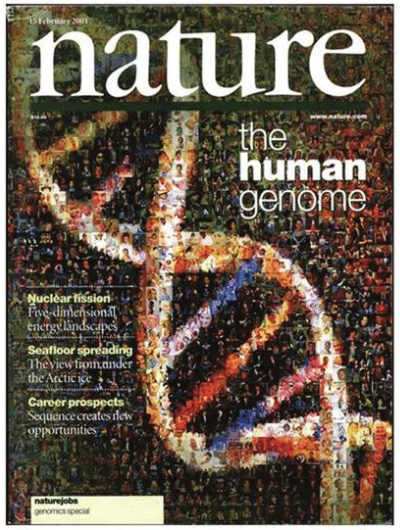
Molecular Biology (2)

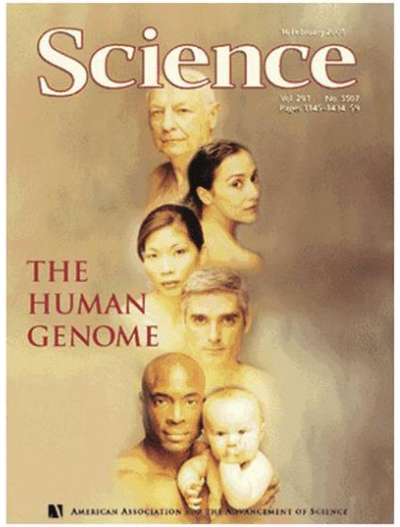
The human genome

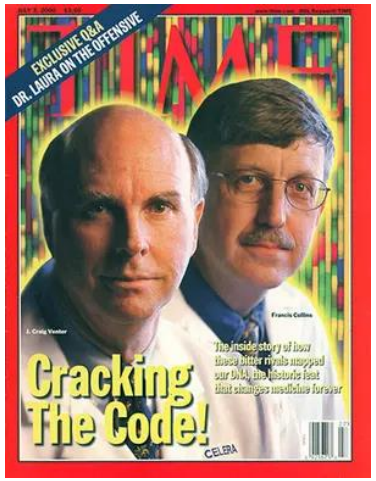
Prof. Mamoun Ahram
School of Medicine
Second year, First semester, 2025-2026

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 and bioinformatic tools
- Completed sequences of other genomes
- Open discussion of legal and ethical issues

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
viruses			
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
organelles			
mitochondria - *H. sapiens*	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
chloroplast – *A. thaliana*	150 kbp	100	1
eukaryotes - multicellular			
dog *C. familiaris*	2.4 Gbp	19,000	40
chimpanzee *P. troglodytes*	3.3 Gbp	19,000	48 (2n)

The ENCODE project (2003-on)

- ENCODE: Encyclopedia of DNA Elements (ENCODE)
- ~75% 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	9640
Pseudogenes	11,224
Percentage of genome transcribed into RNA	74.7%
Percentage of genome-binding transcription factors	8.1%

On March 31, 2022…

RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

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

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.

On August 23, 2023. It is finally done.

nature

Article | Published: 23 August 2023

The complete sequence of a human Y chromosome

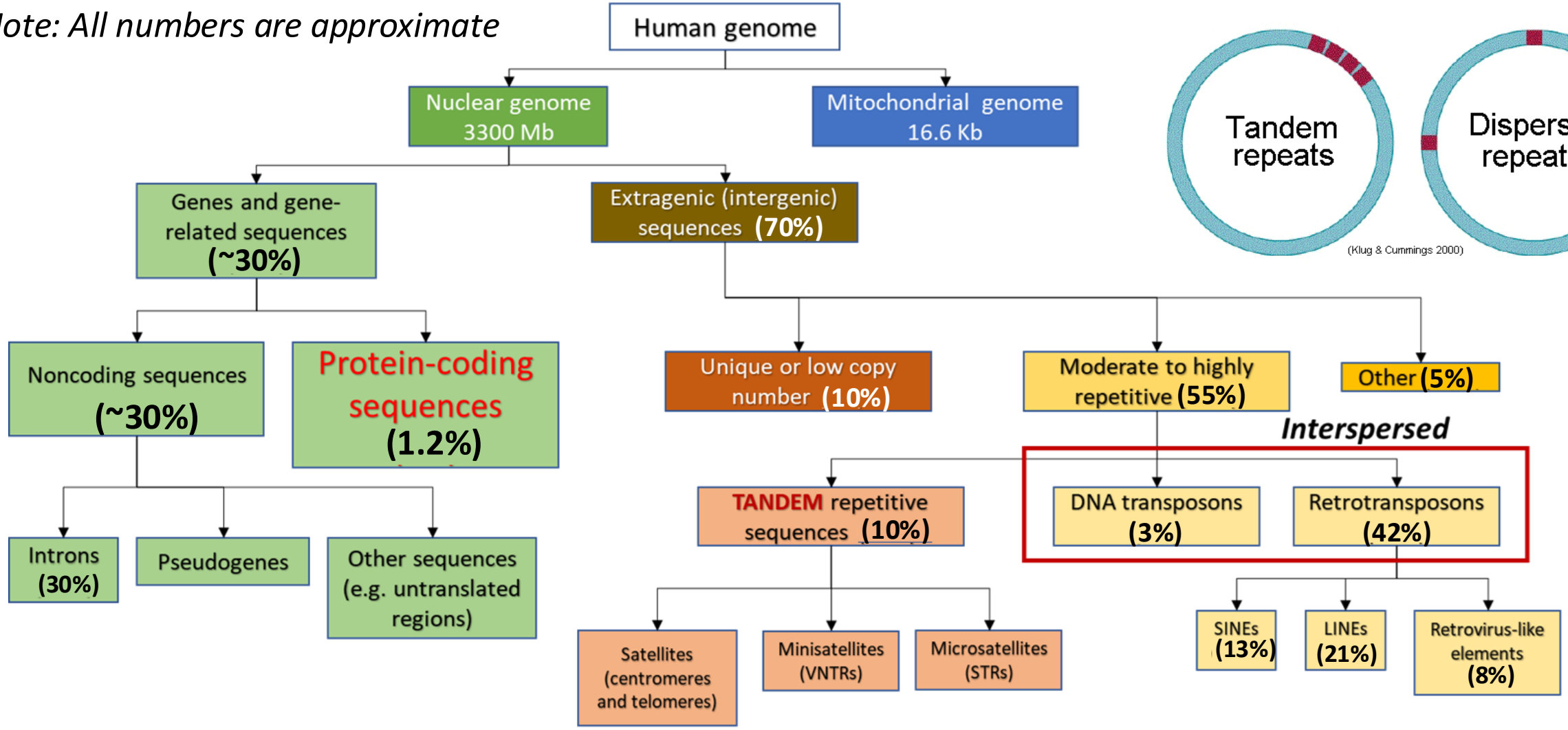
Arang Rhie, Sergey Nurk, Monika Cechova, Savannah J. Hoyt, Dylan J. Taylor, Nicolas Altemose, Paul W.

