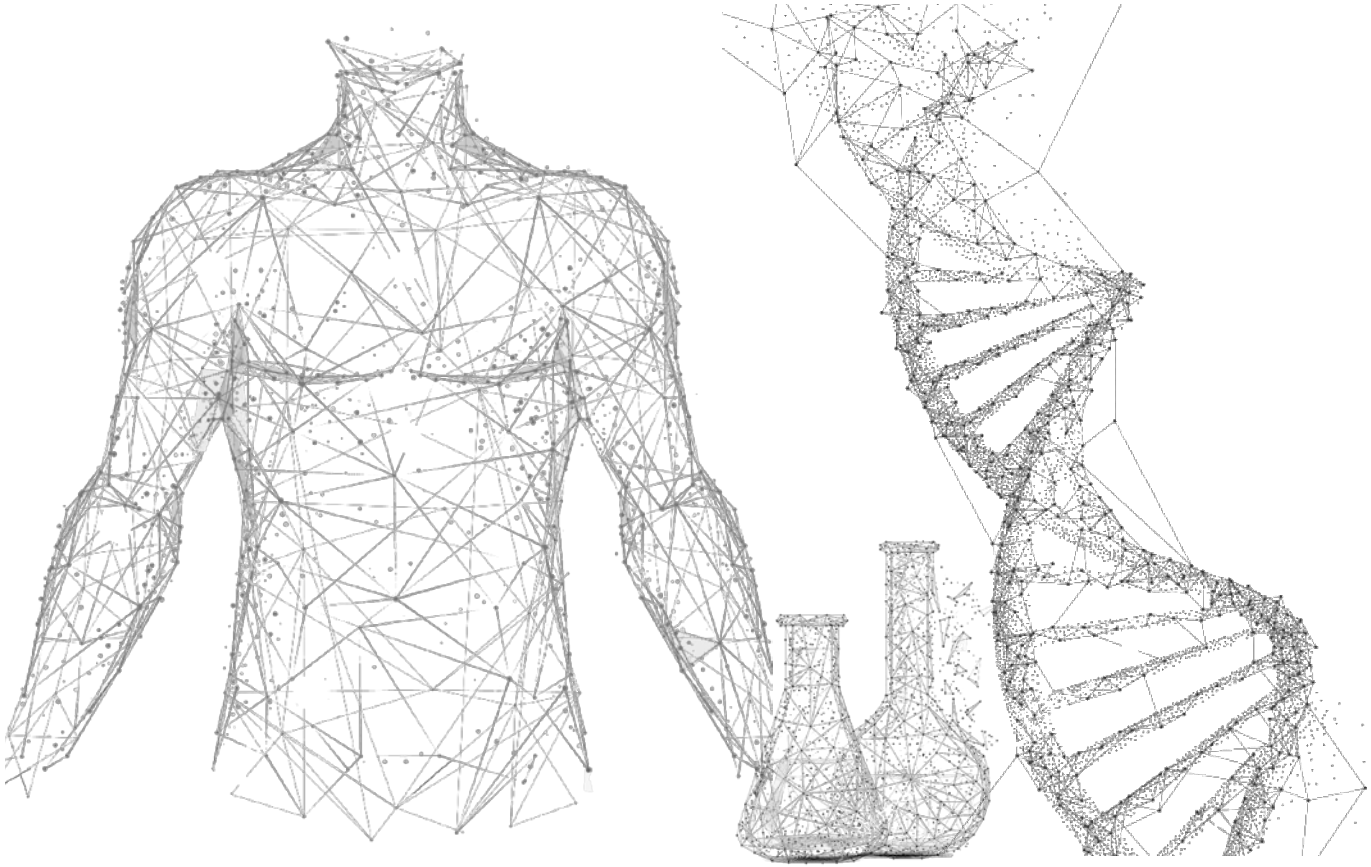


Genomes

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

Dr. Christian A. García-Sepúlveda

Viral & Human Genomics BSL-3 Laboratory

Last updated January 22, 2025 v3

Genomes

A genome is all the genetic information of an organism.

A genome is an organism's complete set of DNA, including all of its genes as well as its hierarchical, three-dimensional structural configuration.

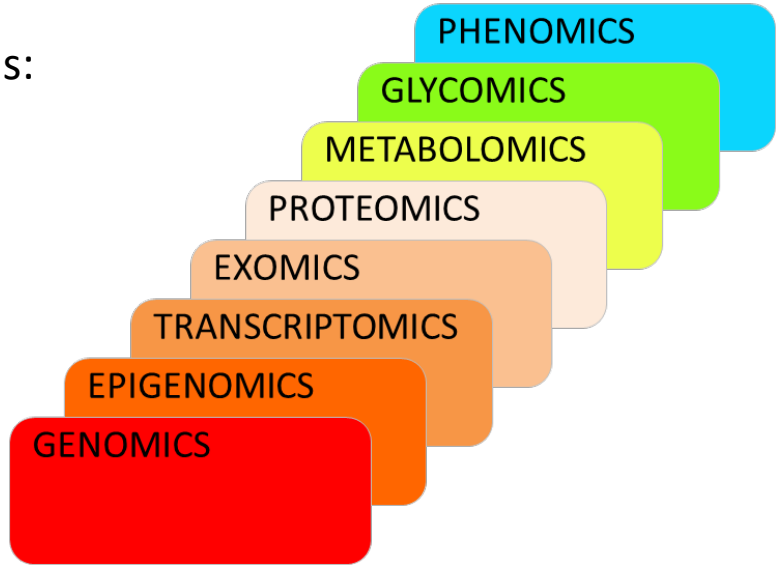
Consists of nucleotide sequences of DNA or RNA.

Term created in 1920 by Hans Winkler a German botanist.

Oxford Dictionary suggest the name is a blend of the words gene and chromosome.

Fits systematically with a few related -ome words:

- Biome
- Rhysome
- Nucleosome
- Chromosome
- Ribosome
- Replisome
- Etceterasome



Genomes

DNA genomes

10,000 bacterial species
500 archaeal species
611,000 fungal species, including yeasts
298,000 plant species
7.8 million animal species.
58,000 to 77,000 viral species

RNA genomes

161,979 RNA virus species
44 viroids

Genomes

Viroid and viral genomes are either DNA or RNA based, compact and overlapped.

Prokaryote (Bacterial) genomes are DNA based, mainly circularized and single.

Bacteria have one or two chromosomes containing all essential genetic material.

Bacteria also contain smaller extrachromosomal plasmid molecules that carry additional non-essential genetic information.

Eukaryote genomes are diploid and have nuclear and endosymbiont components.

- All eukaryotes have mitochondrial genome.
- Algae and plants also contain chloroplast genome.

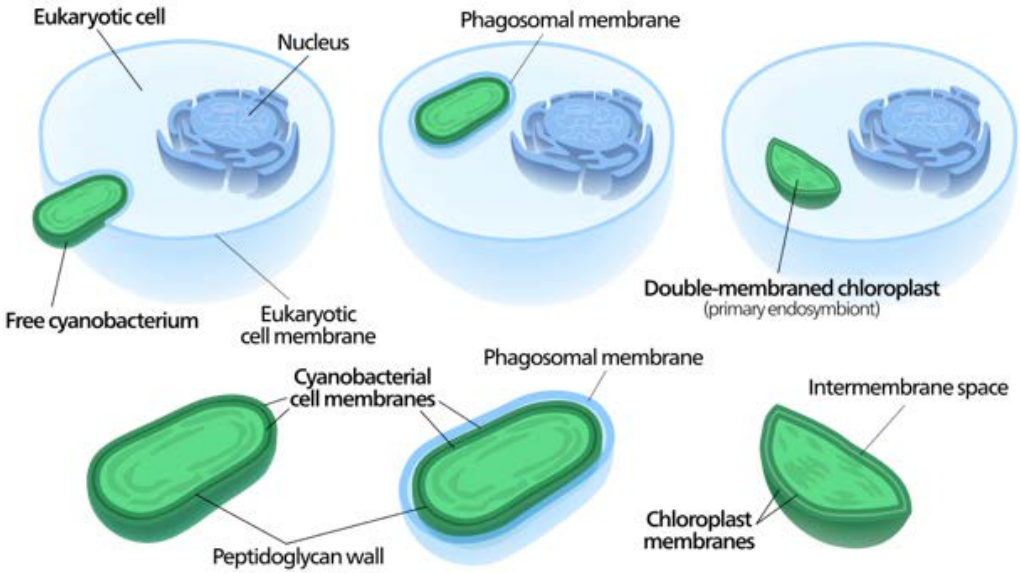
The scientific literature term 'genome' is usually restricted to the large chromosomal DNA molecules in bacteria and eukaryotes.

