Introduction to Molecular Biology

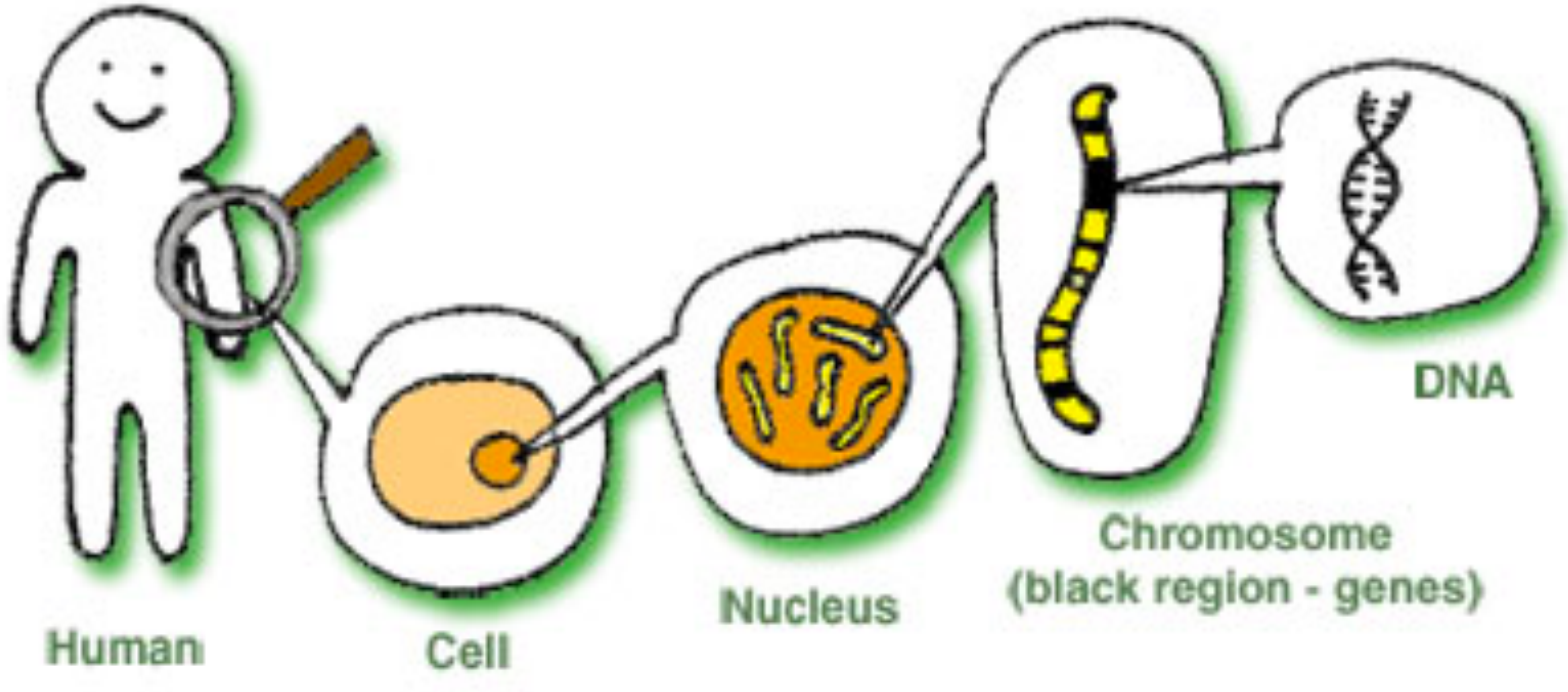
CS 4364 & 5364

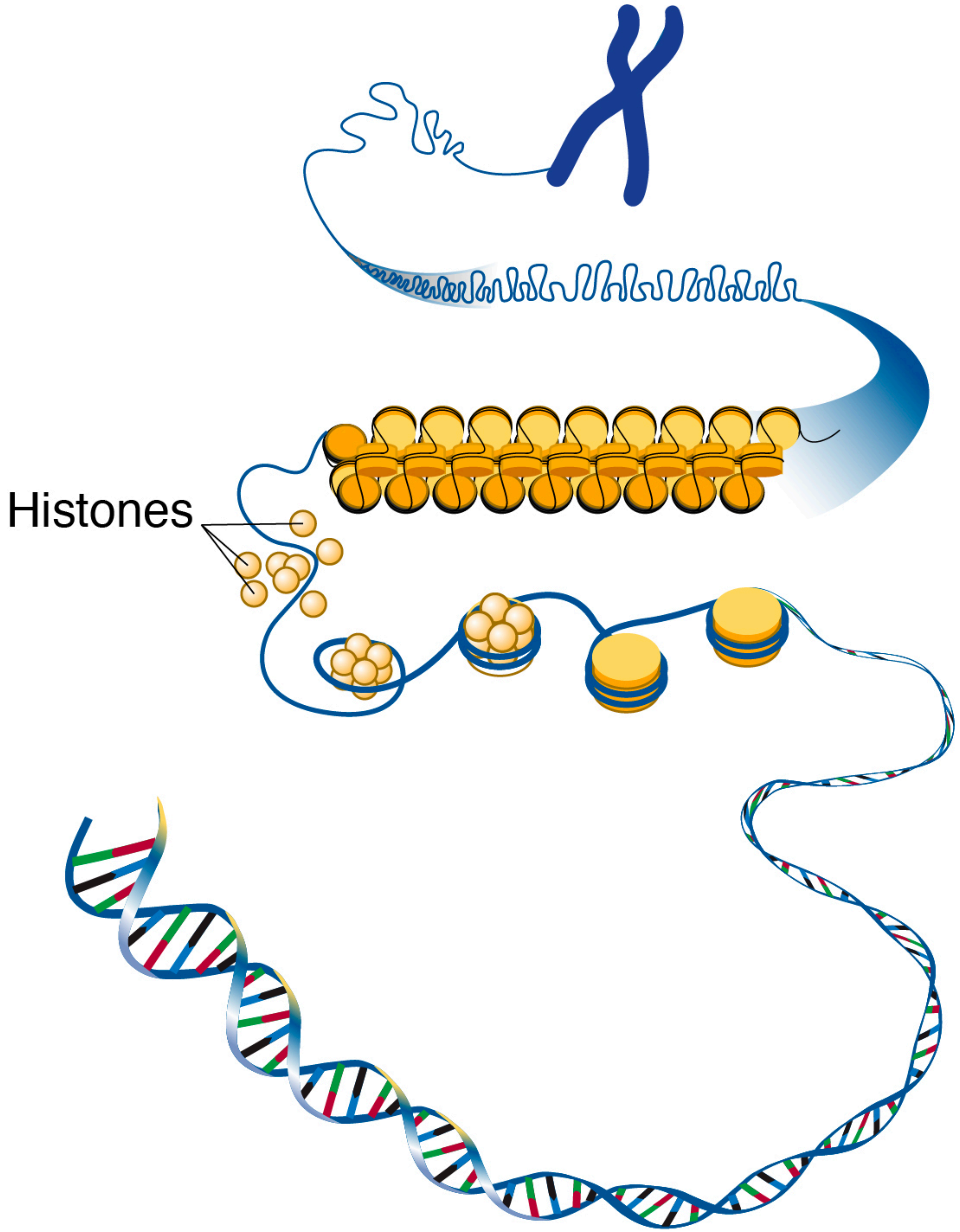
At the highest level

Organism are made up of one or multiple cells

inside the cell is the nucleus, which contains the DNA

humans are *diploid* meaning we have 2 copies of each chromosome (one from each parent)





Histones

Chromosome

Chromatin fiber

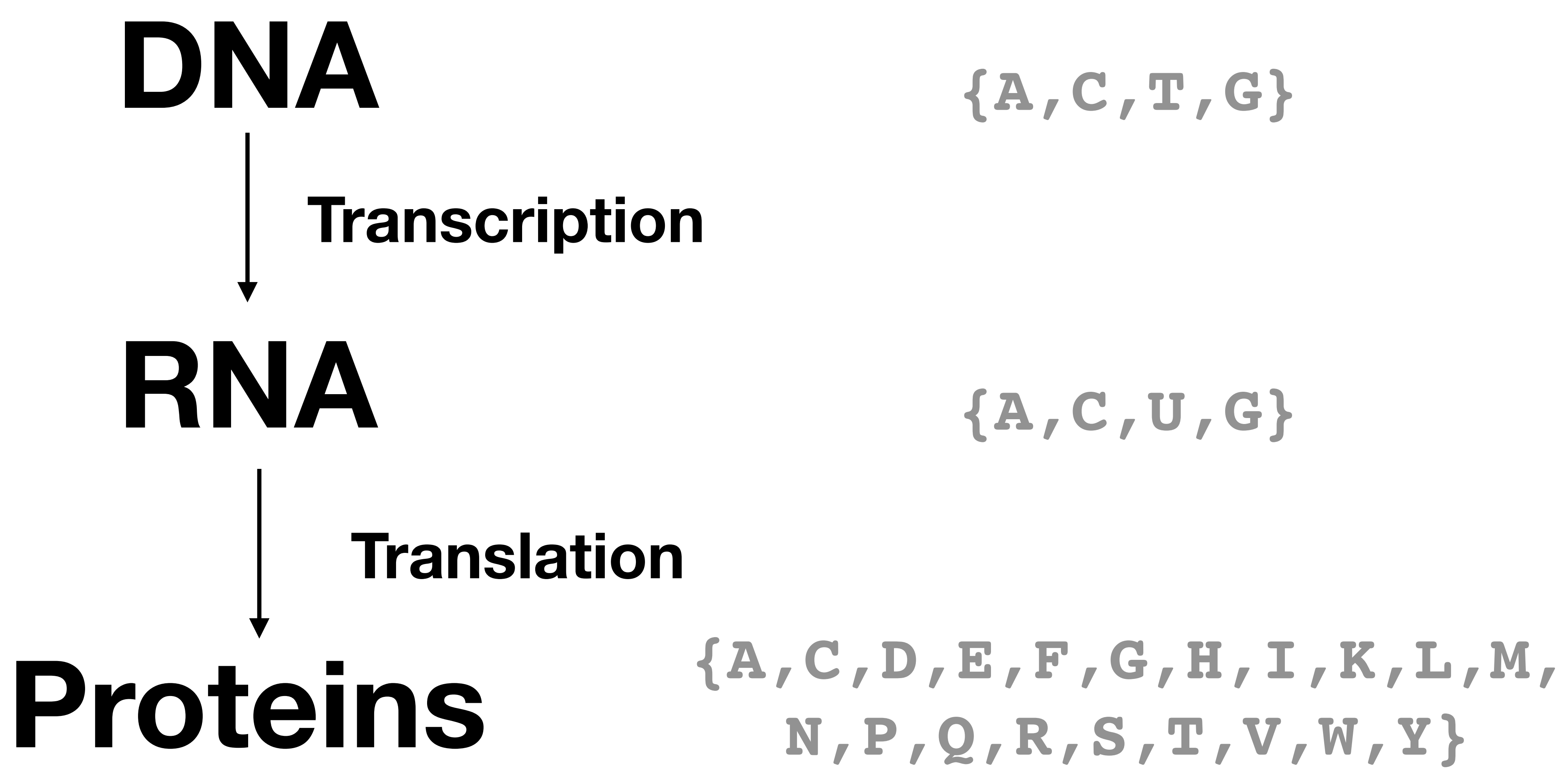
"Beads on a string"
DNA wound on
nucleosomes