Abstract

The human Y chromosome has been notoriously difficult to sequence and assemble because of its complex repeat structure that includes long palindromes, tandem repeats and segmental duplications[1,2,3]. As a result, more than half of the Y chromosome is missing from the GRCh38 reference sequence and it remains the last human chromosome to be finished[4,5]. Here, the Telomere-to-Telomere (T2T) consortium presents the complete 62,460,029-base-pair sequence of a human Y chromosome from the HG002 genome (T2T-Y) that corrects multiple errors in GRCh38-Y and adds over 30 million base pairs of sequence to the reference, showing the complete ampliconic structures of gene families *TSPY*, *DAZ* and *RBMY*; 41 additional protein-coding genes, mostly from the *TSPY* family; and an alternating pattern of human satellite 1 and 3 blocks in the heterochromatic Yq12 region. We have combined T2T-Y with a previous assembly of the CHM13 genome[4] and mapped available population variation, clinical variants and functional genomics data to produce a complete and comprehensive reference sequence for all 24 human chromosomes.

Components of the human genome

Note: All numbers are approximate

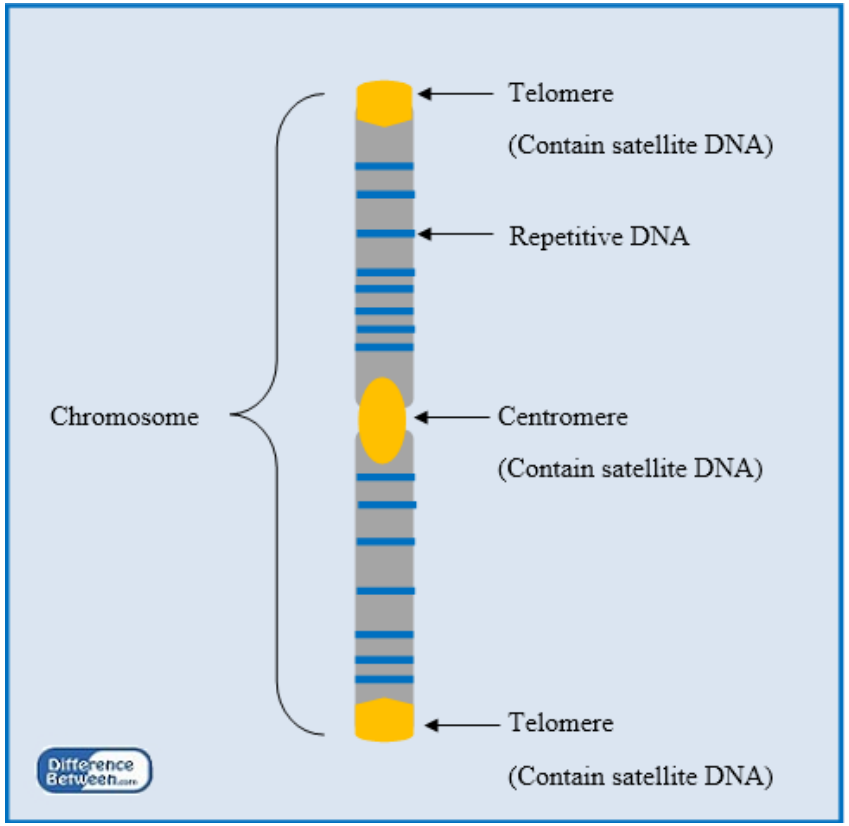


~5% of the genome contains sequences of noncoding DNA that are highly conserved (critical to survival).

Tandem repeats

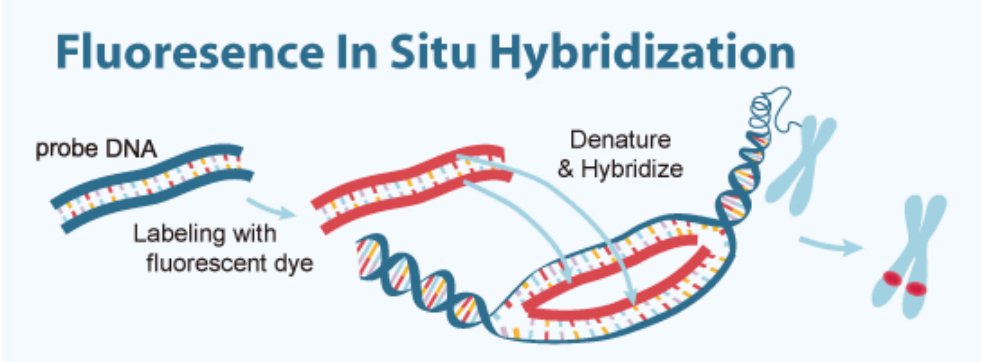
Satellite (macro-satellite) DNA

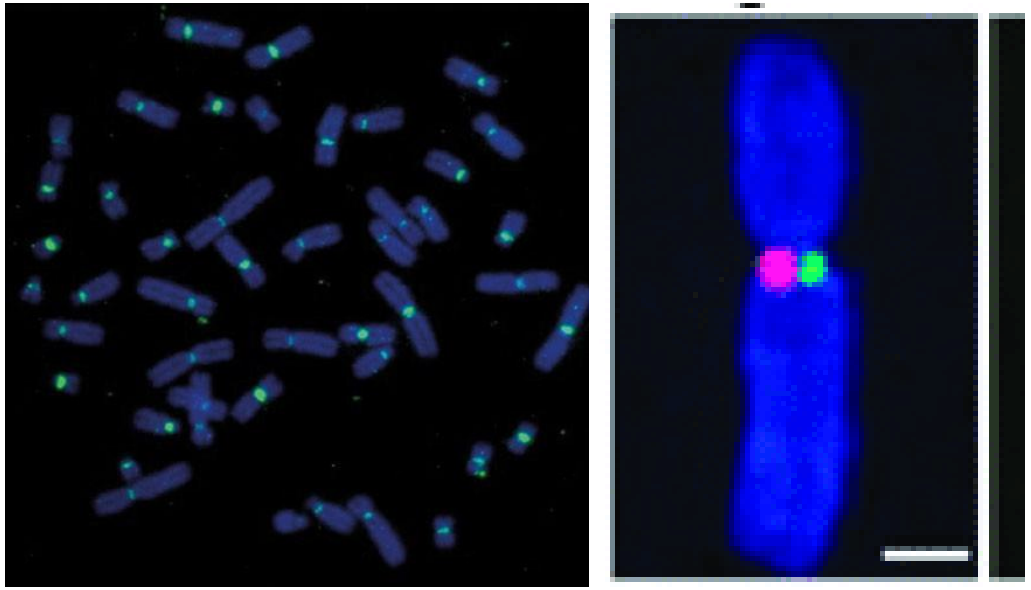
- Regions of 5-300 bp repeated 10^6-10^7 times
  - Centromeres and telomeres
- Centromeric A/T-rich repeats (171 bp) called α-satellite unique to each chromosome (you can make chromosome-specific probes) by fluorescence in situ hybridization (FISH).



