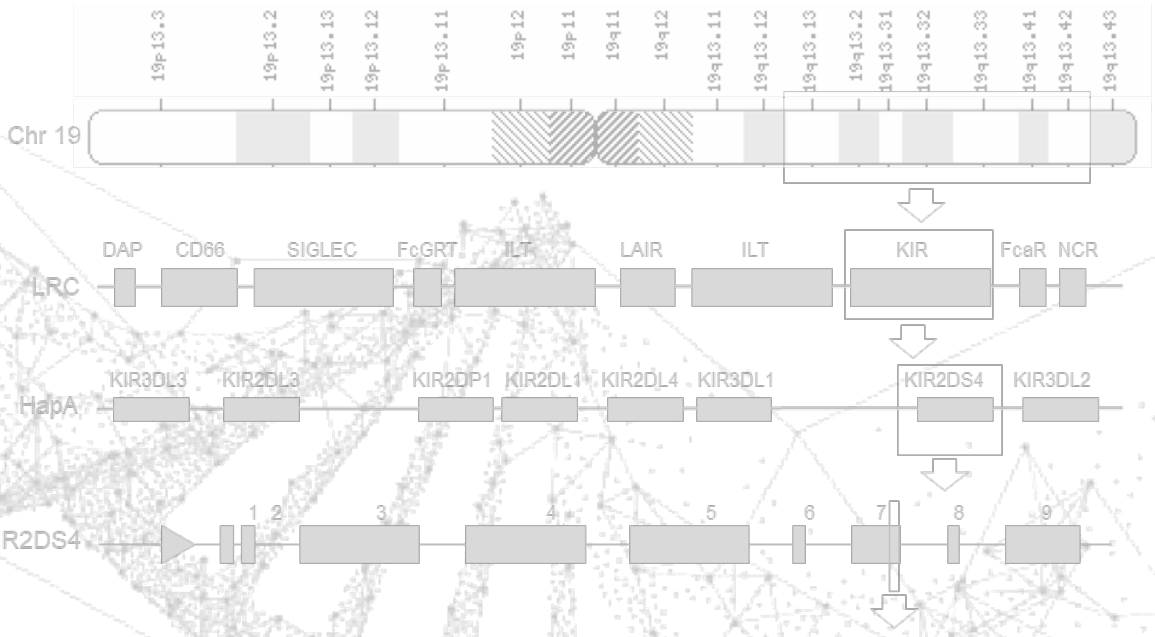


Genes and genomic organization

San Luis Potosi State University (UASLP) Mexico
Molecular Biology Course, Faculty of Medicine post-graduate program

Dr. Christian A. García-Sepúlveda

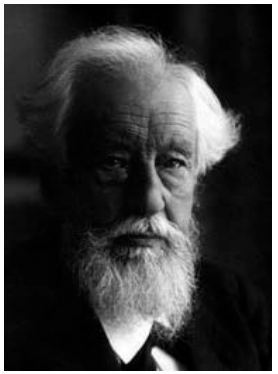
Viral & Human Genomics BSL-3 Laboratory

Last updated November 11, 2024 v1

History

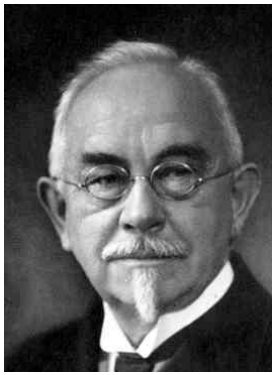
The existence (but not the name) of **genes** was proposed by Gregory Mendel (1882-1884).

Mendel had crucial concepts: **independent segregation** of chromosomes, distinction between **recessive** and **dominant** traits, difference between **homozygotes** and **heterozygotes** as well as the difference between **genotype** and **phenotype**.



Hugo de Vries in 1889 had the unusual decency of today to give credit to Mendel but proposed the term "**pangene**" to describe the minimal unit of inheritance.

William Johannsen abbreviated the term to **gene** two decades later.



In 1920, Thomas Hunt Morgan demonstrated that genes resided on chromosomes (and at specific sites within them called **locus**).

Definition of a gene

The Mendelian gene is a basic unit of heredity.

The Mendelian gene is the classical gene of genetics and refers to any heritable trait.

The Selfish Gene of the gene-centered view of evolution.

The molecular gene is a sequence of nucleotides in DNA that is transcribed to produce a functional RNA.

There are two types of molecular genes: protein-coding genes and non-coding genes.

During gene expression (the synthesis of RNA or protein from a gene), DNA is first copied into RNA.

RNA can be directly functional or be the intermediate template for the synthesis of a protein.

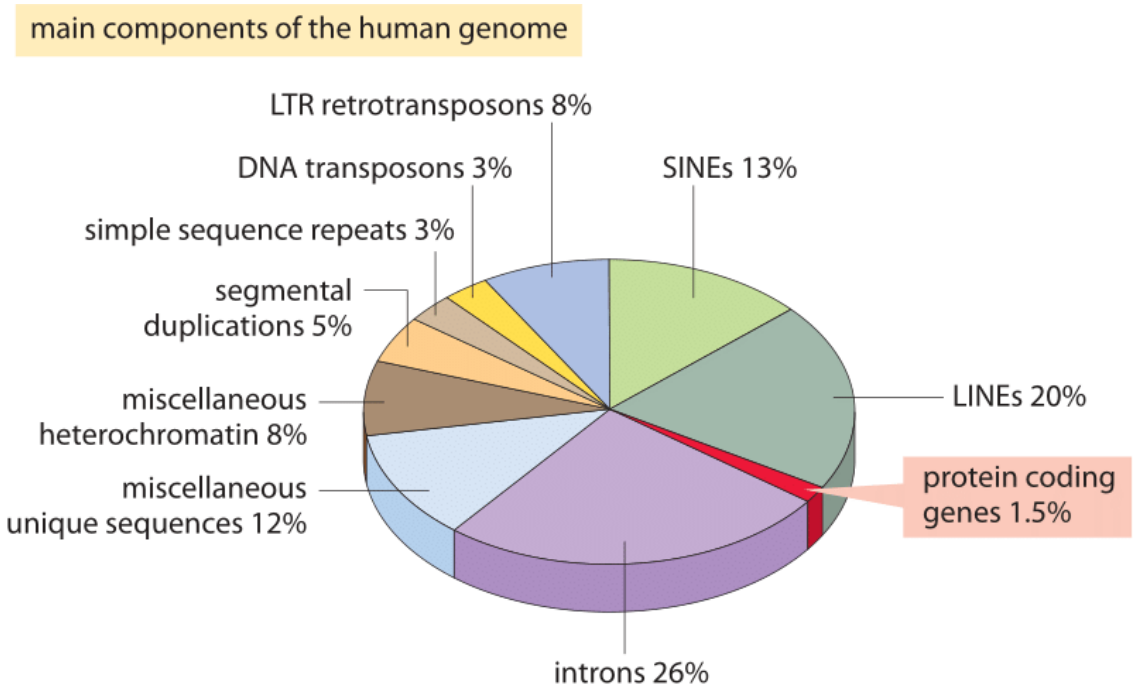
Molecular genes

Typical mammalian transcribed regions are ~ 62,000 bp in length

As HoSa has ~ 20,000 genes.

Transcribed regions make up approximately 1.5% of the HoSa genome

Interspersed by the non coding introns, making up about 26%.



Transposable elements are the largest fraction (40-50%) and include:

Long interspersed nuclear elements (LINEs)
Short interspersed nuclear elements (SINEs).

Most transposable elements are defunct genomic remnants

T. R. Gregory Nat Rev Genet. 9:699-708, 2005

Molecular gene

Genome sizes differ by as much as 8 orders of magnitude (from <2 kb for Hepatitis D virus to >100 Gbp for the Marbled lungfish.

Gene numbers differ by less than 5 orders of magnitude (from 4 in bacteriophages to 100,000 in wheat).

In prokaryotes gene content is proportional to the genome size.

In eukaryotes this is not true.

The inability to successfully estimate the number of genes in eukaryotes based on knowledge of the gene content of prokaryotes was one of the unexpected twists of modern biology.

	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

Gene numbers and complexity

The organism with the most genes known to date is a small water flea called *Daphnia pulex*, with ~ 31,000 genes.

This count surpasses that of humans, who have around 20,000–25,000 genes.

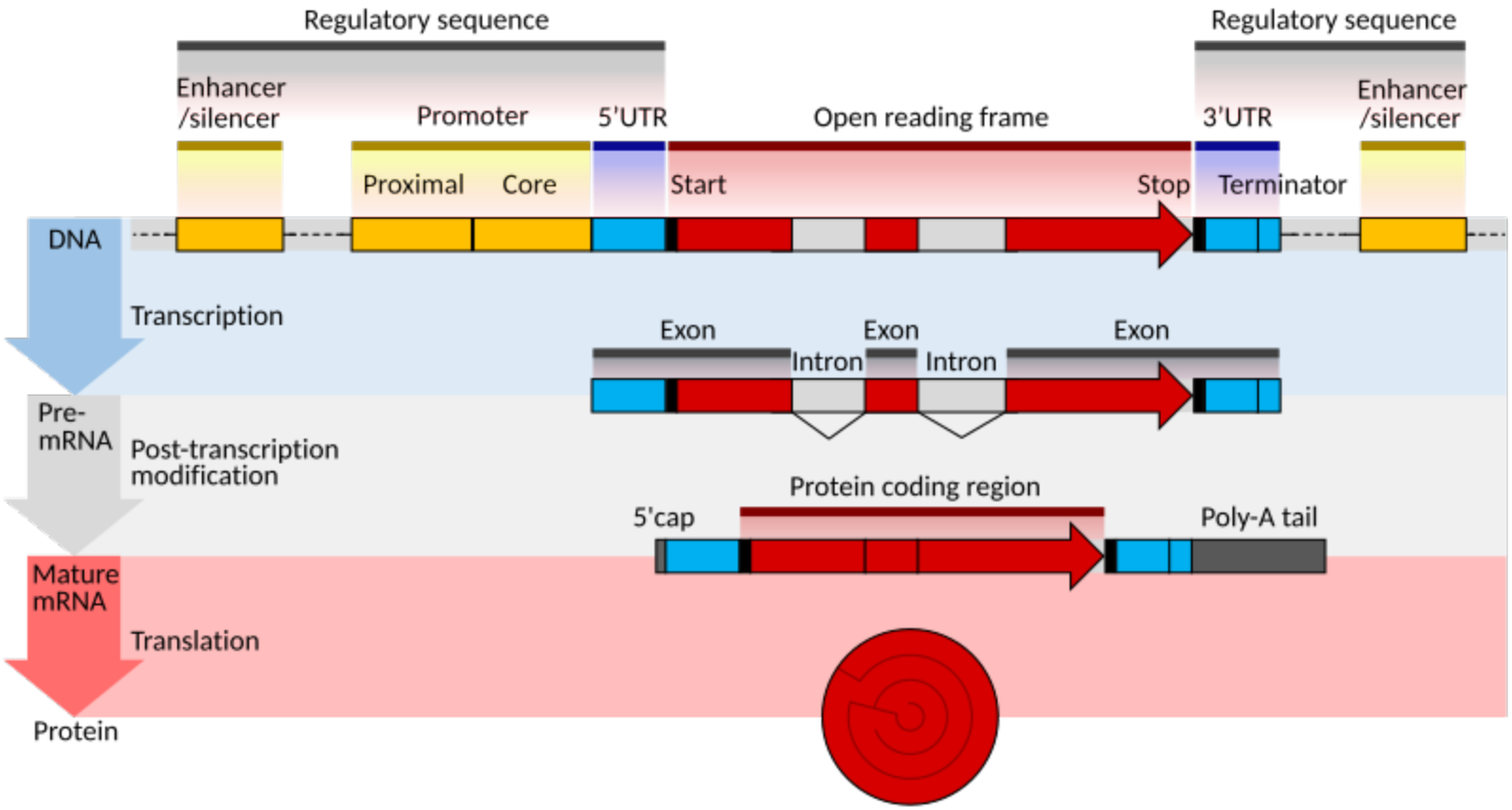
Some plants such as wheat (*Triticum aestivum*) have high gene count ~ 95,000 though many of these are duplicates due to the plant's polyploid nature (multiple sets of chromosomes).

High gene counts don't correlate with complexity, gene number can increase due to genome duplication events, common in plants.

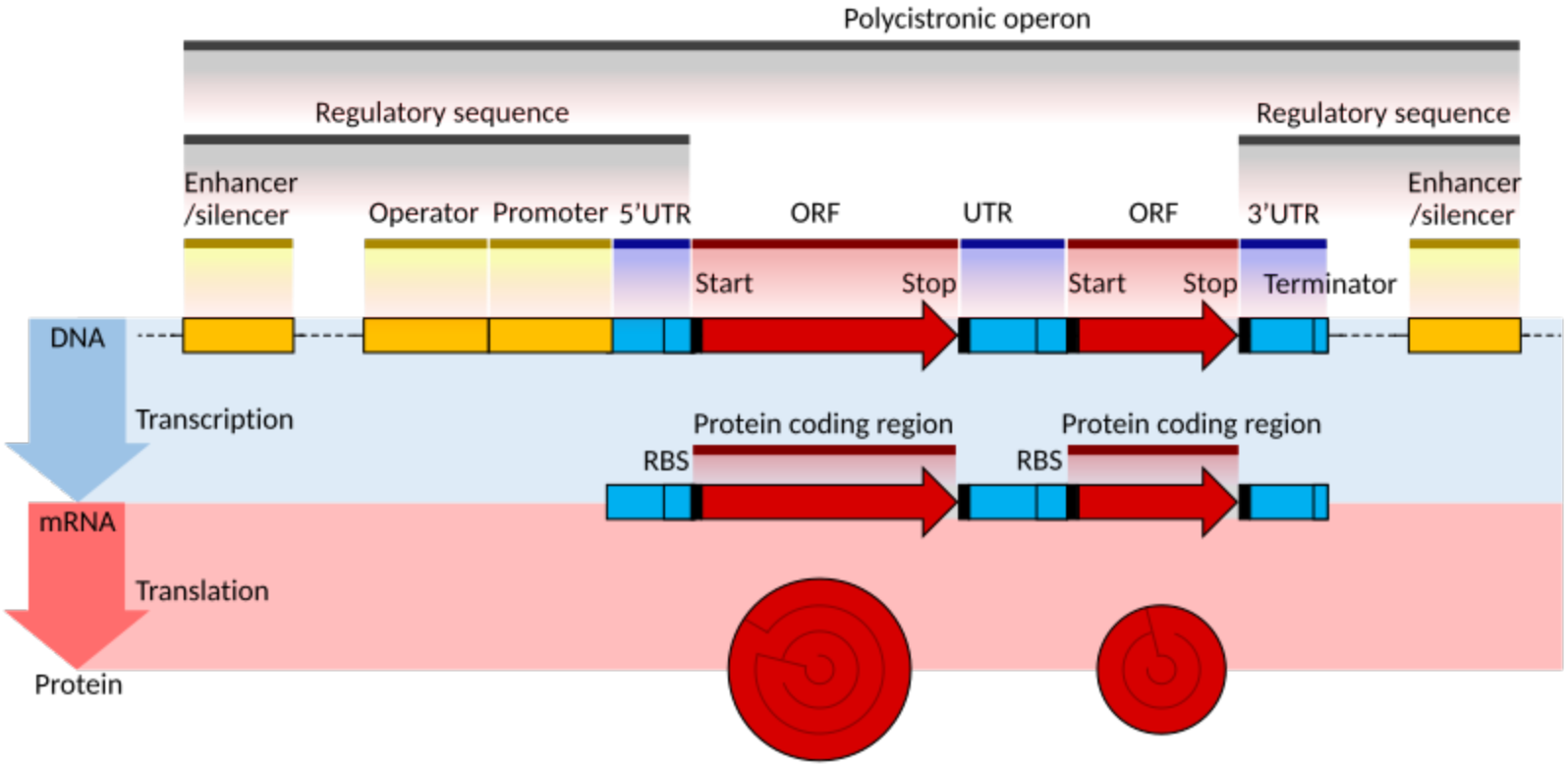
Duplications create more genetic material, often without adding new functions.


