Molecular Biology (3)

The human genome

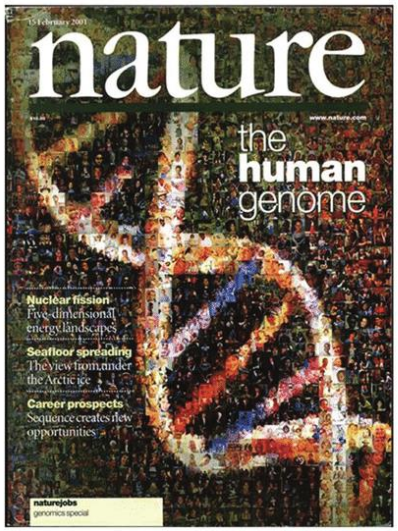
Mamoun Ahram, PhD

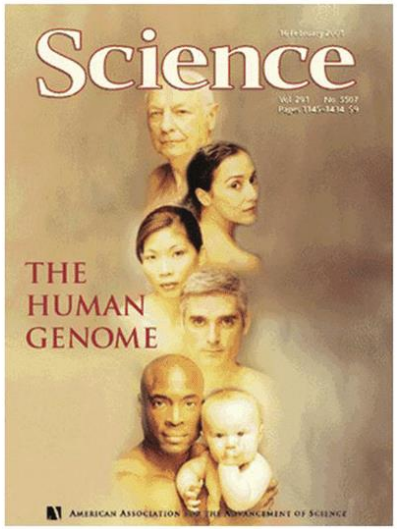
Resources

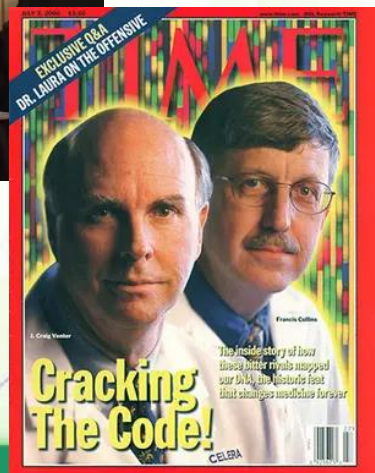
- This lecture
- Cooper, Ch. 6, pp. 157-160, 195-205, 209-212

The human genome project

- A $3 billion, 13-year, multi-national project launched in 1990 led by the US government to (know the) sequence the human genome and to map and identify the genes (a draft was published in 2001 and 92% was completed in 2004).







Major outcomes

- Determination of the number of human genes
- Development of major technologies
- Completed sequences of other genomes
- Open discussion of legal and ethical issues

SPECIES	BASE PAIRS (estimated)	GENES (estimated)	CHROMOSOMES
Human *(Homo sapiens)*	3.2 billion	X ~ 25,000	46
Mouse *(Mus musculus)*	2.6 billion	X ~ 25,000	40
Fruit Fly *(Drosophilia melanogaster)*	137 million	13,000	8
Roundworm *(Caenorhabditis elegans)*	97 million	19,000	12
Yeast *(Saccharomyces cerevisia)*	12.1 million	6,000	32
Bacteria (Escherichia coli)	4.6 million	3,200	1
Bacteria (H. influenzae)	1.8 million	1,700	1

Nucleotides per genomes

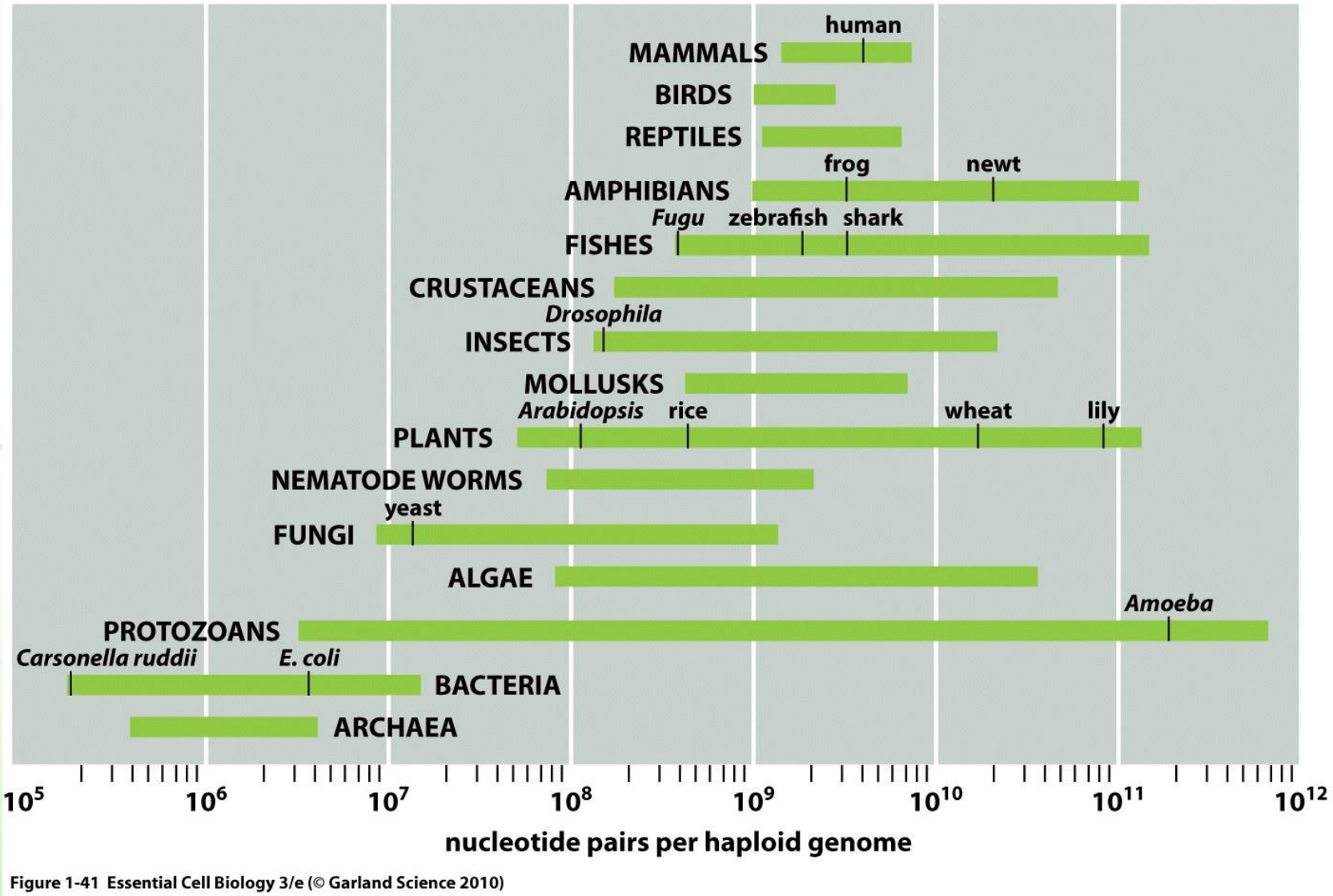
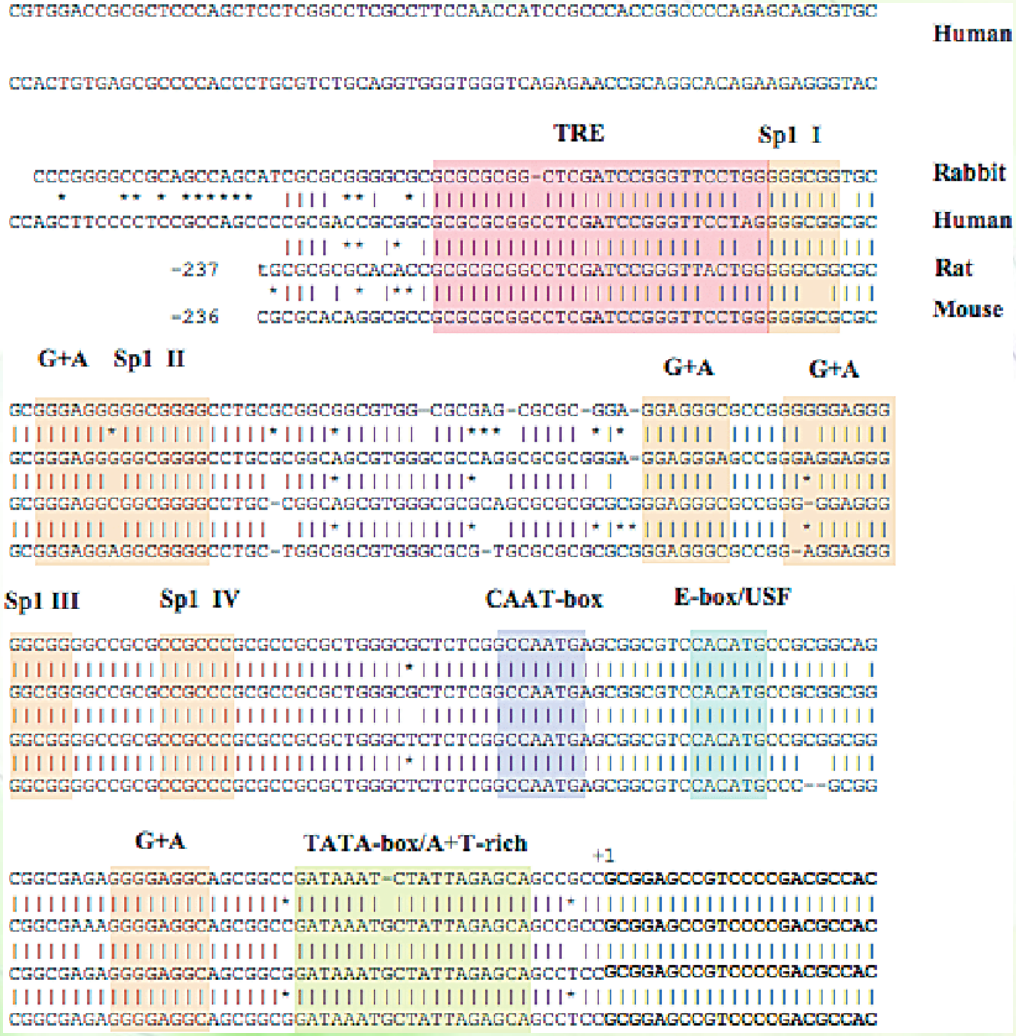
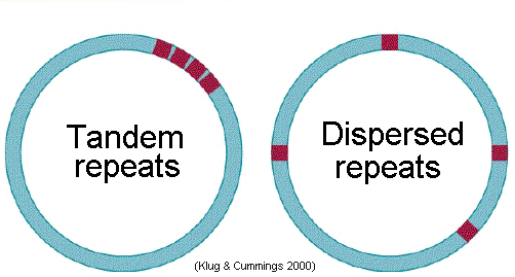


