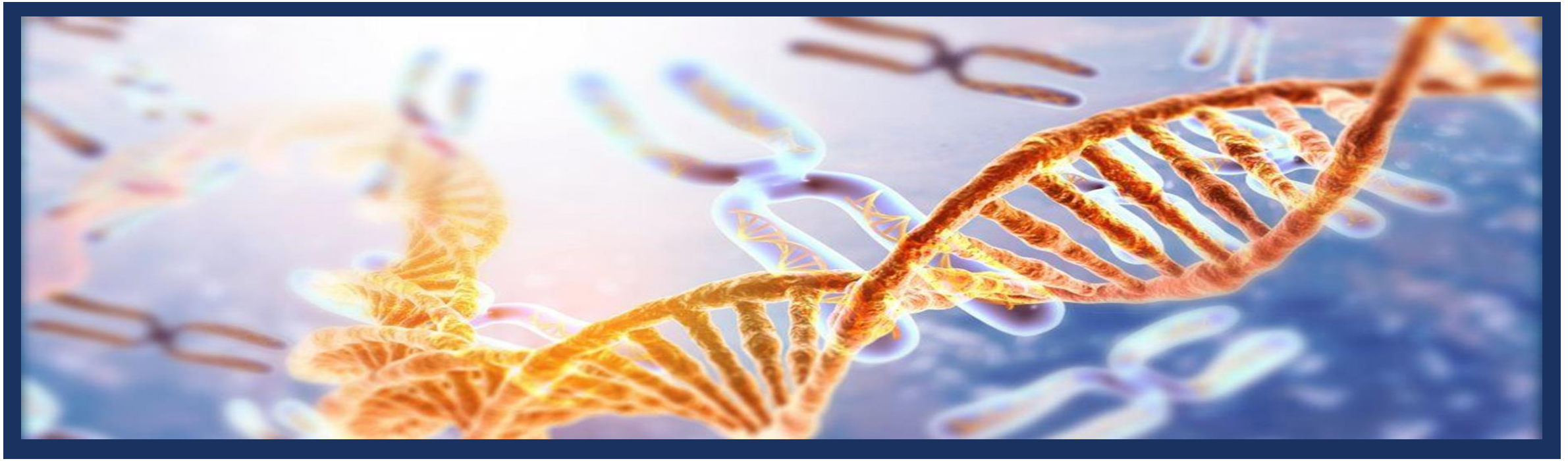
HUMAN GENOME

PRESENTATION OF ADAMO FEDERICA
UNIVERSITY OF RZESZOW



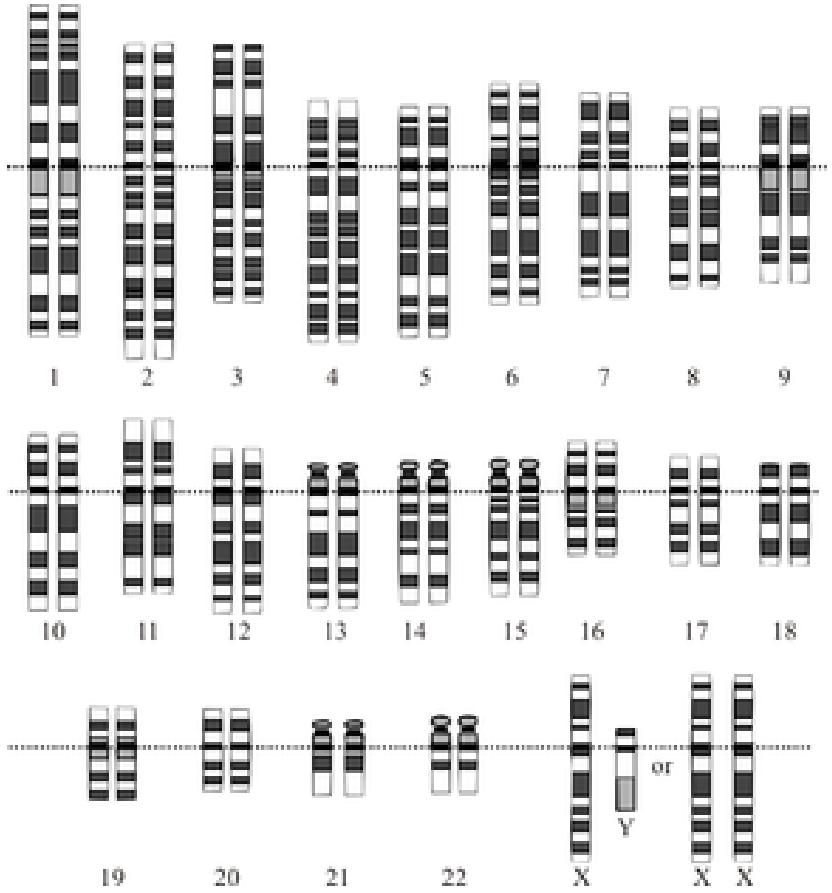
- The human genome is a complete set of nucleic acid sequences for humans, encoded as DNA within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria.
- Human genomes include both protein-coding DNA genes and noncoding DNA.
- Haploid human genomes, which are contained in germ cells (the egg and sperm gamete cells) consist of 3,054,815,472 DNA base pairs (if X chromosome is used), while female diploid genomes (found in somatic cells) have twice the DNA content.
- Size in basepares can vary too: telomeres size is decreasing after every duplication of chromosomes.
- In 2021, scientists reported sequencing the complete "female" genome, without Y chromosome (that nevertheless allowed to achieve "complete status")
- The Human Genome Project has identified a euchromatic reference sequence, which is used globally in the biomedical sciences. The study also found that non-coding DNA totals 98.5 percent, more than had been predicted, and thus only about 1.5 percent of the total length of DNA is based on coding sequences.

CHROMOSOMES

Human nuclear DNA is grouped into 24 types of chromosomes: 22 autosomes, plus two sex-determining chromosomes (X chromosome and Y chromosome).

Chromosomes 1-22 are numbered in order of decreasing length.

Somatic cells have two copies of chromosomes 1-22 each from a parent, plus an X chromosome from the mother and an X or Y chromosome (in female and male, respectively) from the father, for a total of 46 chromosomes distributed in 23 pairs, 22 of homologous chromosomes (autosomes) and one of sex chromosomes (heterosomes).



