Human genome-

Organization, distribution of genes.

The single chromosome (DNA + proteins) of a prokaryote contains all the directions for making a living cell. The chromosome of a prokaryopte is not organized into a nucleus, although it is generally located in one part of the cell).

All the strands are part of the one chromosome, making a loop.

Prokaryote genomes

- Example: E. coli
- 89% coding
- 4,285 genes
- 122 structural RNA genes
- Prophage remains
- Insertion sequence elements
- Horizontal transfers

Prokaryotic genome organization:

- Haploid circular genomes (0.5-10 MB, 500-10000 genes)
- Operons: polycistronic transcription units
- Environment-specific genes on plasmids and other types of mobile genetic elements
- Usually asexual reproduction, great variety of recombination mechanisms
- Transcription and translation take place in the same compartment

Eukaryotic genome

- Example: C. elegans
- 10 chromosomes
- 19,099 genes
- Coding region – 27%
- Average of 5 introns/gene
- Both long and short duplications

Eukaryotic genome organization

- Multiple genomes: nuclear, plastid genomes: mitochondria, chloroplasts
- Plastid genomes resemble prokaryotic genomes

More about the nuclear genome:

- Multiple linear chromosomes, total size 5-10'000 MB, 5000 to 50000 genes
- Monocistronic transcription units
- Discontinuous coding regions (introns and exons)
- Large amounts of non-coding DNA
- Transcription and translation take place in different compartments
- Variety of RNA genes: rRNA, tRNA, snRNA (small nuclear), sno (small nucleolar), microRNAs, etc.
- Often diploid genomes and obligatory sexual reproduction
- Standard mechanism of recombination: meiosis