Figure 1-41 Essential Cell Biology 3/e (© Garland Science 2010)

DNA homology
(i.e. sequence similarity)



Components of the human genome



- Human genome
  - Nuclear genome 3300 Mb
    - Genes and gene-related sequences (30%)
      - Noncoding sequences (28%)
        - Introns (26%)
        - Pseudogenes
        - Other sequences (e.g. untranslated regions)
      - Protein-coding sequences (2%)
    - Extragenic (intergenic) sequences (70%)
      - Unique or low copy number (5%)
      - Moderate to highly repetitive (55%)
        - TANDEM repetitive sequences (10%)
          - Satellites (centromeres and telomeres)
          - Minisatellites (VNTRs)
          - Microsatellites (STRs)
        - *Interspersed*
          - DNA transposons (3%)
          - Retrotransposons (42%)
            - SINEs (13.%)
            - LINEs (21%)
            - Retrovirus-like elements (8%)
      - Other (10%)
  - Mitochondrial genome 16.6 Kb

~5% of the genome contains sequences of noncoding DNA that are highly conserved indicating that they are critical to survival.

Note: All numbers are approximate

The ENCODE project (2003-on)

- ENCODE: Encyclopedia of DNA Elements (ENCODE)
- 80% of the entire human genome is relevant (either transcribed, binds to regulatory proteins, or is associated with some other biochemical activity).

Summary of ENCODE Results	
Protein-coding genes	20,687
Short noncoding RNAs	8801
Long noncoding RNAs	
Pseudogenes	11,224
Percentage of genome transcribed into RNA	74.7%
Percentage of genome-binding transcription factors	8.1%

On March 31, 2022…

A gene: a region of DNA that is transcribed.

A transcript: a RNA molecule that is produced by transcription

	Gene annotation
Number of genes	63,494
Protein coding	19,969
Number of exclusive genes	3,604
Protein coding	140
Number of transcripts	233,615
Protein coding	86,245
Number of exclusive transcripts	6,693
Protein coding	2,780

