Molecular Biology (3)

The human genome

Mamoun Ahram, PhD

Resources

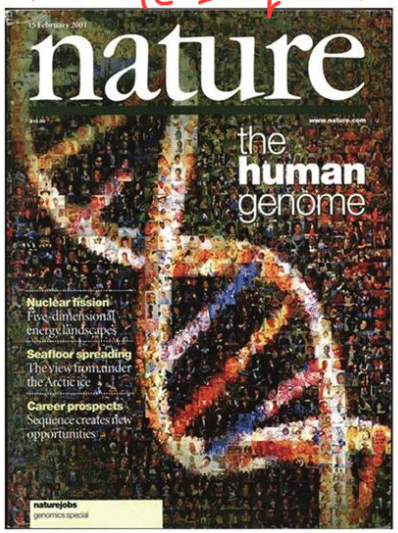
- This lecture
- Cooper, Ch. 6, pp. 157-160, 195-205, 209-212

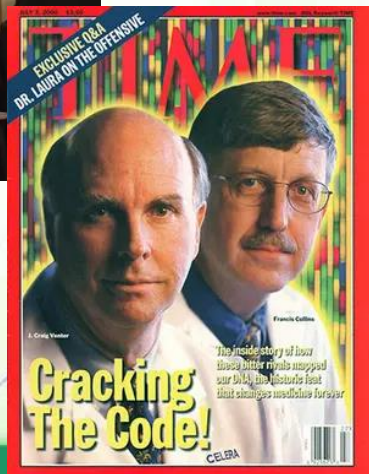
The human genome project

→ total collection of DNA

- 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).

→ the sequence of Nitrogen bases





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

absolutly not important
(now we can know the number of gens in about 24 hours and it cost about 900 $)

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

the doctor did not said memorise it and it is old numbers too (I will not even if he said that 😂)

Nucleotides per genomes

He said not for memorizing

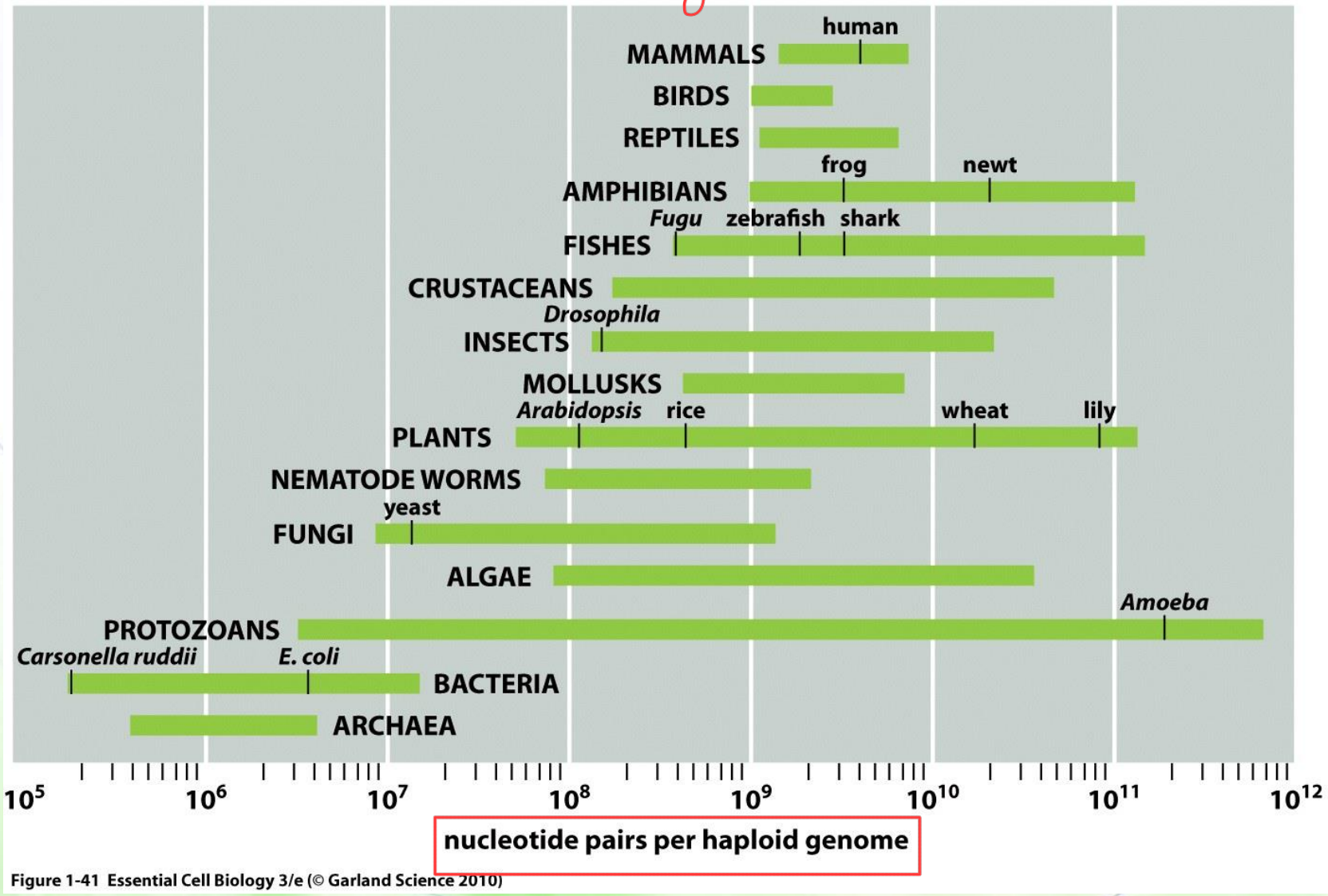
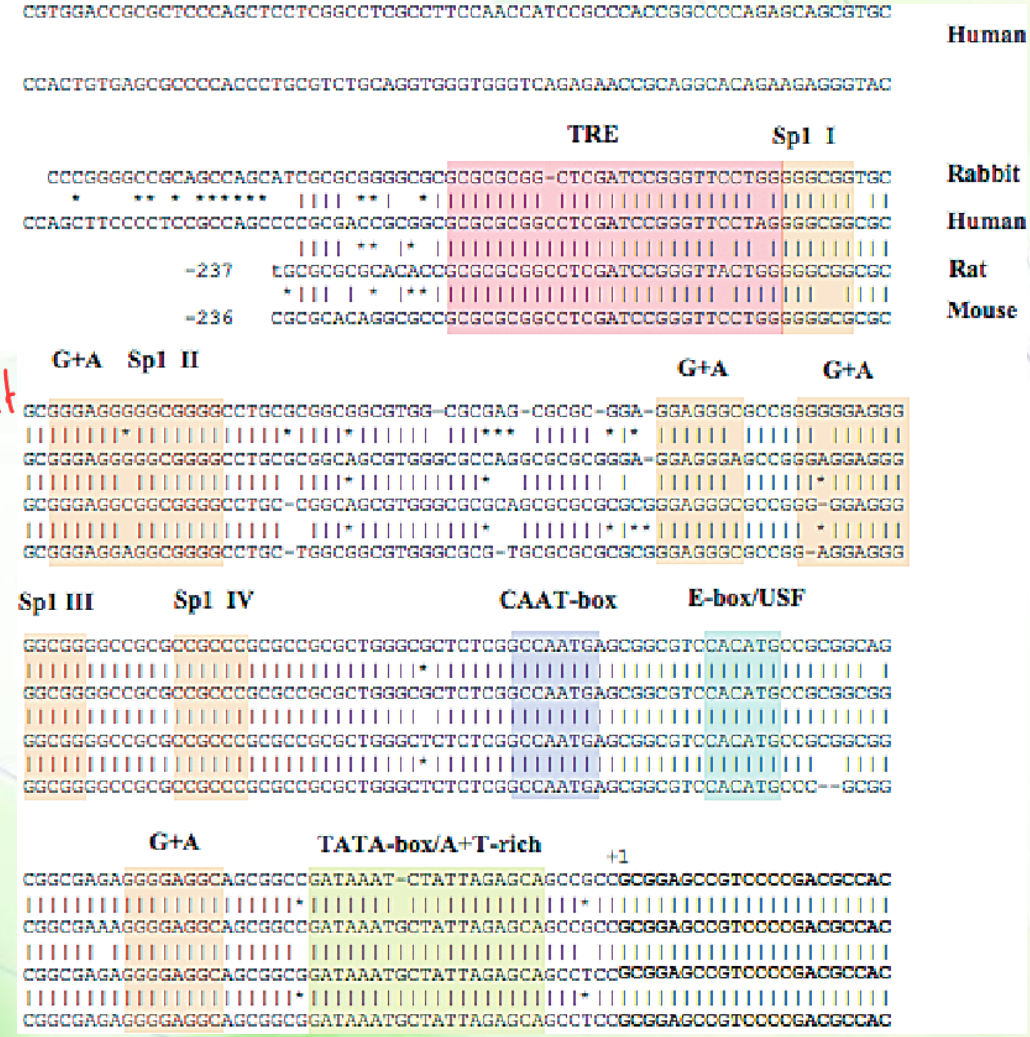


