ORGANIZATION OF EUKARYOTIC GENOME

A genome is an organism's complete set of DNA, comprising of nuclear and mitochondrial DNA. Each genome contains all of the information needed to build and maintain that organism. A human haploid cell, consist of 23 nuclear chromosome and one mitochondrial chromosome, contains more than 3.2 billion DNA base pairs.

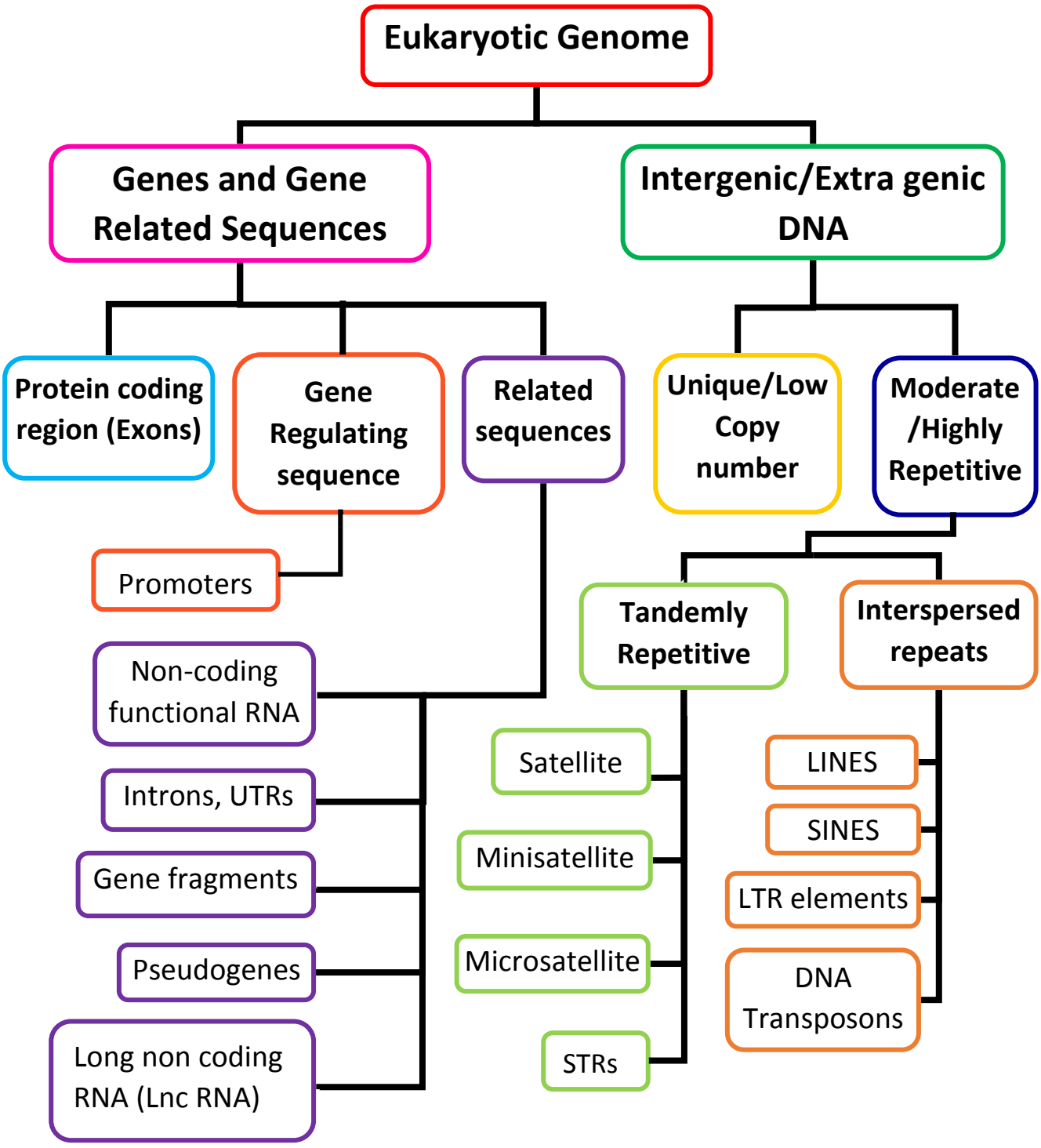
Eukaryotic genome is linear and conforms the Watson-Crick Double Helix structural model. Embedded in Nucleosome-complex DNA & Protein (Histone) structure that pack together to form chromosomes. Eukaryotic genome have unique features of Exon - Intron organization of protein coding genes, representing coding sequence and intervening sequence that represents the functionality of RNA part inside the genome.