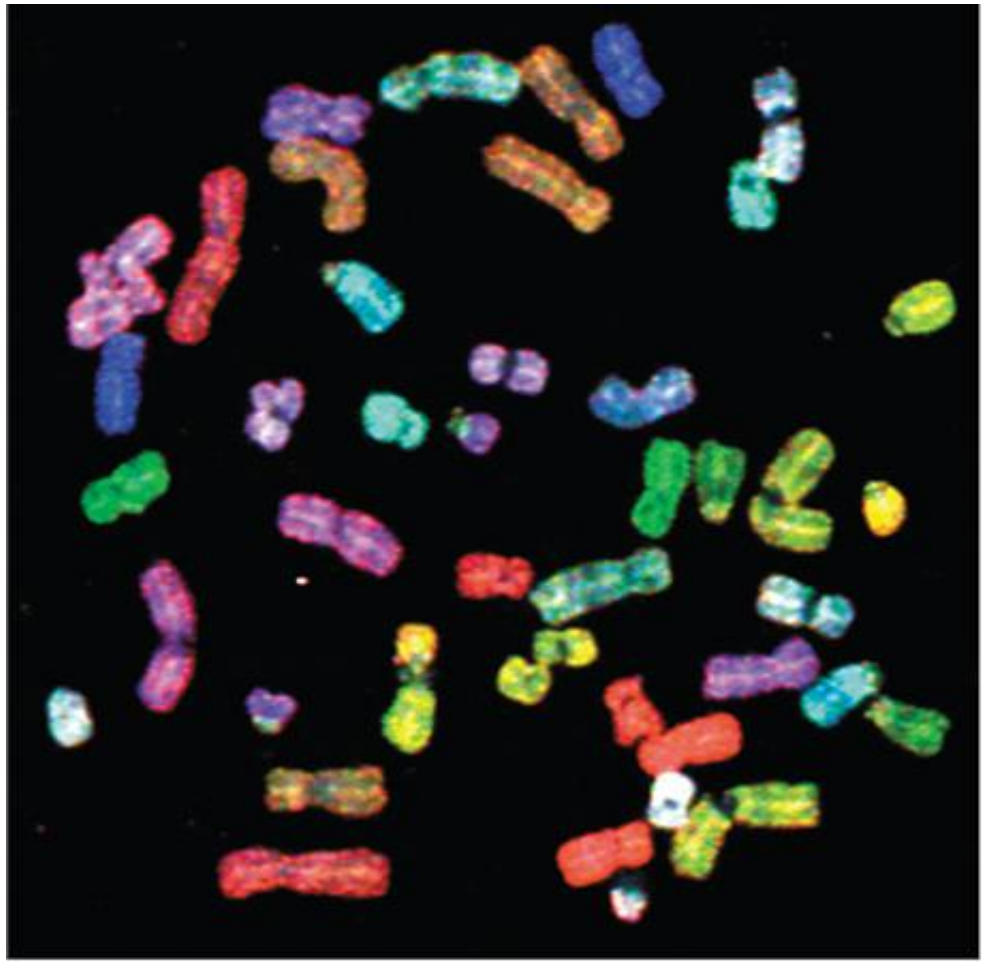
Fluorescence in situ hybridization (FISH)

- FISH is a technique that uses fluorescently labeled oligonucleotide probes that are complementary to specific repeated sequences on individual chromosomes to visualize the location of chromosomes.
- Uses:
  - Locate a gene on a chromosome
  - Determine chromosomal and genetic anomalies like:
    - Duplication, deletion, translocation, and amplification



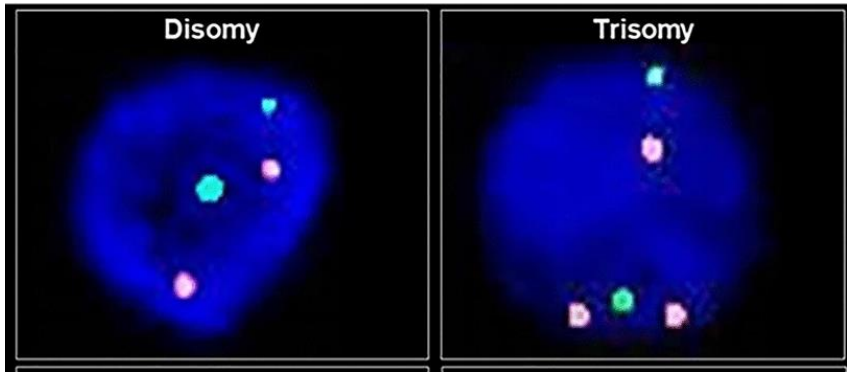


Fluorescence in situ hybridization (FISH)



Courtesy of Thomas Ried and Hesed Padilla-Nash, National Cancer Institute

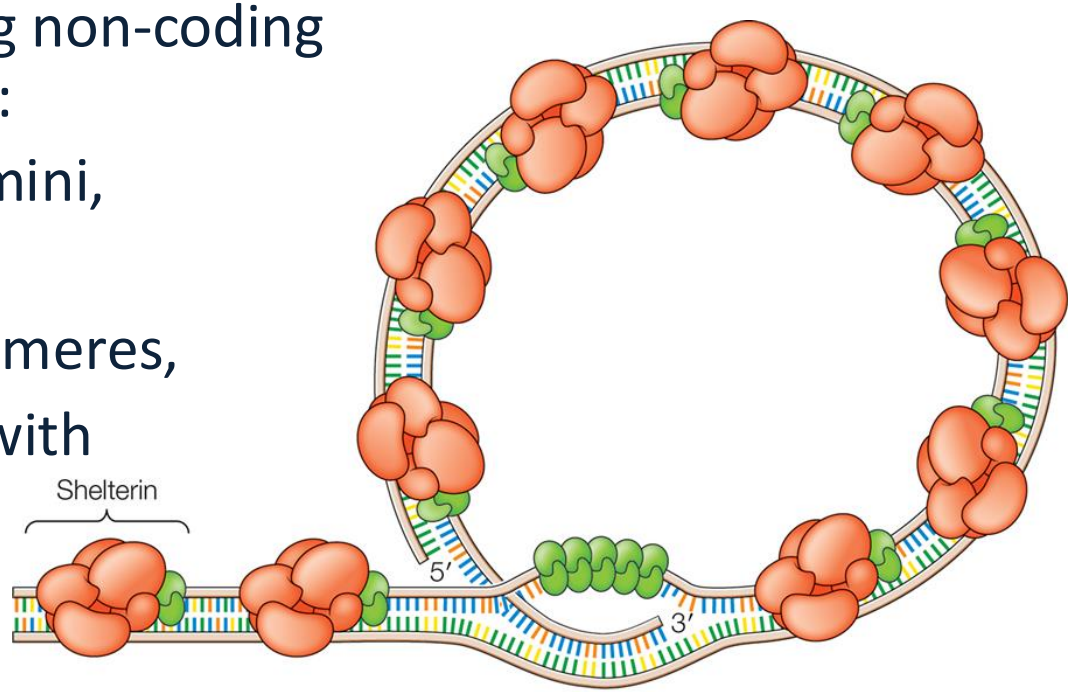
- Hybridization of human chromosomes with chromosome-specific fluorescent probes that label each of the chromosomes a different color.



Telomeric repeats

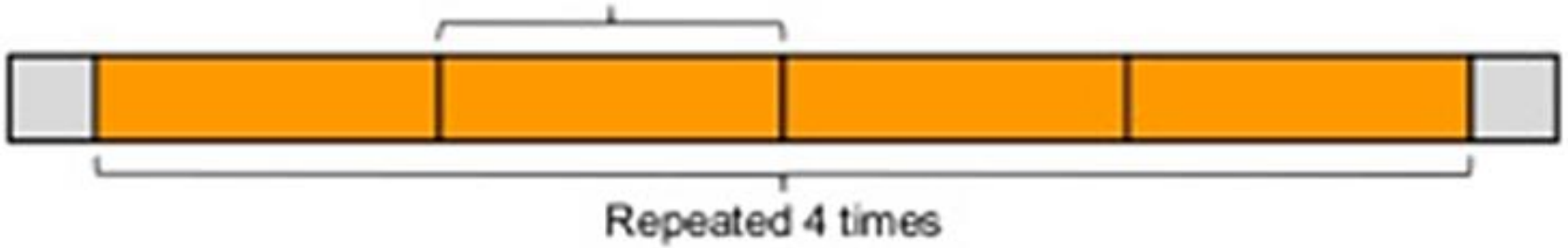
- (TTAGGG) is repeated hundreds to thousands of times at the termini of human chromosomes with a 3′ overhang of single-stranded DNA.
- The repeated sequences form loops that bind a protein complex called **shelterin**, which protects the chromosome termini from degradation.