GENES

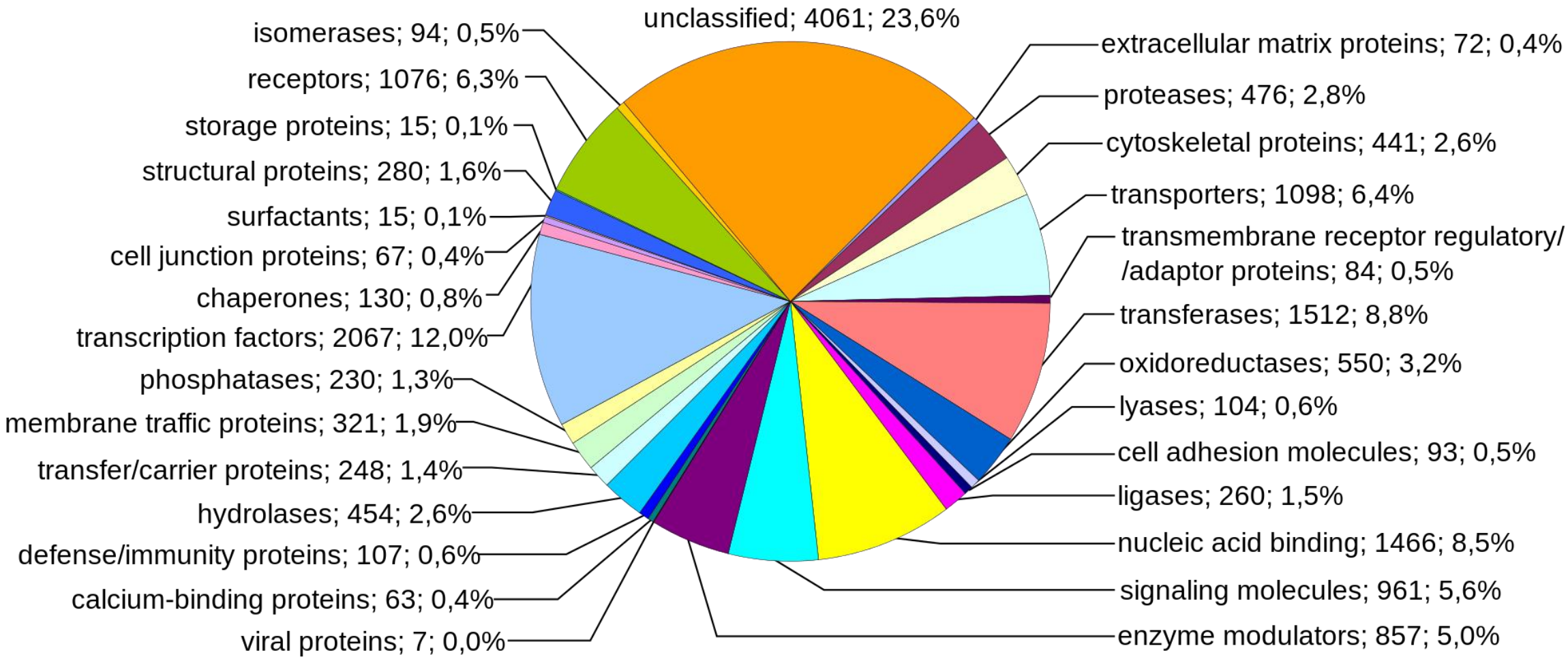
- The existence of approximately 20,000 protein-coding genes has been hypothesized.
- Surprisingly, the number of human genes appears to be only slightly more than twice that of much simpler organisms, such as Drosophila melanogaster. In any case, human cells massively use alternative splicing to produce a large number of different proteins from a single gene, and the human proteome is thought to be much larger than that of the aforementioned organism.
- Most human genes have multiple exons and introns, which are frequently much longer than the flanking exons.
- Human genes are unevenly distributed along chromosomes. Each chromosome contains several gene-rich and gene-poor regions, which appear to correlate with chromosome banding and GC content. The significance of this nonrandom alternation of gene density is not well understood in the current state of scientific knowledge.
- In addition to protein-coding genes, the human genome contains several thousand genes encoding an RNA, including tRNA, ribosomal RNA, and microRNA, as well as other non coding RNA genes.