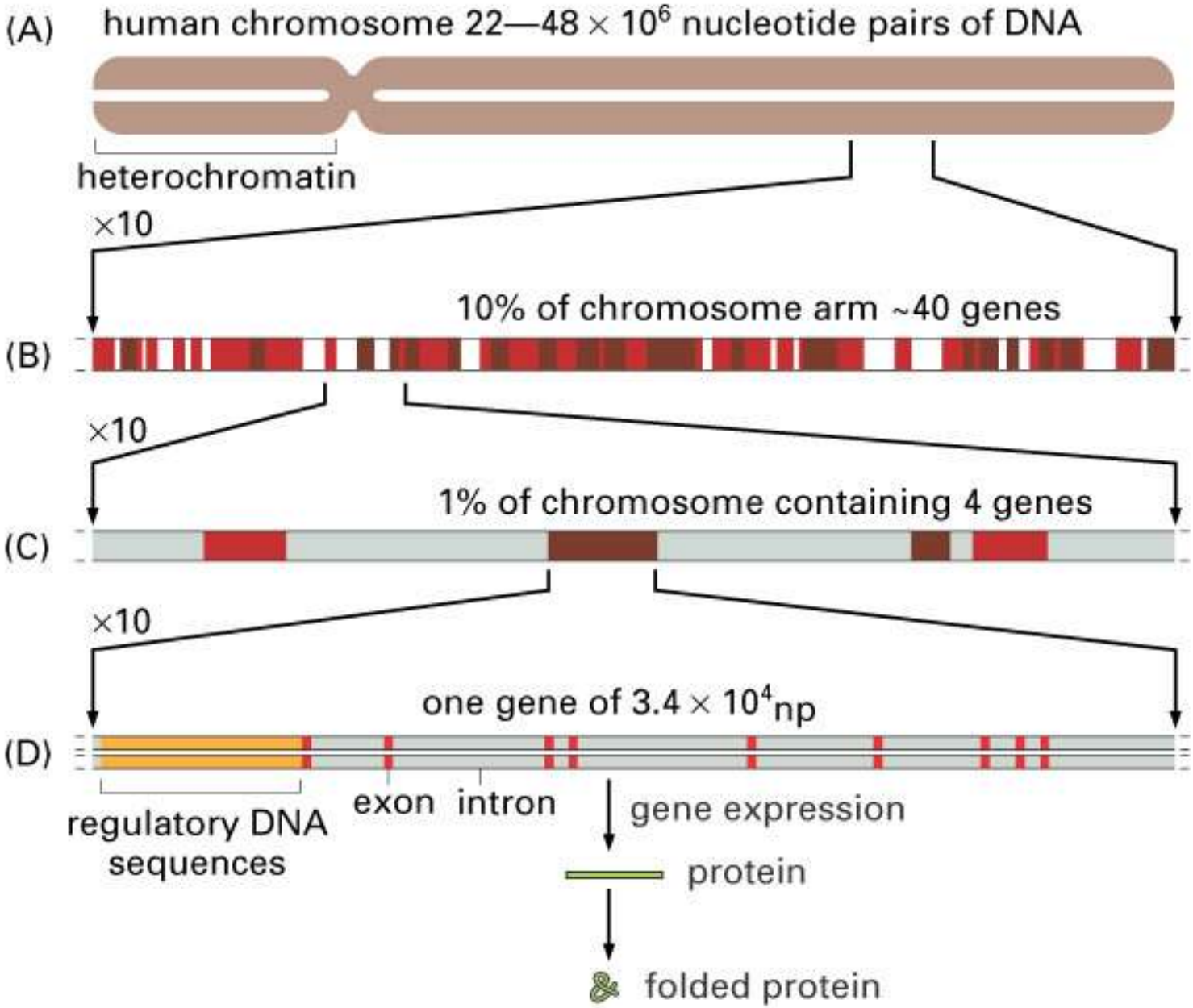
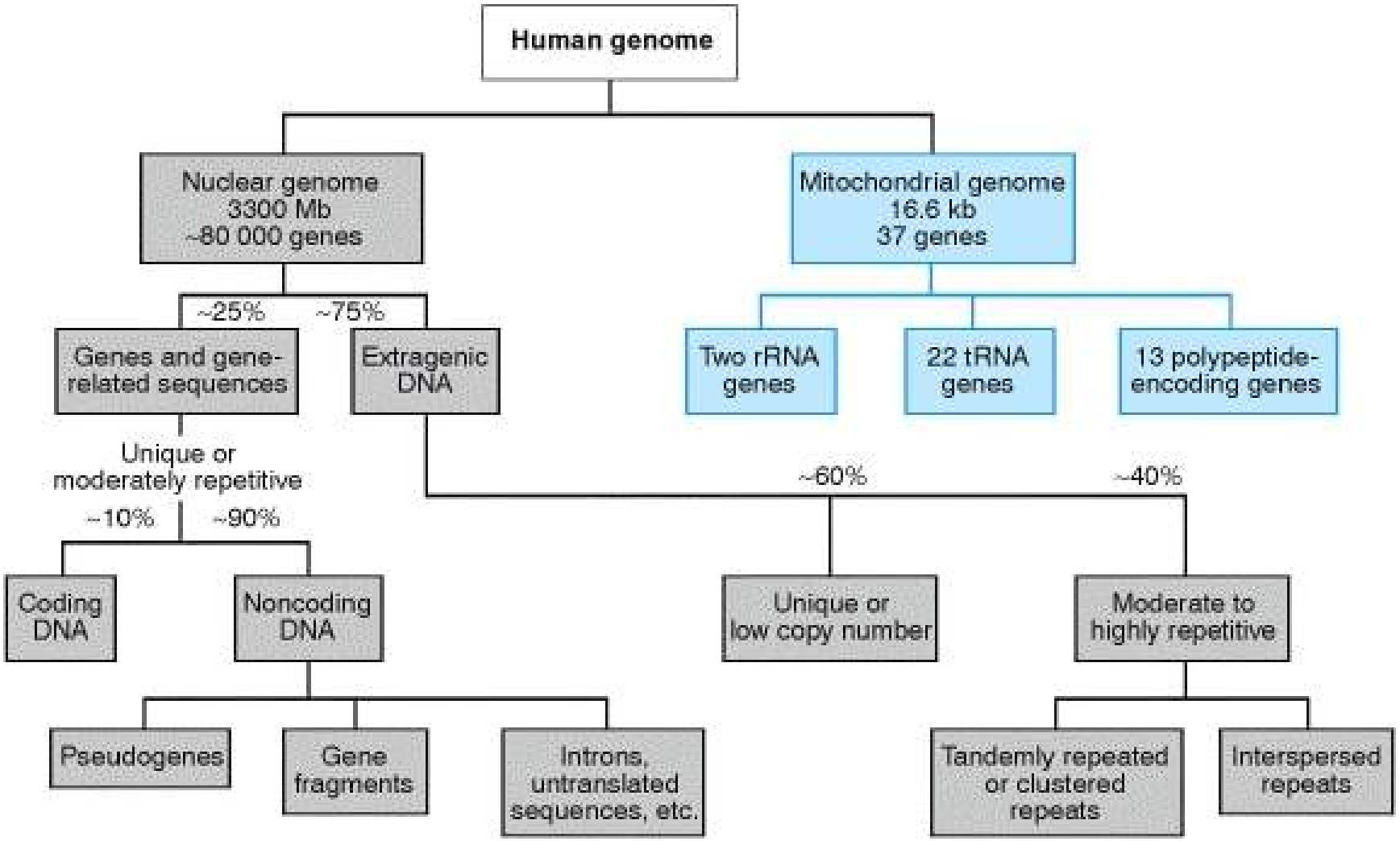