Daphnia pulex

Eukaryote genes



Prokaryote genes



Eukaryotes vs prokaryote genomes

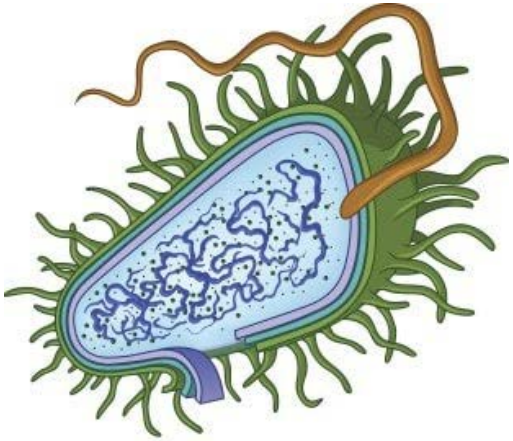
Two different priorities distinguish the forms and habits of prokaryotes and eukaryotes.

Prokaryotic philosophy is minimalist, they possess only what is essential to propagate the genes of isolated individuals (selection of the fittest).

Eukaryotic philosophy is redundant, they possess complex, functionally linked mechanisms and certain luxuries that allow them to ensure not only the good of the isolated organism but also of entire populations (cooperative evolution).

Prokaryotes are therefore designed to fit as much machinery, information and qualities as they can fit into a small space.

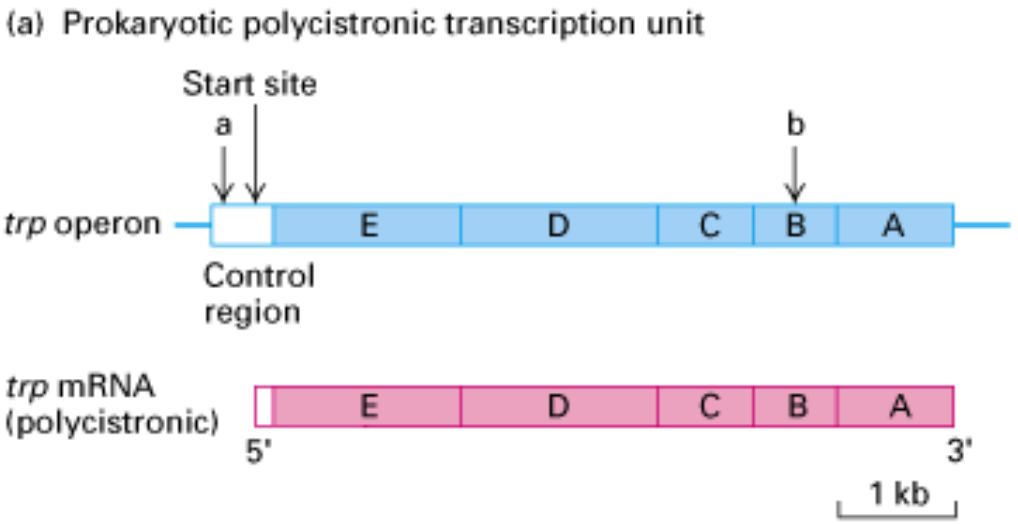
Eukaryotes in general are designed to adopt biological innovations that make them survive more successfully.



Transcriptional units in eukaryotes vs prokaryotes

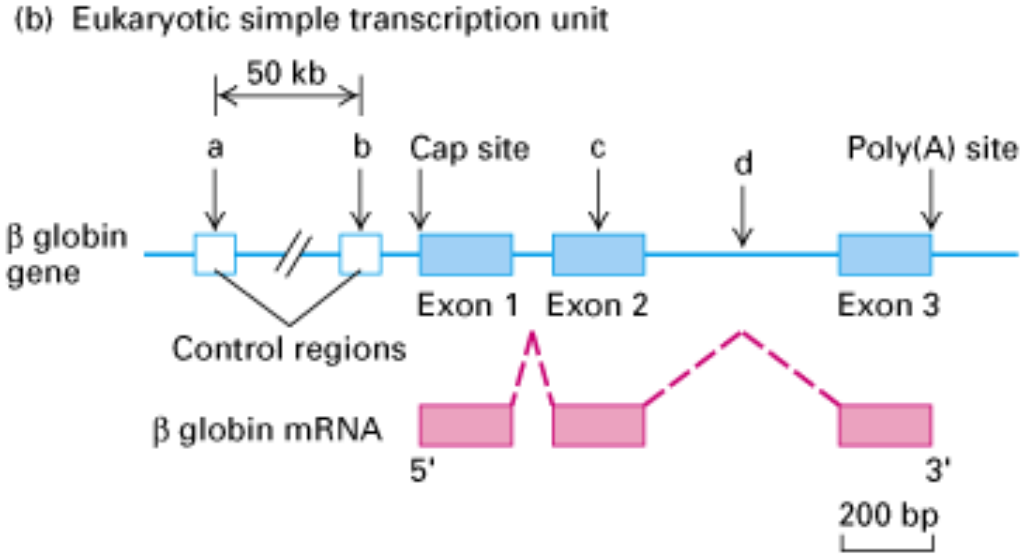
Prokaryotes

Polycistronic, monoexonic



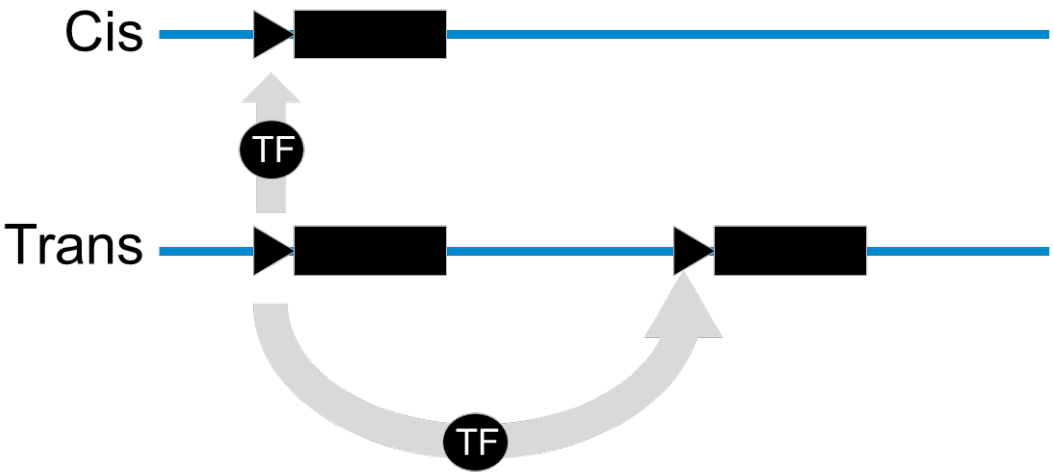
Eukaryotes

Monocistronic, polyexonic



Cis and Trans effects

An sequence has a *Cis* effect when it affects another sequence which is physically concatenated (same molecule.



An sequence has a *Trans* effect when it affects another sequence (physically concatenated or not) through an intermediary (RNA or protein).

Cis-regulatory element

Cis-regulatory elements (CREs) or modules (CRMs), 100 to 1000 bo long.

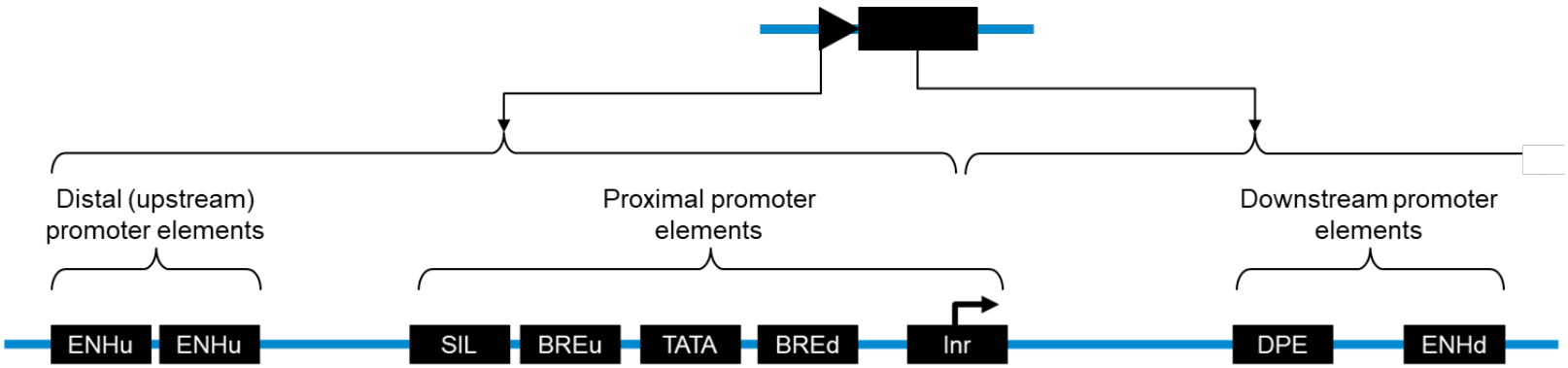
Regulate gene transcription by binding to transcription factors.

Found in the vicinity of the genes that they regulate.

A single transcription factor may bind to many CREs of many genes (pleiotropy).

Conversely, one gene can have several cis-regulatory modules.

Cis because they are located on the same DNA strand as the genes they control.



Cis-regulatory elements (Promoter)

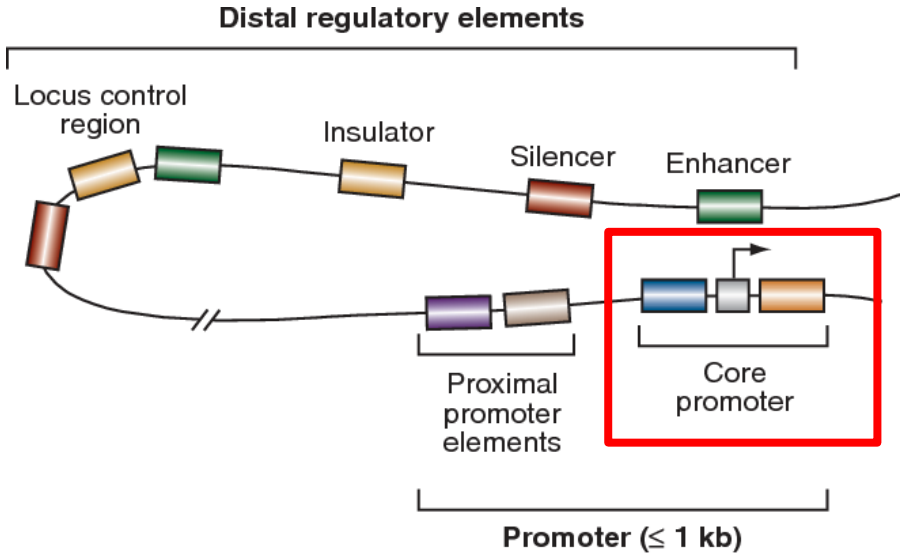
Short CREs which include transcription initiation site (Inr) and the 35 bp region upstream or downstream from Inr

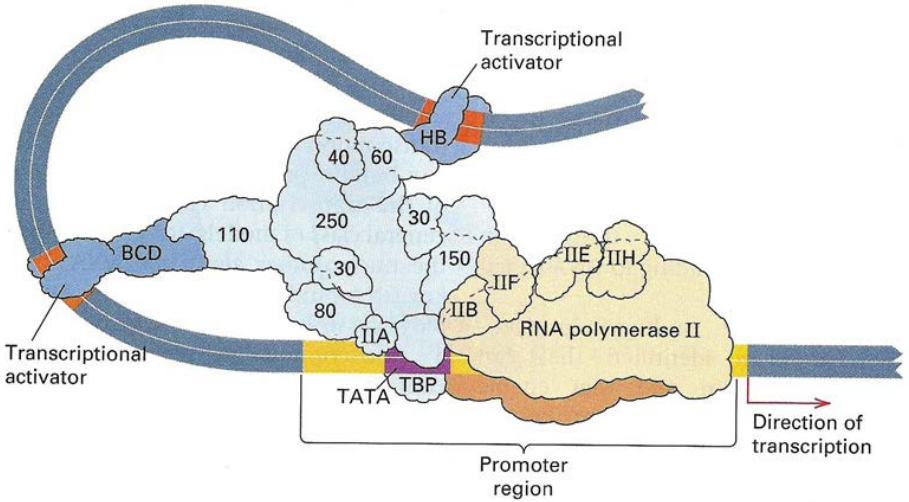
In eukaryotes, usually have:

TATA (Goldberg–Hogness) box
TFIIB recognition site
Initiator
Downstream core promoter element.

A single gene can contain multiple promoters.

Transcription factors (TFs) must bind sequentially to this region before RNA polymerase can begin transcription.