Endosymbionts

An organism that lives within another organism in a typically mutualistic relationship.

2.2 billion years ago an archaeon phagocytosed a bacterium which eventually became the mitochondria which provides eukaryotes with energy.

1 billion years ago, some of those mitochondria-bearing eukaryote cells absorbeda cyanobacteria that eventually became a chloroplasts, which produces energy from sunlight



Viroids

Naked circular ssRNA, 246 to 467 bp

Do not encode proteins.

Phytopathogenic and candidates for the primordial genetic material.

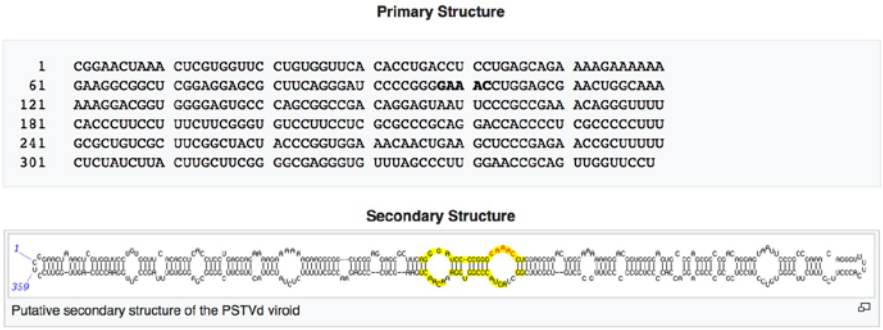
They accumulate mutations and can recombine.

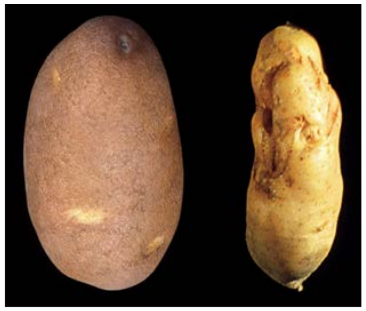
Autonomous replication.

Transmitted mechanically through cellular debris.

33 viroids have been identified.

Extensive intra-strand base pairings with unpaired loops protect the viroid from degradation by ribonuclease.





Potato spindle tuber viroid (PSTVd)

Rod Type Viroid Structure

Branched Type Viroid Structure

www.geeksforgeeks.org/viroid-definition-structure/ ⇦ Get to know a lot more on viroids here

Virusoids

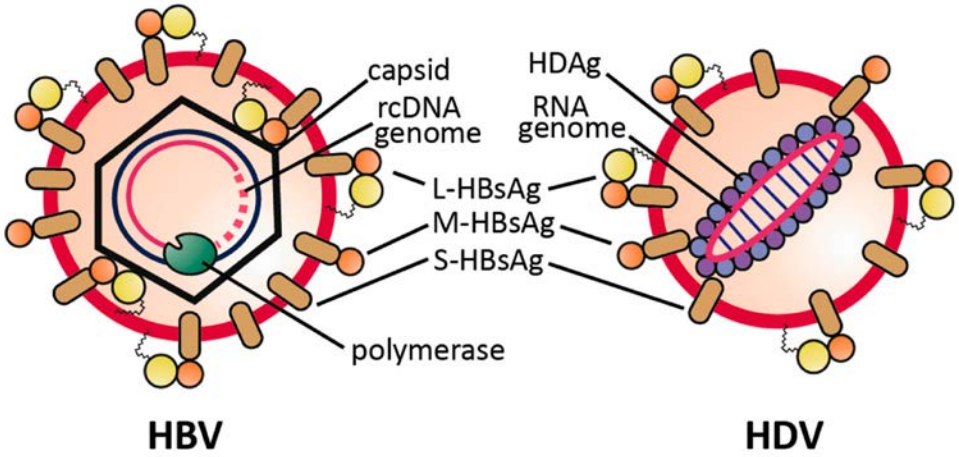
Small (220 to 388 bp), circular, non-self-replicating single-stranded RNA molecules.

Also do not code for any proteins.

Also phytopathogenic.

Need “helper virus” for replication.

Also called satellite RNAs.



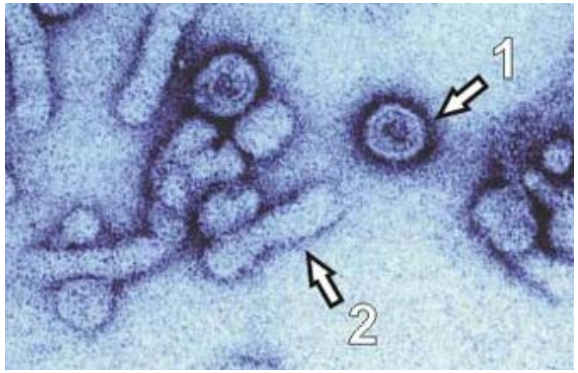
Virusoids are RNA molecules which use the capsids of other viruses.

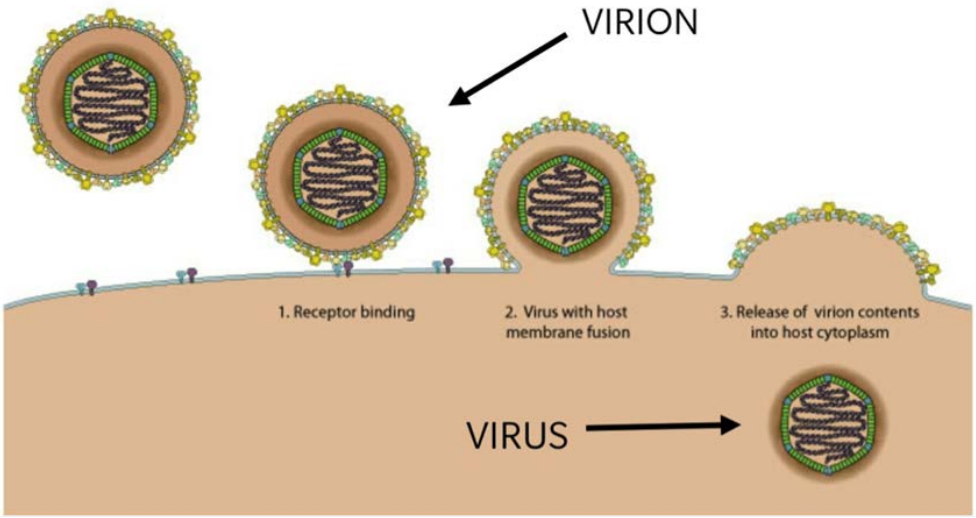
Human hepatitis D agent (HDV) is a virusoid and requires HBV to replicate and to encapsidate.

Subviral Particles

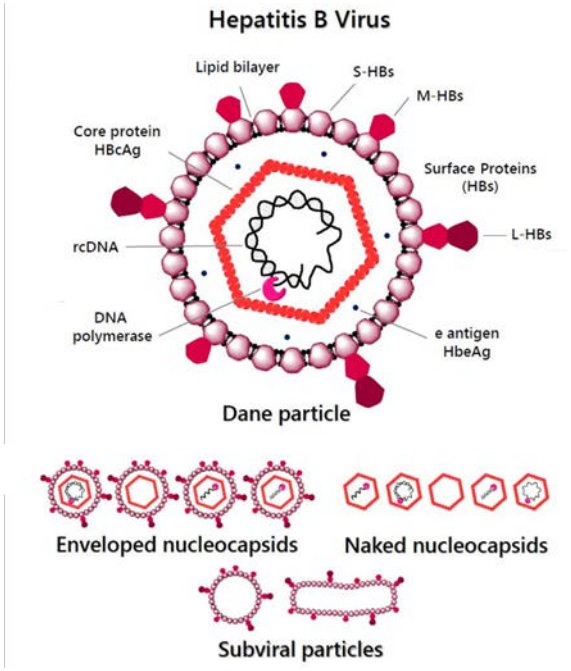
Hepadnaviruses (HBV) are the smallest enveloped (membrane) viruses observed in animals (42 nm).

Infected hepatocytes secrete non-infectious HBV subviral particles that lack genetic material (> 100,000 or 1,000,000 particles per cell).





