NUCLEAR GENES & EXTRAGENIC DNA

Learning Objectives

- To be familiar with the concept of RNA-directed DNA synthesis
- To understand the difference between nuclear genes and extragenic DNA.
- To learn about unique single copy genes and multigene families.
- To understand difference between gene and pseudogene.
- To know the meaning of Tandemly Repeated DNA Sequences.
- To distinguish between satellite, minisatellite, and microsatellite DNA sequences.
- To learn abouts short & long interspersed nuclear elements

RNA-DIRECTED DNA SYNTHESIS

- The process of the transfer of the genetic information from DNA to RNA to protein has been called the **central dogma**.
- It was initially believed that genetic information was transferred only from DNA to RNA and thence translated into protein.
- However, there is evidence from the study of certain types of viruses - **retroviruses** - that genetic information can occasionally flow in the reverse direction, from RNA to DNA using the enzyme **reverse transcriptase**. This is referred to as **RNA-directed DNA synthesis**.
- It has been suggested that regions of DNA in normal cells serve as templates for the synthesis of RNA, which in turn then acts as a template for the synthesis of DNA that later becomes integrated into the nuclear DNA of other cells.
- *Homology between human and retroviral oncogene sequences could reflect this process, which could be an important therapeutic approach for the treatment of inherited disease in humans.*

Types of DNA Sequence

Analysis of the human DNA have shown the presence of two types of DNA sequences:

1) **single-** or **low-copy number DNA sequences**: represents about 60% to 70% of the human genome.
2) moderately or highly **repetitive DNA sequences**: consists about 30% to 40% of human genome. It represents DNA sequences that are not transcribed and consists of mainly satellite DNA and interspersed DNA sequences.

Nuclear Genes

- It is estimated that there are around 21,600 coding genes in the nuclear genome.
- The distribution of these genes varies greatly between chromosomal regions.
- For example, heterochromatic and centromeric regions are mostly noncoding, with the highest gene density observed in subtelomeric regions.
- Chromosomes 19 and 22 are gene rich, whereas chromosomes 4 and 18 are relatively gene poor.
- The size of genes also shows great variability: from small genes with single exons to the *TTN* gene which encodes the largest known protein in the human body (titin protein) and has the largest number of exons (363) in any known gene, and the single largest exon (17,106 bp).

Nuclear genes are divided into unique single copy genes and multigene families:

A) Unique Single-Copy Genes

Most human genes are unique single-copy genes coding for polypeptides that are involved in or carry out a variety of cellular functions. These polypeptides and proteins include enzymes, hormones, receptors, and structural and regulatory proteins.

B) Multigene Families

Multigene families refer to genes that have similar functions, having developed through gene duplication events with subsequent evolutionary differences**.** Some are found physically close together in clusters (for example, the α- and β-globin gene clusters on chromosomes 16 and 11, respectively), whereas others are widely dispersed throughout the genome, occurring on different chromosomes (such as the *HOX* homeobox gene family).

Multigene families can be divided into two types:

1) Classic Gene Families

They consist of genes that show a high degree of sequence homology (similarity). Examples of classic gene families include the **ribosomal RNAs gene families** (clustered as tandem arrays at the nucleolar organizing regions on the short arms of the five acrocentric chromosomes), and the different **transfer RNA gene families** (dispersed in numerous clusters throughout the human genome).

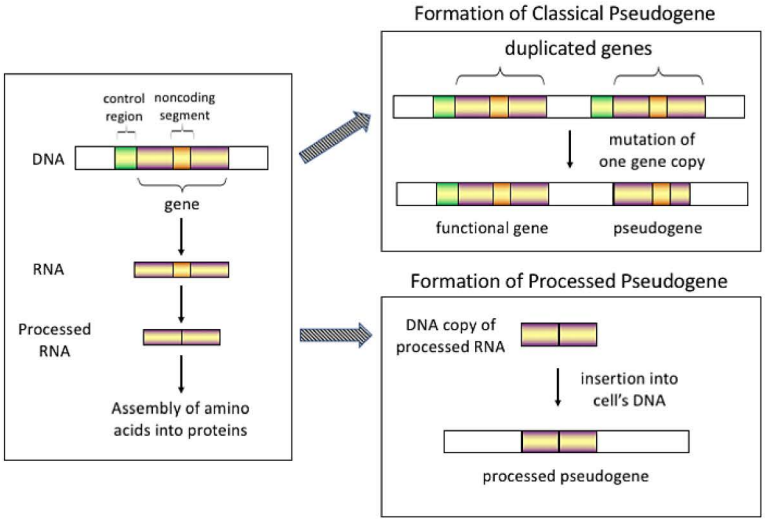
2) Gene Superfamilies

They consist of genes that have limited sequence homology but are functionally related, having similar structural domains. Examples of gene superfamilies include the **HLA** (human leukocyte antigen) **genes** on chromosome 6 and the **T-cell receptor genes** (which have structural homology with the immunoglobulin genes) on chromosomes 11 and 14.