RESEARCH ARTICLE

HUMAN GENOMICS

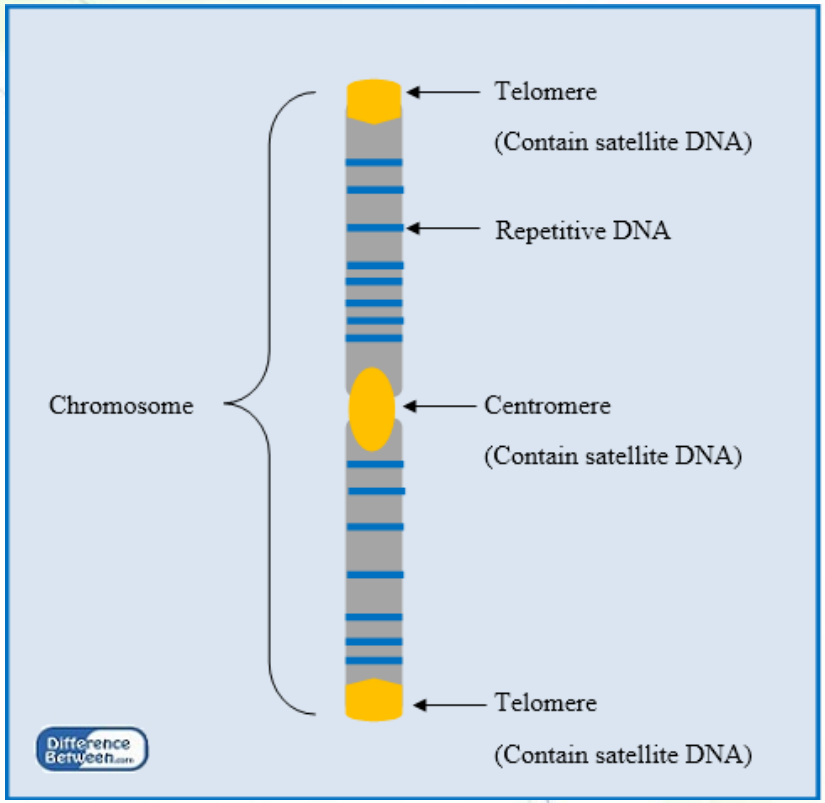
The complete sequence of a human genome

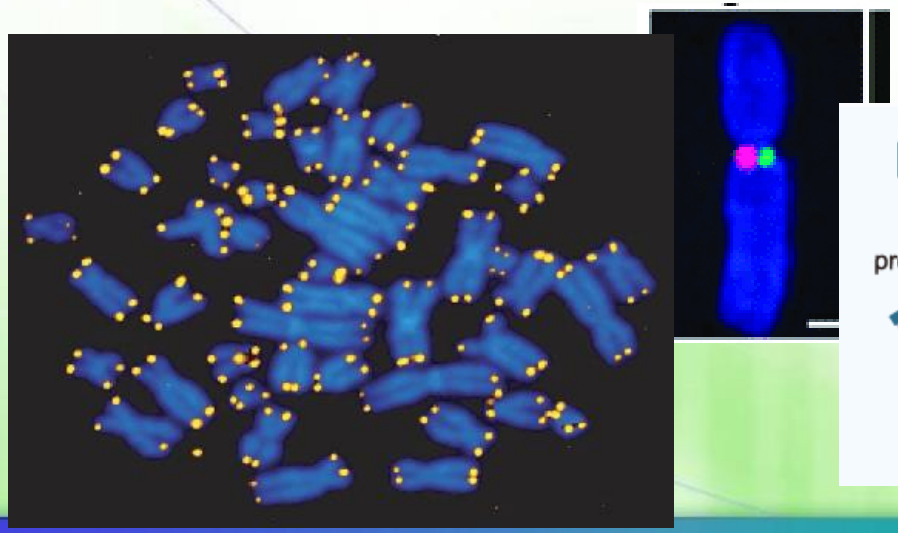
Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

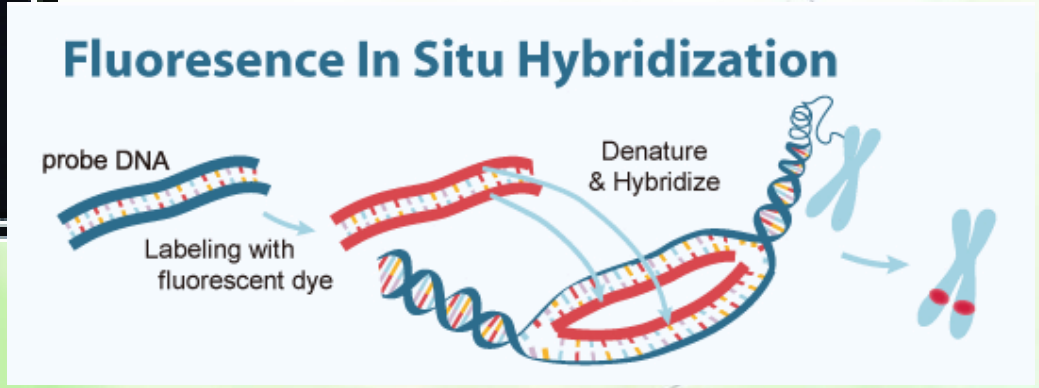
Tandem repeats

Satellite (macro-satellite) DNA

- Regions of 5-300 bp repeated 10^6-10^7 times
  - Centromeres and telomeres
- Centromeric A/T-rich repeats (171 bp) called α-satellite unique to each chromosome (you make chromosome-specific probes) by **fluorescence in situ hybridization (FISH)**.

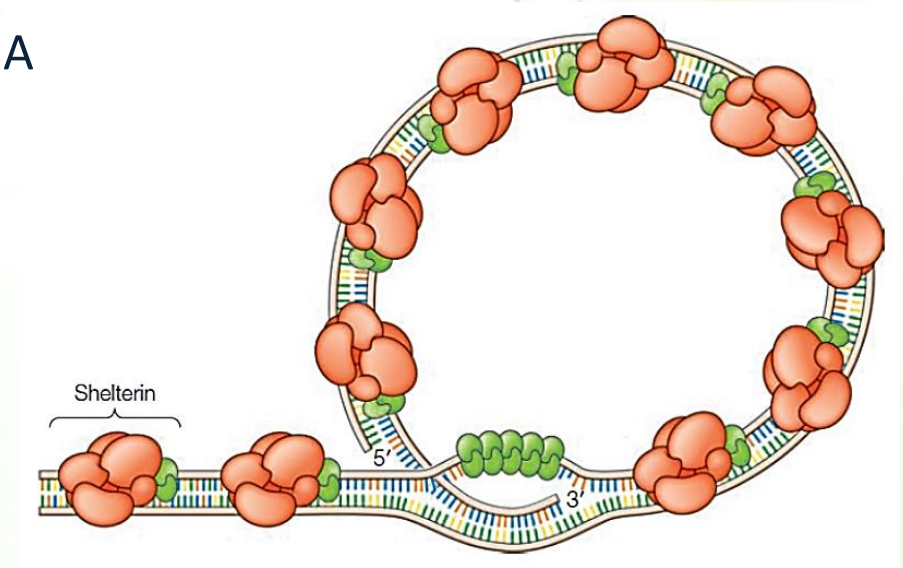






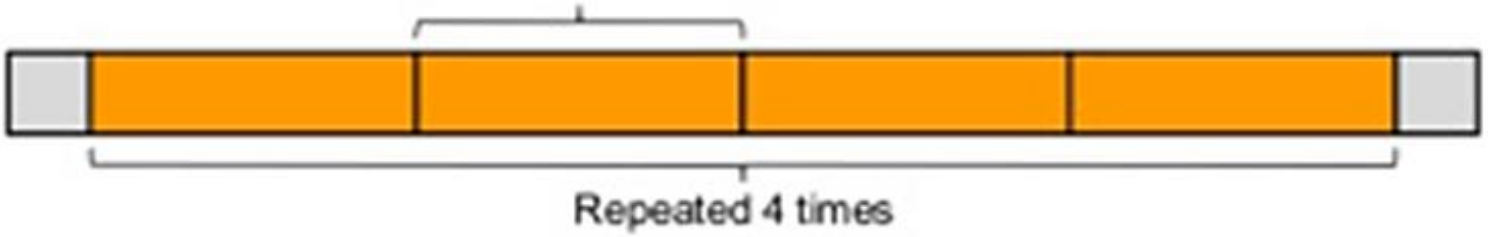
Telomeric repeats

- (TTAGGG) is repeated hundreds to thousands of times at the termini of human chromosomes with a 3′ overhang of single-stranded DNA.
- The repeated sequences form loops that bind a protein complex called **shelterin**, which protects the chromosome termini from degradation.
- **Telomeric repeat-containing RNA (TERRA):** a long non-coding RNA transcribed from telomeres and functions in:
  - maintaining the integrity of chromosome termini,
  - regulating telomerase activity,
  - maintaining the heterochromatic state of telomeres,
  - protecting DNA from deterioration or fusion with neighboring chromosomes