- **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

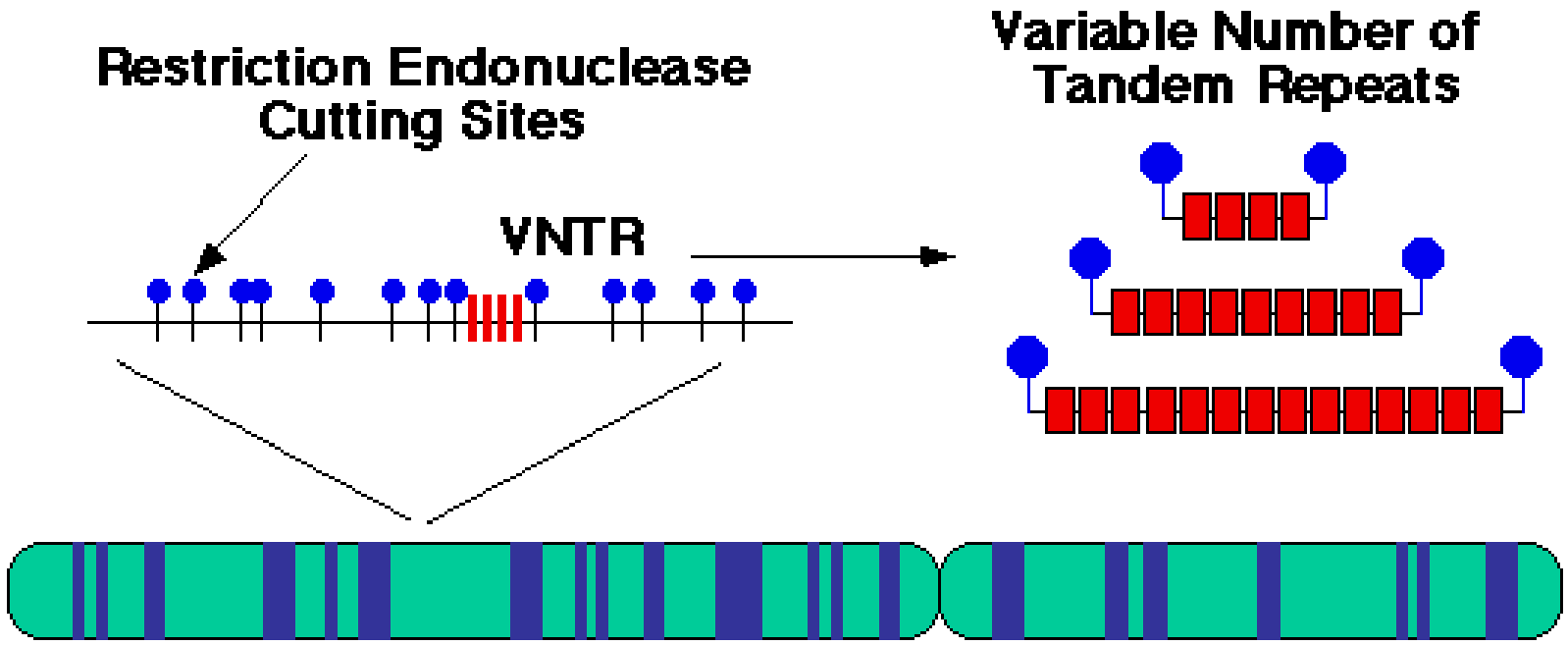
Minisatellite: Variable Number Tandem Repeats (VNTR)



