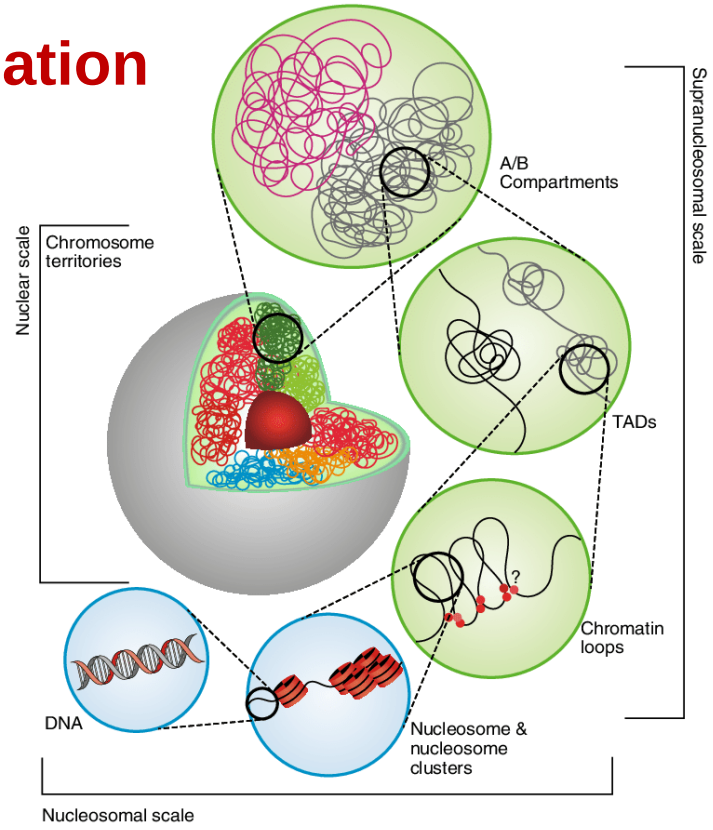
University of Duhok
College of Sciences
Department of Biology
4th Year Class

Lecture2. Genome Organization

Lecture outlines:

- ✓ What is Genome ?
- ✓ The C-value Paradox
- ✓ The Human Genome
- ✓ Bacterial Genomes
- ✓ Viral Genomes
- ✓ Types of Chromatin

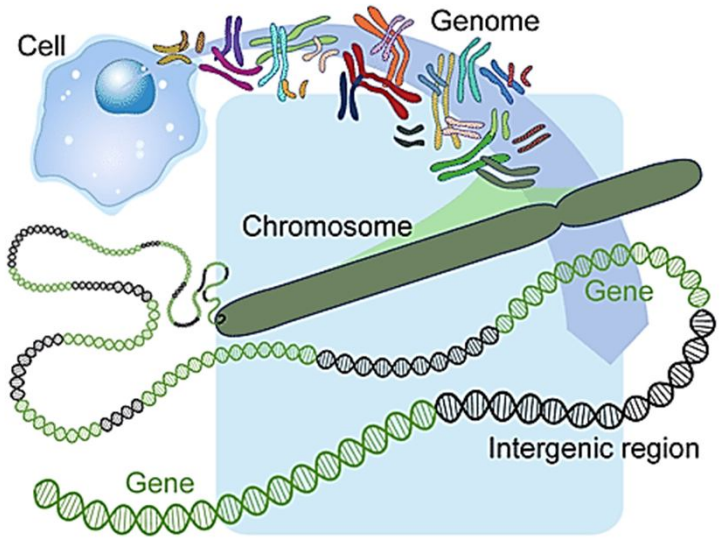


By Dr. Shaymaa Hadi Ali
Assistant Professor of Molecular Biology

Academic Year 2023-2024

What is Genome ?

- **Genome** is the entirety of an organism's hereditary information.
- It is encoded either in DNA or, for many types of virus, in RNA.
- The genome includes both the genes and the non-coding sequences of the DNA.
- Genome is thus the entire collection of genes and all other functional and non functional DNA sequences in an organism in a haploid set of chromosomes.
- It includes structural genes, regulatory genes and non functional nucleotide sequences as follow.



- **Structural genes:** DNA segments that code for some specific RNAs or proteins. Encode for mRNAs, tRNAs, snRNAs, scRNAs.
- **Functional sequences:** Regulatory sequences- occur as regulatory elements (initiation sites, promoter sites, operator sites,..etc.)
- **Nonfunctional sequences:** Introns and repetitive sequences. Needed for coding, regulation and replication of DNA.