Pseudogenes

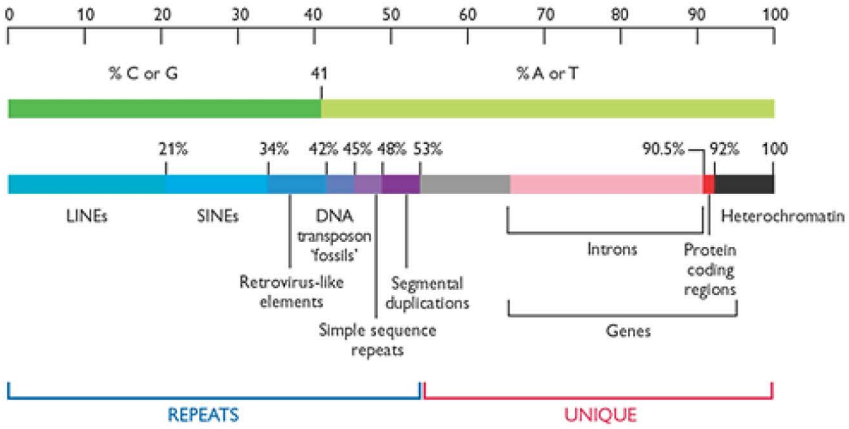
- **Pseudogenes** are genes that closely resemble known structural genes but which, in general, are not functionally expressed.
- These are thought to have arisen in two main ways:

1) genes undergoing duplication events that are rendered silent through the acquisition of mutations in coding or regulatory elements (**classical pseudogenes**),
2) as the result of the insertion of complementary DNA sequences, produced by the action of the enzyme **reverse transcriptase** on a naturally occurring messenger RNA transcript, that lack the promoter sequences necessary for expression (**processed pseudogenes**).



Extragenic DNA

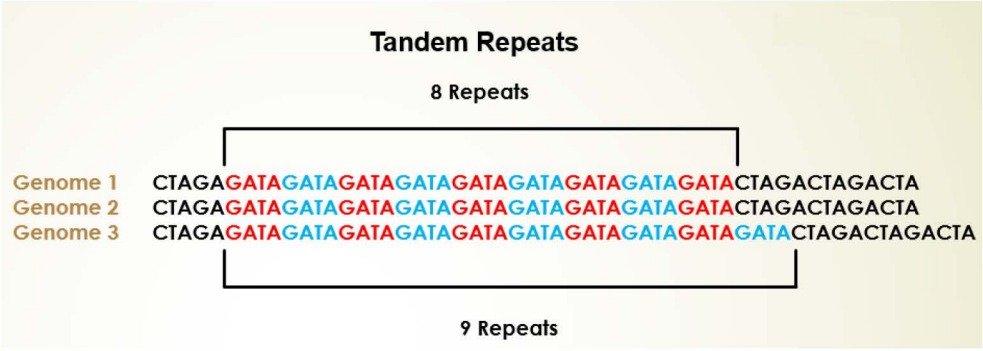
- The unique single-copy genes that encode proteins represent about 2% of the human genome.
- The remainder of the human genome is made up of repetitive DNA sequences that are predominantly transcriptionally inactive (non-coding), known as extragenic DNA.
- Extragenic DNA has been described previously as **junk** DNA, but recent studies have shown that extragenic DNA has multiple functions (most of them are still unknown).
- Some regions in the extragenic DNA show evolutionary conservation and play a critical role in the regulation of gene expression, such as enhancers and silencers sequences of DNA.



Components of Extragenic DNA

Tandemly Repeated DNA Sequences

- Tandemly repeated DNA sequences (pattern of two or more nucleotides is repeated and the repetitions are adjacent to each other) consist of blocks of tandem repeats of non-coding DNA that can be either highly dispersed or restricted in their location in the genome.



- Tandemly repeated DNA sequences can be divided into three subgroups: satellite, minisatellite, and microsatellite DNA.

1) *Satellite DNA:*

Satellite DNA accounts for approximately 10% to 15% of the repetitive DNA sequences of the human genome and consists of very large series of simple or moderately complex, short, tandemly repeated DNA sequences that are transcriptionally inactive and are clustered around the centromeres of certain chromosomes.