Table 1: DNA content of Human Chromosomes

Chromosome	Total amount of DNA (Mb)	Amount of heterochromatin (Mb)	Chromosome	Total amount of DNA (Mb)	Amount of heterochromatin (Mb)
1	279	30	13	118	16
2	251	3	14	107	16
3	221	3	15	100	17
4	197	3	16	104	15
5	198	3	17	88	3
6	176	3	18	86	3
7	163	3	19	72	3
8	148	3	20	66	3
9	140	22	21	45	11
10	143	3	22	48	13
11	148	3	X	163	3
12	142	3	Y	51	27

Configuration of Eukaryotic genome:

The configuration of eukaryotic genome includes protein coding region, gene regulating region, gene related sequence and intergenic DNA or extra genic DNA which includes low copy number and moderate or high copy number repetitive sequence, the flow chart representation of configuration is given below:



1. GENES AND GENE RELATED SEQUENCES

A. PROTEIN CODING REGION (EXONS)

- Protein coding sequences are the DNA sequences that are transcribed into mRNA later translated to proteins.
- The complete protein coding genes capacity of the genome is contained within the exomes (the part of the genome formed by exons, the sequences which when transcribed remain within the mature RNA sequence after introns are removed using RNA splicing)and consists of DNA sequences encoded by exons that can be later translated into proteins.
- It consists of ORF (Open reading frame). These are the reading frame that has the potential to code for the proteins/peptide. It is stretch of codons that do not contain a stop codon (UAA, UAG, and UGA). An AUG with the ORF may indicate where translation starts.

B. GENE REGULATING SEQUENCES

Promoters are combinations of short sequence elements (usually located in the immediate upstream region of the gene often within 200 bp of the transcription start site) which serve to initiate transcription. They can be subdivided into different components.

- The **Core promoter** directs the basal transcription complex to initiate transcription of the gene. In the absence of additional regulatory elements it permits constitutive expression of the gene, but at very low (basal) levels. Core promoter elements are typically located very close to the transcription initiation site, at about nucleotide position -45 to +40. They include: the **TATA box** located at position ca. -25, surrounded by GC-rich sequences and recognized by the TATA- binding protein subunit of TFIID; the **BRE sequence** located immediately upstream of the TATA elements at around -35 and recognized by the TFIIB component; the **lnr (initiator) sequence** located at the start site of transcription and bound by TFIID; the **DPE** or **Downstream Promoter Element**, located at about position +30 relative to transcription and recognized by TFIID.
- The **proximal promoter region** is the sequence located immediately upstream of the core promoter, usually from -50 to -200 bp (promoter elements found further upstream would be said to map to the distal promoter region). They include: GC boxes (also called Sp1

boxes, the consensus sequence is GGGCGG which is often found in multiple copies within 100 bp of the transcription initiation site); CCAAT boxes typically located at position -75.