Figure 9-25 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Human genome organization



- The human genome is the term used to describe the total genetic information (DNA content) in human cells.
- It really comprises two genomes: a complex nuclear genome which accounts for 99.9995% of the total genetic information, and
- a simple mitochondrial genome which accounts for the remaining 0.0005% .

nuclear genome

- accounts for 99.9995% of the total genetic information
- polypeptide synthesis on cytoplasmic ribosomes
- sizeable component of the human genome- non coding DNA
- repetitive DNA, including both noncoding repetitive DNA and multiple copy genes and gene fragments.

Nuclear genome

- The nucleus - 99% of the cellular DNA.
- The nuclear genome is distributed between 24 different types of linear double-stranded DNA molecule, each of which has histones and other nonhistone proteins bound to it, constituting a chromosome.
- The 24 different chromosomes (22 types of autosome and two sex chromosomes, X and Y) can easily be differentiated by chromosome banding techniques and have been classified into groups largely according to size and, to some extent, centromere position
- In addition to the primary constriction (centromere) present on each chromosome, the long arms of chromosomes 1, 9 and 16 possess so-called secondary constrictions (light staining, apparently uncoiled chromosomal regions) which, like the centromeres, are composed of constitutive heterochromatin
- By comparison with the size of a mitochondrial DNA molecule, an average size human chromosome has an enormous amount of DNA, approximately 130 Mb on average, but varying between approximately 50 and 260 Mb.
- Base composition of nuclear genome is approximately GC 42%