2) *Minisatellite DNA:*

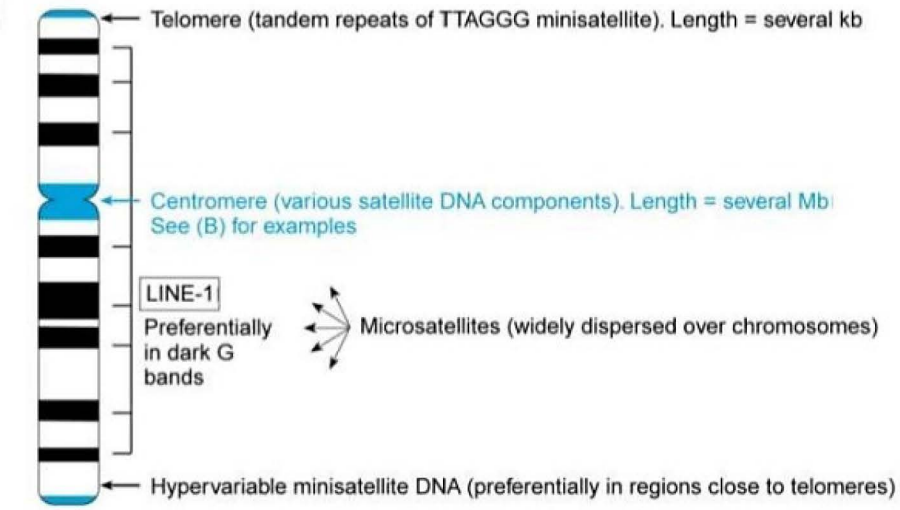
Minisatellite DNA consists of two families of tandemly repeated short DNA sequences: telomeric and hypervariable minisatellite DNA sequences that are transcriptionally inactive.

a) **Telomeric DNA:** The terminal portion of the telomeres of the chromosomes contains 10 to 15 kb of tandem repeats of a 6-base-pair (bp) DNA sequence (TTAGGG) known as telomeric DNA. The telomeric repeat sequences are necessary for chromosomal integrity in replication and are added to the chromosome by an enzyme known as telomerase.

b) **Hypervariable minisatellite DNA:** this is made up of highly polymorphic DNA sequences consisting of short tandem repeats of a common core sequence usually located near the telomeric region. They are also called **VNTRs** (variable number tandem repeats).

3) *Microsatellite DNA (or short tandem repeats – STRs):*

Microsatellite DNA consists of tandem single, di-, tri-, and tetra-nucleotide repeat base-pair sequences located throughout the genome. Microsatellite repeats rarely occur within coding sequences but trinucleotide repeats in or near genes are associated with certain inherited disorders. Nowadays DNA microsatellites are used for forensic and paternity tests. They can also be helpful for gene tracking in families with a genetic disorder but no identified mutation.



Highly Repeated Interspersed Repetitive DNA Sequences

- Approximately one-third of the human genome is made up of two main classes of short and long repetitive DNA sequences that are interspersed throughout the genome (DNA sequence repeats that are not adjacent to each other).

  1) ***Short Interspersed Nuclear Elements:***

     Approximately 12% of the human genome consists of some 750,000 copies of **short interspersed nuclear elements (SINEs).** The most common are DNA sequences of approximately 300 bp that have sequence similarity to a signal recognition particle involved in protein synthesis. They are called **Alu repeats** because they contain an *AluI* restriction enzyme recognition site.

  2) ***Long Interspersed Nuclear Elements:***

     Approximately 20% of the DNA of the human genome is made up of **long interspersed nuclear elements (LINEs)**. The most commonly occurring LINE, known as LINE-1 or an L1 element, consists of more than 100,000 copies of a DNA sequence of up to 6000 bp that encodes a reverse transcriptase.

- The function of these interspersed repeat sequences is not clear.
- The majority of this repetitive sequence is derived from unstable DNA sequences called transposable elements or **transposons.** These sequences insert additional copies of themselves randomly throughout the genome.
- These repetitive sequences permit, through the process of recombination, the rearrangement of parts of the genome, over time modifying the properties of existing genes and even creating new genes.
- Transposons move spontaneously throughout the genome from one chromosome location to another and appear to be universal in the plant and animal kingdoms.
- It has been postulated that Alu repeats could promote unequal recombination, which could lead to pathogenic mutations or provide selective advantage in evolution by gene duplication.
- Both Alu and LINE-1 repeat element insertions are reported causes of mutation in inherited human disease.