- **ENHANCERS** are positive transcriptional control elements which are particularly prevalent in the cells of complex eukaryotes such as mammals but which are absent or very poorly represented in simple eukaryotes such as yeast. They serve to increase the basal level of transcription which is initiated through the core promoter elements. Their function, unlike those of the core promoter, are independent of both their orientation and, to some extent, their distance from the genes they regulate. Enhancers often contain within a span of only 200-300 bp.
- **SILENCERS** serve to reduce transcription levels. Two classes have been distinguished: **classical silencers** (also called silencer elements) are position independent elements that direct an active transcriptional repression mechanism; **negative regulatory elements** are position-dependent elements that result in a passive repression mechanism. Silencer elements have been reported in various position: close to the promoter, some distance upstream and also within introns.
- **BOUNDARY ELEMENTS (INSULATORS)** are regions of DNA, often spanning from 0.5 kb to 3 kb, which function to block the spreading of the influence of agents that have a positive effect on transcription (enhancers) or negative one (silencers, hetero-chromatin-like repressive effects).
- **RESPONSE ELEMENTS** modulate transcription in response to specific external stimuli. They are usually located a short distance upstream of the promoter elements (often within 1kb of the transcription start site). A variety of such elements respond to specific hormones or to intracellular second messengers such as cyclic AMP.

C. RELATED SEQUENCES

1) NON CODING FUNCTIONAL RNA:

A non-coding RNA (ncRNA) is an RNA molecule that is not translated into a protein. Less-frequently used synonyms are non-protein-coding RNA (npcRNA), non-messenger RNA (nmRNA) and functional RNA (fRNA). The DNA sequence from which a functional non-coding RNA is transcribed is often called an RNA gene.

- **Below table represents the Human ncRNA**

MAJOR CLASSES OF NONCODING RNA				
RNA Class	**Subclass/ evolutionary or Functional sub family**	**No. Diff types**	**Function**	**Gene organization/biogenesis**
Ribosomal RNA (rRNA), ~120–5000 nucleotides	12S rRNA, 16S rRNA	1 of each	Components of mitochondrial ribosomes	Cleaved from multi genic transcripts produced by Heavy strand of mitochondrial DNA (mt-DNA)
	5SrRNA, 5.8S rRNA, 18S rRNA, 28S rRNA	1 of each	Components of cytoplasmic ribosomes	5S rRNA is present in chr1 is encoded by multiple genes in various gene clusters. 5.8S, 18S, and 28S rRNA are cleaved from multigenic transcripts .The multigenic 5.8S–18S–28S transcription units are tandemly repeated on each of 13p,14p,15p, 21p, and 22p chromosome (rDNA clusters)
Transfer RNA (tRNA), ~70–80 nucleotides	mitochondrial family	22	decode mitochondrial mRNA to make proteins on mitochondrial ribosomes	Single-copy genes. tRNA are cleaved from multigenicmt-DNA transcripts
	cytoplasmic family	49	decode mRNA produced by nuclear genes	700 tRNA genes and pseudogenes dispersed at multiple chromosomal locations with some large gene clusters
Small nuclear RNA (snRNA), ~60–360 nucleotides	spliceosomal family with subclasses Sm (smith antigen) and LSm*****	9	U1, U2, U4, U5, and U6 snRNA process standard GU–AG introns ; U4atac, U6atac, U11,	about 200 spliceosomal snRNA genes are found at multiple locations but there are moderately large clusters of U1 and

			and U12 snRNA process rare AU–AC introns	U2 snRNA genes; most are transcribed by RNA polymerase II
	non-spliceosomal snRNAs	several	U7 snRNA: 3' processing of histone mRNA; 7SKRNA*: general transcription regulator; Y RNA family: involved in chromosomal DNA replication and regulators of cell proliferation	mostly single-copy functional genes
Small nucleolar RNA (snoRNA), ~60–300 nucleotides	C/D box* class	246	maturation of rRNA, mostly nucleotide site-specific 2'-O-ribose methylations	Usually within introns of protein-coding genes; multiple chromosomal locations, but some genes are found in multiple copies in gene clusters (such as the HBII-52and HBII-85 clusters.(a ncRNA belongs to C/D box))
	H/ACA* class	96	maturation of rRNA by modifying uridines at specific positions to give pseudouridine	
Small Cajal body RNA (scaRNA)		25	maturation of certain snRNA classes in Cajal bodies (coiled bodies) in the nucleus	usually within introns of protein-coding genes
RNA ribonucleases, ~260–320 nucleotides		2	RNase P cleaves pre-tRNA in nucleus + mitochondria; RNase MRP cleaves rRNA in nucleolus and is involved in initiating mt-DNA replication	single-copy gene
Miscellaneous small cytoplasmic RNAs, ~80–500 nucleotides	BC200	1	neural RNA that regulates dendritic protein biosynthesis; originated from Alu repeat (short stretch of DNA characterized by action of Arthrobacter luteus restriction endonuclease)	1 gene, BCYRN1, at 2p16