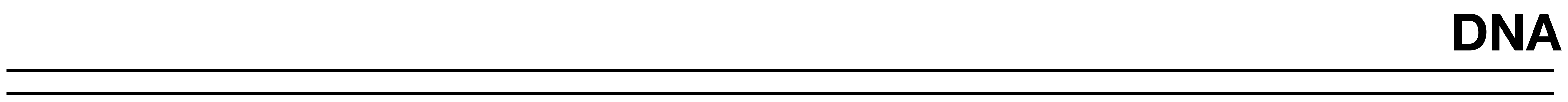
Double helix

The Central Dogma

DNA

- double stranded
- contains all of the information for "you"
- only about 1.5% of the human genome encodes proteins

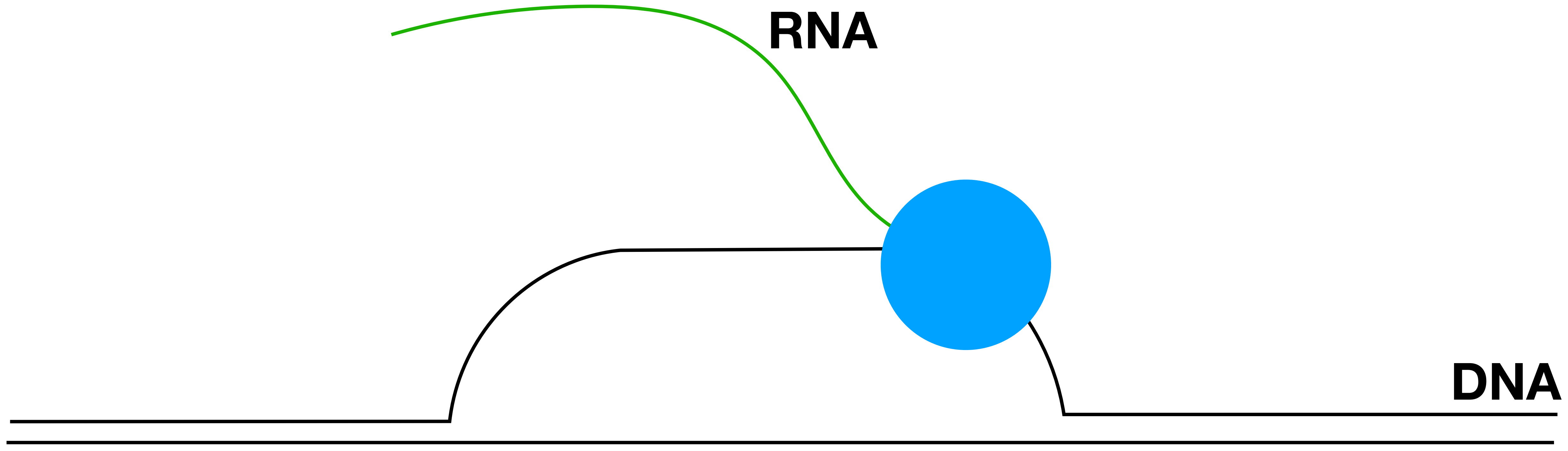


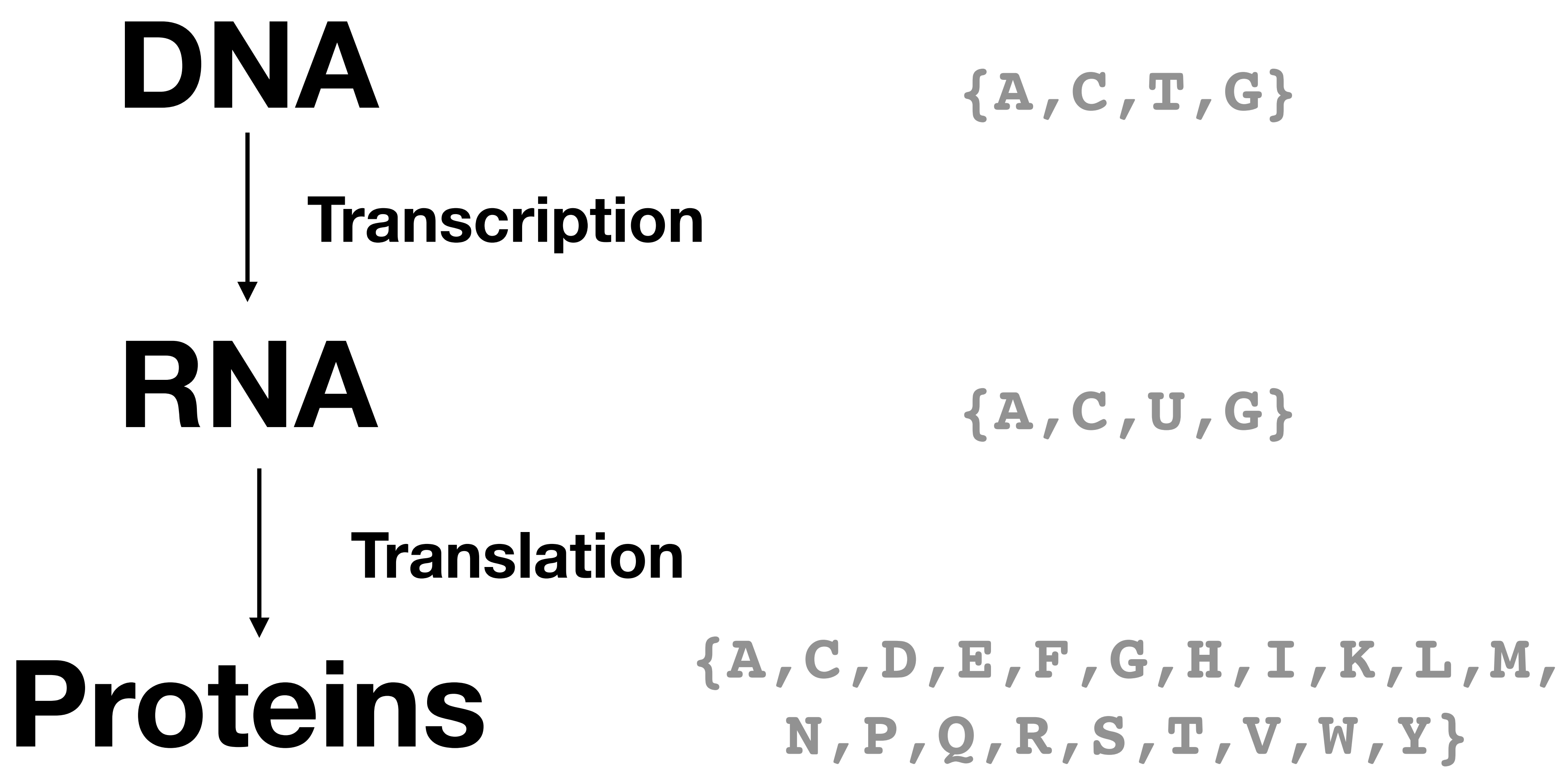


The Central Dogma

Transcription

- process of uncoiling, seperating, and copying DNA into RNA
- first stage is called "pre-mRNA" in the case of protein coding genes