DNA content of human chromosomes

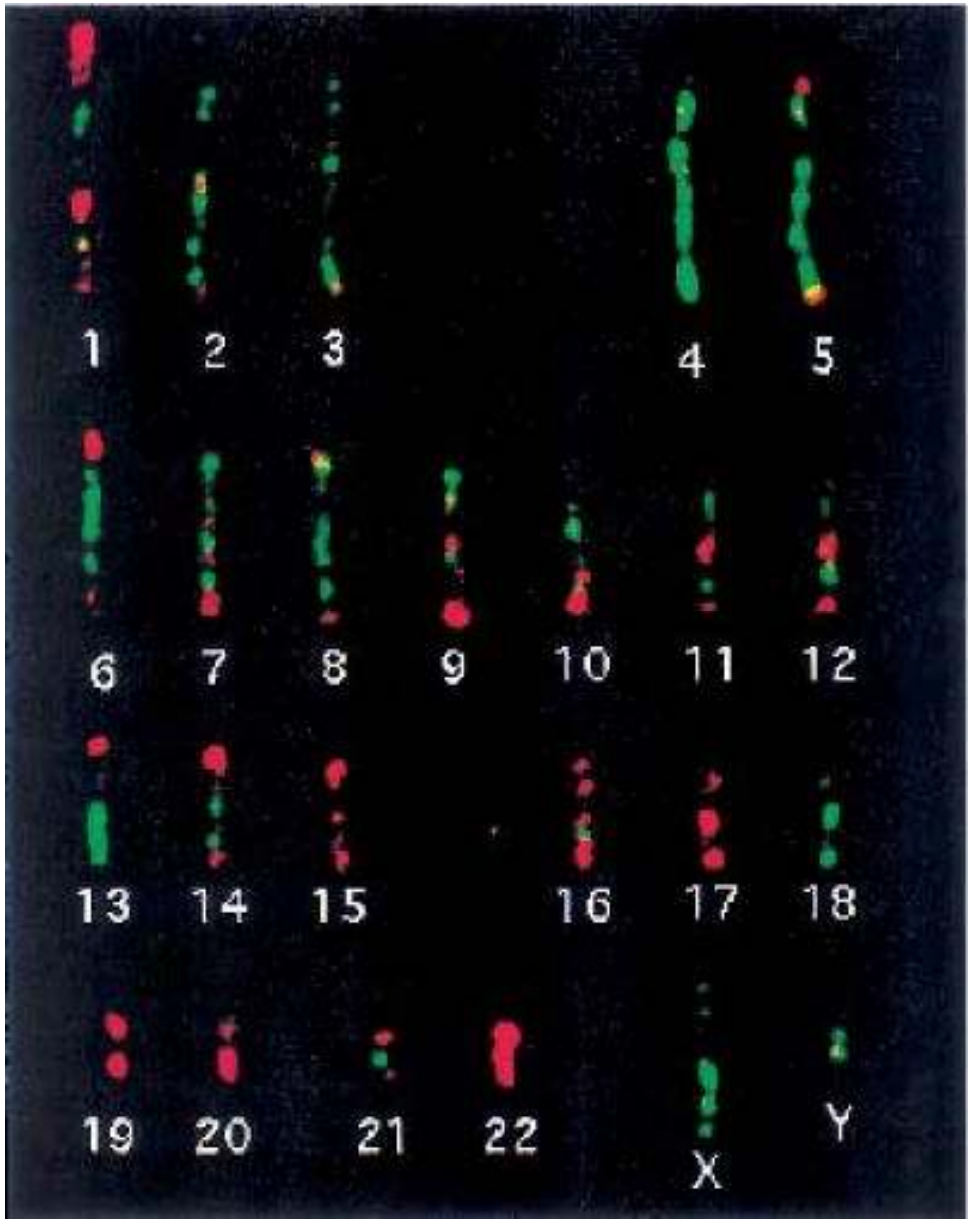
Chromosome	Amount of DNA (Mb)	Chromosome	Amount of DNA (Mb)
1	263	13	114
2	255	14	109
3	214	15	106
4	203	16	98
5	194	17	92
6	183	18	85
7	171	19	67
8	155	20	72
9	145	21	50
10	144	22	56
11	144	X	164
12	143	Y	59

Properties of chromosome bands seen with standard Giemsa staining

Dark bands (G bands)	**Pale bands (correspond to R bands)**
Stain strongly with dyes that bind preferentially to AT-rich regions, such as Giemsa and Quinacrine	Stain weakly with Giemsa and Quinacrine
May be comparatively AT-rich	May be comparatively GC-rich
DNase insensitive	DNase sensitive
Condense early during the cell cycle but replicate late	Condense late during cell cycle but replicate early
Gene poor. Genes may be large because exons are often separated by very large introns	Gene rich. Genes are comparatively small because of close clustering of exons
LINE rich, but may be poor in *Alu* repeats	LINE poor, but may be enriched in *Alu* repeats

Nuclear genome contd..

- The total number of genes in the human genome has been estimated to be about 70 000–80 000.
- As all but 37 of these genes are located in the nuclear genome, this gives a rough estimate of about 3000 genes per chromosome.
- However, gene density can vary substantially between chromosomal regions and also between whole chromosomes.
- For example, heterochromatic regions are known to be very largely composed of repetitive noncoding DNA, and the centromeres and large regions of the Y chromosome, in particular, are notably devoid of genes.



FISH of a CpG (that is, neighboring cytosine and guanine residues on the same DNA strand in the 5′ → 3′ direction) island fraction from human DNA

The texas red signal is derived from the CpG island probe while the fluorescein isothiocyanate (FITC) green signal represents late replicating regions (which are mostly transcriptionally inactive), Black regions represent overlap of signals

Gene distribution

- Recently, insight into gene distribution along the lengths of the different chromosomes has been obtained by hybridizing purified CpG island fractions of the genome (which are associated with perhaps about 56% of human genes).
- On this basis, it is clear that gene density is high in subtelomeric regions and that some chromosomes (e.g. 19 and 22) are gene rich while others (e.g. 4 and 18) are gene poor.