Difference between virus and virion



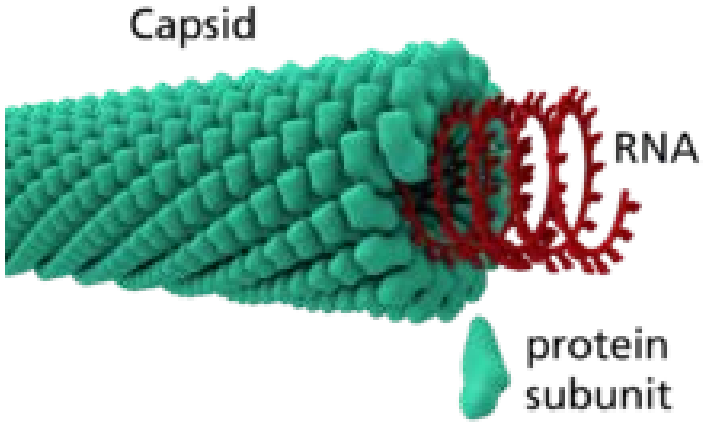
Viral genomes

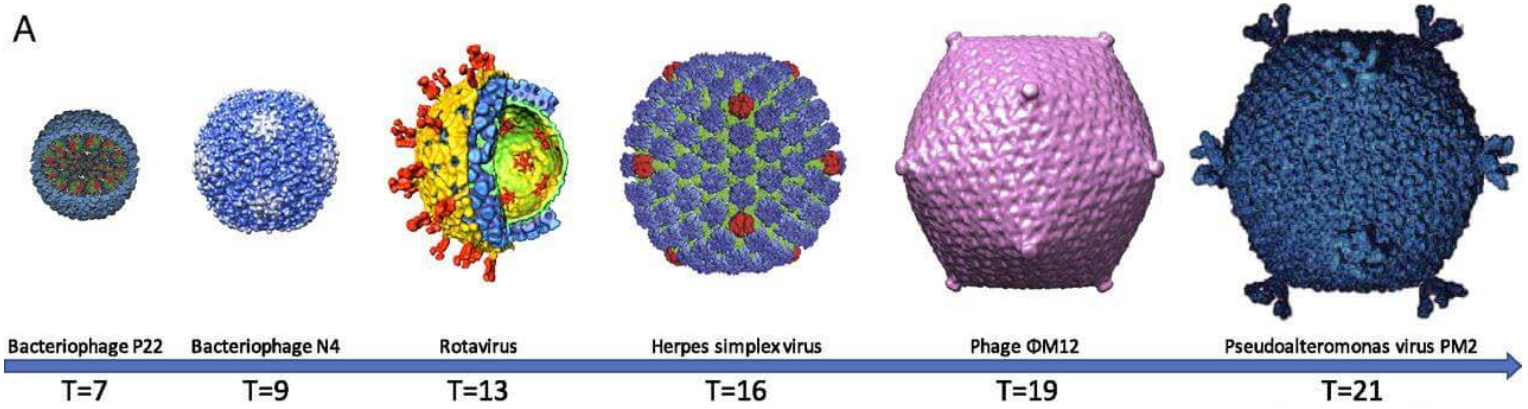
The more ergonomic extreme leads to the viral capsid being made up of a single type of protein subunit.

Two essential types of capsids:

1.- **Filamentous or helical.**

2.- **Icosahedral.**





Prions

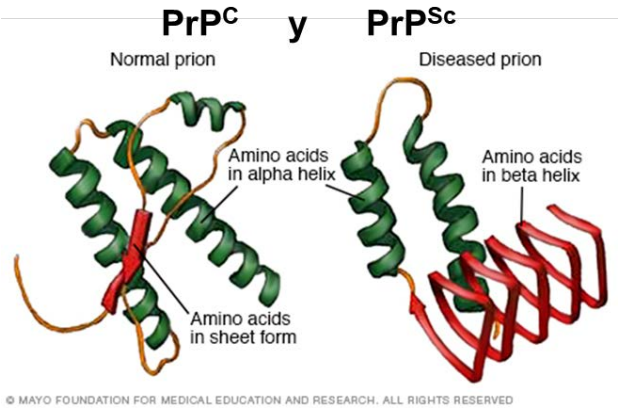
Stanley Prusiner coined the term in the early 1980s.

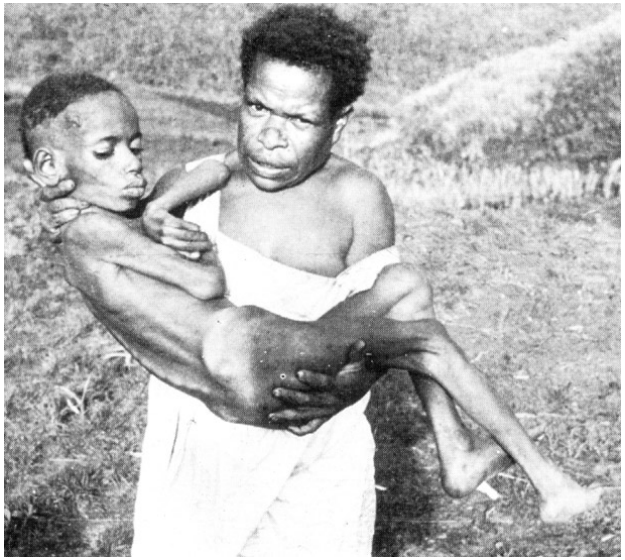
Neurological diseases caused by infectious agents resistant to nucleic acid destruction processes.

Initially controversial, won the 1997 Nobel Prize in Physiology or Medicine.

Diseases called spongiform enceflaopathies due to histological appearance.

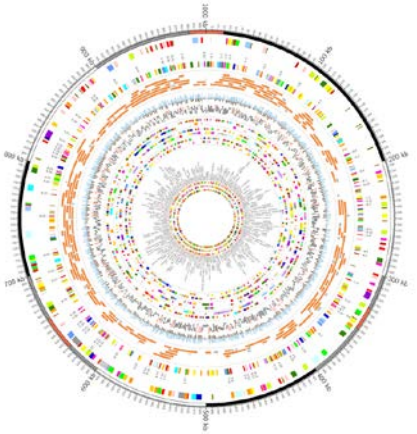
Protein particles that catalyze irreversible changes on similar proteins, leading to their accumulation in cells...



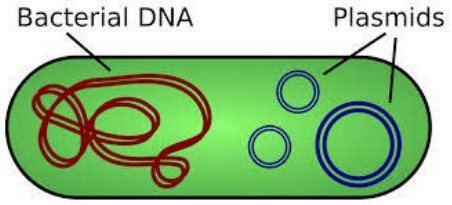


Prokaryotic genomes

Compact and Efficient:
Prokaryotic genomes are typically smaller and more compact than eukaryotic genomes, ranging from about 0.5 to 10 Mb in size.

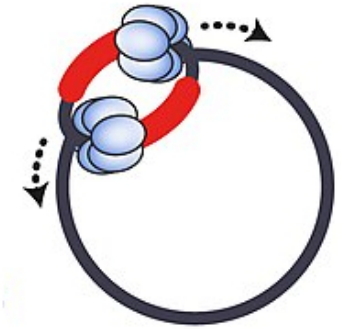


Circular DNA
Most prokaryotes have a single, circular chromosome, although some species may possess multiple chromosomes or linear chromosomes.



Replication Origin
Replication usually starts at a single origin of replication (OriC) and proceeds bidirectionally.

Plasmids
In addition to the main chromosome, prokaryotes often carry extrachromosomal DNA in the form of plasmids, which can confer advantageous traits such as antibiotic resistance.