Figure 1-41 Essential Cell Biology 3/e (© Garland Science 2010)

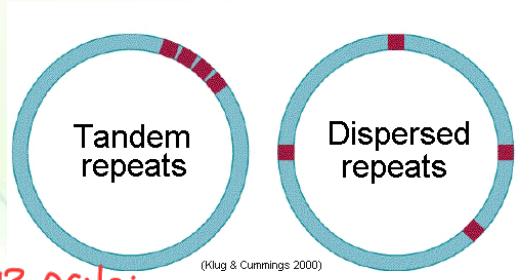
DNA homology

(i.e. sequence similarity)

there are similer to each other in some regions which is most important region



Components of the human genome



- Human genome
  - Nuclear genome 3300 Mb (each strand has 3300 Mb)
    - Genes and gene-related sequences (30%)
      - Noncoding sequences (28%) — but it still important
        - Introns (26%) — transcripted but not translated
        - Pseudogenes — fake (his function is lost)
        - Other sequences (e.g. untranslated regions) — do not transfered to protein
      - Protein-coding sequences (2%) — which will encode proteins
    - Extragenic (intergenic) sequences (70%) — outside — in between
      - Unique or low copy number (5%)
      - Moderate to highly repetitive (55%) — element: in DNA = sequence, in factors = proteins
        - TANDEM repetitive sequences (10%) — repetited after themselves (abcabcabc)
          - Satellites (centromeres and telomeres) — on the poles (يعني بجاد او مطرفين)
          - Minisatellites (VNTRs) — will talk about them
          - Microsatellites (STRs)
        - Interspersed (متناثر)
          - DNA transposons (3%) — moveable DNA elements
          - Retrotransposons (42%)
            - SINEs (13.%) — short
            - LINEs (21%) — long
            - Retrovirus-like elements (8%) — we don't know what they do
      - Other (10%) — DNA that come from RNA

Mitochondrial genome 16.6 Kb — encode 13 protein

he said that the transposons and retrotransposon is from viruses but I doubt that

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

an agelin he mention the evlutionary method

Note: All numbers are approximate

So I advise you to watch الدكتور اياد القيسي

رحلة ، ليقين

The ENCODE project (2003-on)

all the numbers all approximate (don't memorize it significant number) but memorize it as approximatly number so it maybe asked in the exam

- 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%

but 2% transfer to protein

will talk about it.

