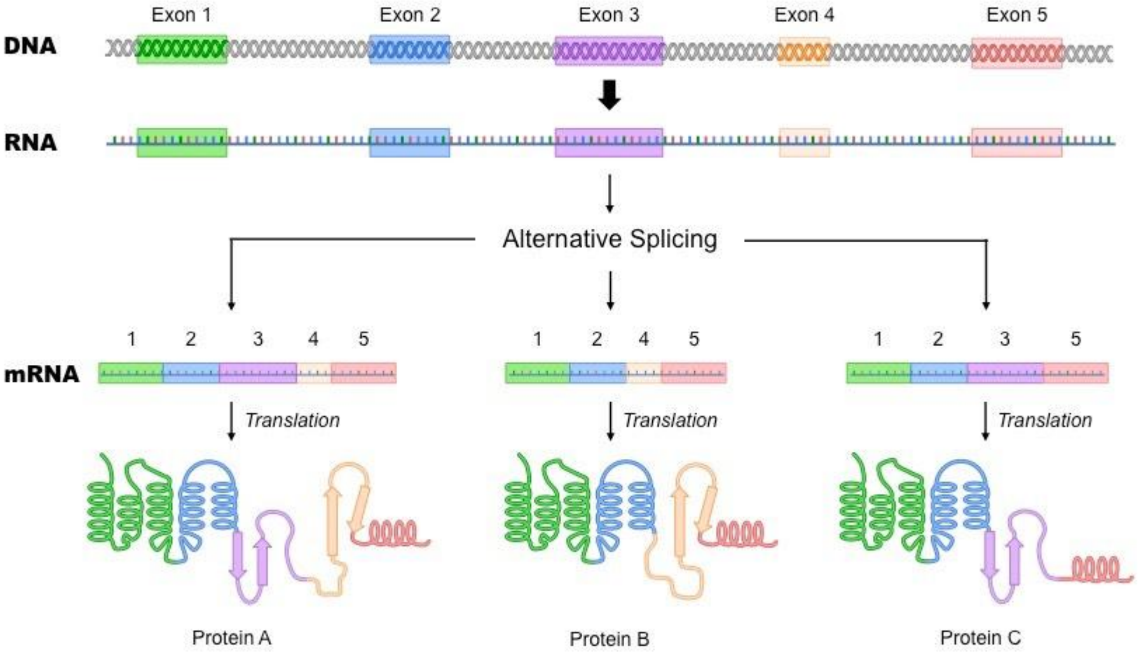




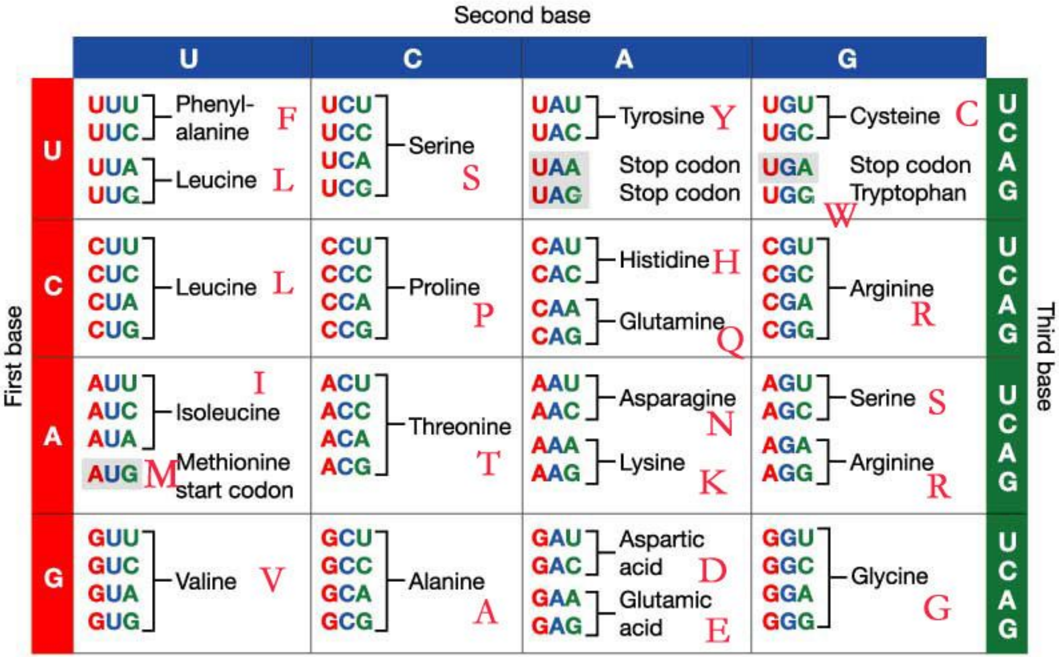
mRNAs: alternative splicing

transcripts from same gene: *isoforms*

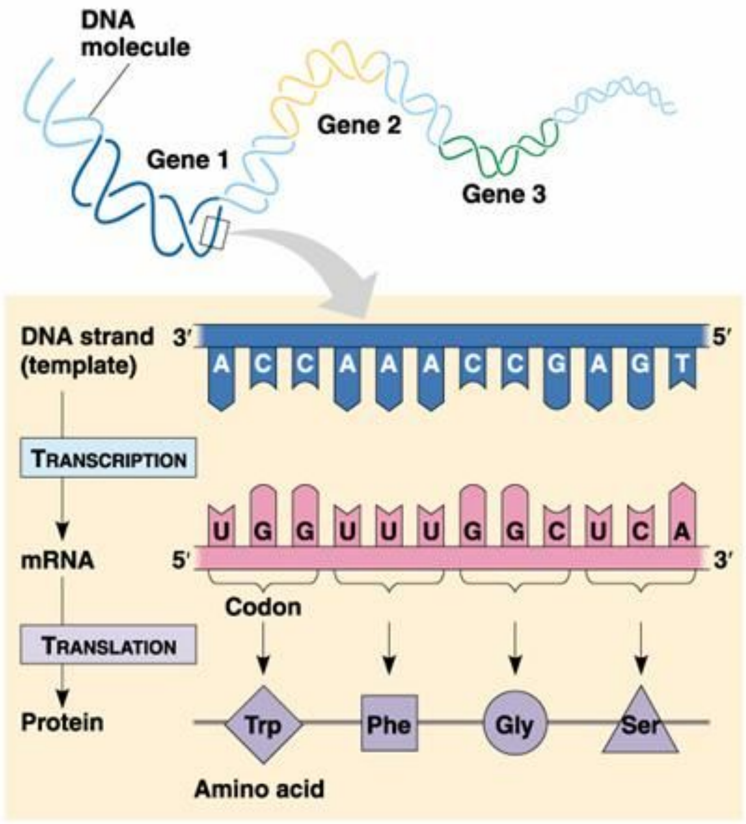




mRNAs: translation to proteins

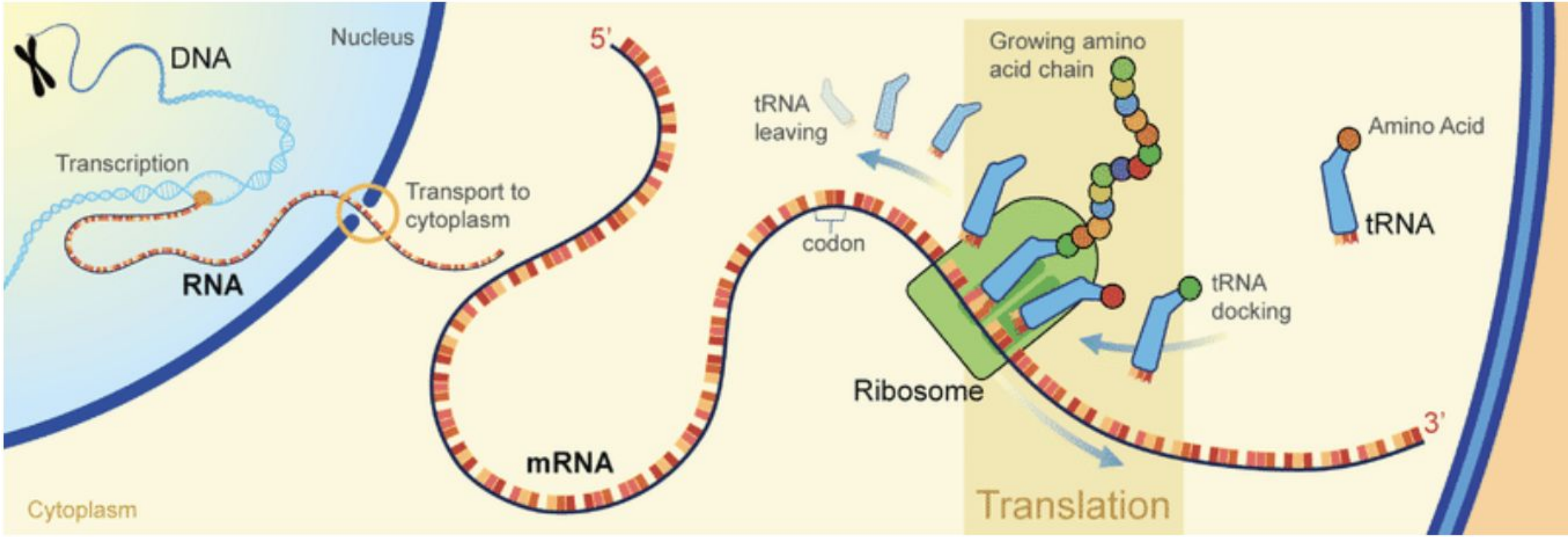


First base	Second base: U	Second base: C	Second base: A	Second base: G	Third base
U	UUU Phenyl-alanine F	UCU Serine S	UAU Tyrosine Y	UGU Cysteine C	U