	7SL RNA	3	component of the signal recognition particle (SRP) that mediates insertion of secretory proteins into the lumen of the endoplasmic reticulum	three closely related genes clustered on 14q22
	TERC (telomerase RNA component)	1	component of telomerase, the ribonucleoprotein that synthesizes telomeric DNA, using TERC as a template	single-copy gene at 3q26
	Vault RNA	3	components of cytoplasmic vault RNPs that have been thought to function in drug resistance	VAULTRC1, VAULTRC2, and VAULTRC3 are clustered at 5q31 and share ~84% sequence identity
	Y RNA	4	components of the 60 kD Ro ribonucleoprotein, an important target of humoral autoimmune responses	RNY1, RNY3, RNY4, and RNY5 are clustered at 7q36
MicroRNA (miRNA), ~22 nucleotides	> 70 families of related miRNAs	~2000	Multiple important roles in gene regulation, notably in development, and implicated in post-transcriptional gene regulation.	Interspersed in whole genome, pre miRNA is converted to miRNA by several enzyme and act on 3' UTR region of protein coding gene.
Piwi-binding RNA (piRNA), ~24–31 nucleotides	89 individual clusters	> 15,000	often derived from repeats; expressed only in germline cells, where they limit excess transposon activity	89 large clusters distributed across the genome; individual clusters span from 10 kb to 75 kb with an average of 170 piRNAs per cluster
Endogenous short interfering RNA (endosiRNA),	many	Endogenous short interfering RNA (endosiRNA),	often derived from pseudogenes, inverted repeats, etc.; involved in gene regulation in somatic cells and may also be involved in	clusters at many locations in the genome

~21–22 nucleotides		~21–22 nucleotides	regulating some types of transposon	
Long noncoding regulatory RNA, often > 1 kb	many	> 3000	involved in regulating gene expression; some are involved in monoallelic expression (X-inactivation, imprinting), and/or as antisense regulators	usually individual gene copies; transcripts often undergo capping, splicing, and polyadenylation but antisense regulatory RNAs are typically long transcripts that do not undergo splicing

***LSm =** These are a family of RNA-binding proteins. The various kinds of LSm rings function as scaffolds or chaperones for RNA oligonucleotides, assisting the RNA to assume and maintain the proper three-dimensional structure.

***7SK =** It plays a role in regulating transcription by controlling the positive transcription elongation factor P-TEFb. 7SK is found in a small nuclear ribonucleoprotein complex (snRNP) with a number of other proteins that regulate the stability and function of the complex.

***C/D box =** C/D box snoRNAs contain two short conserved sequence motifs, C (RUGAUGA) and D (CUGA), located near the 5' and 3' ends of the snoRNA, respectively. Short regions (~ 5 nucleotides) located upstream of the C box and downstream of the D box are usually base complementary and form a stem-box structure, which brings the C and D box motifs into close proximity.

***H/ACA =** H/ACA box snoRNAs have a common secondary structure consisting of a two hairpins and two single-stranded regions termed a hairpin-hinge-hairpin-tail structure. H/ACA snoRNAs also contain conserved sequence motifs known as H box (consensus ANANNA) and the ACA box (ACA). Both motifs are usually located in the single-stranded regions of the secondary structure.

1) INTRONS, UTRS

The term intron refers to both the DNA sequence within a gene and the corresponding sequence in RNA transcripts. At least four distinct classes of introns have been identified.

- Introns in nuclear protein-coding genes that are removed by spliceosomes (spliceosomal introns)
- Introns in nuclear and archaeal transfer RNA genes that are removed by proteins (tRNA introns)
- Self-splicing group I introns that are removed by RNA catalysis.
- Self-splicing group II introns that are removed by RNA catalysis

- Group III introns are proposed to be a fifth family, but little is known about the biochemical apparatus that mediates their splicing. They appear to be related to group II introns, and possibly to spliceosomal intron.