→ to transfer Data to information

On March 31, 2022...

remember in 2004 they have finished 92 % of 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 (number of genes that make protein)	19,969
Number of exclusive genes	3,604
Protein coding	140
Number of transcripts	233,615
Protein coding (protein number)	86,245
Number of exclusive transcripts	6,693
Protein coding	2,780

after proccesses they could be 200k

RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

to understand the repetitive sequence of DNA

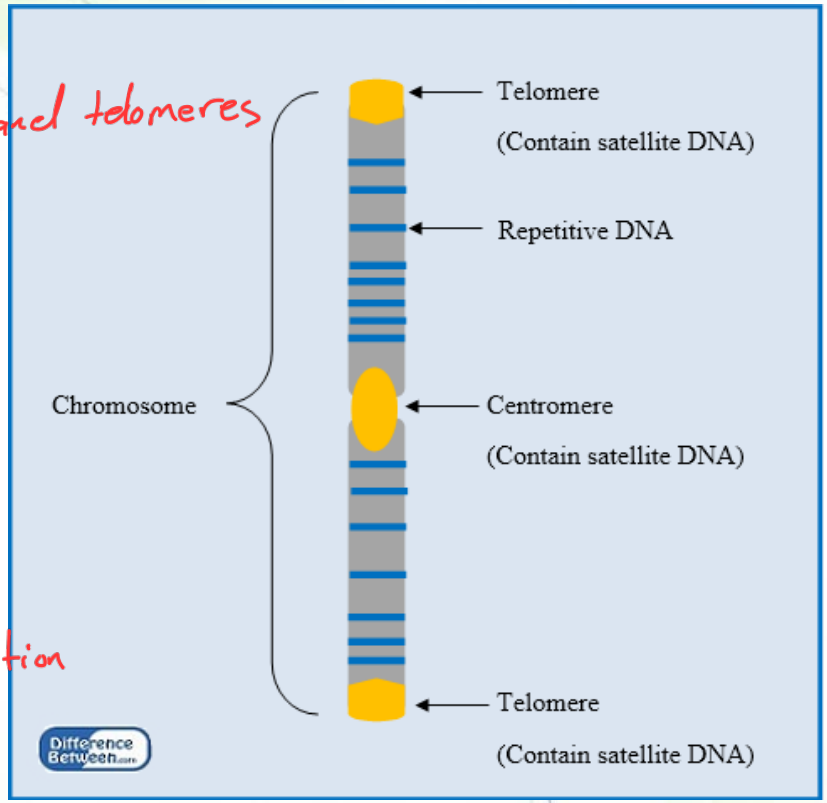
Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair 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.

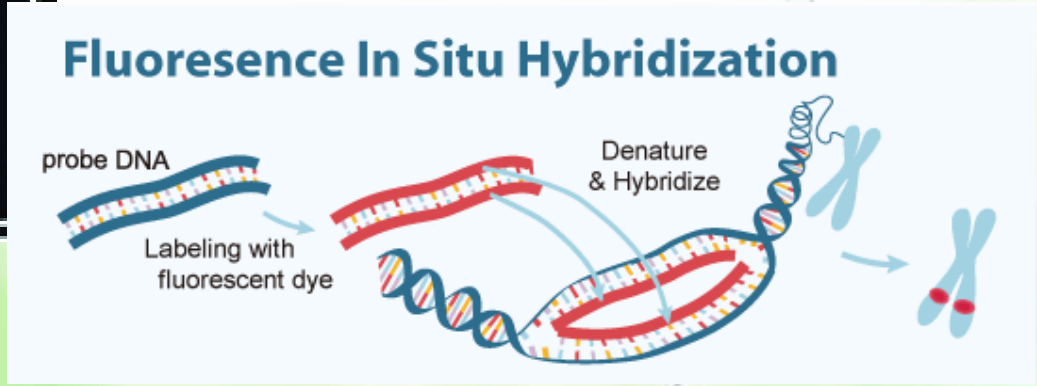
Tandem repeats

Satellite (macro-satellite) DNA

- Regions of 5-300 bp repeated 10^6-10^7 times and found in centromeres and telomeres
  - 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)**. in specific location

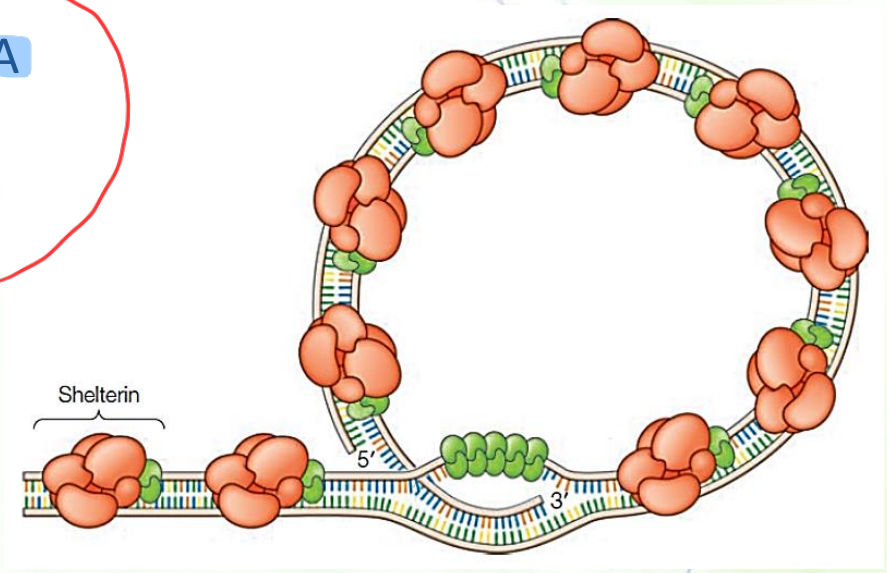
→ I will hybridize a prob that move to specific location in the chromosome