	UUC	UCC	UAC	UGC	C
	UUA Leucine L	UCA	UAA Stop codon	UGA Stop codon	A
	UUG	UCG	UAG Stop codon	UGG Tryptophan W	G
C	CUU Leucine L	CCU Proline P	CAU Histidine H	CGU Arginine R	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Glutamine Q	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Isoleucine I	ACU Threonine T	AAU Asparagine N	AGU Serine S	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lysine K	AGA Arginine R	A
	AUG M Methionine start codon	ACG	AAG	AGG	G
G	GUU Valine V	GCU Alanine A	GAU Aspartic acid D	GGU Glycine G	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glutamic acid E	GGA	A
	GUG	GCG	GAG	GGG	G

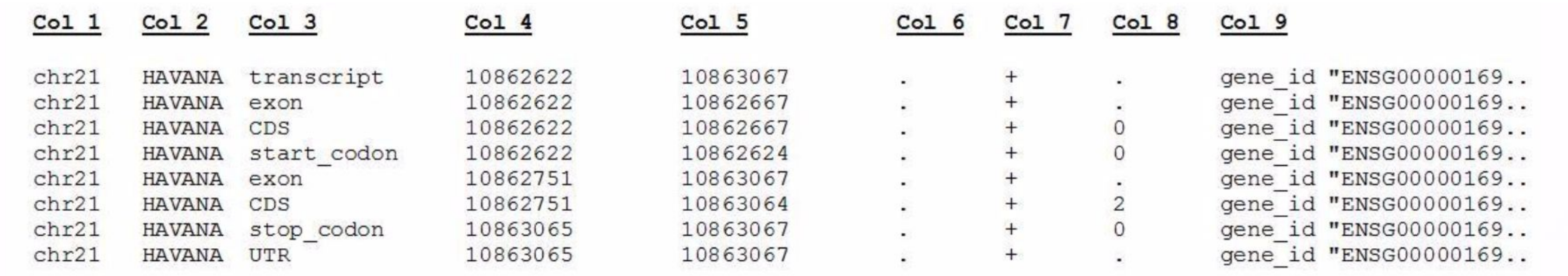


©1999 Addison Wesley Longman, Inc.

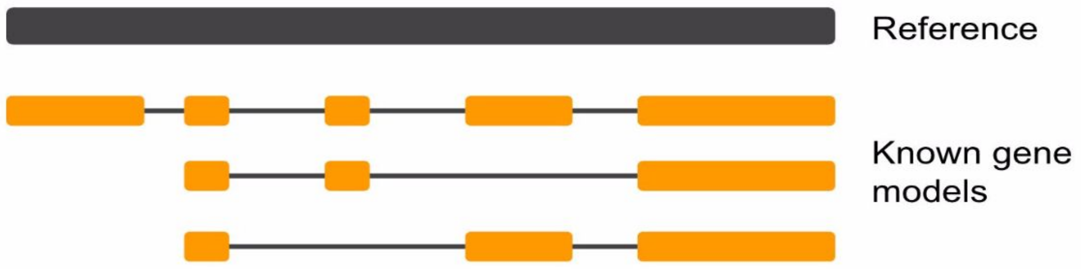
mRNAs: translation to proteins



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..