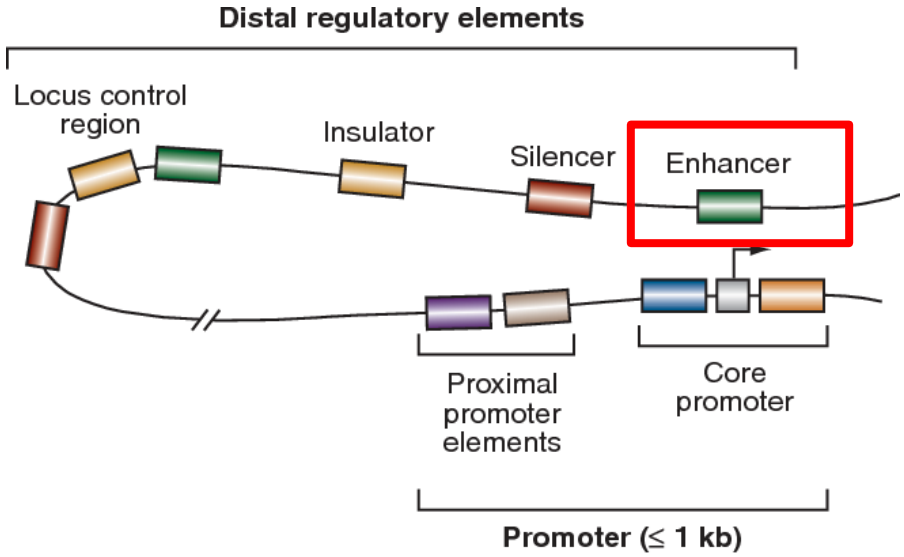
Cis-regulatory elements (Enhancer)

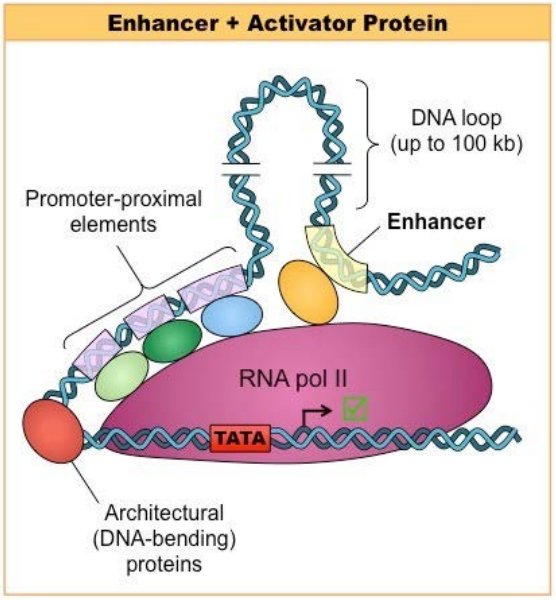
CREs that promote the transcription of genes on the same molecule of DNA.

Can be upstream, downstream, within the introns, or even in other distant genes.

Multiple enhancers can coordinate to regulate gene transcription.

Often transcribed to long non-coding RNA (lncRNA) or enhancer RNA (eRNA), whose quantity correlates with those of the target gene mRNA.



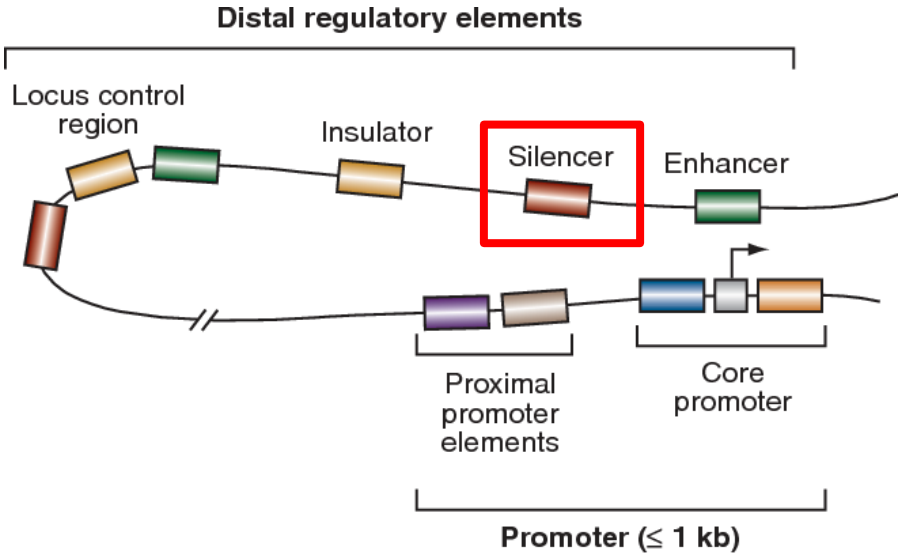


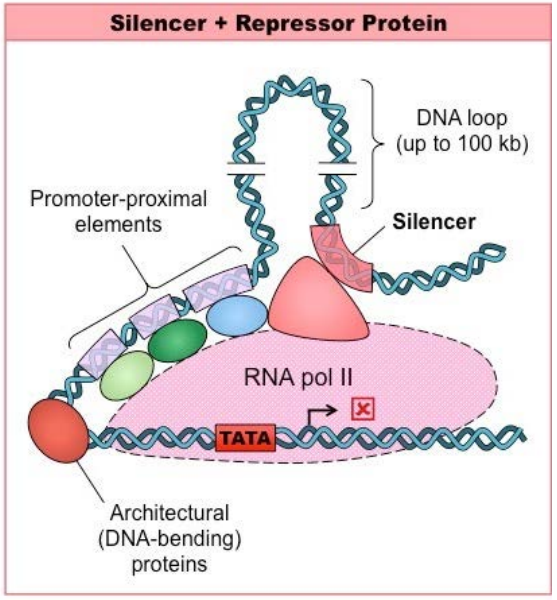
Cis-regulatory elements (Silencers)

CREs that can bind transcription regulation factors (proteins) called repressors.

Preventing transcription of a gene.

The term "silencer" can also refer to a region in the 3' untranslated region of messenger RNA, that binds proteins which suppress translation of that mRNA molecule.



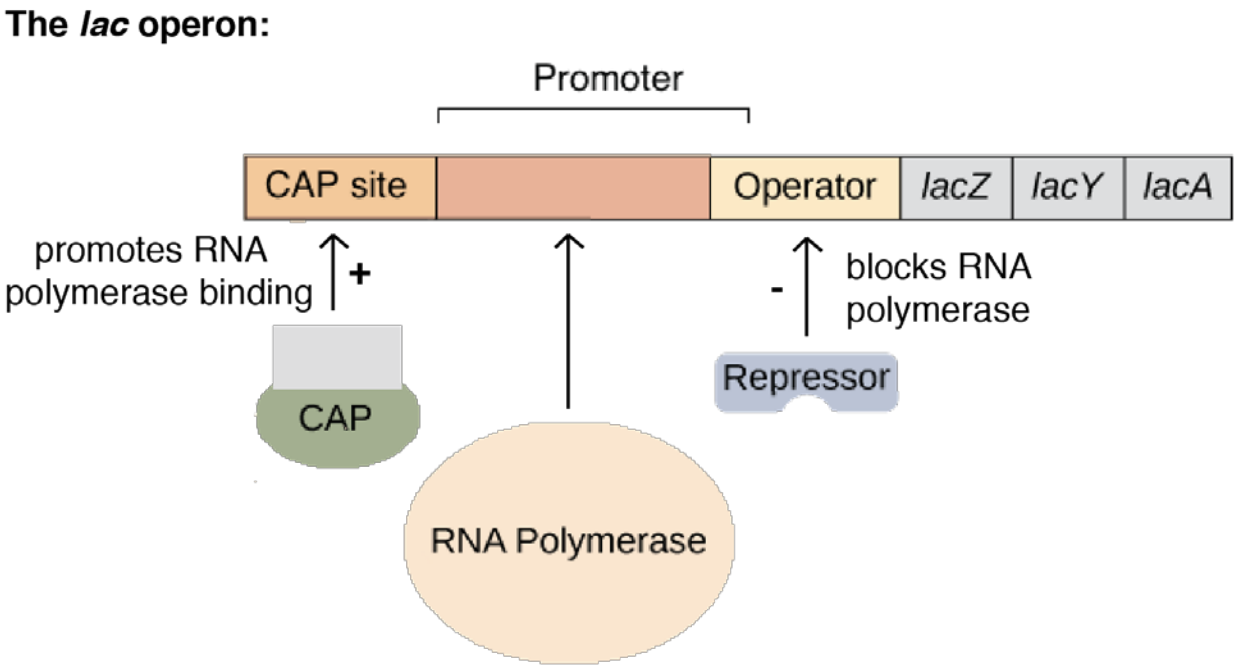


Cis-regulatory elements (Operators)

CREs in prokaryotes and some eukaryotes.

Exist within operons.

Can bind proteins called repressors to affect transcription.



Gene density

Average human gene density is low (6-10 genes/Mbp) compared to other organisms.

E. coli has a gene density of 950 genes/Mbp due to few non-coding regions.

Gene density is not uniform across the human genome.

Chromosomes 19 and 22 rich in euchromatin have a high gene density.

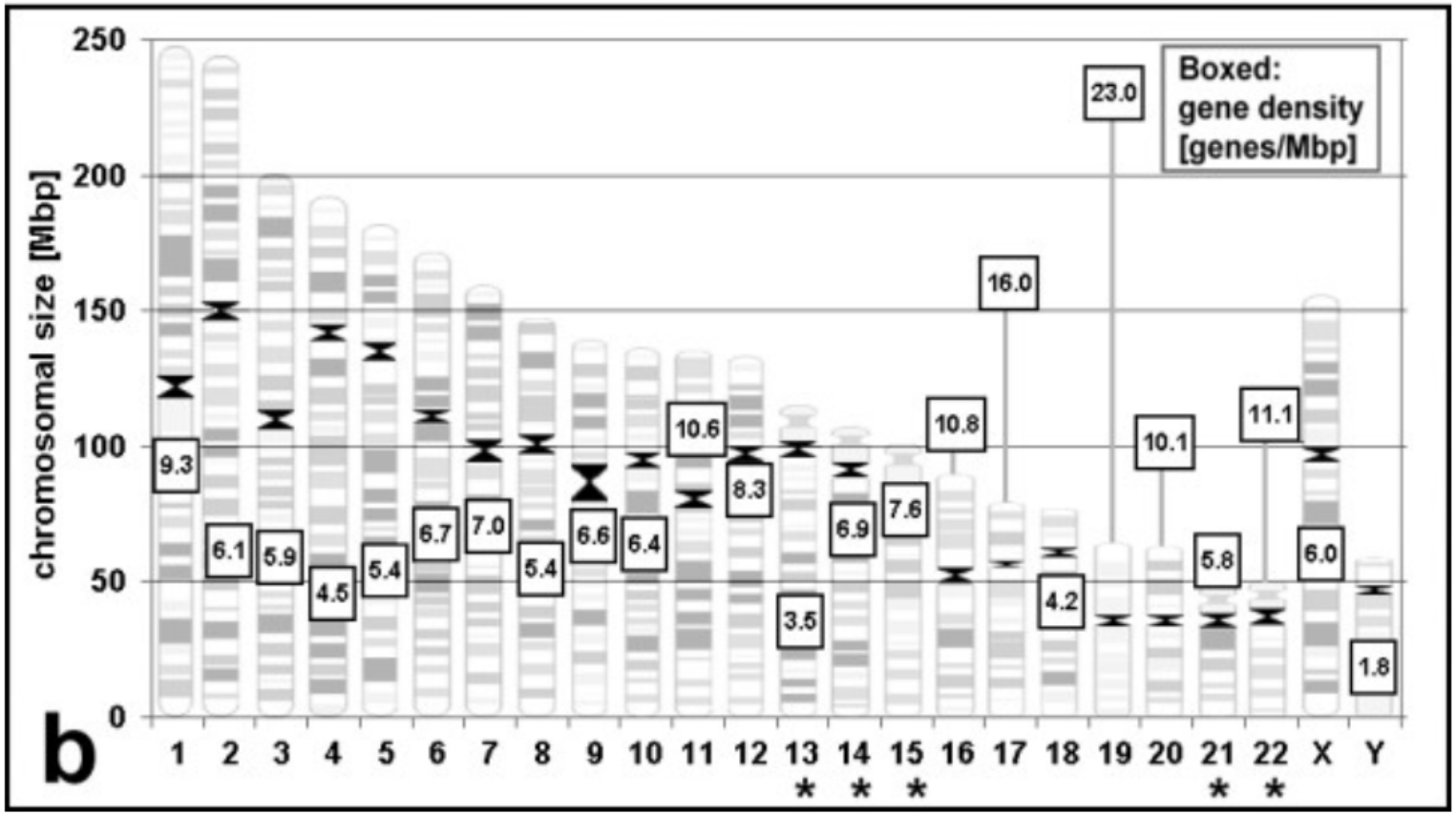
Chromosome 13 and Y have lower gene densities and more repetitive DNA.

The human genome has large intergenic (between genes) and intronic (within genes) regions, contributing to its lower gene density.