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

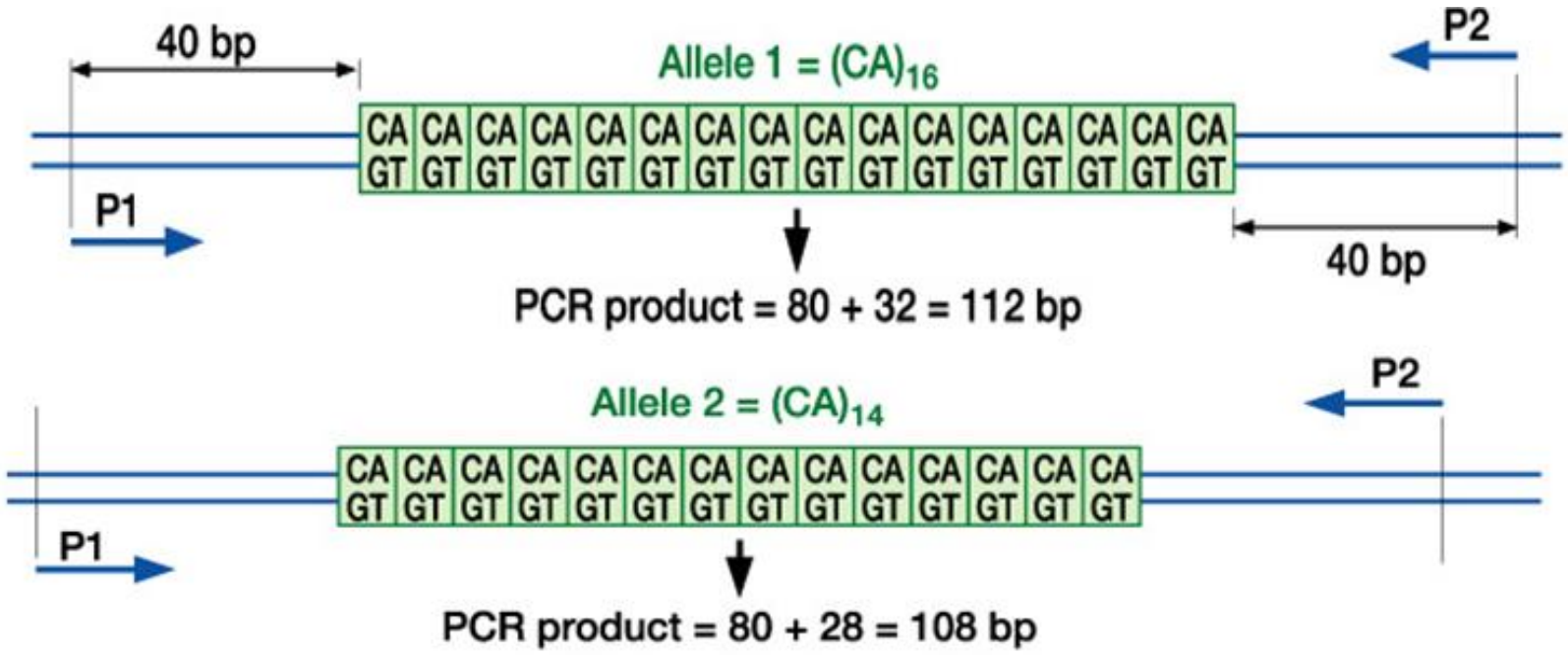
Mini-satellite DNA

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



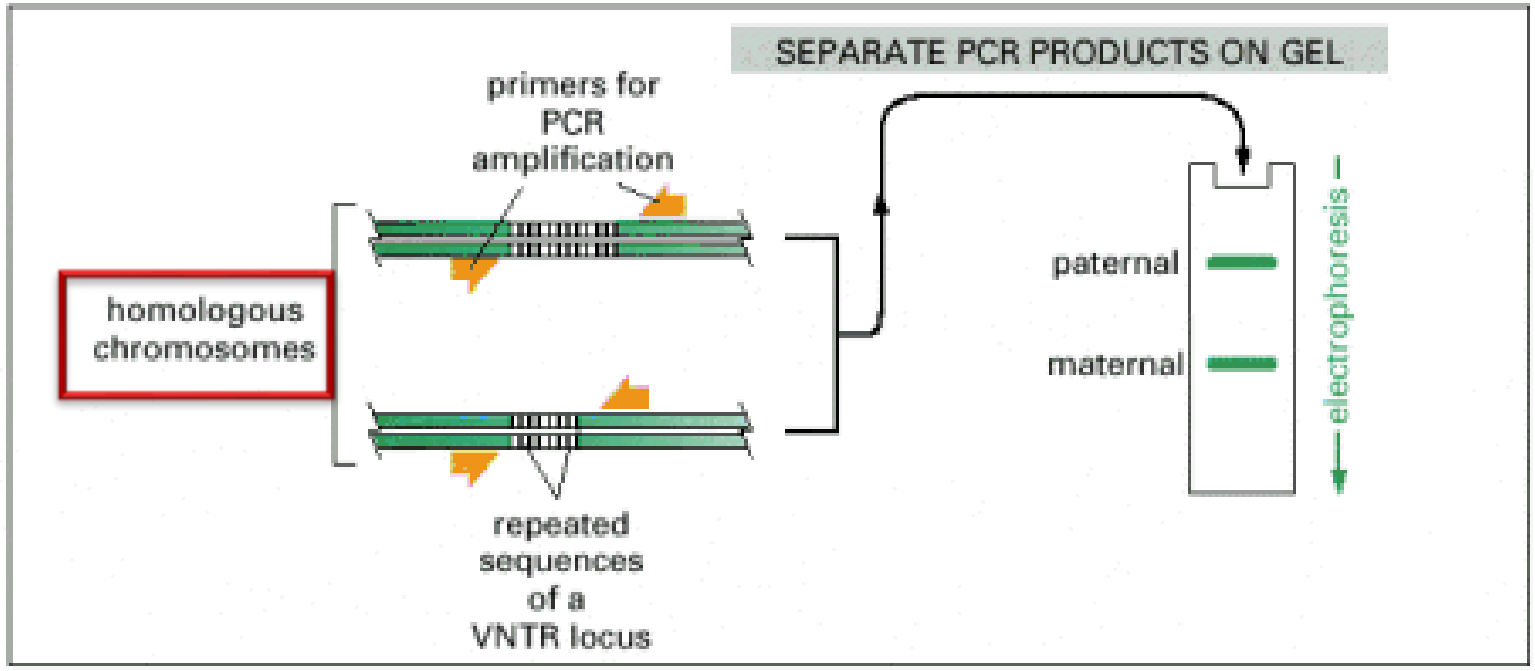
Micro-satellite DNA

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



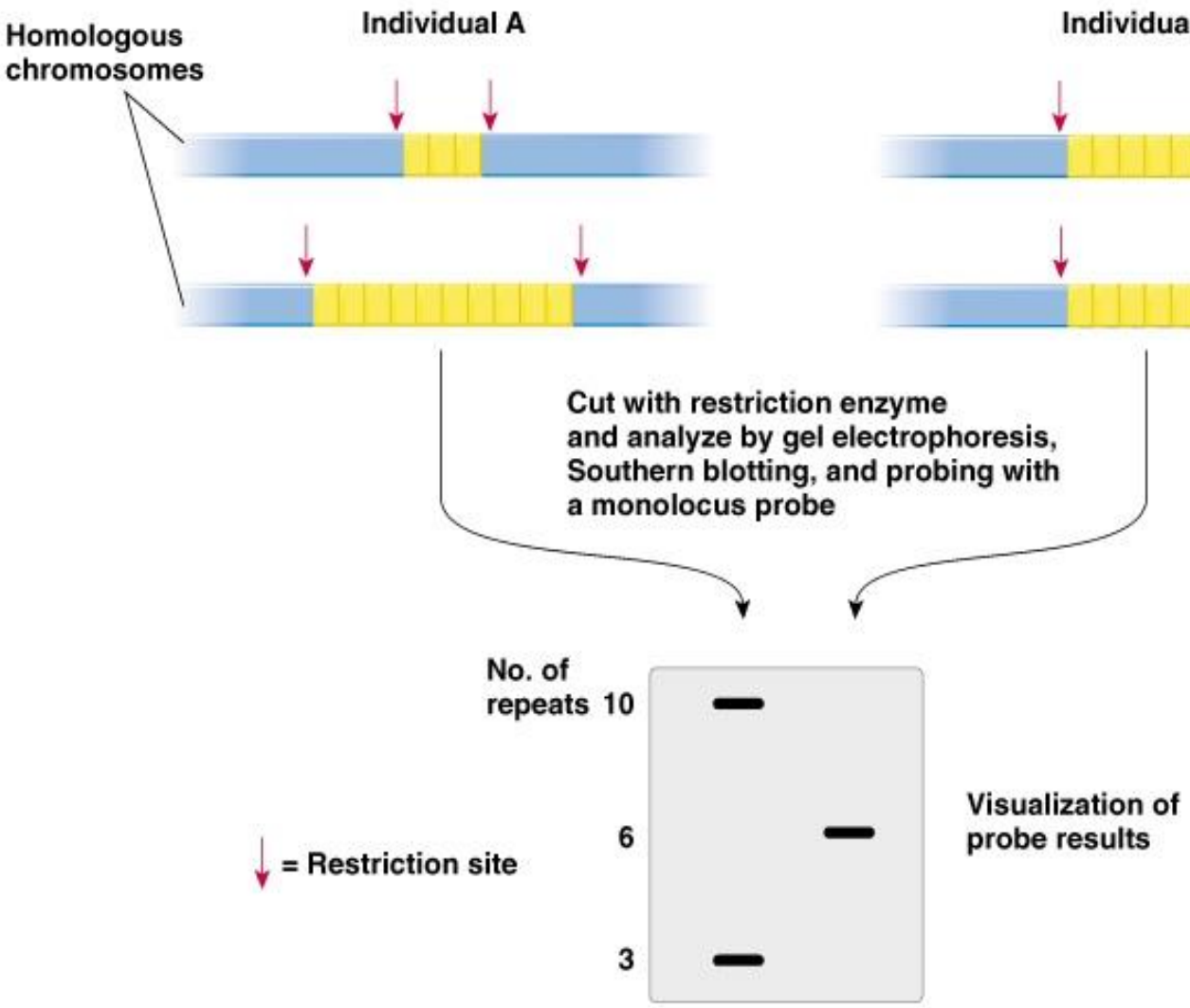
Polymorphisms of VNTR and STR

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

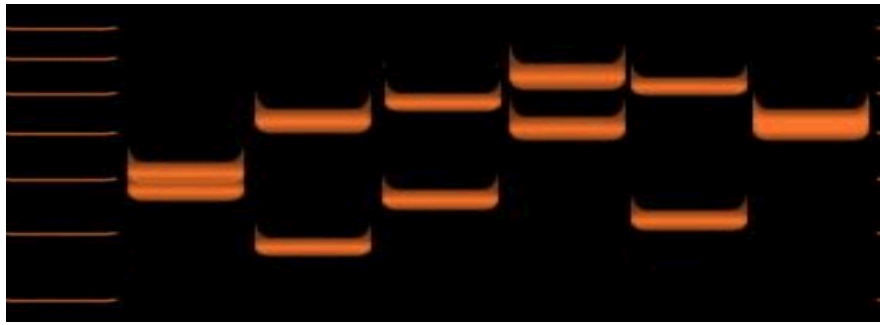


Homologous chromosome (or **homologs) are the** set of one maternal and one paternal chromosome somatic diploid cells.