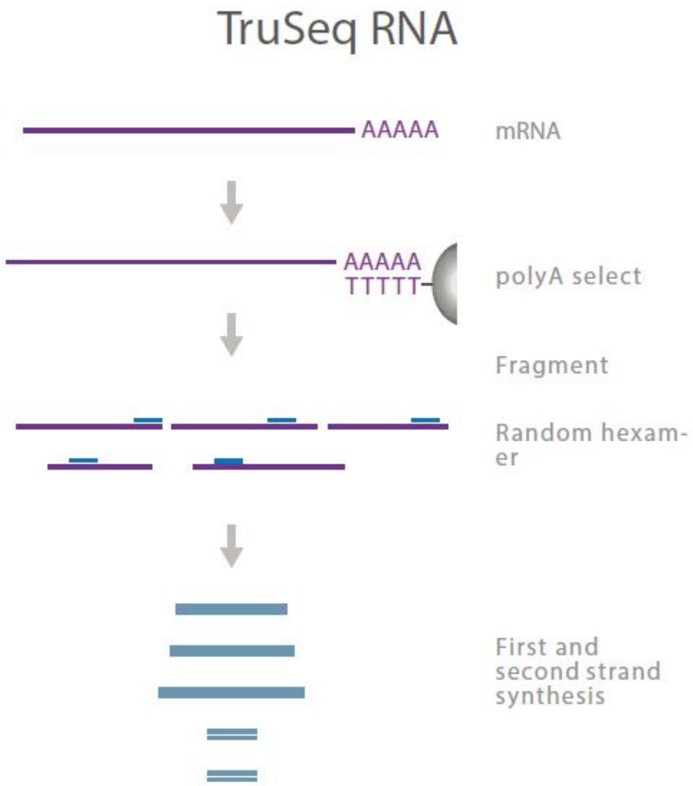
RNA-seq

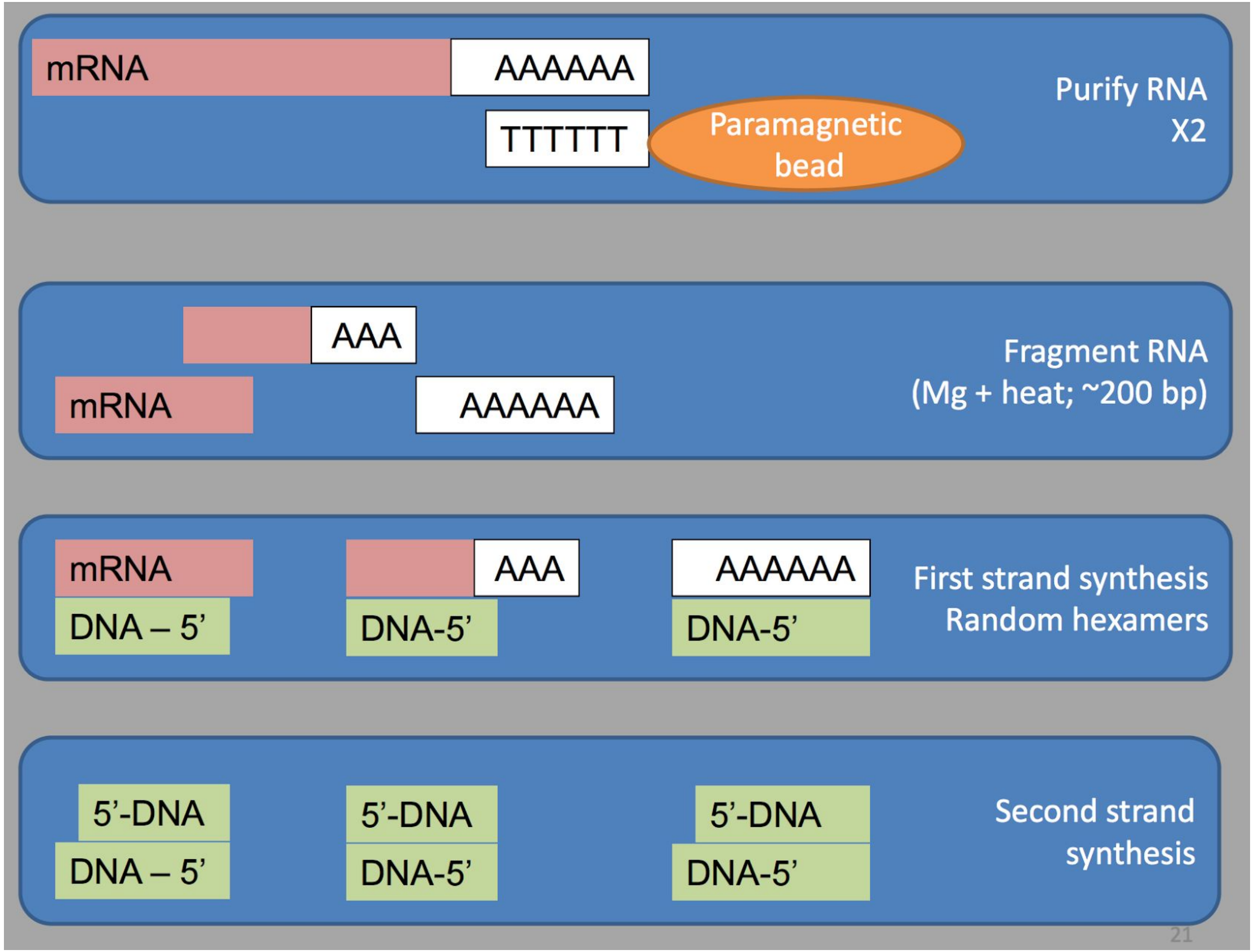
Library preparation

RNA-seq library prep (1/2)

Protocols differ in :

- what types of RNA they target (total RNA, mRNA)
- on fragment sizes
- strand specificity
- bulk or single-cell