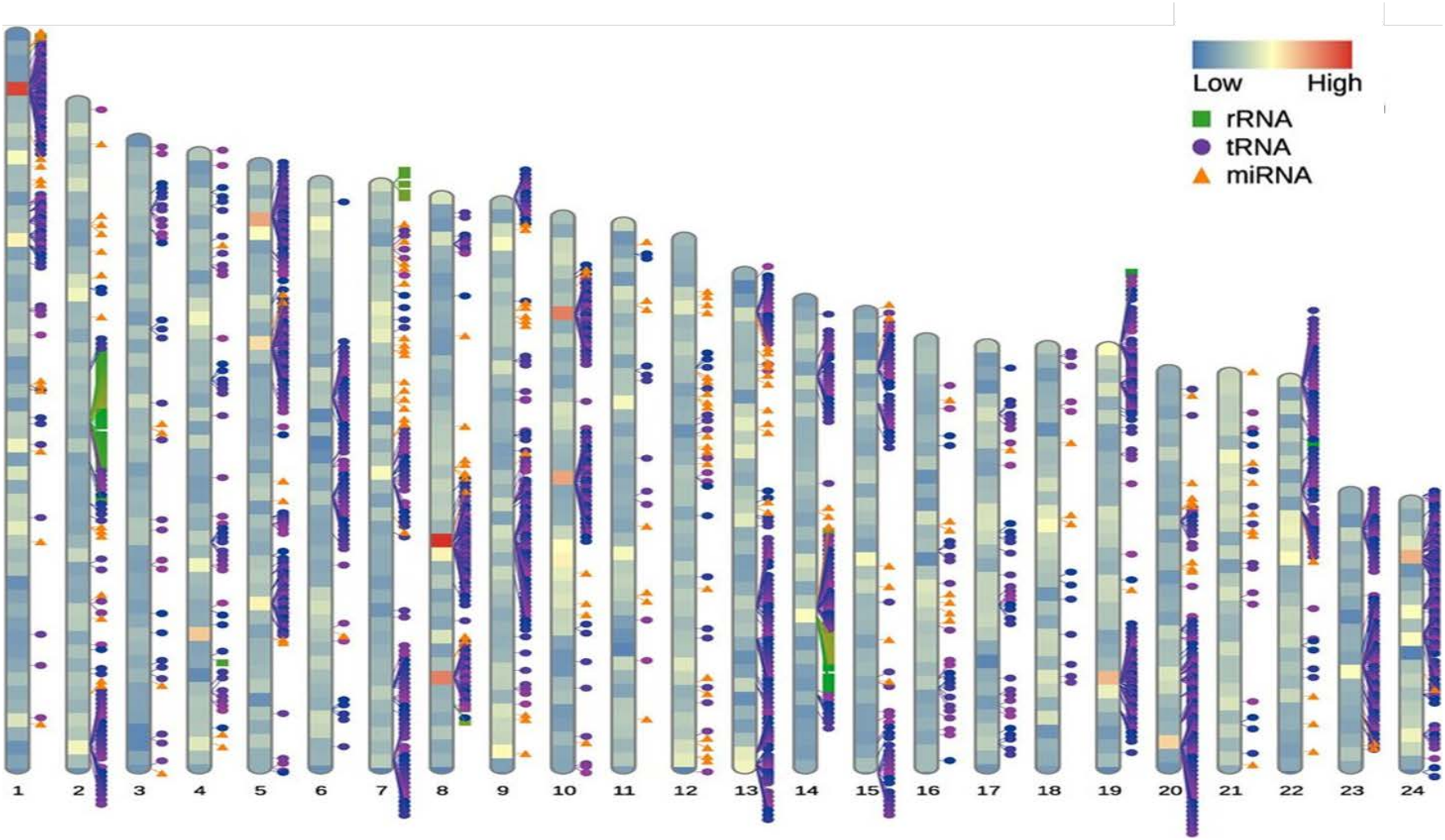
The lower gene density in the human genome reflects the complexity of regulatory mechanisms and the importance of non-coding DNA for gene regulation, cellular differentiation, and organismal development

Gene density

Human gene density by chromosome as genes/Mbp (indicated in a box).



rRNA, tRNA and mRNA gene density



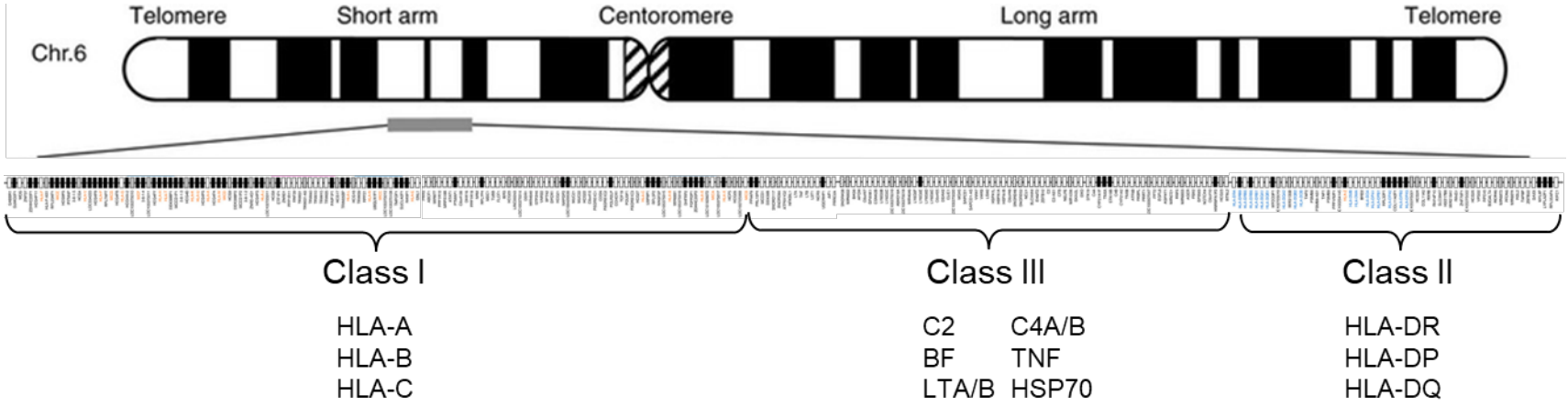
Major Histocompatibility Complex (MHC)

The MHC has one of the highest gene densities in our genome.

Located on 6p21.3 spans ~ 3.6 million base pairs.

Gene Count over 200 genes, most involved in immune system function, especially antigen processing and presentation.

Gene density of ~ 55 genes per Mb.

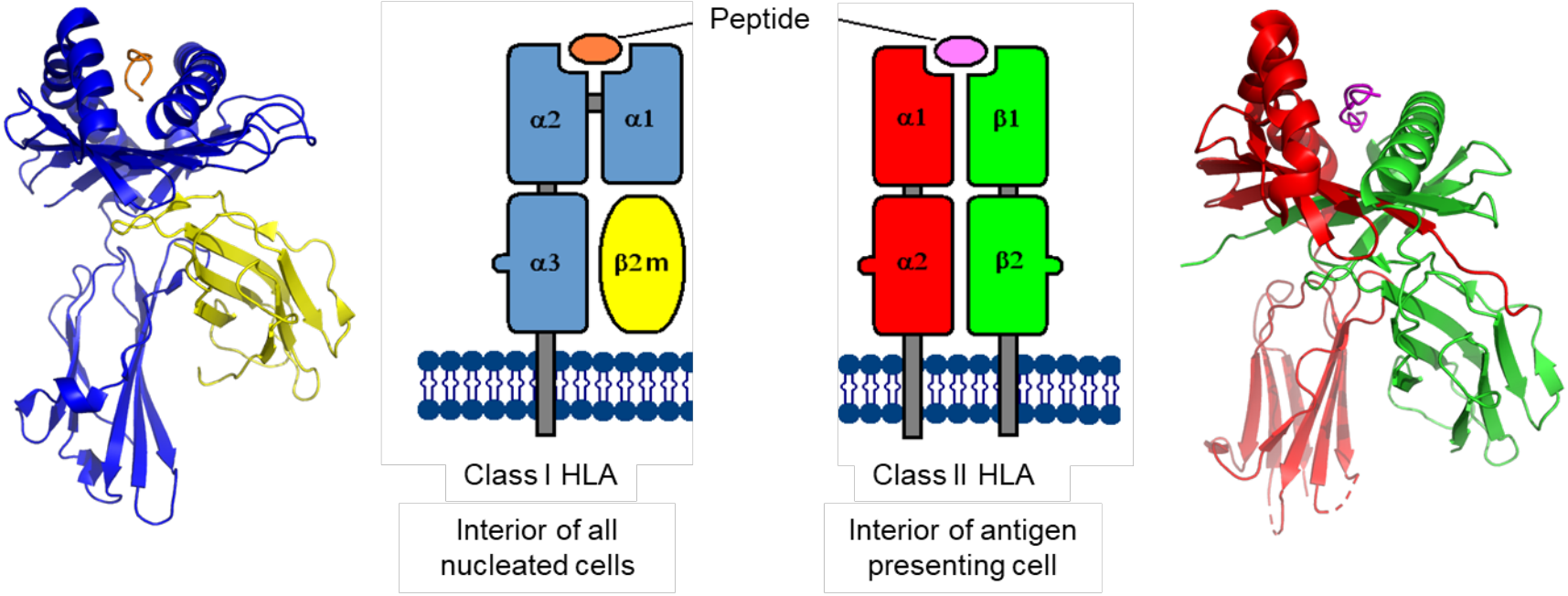


Major Histocompatibility Complex (MHC)

Class I MHC genes (HLA-A, HLA-B, HLA-C)
Present antigens from nucleated cells to cytotoxic T cells.

Class II MHC genes (HLA-DP, HLA-DQ, HLA-DR)
Present extracellular antigens to helper T cells.

Class III MHC genes (TNF-a/-b, C4 and C2)
Includes genes involved in the complement system and other immune functions.



Major Histocompatibility Complex (MHC)

The MHC is polygenic, polymorphic and codominant.

One of the most polymorphic areas of the genome, high rate of genetic variation between individuals.

Diversity enhances immune system adaptability.

The high gene density and polymorphism in the MHC region reflect its importance in immune surveillance and response.

The dense clustering of immune-related genes allows for coordinated expression and regulation, essential for rapid and effective immune responses.

IPD-IMGT/HLA Release 3.58 (Oct 2024)

The IPD-IMGT/HLA Database provides a specialist database for MHC sequences

Official WHO Nomenclature Committee For Factors of the HLA System.

Number of HLA Alleles	
HLA class I alleles	28,062
HLA class II alleles	12,375
All HLA alleles	40,437
Other non-HLA alleles	996

Other non-HLA Genes					
Gene	*HFE*	*MICA*	*MICB*	*TAP1*	*TAP2*
Alleles	6	570	295	19	106
Proteins	4	281	49	11	13

HLA class I

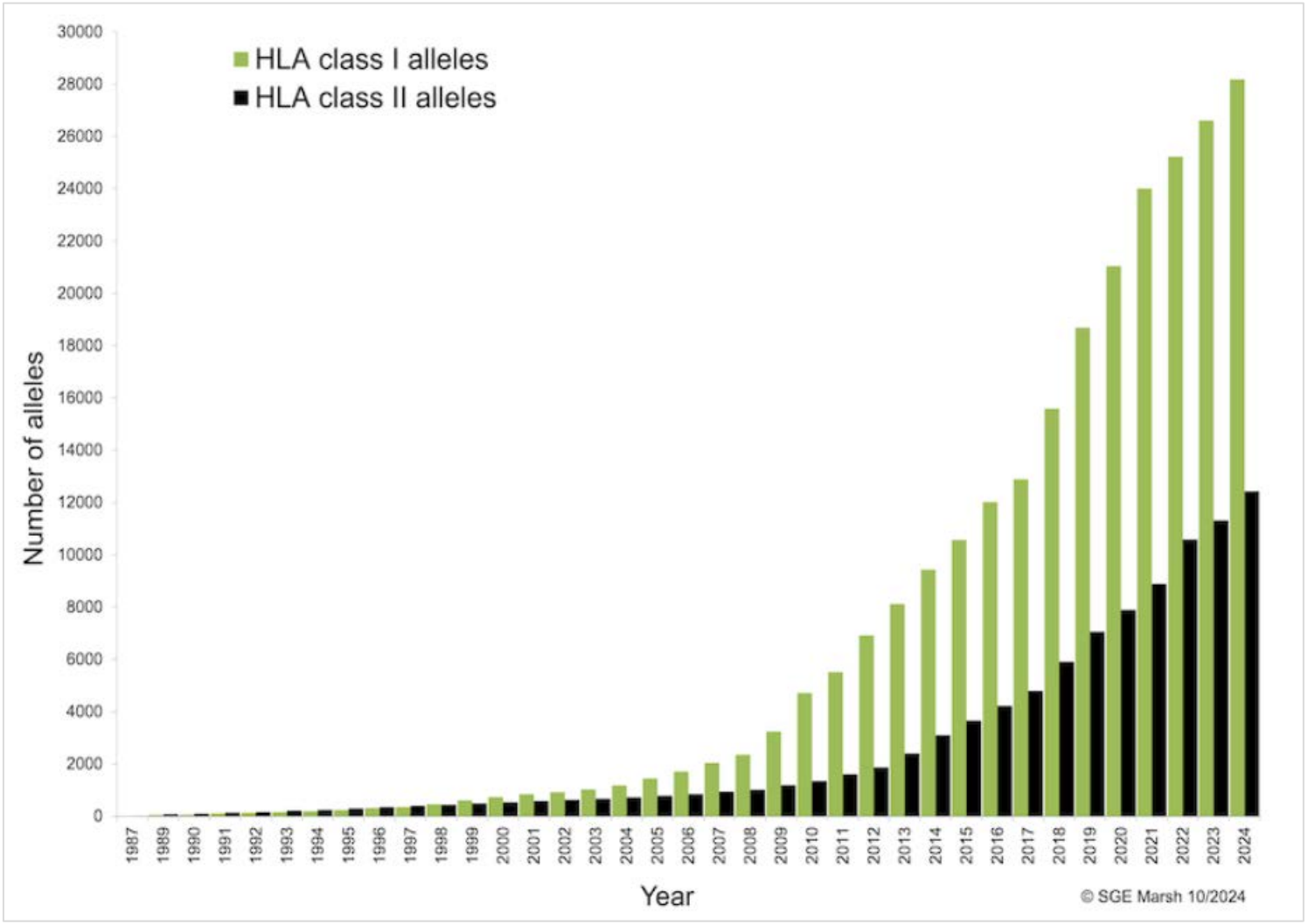
	Classical			Non-Classical		
Gene	*A*	*B*	*C*	*E*	*F*	*G*
Alleles	8,472	10,183	8,570	372	106	176
Proteins	4,949	6,073	4,723	141	19	52

HLA class II

Gene	*DRA*	*DRB*	*DQA1*	*DQA2*	*DQB1*	*DQB2*	*DPA1*	*DPA2*	*DPB1*	*DPB2*	*DMA*	*DMB*	*DOA*	*DOB*
Alleles	78	4,752	847	42	2,771	41	740	6	2,762	7	60	85	113	71
Proteins	17	3,139	431	11	1,657	9	361	0	1,583	0	9	9	16	17