DNA {A,C,T,G}

Transcription

RNA {A,C,U,G}

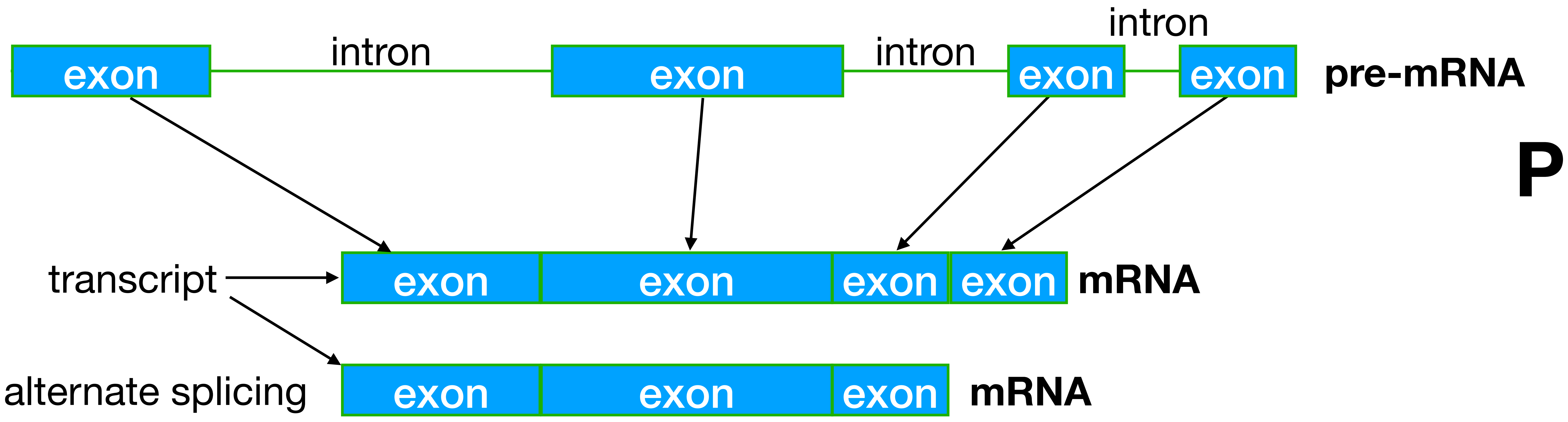
Translation

Proteins {A,C,D,E,F,G,H,I,K,L,M, N,P,Q,R,S,T,V,W,Y}

The Central Dogma

RNA

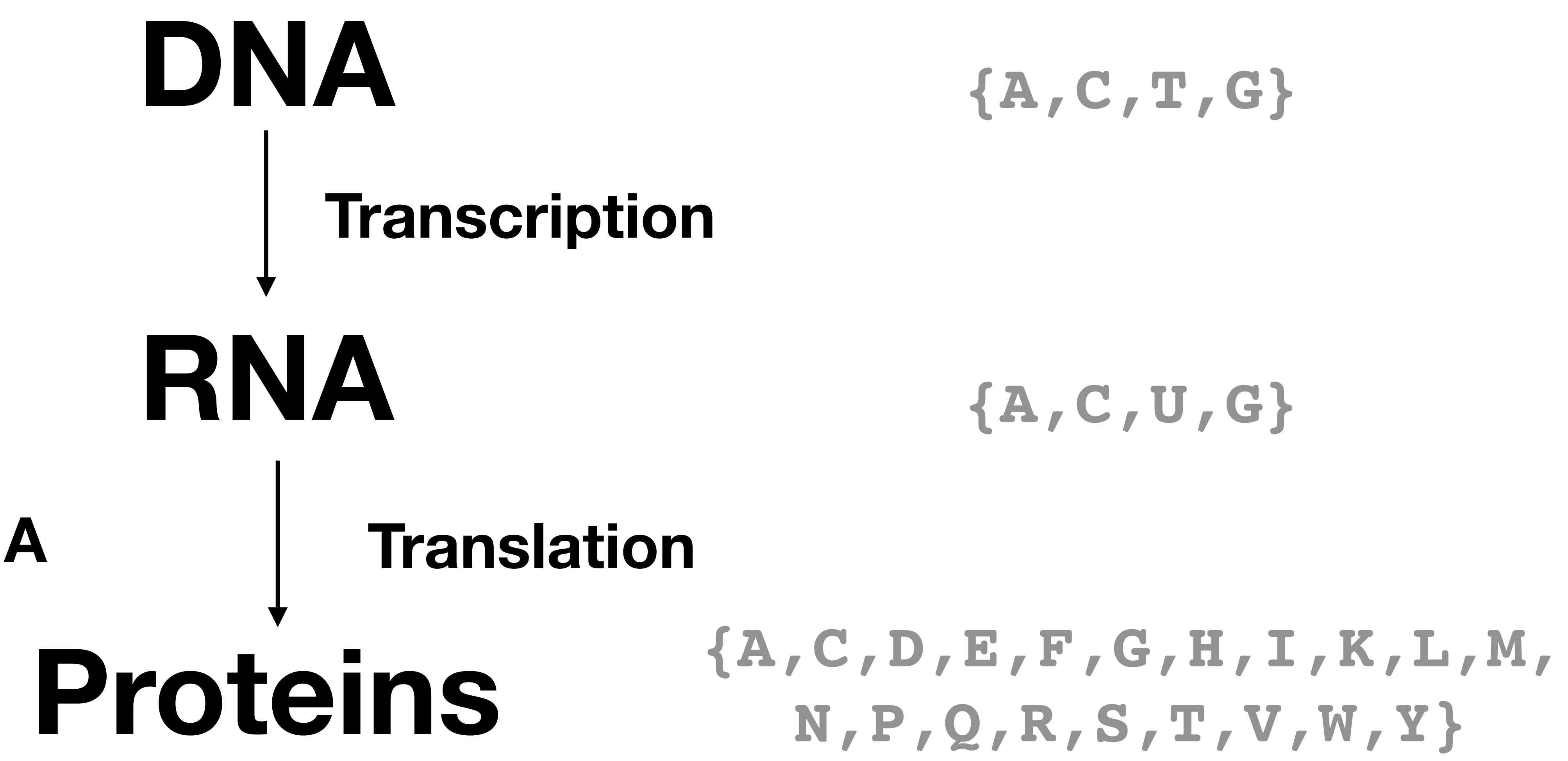
- pre-mRNA undergo splicing to remove the *introns* and leave only (some) *exons*
- some RNA perform functions on their own and are not spliced, called ncRNA (non-coding RNA)



exon intron exon intron exon intron exon **pre-mRNA**

transcript → exon exon exon exon **mRNA**

alternate splicing exon exon exon **mRNA**



DNA {A,C,T,G}

Transcription

RNA {A,C,U,G}

Translation

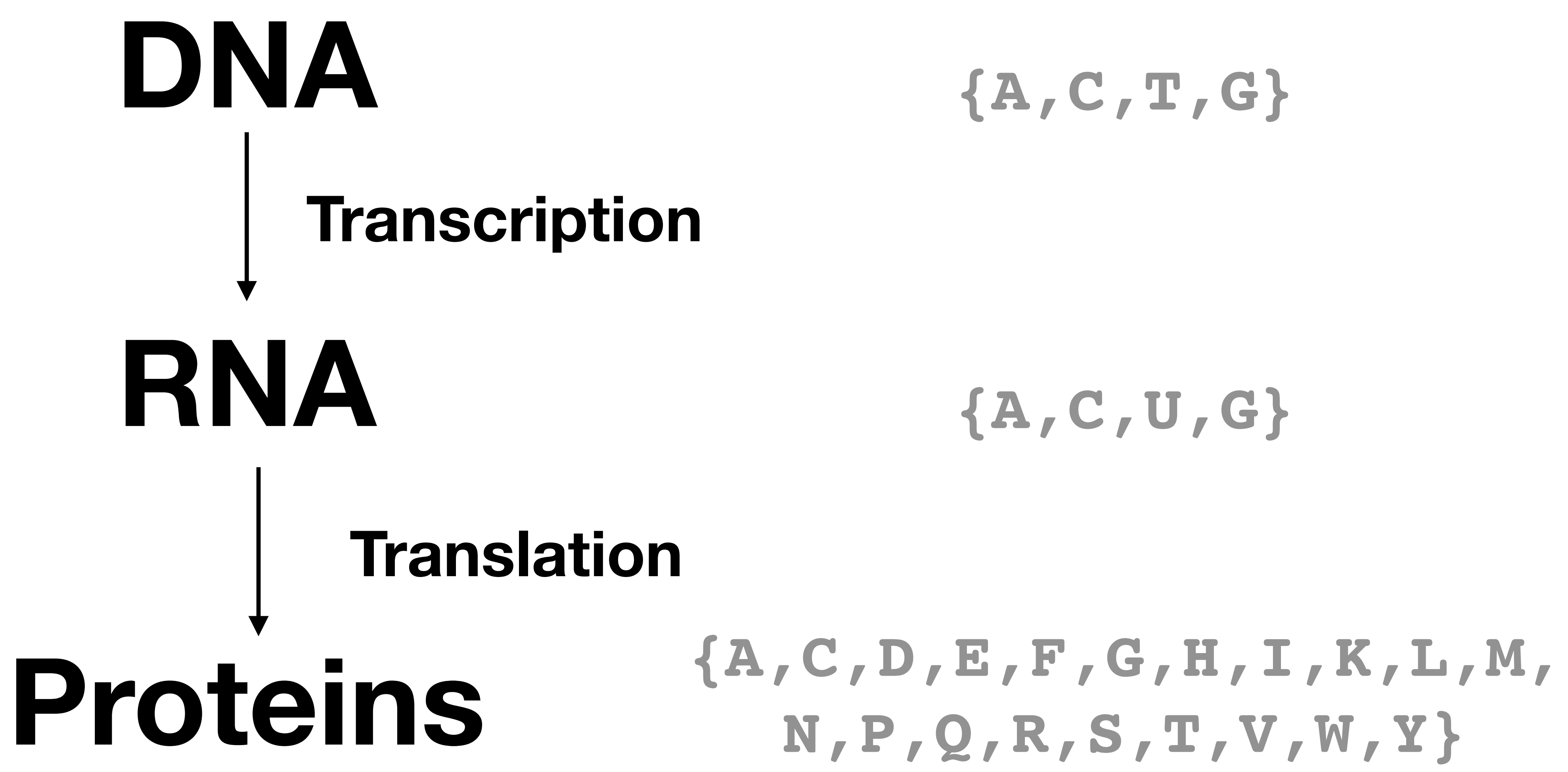
Proteins {A,C,D,E,F,G,H,I,K,L,M, N,P,Q,R,S,T,V,W,Y}

The Central Dogma

Translation

- 3-letter groups of RNA characters, *codons,* are converted to amino acids, the building blocks for proteins

	Second Character								
First Char.	A		C		U		G		Third .Char
A	AAC	N	ACC	T	AUC	I	AGC	S	C
	AAU		ACU		AUU		AGU		U
	AAA	K	ACA		AUA		AGA	R	A
	AAG		ACG		AUG	M/start	AGG		G
C	CAC	H	CCC	P	CUC	L	CGC	R	C
	CAU		CCU		CUU		CGU		U
	CAA	Q	CCA		CUA		CGA		A
	CAG		CCG		CUG		CGG		G
U	UAC	Y	UCC	S	UUC	F	UGC	C	C
	UAU		UCU		UUU		UGU		U
	UAA	stop	UCA		UUA	L	UGA	stop	A
	UAG		UCG		UUG		UGG	W	G
G	GAC	D	GCC	A	GUC	V	GGC	G	C
	GAU		GCU		GUU		GGU		U
	GAA	E	GCA		GUA		GGA		A
	GAG		GCG		GUG		GGG		G