Telomeric repeats

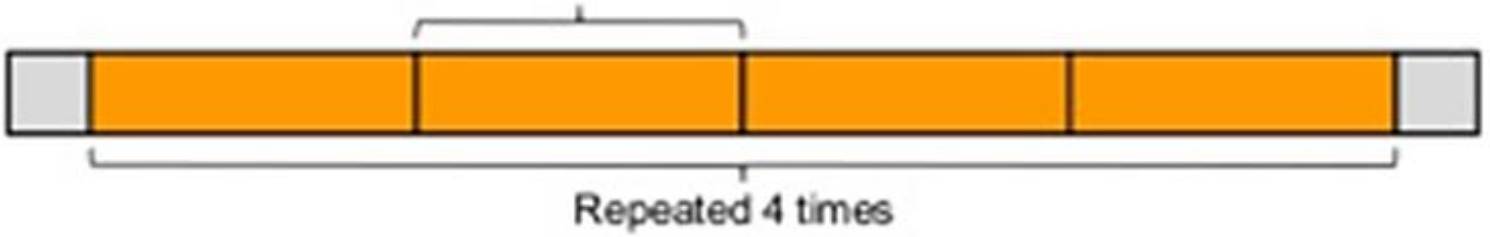
- (TTAGGG) is repeated hundreds to thousands of times at the termini of human chromosomes with a 3′ overhang of single-stranded DNA.
- The repeated sequences form loops that bind a protein complex called **shelterin**, which protects the chromosome termini from degradation.
- **Telomeric repeat-containing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

another name

Minisatellite: Variable Number Tandem Repeats (VNTR)

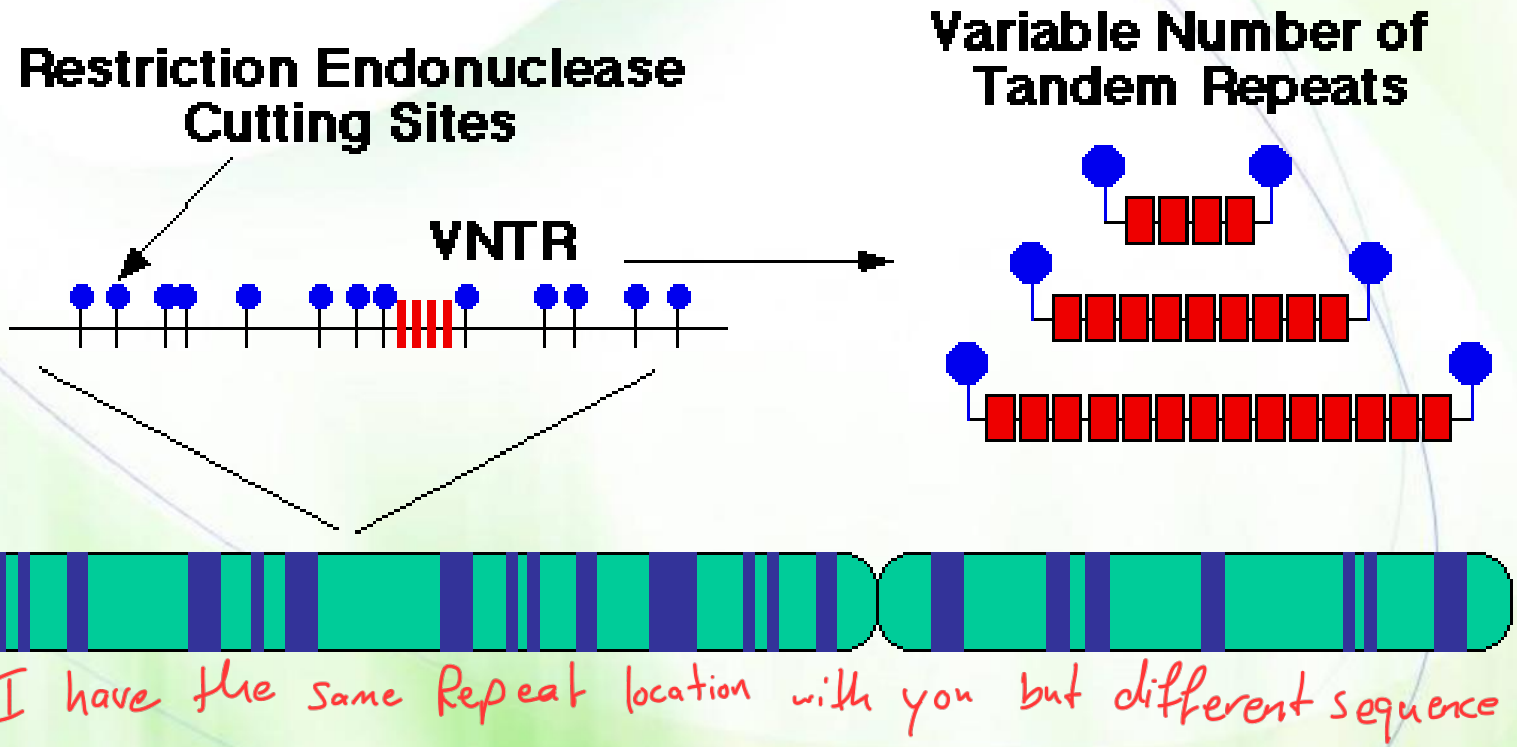


Microsatellite: Short Tandem Repeats (STR) – Simple Sequence Repeats (SSR)

