INSTITUTE OF PLANT SCIENCES

Sant'Anna
School of Advanced Studies – Pisa

Advanced Genomics

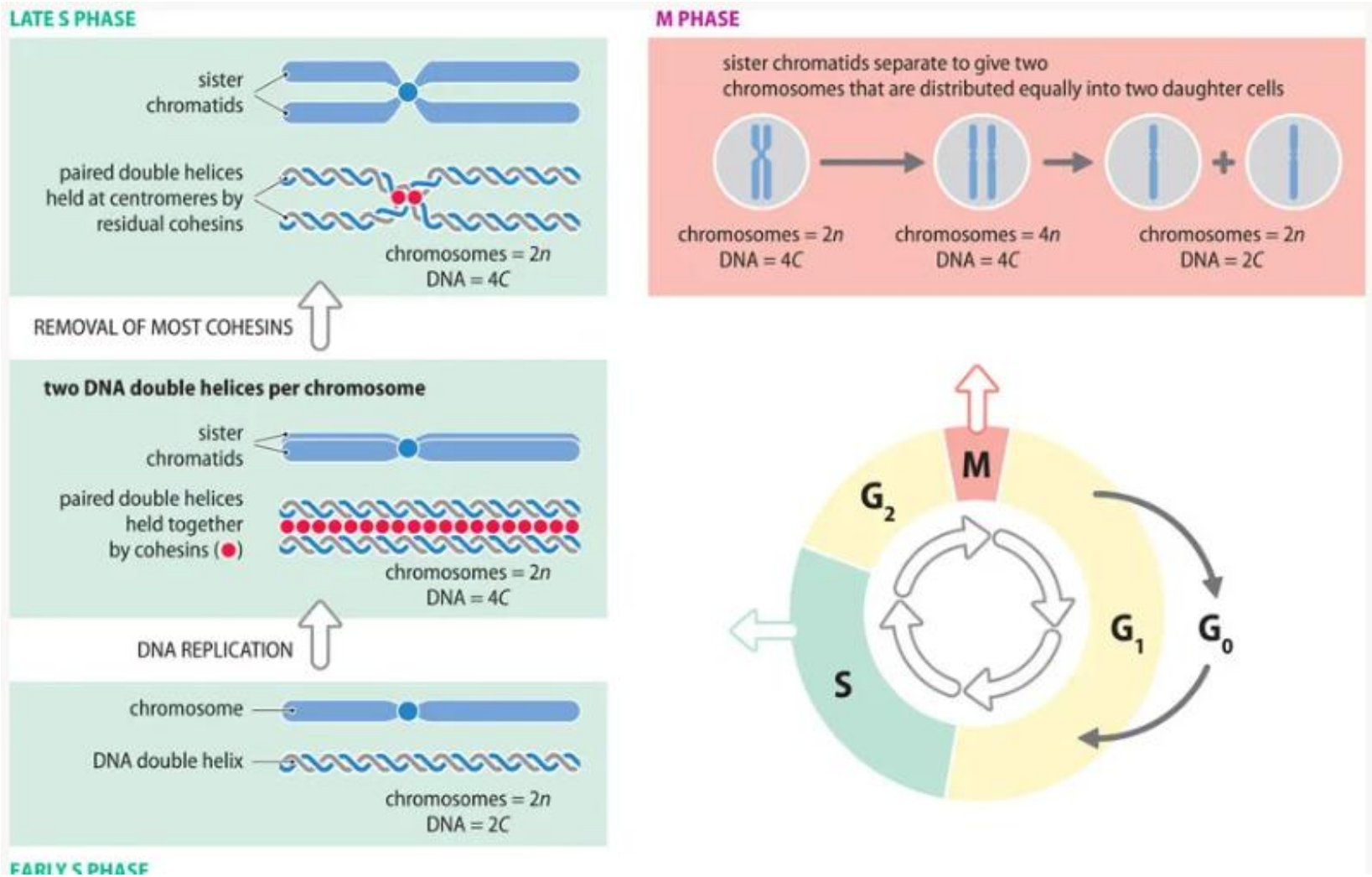
Genome contents

Genomic content across the cell life

- As a rule of thumb, all cells of an organism have the same genomic content; it is the activation/inactivation of genomic regions that makes a difference
- Another thing that may change is polidy: that is, the number of copies of the basic chromosome set. Diploid somatic cells have 2n chromosomes. Germ cells (gametes) have n. Some cells can be nulliploids (e.g. red blood cells)
- It is very important to mantain the genomic content as intended
- Let's not forget plastid DNA; in humans, from 100s to 1000s copies. Male sperm cells have about 100. Egg cells have about 250K

The cell cycle consists of four major phases and has a lot to do with DNA content

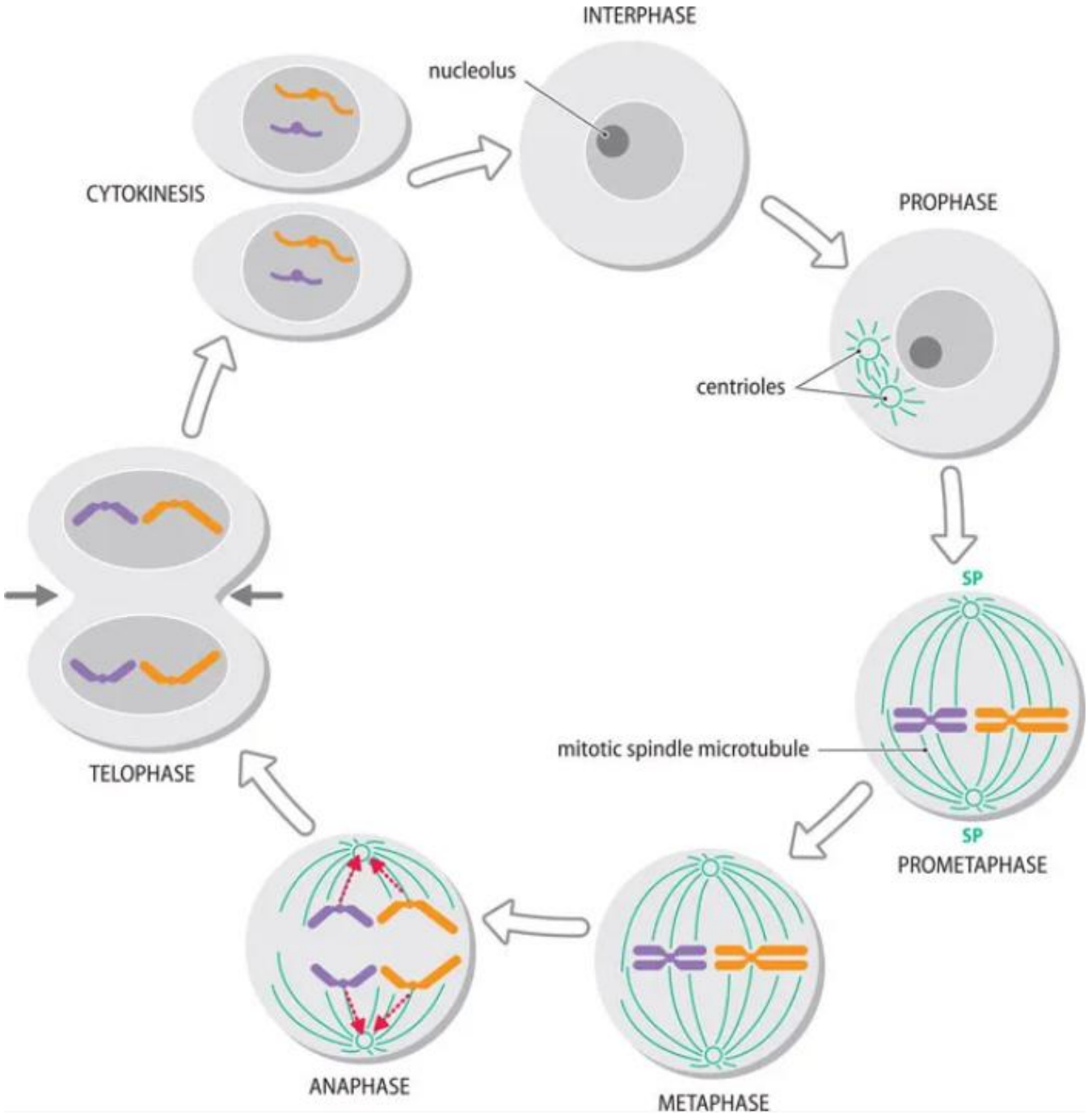
- S phase (synthesis of DNA)
- M phase (cell division)
- G1 phase (gap between M phase and S phase)
- G2 phase (gap between S phase and M phase)



If the cell cycle is 24h, M takes about 1h. It is in G1 and G2 that most proteins are produced

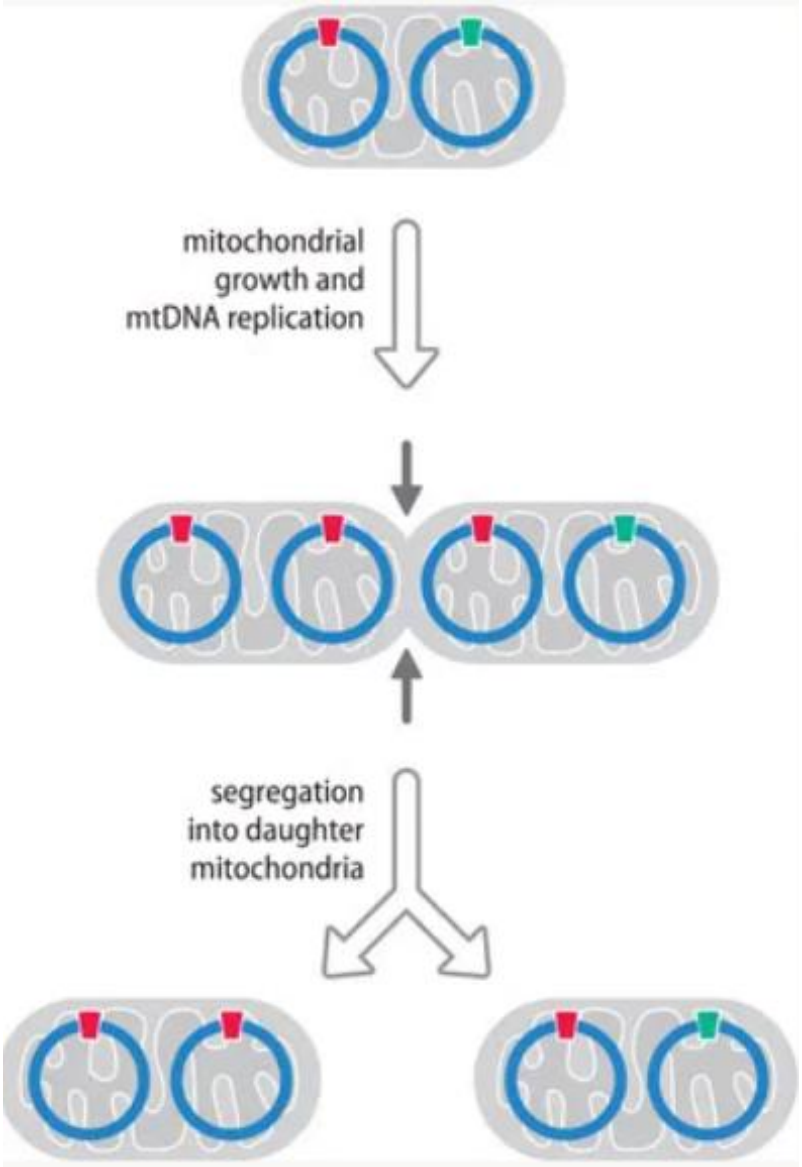
Phase M includes nuclear division (mitosis) and cell division (cytokinesis)

- Scope: multiplication of cells
- Number of divisions: one
- Result: two daughter diploid cells
- Variation: none
- Where: everywhere



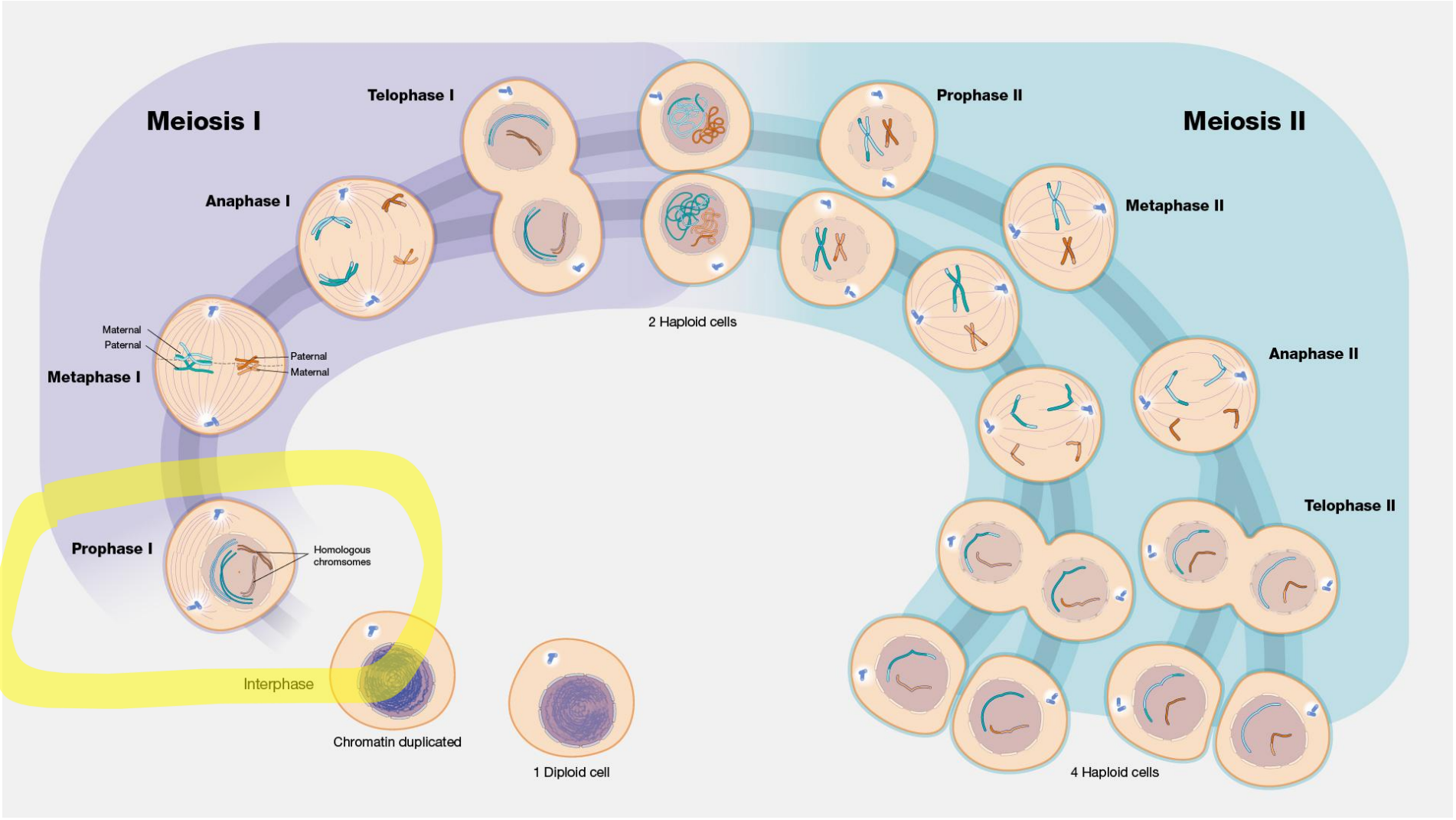
mtDNA and cPDNA replicate too, before segregating in daughter cells

- Each mt and ch can contain tens of genome copies
- However their replication is not synchronized with the cell cycle, nor it is tightly controlled
- Variants in mtDNA/cpDNA may arise and be unevenly spread, resulting in heteroplasmy
- Mitochondria are then segregated in daughter cells stochastically