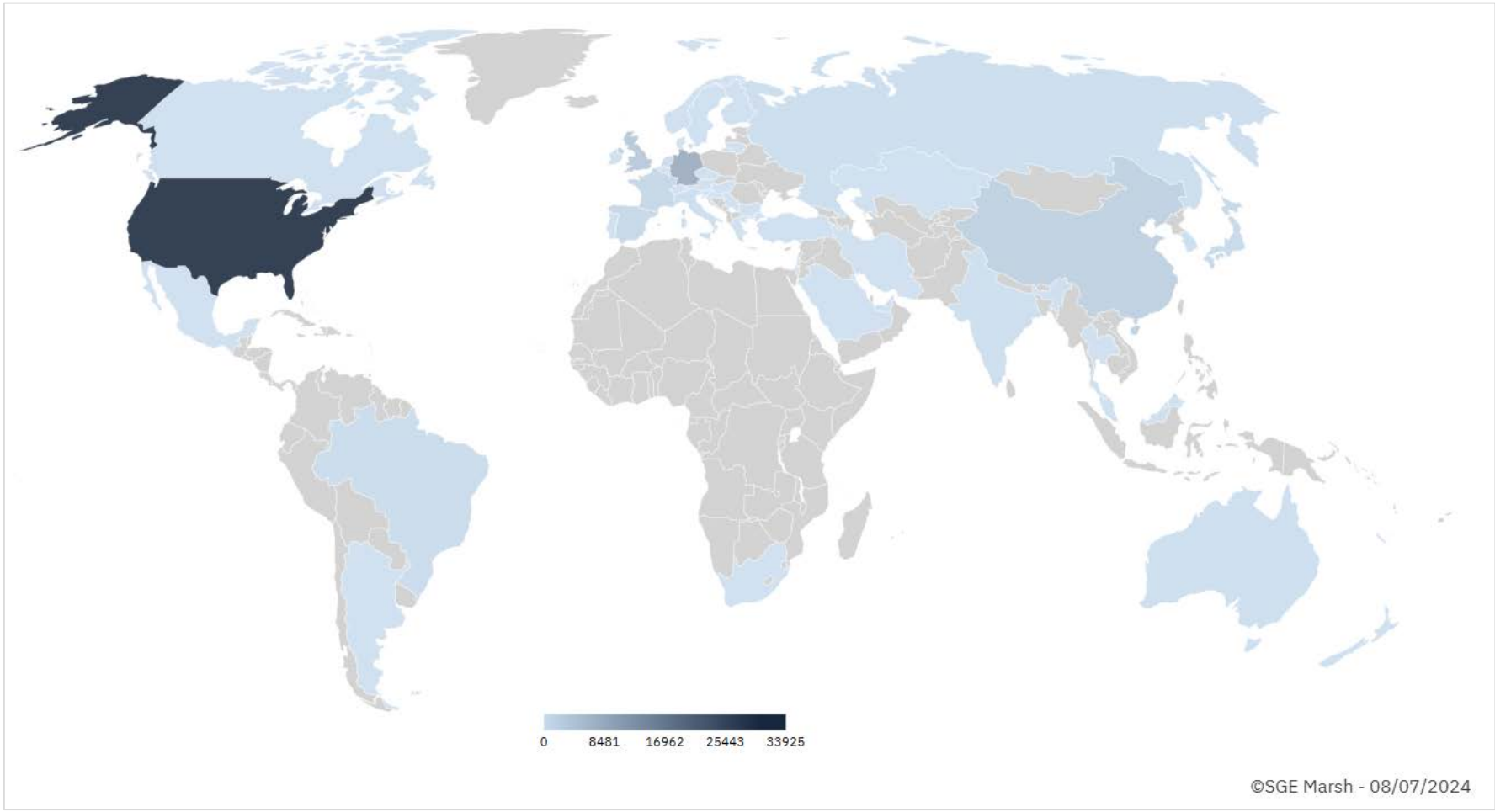
IPD-IMGT/HLA Release 3.58 (Oct 2024)

Number of HLA alleles identified from 1987 up to 2024-10-09.



IPD-IMGT/HLA Release 3.58 (Oct 2024)

Map of IPD-IMGT/HLA database submissions.



Polymorphism

Classical Mendelian genetics only distinguished two types of genes: Wild-type (normally circulating) and Mutant (the least common, initially the one that produced a disease or phenotypic change).

Today we know that some genes have different variants that may or may not produce phenotypic changes or disease, so they are not really mutants = alleles.

In some instances it is not correct to use the term “wild-type” (HLA).

Genetic polymorphism = refers to the existence of multiple alleles for a gene.

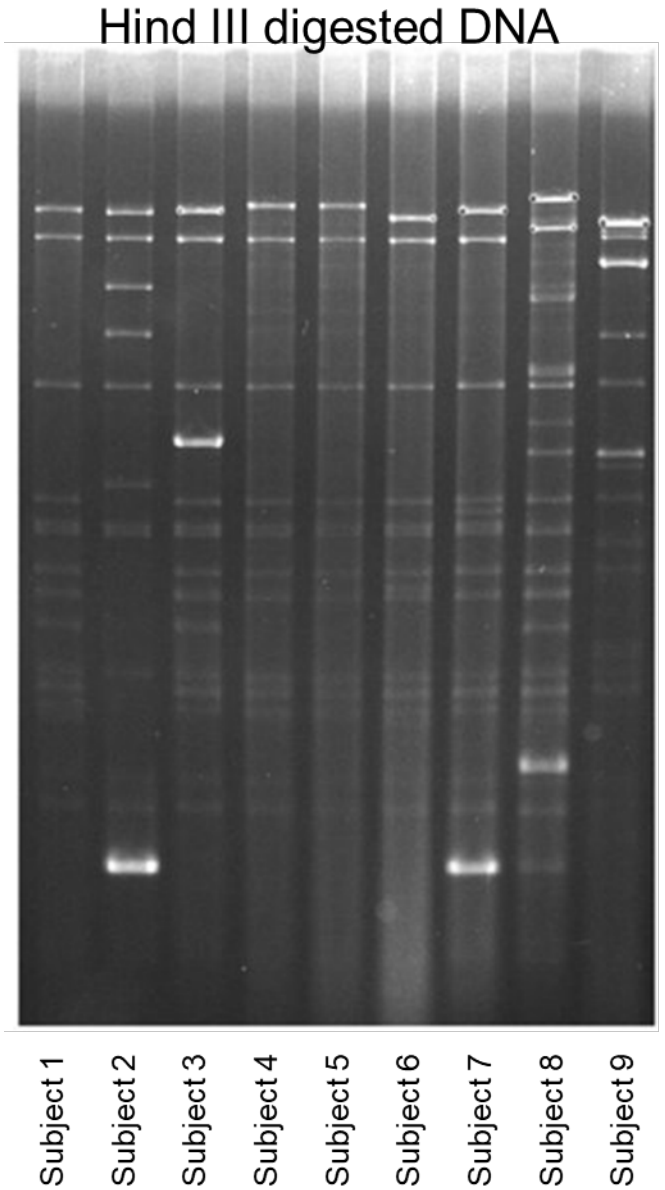
A mutation is considered polymorphism when it is found in more than 1% of the population.

Why more than 1%? Because the genetic drift that governs evolution gives rise to new alleles all the time, not all of them are important because not all of them stabilize their existence in a population (population fixation).

Restriction Fragment Length Polymorphism (RFLP)

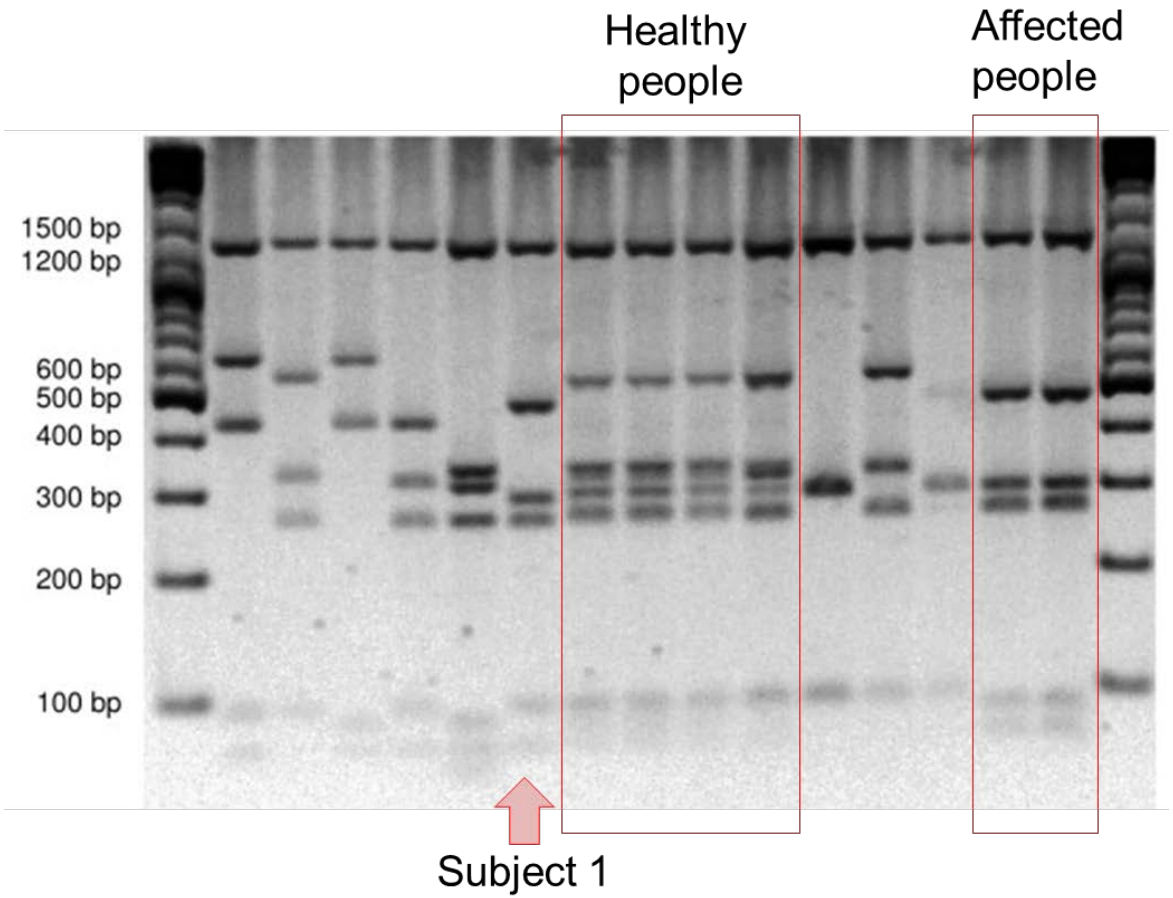
Polymorphisms can modify restriction sites, a fact that is exploited for the production of Restriction Fragment Length Polymorphism (RFLP) Maps.

ENZYME	ORGANISM	RECOGNITION SEQUENCE*
*Eco*RI	*Escherichia coli*	GAATTC
*Bam*HI	*Bacillus amyloliquefaciens*	GGATCC
*Bgl*II	*Bacillus globigii*	AGATCT
*Pvu*I	*Proteus vulgaris*	CGATCG
*Pvu*II	*Proteus vulgaris*	CAGCTG
*Hind*III	*Haemophilus influenzae* R_d	AAGCTT
*Hin*fI	*Haemophilus influenzae* R_f	GANTC
*Sau*3A	*Staphylococcus aureus*	GATC
*Alu*I	*Arthrobacter luteus*	AGCT
*Taq*I	*Thermus aquaticus*	TCGA
*Hae*III	*Haemophilus aegyptius*	GGCC
*Not*I	*Nocardia otitidis-caviarum*	GCGGCCGC
*Sfi*I	*Streptomyces fimbriatus*	GGCCNNNNNGGCC



Restriction Fragment Length Polymorphism (RFLP)

Polymorphisms can modify restriction sites, a fact that is exploited for the production of Restriction Fragment Length Polymorphism (RFLP) Maps.



Interrupted nature of eukaryotic genes

Unlike prokaryotic genes, which are continuous, eukaryotic genes are typically fragmented.

Requiring a process called RNA splicing to remove introns from the pre-RNA.

Machinery must then join exons together to form a mature mRNA.

1. Protects information.

2. Allows for alternative splicing, to produce multiple protein isoforms from a single gene.

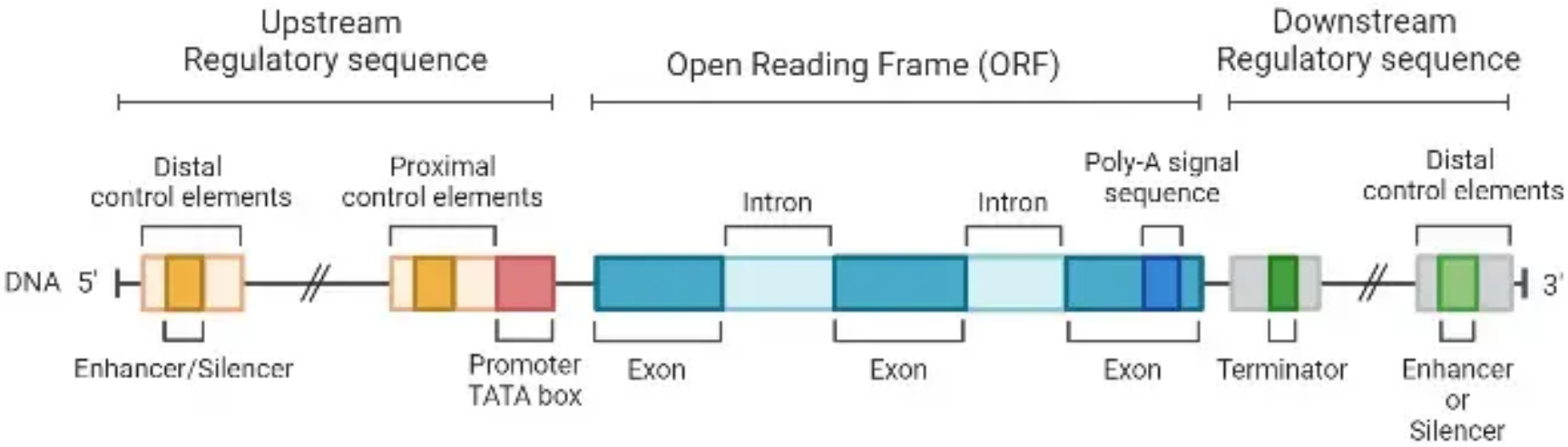
3. Allows exon shuffling and recombination, which can create new gene variants.

Exons and Introns

An exon is any part of a gene that will form a part of the final mature RNA after introns have been removed by RNA splicing.