Mini- and Micro-satellite DNA

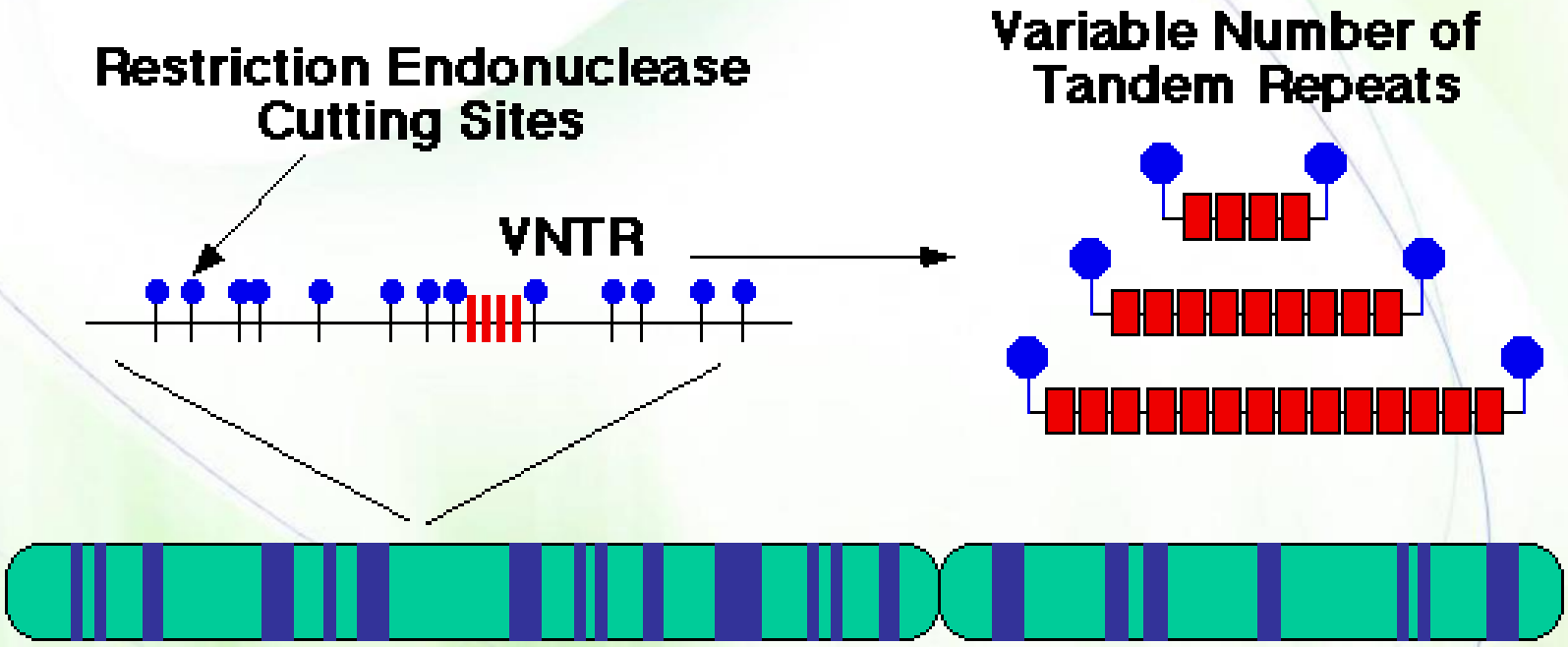
Minisatellite: Variable Number Tandem Repeats (VNTR)



Microsatellite: Short Tandem Repeats (STR) – Simple Sequence Repeats (SSR)

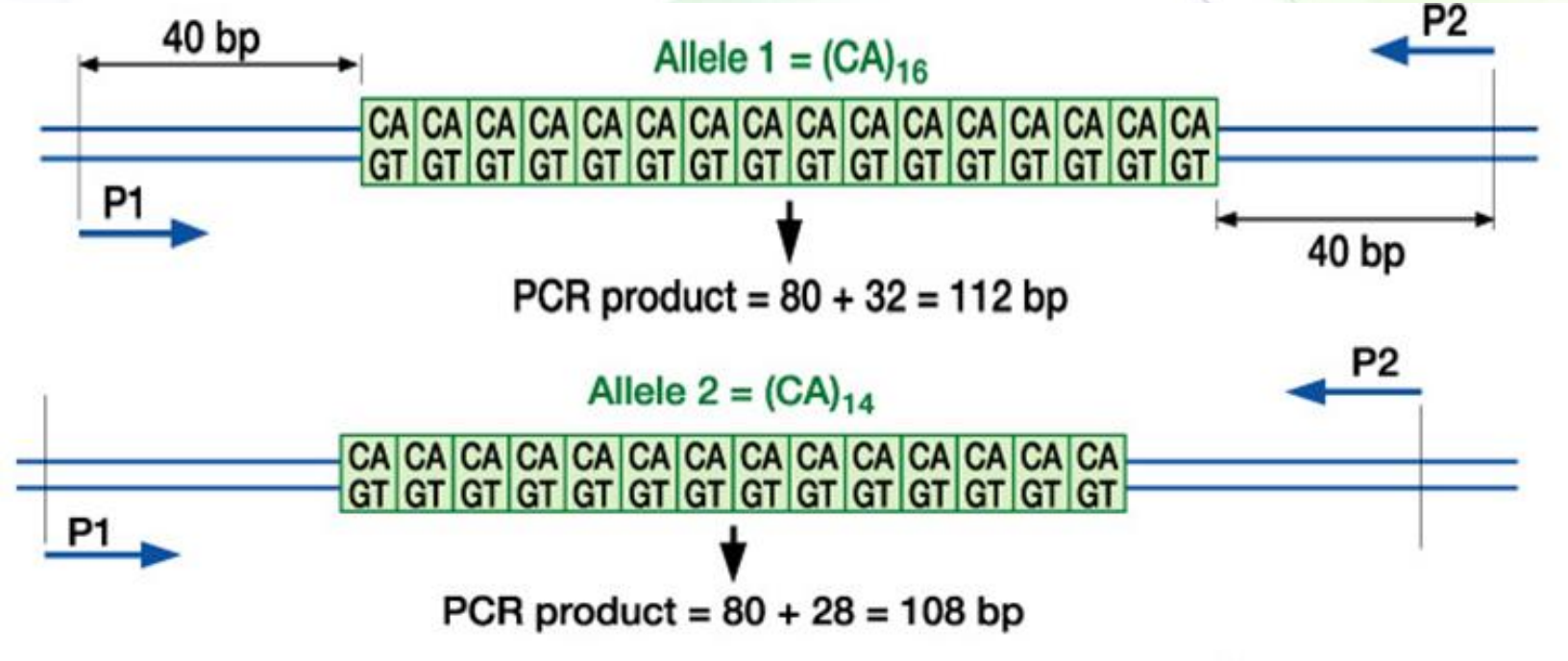
Mini-satellite DNA

- Mini satellite sequences or VNTRs (variable number of tandem repeats) of 20 to 100 bp repeated 20-50 times