DNA {A,C,T,G}

Transcription

RNA {A,C,U,G}

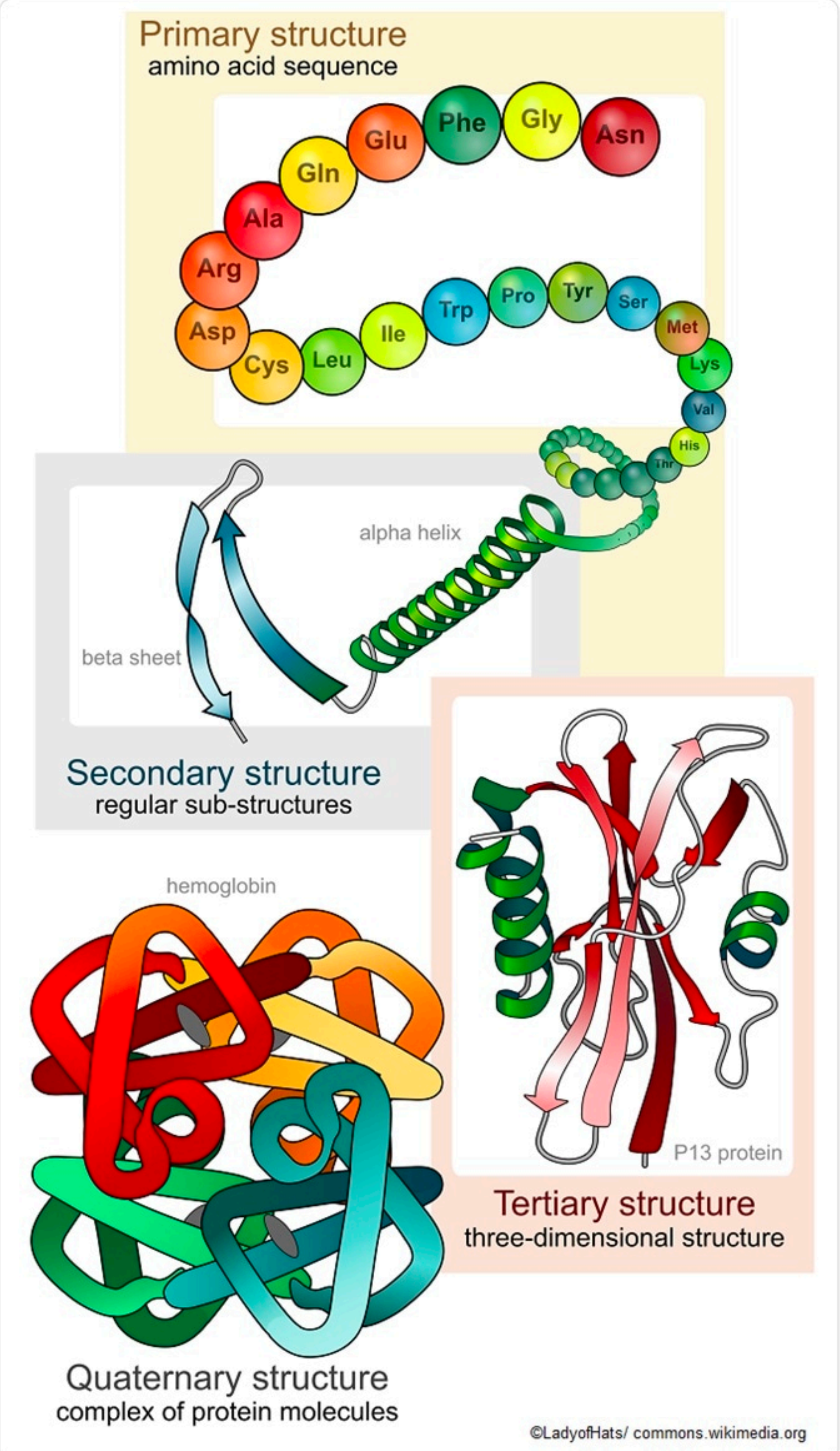
Translation

Proteins {A,C,D,E,F,G,H,I,K,L,M, N,P,Q,R,S,T,V,W,Y}

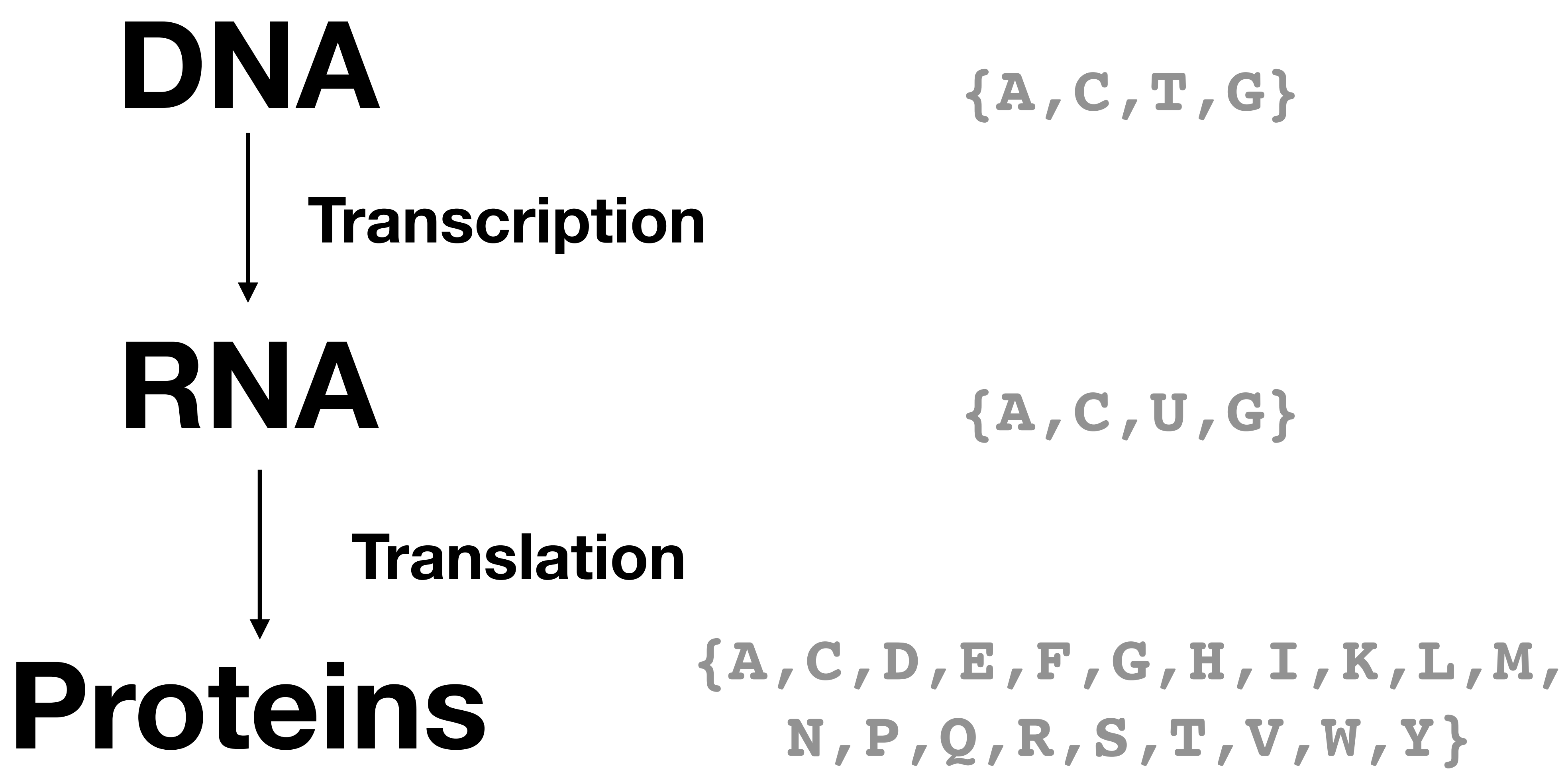
The Central Dogma

Proteins

- Do stuff in the cell, including help with translation and transcription



©LadyofHats/ commons.wikimedia.org



DNA {A,C,T,G}

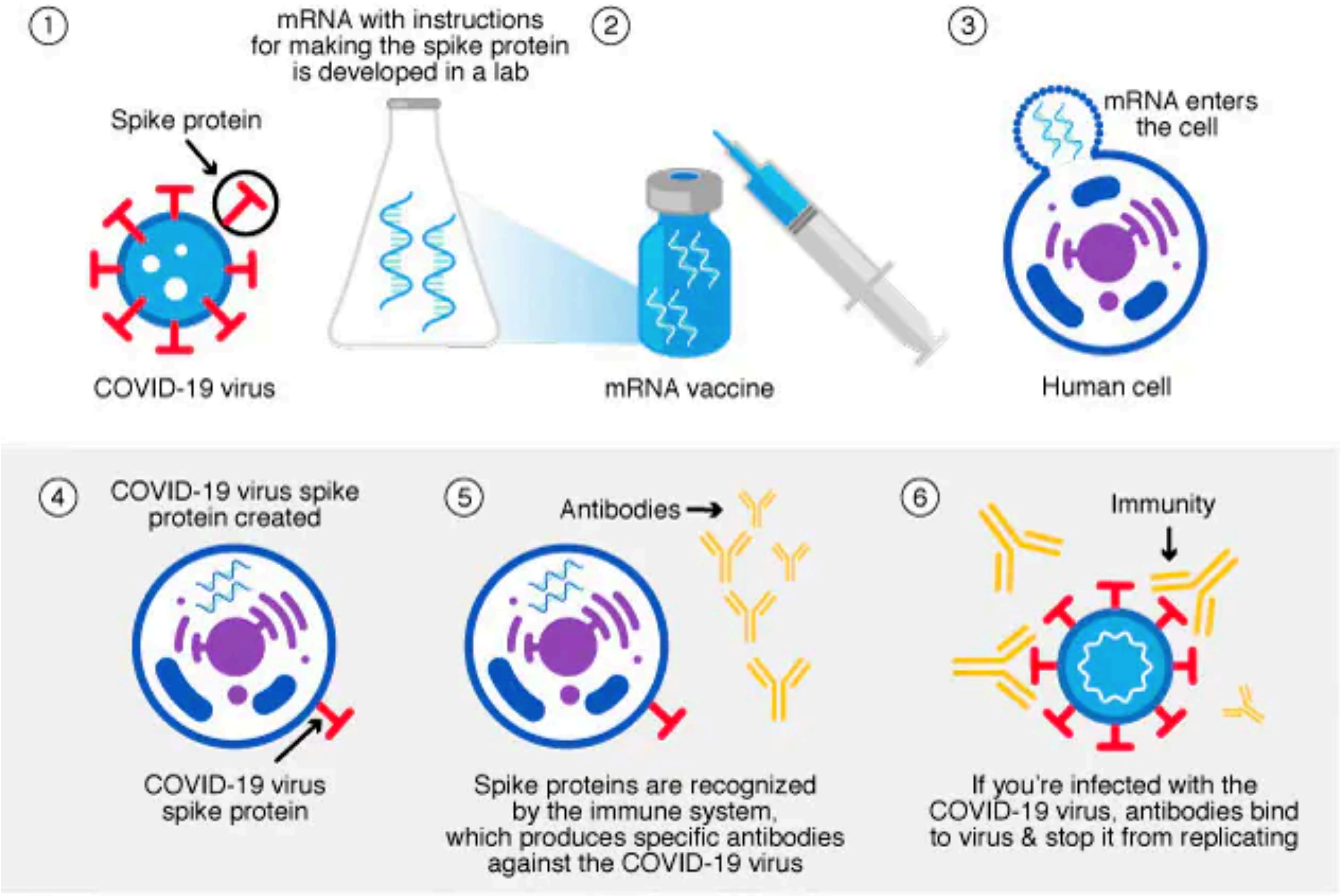
Transcription

RNA {A,C,U,G}

Translation

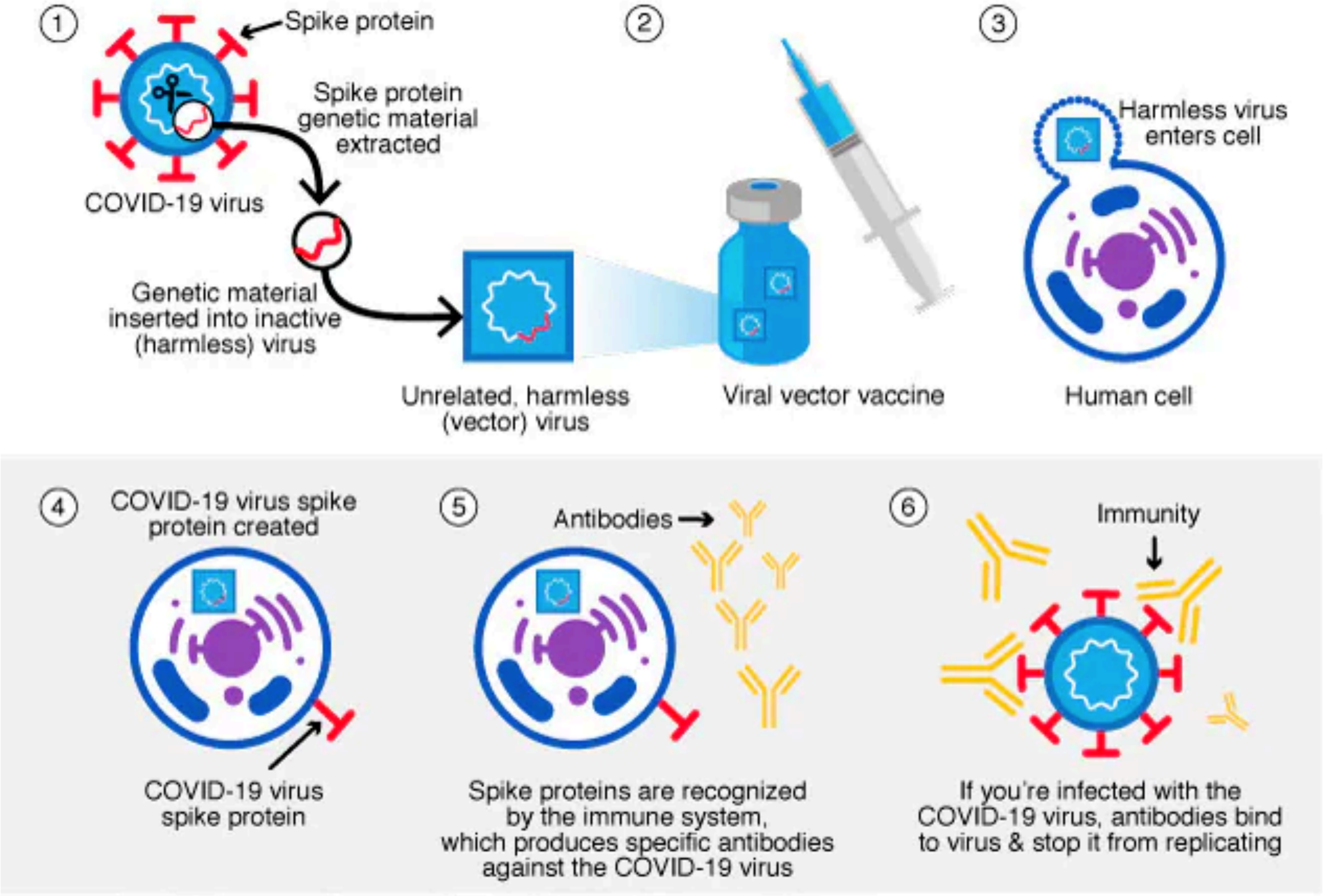
Proteins {A,C,D,E,F,G,H,I,K,L,M,
N,P,Q,R,S,T,V,W,Y}