C-value and the C-value Paradox

- The C-value is a measure of genome size, typically expressed in base pairs of DNA per haploid genome.
- The C-value paradox states that the organism with the largest genome is not necessarily the most complex and that genome size cannot be used as a predictor of genetic or morphological complexity.
- For example human and mice have a genome size of around 3 billion base pairs (3×10^9 bp). However the unicellular protozoa *Amoeba dubia* has a genome size of over 600 billion base pairs (6×10^{11} bp) about 200 times as big.
- The C-value paradox means that organisms with similar complexity may have very different genome sizes and conversely organisms with similar C-values may not be equally complex.
- The C-value in the prokaryotic kingdom, is a good predictor of metabolic complexity.
- In prokaryotic organisms like bacteria, genes are packed tightly together with very little non-coding DNA being present.
- The presence of varying amounts of the non-coding DNA “junk sequences” in different eukaryotic organisms explains how relatively simple organisms can have more DNA in their genomes than more complex ones. In the human genome, for example, there is on average only one gene for every 100 kb of sequence.

Comparative genome sizes and the C-value for a range of different organisms.

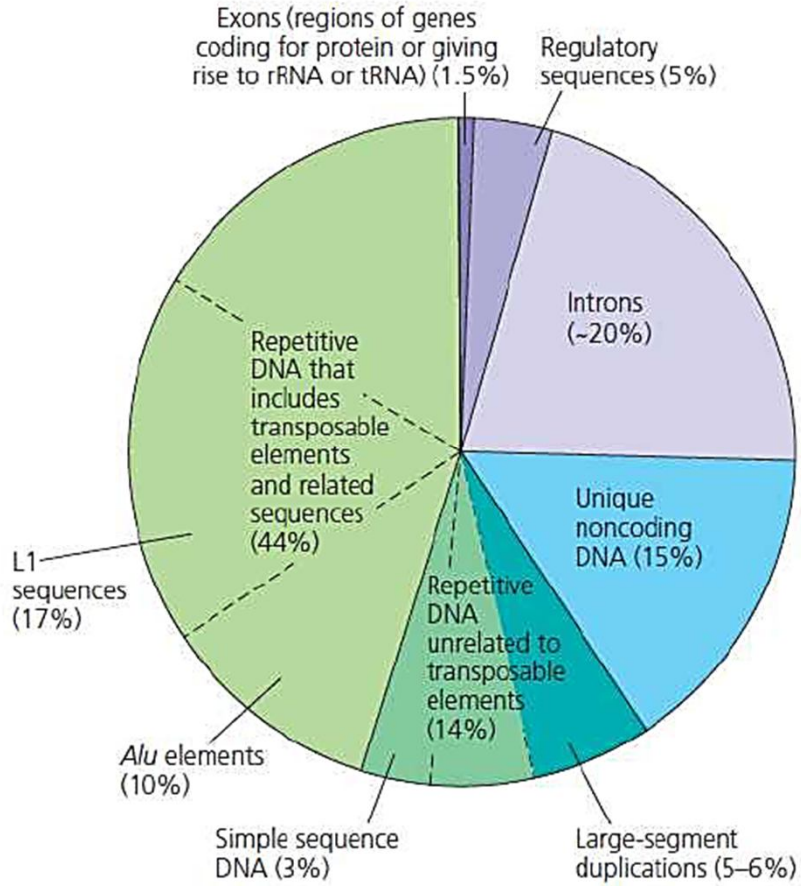
Characteristics of the genomes of example organisms			
Organism	Genome Size (bp)	Chromosome Number (*n*)	Predicted Number of Genes
Mycoplasma genitalium	580,000	1	500
Escherichia coli K12	4,639,000	1	4,500
Saccharomyces cerevisiae (yeast)	12,069,000	16	6,000
Caenorhabditis elegans (worm)	97,000,000	6	20,000
Drosophila melanogaster (fly)	137,000,000	6	15,000
Oryza sativa (rice)	420,000,000	12	40,000
Arabidopsis thaliana (weed)	115,000,000	5	28,000
Fugu rubripes (pufferfish)	390,000,000	22	25,000
Mouse	2,500,000,000	20	25,000
Humans	3,300,000,000	23	25,000