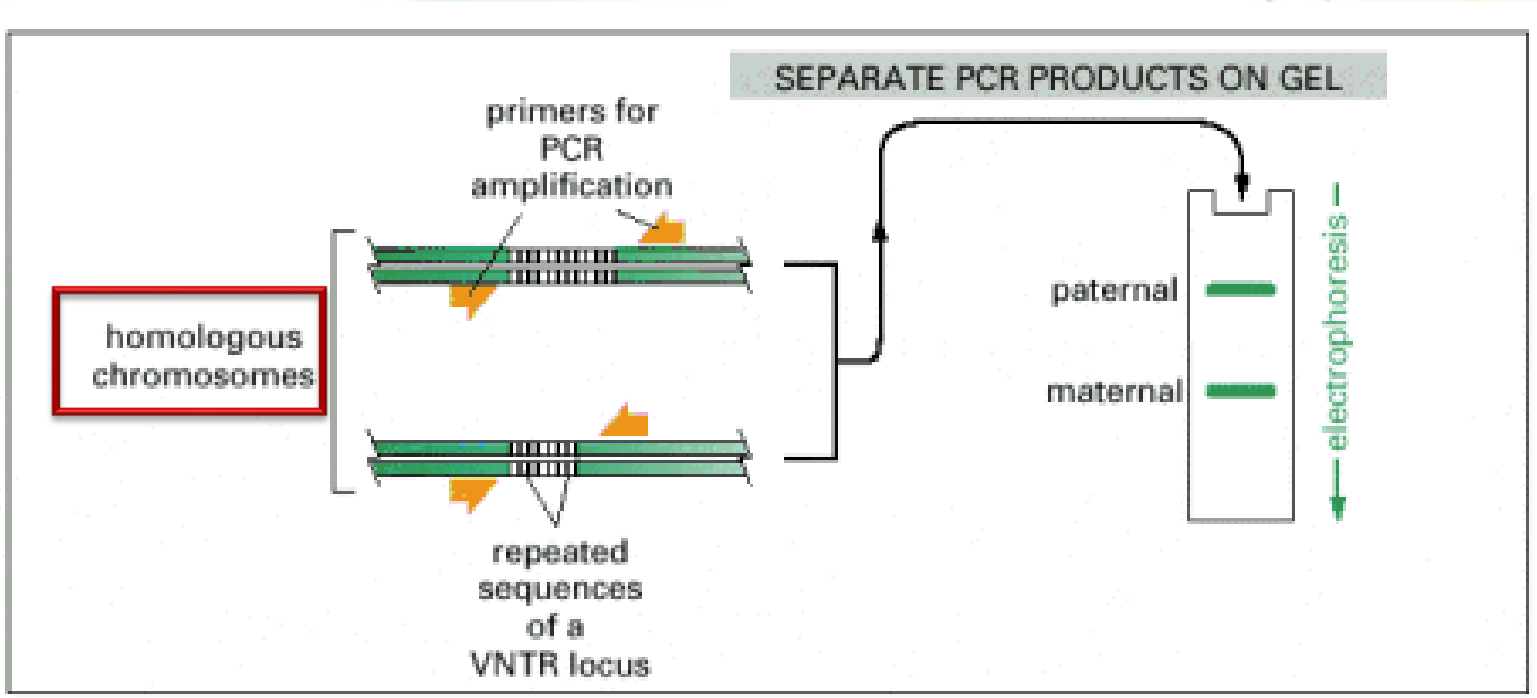
Micro-satellite DNA

- STRs (short tandem repeats) of 2 to 10 bp repeated 10-100 times

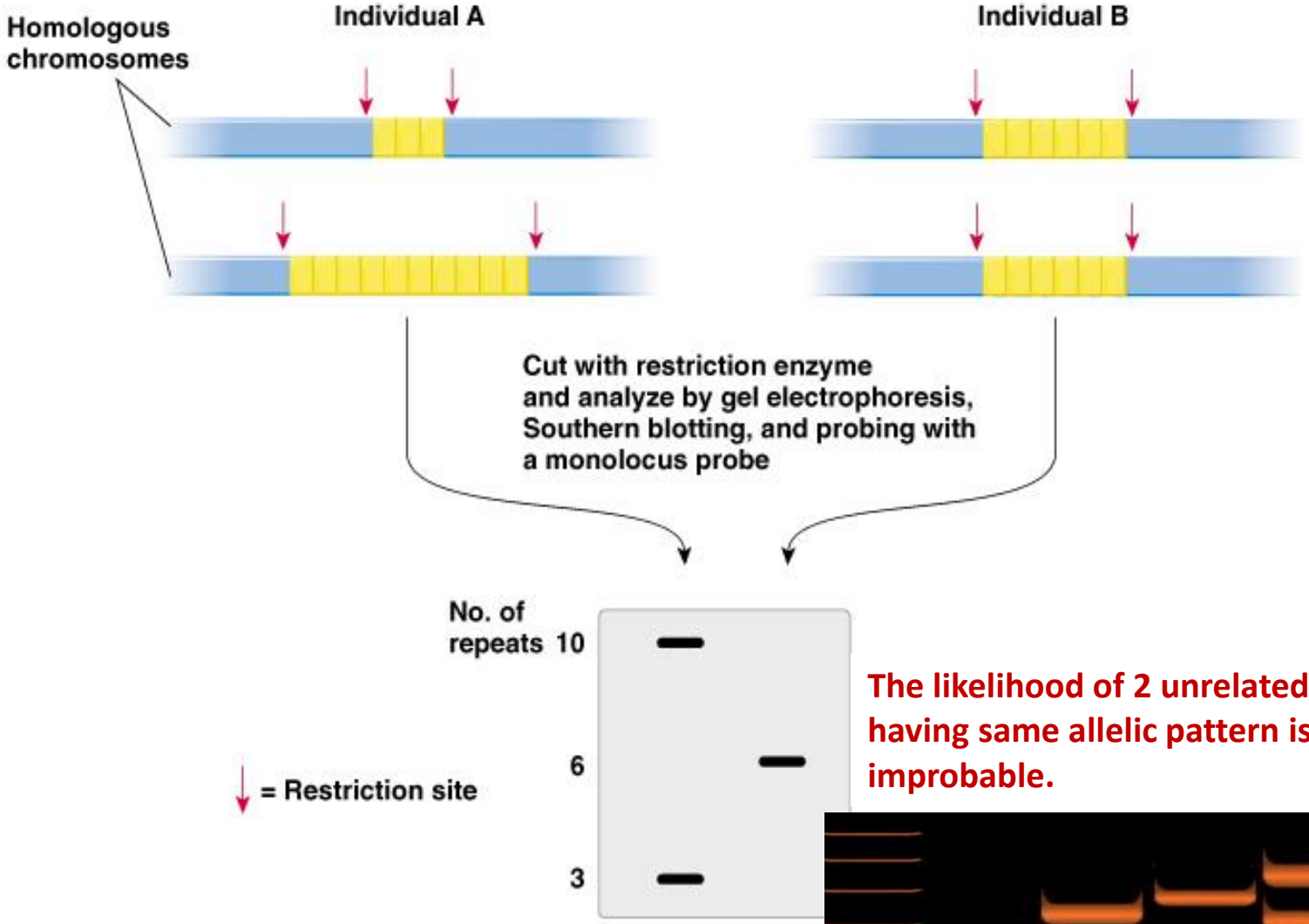


Polymorphisms of VNTR and STR

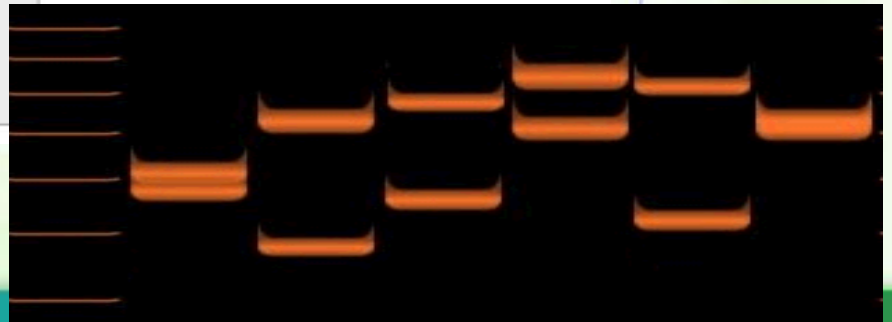
- STRs and VNTRs are highly variable among individuals (polymorphic).