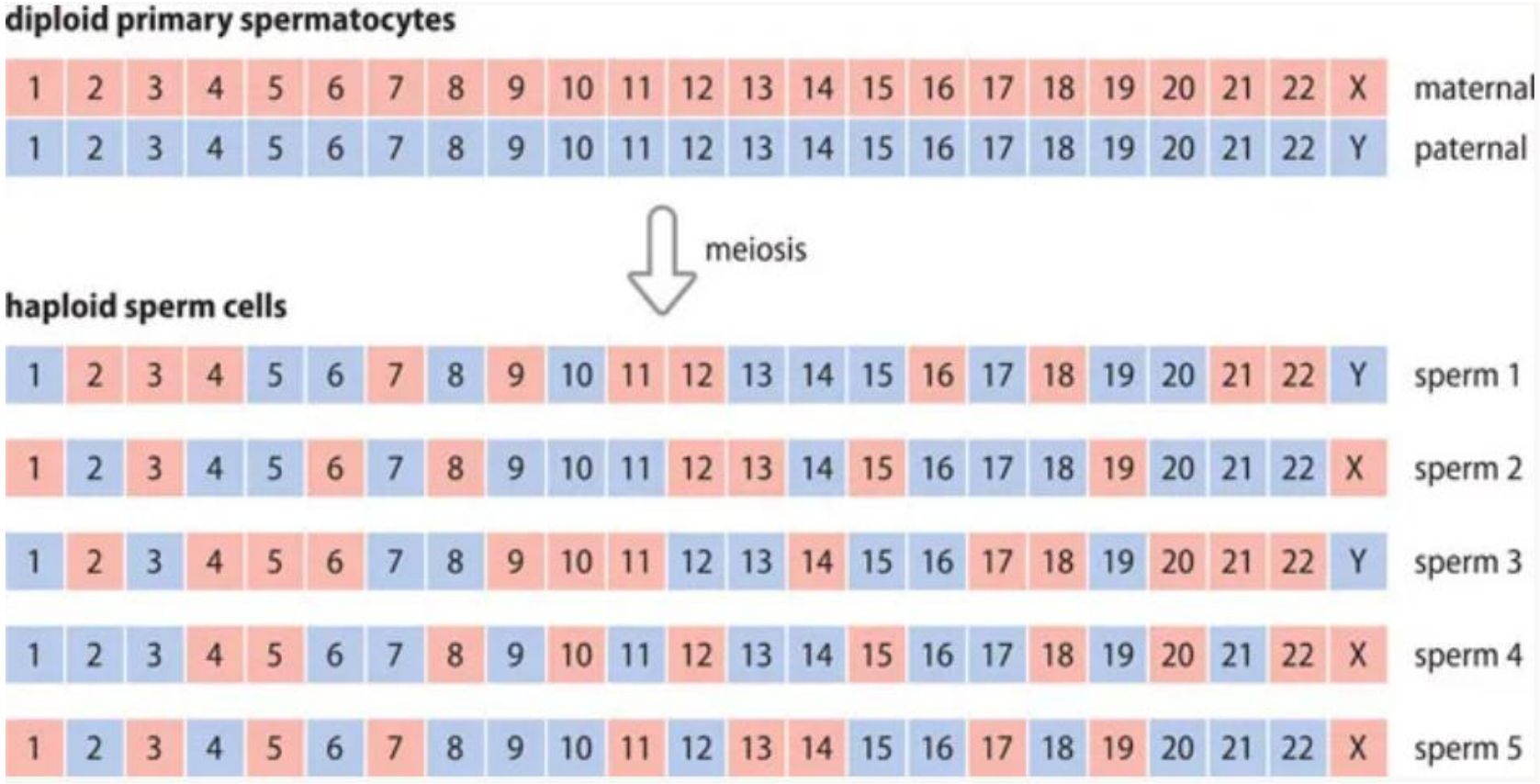
Meiosis is the reductive division that carries (diploid) germ cells to (haploid) gametes. Meiosis I is to divide and reduce; Meiosis II is basically a mitosis but ending with haploid cells

- Scope: production of gametes
- Number of divisions: two
- Result: four daughter haploid cells
- Variation: chr assortment and Xover
- Where: germline



Creating variation

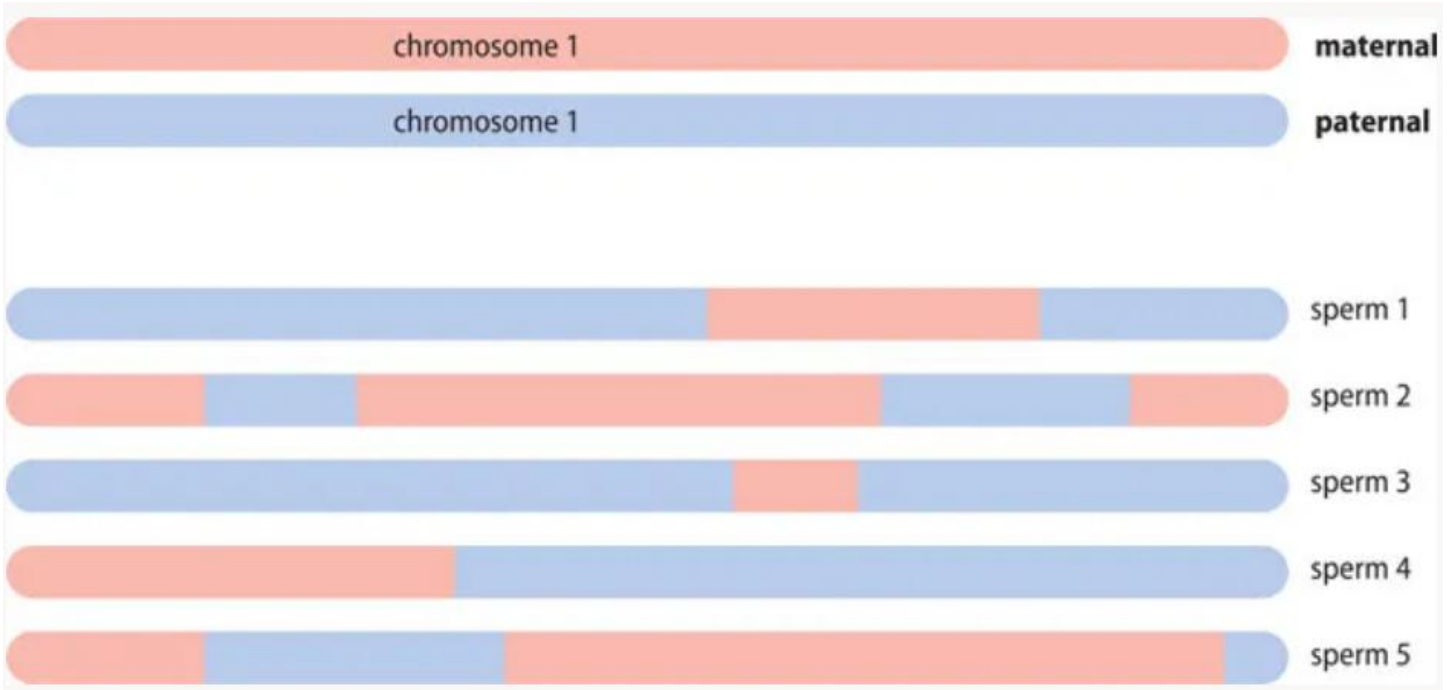
The whole point about sex is the recombination of genetic diversity from one generation to the next



Even not considering recombination, there's 2^{23} (8,388,608) possible combinations of chromosomes from one generation to the next in humans

Creating variation - 2

- In humans, the prophase of meiosis I begins during fetal life
- Paternal and maternal chromatids within each bivalent pair exchange segments of DNA at randomly positioned but matching locations
- This is recombination (or crossover): it involves physical breakage of the DNA in one paternal and one maternal chromatid, and the subsequent joining of maternal and paternal fragments



Genome content and size

- One would expect that the most complex organisms require more complex information to be built the way they are, and hence more DNA content
- The complexity, or size of the genome, can be described in different ways
  - C value, the mass of DNA in pg. 1C is the mass of a haploid nucleus
  - Basepairs (bp), the number of paired nucleotides in a haploid genome

Table 1. Conversion of units commonly used to measure genome sizes.

Unit	Conversion factor		
	Picograms	Daltons	Base pairs
Picogram	1	6.02×10^{11}	0.98×10^{9}
Dalton	1.66×10^{-12}	1	1.62×10^{-3}
Base pair	1.02×10^{-9}	618	1

- Genome size varies wildly across and within taxa
- Some taxa have higher variability
- There seems to be a discrepancy between size of genome and complexity (even among comparable taxa)
- «C-vale paradox»

Search | Results

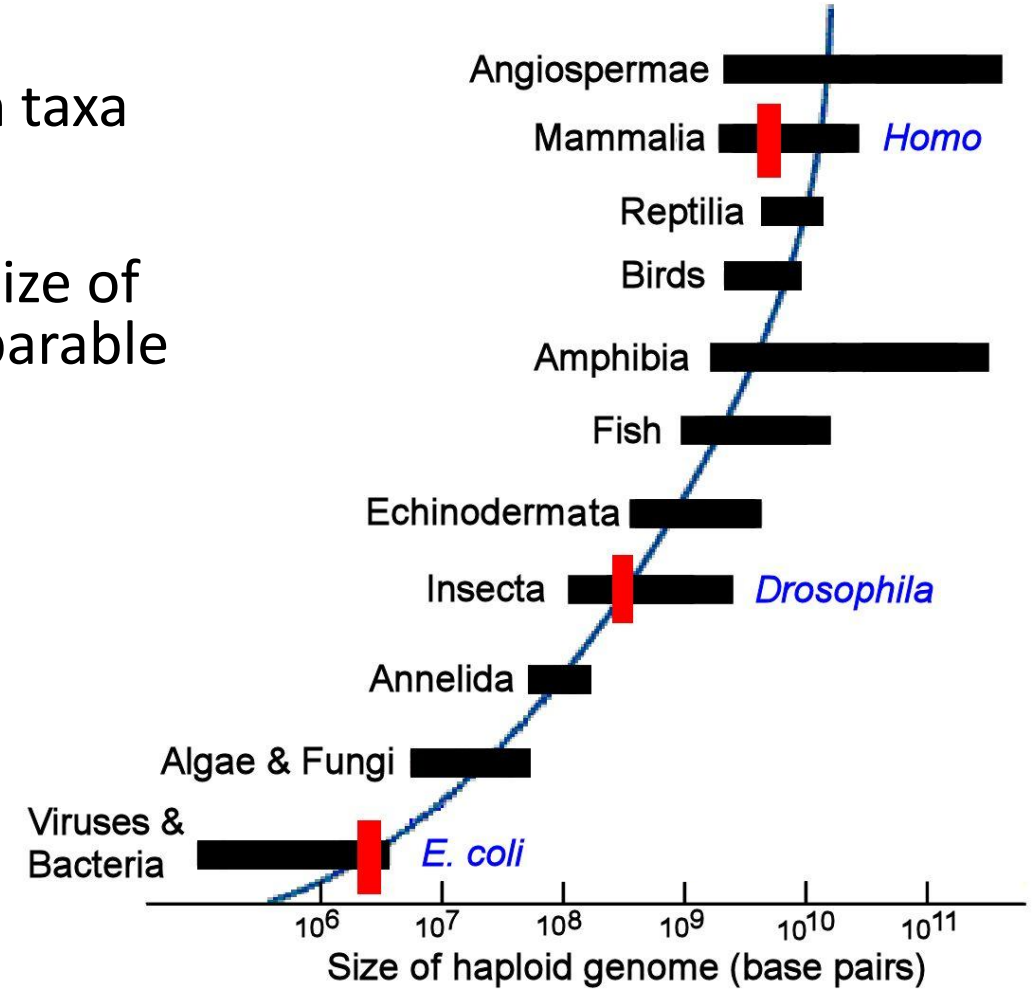
Search Results (showing results 1 to 100 of 421)

1 2 3 4 5

Summary Statistics	Mean	Min	Max	Std Dev
1C (pg)	18.35	2.25	36.00	7.31

Genus	Species	Subspecies	DNA Amount 1C (pg)	Original Reference
Gnetum	ula		2.25	Ohri and Khoshoo,1986
Gnetum	gnemon		3.87	Leitch et al.,2001
Gnetum	costatum		3.98	Leitch et al.,2001
Microcachrys	tetragona		4.15	Zonneveld,2012
Microstrobos (= Pherosphaera)	niphopheles		4.20	Zonneveld,2012
Microstrobos (= Pherosphaera)	fitzgeraldii		4.30	Zonneveld,2012
Lepidothamnus	fonkii		4.75	Zonneveld,2012
Saxegothaea	conspicua		5.10	Zonneveld,2012

https://cvalues.science.kew.org/



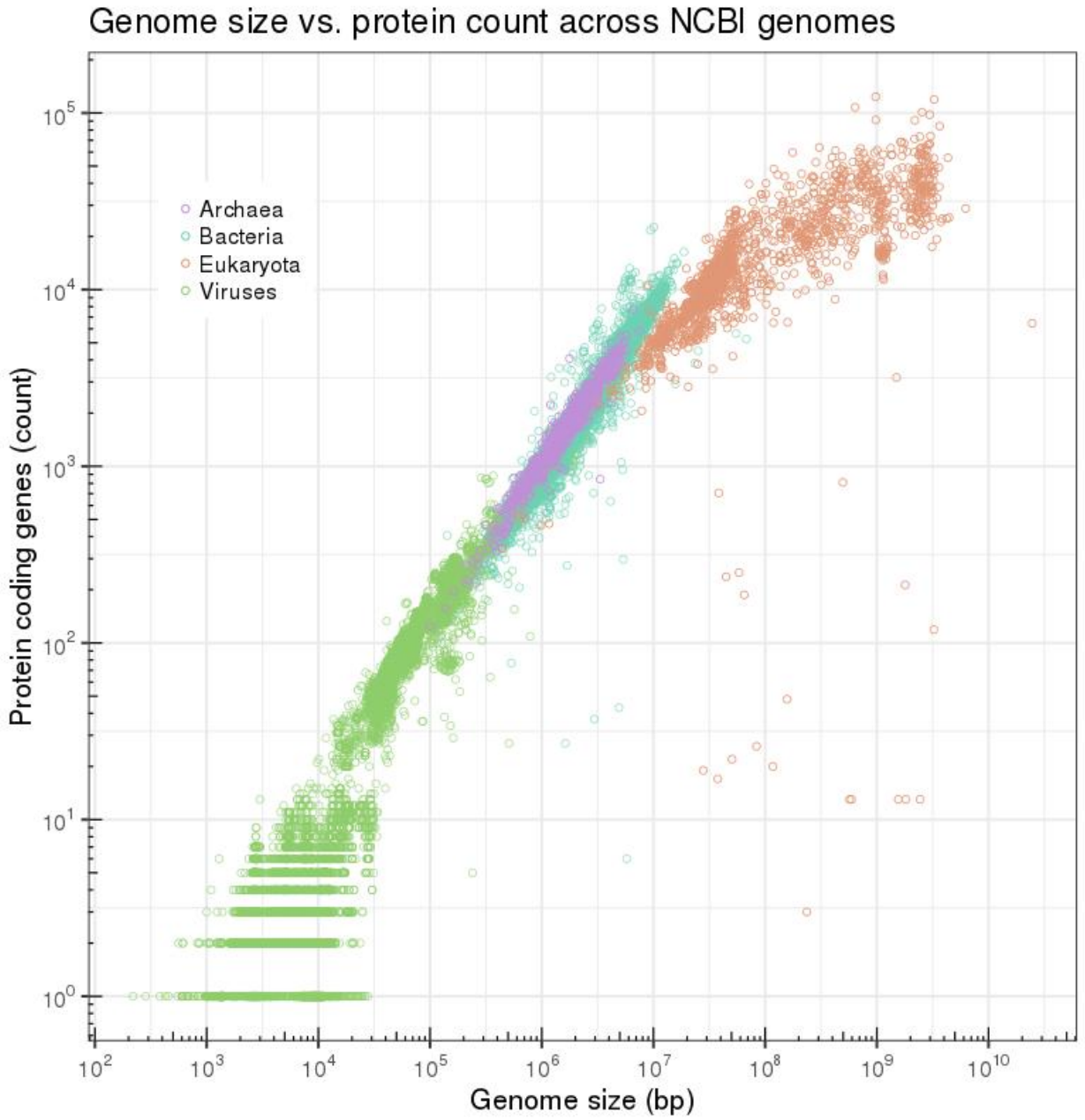
Human DNA has 3 billion base pairs. Each pair can hold one of four values. So, every 2 base pairs would hold 16 values, which is equal to a byte (8 bits) of computer information. Thus, it could hold about 1.5 billion bytes (1.5GB of information).