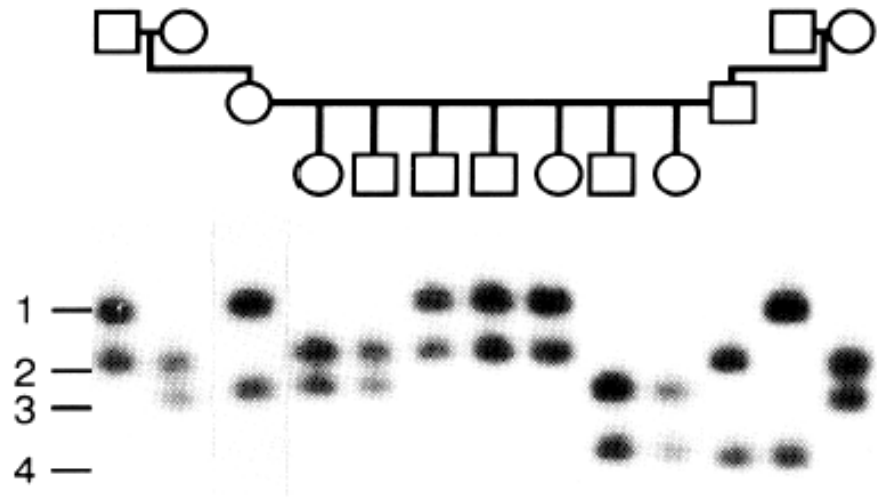
STRs and VNTRs as DNA Markers



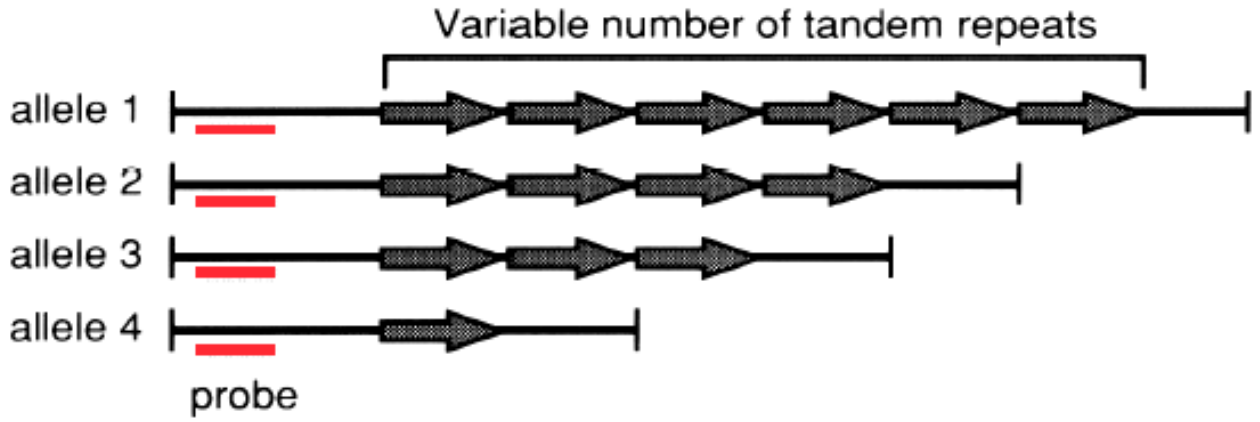
The likelihood of 2 unrelated individuals having same allelic pattern is extremely improbable.



Real example

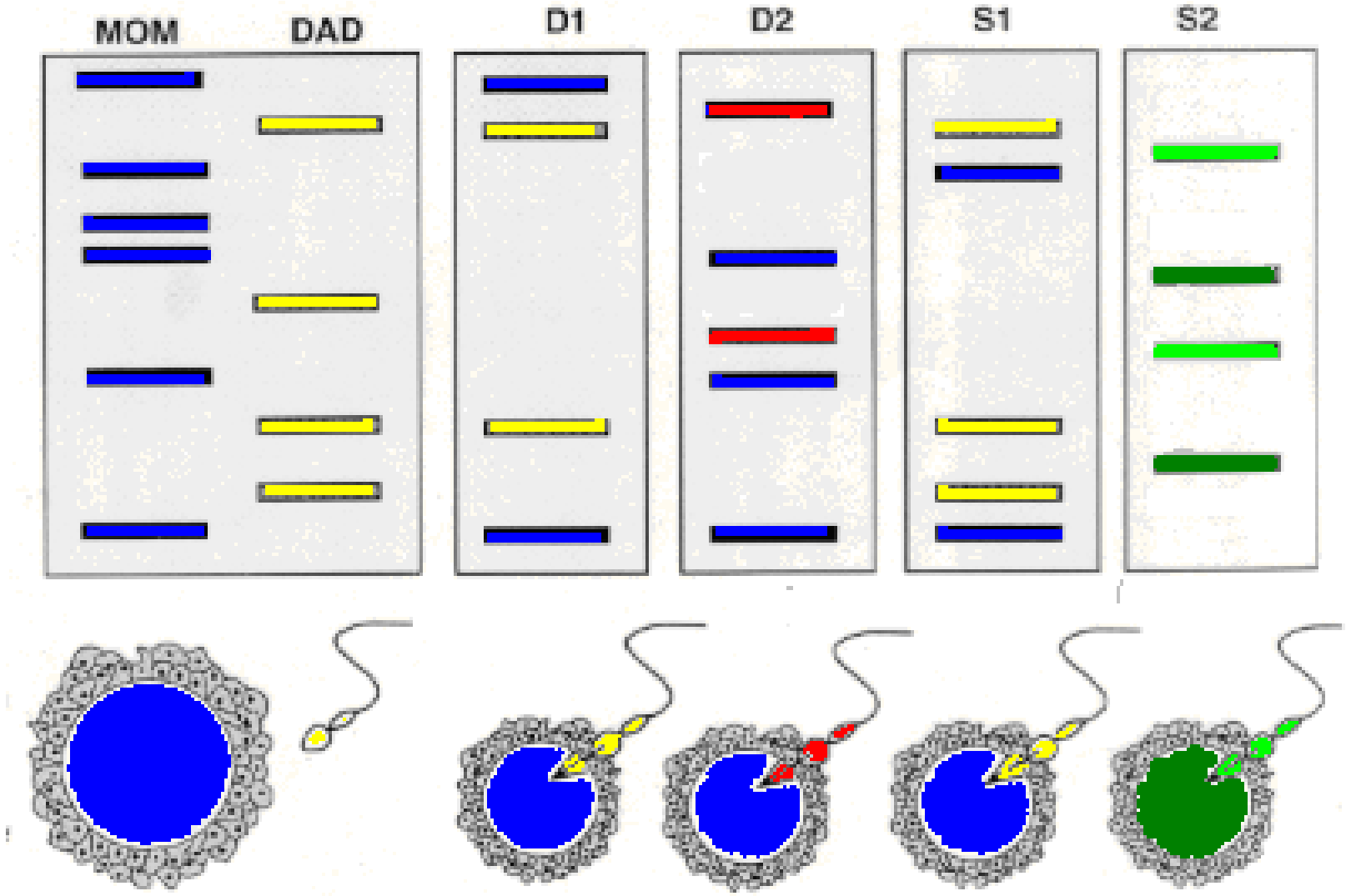


single-locus probe but multiple alleles



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

Paternity testing



Single nucleotide polymorphism (SNPs)

- Another source of 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 ...CGATATTCC**T**ATCGAATGTC...
copy1 ...GCTATAAGG**A**TAGCTTACAG...