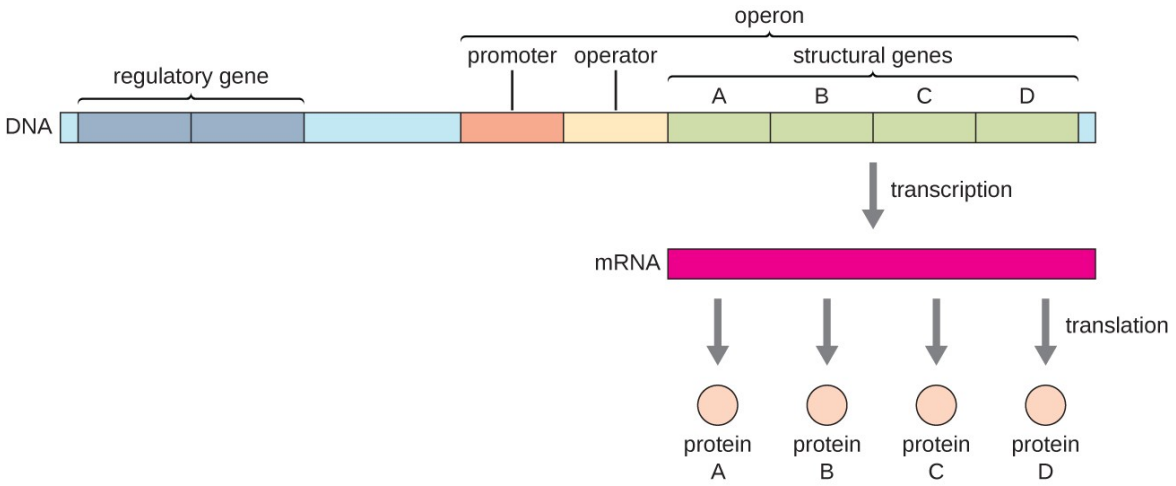
Prokaryotic genomes

Minimal Non-Coding Regions

Non-coding DNA is minimal compared to eukaryotic genomes, optimizing genetic information density.

Operons

Genes are often organized into operons, allowing coordinated expression of functionally related genes.



Functional Redundancy

Despite their compact size, prokaryotic genomes often encode redundant systems for critical functions, enhancing survival under diverse conditions.

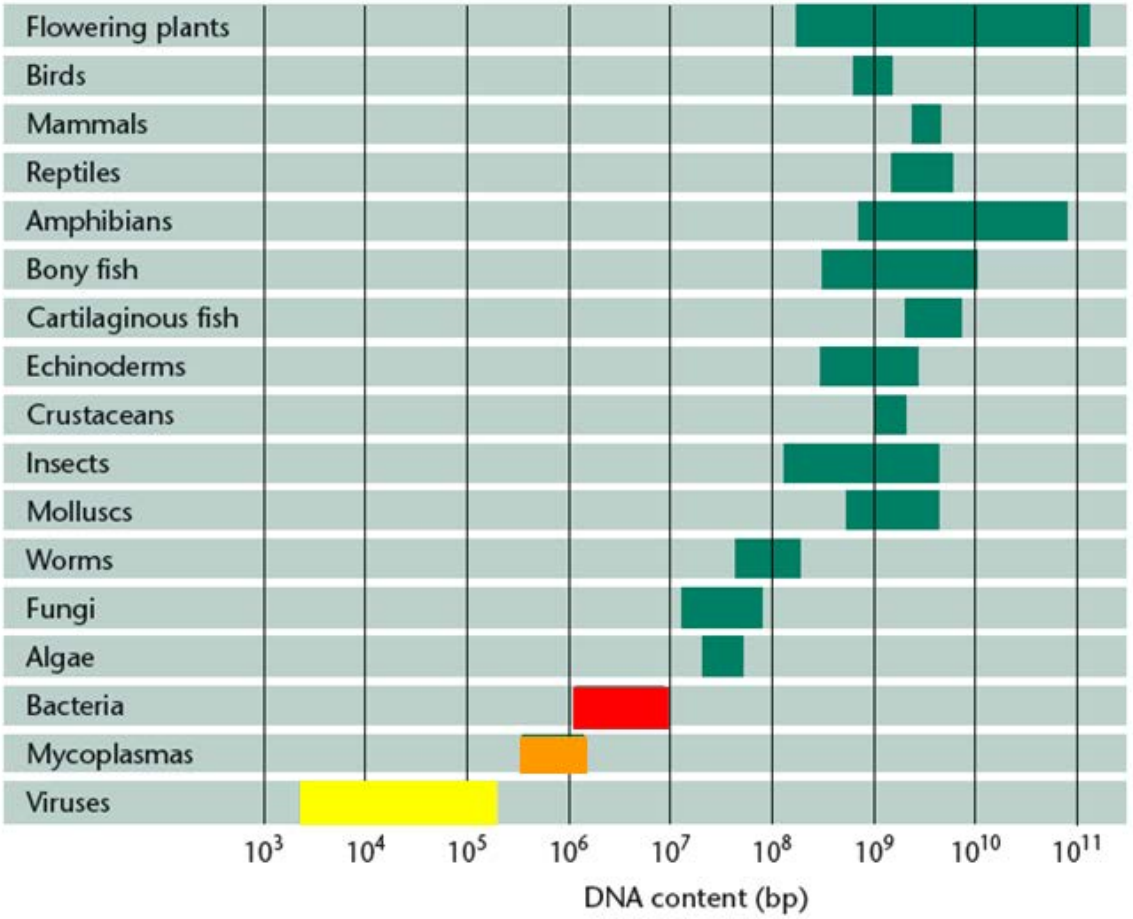
Eukaryotic genomes

Large and Complex

Significantly larger and more complex than prokaryotic genomes, ranging from tens of millions to billions of base pairs.

Linear Chromosomes

Eukaryotic genomes are organized into multiple linear chromosomes housed within a nucleus.



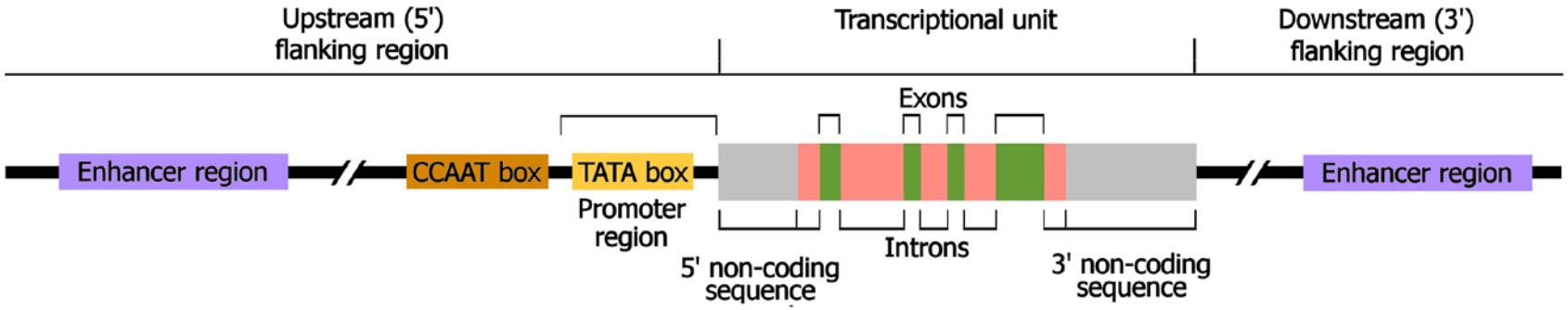
Eukaryotic genomes

Introns and Exons

Genes are interrupted by non-coding sequences (introns), requiring splicing to generate functional mRNAs.

Regulatory Elements

Eukaryotic genomes contain extensive regulatory elements, including promoters, enhancers, and silencers, to finely tune gene expression.



Eukaryotic genomes

Repetitive DNA

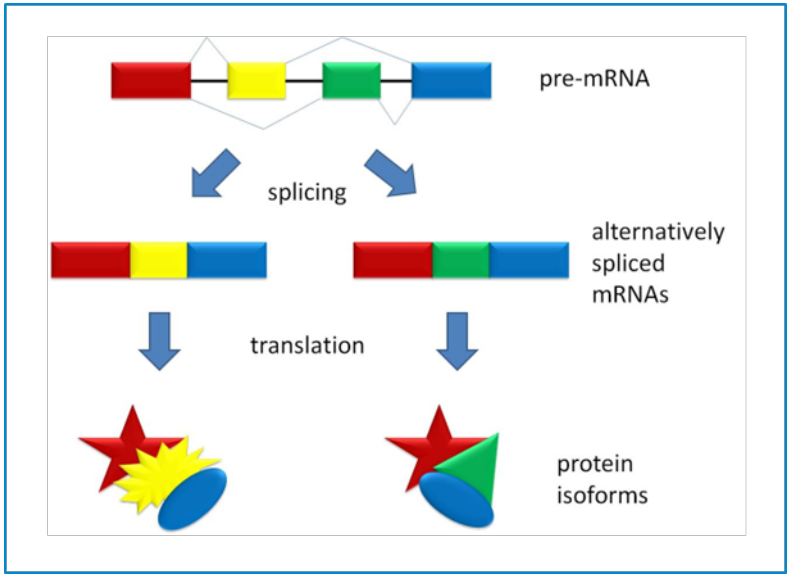
A large portion of eukaryotic genomes consists of repetitive DNA sequences, including tandem repeats and transposable elements.