Gene density

- The number of genes in the human genome has been the subject of much speculation; while the small mitochondrial genome is known to have precisely 37 genes, the number in the nuclear genome remains unknown.

- Theoretical calculations based on the *mutational load* that a genome can tolerate and observed average mutation rates of human genes (~10^{-5} per gene per generation) suggest an upper limit of about 100,000.

- A variety of different approaches have been used to obtain more precise estimates of the total gene number.

Three approaches have suggested a best estimate of about 65 000–80 000 genes:

- **Genomic sequencing**.

  Extrapolation from sequencing of large chromosomal regions may suggest that there are about 70,000 genes (Fields *et al.*, 1994). This is based on the observation that gene-rich regions have an average gene density of close to one per 20 kb, but gene-poor regions have a much lower density, say one-tenth of this density, and that the genome is split 50:50 into gene-rich and gene-poor regions.

- **CpG island number**.

  Restriction enzyme analysis using the methylation-sensitive enzyme *Hpa*II suggests that the total number of CpG islands (see Box 8.5) in the human genome is 45 000 (Antequara and Bird, 1993). Using an estimate that approximately 56% of genes are associated with CpG islands, these authors have suggested a total of about 80,000 human genes.

- **EST analysis**.

  Large-scale random sequencing of cDNA clones provides so-called expressed sequence tags (ESTs, see Section 13.2.3). Comparison of known human EST sequences with a large set of different human genomic coding DNA sequences listed in sequence databases has suggested a figure of about 65 000 human genes (Fields *et al.*, 1994).

Gene families

- While the great majority of human genes are expected to encode polypeptides, a significant minority encode mature RNA molecules of diverse function.
- 5% of the genes, perhaps 3000–4000 genes in all, are expected to encode RNA molecules

Ribosomal RNA (rRNA) genes

- There are multiple rRNA genes.

- The 28S, 18S and 5.8S cytoplasmic rRNAs are encoded by a single transcription unit, which is tandemly repeated about 250 times, comprising five clusters of about 50 tandem repeats located on the short arms of human chromosomes 13, 14, 15, 21 and 22.

- In addition, the 5S cytoplasmic rRNA is encoded by several hundred gene copies in at least three clusters on the long arm of chromosome 1.

- The major rationale for the repetition of cytoplasmic rRNA genes is likely to be based on gene dosage: by having a comparatively large number of these genes, the cell can satisfy the huge demand for cytoplasmic ribosomes needed for protein synthesis

Transfer RNA (tRNA) genes

- These belong to a very large dispersed gene family, comprising more than 40 different subfamilies each with several members which encode the different species of cytoplasmic tRNA. In addition to multiple copies of genes specifying the individual cytoplasmic tRNA molecules, there are several defective gene copies (*pseudogenes*).

Small nuclear RNA (snRNA) genes

- A heterogeneous collection of several hundred **small nuclear RNA** species are encoded by a large dispersed family of genes.

- Many of the snRNA species are uridine-rich and are named accordingly, e.g. U3 snRNA means the third uridine-rich small nuclear RNA to be classified.

- Individual species of RNA are associated with specific proteins to form ribonucleoprotein particles (RNPs). Some are known to be important in RNA splicing.

- A large subfamily of perhaps about 200 genes are present in the nucleolus, and have been termed small nucleolar RNA (snoRNA). They have important roles in specific cleavage reactions and base-specific modifications during maturation of ribosomal RNA.

Other RNA genes

- Additional RNA genes encode functionally diverse products, including the 7SL RNA component of the signal recognition particle which is required for protein export and the RNA component of telomerase, the enzyme required to synthesize DNA at the telomeres.

- More recently, evidence has been obtained suggesting that certain RNA genes encode products that are important in gene regulation.

- An important example is the *XIST* gene. This gene is thought to be the major gene involved in initiating the process of X chromosome inactivation, being expressed exclusively from inactivated X chromosomes. No long open reading frames can be identified, and gene function is thought to be carried out through an RNA product by a mechanism that remains obscure.