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy2 ...GCTATAAGG**G**TAGCTTACAG...

Individual 2

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy1 ...GCTATAAGG**G**TAGCTTACAG...

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy2 ...GCTATAAGG**G**TAGCTTACAG...

Individual 3

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy1 ...GCTATAAGG**A**TAGCTTACAG...

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy2 ...GCTATAAGG**A**TAGCTTACAG...

Individual 4

Chr 2 ...CGATATTCC**T**ATCGAATGTC.
copy1 ...GCTATAAGG**A**TAGCTTACAG.

Chr 2 ...CGATATTCC**C**ATCGAATGTC.
copy2 ...GCTATAAGG**G**TAGCTTACAG.

Individual 5

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy1 ...GCTATAAGG**G**TAGCTTACAG...

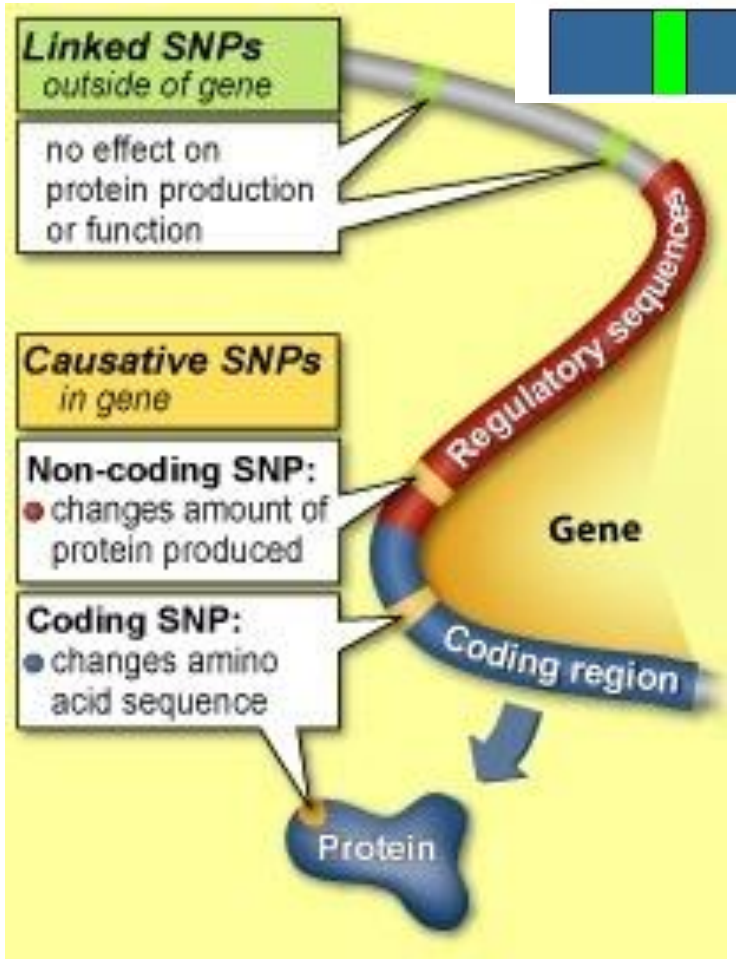
Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy2 ...GCTATAAGG**A**TAGCTTACAG...

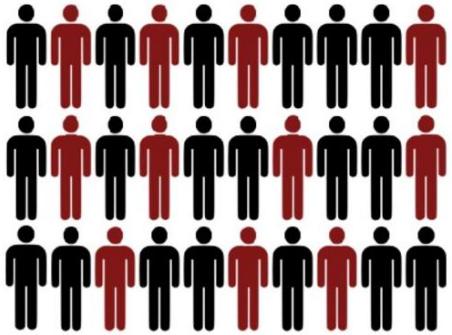
Individual 6

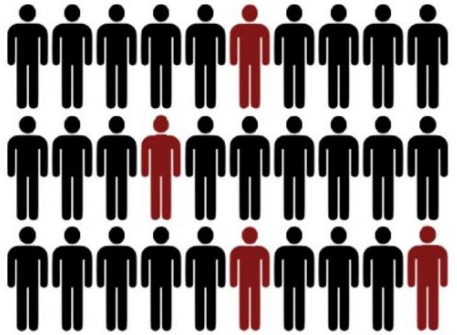
Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy1 ...GCTATAAGG**G**TAGCTTACAG...

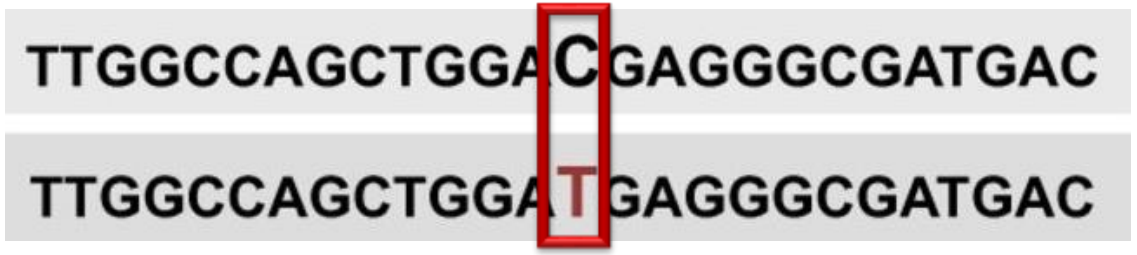
Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy2 ...GCTATAAGG**A**TAGCTTACAG...