Is it gene number driving genome size variation?

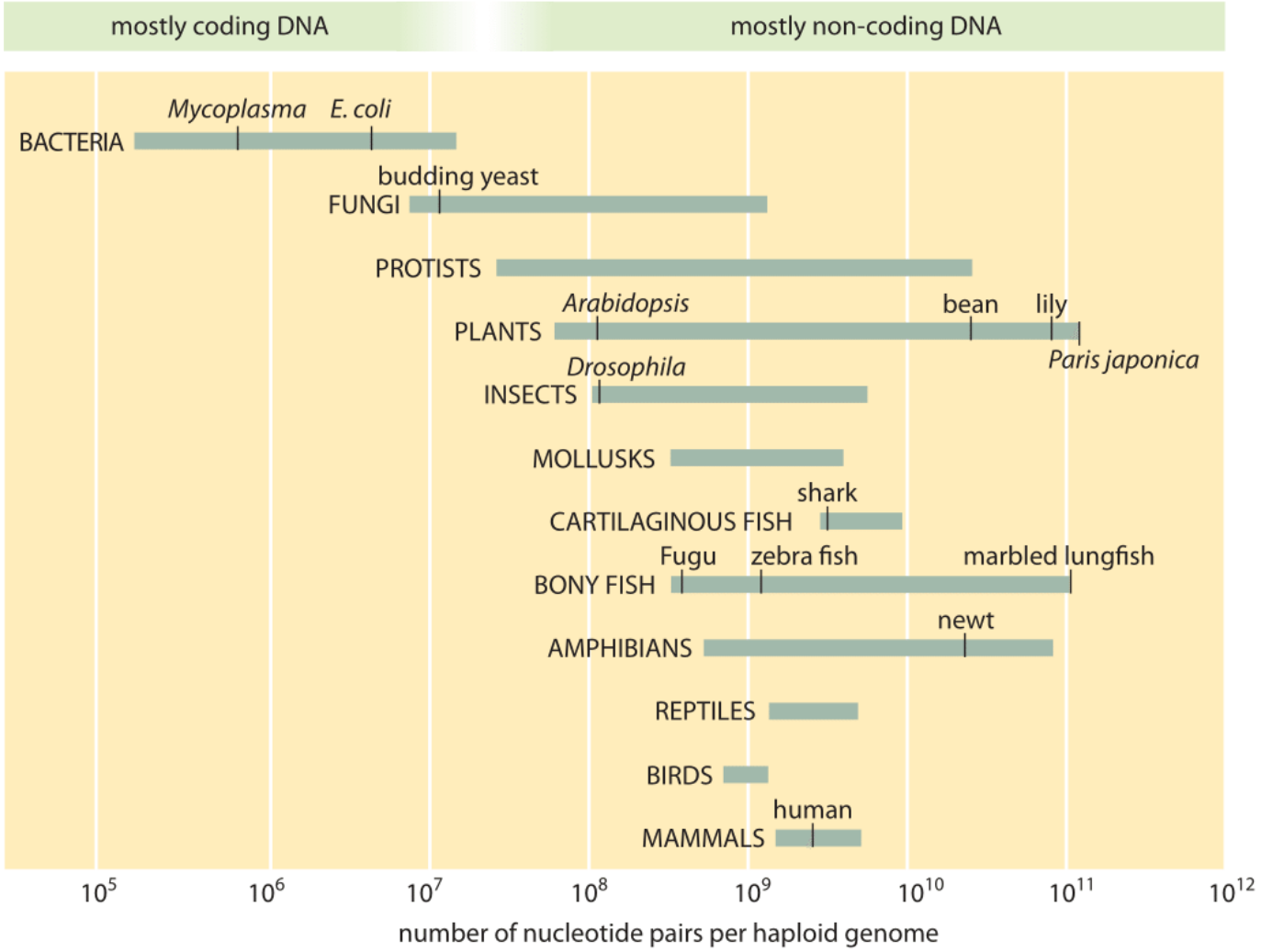
- Initially, it was believed that *H. sapiens* had 80K to 100K protein coding genes, based on the assumption that each protein derived from a different sequence (we now know this is not true)
- The current annotation of the human genome contains 20,442 protein-coding genes and 23,982 genes for noncoding RNAs

	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)
viruses	HIV 1	9	10
	Influenza A virus	10-11	14
	Bacteriophage λ	66	49
	Epstein Barr virus	80	170
prokaryotes	*Buchnera sp.*	610	640
	T. maritima	1,900	1,900
	S. aureus	2,700	2,900
	V. cholerae	3,900	4,000
	B. subtilis	4,400	4,200
	E. coli	4,300	4,600
eukaryotes	*S. cerevisiae*	6,600	12,000
	C. elegans	20,000	100,000
	A. thaliana	27,000	140,000
	D. melanogaster	14,000	140,000
	F. rubripes	19,000	400,000
	Z. mays	33,000	2,300,000
	M. musculus	20,000	2,800,000
	H. sapiens	21,000	3,200,000
	T. aestivum (hexaploid)	95,000	16,800,000

- In prokaryotes, gene number varies linearly with genome size
- In eukaryotes the relation is broken, and gene number plateaus around 30K

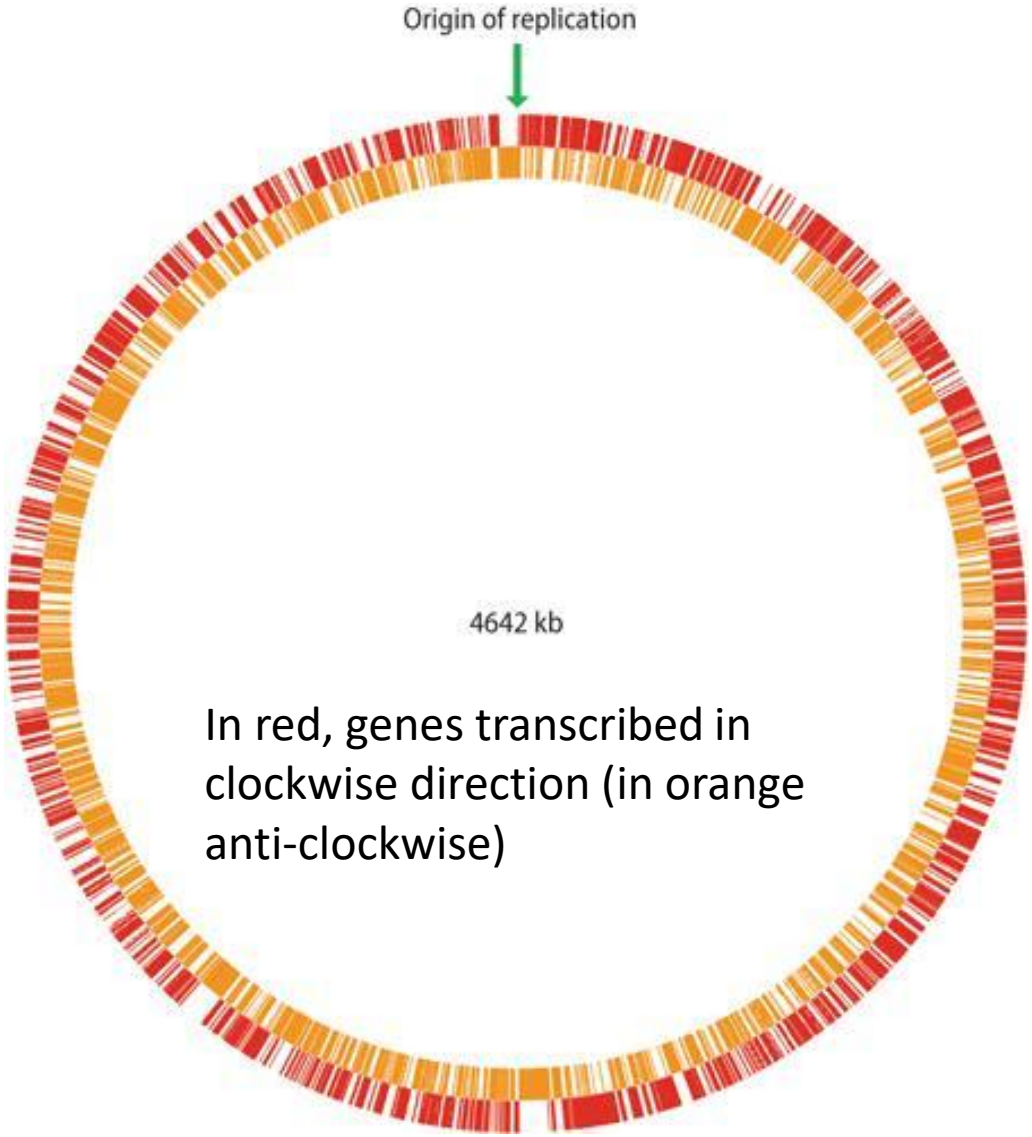


- What remains is non-coding DNA, once called Junk DNA
- The sum of all DNA that is not genes, including repetitive DNA

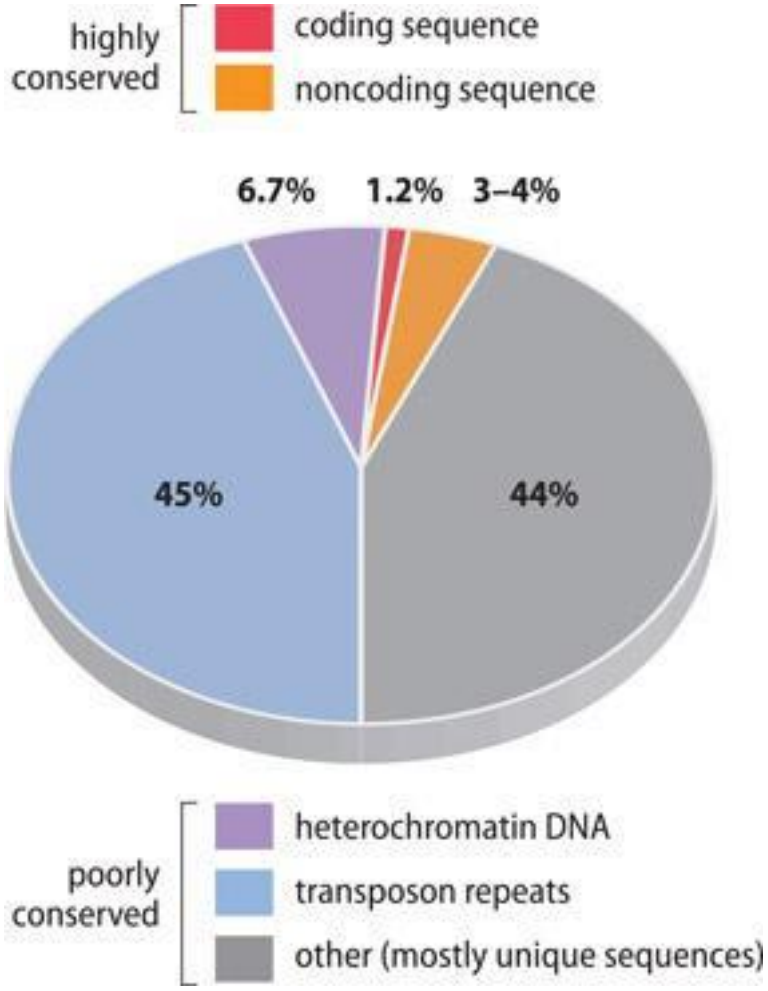


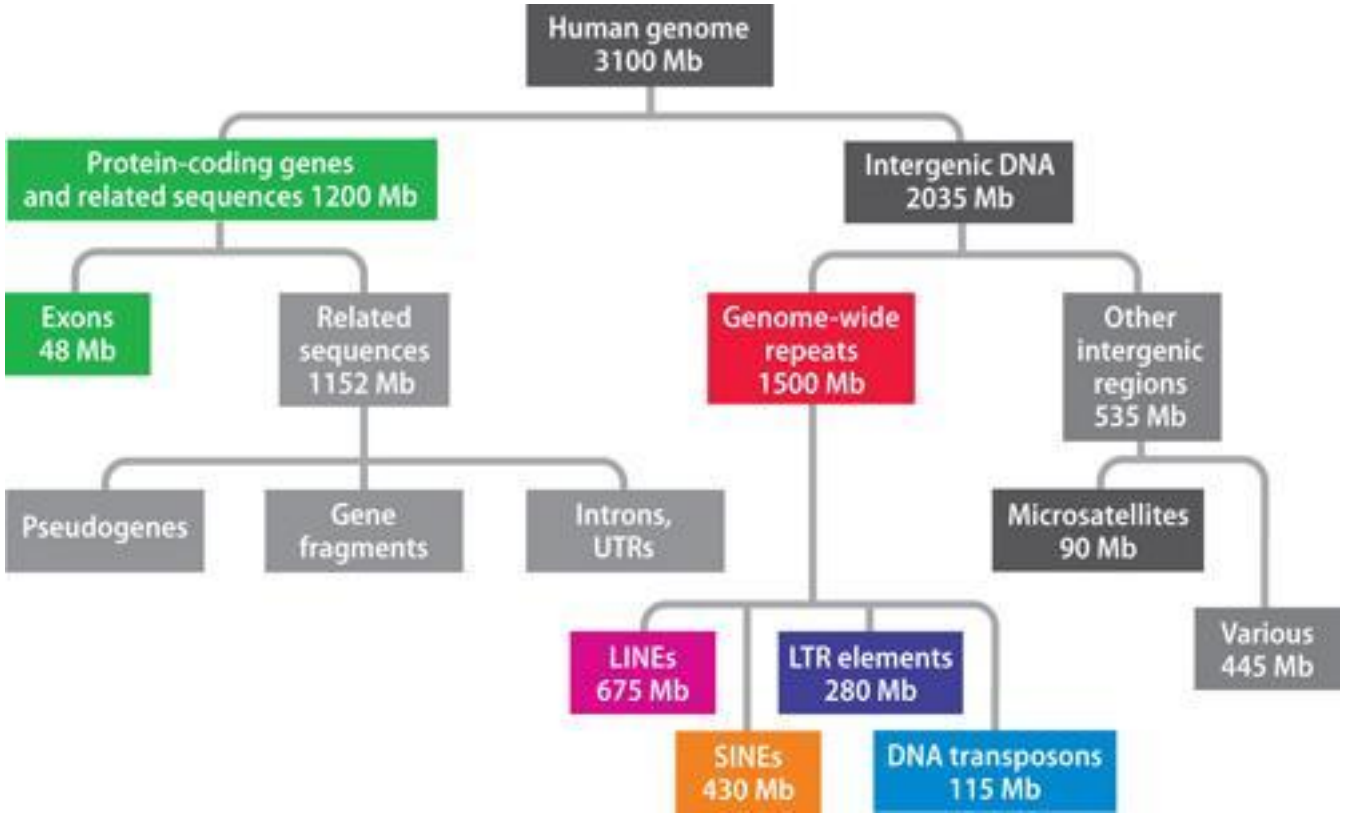
Content of prokaryotic genomes

- Prokaryotic genomes are very compact
- Only about 11% of *E coli* genomic DNA is intergenic at it is interspersed around the genome in bits and pieces. Some genes are literally 1bp apart
- Gene length is about 2/3 of eukaryotic genes; very few introns (and different from eukaryotic introns)