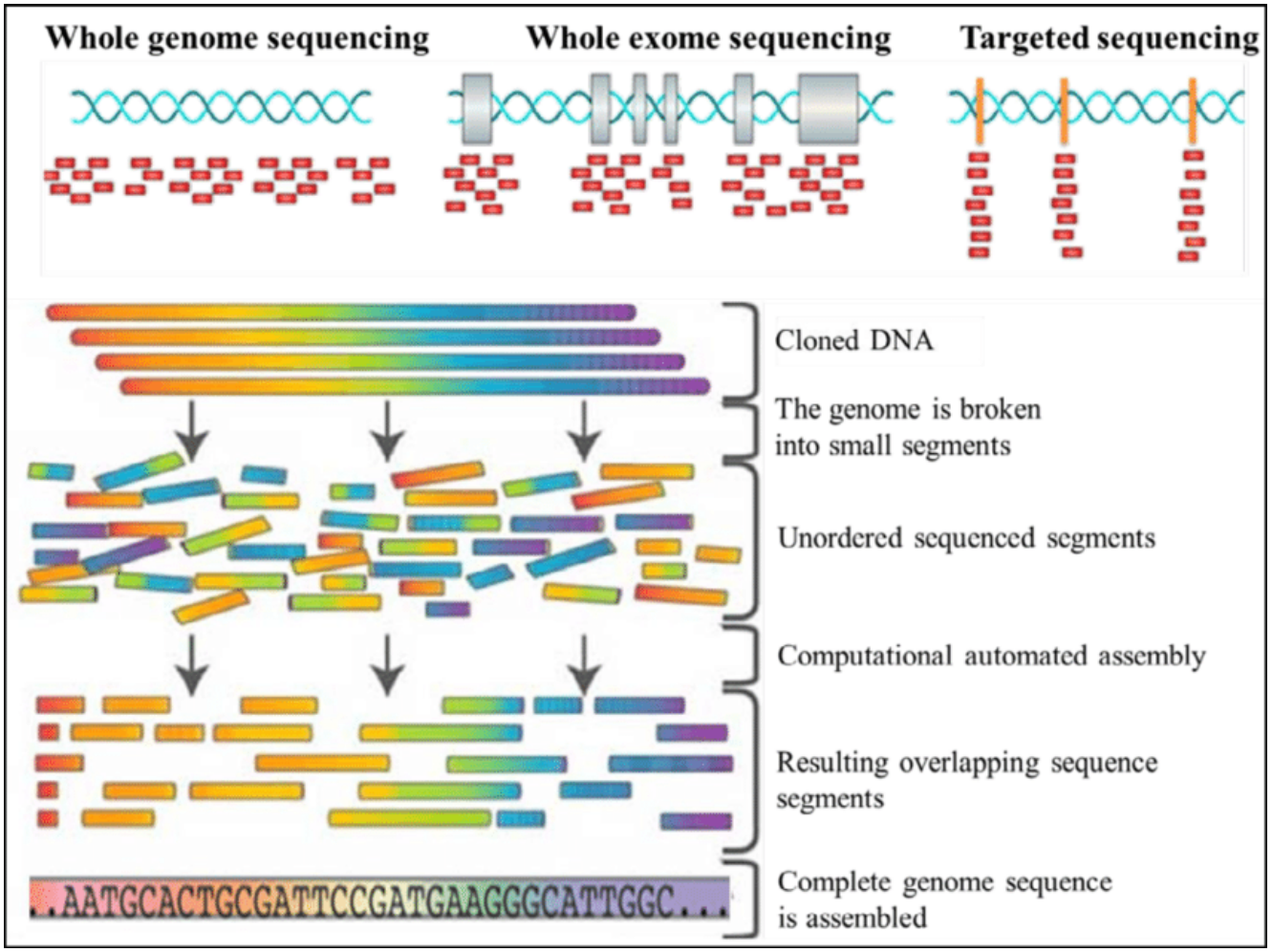
Applies to both the DNA sequence within a gene and the corresponding RNA sequence.

In RNA splicing, introns are removed and exons are covalently joined to one another.



Exome

Just as the entire set of genes for a species constitutes the genome, the entire set of exons constitutes the exome.



Exons and Introns

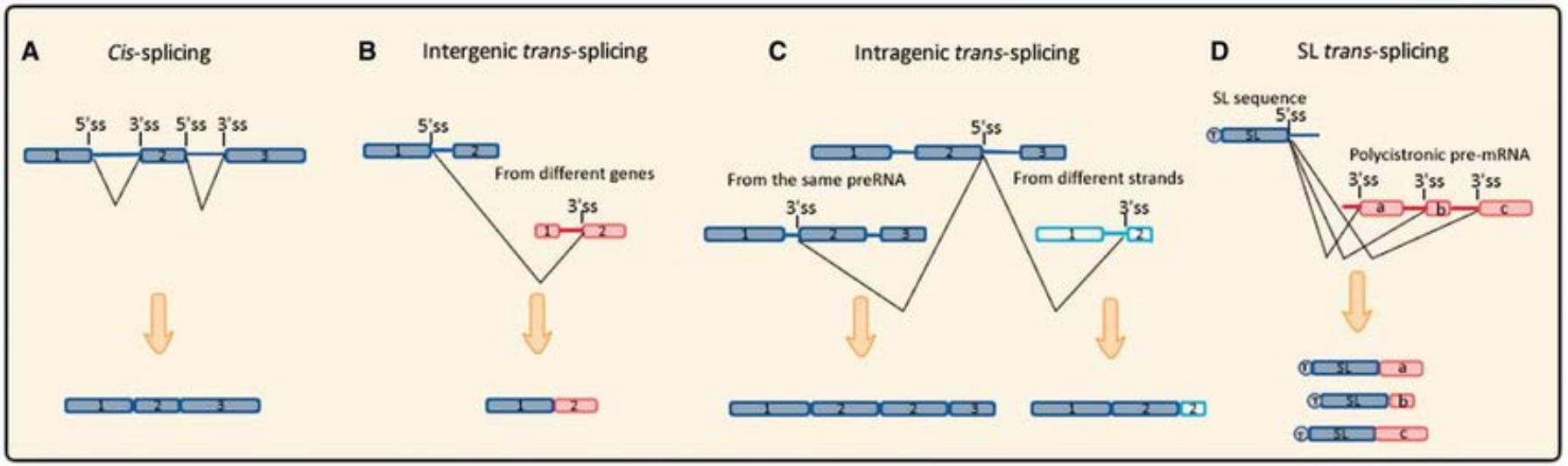
Terms coined by American biochemist Walter Gilbert in 1978.

"…a transcriptional unit contains regions which will be lost in the mature messenger – which I suggest we call introns (for intragenic regions) – alternating with regions which will be expressed – exons."

Originally used for protein-coding transcripts (mRNA) that are translated.

Later included sequences removed from rRNA and tRNA, and other ncRNA.

Currently includes RNA molecules originating from trans-splicing.



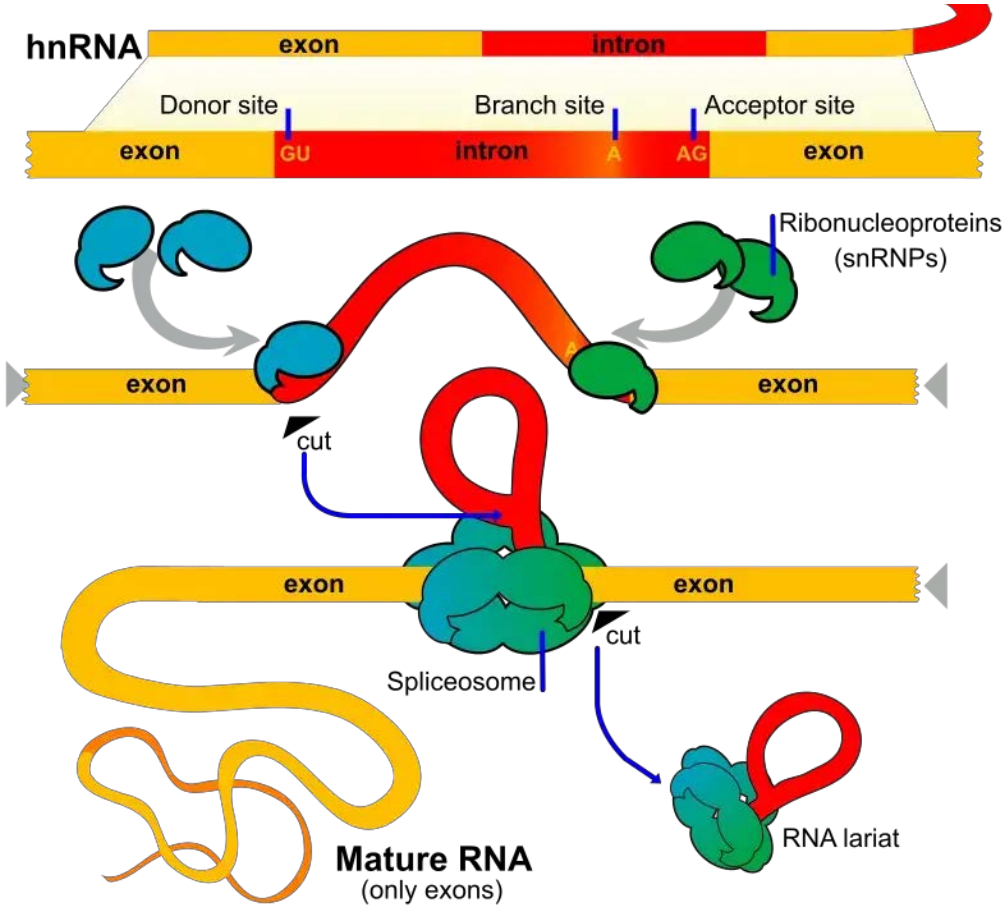
RNA *cis*-splicing

Gene transcription generates pre-mRNAs containing exons & introns.

Introns are removed from the pre-mRNA and adjacent exons are joined together in a highly orchestrated fashion.

When this process occurs within one pre-mRNA, it is called pre-mRNA cis-splicing.

Catalyzed by the spliceosome, an enzymatic complex of many proteins and four small ribonucleoprotein particles (snRNPs): U1 snRNP, U2 snRNP, U5 snRNP and U4/U6 snRNP.



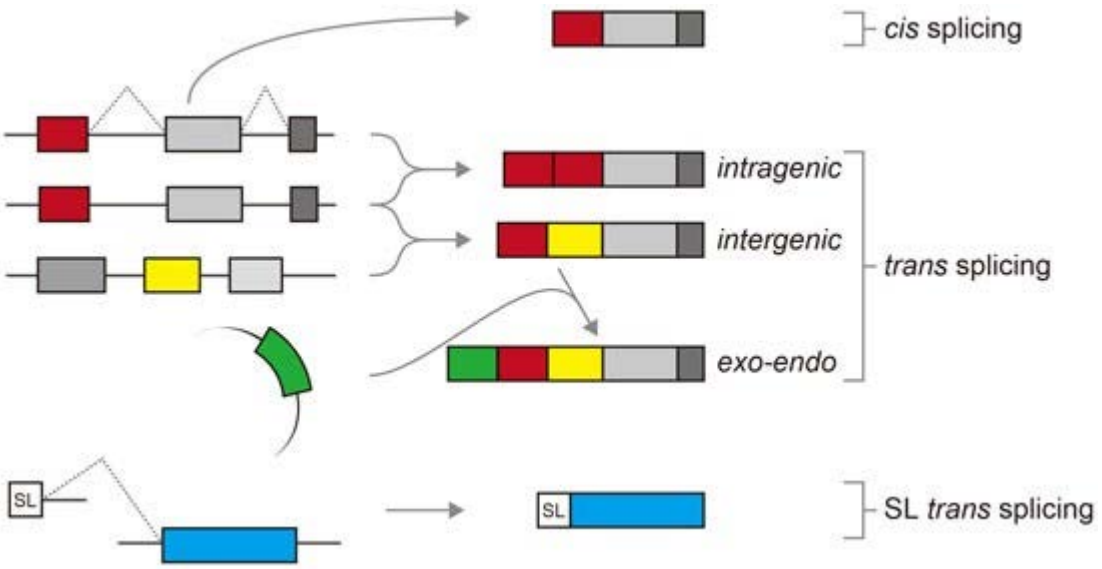
RNA *trans*-splicing

A second, and less well-known, form of natural RNA splicing.

Two independently transcribed RNA precursors form a composite mRNA.

Can expand the diversity of the transcriptome by creating novel mRNA and protein variants not encoded by a single gene.

Rare in mammals, common in trypanosomes and some nematodes,



SL (Spliced Leader) Trans-Splicing: A spliced leader (SL) RNA is added to the 5' end of various mRNAs.

Nuclear splicing

Process by which introns are removed from a pre-mRNA transcript and exons are joined together to produce a mature mRNA molecule.

Occurs within the nucleus of eukaryotic cells.

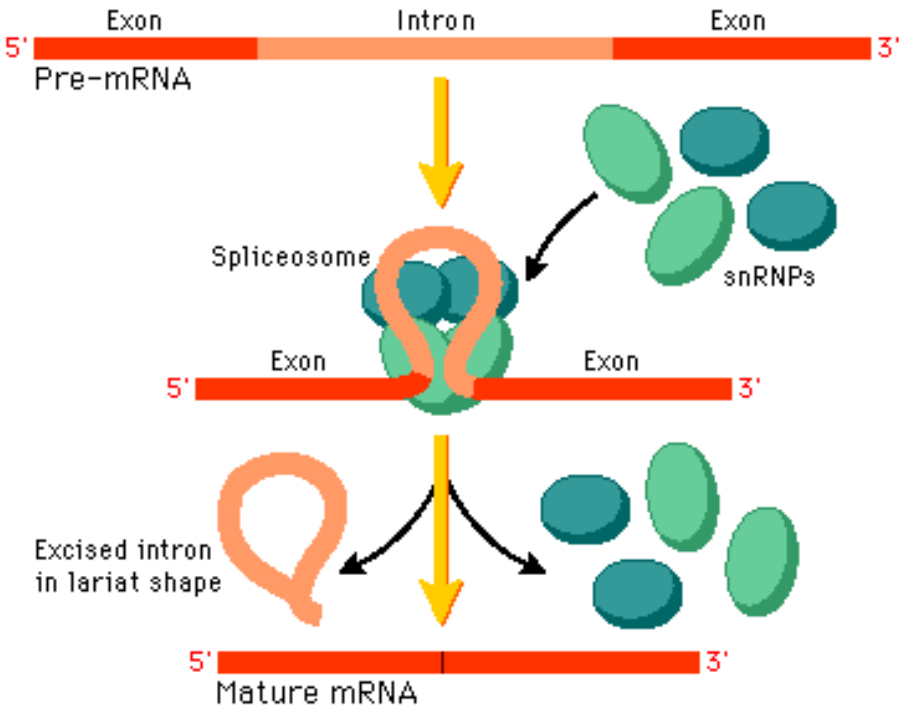
Essential for producing mRNA that can be translated into functional proteins.