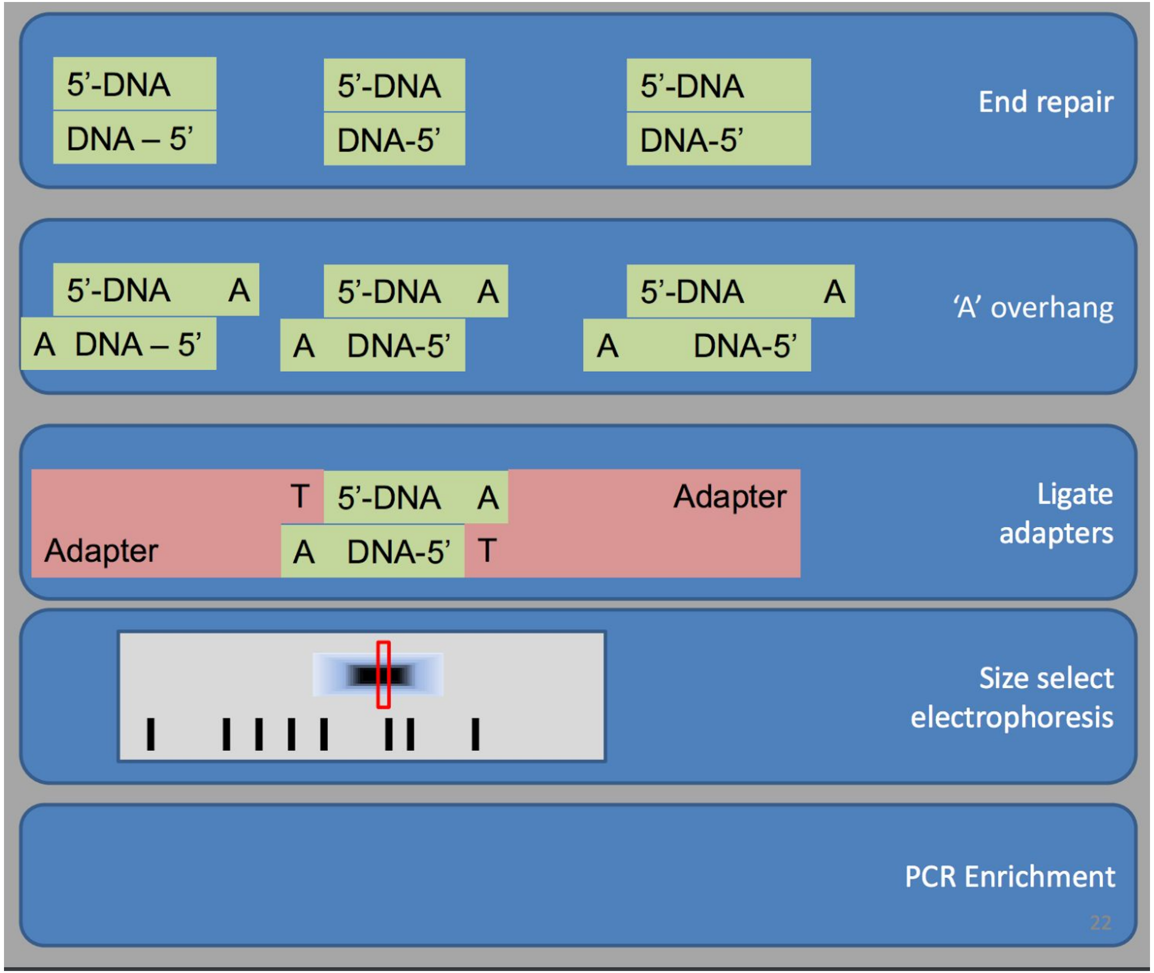


Purify RNA
X2

Fragment RNA
(Mg + heat; ~200 bp)

First strand synthesis
Random hexamers

Second strand synthesis



End repair

'A' overhang

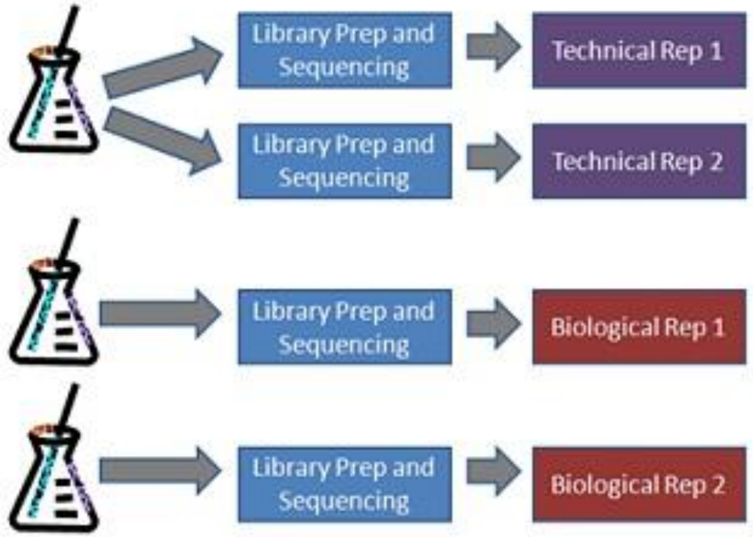
Ligate adapters

Size select electrophoresis

PCR Enrichment

RNA-seq library prep (2/2)

- two sources of variation in RNA-seq
- technical or biological replicates
- Important to estimate # replicates (power analysis)


Library Prep and Sequencing
Technical Rep 1
Library Prep and Sequencing
Technical Rep 2
Library Prep and Sequencing
Biological Rep 1
Library Prep and Sequencing
Biological Rep 2