Content of an eukariotic genome





Not all eukaryotes are alike

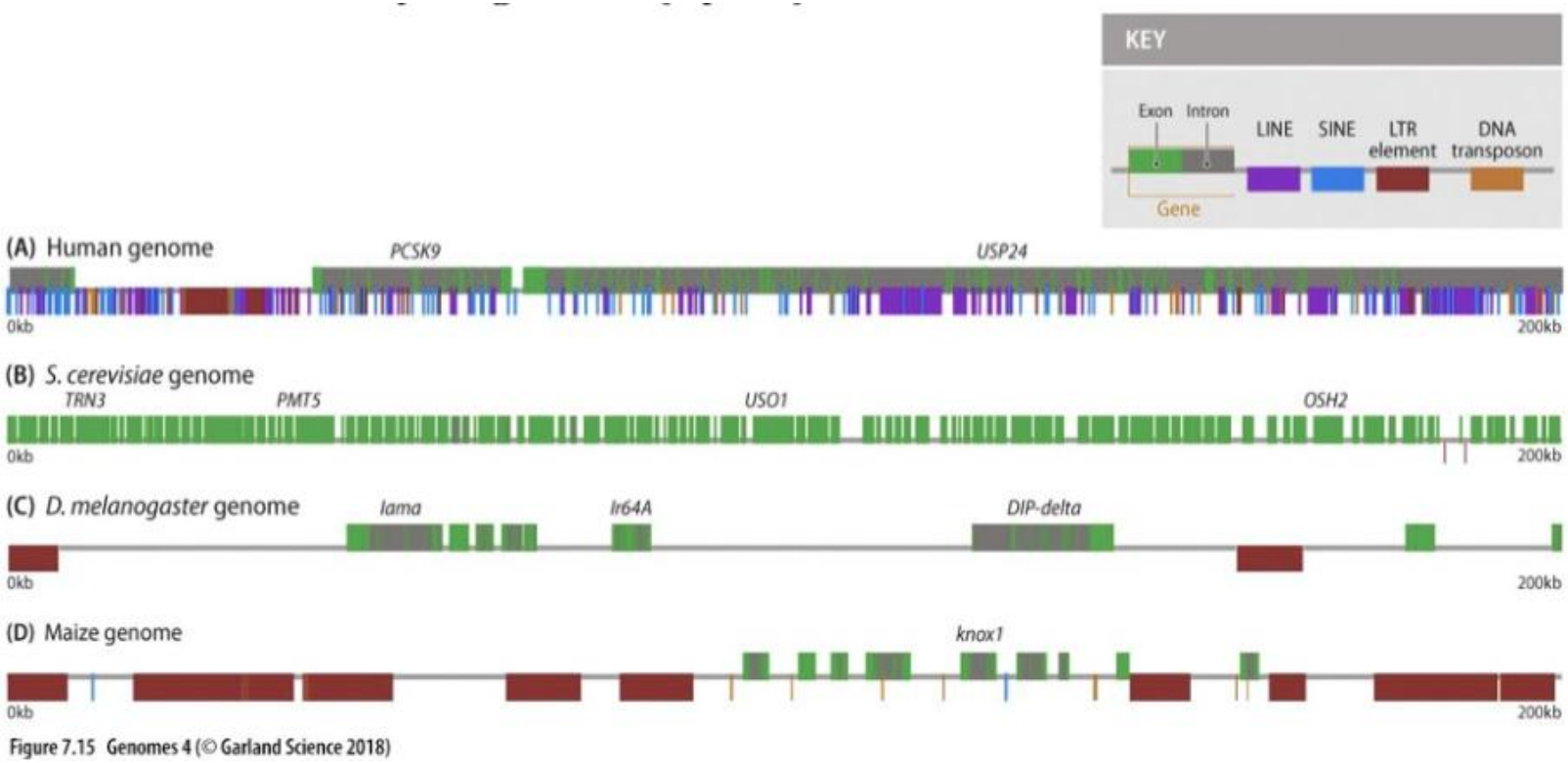
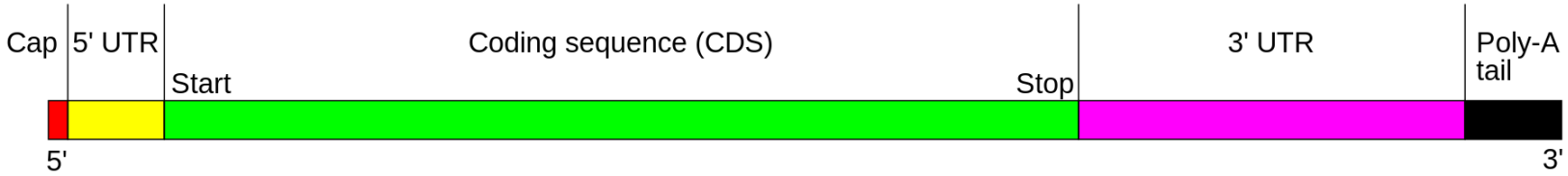
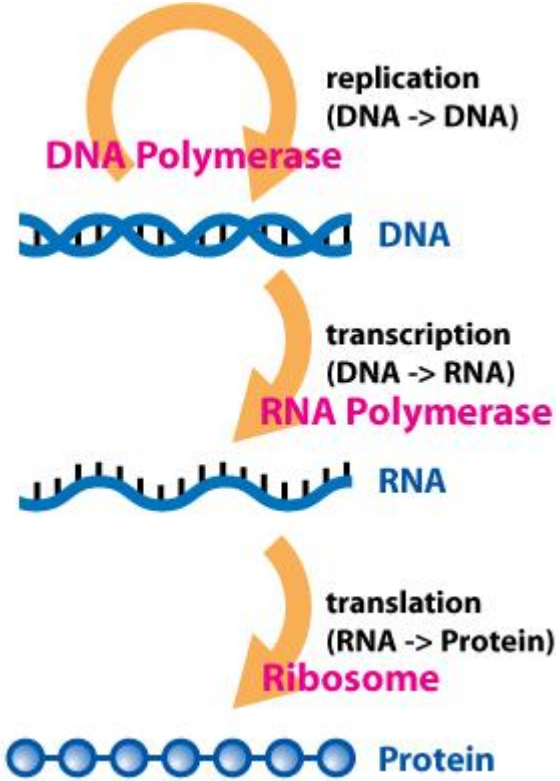


Figure 7.15 Genomes 4 (© Garland Science 2018)

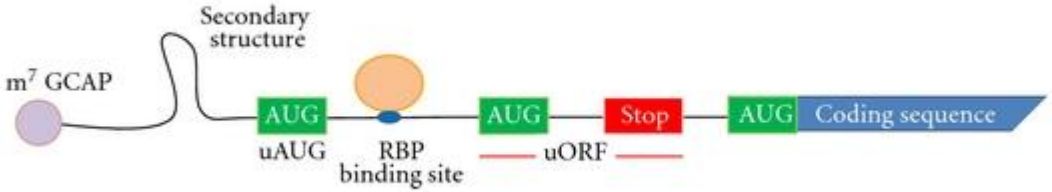
Coding genes

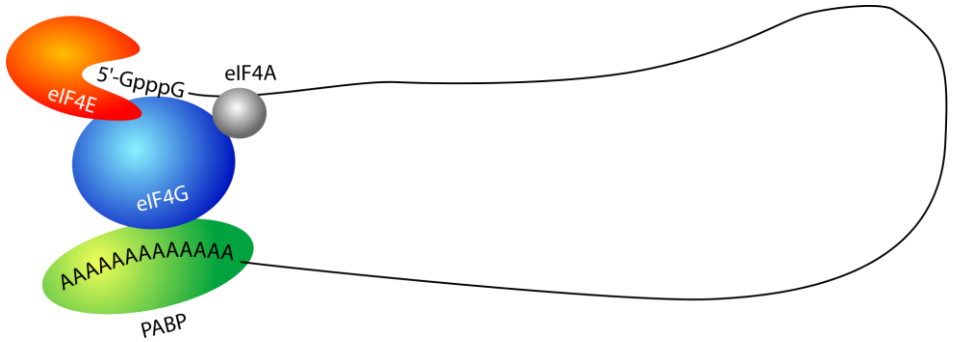
A «coding gene» is a sequence of DNA that contains information that is later translated to proteins (through the intermediate of RNA)





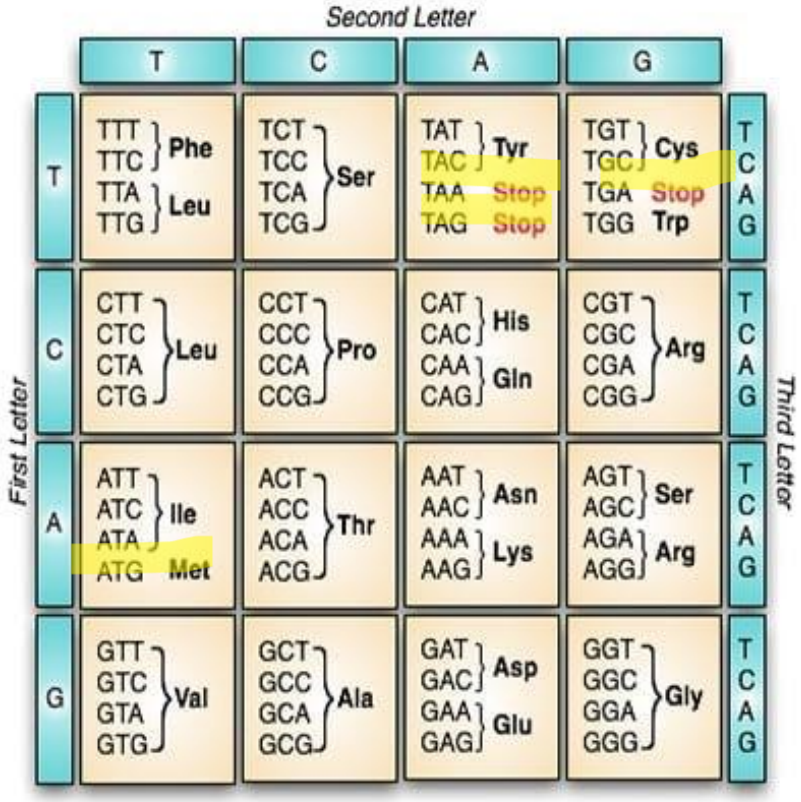
Untranslated region (UTR) refers to either of two sections of the mRNA that are not translated in the protein (however they can produce peptides and/or have regulatory functions)





Open reading frames (ORFs) are the DNA sequence between start and stop codons. Coding DNA!

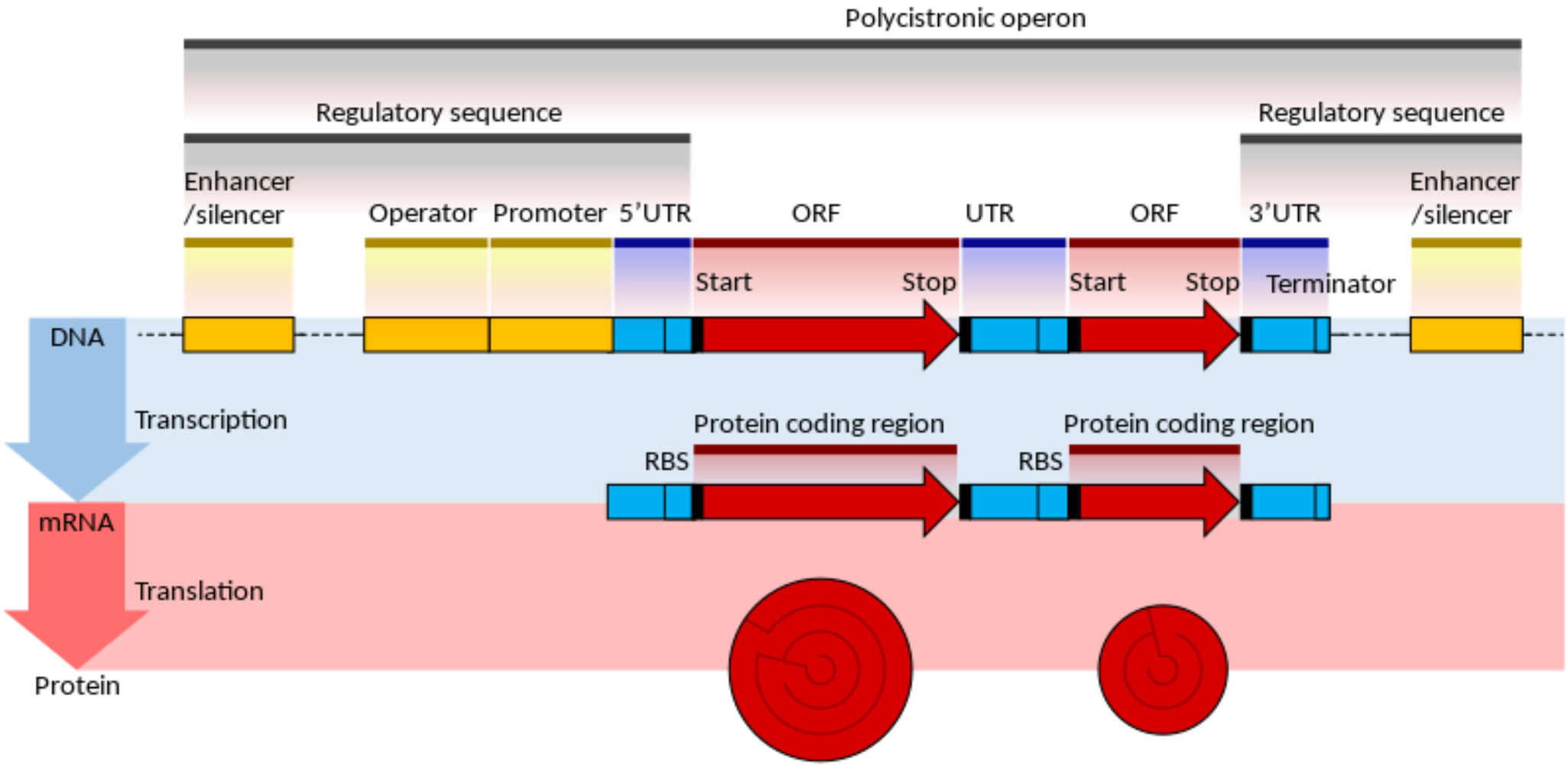
- Since DNA is interpreted in groups of three nucleotides (codons), a DNA strand has three distinct reading frames (six if you consider the two strands)
- In eukaryotes the ORF can only be referred to mature RNA (due to splicing); the more general coding sequence (CDS) term is used