Quick Diversion

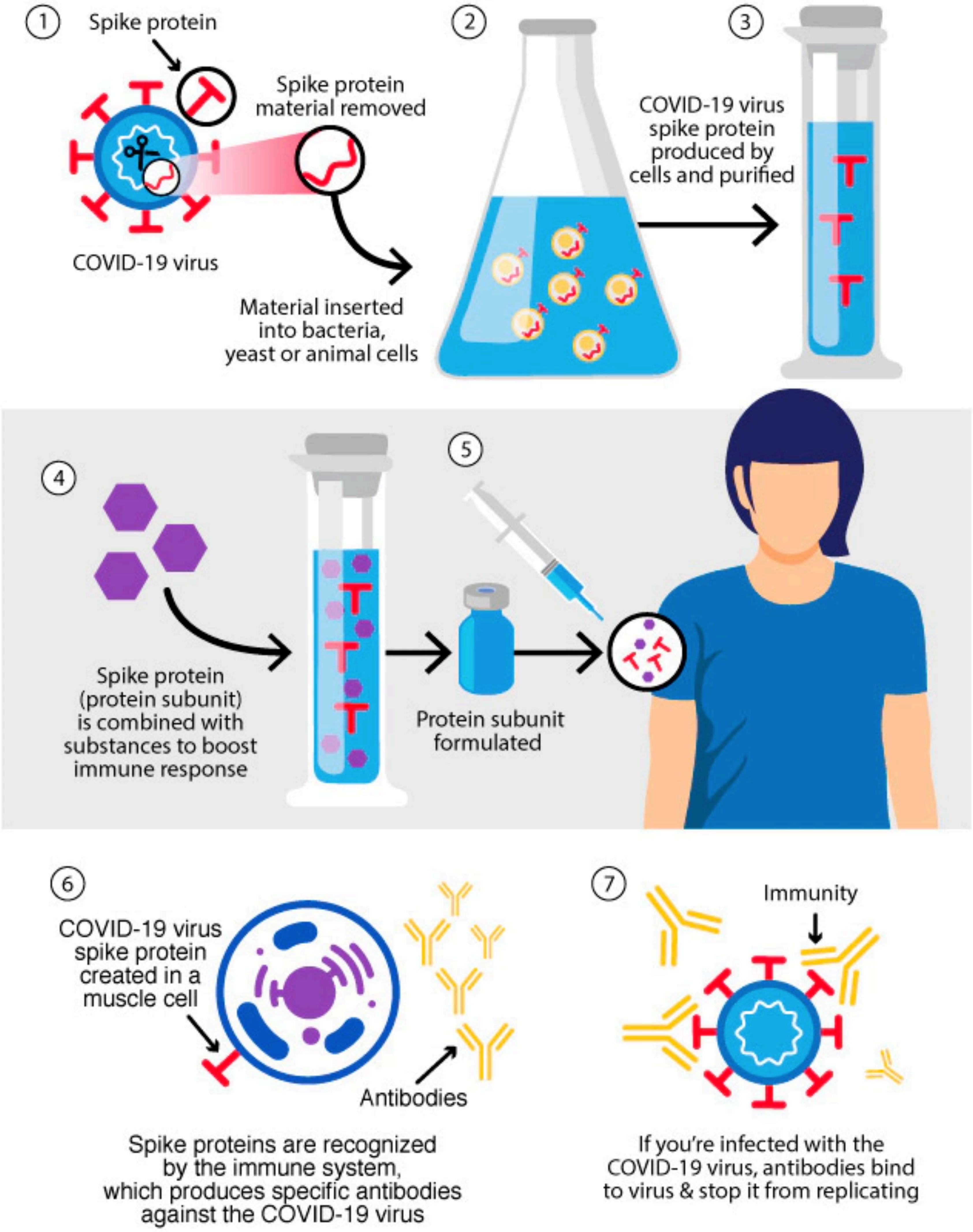


1
mRNA with instructions for making the spike protein is developed in a lab
2
3
Spike protein
COVID-19 virus
mRNA vaccine
mRNA enters the cell
Human cell
4
COVID-19 virus spike protein created
COVID-19 virus spike protein
5
Antibodies
Spike proteins are recognized by the immune system, which produces specific antibodies against the COVID-19 virus
6
Immunity
If you're infected with the COVID-19 virus, antibodies bind to virus & stop it from replicating

Quick Diversion


1
Spike protein
Spike protein genetic material extracted
COVID-19 virus
Genetic material inserted into inactive (harmless) virus
Unrelated, harmless (vector) virus
2
Viral vector vaccine
3
Harmless virus enters cell
Human cell
4
COVID-19 virus spike protein created
COVID-19 virus spike protein
5
Antibodies
Spike proteins are recognized by the immune system, which produces specific antibodies against the COVID-19 virus
6
Immunity
If you're infected with the COVID-19 virus, antibodies bind to virus & stop it from replicating

Quick Diversion

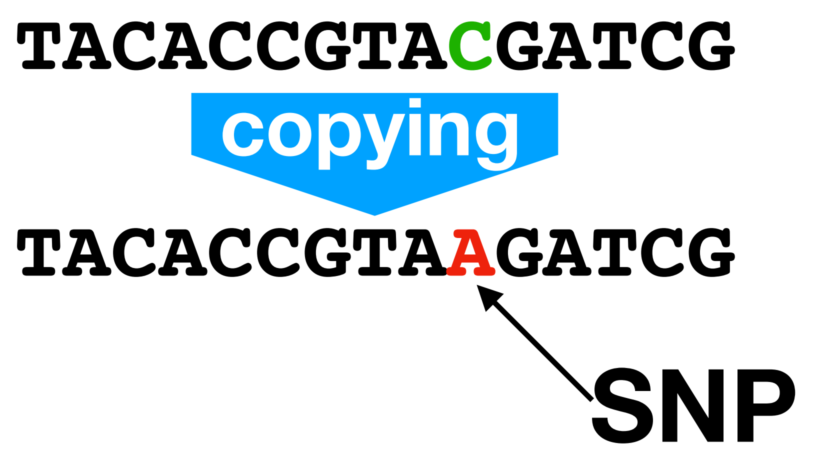


1
Spike protein
Spike protein material removed
COVID-19 virus
Material inserted into bacteria, yeast or animal cells
2
3
COVID-19 virus spike protein produced by cells and purified
4
Spike protein (protein subunit) is combined with substances to boost immune response
5
Protein subunit formulated
6
COVID-19 virus spike protein created in a muscle cell
Antibodies
Spike proteins are recognized by the immune system, which produces specific antibodies against the COVID-19 virus
7
Immunity
If you're infected with the COVID-19 virus, antibodies bind to virus & stop it from replicating

Genetic Variants

When copying a genome "errors" may occur, these changes are what make people different

- 99.99% of our genomes are identical
- **Single Nucleotide Polymorphism (SNP)** -- a change at a single base
- **Structural Variants (SV)** -- large scale changes