  - They are useful in DNA profiling for forensic testing.



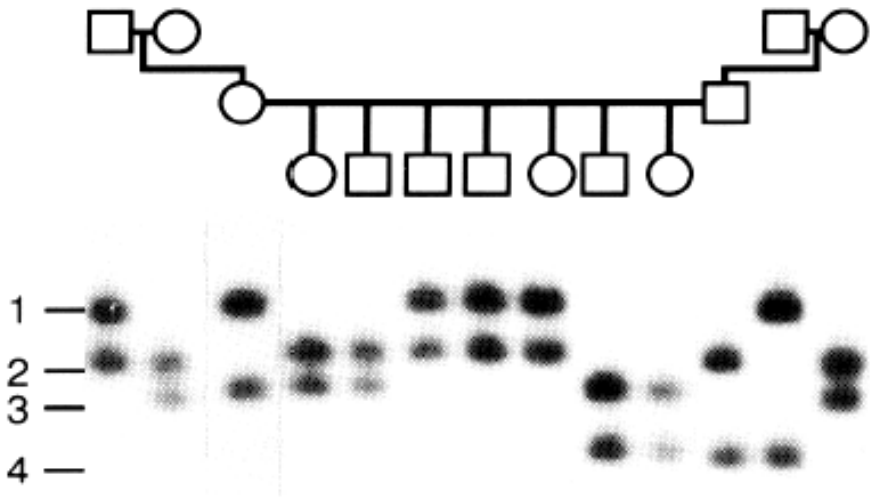
STRs and VNTRs as DNA Markers



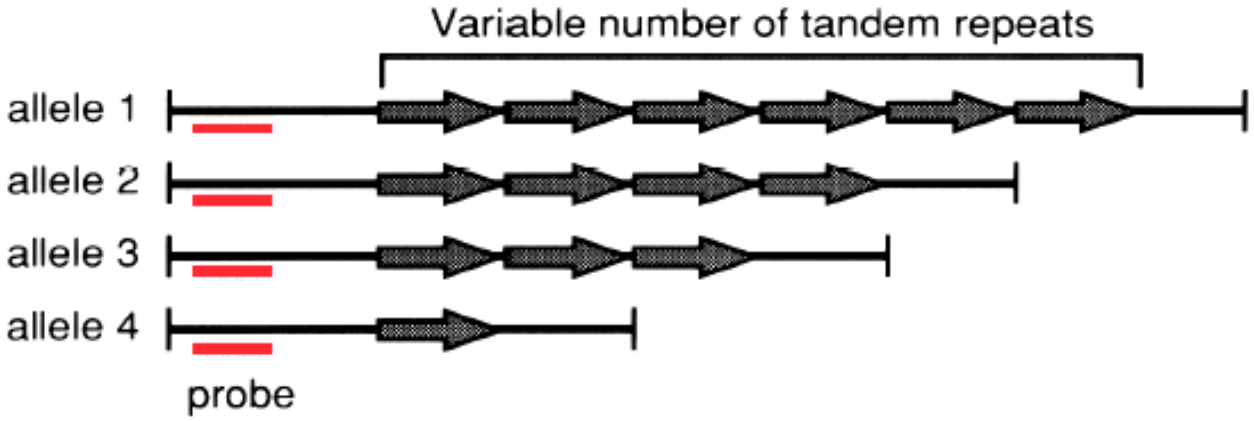
The likelihood of 2 unrelated individuals having same allelic pattern is extremely improbable.