The Human Genome

- The human haploid nuclear genome contains $3x10^9$ base pairs, a vast amount of DNA.
- This DNA is organized into linear segments called the chromosomes (22 autosomal chromosomes and one sex chromosomes either X or Y), which vary in length from 47×10^6 bp to 246×10^6 bp.
- In addition to the nuclear genome, mitochondria also contain DNA: in humans, the mitochondrial genome is about 17×10^5 bp in length.
- Less than 2% of the human genome actually encodes protein. The remaining non-coding DNA is often referred to as junk DNA and in most cases probably serves no purpose.
- However, this junk DNA, some of which has arisen due to DNA duplication, has allowed the evolution of new genes and hence the generation of genetic variation.
- Approximately 5% of the mammalian genome is under selective pressure, i.e. 5% of their genome sequences have not changed significantly since mice, dogs and humans diverged from a common ancestor; this suggests that at least 3% of the non-coding DNA is conserved.
- The conserved noncoding DNA will include regulatory sequences required to control gene expression.

The Human Genome ……. continue

How is the genome organized?



Non-coding DNA

Non-coding DNA can be grouped into four main categories, namely:

1. **Introns:** Most eukaryotic genes are made up of stretches of DNA that code for amino acids, interrupted by non-coding sequences. The coding and noncoding sections are called exons and introns respectively.
2. **Simple sequence repeats:** A region of the genome where a DNA sequence is repeated many times in tandem is called a simple sequence repeat (SSR). These regions are described as satellite DNA, mini-satellite DNA and micro-satellite DNA depending on the number and length of the repeats. SSRs, make up 3% of the human genome.
3. **Interspersed transposon-derived repeats:** These repeated sequences are derived from transposons, mobile genetic elements capable of being copied and inserted into new sites in the genome which account for 45% of the human genome.
4. **Non-repeat, non-coding DNA:** approximately 50% of the genome made up of “unique” non-coding DNA, half of which will be present in introns. A small percentage of this non-coding DNA will contain regulatory sequences such as enhancers and promoters that are required for the control of gene transcription

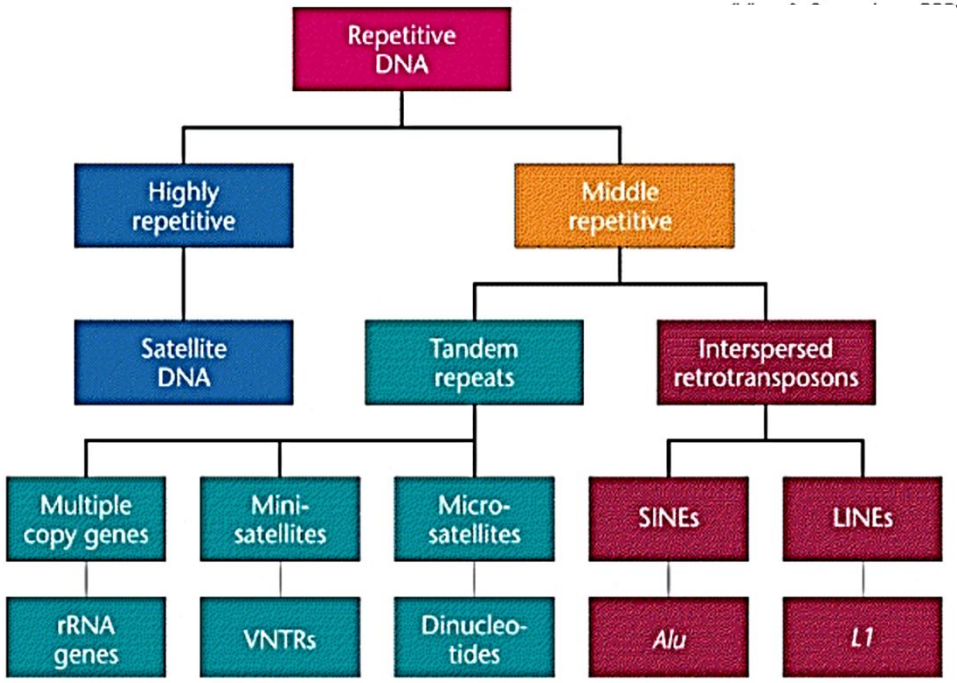
Types of the repeat regions

Depending on the size of the repeat, the repeat regions are classified into two groups:

1. Short tandem repeats (STRs): contain 2-5 base pair repeats.
2. Variable number of tandem repeats (VNTRs): have repeats of 9-80 base pairs.