the only different is the size

Mini-satellite DNA

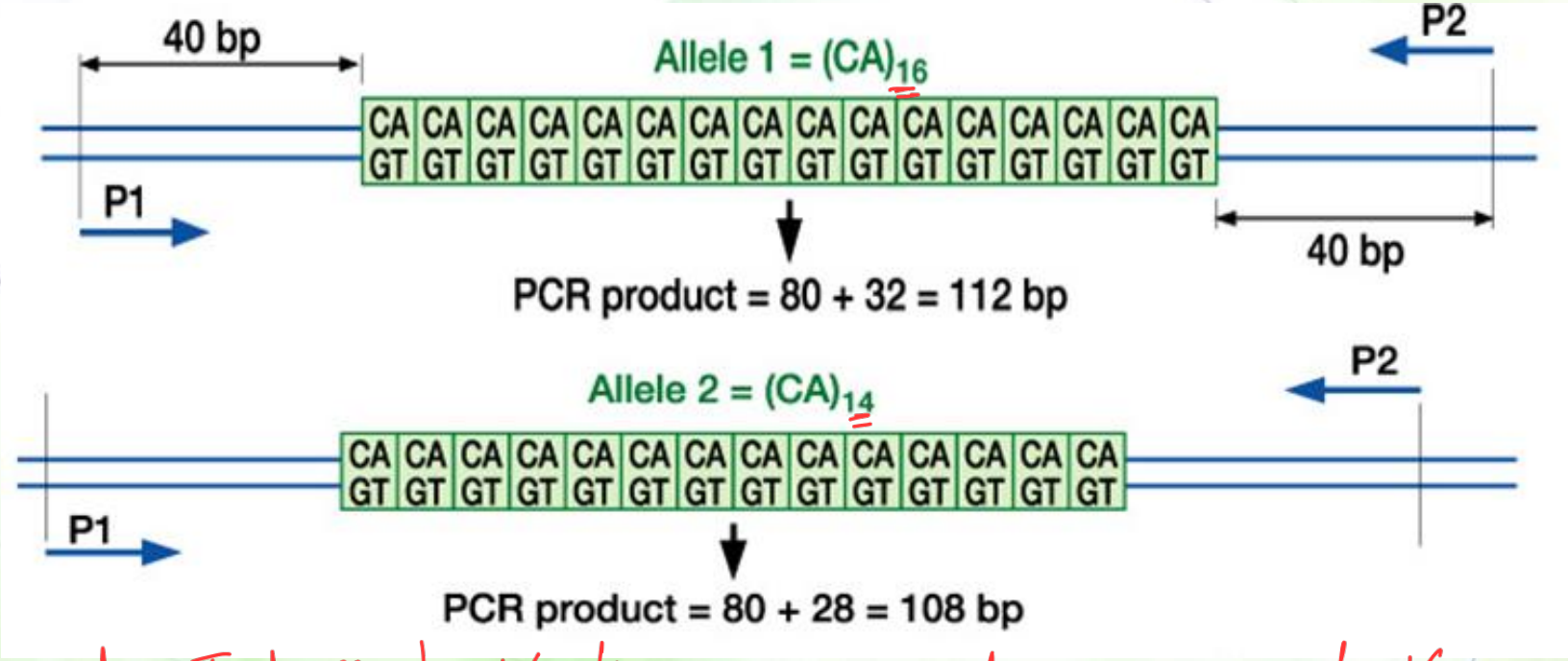
- Mini satellite sequences or VNTRs (variable number of tandem repeats) of 20 to 100 bp repeated 20-50 times



Micro-satellite DNA

Same the mini-satellite ⇒ Same location but different sequence

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



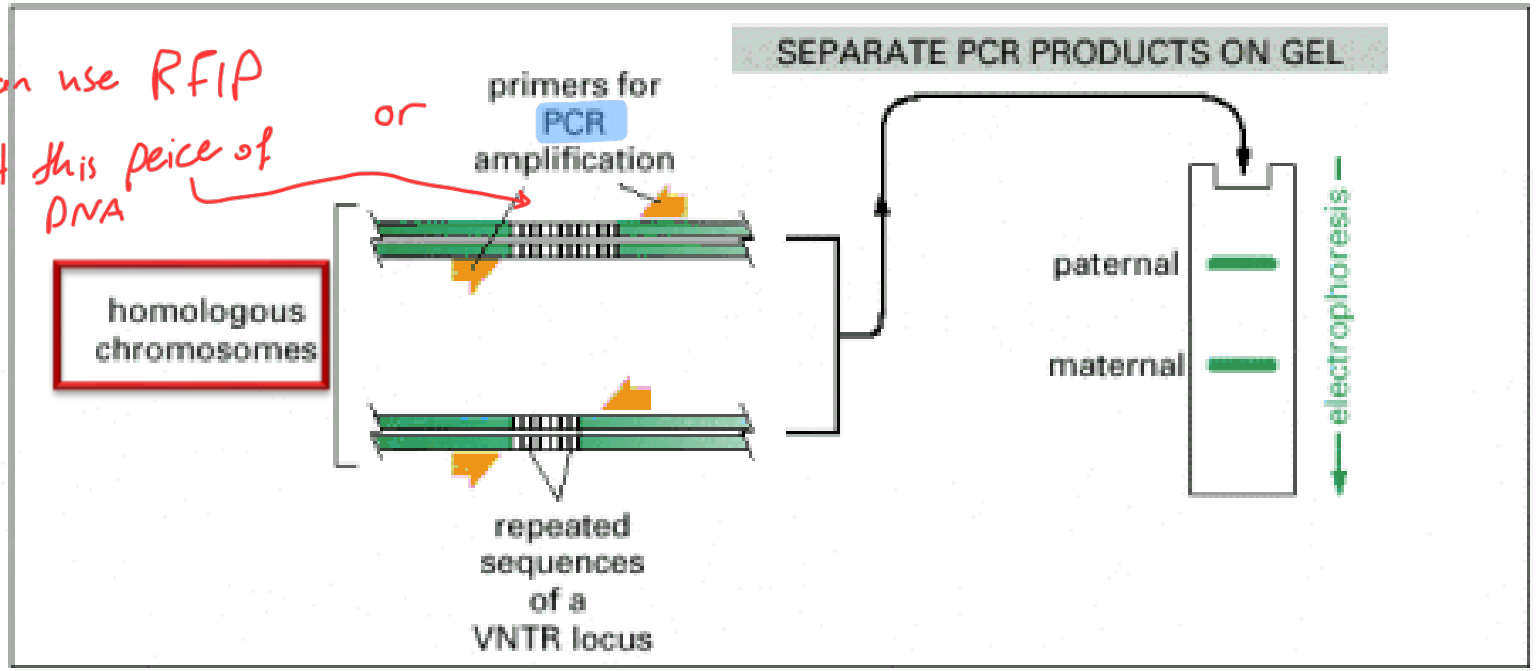
may be I have it 16 times on one chromosome and 16 on another because it is diploid chromosomes