Epigenetics

DNA methylation and histone modifications regulate gene expression without altering the underlying DNA sequence.

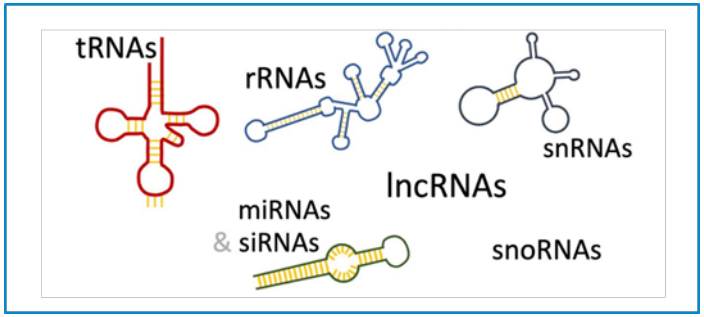
Alternative Splicing

Eukaryotes utilize alternative splicing to produce multiple protein isoforms from a single gene, increasing proteome diversity.



Non-Coding RNA

Significant portions of eukaryotic genomes transcribe non-coding RNAs, such as microRNAs and long non-coding RNAs, with regulatory roles.



Transposons

Two types of transposable elements (TEs):

- Class 1: Retrotransposons (DNA-RNA-DNA)
- Class 2: DNA transposons (Alu)

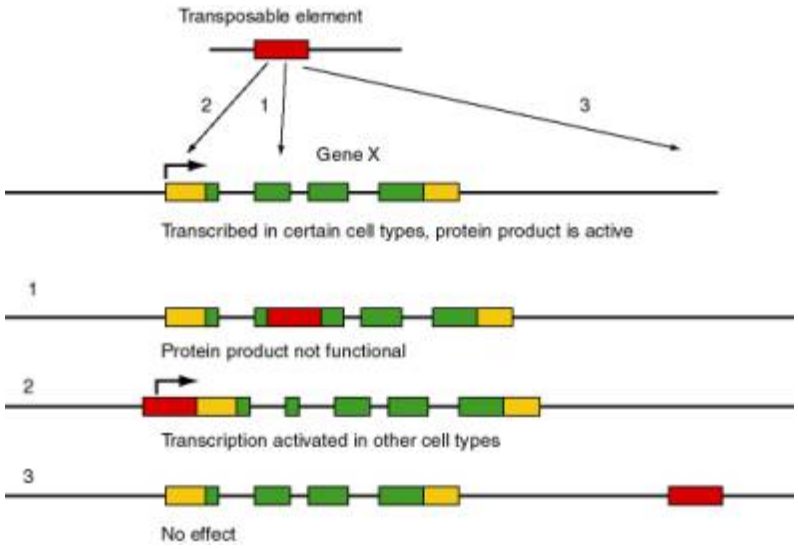
Between 300,000 and 1 million copies of 300 bp Alu repeats in human genome (15-17% of genome).

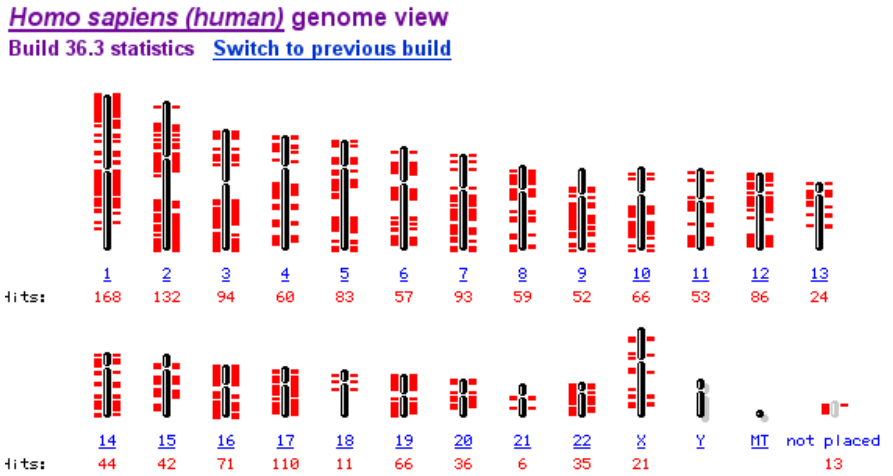
85% of Maize's genome consists of TEs.

TEs in bacteria carry genes for antibiotic resistance.

Diseases can be caused by TEs including:

- Hemophilia A and B,
- Severe combined immunodeficiency (SCID),
- Porphyria,
- Predisposition to cancer, and
- Duchenne muscular dystrophy





Retrotransposons

Retrovirus that have been incorporated to the genome are called Endogenous retroviral sequences (ERV).

They are derived from ancient infections of germ cells in humans, mammals & other vertebrates.

ERVs make up 5-8% of the human genome (98,000 elements).

Most insertions have no known function (junk DNA) but some play important roles in host biology:

Control of gene transcription.

Control of cell fusion during placental development.

Resistance to exogenous retroviral infection.

Immunosuppresion

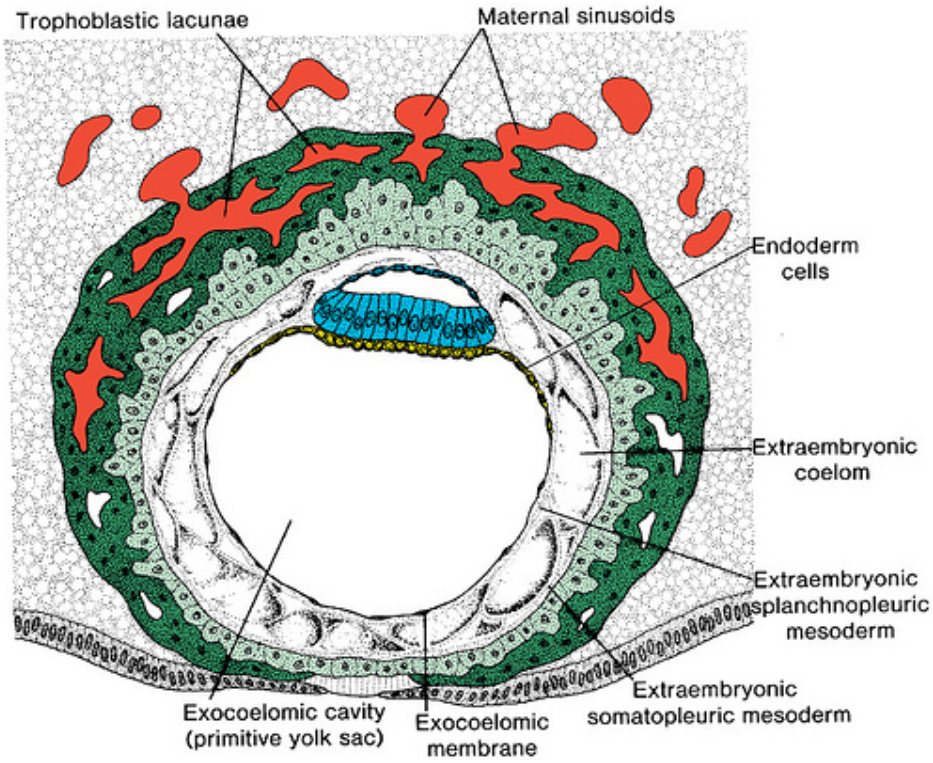
Endoretroviral sequences

ERVs are activated during pregnancy in viviparous mammals (Eutremes).

Monotreme mammals still come from eggs.

They act as immunosupresors protecting the embryo from its mother's immune system.

Viral fusion proteins involved in the formation of placental syncytium limits cell migration (something an epithelium will not do well, as certain blood cells are able to diapedize).