In molecular genetics, an untranslated region (or UTR) refers to either of two sections, one on each side of a coding sequence on a strand of mRNA. If it is found on the 5' side, it is called the 5' UTR (or leader sequence), or if it is found on the 3' side, it is called the 3' UTR (or trailer sequence).

2) GENE FRAGMENTS

Gene fragments are pieces of genes containing only the exons (those parts of the gene which actually encode the protein sequence). They are composed of cDNA.

3) PSEUDOGENES:

Pseudogenes are dysfunctional relatives of genes that have lost their gene expression in the cell or their ability to code protein. Pseudogenes often result from the accumulation of multiple mutations within a gene whose product is not required for the survival of the organism. Depending on their DNA sequence characteristics pseudogenes are mainly of two types:

- **Processed pseudogene:** They have all the normal parts of a protein-coding gene, but was originally thought to be incapacitated based on presumed DNA code errors.
- **Unprocessed pseudogene:** They lacks the intervening non-protein coding sequences called introns, which are typically spliced out when a messenger RNA (mRNA transcript) is produced from a gene.

4) LONG NON CODING RNA (LNC RNA)

Long non-coding RNAs (lncRNA) are a type of non-coding RNAs (ncRNAs) that exceed 200 nucleotides in length. lncRNAs are a relatively abundant component of the mammalian transcriptome and have been implicated in several cellular functions, including the regulation of gene transcription through the recruitment of chromatin-modifying enzymes.

Human genome contains many thousands of Lnc-RNA these is again divided into following categories:

a. **MACRO Lnc-RNA**: Lnc-RNA particularly lying in imprinted clusters, predominantly unspliced “macro” transcripts that can also show a low level of splicing, and both the unspliced and spliced transcript have the potential to be functional. Three known imprinted macro lncRNAs have been shown to silence from three to ten genes in cis in imprinted gene clusters.
b. **BIDIRECTIONAL Lnc-RNA:** It is oriented head to head with a protein coding gene within 1 kb. It exhibits a similar expression patterns to its protein coding counterpart that suggests that may be subject to share a regulatory pressure. Non coding locus that originates from within the promoter region of a protein-coding gene, with transcription proceeding in the opposite direction on the other strand.
c. **SENSE- OVERLAPPING LncRNA:** These can be considered transcript variants of protein - coding RNA as they overlap with a known annotated gene on the same genomic strand. majority of these lack substantial orf for protein translation while others have orfs that shares the same start codon as protein coding transcript for that gene, but encode protein for several reasons including non-sense mediated decay issues that limit the translation of mRNAs with premature termination stop codons and triggered NMD mediated destruction of the m RNA or an upstream alternative open reading frame which inhibits the translation of the predicted ORF.
d. **SENSE- INTRONIC LncRNA:** Sense intronic reside within of a coding gene, but do not intersect any exons.
e. **RETAINED- INTRON:** An Alternative spliced transcript that contain transcribed intronic sequence with respect to isoform of the locus, coding or other variants.
f. **NC_RNA HOST:** It has separate non coding transcript that host non coding micro RNA. Such as MIRs.
g. **LINC RNA (LONGINTERGENIC_RNA**): "Intergenic" refers to long non-coding RNAs that are transcribed from non-coding DNA sequences between protein-coding genes. Requires lack of coding potential and may not be conserved between species. These have length >200bp. they are named according to the 3- protein coding gene nearby.
h. **ANTISENSE:** RNA molecules that are transcribed from the antisense strand and overlap in part well defined spliced sense or intron less sense RNA. Antisense -overlapping lncRNA have the tendency to undergo fewer splicing events and typically show lower abundance than sense transcript.
i. **AMBIGUOUS _ORF:** Has transcripts that are believed to be protein coding, but have more than one possible open reading frame.

j. **3 PRIME_OVERLAPPING_NCRNA:** Has transcripts where ditag and/or published experimental data strongly supports the existence of long (>200bp) non-coding transcripts that overlap the 3'UTR of a protein-coding locus on the same strand.

k. **NONCODING:** Contains transcripts which are known from the literature to not be protein coding.