Real example

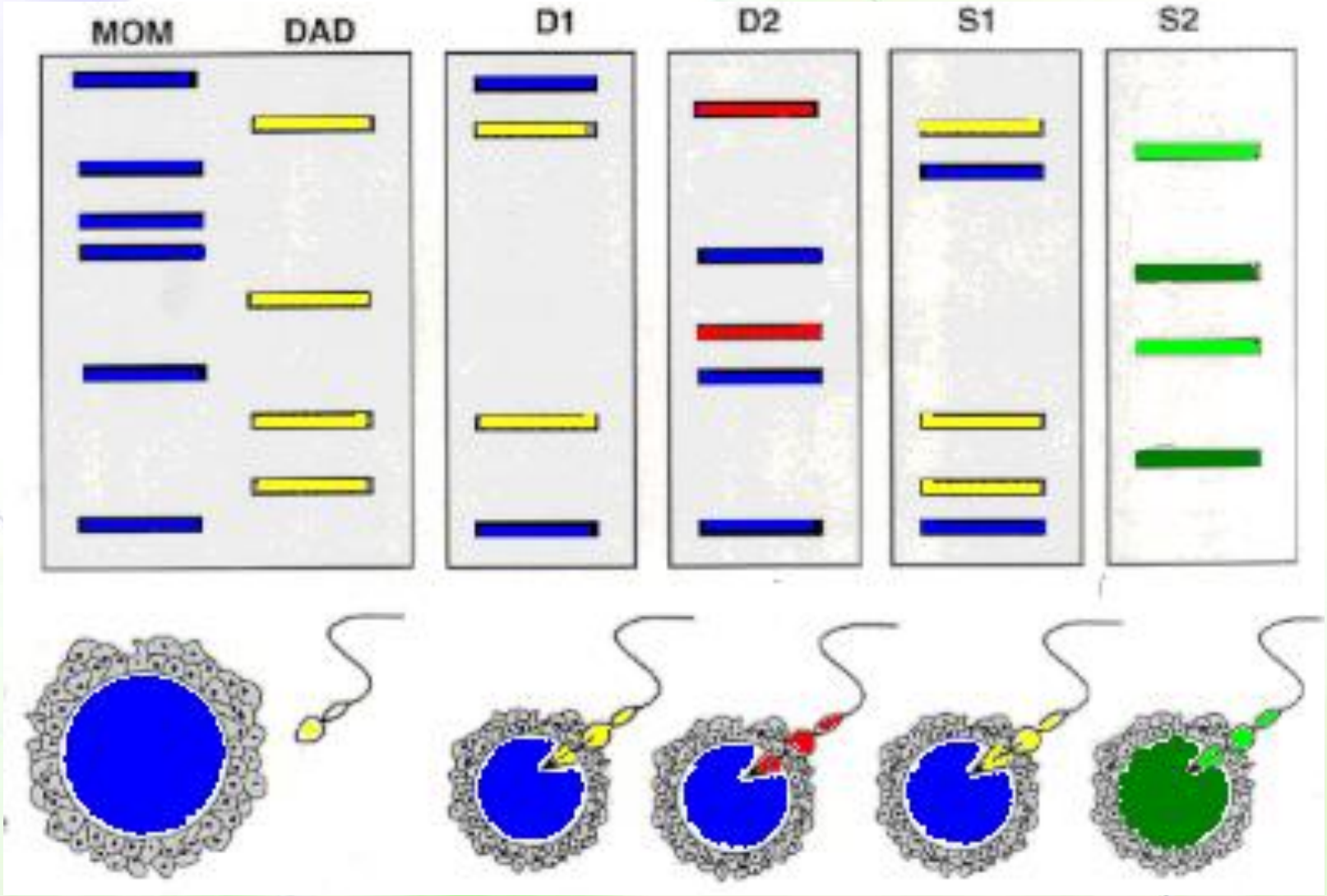


single-locus probe but multiple alleles



Thompson & Thompson Genetics in Medicine, p. 130, 1991

Paternity testing

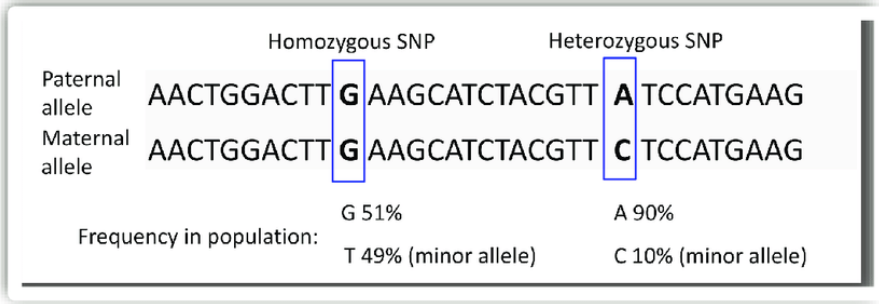


Single nucleotide polymorphism (SNPs)

Another source of polymorphism

- Another source of genetic variation
- Single-nucleotide substitutions of one base for another
- Two or more versions of a sequence must each be present in at least one percent of the general population
- SNPs occur throughout the human genome - about one in every 300 nucleotide base pairs.
  - ~10 million SNPs within the 3-billion-nucleotide human genome
  - Only 500,000 SNPs are thought to be relevant

Examples



Individual 1

Chr 2 ...CGATATTCCTATCGAATGTC...
copy1 ...GCTATAAGGATAGCTTACAG...

Chr 2 ...CGATATTCCCATCGAATGTC...
copy2 ...GCTATAAGGGTAGCTTACAG...

Individual 2

Chr 2 ...CGATATTCCCATCGAATGTC...
copy1 ...GCTATAAGGGTAGCTTACAG...

Chr 2 ...CGATATTCCCATCGAATGTC...
copy2 ...GCTATAAGGGTAGCTTACAG...

Individual 3

Chr 2 ...CGATATTCCTATCGAATGTC...
copy1 ...GCTATAAGGATAGCTTACAG...

Chr 2 ...CGATATTCCTATCGAATGTC...
copy2 ...GCTATAAGGATAGCTTACAG...

Individual 4

Chr 2 ...CGATATTCCTATCGAATGTC...
copy1 ...GCTATAAGGATAGCTTACAG...

Chr 2 ...CGATATTCCCATCGAATGTC...
copy2 ...GCTATAAGGGTAGCTTACAG...

Individual 5

Chr 2 ...CGATATTCCCATCGAATGTC...
copy1 ...GCTATAAGGGTAGCTTACAG...

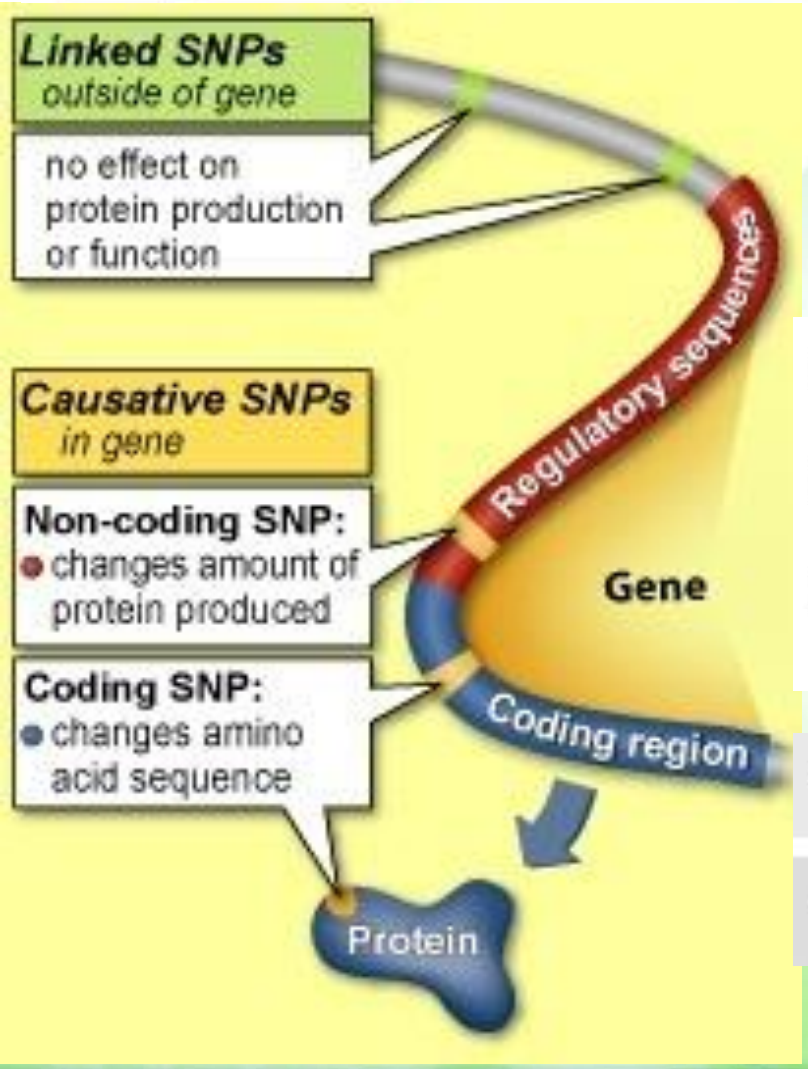
Chr 2 ...CGATATTCCTATCGAATGTC...
copy2 ...GCTATAAGGATAGCTTACAG...