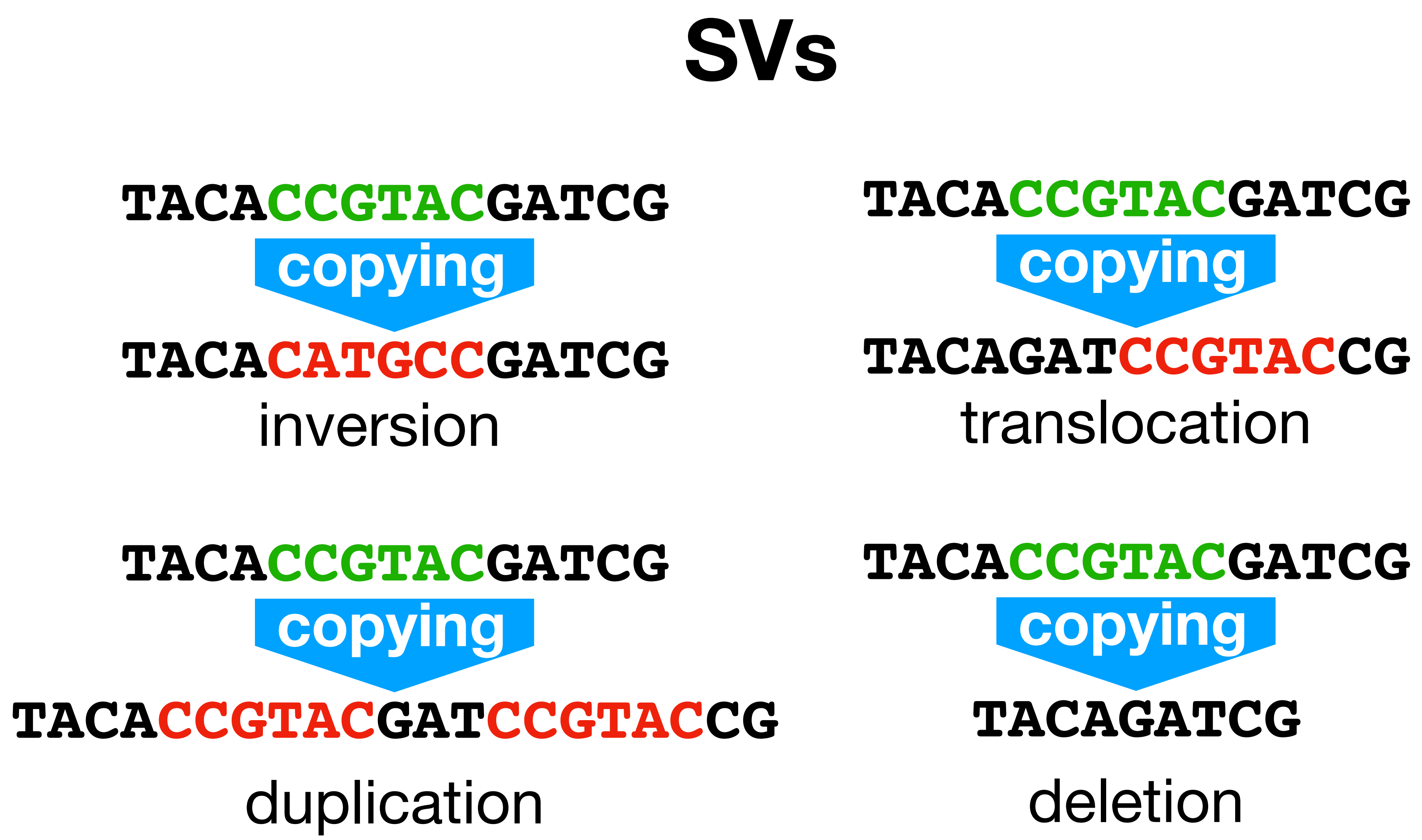


TACACCGTACGATCG

copying

TACACCGTAAGATCG

SNP



SVs

TACACCGTACGATCG

copying

TACACATGCCGATCG

inversion

TACACCGTACGATCG

copying

TACAGATCCGTACCG

translocation

TACACCGTACGATCG

copying

TACACCGTACGATCCGTACCG

duplication

TACACCGTACGATCG

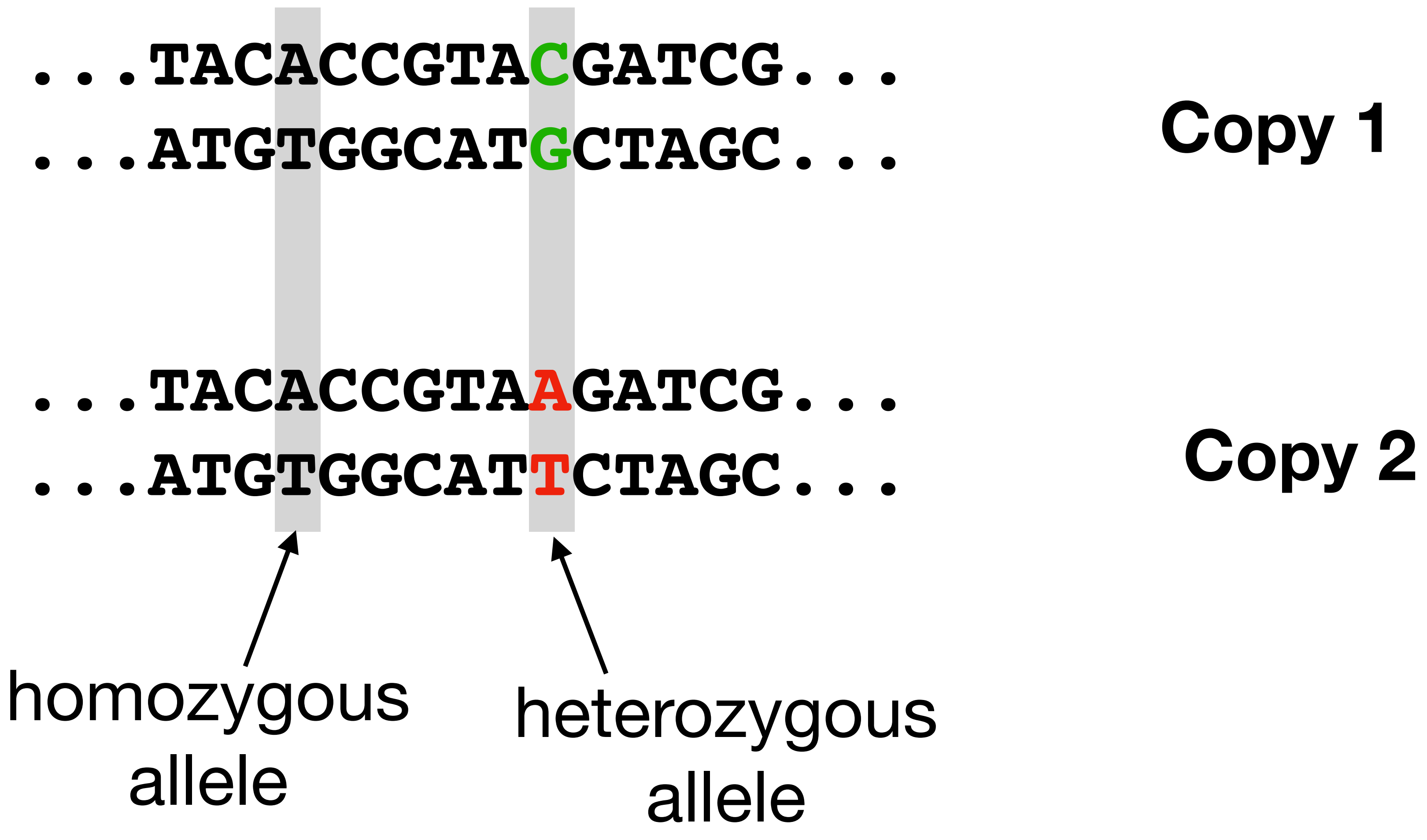
copying

TACAGATCG

deletion

Genetic Variants

- **Deleterious Mutations** -- changes that are harmful (lethal) to a cell
- **Germline Mutations** -- changes passed to offspring
- **Somatic Mutations** -- those not passed down
- **Heterozygous** -- different beween copies
- **Homozygous** -- same on both copies
- **Allele** -- specific position on a chromosome



...TACACCGTACGATCG...
...ATGTGGCATGCTAGC...
Copy 1

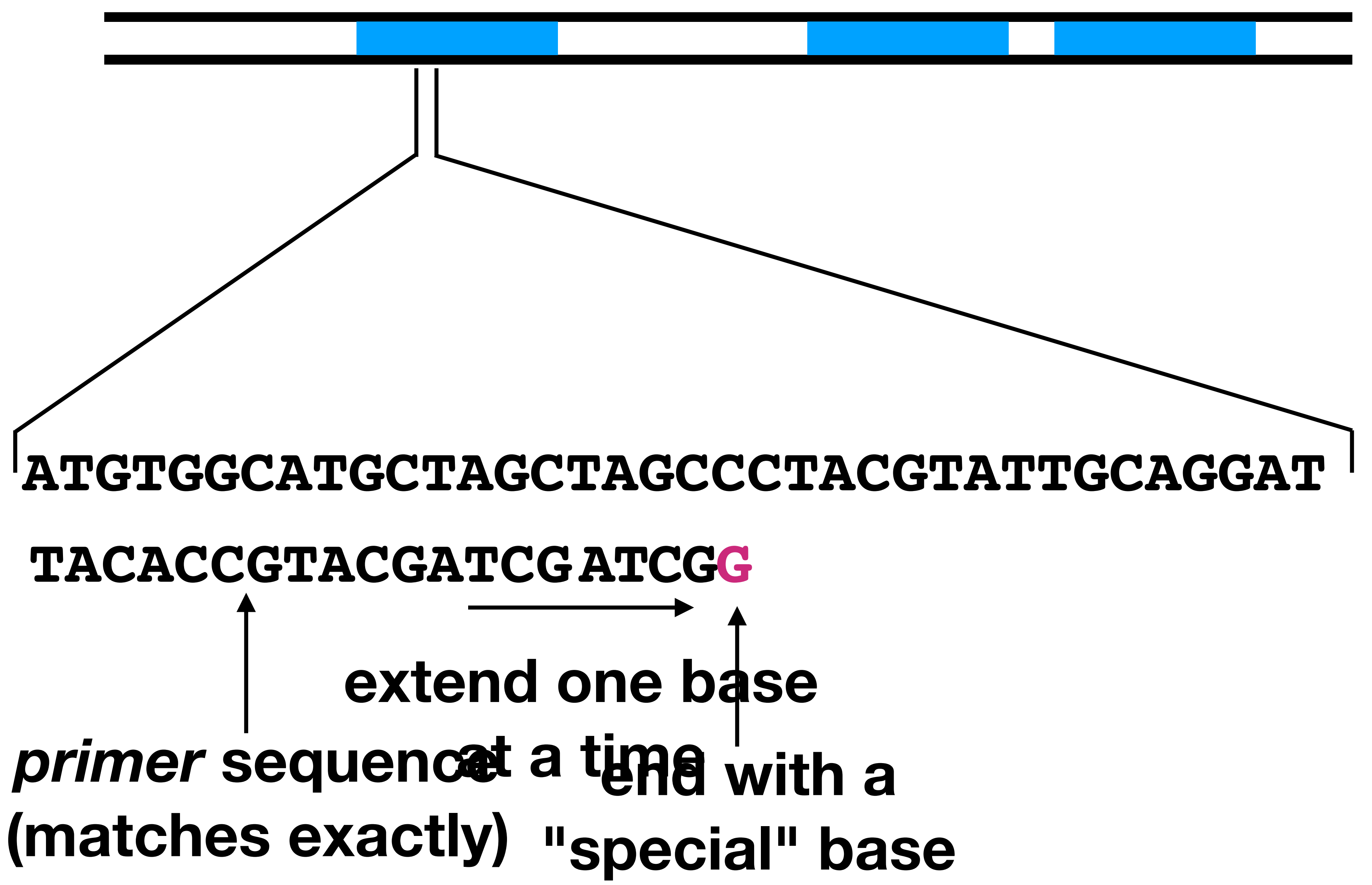
...TACACCGTAAGATCG...
...ATGTGGCATTCTAGC...
Copy 2

homozygous allele

heterozygous allele

Sanger Sequencing

The basis of all modern sequencing.



ATGTGGCATGCTAGCTAGCCCTACGTATTGCAGGAT

TACACCGTACGATCGATCGG

extend one base at a time

primer sequence (matches exactly)

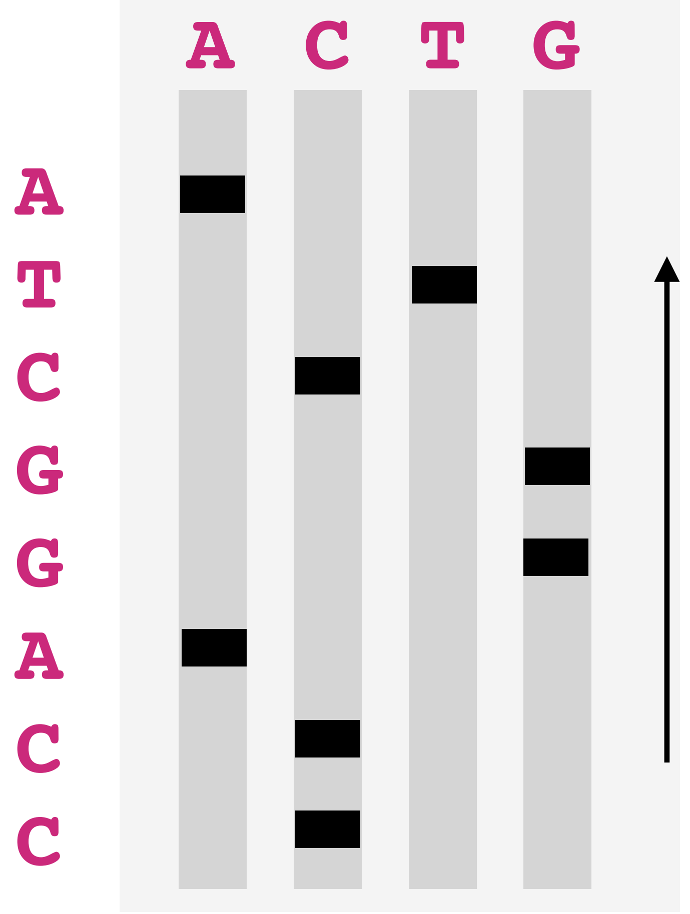
end with a "special" base

figure adapted from Gibson and Muse, 3rd Edition (2009)

Sanger Sequencing

The basis of all modern sequencing.