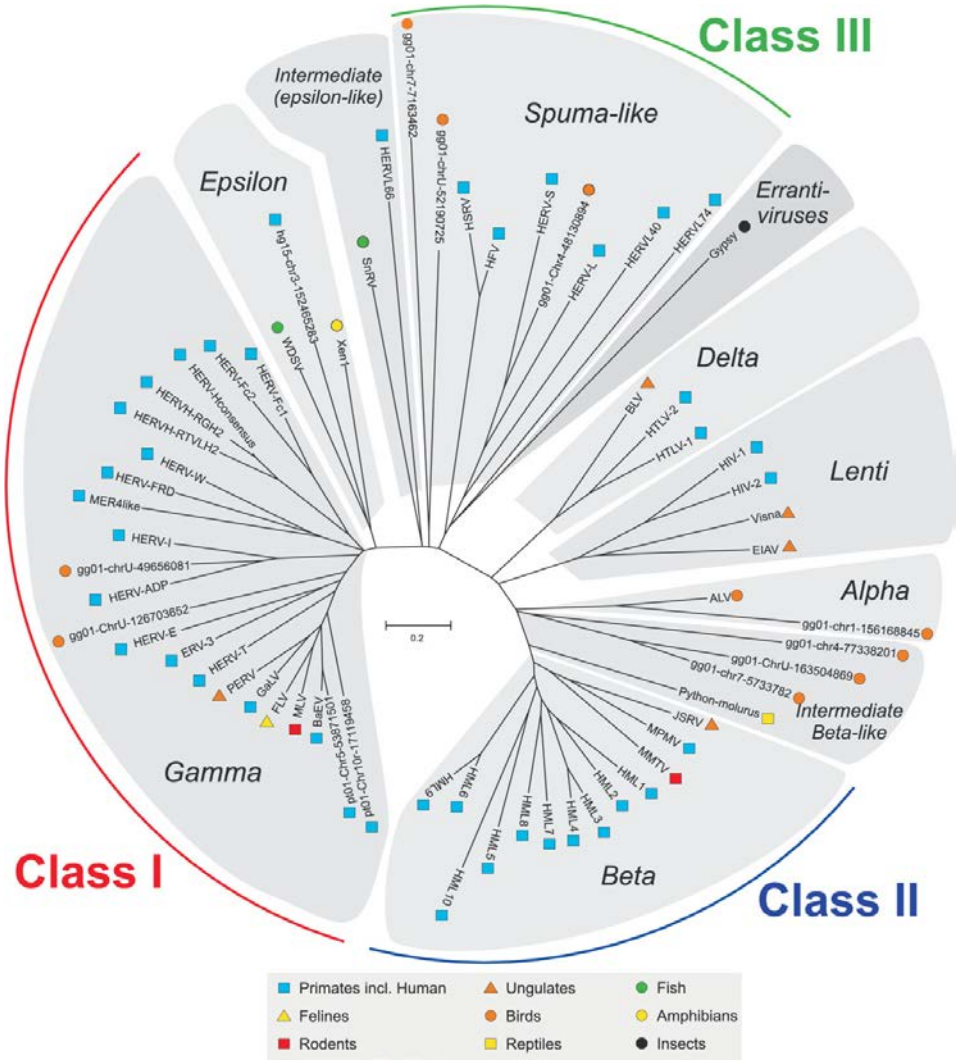


Endoretroviral sequences

24 ERV families identified by Human Genome Project (HGP).

Broadly classified into 3 classes, on the basis of relatedness to exogenous genera:

- **Class I** are similar to the gammaretroviruses
- **Class II** are similar to the betaretroviruses & alpharetroviruses
- **Class III** are similar to the spumaviruses



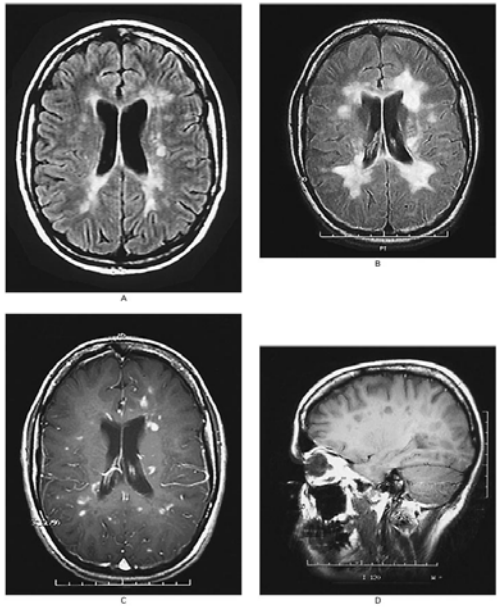
Endoretroviral sequences

Human ERVs (hERVs) are suspected of involvement in some autoimmune diseases (multiple sclerosis).

Especially human endogenous retrovirus W known (MSRV).

Also a possible hERV involvement in the HELLP (Hemolytic anemia, Elevated Liver enzymes & Low Platelet count) syndrome & pre-eclampsia.

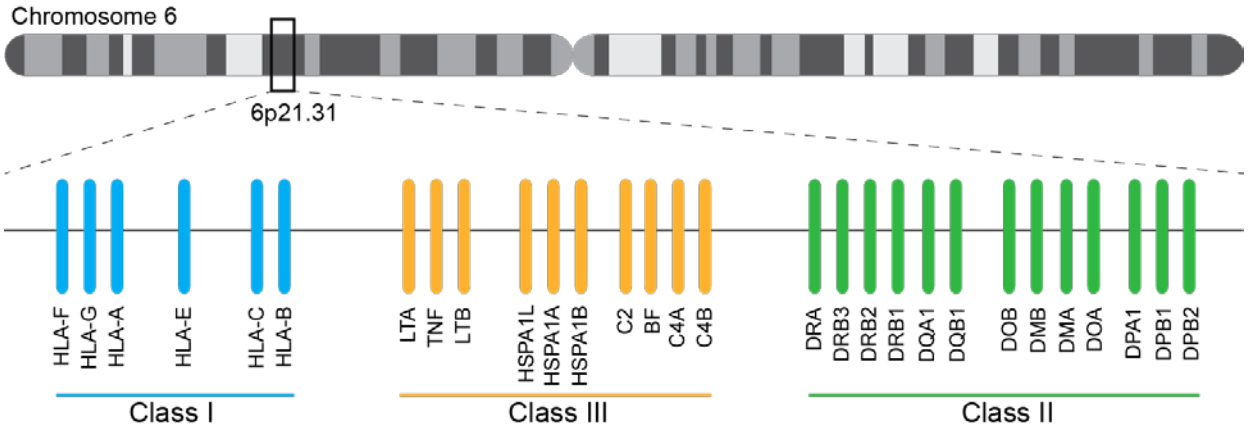
hERVs very likely associated with some types of schizophrenia.



Major Histocompatibility Complex (MHC)

The Major Histocompatibility Complex (MHC) constitutes the most polymorphic genetic system in animals.

It was the first region to be exhaustively studied and the first to be sequenced by the HGP. The MHC constitutes a large genomic region (3.6 Mbp) present in most vertebrates.