The spliceosome recognizes branch-point nucleotide within the intron and forms a looped lariat structure.

Spliceosome excises introns.

Forms larita structure out excised intron.

Joins exons to create a mature RNA.



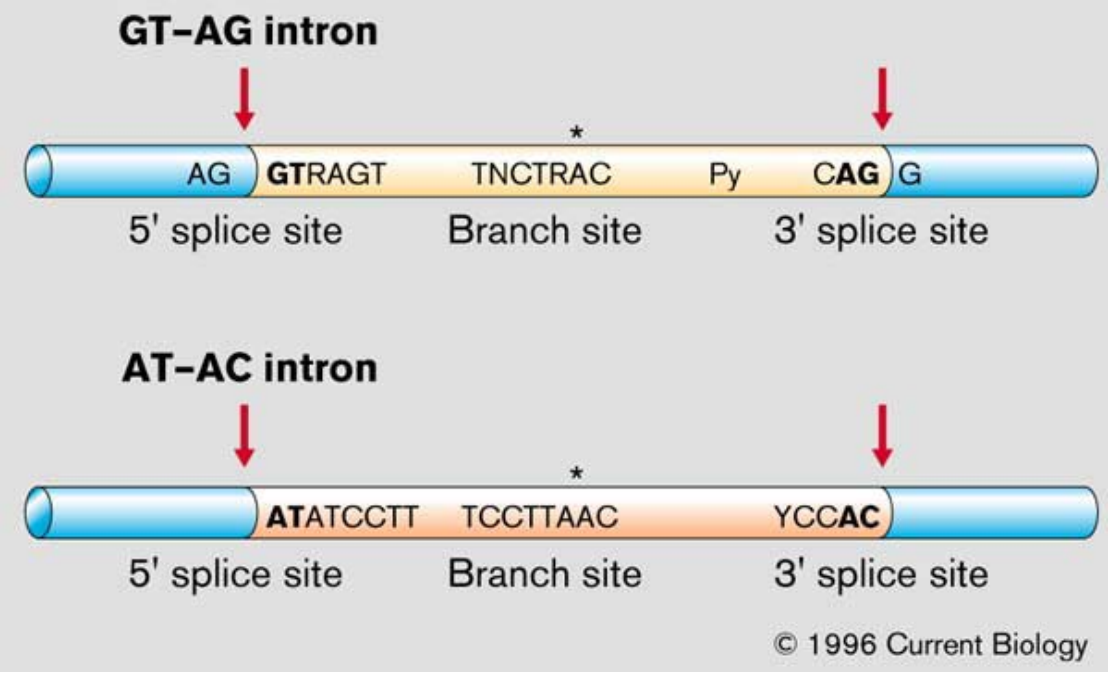
Splice sites and intron-exon boundaries

5' Splice Site (Donor Site) Located at the exon-intron boundary (GT or AT).

3' Splice Site (Acceptor Site): Located at the intron-exon boundary (AG or AC).

Branch Point: Adenine nucleotide located 20 to 50 bp upstream of the 3' splice site.

This branch-point A is essential for forming a lariat structure during splicing, which is an intermediate loop that helps the intron to be excised.



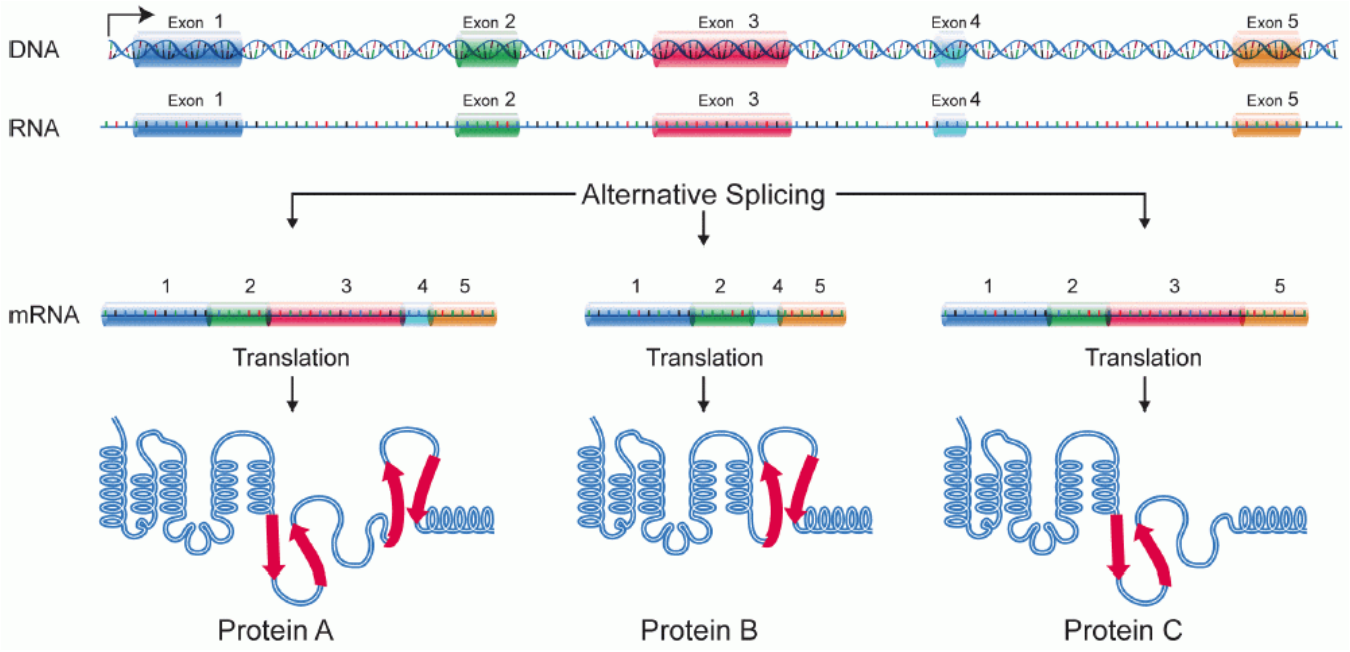
Alternative splicing

Normal eukaryote post-transcriptional process allowing a single gene to produce different variants.

Exons are joined in different combinations, leading to different splice variants.

Splice variants contain amino acid sequence and biological function differences.

In humans, ~95% of multi-exonic genes are alternatively spliced.



Isoforms:

A protein produced from the same gene but different in structure due to alternative splicing, post-translational modifications, or use of alternative promoters or start codons.

Alternative splicing modes

There are five alternative splicing modes.

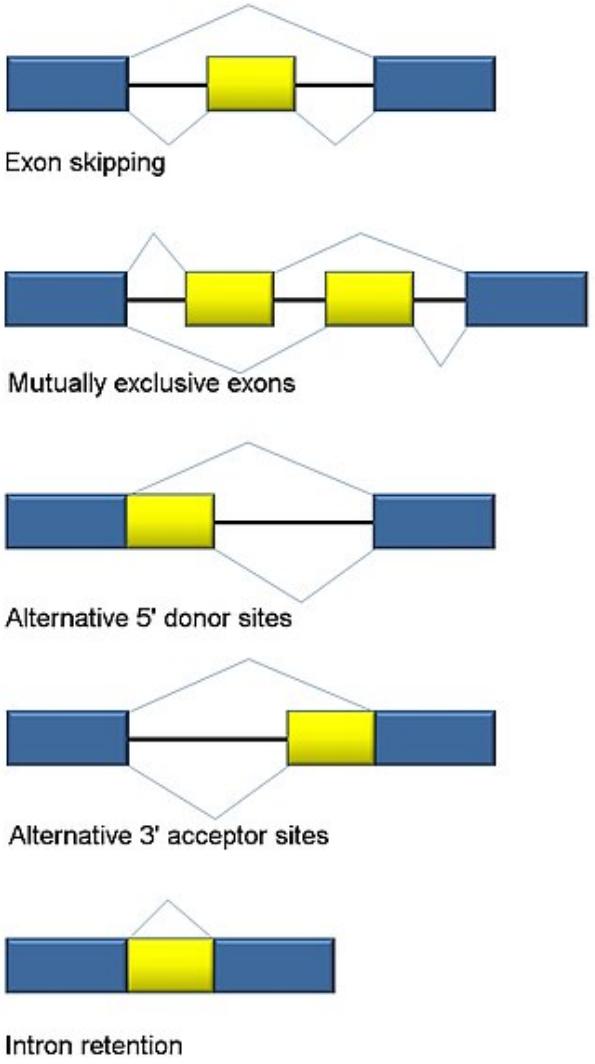
Exon skipping (cassette exon): An exon can either be included or excluded in the final mRNA. This is the most common type in mammals.

Mutually exclusive exons: Only one of two possible exons is included in the final mRNA, not both.

Alternative donor site: An alternative 5' splice site changes the 3' boundary of the upstream exon.

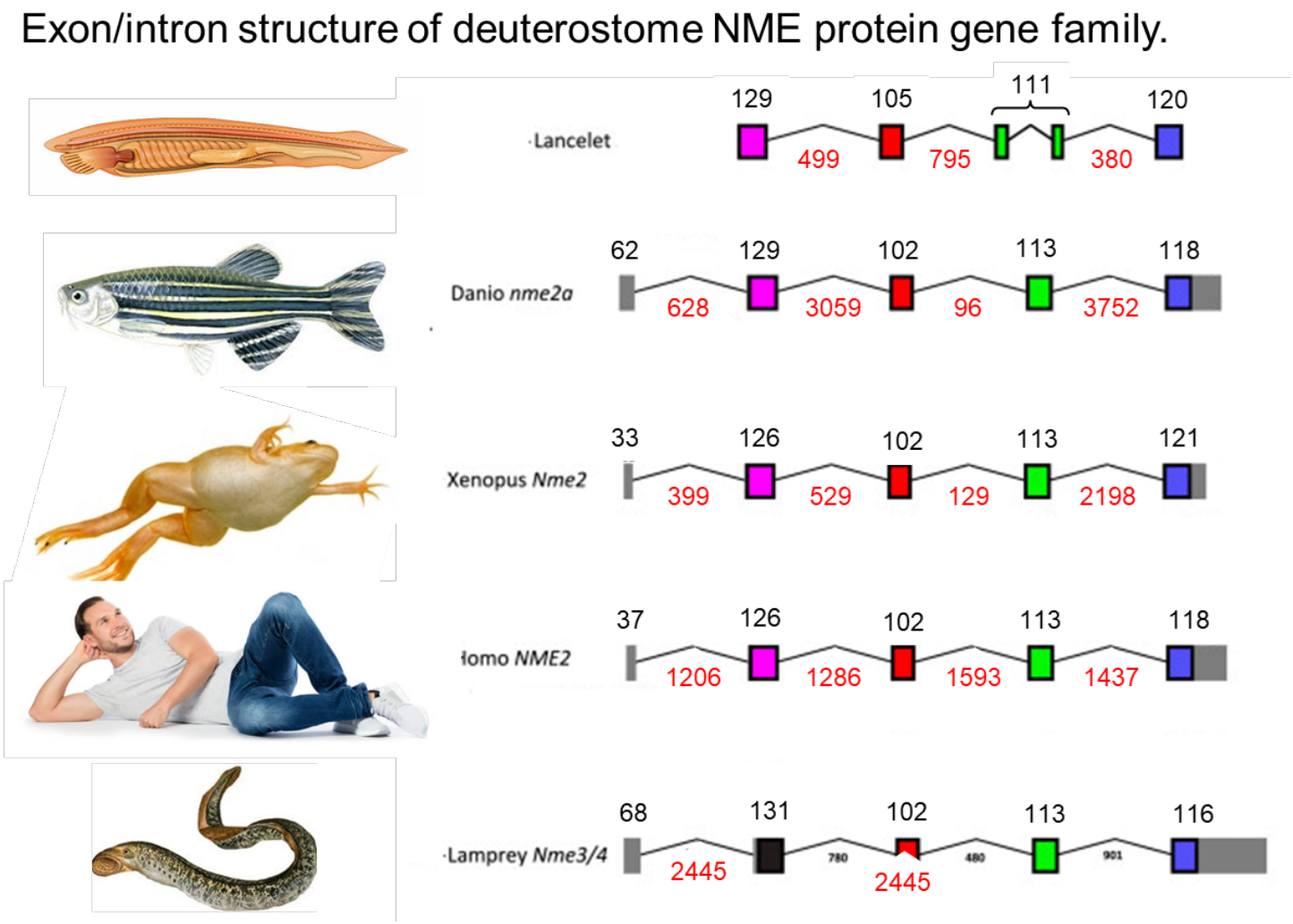
Alternative acceptor site: An alternative 3' splice site changes the 5' boundary of the downstream exon.

Intron retention: An intron may be retained rather than spliced out. If it's in a coding region, the sequence must remain in frame. This mode is rare in mammals but common in plants.



Wikimedia Common Alt splicing bestiary2.jpg

Exon conservation and intron variability



Desvignes T, *et al*. BMC Evol Biol. 2009 Oct 23;9:256.

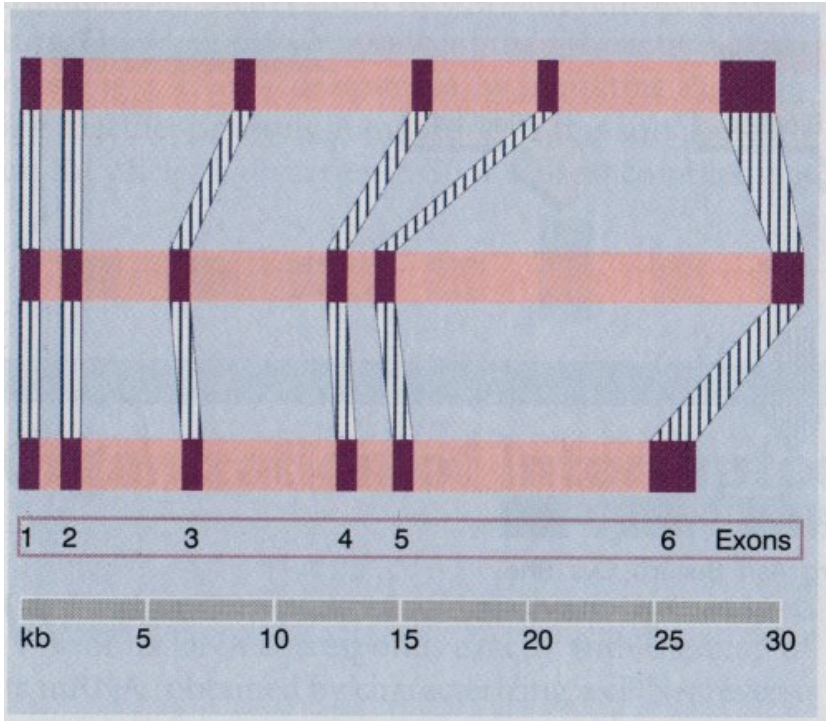
Exon conservation and intron variability

A classic example is the Dihydrofolate Reductase (DHFR) gene of several mammals.