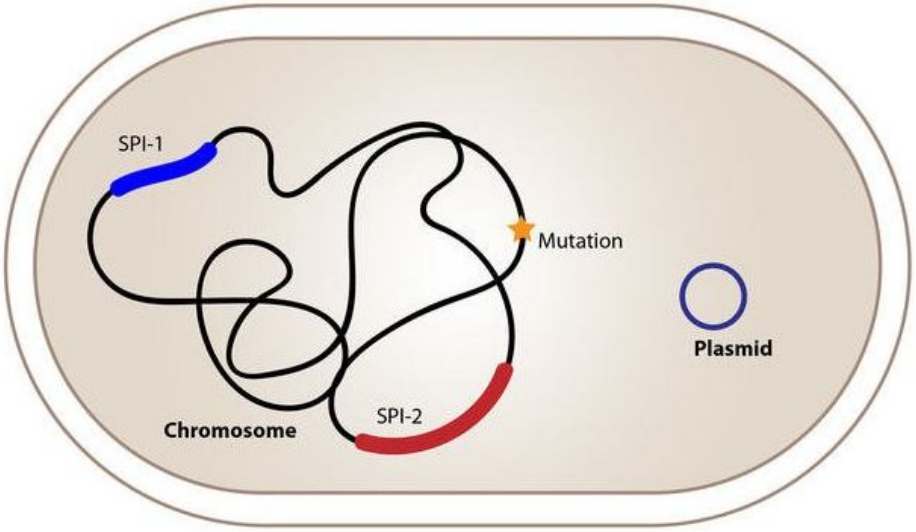
Categories of repetitive DNA in the human genome

- More than half of the human genome consists of non-protein-coding, repetitive DNA elements. These occur in several classes with specific characteristics. The most important of these are:
- Highly-repetitive DNAs that are detected as satellite DNA.
- Middle repetitive DNAs may be arranged either as **tandem repeats** where specific short DNA sequences are repeated in end-to-end arrays from a few to many tens of times, or as **interspersed repeats** in which a particular sequence occurs at hundreds or thousands of separate locations.



- Tandem microsatellites comprise two-letter motifs, minisatellites typically comprise four- or six-letters motifs; both occur in variable numbers per locus. Interspersed repeats may be short (**SINEs**) or long (LINEs), for which the characteristic human families are Alu (200~300 bp) and L1 (6,400 bp), respectively.

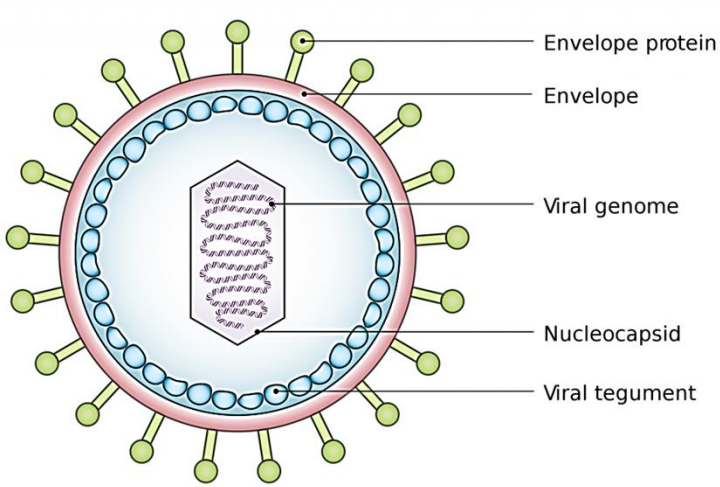
Bacterial Genomes



- Most bacterial genomes are single circular DNA molecules that are 0.5–10 Mb in size.
- The size of bacterial genomes varies considerably. However, for bacteria, genome size correlates well with gene number, which in turn correlates with morphological, physiological or metabolic complexity.
- Bacteria with small genomes encode a small number of genes and tend to be restricted to growth in relatively few specialized niches; they are often parasites.
- The average size of a gene in all bacteria is around 1000 bp.
- In Bacteria, Only about 10% of the genome is non-coding and even this 10% in most cases plays a critical role to the cell since it contains the control signals required to co-ordinate patterns of gene expression.
- Bacterial genes do not contain introns and the genes are much more closely packed