Chromo-some	Length	Base pairs	Varia-tions	Protein-coding genes	Pseudo-genes	Total long ncRNA	Total small ncRNA	miRNA	rRNA	snRNA	snoRNA	Misc ncRNA	Links	Centromere position (Mbp)	Cumu-lative (%)
1	8.5 cm	248,387,328	12,151,146	2058	1220	1200	496	134	66	221	145	192	EBI	125	7.9
2	8.3 cm	242,696,752	12,945,965	1309	1023	1037	375	115	40	161	117	176	EBI	93.3	16.2
3	6.7 cm	201,105,948	10,638,715	1078	763	711	298	99	29	138	87	134	EBI	91	23
4	6.5 cm	193,574,945	10,165,685	752	727	657	228	92	24	120	56	104	EBI	50.4	29.6
5	6.2 cm	182,045,439	9,519,995	876	721	844	235	83	25	106	61	119	EBI	48.4	35.8
6	5.8 cm	172,126,628	9,130,476	1048	801	639	234	81	26	111	73	105	EBI	61	41.6
7	5.4 cm	160,567,428	8,613,298	989	885	605	208	90	24	90	76	143	EBI	59.9	47.1
8	5.0 cm	146,259,331	8,221,520	677	613	735	214	80	28	86	52	82	EBI	45.6	52
9	4.8 cm	150,617,247	6,590,811	786	661	491	190	69	19	66	51	96	EBI	49	56.3
10	4.6 cm	134,758,134	7,223,944	733	568	579	204	64	32	87	56	89	EBI	40.2	60.9
11	4.6 cm	135,127,769	7,535,370	1298	821	710	233	63	24	74	76	97	EBI	53.7	65.4
12	4.5 cm	133,324,548	7,228,129	1034	617	848	227	72	27	106	62	115	EBI	35.8	70
13	3.9 cm	113,566,686	5,082,574	327	372	397	104	42	16	45	34	75	EBI	17.9	73.4
14	3.6 cm	101,161,492	4,865,950	830	523	533	239	92	10	65	97	79	EBI	17.6	76.4
15	3.5 cm	99,753,195	4,515,076	613	510	639	250	78	13	63	136	93	EBI	19	79.3
16	3.1 cm	96,330,374	5,101,702	873	465	799	187	52	32	53	58	51	EBI	36.6	82
17	2.8 cm	84,276,897	4,614,972	1197	531	834	235	61	15	80	71	99	EBI	24	84.8
18	2.7 cm	80,542,538	4,035,966	270	247	453	109	32	13	51	36	41	EBI	17.2	87.4
19	2.0 cm	61,707,364	3,858,269	1472	512	628	179	110	13	29	31	61	EBI	26.5	89.3
20	2.1 cm	66,210,255	3,439,621	544	249	384	131	57	15	46	37	68	EBI	27.5	91.4
21	1.6 cm	45,090,682	2,049,697	234	185	305	71	16	5	21	19	24	EBI	13.2	92.6
22	1.7 cm	51,324,926	2,135,311	488	324	357	78	31	5	23	23	62	EBI	14.7	93.8
X	5.3 cm	154,259,566	5,753,881	842	874	271	258	128	22	85	64	100	EBI	60.6	99.1
Y	2.0 cm	62,460,029	211,643	71	388	71	30	15	7	17	3	8	EBI	10.4	100
mtDNA	5.4 μm	16,569	929	13	0	0	24	0	2	0	0	0	EBI	N/A	100

CODING SEQUENCES (PROTEIN-CODING GENES)

- Protein-coding sequences represent the most widely studied and best understood component of the human genome., that lead to the production of all human proteins, although several biological processes (e.g. DNA rearrangements and alternative pre-mRNA splicing) can lead to the production of many more unique proteins than the number of protein-coding genes.
- The complete modular protein-coding capacity of the genome is contained within the exome, and consists of DNA sequences encoded by exons that can be translated into proteins. Because of its biological importance, and the fact that it constitutes less than 2% of the genome, sequencing of the exome was the first major milepost of the Human Genome Project.
- About 20,000 human proteins have been annotated in databases such as Uniprot.
- Protein-coding genes are distributed unevenly across the chromosomes, ranging from a few dozen to more than 2000, with an especially high gene density within chromosomes 1, 11, and 19.

Human genes categorized by function of the transcribed proteins, given both as number of encoding genes and percentage of all genes



NONCODING DNA (NCDNA)

- Noncoding DNA is defined as all of the DNA sequences within a genome that are not found within protein-coding exons, and so are never represented within the amino acid sequence of expressed proteins. By this definition, more than 98% of the human genomes is composed of ncDNA.
- Numerous classes of noncoding DNA have been identified, including genes for noncoding RNA (e.g. tRNA and rRNA), pseudogenes, introns, untranslated regions of mRNA, regulatory DNA sequences, repetitive DNA sequences, and sequences related to mobile genetic elements.
- Many of these sequences regulate the structure of chromosomes by limiting the regions of heterochromatin formation and regulating structural features of the chromosomes, such as the telomeres and centromeres.
- Other noncoding regions serve as origins of DNA replication, or to regulate the expression of protein-coding genes (for example mRNA translation and stability), chromatin structure (including histone modifications), DNA methylation, DNA recombinatio, and cross-regulate other noncoding RNAs.
- It is also likely that many transcribed noncoding regions do not serve any role and this transcription is the product of non-specific RNA Polymerase activity.