Eukaryotic Genome

- Vast majority of eukaryotic genome does not encode functional genes
  - many single copy DNA sequences appear to be non-coding = pseudogenes – evolutionary vestiges of duplicated copies, undergone significant mutational alterations
- Only small portion of genome actually codes for proteins

3 Categorie

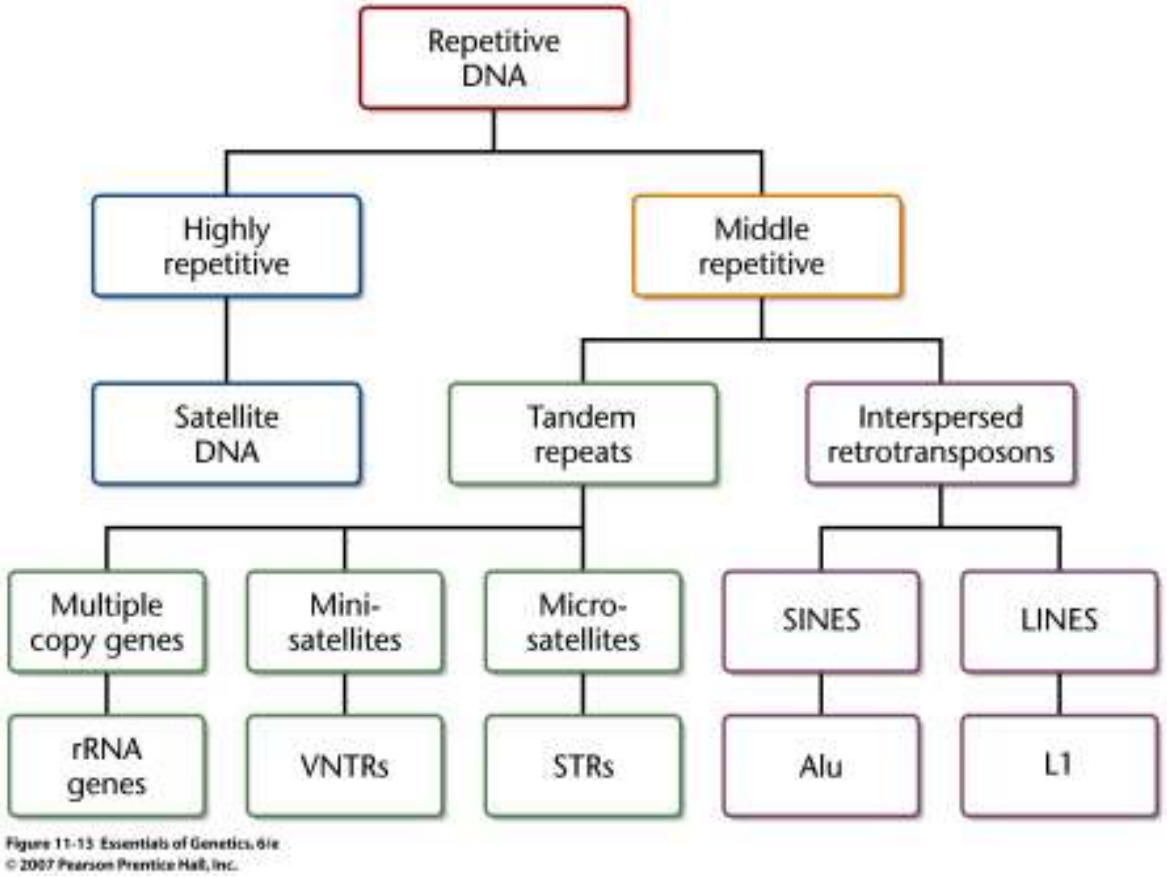


Figure 11-15 Essentials of Genetics, 6/e
© 2007 Pearson Prentice Hall, Inc.

- Heterochromatin associated with centromere and telomeres
- Tandem repeats of both short and long DNA sequence
- Transposable sequences that are interspersed throughout the genome of eukaryotes

Repetitive DNA

- DNA has areas of unique DNA in single copies that make genes - Unique sequences
- Most DNA is actually repetitive in mature and at various levels of repeats
  - various classes and organization within the genome
  - may be functional genes in more than 1 copy
  - much is non-genic

Repetitive DNA

- **Highly repetitive:** About 10-15% of mammalian DNA reassociates very rapidly. This class includes **tandem repeats.**
- **Moderately repetitive:** Roughly 25-40% of mammalian DNA reassociates at an intermediate rate. This class includes **interspersed repeats.**
- **Single copy (or very low copy number):** This class accounts for 50-60% of mammalian DNA.