Polymorphisms of VNTR and STR

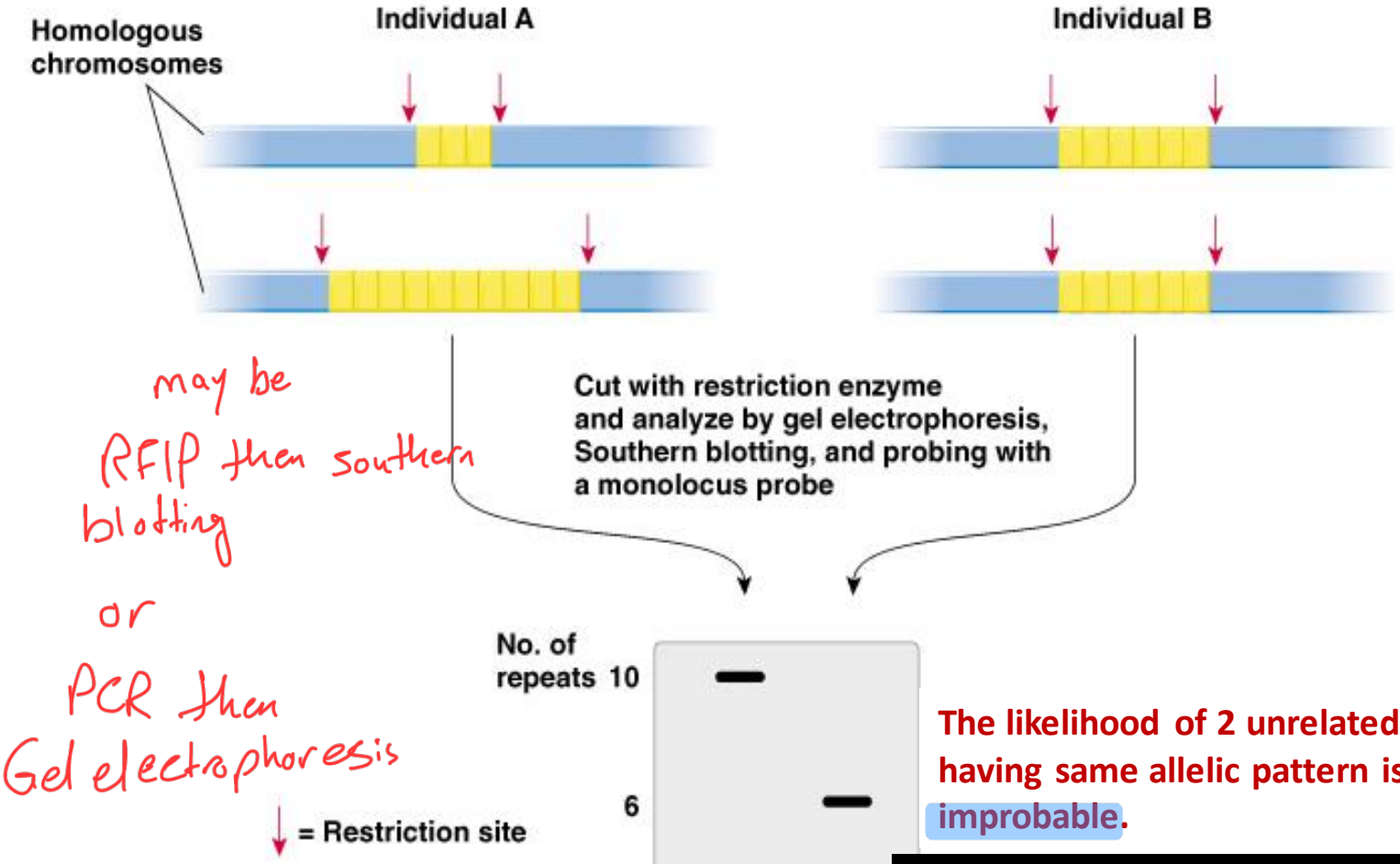
differences in DNA

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

we can use RFLP to cut this peice of DNA

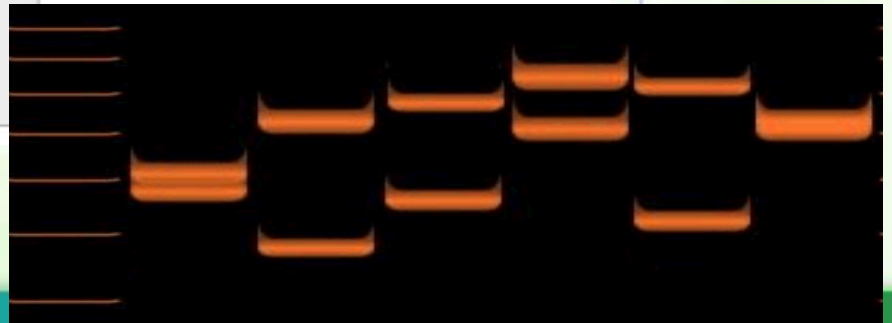


STRs and VNTRs as DNA Markers



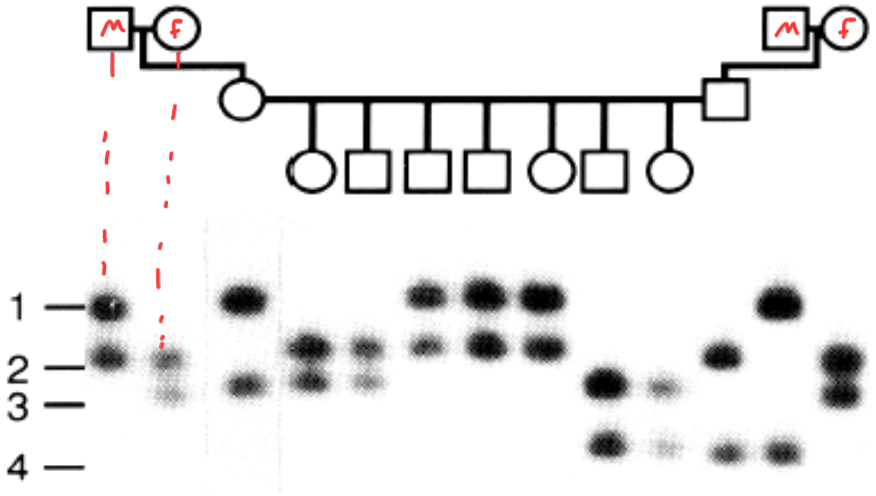
may be
RFLP then southern blotting
or
PCR then Gel electrophoresis