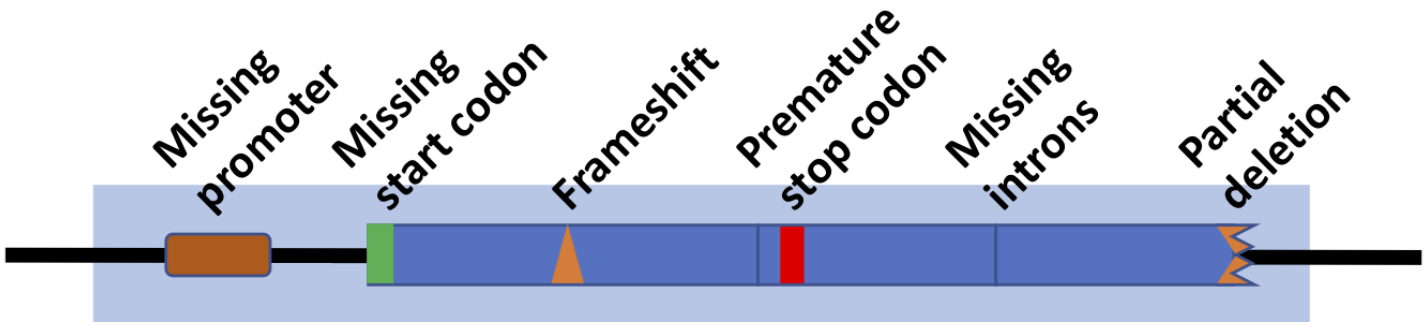
PSEUDOGENES

Pseudogenes are inactive copies of protein-coding genes, often generated by gene duplication, that have become nonfunctional through the accumulation of inactivating mutations.

The number of pseudogenes in the human genome is on the order of 13,000, and in some chromosomes is nearly the same as the number of functional protein-coding genes.

Gene duplication is a major mechanism through which new genetic material is generated during molecular evolution.

Common defects of pseudogenes:



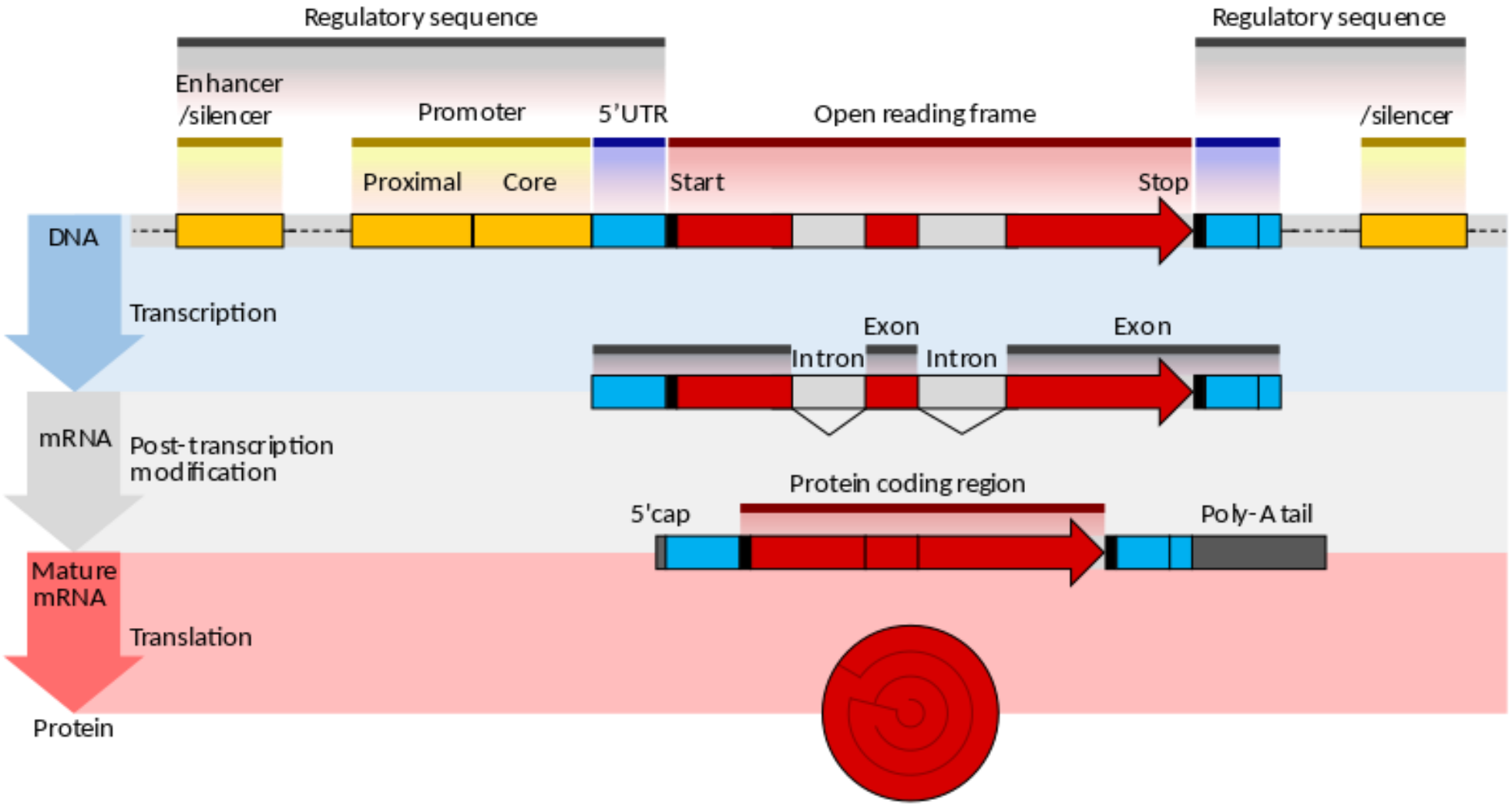
REGULATORY DNA SEQUENCES

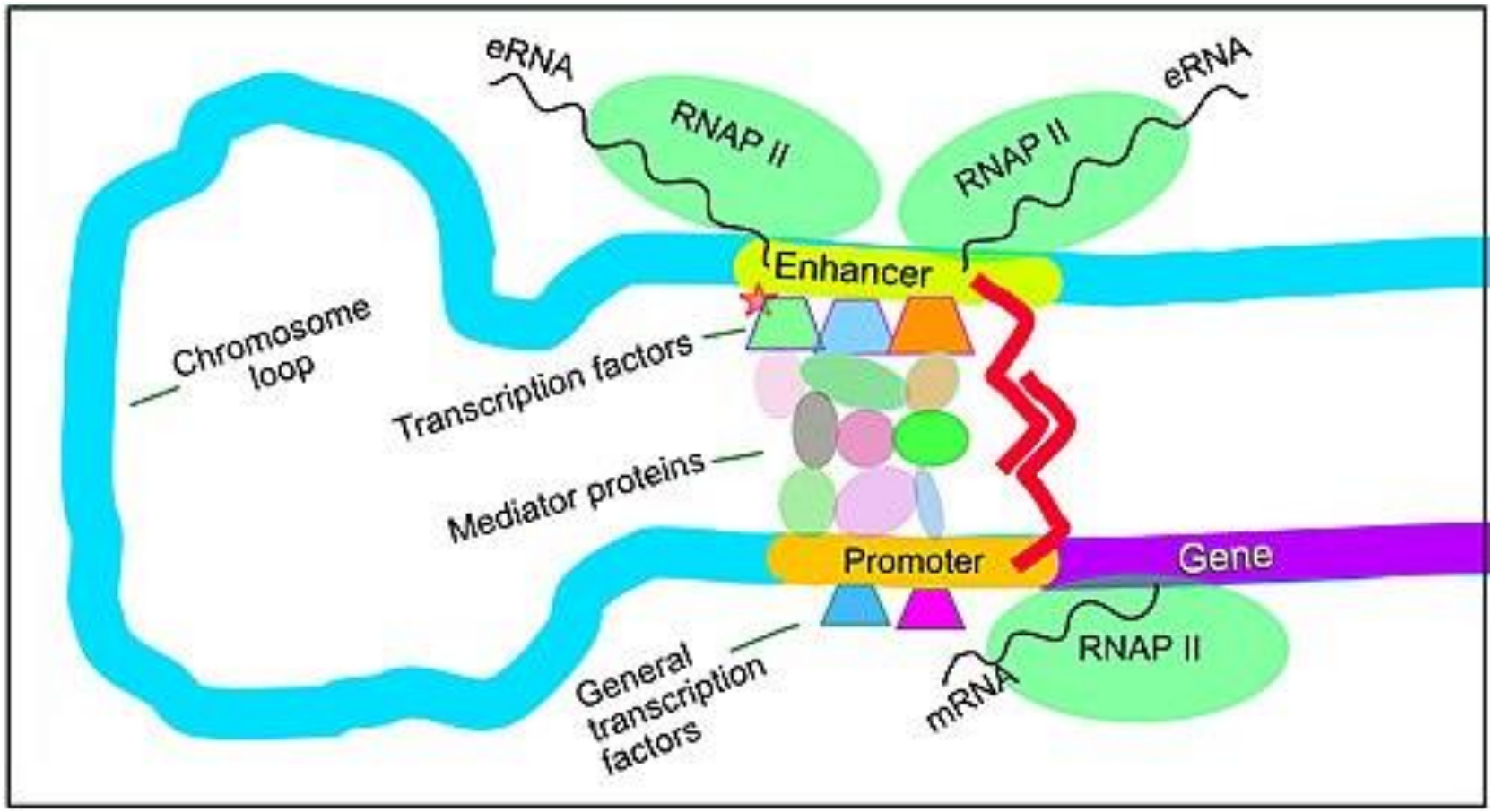
A regulatory sequence is a segment of a nucleic acid molecule which is capable of increasing or decreasing the expression of specific genes within an organism. Regulation of gene expression is an essential feature of all living organisms.

Conservative estimates indicate that these sequences make up 8% of the genome, however extrapolations from the ENCODE project give that 20-40% of the genome is gene regulatory sequence.

Some types of non-coding DNA are genetic "switches" that do not encode proteins, but do regulate when and where genes are expressed (called enhancers)

- CAAT box
- CCAAT box
- Operator (biology)
- Pribnow box
- TATA box
- SECIS element, mRNA
- Polyadenylation signal, mRNA
- A-box
- Z-box
- C-box
- E-box
- G-box


Regulatory sequence
Regulatory sequence
Enhancer
/silencer
Promoter
5'UTR
Open reading frame
/silencer
Proximal
Core
Start
Stop
DNA
Transcription
Exon
Exon
Intron
Intron
mRNA
Post-transcription
modification
Protein coding region
5'cap
Poly-A tail
Mature
mRNA
Translation
Protein


eRNA
RNAP II
eRNA
RNAP II
Enhancer
Chromosome loop
Transcription factors
Mediator proteins
Promoter
Gene
General transcription factors
mRNA
RNAP II