The same piece of DNA can contain different ORFs

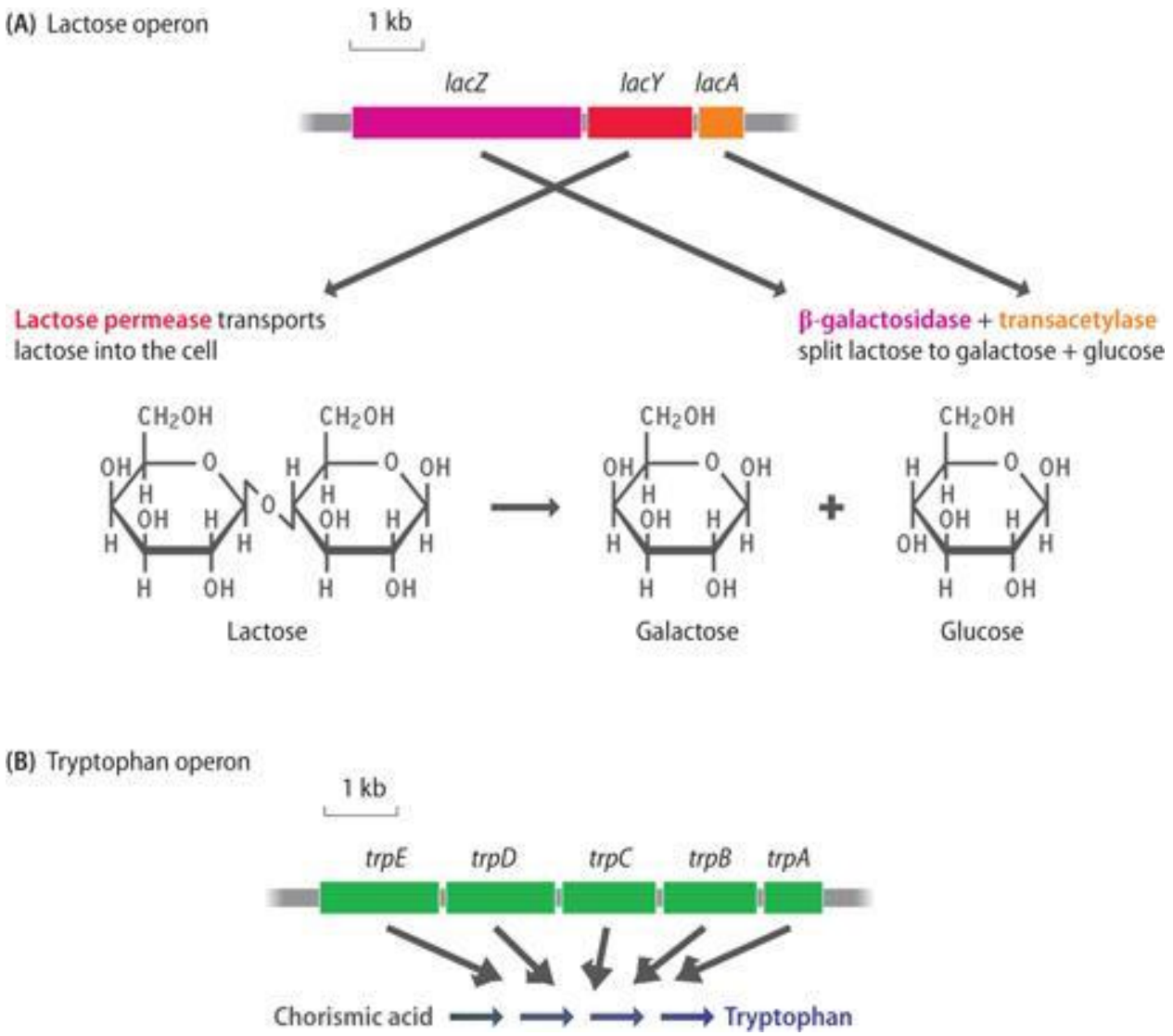
1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

Gene structure in prokaryotes

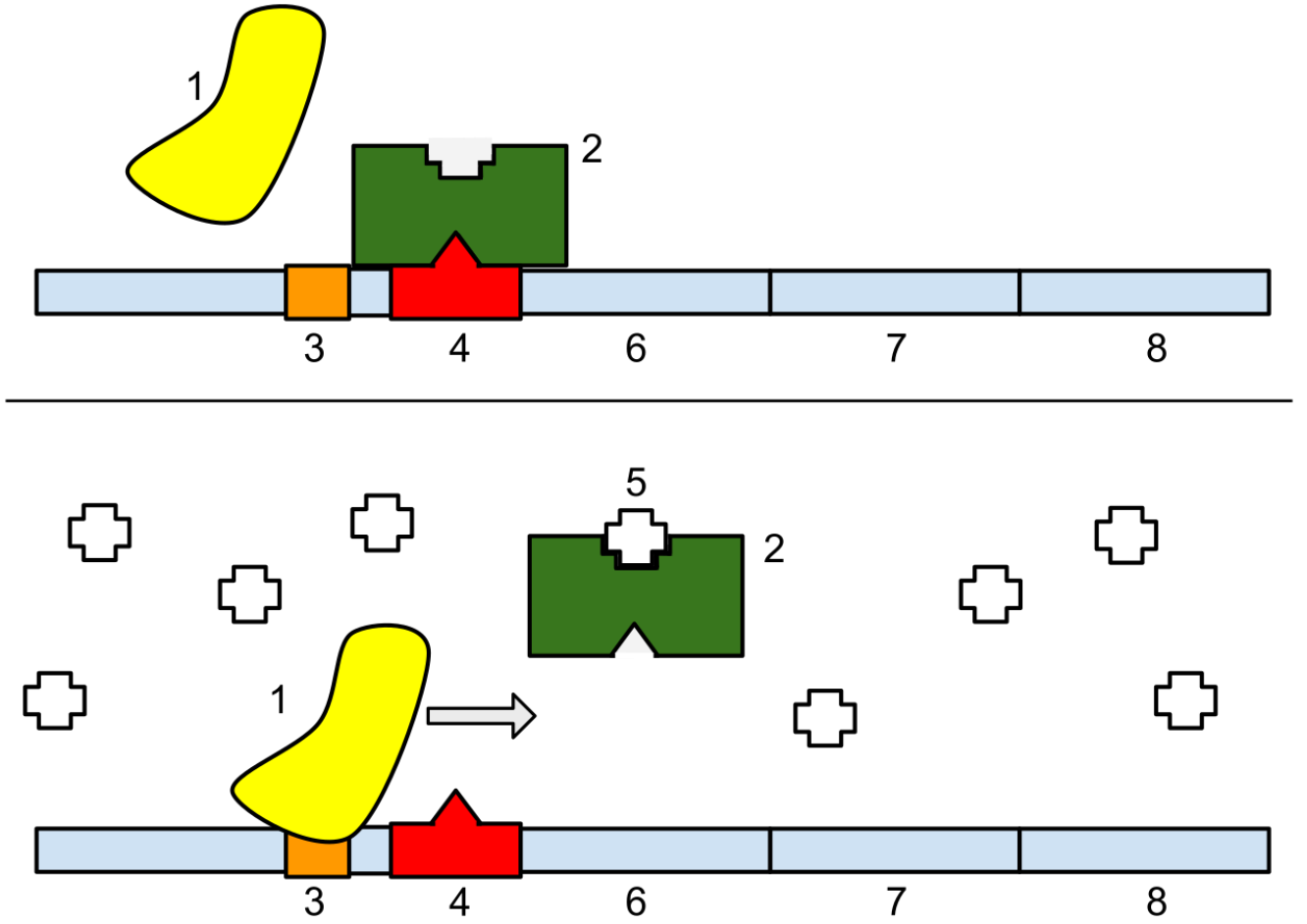


Polycistronic= more than one protein encoded by a single mRNA

- Prokaryotic genomes are characterized by **operons**
- Group of genes adjacent to one another
- Encoding proteins with related functions
- Operons ensure coordinated synthesis of proteins on pathways
- *E. coli* contains 239 operons comprising 2620 genes (64% of the total)

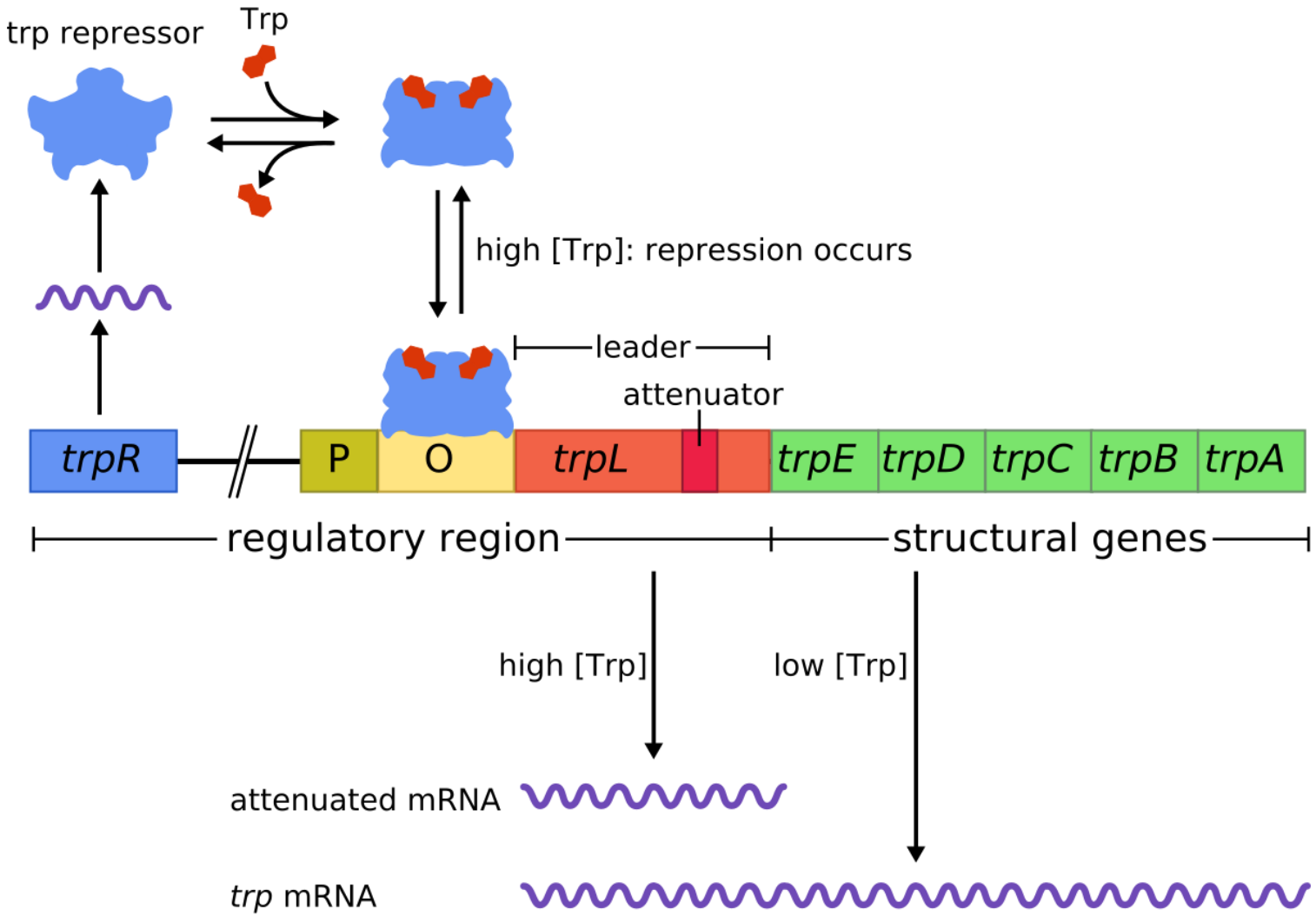


Typical mechanisms of operon regulation



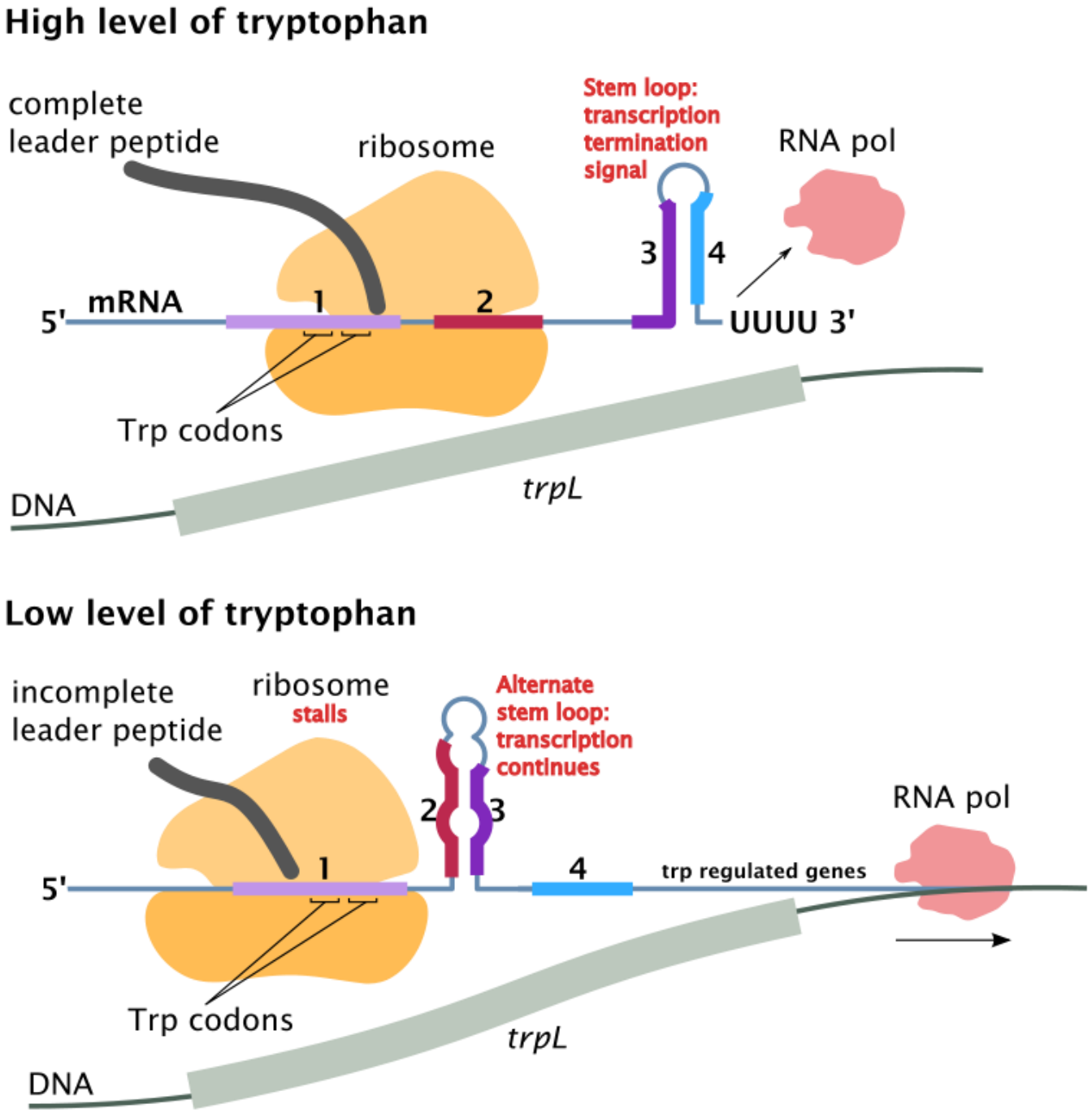
1: RNA Polymerase, 2: Repressor, 3: Promoter, 4: Operator, 5: Lactose, 6: lacZ, 7: lacY, 8: lacA.