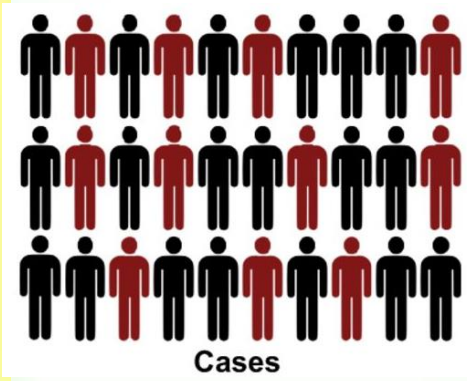
Individual 6

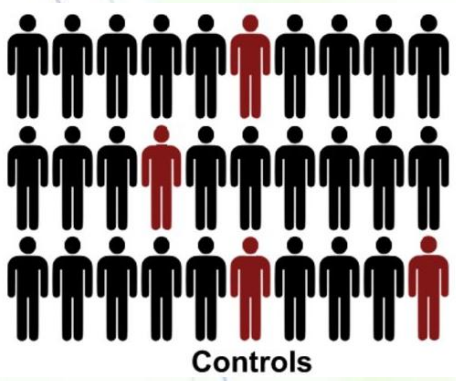
Chr 2 ...CGATATTCCCATCGAATGTC...
copy1 ...GCTATAAGGGTAGCTTACAG...

Chr 2 ...CGATATTCCTATCGAATGTC...
copy2 ...GCTATAAGGATAGCTTACAG...

Categories of SNPs







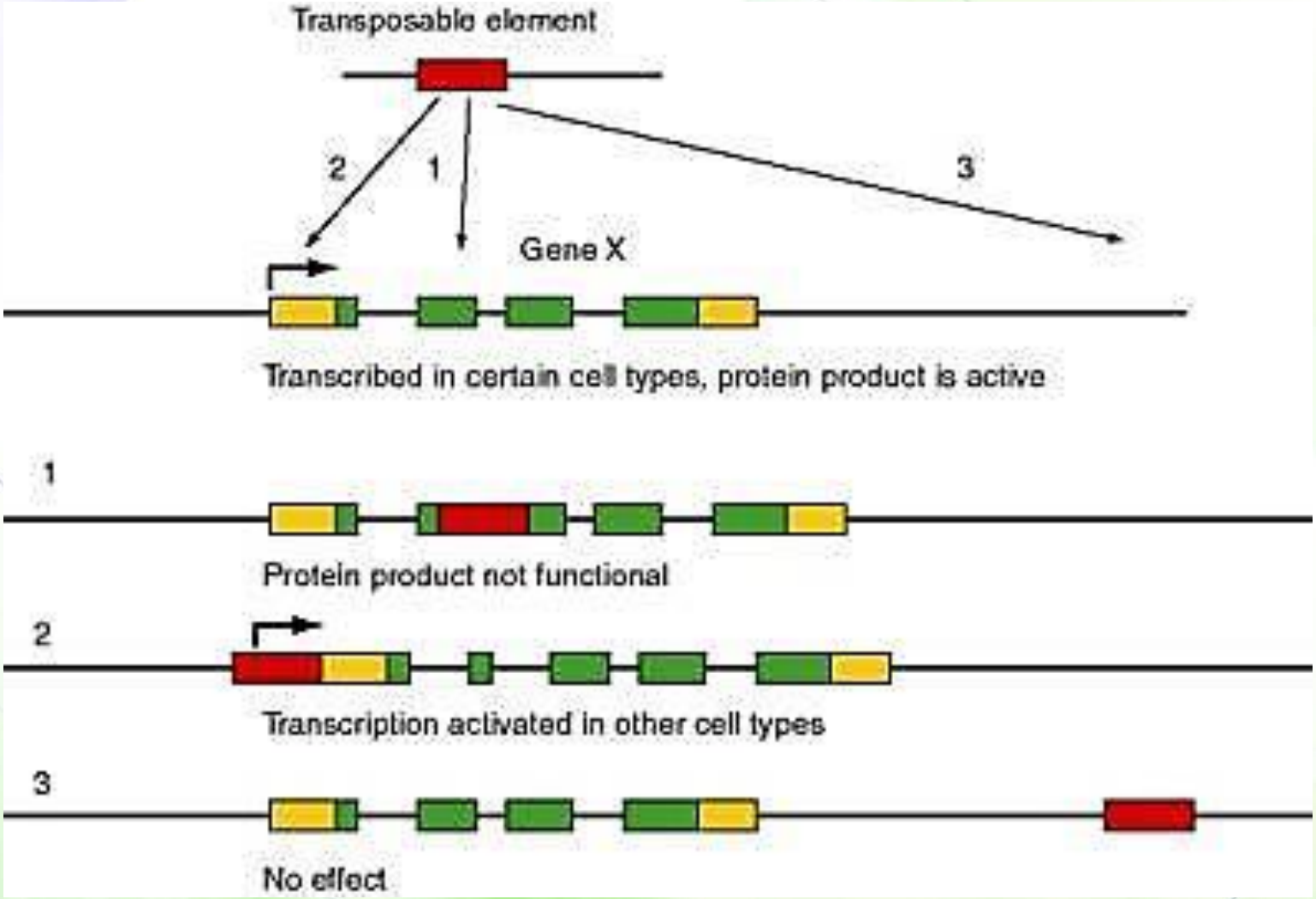
TTGGCCAGCTGGA**C**GAGGGCGATGAC

TTGGCCAGCTGGA**T**GAGGGCGATGAC

Interspersed repeats

Transposons (jumping genes)

- They are segments of DNA that can move from their original position in the genome to a new location.
- Two classes:
  - DNA transposons (3% of human genome)
  - RNA transposons or retrotransposons (42% of human genome).
    - Long interspersed elements (LINEs, 21%)
    - Short interspersed elements (SINEs, 13%)
      - An example is Alu (300 bp)
    - Retrovirus-like elements (8%)
- Over 99% of the transposons in the human genome lost their ability to move, but we still have some active transposable elements that can sometimes cause disease.
  - Hemophilia A and B, severe combined immunodeficiency, porphyria, predisposition to cancer, and Duchenne muscular dystrophy.


Transposable element
2
1
3
Gene X
Transcribed in certain cell types, protein product is active
1
Protein product not functional
2
Transcription activated in other cell types
3
No effect