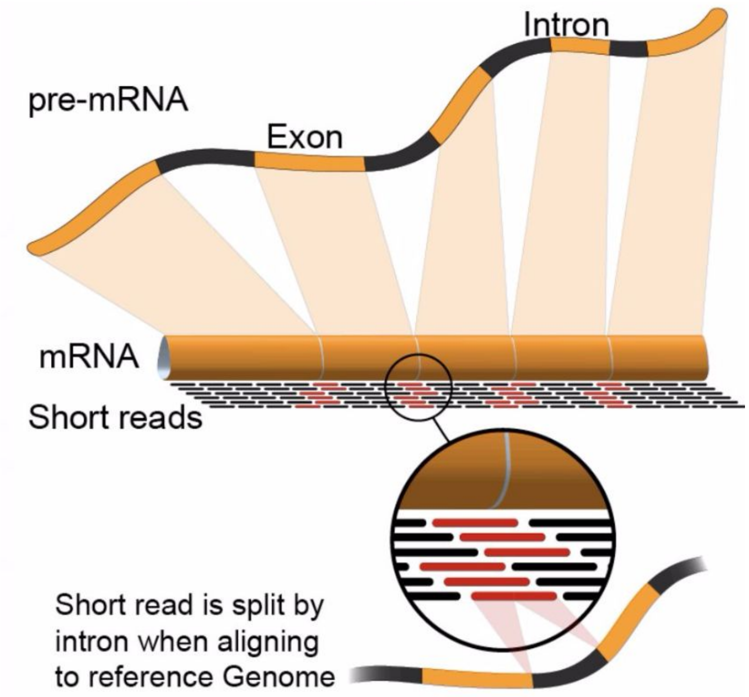
Source: http://hdl.handle.net/2345/3145

RNA-seq

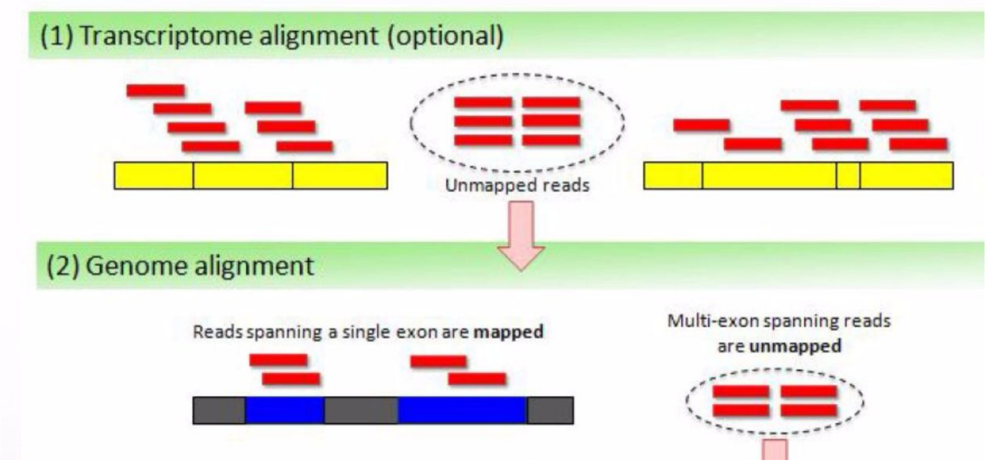
Splice-aware alignment

Splice-aware alignment

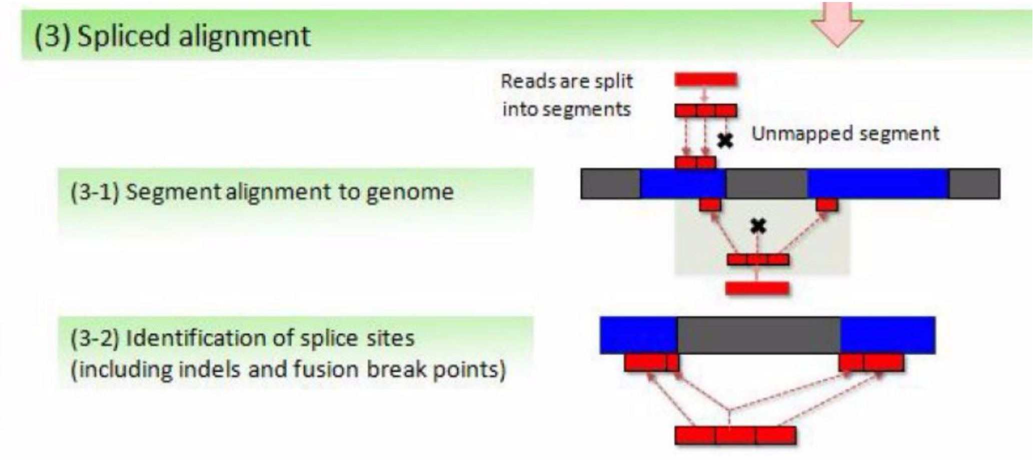
- Average gene size ~ 10-15 kbp
- Average length of mRNA ~ 2200bp
- Average exon ~ 230bp
- Average number of exons ~ 9.5
- For 100bp reads ~ 35% of reads would span exons


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

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)

Questions?