Example given: the *trp* (tryptophan) operon



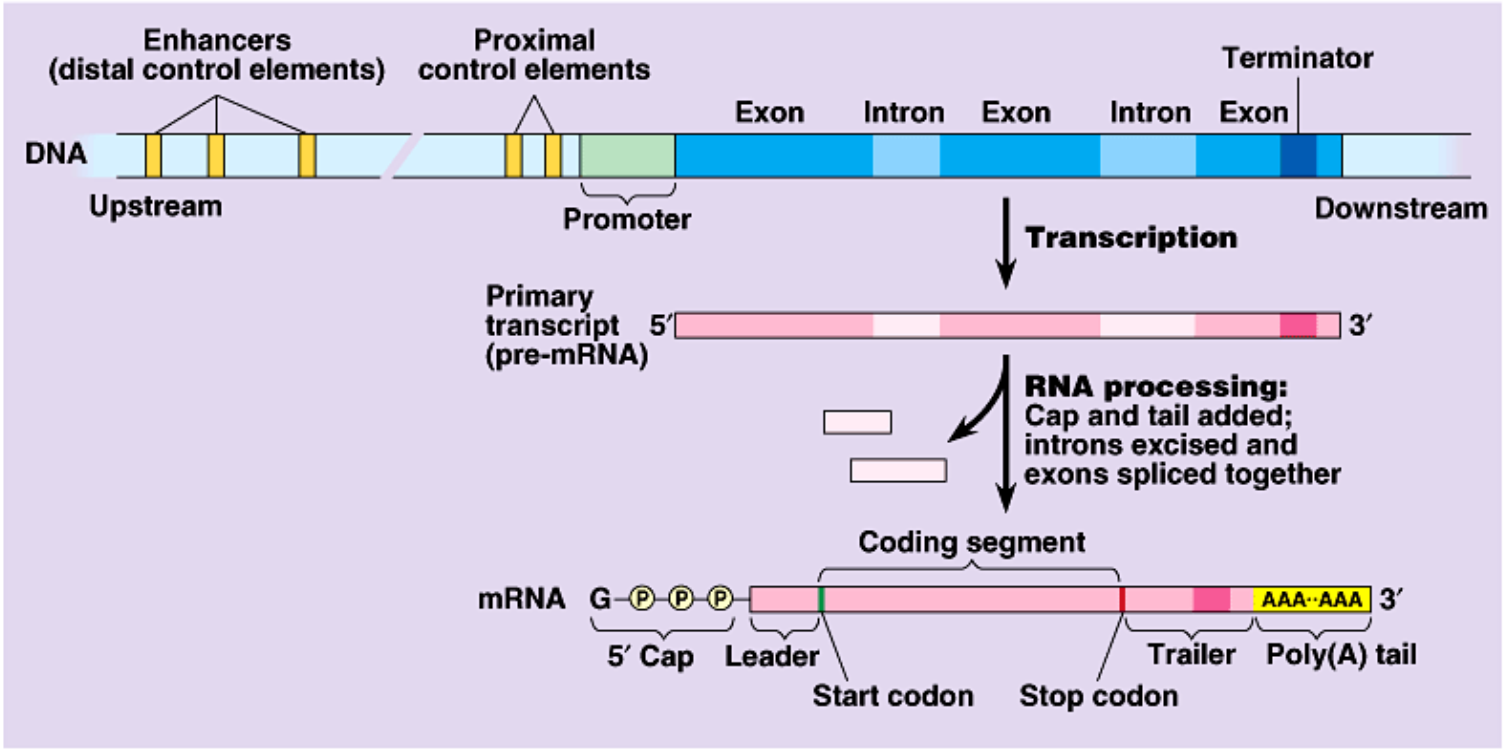
- Five structural genes: enzymes needed to produce tryptophan
- Repressor gene *trpR*
- When tryptophan is present, it activates the repressor, which binds to the operator an inhibits transcription
- Controls expression levels by 70 folds

- Attenuation is a second mechanism exploiting the concurrent transcription and translation
- The attenuator sequence has 4 domains capable of pairing
- The first domain encodes a peptide that needs trp
- if trp is not present, the ribosome stalls and transcription is enabled
- if trp is present, the hairpin loop creates a transcription termination signal (rho-independent termination)
- Controls expression by 10 fold



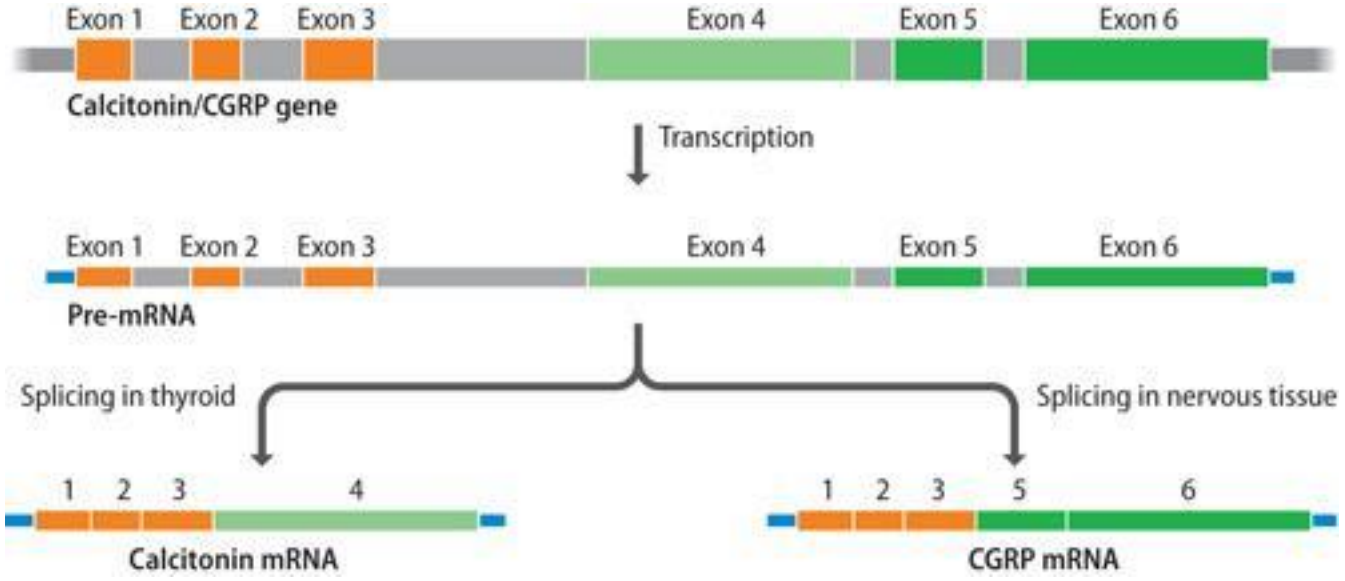
Gene structure in eukaryotes

- In eukarya, the entire gene is transcribed in mRNA but exons are the only components ending up in proteins
- Alternative splicing may end up in alternative mature mRNAs

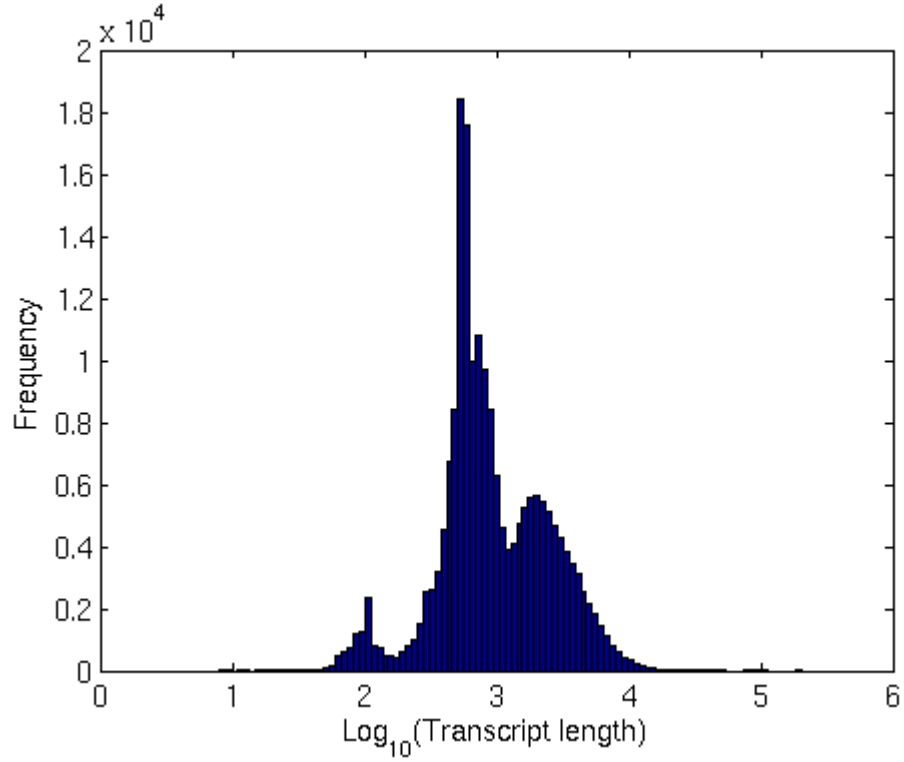


Alternative splicing is relatively common in vertebrates

- 75% of all human protein-coding genes (95% of those with two or more introns) undergo alternative splicing
- With an average of four different spliced mRNAs per gene, 20,442 human genes specify a total of 78,120 proteins
- Alternative splicing also occurs in lower eukaryotes, but it is less prevalent (in *C. elegans* only about 25% of the protein-coding genes have alternative splicing pathways, on average 2.2 variants per gene)



Median gene length in humans is 25Kb



Five longest genes in *H. sapiens*

Transcript stable ID	Gene ID	Gene name	Transcript length	Gene length	Exon count	Intron count	Number of SNPs	Protein size
Longest genes								
ENST00000589042	ENSG00000155657	*TTN*	109,224	118,976	363	362	69,258	35,991
ENST00000397910	ENSG00000181143	*MUC16*	43,816	43,830	84	83	38,498	14,507
ENST00000262160	ENSG00000175387	*SMAD2*	34,626	36,426	11	10	26,668	467
ENST00000330753	ENSG00000185070	*FLRT2*	33,681	34,901	2	1	25,451	660
ENST00000609686	ENSG00000273079	*GRIN2B*	30,355	30,941	13	12	90,195	1,484

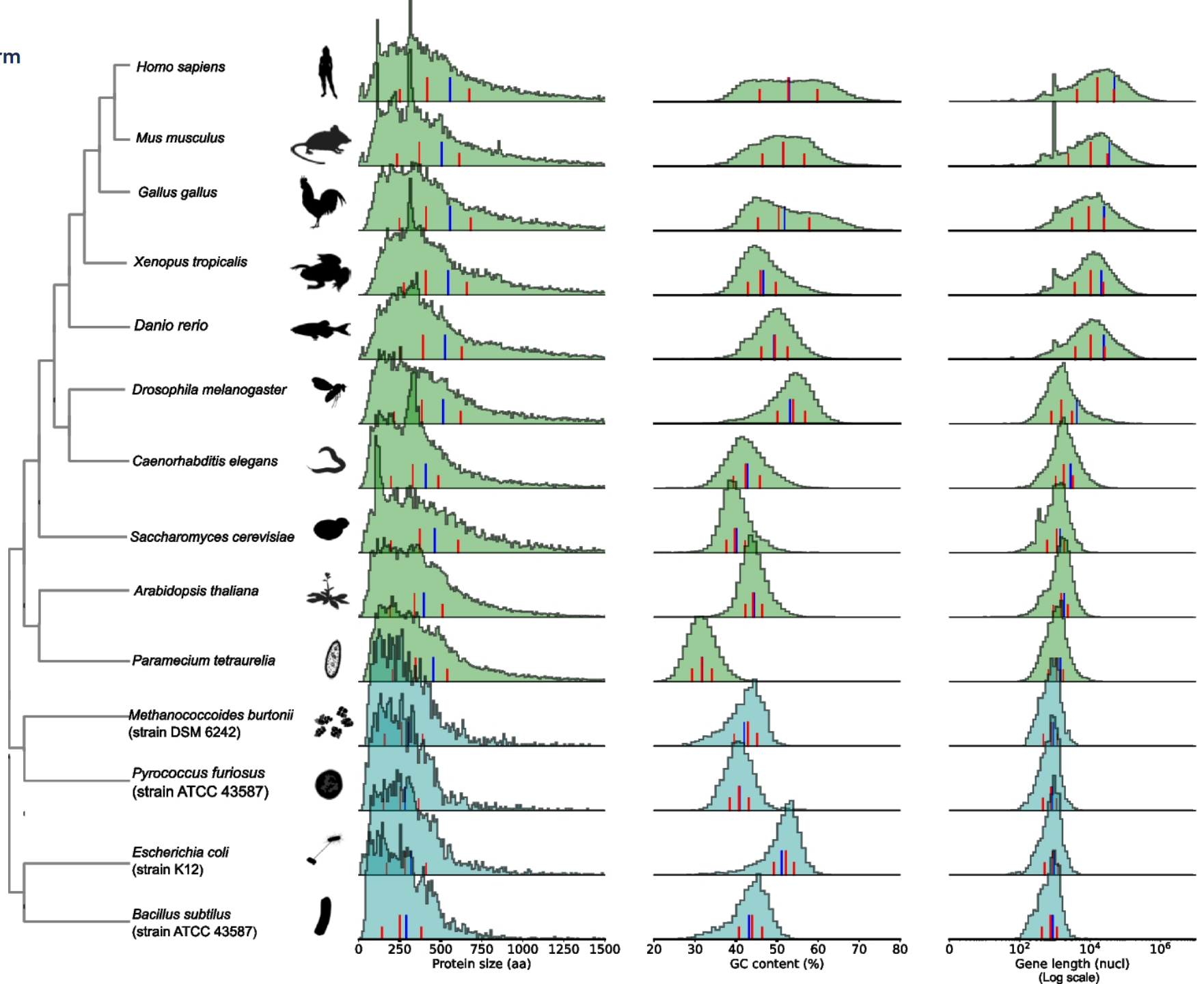
Research | Open access | Published: 08 June 2023

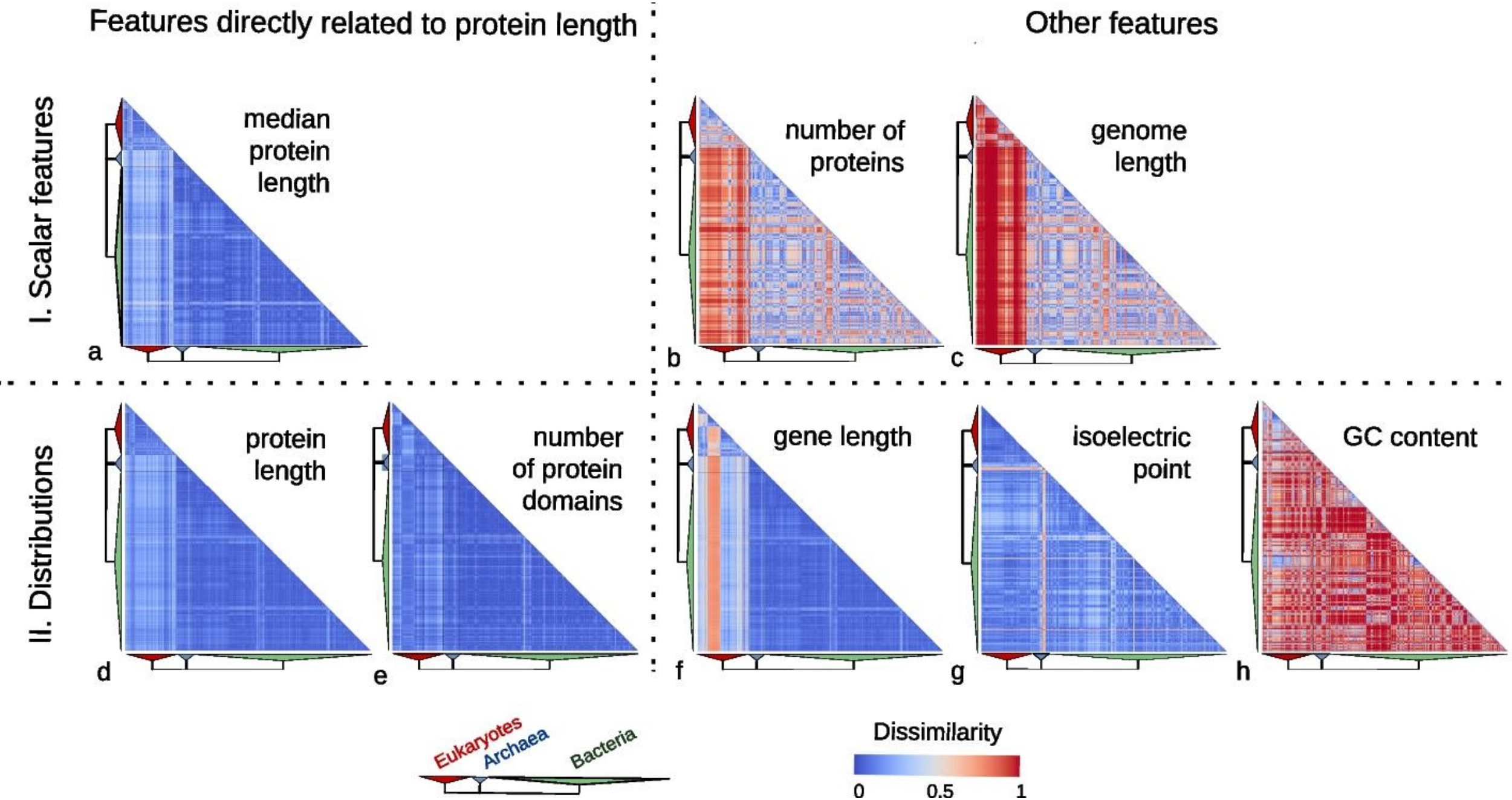
Protein length distribution is remarkably uniform across the tree of life