Satellite DNA

- **Satellites**
- The size of a satellite DNA ranges from 100 kb to over 1 Mb. In humans, a well known example is the **alphoid DNA** located at the centromere of all chromosomes.
- Its repeat unit is 171 bp and the repetitive region accounts for 3-5% of the DNA in each chromosome. Other satellites have a shorter repeat unit.
- Most satellites in humans or in other organisms are located at the centromere.

Telomeres

- Another type of minisatellites is the telomere. In a human germ cell, the size of a telomere is about 15 kb. In an aging somatic cell, the telomere is shorter. The telomere contains tandemly repeated sequence GGGTTA.

Telomeres

- Adds stability to the end of chromosomes
- Keeps the ends inert so they do not interact with other chromosomes and with enzymes that use dsDNA ends as substrates
  - telomere sequences are repetitive in nature
  - use the mechanism of finishing the chromosome to make a loop that protects the ends of chromosomes

Minisatellites

- The size of a minisatellite ranges from 1 kb to 20 kb. One type of minisatellites is called **variable number of tandem repeats (VNTR)**. Its repeat unit ranges from 9 bp to 80 bp. They are located in non-coding regions. The number of repeats for a given minisatellite may differ between individuals. This feature is the basis of **DNA fingerprinting**.

VNTR- individual differences

- **Variable Number of Tandem Repeat (VNTR) Polymorphism**
- VNTR may result from unequal crossover. It is the molecular basis of DNA fingerprinting which has many practical applications

Microsatellites

- Microsatellites are also known as **short tandem repeats** (STR), because a repeat unit consists of only 1 to 6 bp and the whole repetitive region spans less than 150 bp.
- Similar to minisatellites, the number of repeats for a given microsatellite may differ between individuals. Therefore, microsatellites can also be used for DNA fingerprinting

Miniature Inverted-repeat Transposable Elements (MITES)

- almost identical sequences of about 400 base pairs flanked by
- characteristic **inverted repeats** of about 15 base pairs such as

5' GGCCAGTCACAATGG..~400 nt..CCATTGTGACTGGCC 3'
3' CCGGTCAGTGTTACC..~400 nt..GGTAACACTGACCGG 5'

Transposons

- Transposons are segments of DNA that can move around to different positions in the genome of a single cell. In the process, they may
- cause mutations
- increase (or decrease) the amount of DNA in the genome.
- These mobile segments of DNA are sometimes called **"jumping genes"**.

Repetitive Transposed Sequences

- 2 types – SINEs and LINEs
- Not tandem repeats – dispersed thru the genome, long or short
- Transposable elements – mobile and can potentially move to different location within genome
- Large portion of human genome

SINEs

- Short interspersed elements – 100-500 bp, present 1.5 million times in humans
  - alu family (named after the enzyme that cuts the DNA, AluI)
    - 200-300 bp – dispersed throughout genome, between and within genes, may transcribe into RNA of unknown function – may help to move around genome
- ~13% of genome

LINEs

- Long interspersed elements – transposable elements
  - ~6 kb in length, present 850,000 times
  - L1 family – 6400 bp and 100,000 repeats
    - transcribed into an RNA, serves as template to make DNA complement by reverse transcriptase, encoded by part of L1 gene
    - new copy inserts (integrates) into DNA at new site
    - similar to retroviruses so also called retrotransposons
- ~21% of genome

Alus

- Alu elements consist of a sequence of 300 base pairs containing a site that is recognized by the restriction enzyme AluI. They appear to be reverse transcripts of 7S rRNA, part of the signal recognition particle.
- Most SINEs do not encode any functional molecules and depend on the machinery of active L1 elements to be transposed; that is, copied and pasted in new locations.

PALINDROME:

sequence that reads the same way in both directions
ie. complementary sequence is identical to the other side

GGGCCC **restriction site**
CCCGGG **restriction site**

most restriction enzymes are protein homodimers, forcing them to cut palindromic sequences