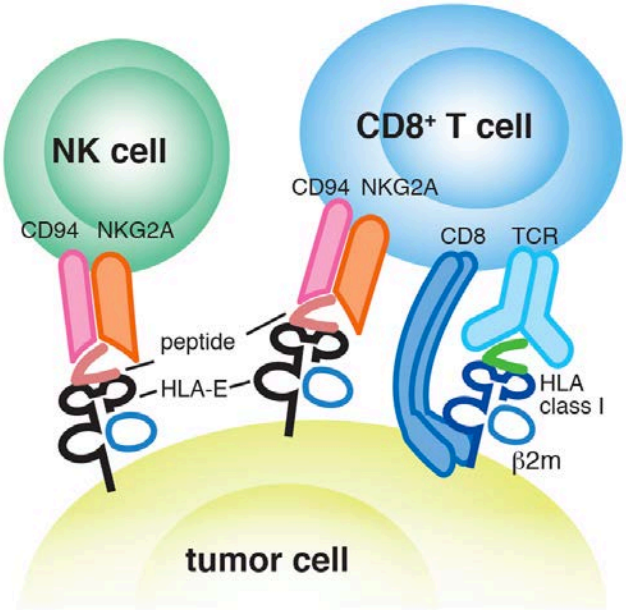


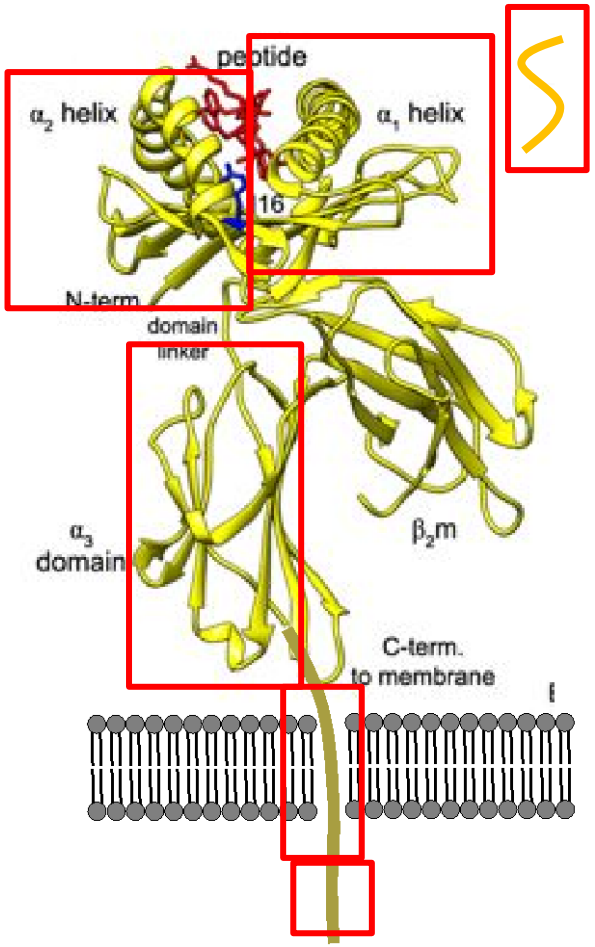
It constitutes the most genetically dense region of the mammalian genome (> 150 genes). Average density of eukaryotes is 14 genes per Mbp, MHC density = ca 42 genes/Mbp.

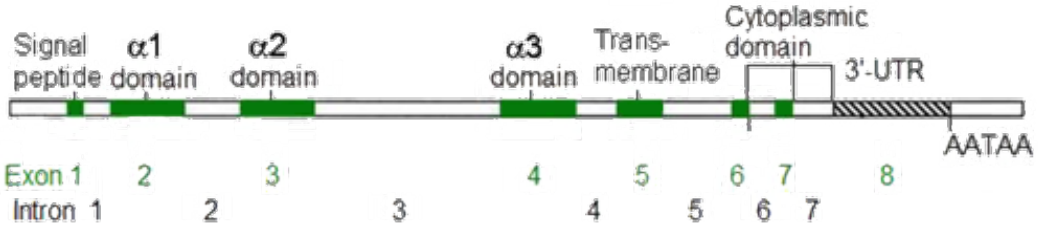
It contains genes involved in the innate and adaptive immune response, with immune, reproductive and inflammatory functions.

HLA molecules

MHC-encoded HLA proteins present peptides to cells responsible for immune surveillance.







HLA polymorphism

The MHC is divided into three functionally distinct regions:

MHC class I, MHC class III and MHC class II

Polymorphic (many alleles), Complex (many genes) and Codominant (all genes expressed) genetic system.

And the number continues to grow...

Numbers of HLA Alleles	
HLA class I alleles	28409
HLA class II alleles	12594
HLA alleles	41003

HLA class I						
Gene	*A*	*B*	*C*	*E*	*F*	*G*
Alleles	8556	10346	8657	376	115	176
Proteins	5004	6172	4776	141	22	52
Nulls	449	370	388	10	3	6

HLA class II														
Gene	*DRA*	*DRB*	*DQA1*	*DQA2*	*DQB1*	*DQB2*	*DPA1*	*DPA2*	*DPB1*	*DPB2*	*DMA*	*DMB*	*DOA*	*DOB*
Alleles	78	4812	872	42	2813	41	765	6	2795	7	62	100	121	80
Proteins	17	3168	448	11	1682	9	373	0	1611	0	9	9	16	17
Nulls	0	215	21	0	122	1	33	0	145	0	0	0	1	1

IPD-IMGT/HLA [IPD-IMGT/HLA Release 3.59.0 (2025-01)]

Polymorphism

Classic Mendelian genetics only distinguished two types of genes: the Wild-type (normally circulating) and the 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, which is why they are not actually mutants = alleles.

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

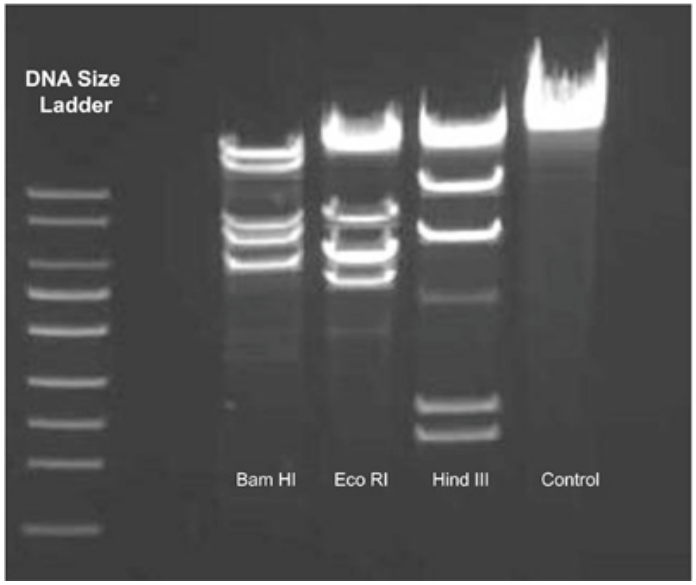
Genetic polymorphism = refers to the existence of multiple alleles for a gene.

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).

Polymorphism

On the other hand, in those systems in which a wild-type allele does exist, a more detailed scrutiny (nt sequence could reveal that even the WT is itself polymorphic).



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

Normally digestion by an enzyme produces specific electrophoretic migration patterns that depend on the existence of specific sequences for each enzyme (restriction sites).

Some polymorphisms (mutations) modify these restriction sites and the electrophoretic pattern generated.

Originally this was used for paternity testing and identity authentication (because we have accumulated different types of mutations that make us different individuals).

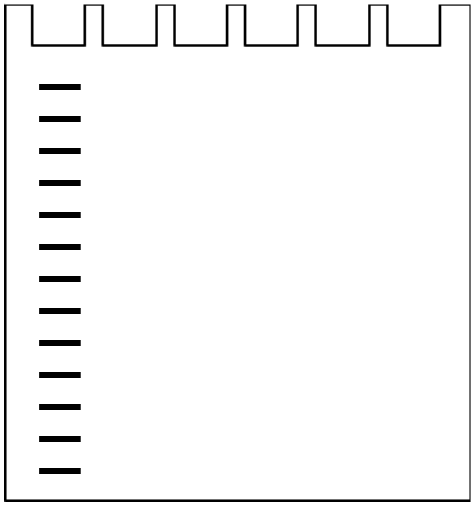
Polymorphism

Dr. Tano

EcoRI GAATTC / CTTAAG

```
5′- ATGCGAATTCCGTTAAGCAGTGAGCTAGGCATGAATTCGTGCGATGCGTA -3′
3′- TACGCTTAAGGCAATTCGTCACTCGATCCGTACTTAAGCACGATACGCAT -5′

5′- ATGCGAATTCCGTTAAGCAGTGAGCTAGGCATGAGTTCGTGCGATGCGTA -3′
3′- TACGCTTAAGGCAATTCGTCACTCGATCCGTACTCAAGCACGATACGCAT -5′
```



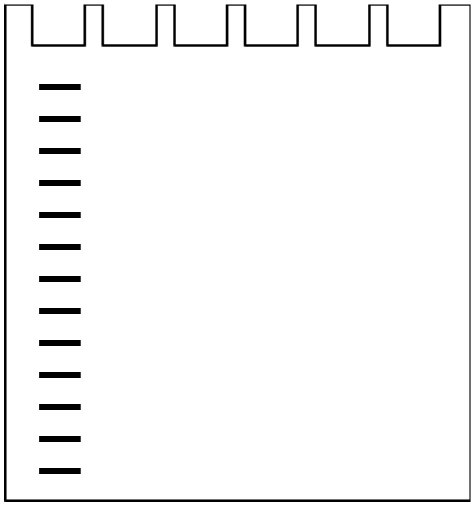
Polymorphism

Dr. Tano

EcoRI G|AATTC / CTTAA|G

```
5′- ATGCG|AATTCCGTTAAGCAGTGAGCTAGGCATG|AATTCGTGCGATGCGTA -3′
3′- TACGCTTAA|GGCAATTCGTCACTCGATCCGTACTTAA|GCACGATACGCAT -5′

5′- ATGCG|AATTCCGTTAAGCAGTGAGCTAGGCATGAGTTCGTGCGATGCGTA -3′
3′- TACGCTTAA|GGCAATTCGTCACTCGATCCGTACTCAAGCACGATACGCAT -5′
```