Yannis Nevers, Natasha M. Glover, Christophe Dessimoz & Odile Lecompte

Genome Biology **24**, Article number: 135 (2023) | Cite this article

- Proteins tend to be longer in eukaryotes than in bacteria or archaea, but variation of length distribution across species is low, especially compared to the variation of genome size, number of proteins, gene length, GC content, isoelectric points of proteins
- evidence for a universal selection on protein length (albeit why is unclear)




Features directly related to protein length
Other features
I. Scalar features
II. Distributions
median protein length
number of proteins
genome length
protein length
number of protein domains
gene length
isoelectric point
GC content
a
b
c
d
e
f
g
h
Eukaryotes
Archaea
Bacteria
Dissimilarity
0
0.5
1

Intronic sequences are not encoded in proteins; their evolution is not constrained by coding function (though genetic code)

- They tend to vary more across individuals and species

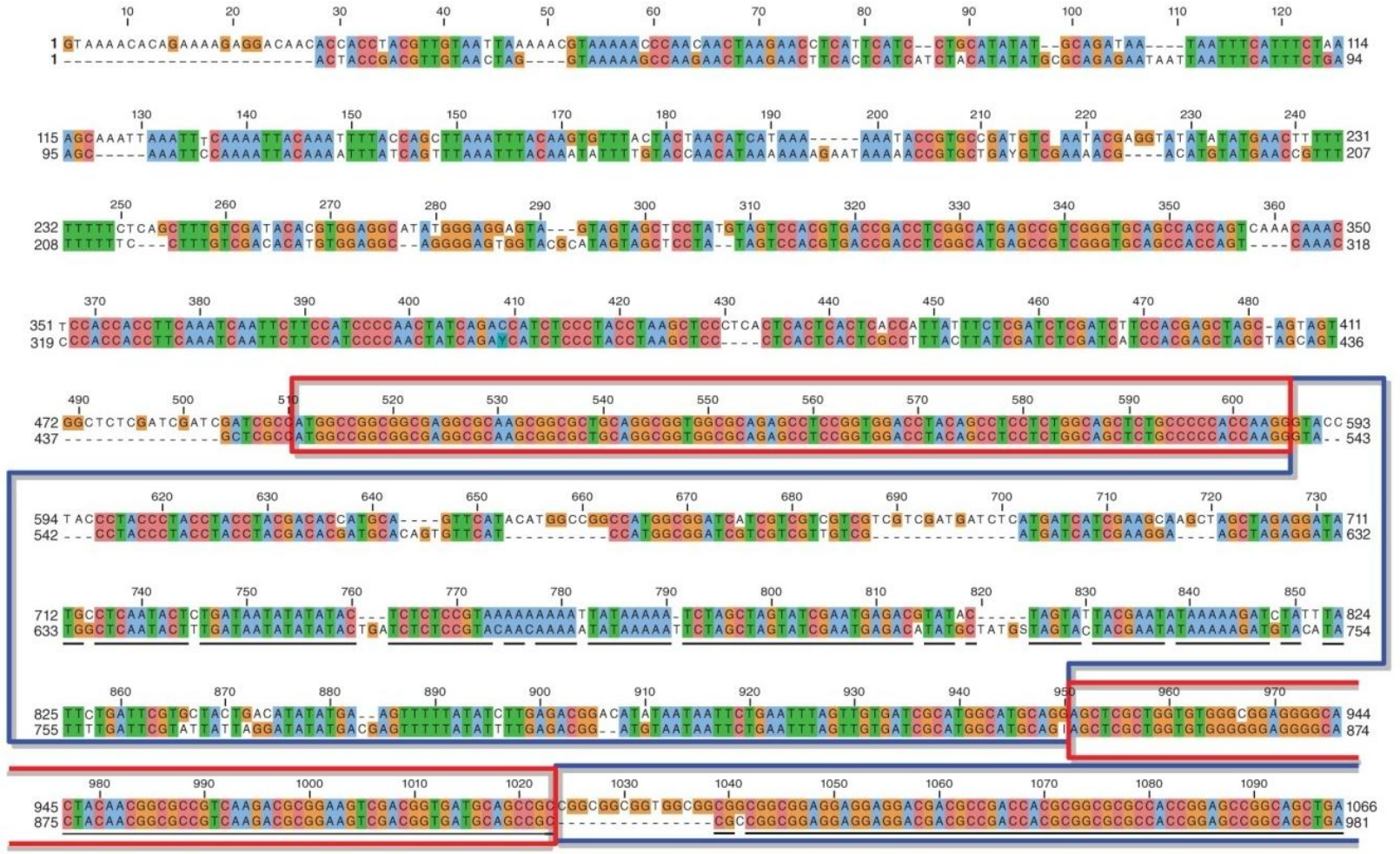
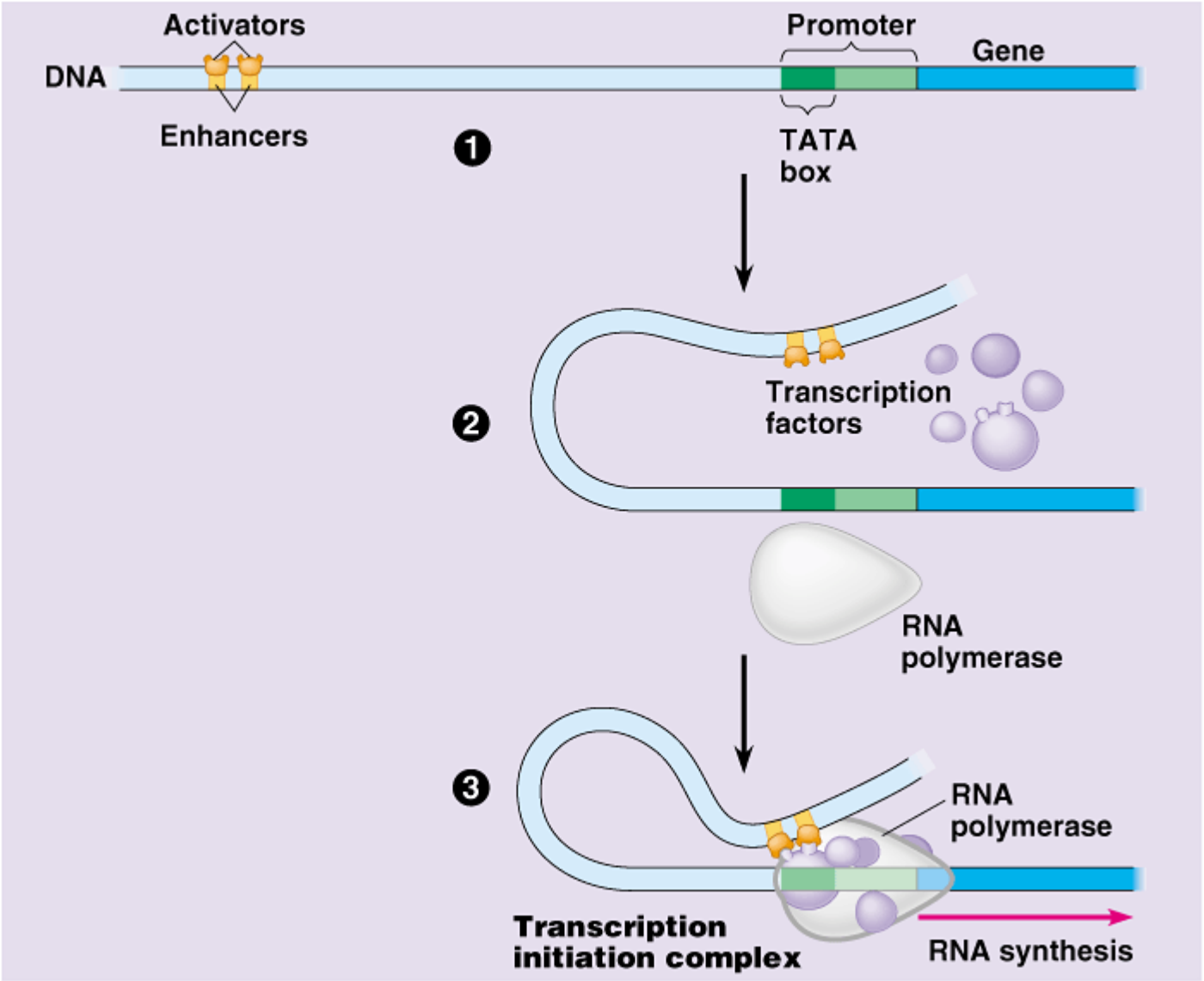


Figure 2.3 Part of the transcribed sequence of the Rc gene in two closely related species of *Oryza*. All sequences shown are transcribed, but only the sequences boxed in red are translated. Sequences boxed in blue are spliced out to form the mRNA. The first exon includes the 5′ UTR (not boxed) plus the translation start (red box). This is followed by the first intron (blue box), and the second exon (red box), and a portion of the second intron (blue box). Notice the difference in sequence conservation between the various regions. This particular portion of the 5′ UTR is ca. 70% identical between the two species (358/510 bp), and the first intron is 79% identical. The translated portion of the first exon is 100% identical, whereas the second exon is 99% identical. Data from Gross *et al.*, 2010

Upstream and/or downstream genes it is possible to find sequences whose function is to activate/deactivate expression, typically by binding protein complexes

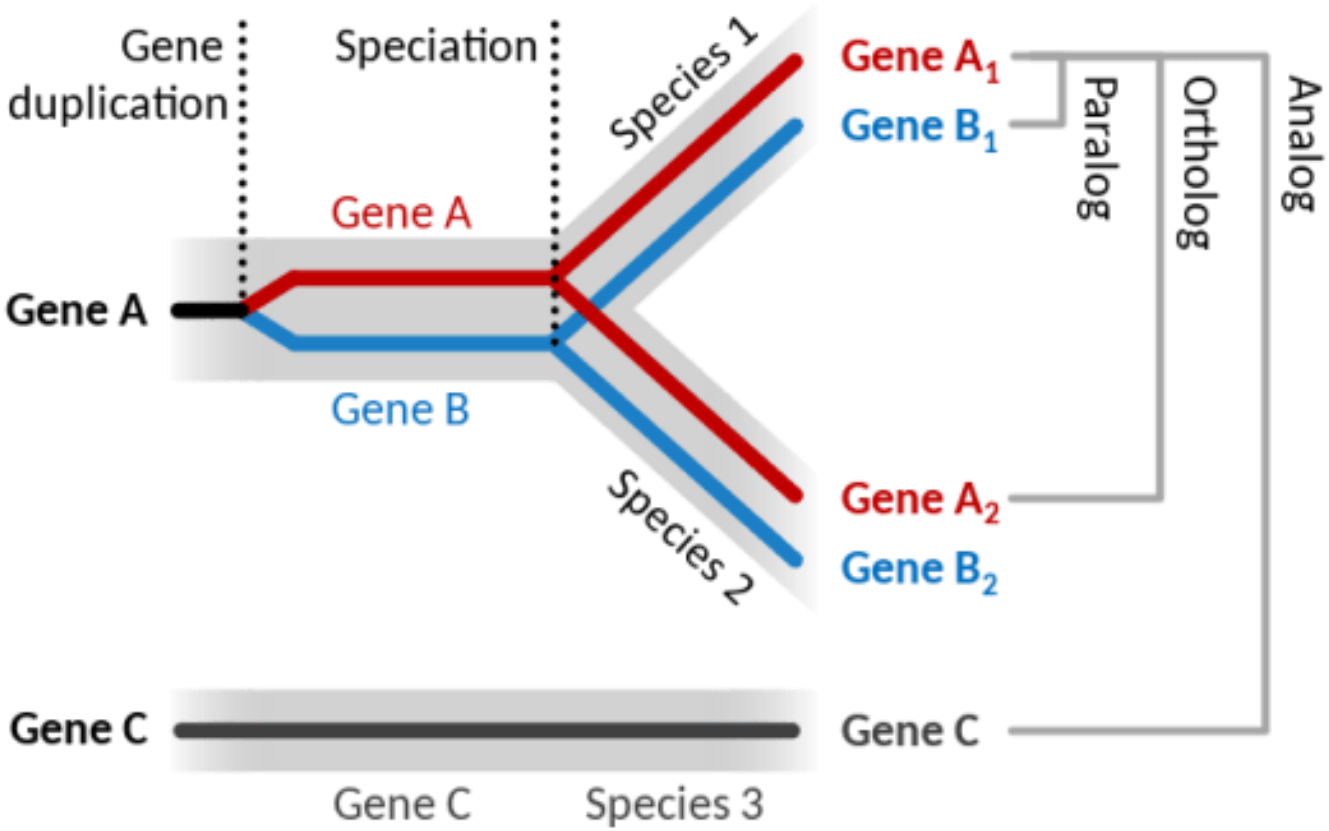
The same holds for sequences promoting histone occupancy



Genes separated by speciation are called orthologs. Genes separated by gene duplication events are called paralogs.

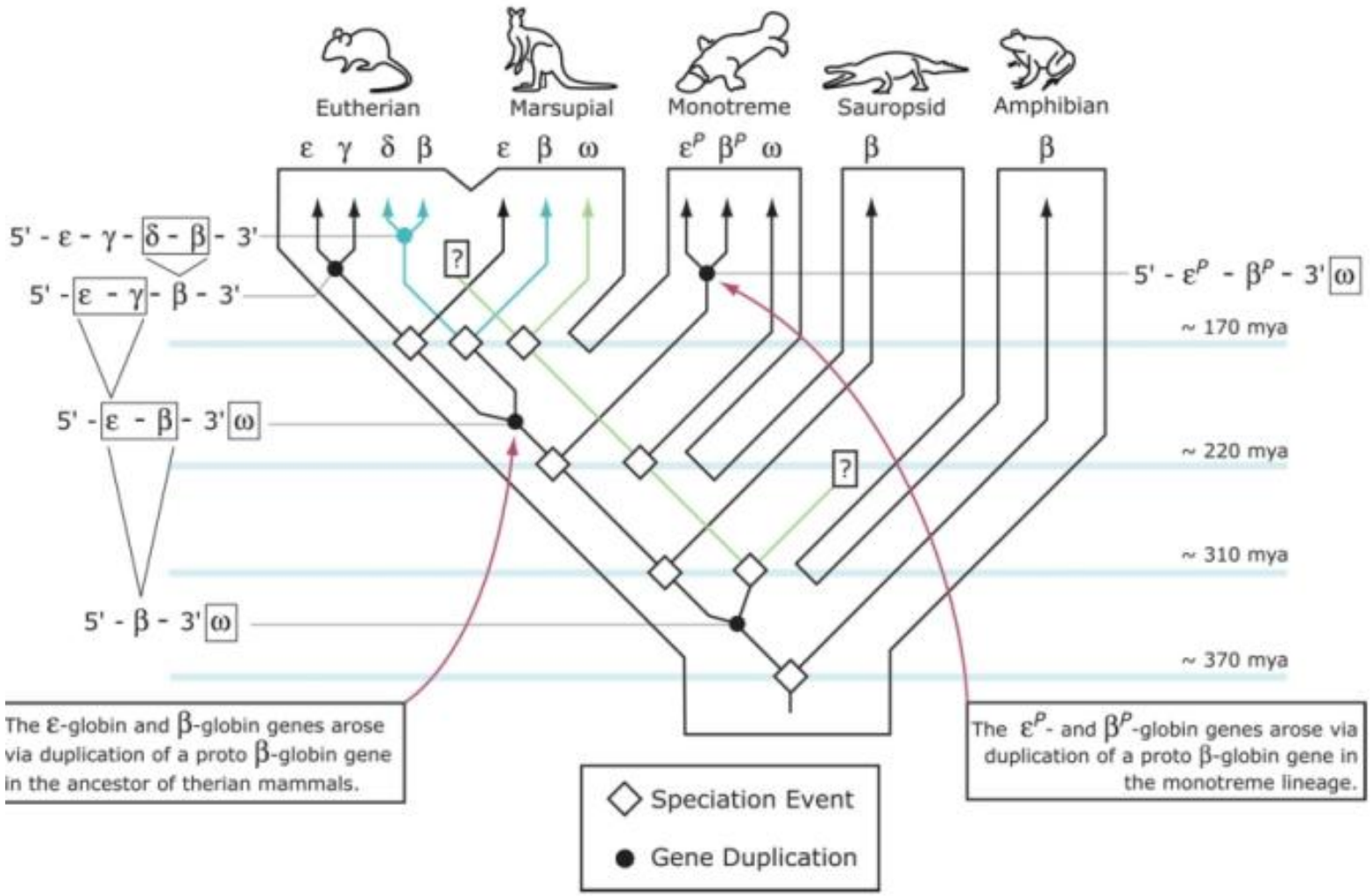
Gene families are groups of similar genes tipically with common orgin