Categories of SNPs









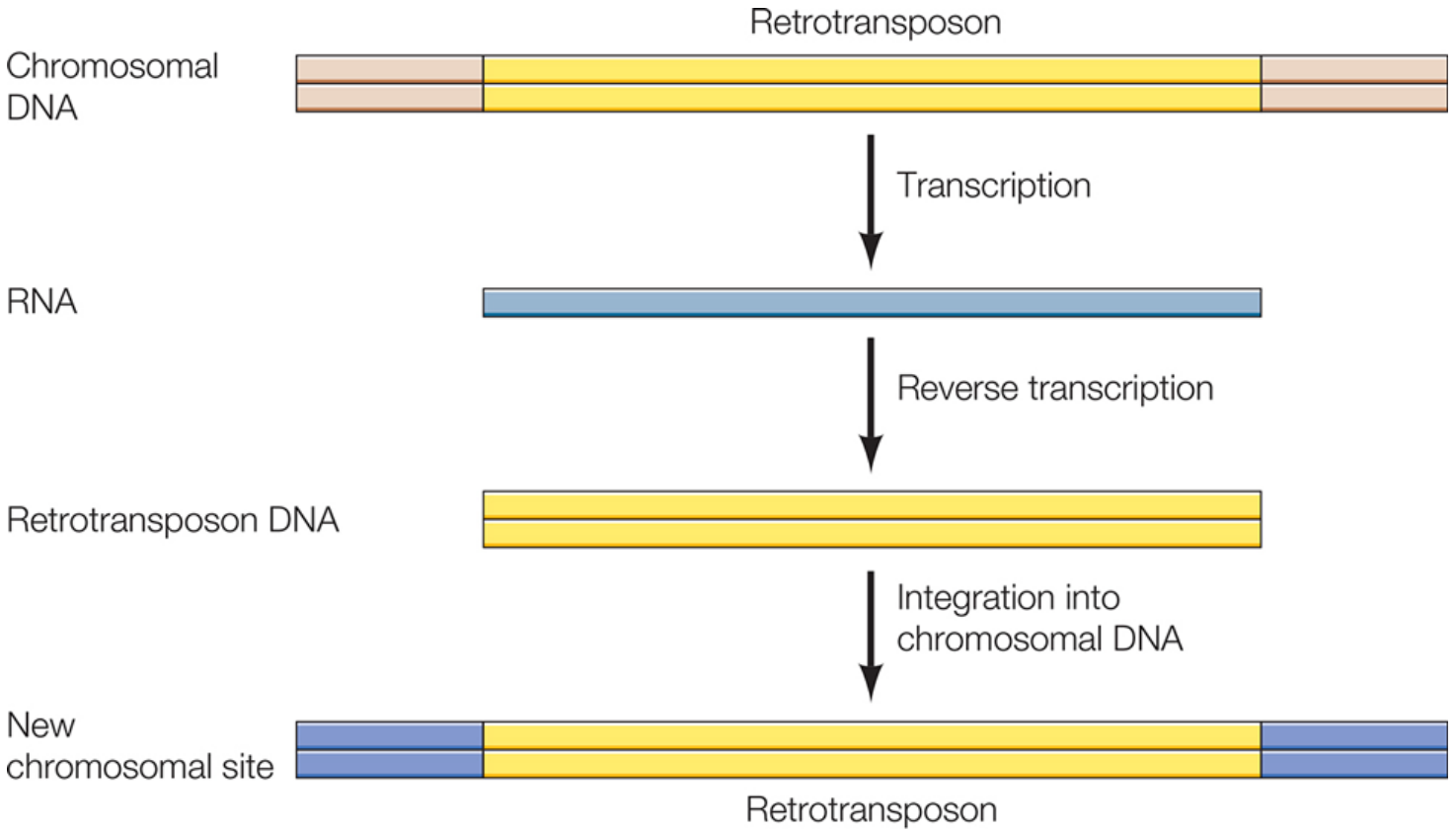
Interspersed repeats

Transposons (jumping genes)

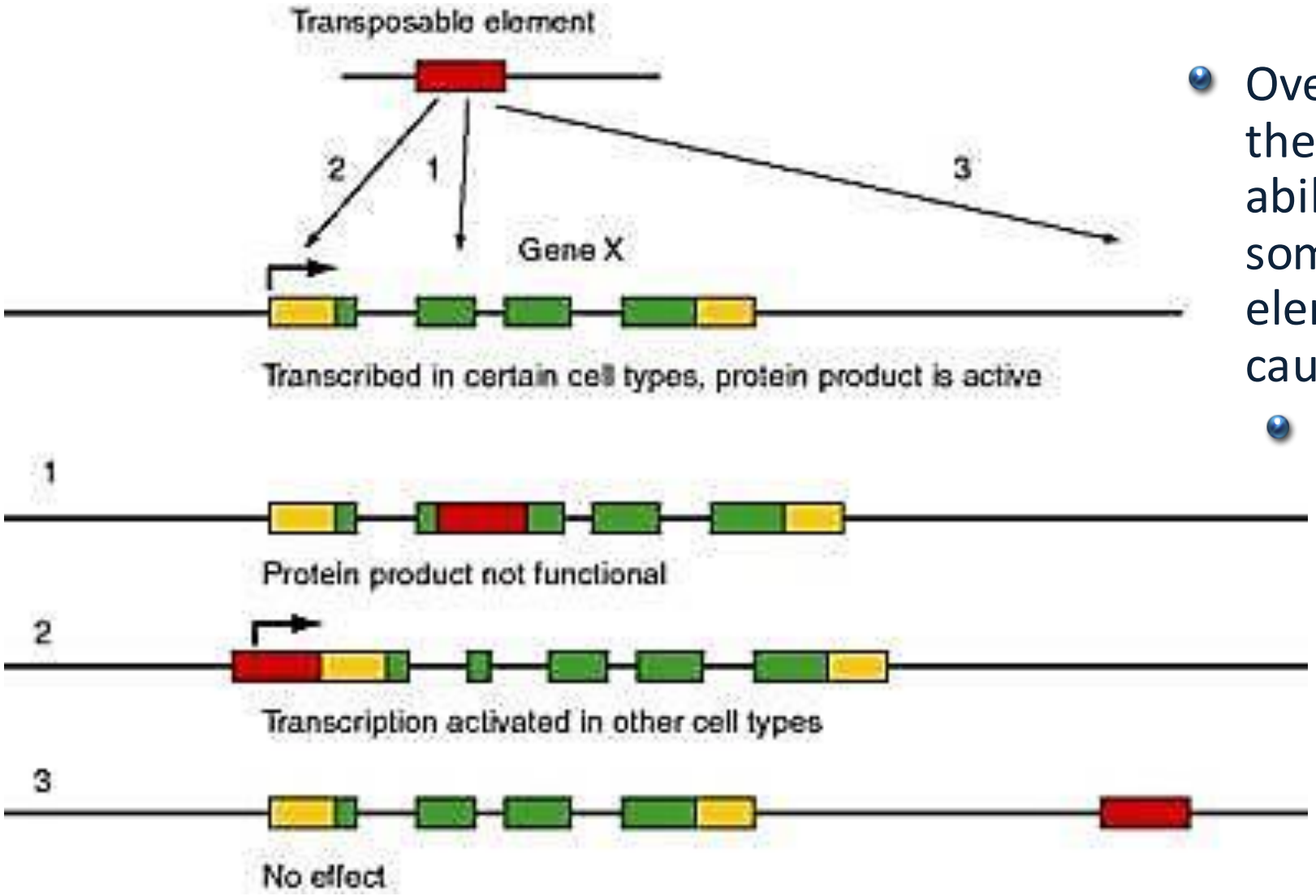
- They are segments of DNA that can move from their original position in the genome to a new location.
- Two classes:
  - DNA transposons (3% of the human genome)
  - RNA transposons or retrotransposons (42% of human genome).
    - Long interspersed elements (LINEs, 21%)
    - Short interspersed elements (SINEs, 13%)
      - An example is Alu (300 bp)
    - Retrovirus-like elements (8%)

How doe retrotransposons move and integrate?

- A retrotransposon present at one site in chromosomal DNA is transcribed into RNA.
- The RNA is converted back into DNA by reverse transcriptase.
- The retrotransposon DNA can then integrate into a new chromosomal site.
- LINEs contain reverse transcriptase genes and the integrase gene that is necessary for integration into cellular DNA.



The outcome of transposition



- 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.