In the vast majority of mammals, the gene has 6 exons corresponding to the 2 kb of mRNA.

Individual genes (including introns) vary greatly between these species (from 25 to 31 kbp).

That is, the coding portion remains relatively the same while the size of the introns varies.



This is a reflection of evolution, phylogenetically related genes tend to have similar organizations: similar sizes and positions of introns and exons.

Exon conservation and intron variability

Animals have more exons per gene, shorter exons and longer introns.

Plants less exons per gene, longer exons and shorter introns.

Plant introns contain fewer transposable elements than animal introns

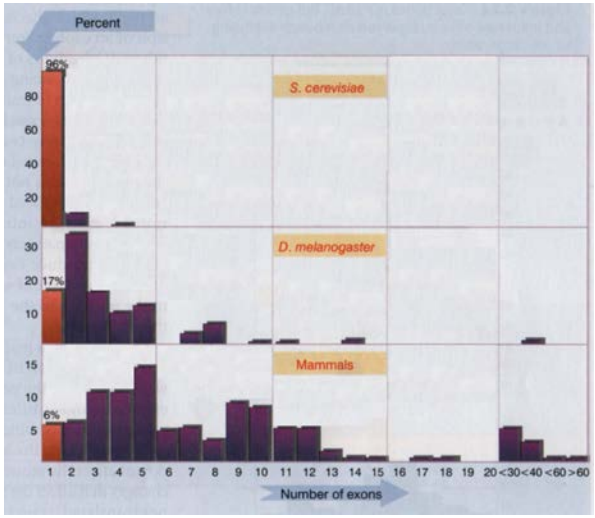
Du Y, *et al*. J Adv Res. 2024 Oct;64:83-98.

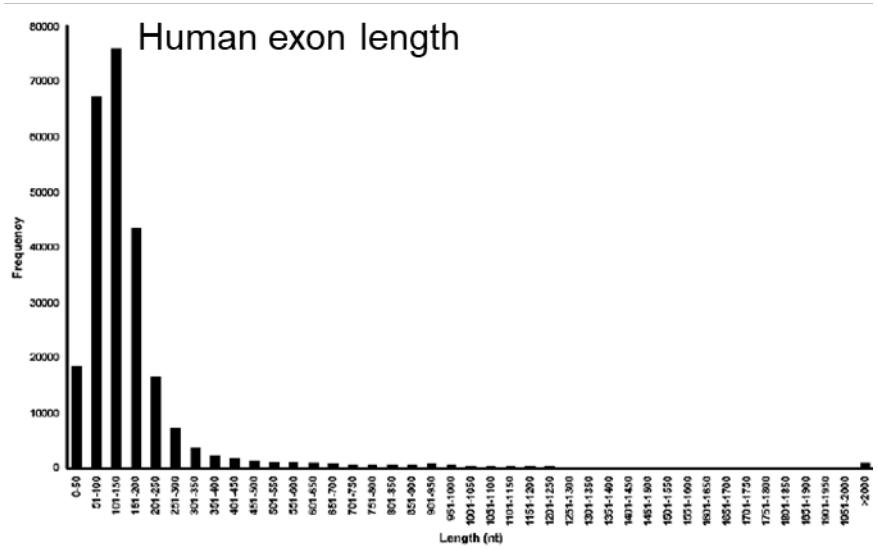
Human exon size and content

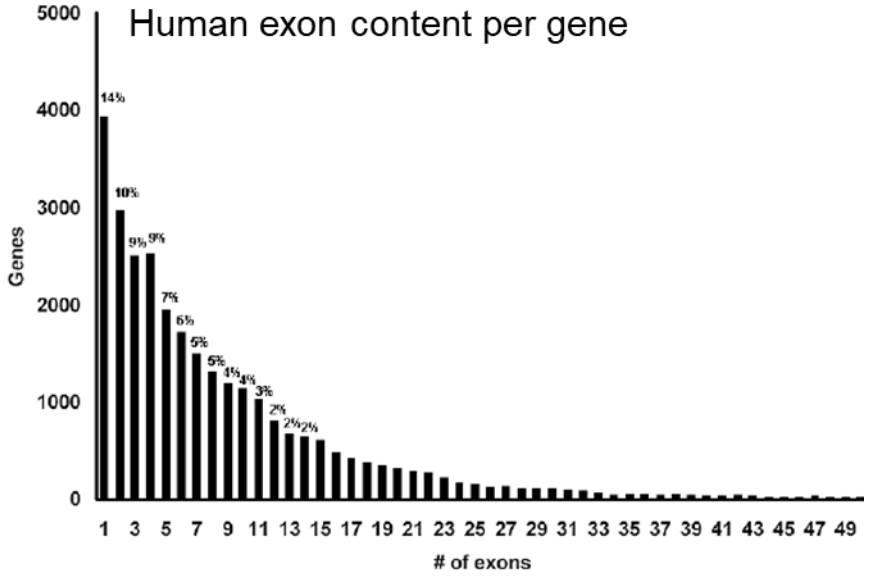
Most human exons between 50 and 350 bp in length.

Most human genes have between 1 and 5 exons.

Yeast have a significant number of monoexonic genes (<95%), this proportion is lower in insects (<20%) and much lower in mammals (5%).







Du Y, *et al*. J Adv Res. 2024 Oct;64:83-98.

Exons and Introns

The number of exons and introns in eukaryotic genes varies.

Eukaryote genes usually contain multiple exons (HoSa average of 8–10 per gene).

In some plants and simpler eukaryotes, genes may have only a single exon (*Saccharomyces cerevisiae*).

Others, like the dystrophin gene (associated with muscular dystrophy), have 79 exons.

Eukaryotic genes have several introns (HoSa average of 7–8 per gene).

In some plants and simpler eukaryotes, there may be fewer introns, or, in some cases, none.

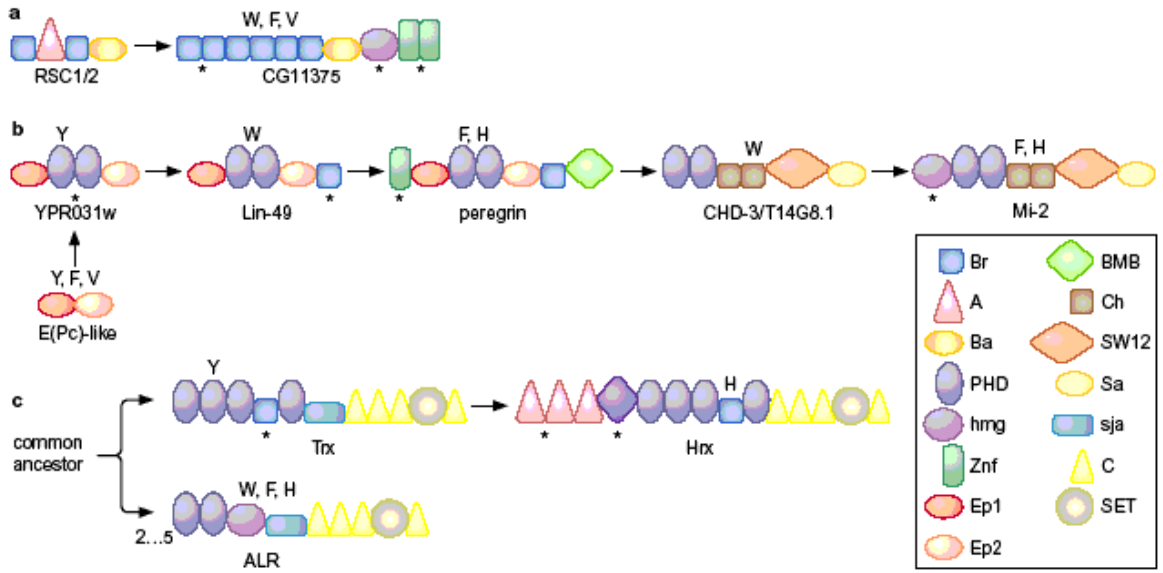
More complex organisms have more introns per gene (SaCe – ArTa- DroMe-HoSa)

Introns early hypothesis

Primordial genomes contained introns, but these were gradually eliminated in prokaryotes, eventually disappearing from this lineage to optimize resources and accelerate growth.

Bright side: Would have allowed highly specialized traits (proteins) to evolve through exon shuffling.

Ugly side: Difficult to explain success of eliminating all evidence of introns among prokaryotes.



INTERSTINGLY

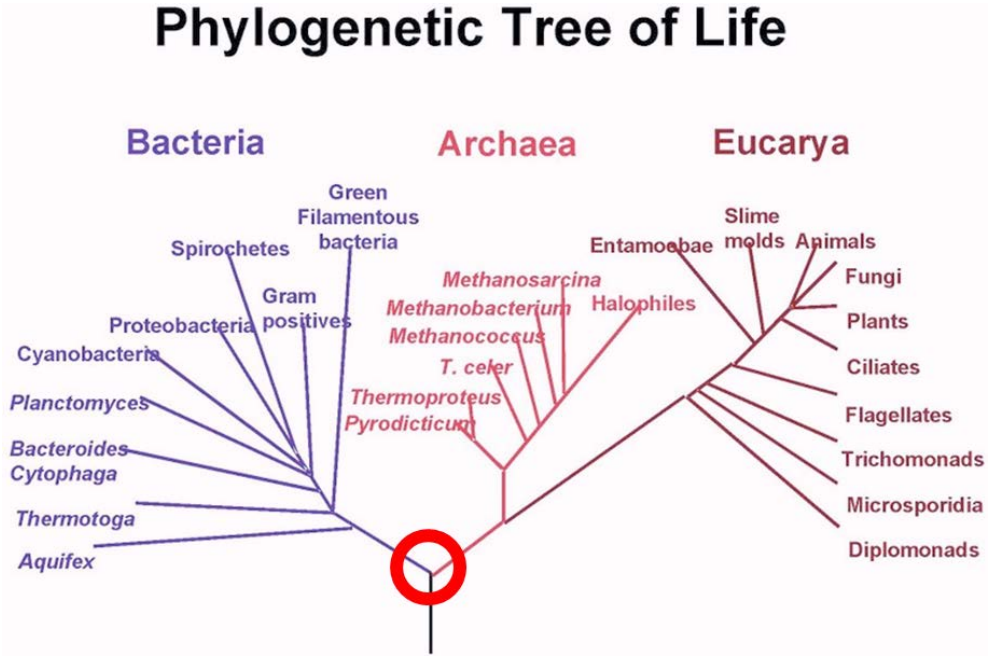
Mitochondria and chloroplasts, endosymbionts have nuclear introns in their genomes.

Introns late hypothesis

Eukaryotes incorporated introns after the prokaryotic/eukaryote speciation.

Theory does not explain the evolutionary process in such a sensual way but fits better than the other to the current evidence (nuclear introns only present in eukaryotes).

Proposes that introns arose from transposons or transferable elements that infected eukaryotic cells taking advantage of the boom in the creation of this new branch of the tree of life.



Overlapping reading frames (ORFs)

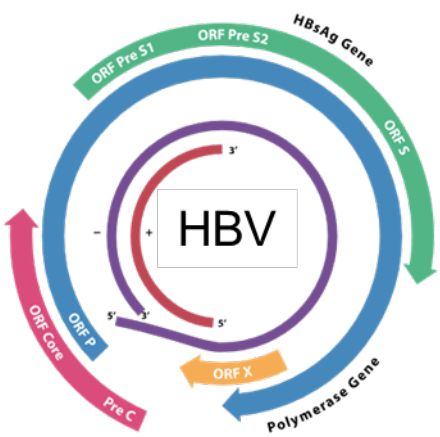
Some viral and mitochondrial genomes make use (due to lack of space) of multiple reading frames to encode more proteins.

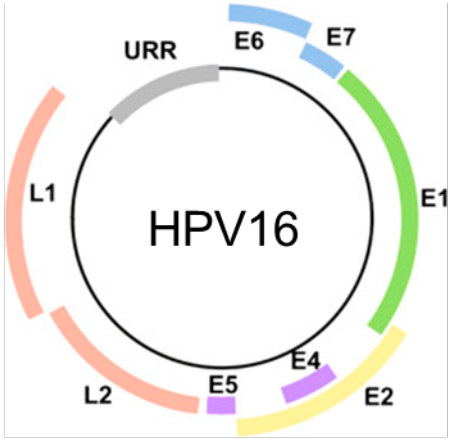
Sequence: `ATGCATGCATGCATGCATGCATCATC`

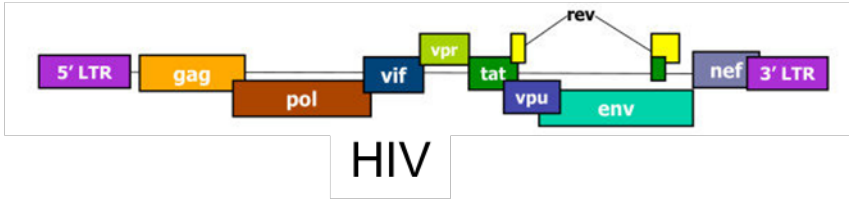
Possible reading frames (remember that they are read in triplets).

```
Frame #1    ATG CAT GCA TGC ATG CAT GCA TGC ATC  = MHACMHACI
Frame #2    TGC ATG CAT GCA TGC ATG CAT GCA      = CMHACMHA
Frame # 3   GCA TGC ATG CAT GCA TGC ATG CAT      = ACMHACMH

Frame #4 is identical to frame #1 minus start codon.        = HACMHACI
```







Laboratorio de Genómica Viral y Humana

Instalaciones de Alta Contención Biológica Nivel de Bioseguridad 3 (BSL-3) CDC-certificadas

Facultad de Medicina UASLP
San Luis Potosí, México

www.genomica.uaslp.mx

GenomicaUASLP

YouTube GenomicaUASLP

Content copyright and license

The Viral and Human Genomics Laboratory is committed to promoting the human rights of free access to knowledge and to receiving the benefits of scientific progress and its applications by providing universal access to all the resources and publications it produces. This is in agreement with article 15 of the United Nations International Covenant on Economic, Social and Cultural Rights published on April 30, 2020.

(Last updated: October 21, © 2024.)