. . .
TACACCGTACGATCGATCGG
TACACCGTACGATCGATCG
TACACCGTACGATCGATC
TACACCGTACGATCGAT
TACACCGTACGATCGA



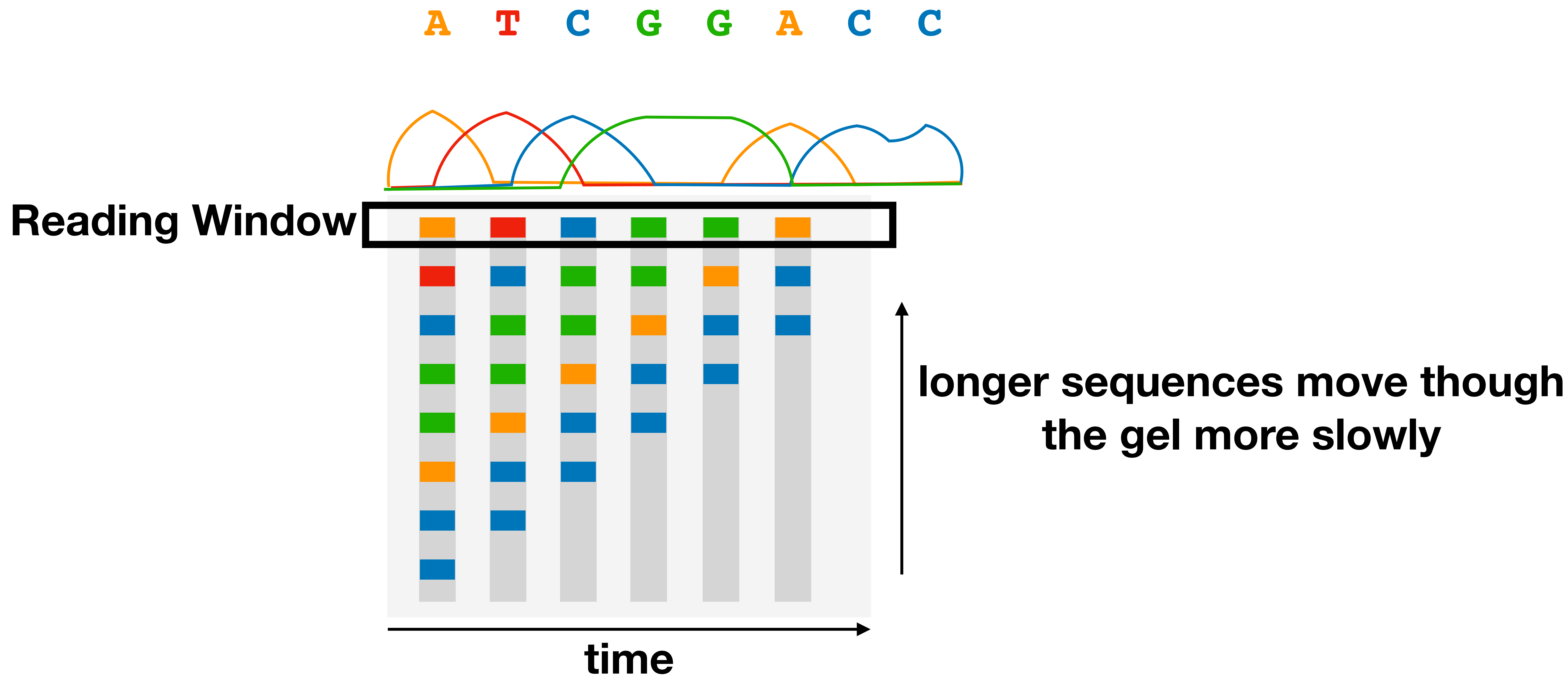
longer sequences move though
the gel more slowly

figure adapted from Gibson and Muse, 3rd Edition (2009)

Sanger Sequencing

The basis of all modern sequencing.

. . .
TACACCGTACGATCGATCGG
TACACCGTACGATCGATCG
TACACCGTACGATCGATC
TACACCGTACGATCGAT
TACACCGTACGATCGA



longer sequences move though the gel more slowly

figure adapted from Gibson and Muse, 3rd Edition (2009)

Second Generation Sequencing

Also called next generation sequencing

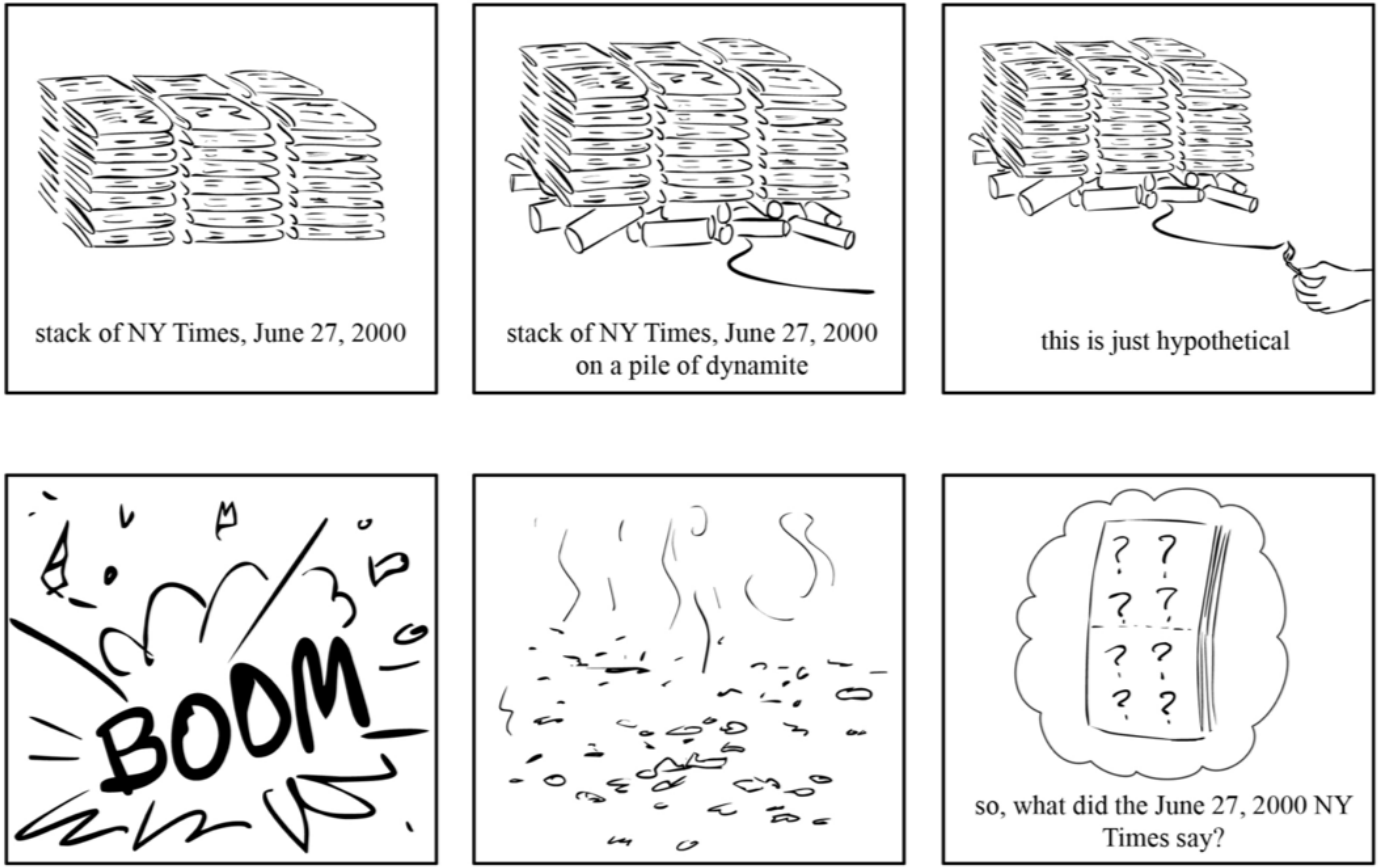
Based on the same principles, but at a much larger scale

Improvements were made in the amplification and reading with better microscopes

With this came shorter sequences

- Sanger could do >1,000 bases (characters) at once but all done by hand, so 10s of sequences, very accurate
- Illumina (current standard) ~250 base reads, 1,000,000s of sequences, some errors

Second Generation Sequencing

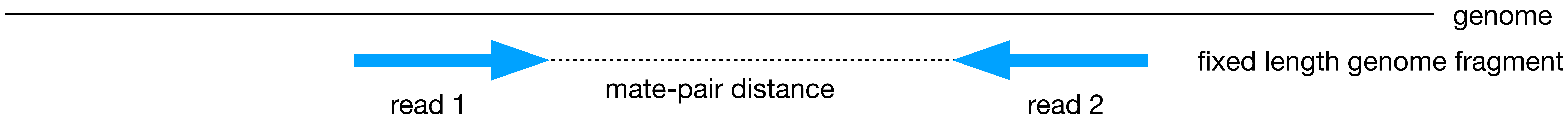

stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM
so, what did the June 27, 2000 NY Times say?

from Compeau and Pevzner 2018

Second Generation Sequencing

NextGen sequencing also introduced paired-end reads

- take a long piece of sequence (much longer than the read size, but predictable size)
- sequence both ends but keep them together
- gives two reads that you know are a certain distance from each other



Third Generation Sequencing

Recently Pacific Biosciences and Oxford Nanopore have introduced new technologies that:

- have long reads
- with high(er) error rates

Clip slide

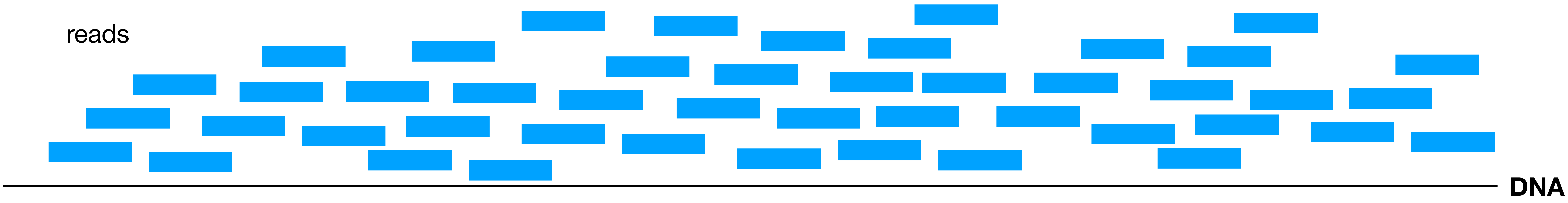
	Sanger	Next-Generation	Third-Generation
Launched	1977 Basic chemistry 1998 Modern form	2005 with significant improvements since	2010 with significant improvements since
Estimated Error Rate	0.001% - 1%	0.46% - 2.4%	11% - 14% (but decreasing)
Cost	🪙🪙🪙	🪙	🪙🪙
Throughput	🧬	🧬🧬🧬	🧬🧬
Currently Available Platforms	Applied Biosystems*	Illumina Ion Torrent* Qiagen (Europe) Complete Genomics (China)**	Pacific Biosciences Oxford Nanopore
Clinical Uses	Many (but dwindling)	Many (and growing)	Niche uses (today)