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

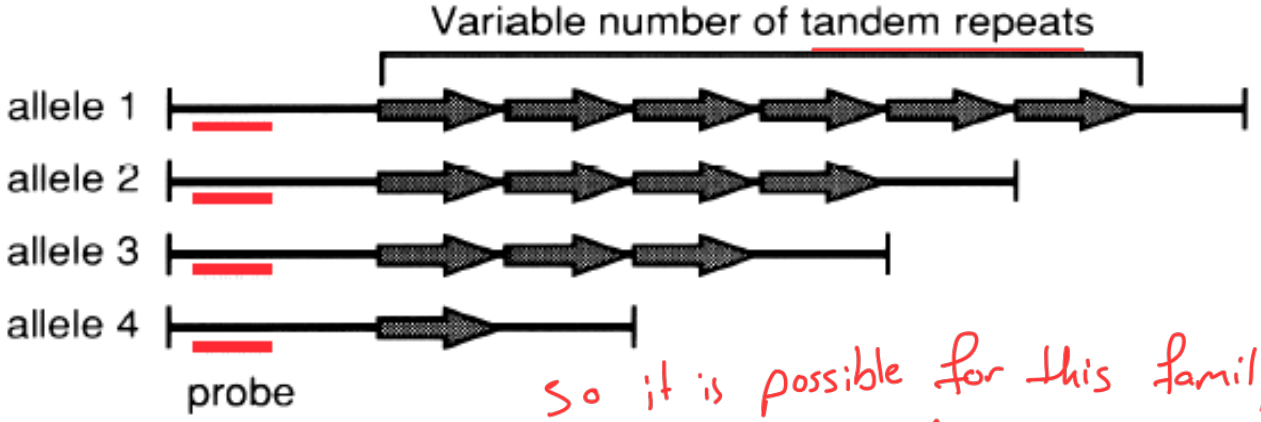


Real example

STR is more used than VNTR



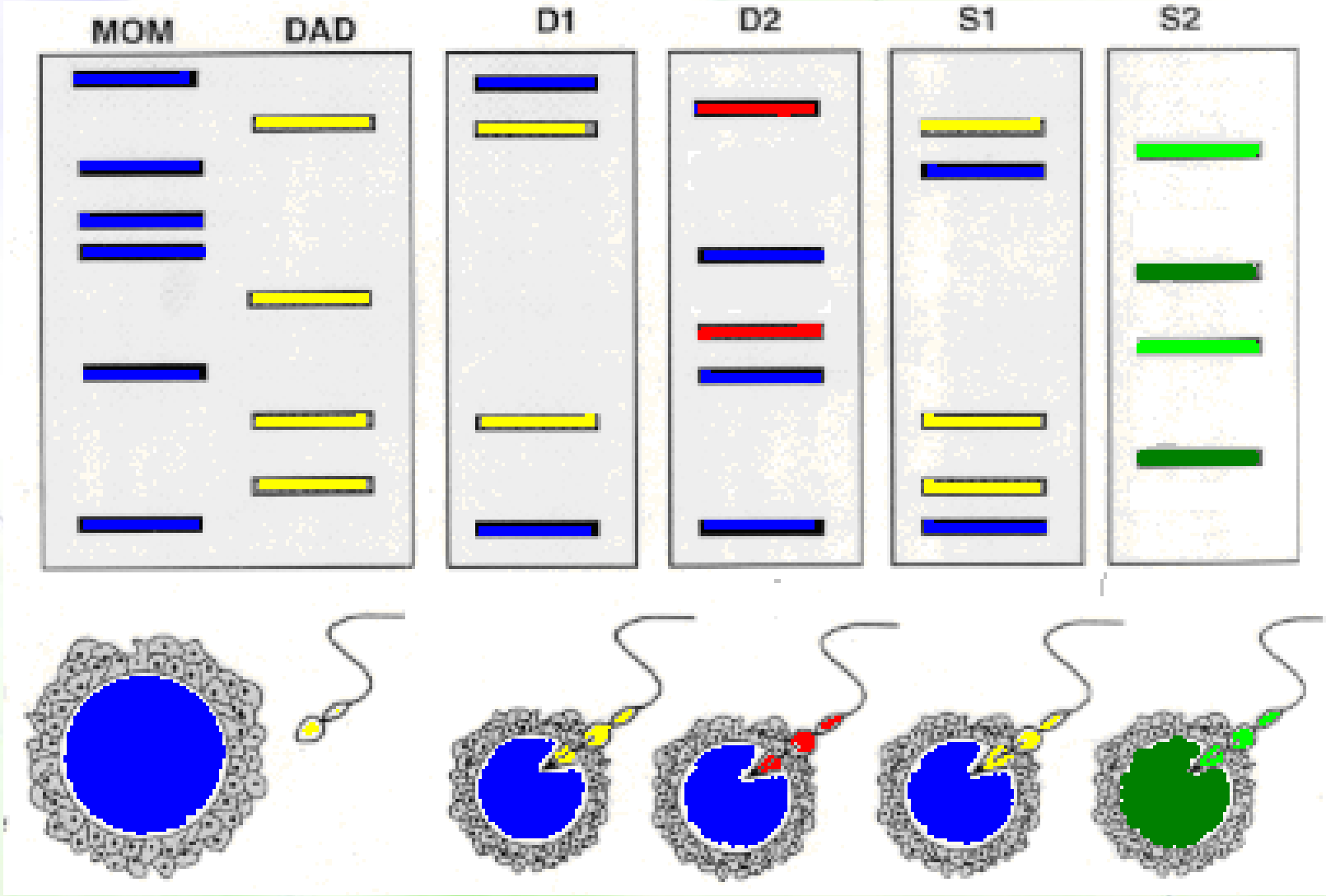
single-locus probe but multiple alleles



So it is possible for this family to repet the tandem 1, 3, 4, 6 times