2. INTERGENIC /EXTRAGENIC DNA

An Intergenic region (IGR) is a stretch of DNA sequences located between genes. Intergenic regions are a subset of Noncoding DNA. Occasionally some intergenic DNA acts to control genes nearby, but most of it has no currently known function. It is sometimes referred to as junk DNA.

In humans, intergenic regions comprise about 75% of the genome, whereas this number is much less in bacteria (15%) and yeast (30%)

Intergenic regions are different from intragenic regions (or introns), which are short, non-coding regions that are found within genes, especially within the genes of eukaryotic organisms.

Do contain functionally important elements such as promoters and enhancers. Also intergenic regions may contain as yet unidentified genes such as noncoding RNAs. They are thought to have regulatory functions.

A. UNIQUE/LOW COPY NUMBER

Low Copy Number (LCN) is a DNA profiling or DNA testing technique developed by the Forensic Science Service (FSS) which has been in use since 1999.

LCN is an extension of Second Generation Multiplex Plus (SGM Plus) profiling technique. It is a more sensitive technique because it involves a greater amount of copying via polymerase chain reaction (PCR) from a smaller amount of starting material, meaning that a profile can be obtained from only a few cells, which may be as small as a millionth the size of a grain of salt, and amount to a just few cells of skin or sweat left from a fingerprint.

B. MODERATE/HIGHLY REPETITIVE

Genome contain some repetitive DNA sequences, including repetitive coding DNA. However, the majority of highly repetitive DNA sequences occur outside genes. Some of the sequences are present at certain sub chromosomal regions as large arrays of tandem repeats. This

type of DNA, known as heterochromatin, remains highly condensed throughout the cell cycle and does not generally contain genes.

1) TANDEMLY REPETITIVE REGION

Highly repeated noncoding human DNA often occurs in arrays (or blocks) of tandem repeats of sequence which may be a simple one (1-10 nucleotides), or a moderately complex one (tens to hundreds of nucleotides). Individual arrays can occur at a few or many different chromosomal locations.

1(a) Satellite DNA: Satellite DNA is transcriptionally inactive as is the vast majority of minisatellite DNA, but in the case of microsatellite DNA a significant percentage is located in coding DNA.

- Tandem repeats occur in DNA when a pattern of one or more nucleotides is repeated and the repetitions are directly adjacent to each other.

Table 2: Major classes of tandemly repeated Human DNA

Class	Size of repeat unit (bp)	Major chromosomal location(s); transcriptional status
Satellite DNA (arrays often within 100kb to several Mb size range	5-171	Especially at centromeres; not transcribed
α (alphoid DNA)	171	Centromeric heterochromatin of all chromosomes
β (Sau3A family)	68	Notably the centromeric heterochromatin of 1, 9, 13, 14, 15, 21, 22 and Y
Satellite 1 (AT-rich)	25-48	Centromeric heterochromatin of most chromosomes and other heterochromatic regions
Satellites 2 and 3	5	Most, possibly all, chromosomes
Minisatellite DNA (arrays often within the 0.1-20 kb range)	9-64	At or close to telomeres of all chromosomes; vast majority not transcribed
Telomeric family	6	All telomeres
Hypervariable family	9-64	All chromosomes, often near telomeres
Microsatellite DNA (= simple sequence repeats, SSR) (arrays typically < 100bp)	12	Dispersed throughout all chromosomes; some small arrays of very simple sequence

1(b) Minisatellites DNA: It comprises of a collection of moderately sized arrays of tandemly repeated DNA sequence which are dispersed over considerable portions of nuclear genome.