REPETITIVE DNA SEQUENCES

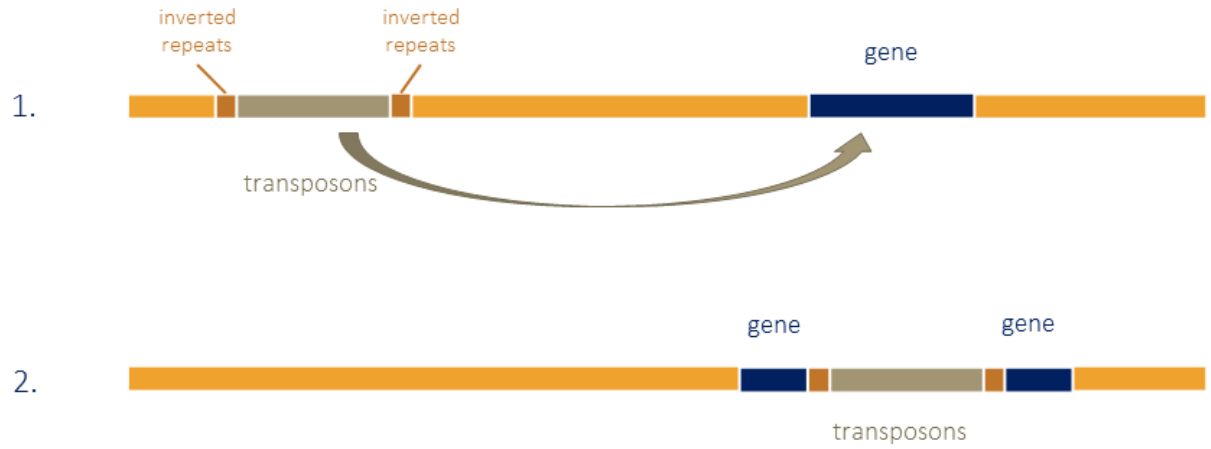
- Repetitive DNA sequences comprise approximately 50% of the human genome.
- About 8% of the human genome consists of tandem DNA arrays or **tandem repeats**, low complexity repeat sequences that have multiple adjacent copies (e.g. "CAGCAGCAG..."). The tandem sequences may be of variable lengths, from two nucleotides to tens of nucleotides. These sequences are highly variable, even among closely related individuals, and so are used for genealogical DNA testing and forensic DNA analysis.
- Repeated sequences of fewer than ten nucleotides (e.g. the dinucleotide repeat (AC)n) are termed **microsatellite** sequences. Among the microsatellite sequences, trinucleotide repeats are of particular importance, as sometimes occur within coding regions of genes for proteins and may lead to genetic disorders. For example, Huntington's disease results from an expansion of the trinucleotide repeat (CAG)n within the Huntingtin gene on human chromosome 4.
- **Telomeres** (the ends of linear chromosomes) end with a microsatellite hexanucleotide repeat of the sequence (TTAGGG)n.
- Tandem repeats of longer sequences (arrays of repeated sequences 10–60 nucleotides long) are termed *minisatellites*.

MOBILE GENETIC ELEMENTS (TRANSPOSONS)

Transposable genetic elements, DNA sequences that can replicate and insert copies of themselves at other locations within a host genome, are an abundant component in the human genome, accounting for over half of total human DNA.

The most abundant transposon lineage, Alu, has about 50,000 active copies, and can be inserted into intragenic and intergenic regions. One other lineage, LINE-1, has about 100 active copies per genome.

Some of these sequences represent endogenous retroviruses, DNA copies of viral sequences that have become permanently integrated into the genome and are now passed on to succeeding generations



REFERENCES

- *"T2T-CHM13v2.0 - Genome - Assembly - NCBI". www.ncbi.nlm.nih.gov.*
- *Brown TA (2002). The Human Genome (2nd ed.). Oxford: Wiley-Liss.*
- *Jump up to: Nurk, Sergey; et al. (April 2022). "The complete sequence of a human genome".*
- *Jump up to: Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (November 2012). "An integrated map of genetic variation from 1,092 human genomes".*
- *Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. (October 2015). "A global reference for human genetic variation".*
- *Chimpanzee Sequencing Analysis Consortium (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome.*
- *en.wikipedia.org/wiki/Human_genome#Mapping_human_genomic_variation*