*Part of Thermo Fisher **Part of BGI

Error rate figures courtesy of G. Corey Shan from https://shanguangyu.com/articles/comparison-of-sanger-NGS-and-SMS/

Adapted from slide by Erica Ramos via SlideShare https://www.slideshare.net/EricaRamos14/acmg-2017-the-data-behind-the-results-bioinformatics-for-clinicians/7

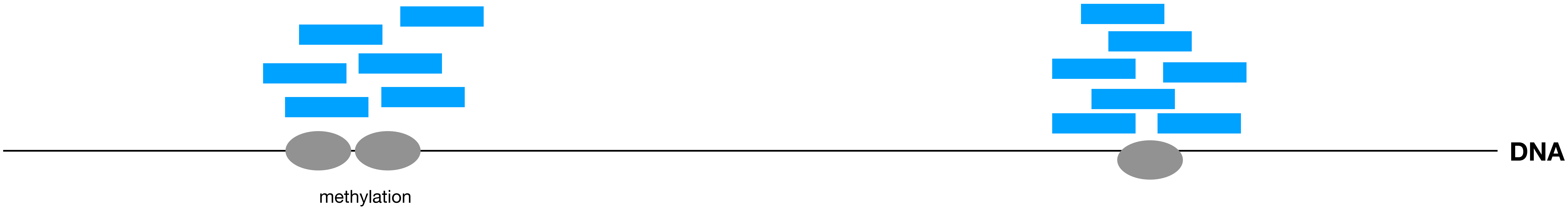
Sequencing Applications



whole genome sequencing

adapted from figure 1.2 in Mäkinen, *et al.* 2015

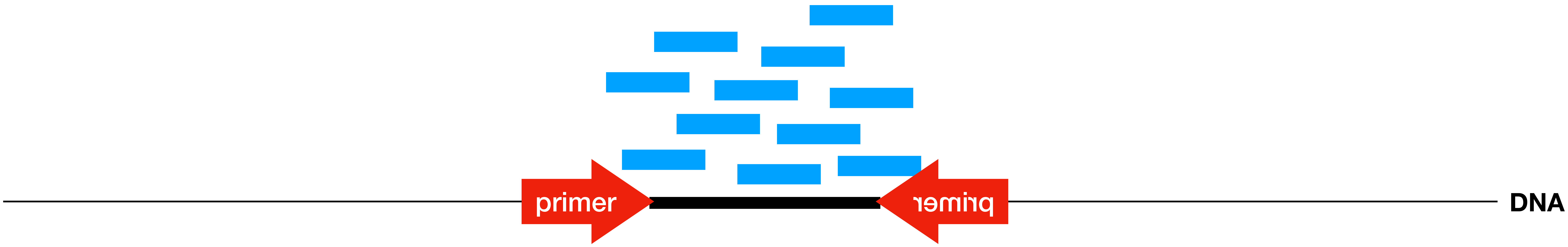
Sequencing Applications



bisulphite sequencing

adapted from figure 1.2 in Mäkinen, *et al.* 2015

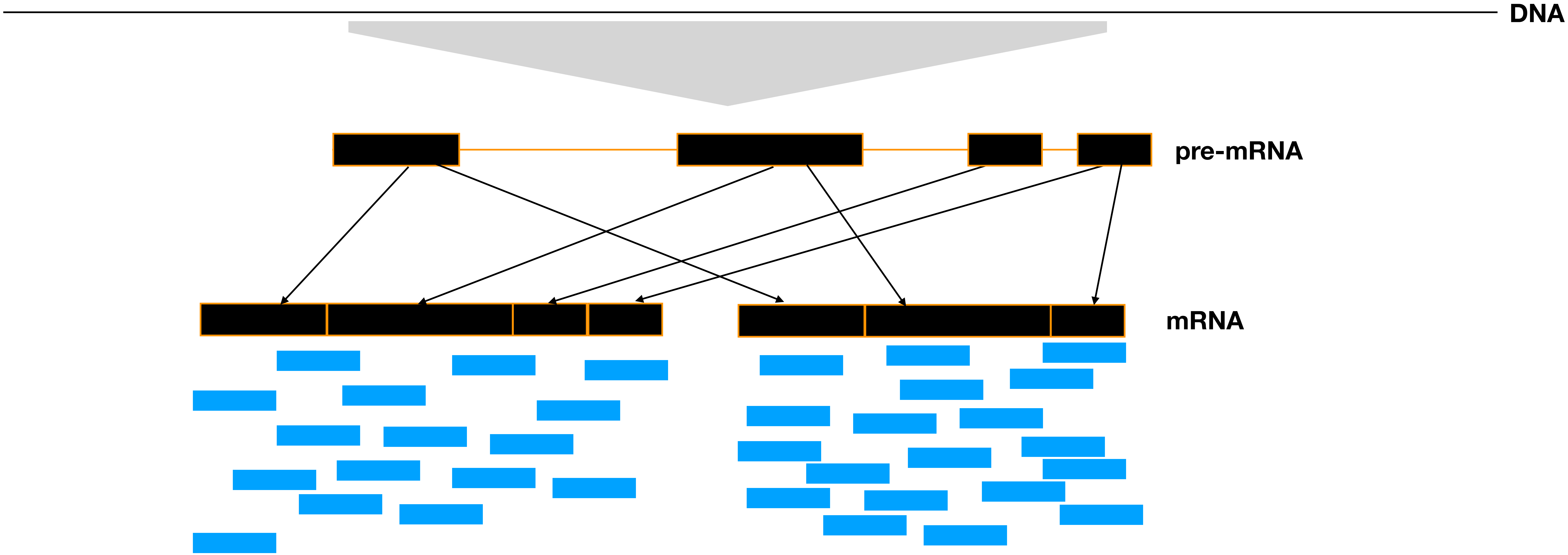
Sequencing Applications



targeted sequencing

adapted from figure 1.2 in Mäkinen, *et al.* 2015

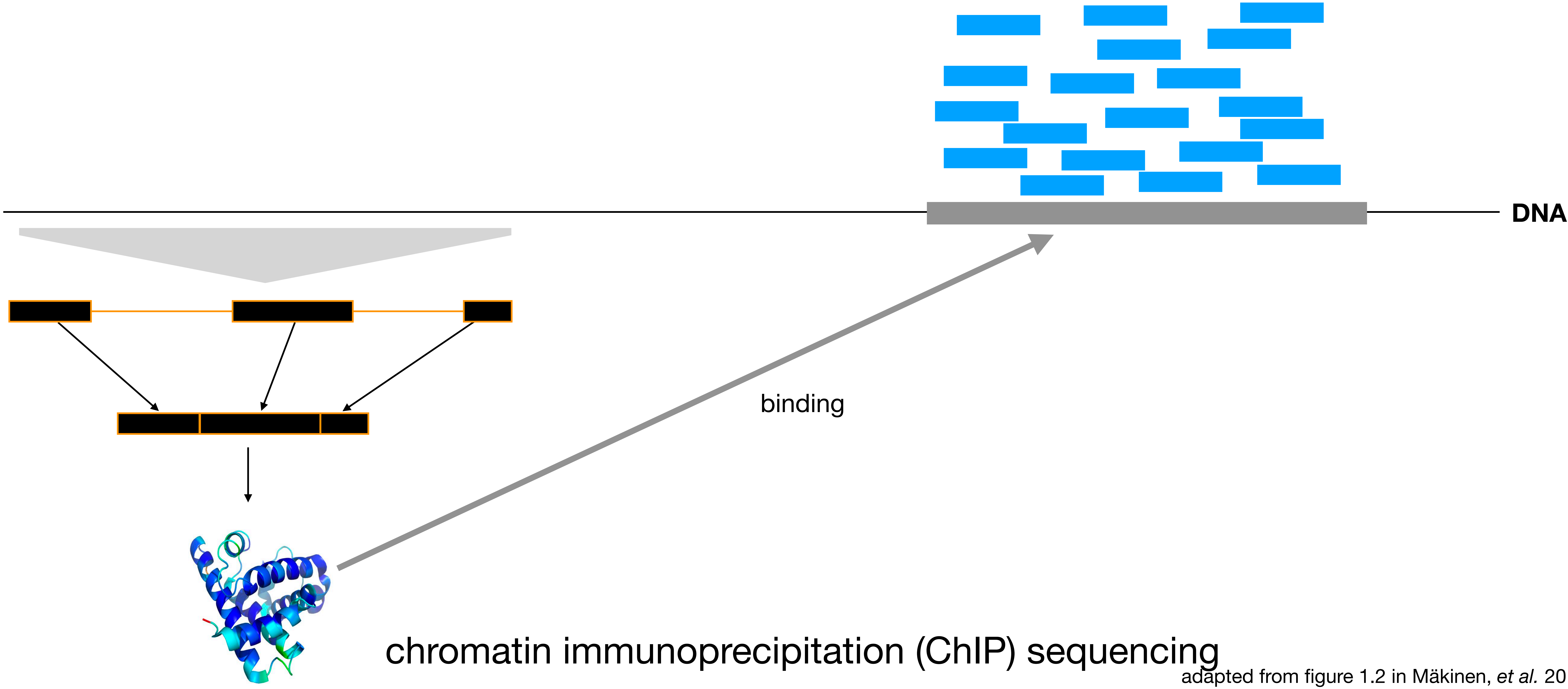
Sequencing Applications



RNA sequencing

adapted from figure 1.2 in Mäkinen, *et al.* 2015

Sequencing Applications



chromatin immunoprecipitation (ChIP) sequencing

adapted from figure 1.2 in Mäkinen, *et al.* 2015

History

1866 -- Gregor Mendel discovers genetics using pea plants

1869 -- DNA was discovered

1944 -- Avert and McCarty show DNA carried genetic information

1953 -- Watson and Crick discovered the 3D structure of DNA

1961 -- Nirenberg maps DNA to proteins

1968 -- Discovery of restriction enzymes

1970s -- Development of the first sequencing techniques

1985 -- Development of PCR

1986 -- Discovery of RNA splicing

1980-1990 -- Complete sequencing of genomes of small organisms

1990 -- Launch of the Human Genome Project

1998 -- Discovery of post-transciption RNA interference

2000 -- Announcement of the draft human genome

Major Ongoing Projects

ENCODE (The **Enc**yclopedia **o**f **D**NA **E**lements)

- Effort to identify all functions elements in the human genome

1000 Genomes Project

- Large sample size will hopefully show all (most) of the variation within the population

UK BioBank

- 500,000 UK genomes in great details

SRA (Sequence Read Archive)

- Public repository of all types of sequencing data

GWAS Catalog (Genome Wide Association Studies)

- Multiple studies for many possible purposes (i.e. cancer, disorders, etc.)