- Similar sequence and similar structure
- Identified by sequence homology (either on DNA or proteins)

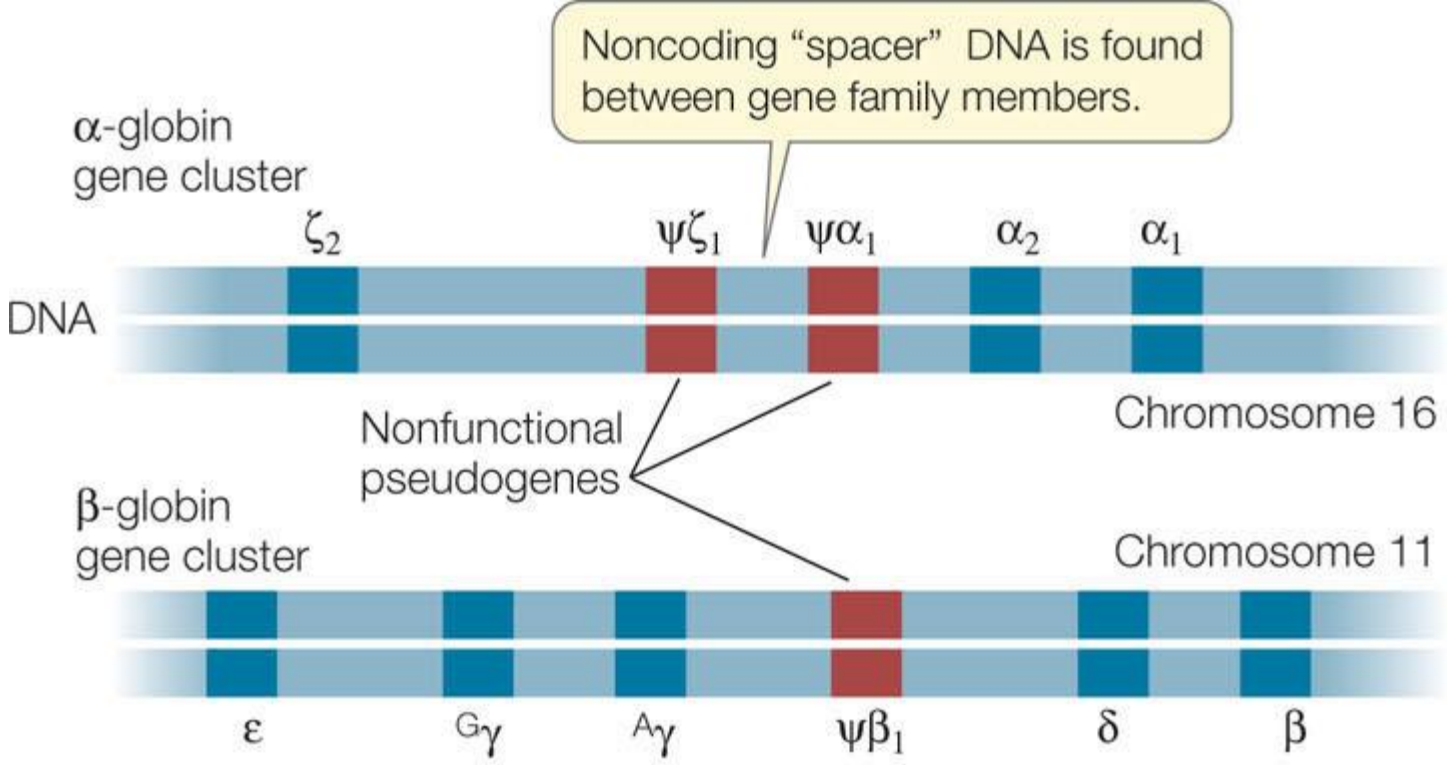


Multigene families –groups of genes of identical or similar sequence – are common features of many genomes

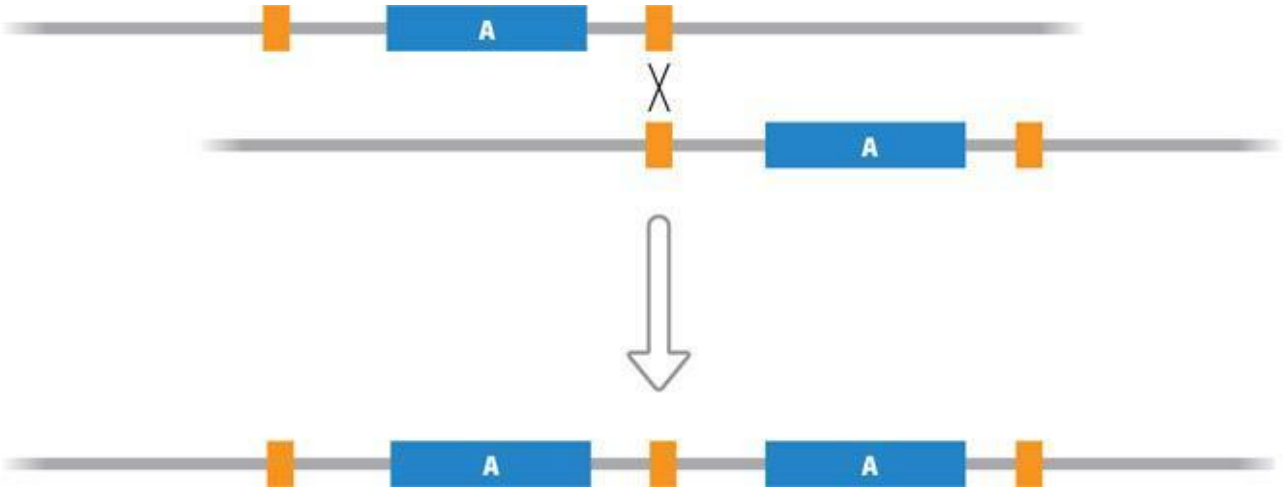
- **Simple**: may arise by gene duplication, classic example is the rRNA genes. Several thousand copies of the 5S rRNA exist on chr 1 in human, plus repeats of other subunits elsewhere
- **Complex**: made of individual members similar in sequence but dissimilar enough to have different properties. Classic example are the mammalian globin genes: ε is expressed in the early embryo, γG and γA (whose protein products differ by just one amino acid) in the fetus, and δ and β in the adult. All do the same stuff, but differ for oxygen affinity



Not by chance gene family clusters are in close vicinty

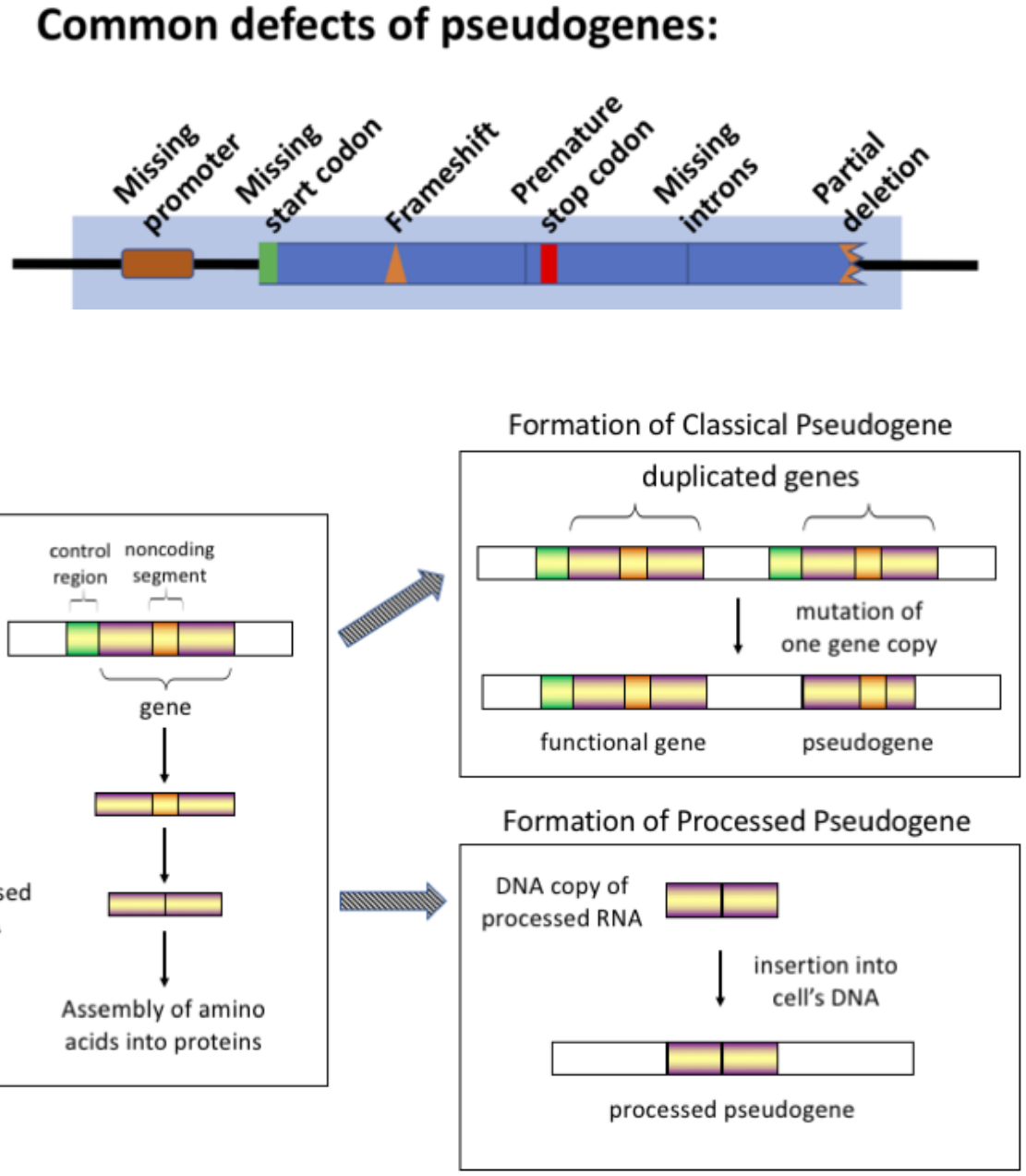


- Exon duplication and shuffling
- Entire gene duplication
- Multigene family duplication
- Whole genome duplication
- TE-mediated relocation



Pseudogenes and **gene fragments** are nonfunctional segments of DNA that resemble functional genes

- Most arise as superfluous copies of functional genes, either directly by gene duplication or indirectly by reverse transcription of an mRNA transcript
- Once the cease to be functional they start accumulating mutations until they are no more recognizable
- Pseudogene sequences may be transcribed into RNA at low levels, due to promoter elements inherited from the ancestral gene or arising by new mutations (see long non-coding RNAs)



Mechanisms of gene duplication and functionalization through reverse transcription

a reverse transcriptase is needed, often derived from retrotransposons (such as LINE-1 elements in mammals).

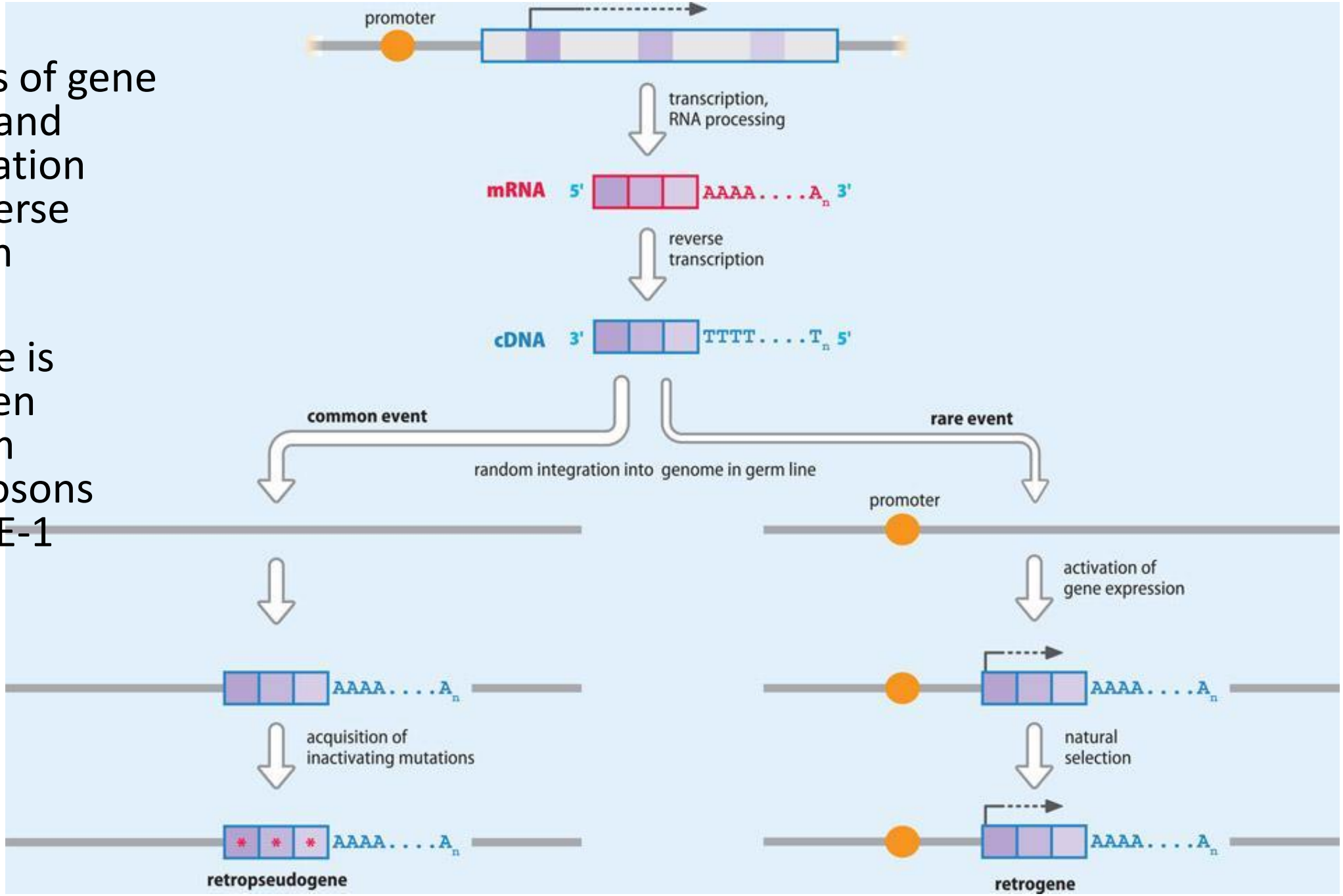


TABLE 2.4 A SNAPSHOT OF THE NUMBERS OF HUMAN GENES AND PSEUDOGENES LISTED BY GENCODE VERSION 40 (RELEASED IN APRIL 2022)

Class	Number
PROTEIN-CODING GENES	19988
RNA GENES	26372
making long ncRNA	18805
making short ncRNA	7567
PSEUDOGENES	14774
processed	10661
unprocessed	3566
other	547

Obtained at http//gencodegenes.org/human/stats.html