Viral Genomes

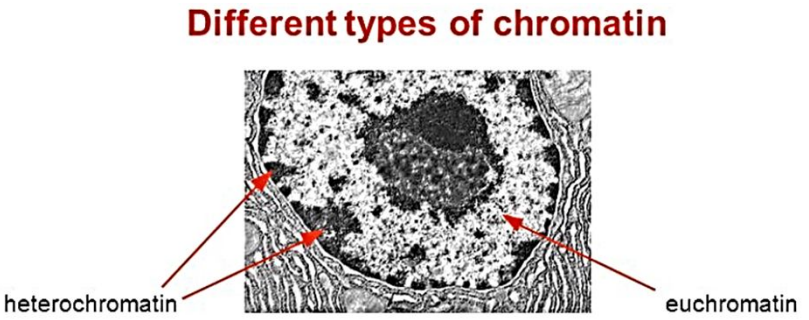
- Viral genomes consist of either RNA or DNA which can be either single- or double-stranded.
- Since viruses are obligate intracellular parasites they do not need to encode all the functions required by a free-living organism and accordingly they typically have very small genomes.
- The genome sizes of viruses, which may infect bacteria or eukaryotic cells, vary from 2 kb to 700 kb.



- Bacteriophage λ is a virus that infects various bacteria including *E. coli*. It has a 48 kb double-stranded DNA genome coding for 46 genes packaged into an icosahedral head.
- Retroviruses are a class of viruses which infect eukaryotic cells that contain a single-stranded RNA genome in a protein capsid surrounded by a membrane envelope. The simplest retroviruses contain an 8 kb genome encoding just three genes.

Types of Chromatin

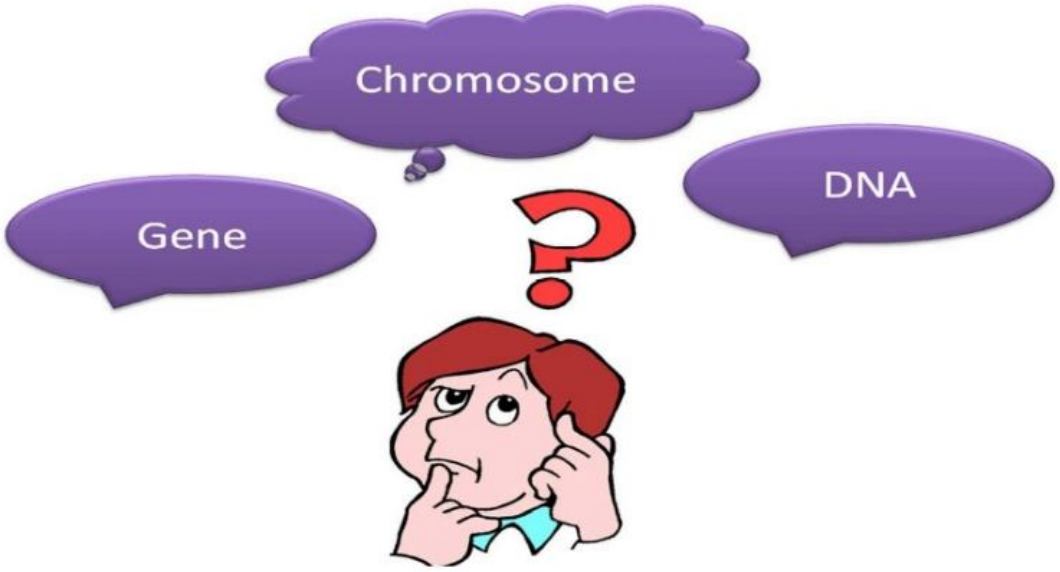
Euchromatin



- Lightly packed form of chromatin that is rich in gene concentration takes up light stain and represent most of the chromatin, that is less condensed and can be transcribed.
- Consists of structural genes which replicate and transcribe during G1 and S phase of the interphase.
- Considered genetically active chromatin, since it has a role in their phenotypic expression of the genes.
- DNA is found packed in 3-8 mm fiber. During metaphase it takes up dark stain.

Heterochromatin

- Tightly packed form of chromatin that takes up deep stain during interphase and prophase but metaphase takes up light stain.
- Heterochromatin consists of highly repetitive DNA sequences. It is late replicating during the s-phase of the cell, it is highly condensed and is typically not transcribed, e.g. centromeric regions .


Chromosome
DNA
Gene
?