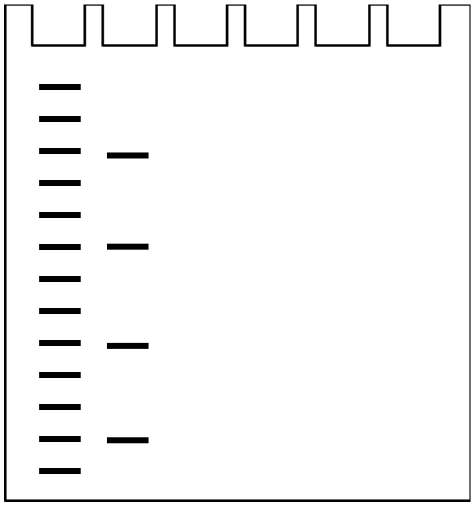
Polymorphism

Dr. Tano

EcoRI G|AATTC / CTTAA|G

```
5′- ATGCGAATTCCGTTAAGCAGTGAGCTAGGCATGAATTCGTGCGATGCGTA -3′
3′- TACGCTTAAGGCAATTCGTCACTCGATCCGTACTTAAGCACGATACGCAT -5′

5′- ATGCGAATTCCGTTAAGCAGTGAGCTAGGCATGAGTTCGTGCGATGCGTA -3′
3′- TACGCTTAAGGCAATTCGTCACTCGATCCGTACTCAAGCACGATACGCAT -5′
```

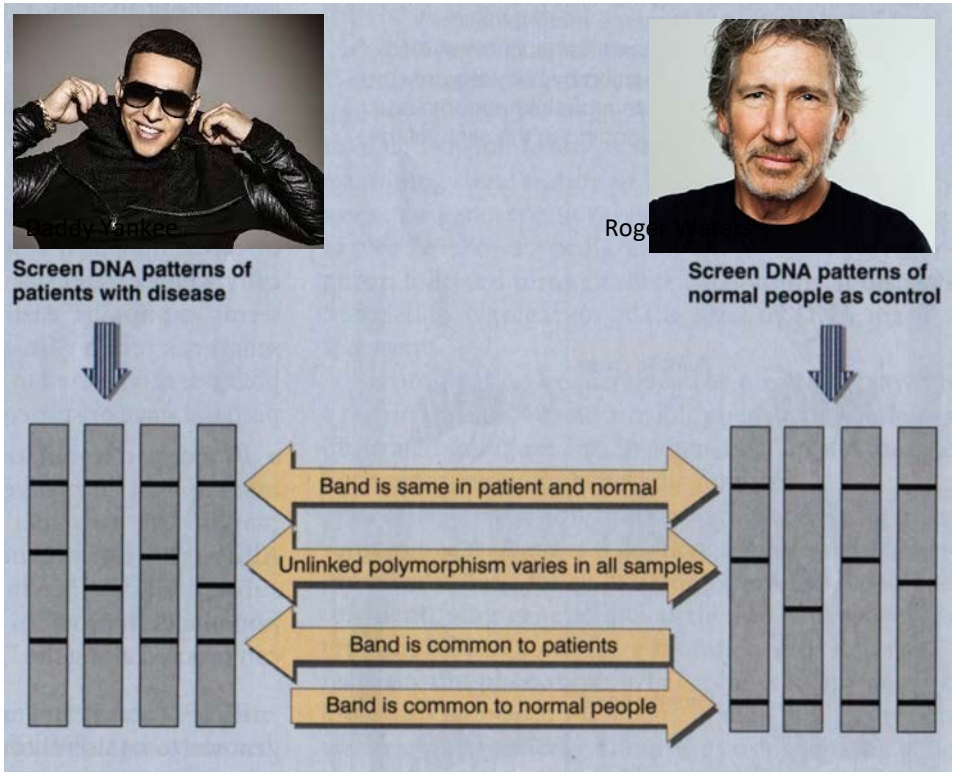


Polymorphism and molecular epidemiology

Thus, the RFLP of a normal person (wild-type) could differ from that of a sick person = a genetic marker for said pathology.

This is the general principle behind molecular epidemiology studies.

In perspective: the human genome has 4-5 million identified SNPs (as of 2018), which are separated by approximately 1 Kbp.



C-value Paradox

Genome size does not correlate with organism complexity.

Non-coding DNA proportion: humans have ~98% non-coding DNA, while some simple organisms have less.

Viral Genomes: Extremely compact, ranging from a few thousand to over a million base pairs, with minimal non-coding DNA.

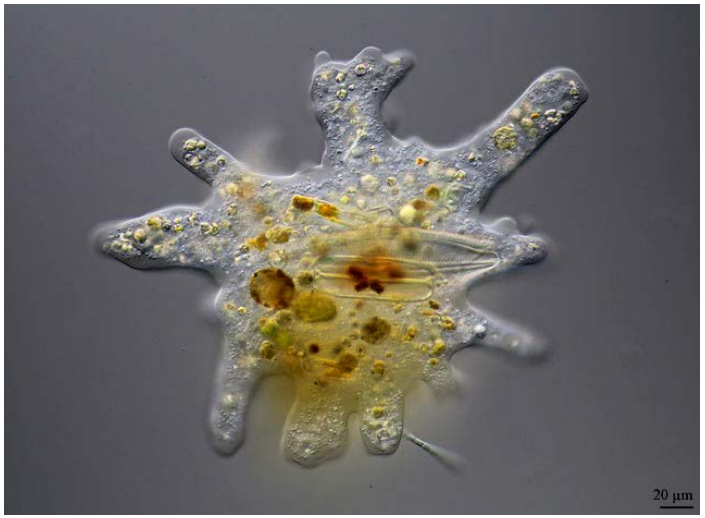
Gene Density: Smaller genomes often have higher gene density, while larger genomes tend to have more repetitive and non-coding sequences.

Comparative Insight: Genome size and structure provide insights into evolutionary history, adaptation, and complexity.

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
viruses			
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500	1 dsDNA
organelles			
mitochondria - *H. sapiens*	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
mitochondria – *S. cerevisiae*	86 kbp	8	1
chloroplast – *A. thaliana*	150 kbp	100	1
bacteria			
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470	1
H. pylori	1.7 Mbp	1,600	1
Cyanobacteria S. elongatus	2.7 Mbp	3,000	1
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700	1
B. subtilis	4.3 Mbp	4,100	1
S. cellulosum (largest known bacterial genome)	13 Mbp	9,400	1
archaea			
Nanoarchaeum equitans (smallest parasitic archaeal genome)	490 kbp	550	1
Thermoplasma acidophilum (flourishes in pH<1)	1.6 Mbp	1,500	1
Methanocaldococcus (Methanococcus) jannaschii (from ocean bottom hydrothermal vents; pressure >200 atm)	1.7 Mbp	1,700	1
Pyrococcus furiosus (optimal temp 100°C)	1.9 Mbp	2,000	1
eukaryotes - multicellular			
pufferfish *Fugu rubripes* (smallest known vertebrate genome)	400 Mbp	19,000	22
poplar *P. trichocarpa* (first tree genome sequenced)	500 Mbp	46,000	19
corn *Z. mays*	2.3 Gbp	33,000	20 (2n)
dog *C. familiaris*	2.4 Gbp	19,000	40
chimpanzee *P. troglodytes*	3.3 Gbp	19,000	48 (2n)
wheat *T. aestivum* (hexaploid)	16.8 Gbp	95,000	42 (2n=6x)
marbled lungfish *P. aethiopicus* (largest known animal genome)	130 Gbp	unknown	34 (2n)
herb plant *Paris japonica* (largest known genome)	150 Gbp	unknown	40 (2n)

C-value Paradox

Adder's-tongue Fern *(Ophioglossum)*
1400 Chromosomes



Amoeba dubia (*Polychaos dubium*)
670 Gbp (200x HoSa)

Tiny fern (*Tmesipteris oblanceolate*)
160 Gbp (50x HoSa)

Genomic complexity paradox

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: August 23, © 2024.)