- A minisatellite, is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 10–60 base pairs) are typically repeated 5-50 times. Minisatellites occur at more than 1,000 locations in the human genome and they are notable for their high mutation rate and high diversity in the population.
- Minisatellites are prominent in the centromeres and telomeres of chromosomes, the latter protecting the chromosomes from damage.
- Hypervariable minisatellite DNA sequences are highly polymorphic and are organized in over 1000 arrays (from 0.1 to 20 kb long) of short tandem repeats. its significance is not clear, although it has been reported to be a 'hotspot' for homologous recombination in human cells.

1(c) Microsatellites DNA: Microsatellite DNA, also called simple sequence repeats (SSR), are small arrays of tandem repeats of a simple sequence (usually less than 10 bp). They are interspersed throughout the genome, accounting for over 60 Mb (2% of genome), and are thought to have arisen mostly by replication slippage.

A microsatellite is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 2–5 base pairs) are repeated, typically 5-50 times. Microsatellites occur at thousands of locations in the human genome and they are notable for their high mutation rate and high diversity in the population. Microsatellites are often referred to as short tandem repeats (STRs) by forensic geneticists, or as simple sequence repeats (SSRs) For example, the sequence TATATATATA is a dinucleotide microsatellite.

1(d) STRS: Short tandem repeat is a microsatellite, consisting of a unit of two to thirteen nucleotides repeated hundreds of times in a row on the DNA strand.

2) INTERSPERSED REPEAT

Interspersed repeats mainly come from transposable elements (TEs), but they also include some protein coding gene families and pseudogenes. Transposable elements are able to integrate into the genome at another site within the cell. It is believed that TEs are an important driving force on genome evolution of higher eukaryotes. TEs can be classified into two categories, Class 1 (retrotransposons) and Class 2 (DNA transposons).

2(a) LINES: LINEs (long interspersed nuclear elements) have been very successful transposons. They have a comparatively long evolutionary history, occurring in other mammals, including mice. As autonomous transposons, they can make all the products needed for retro transposition, including the essential reverse transcriptase. Human LINEs consist of three distantly related families: LINE-1, LINE-2, and LINE-3, collectively comprising about 20% of the genome .They are located primarily in euchromatic regions and are located preferentially in the dark AT-rich G bands (Giemsa-positive) of metaphase chromosomes.

2(b) SINES: SINEs (short interspersed nuclear elements) are retrotransposons about 100–400 bp in length. They have been very successful in colonizing mammalian genomes, resulting in various interspersed DNA families, some with extremely high copy numbers. Unlike LINEs, SINEs do not encode any proteins and they cannot transpose independently. However, SINEs and LINEs share sequences at their 3' end, and SINEs have been shown to be mobilized by neighboring LINEs. By parasitizing on the LINE element transposition machinery, SINEs can attain very high copy numbers. The human Alu family is the most prominent SINE family in terms of copy number, and is the most abundant sequence in the human genome, occurring on average more than once every 3 kb.

2(c) LTR ELEMENTS: LTR transposons include autonomous and non-autonomous retrovirus-like elements that are flanked by long terminal repeats (LTRs) containing necessary transcriptional regulatory elements. Endogenous retroviral sequences contain gag and pol genes, which encode a protease, reverse transcriptase, RNAse H, and integrase. They are thus able to transpose independently. There are three major classes of human endogenous retroviral sequence (HERV), with a cumulative copy number of about 240,000, accounting for a total of about 4.6% of the human genome.

2(d) DNA TRANSPOSONS: The cut-and-paste transposition mechanism of class II TEs does not involve an RNA intermediate. The transpositions are catalyzed by several transposase enzymes. Some transposases non-specifically bind to any target site in DNA, whereas others bind to specific DNA sequence targets. DNA transposons have terminal inverted repeats and encode a transposase that regulates transposition. They account for close to 3% of the human genome. Ritually all the resident human DNA transposon sequences are no longer active; they are therefore transposon fossils. DNA transposons tend to have short lifespans within a species, unlike some of the other transposable elements such as LINEs. However, quite a few functional human genes seem to have originated from DNA transposons, notably genes encoding the RAG1 and RAG2 recombinases and the major centromere-binding protein CENPB.