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

that connected with desease or something important

Examples

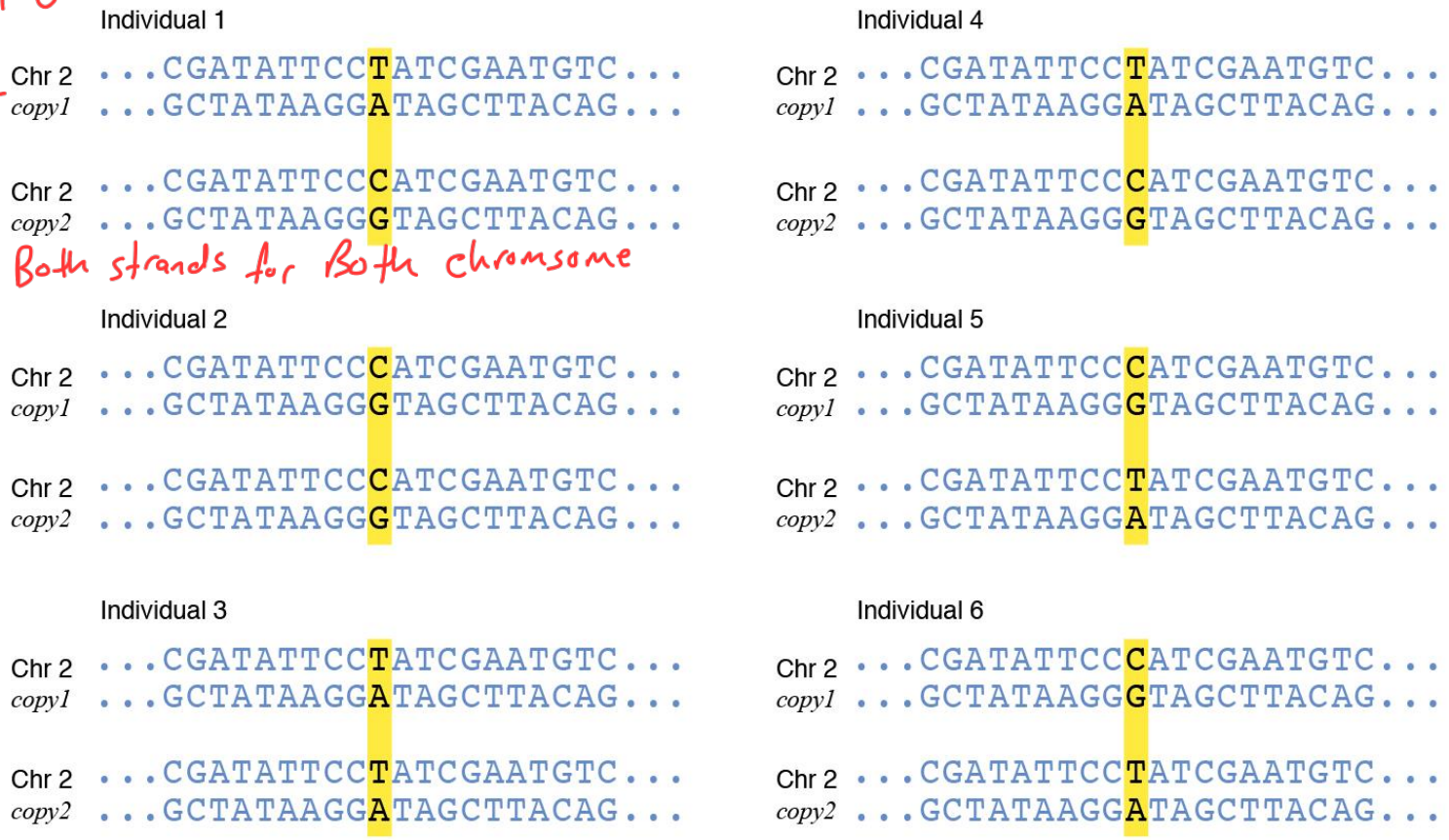
the percent of of SNPs must be more 1% of population otherwise we call it mutation

2 strands for 1 chromosome

Hetrozygous SNPs

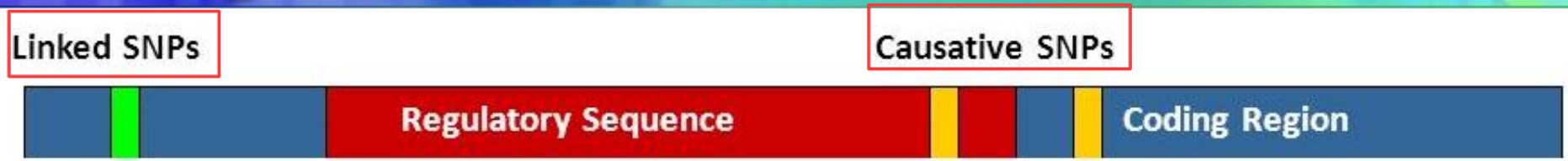
Homozygous SNPs

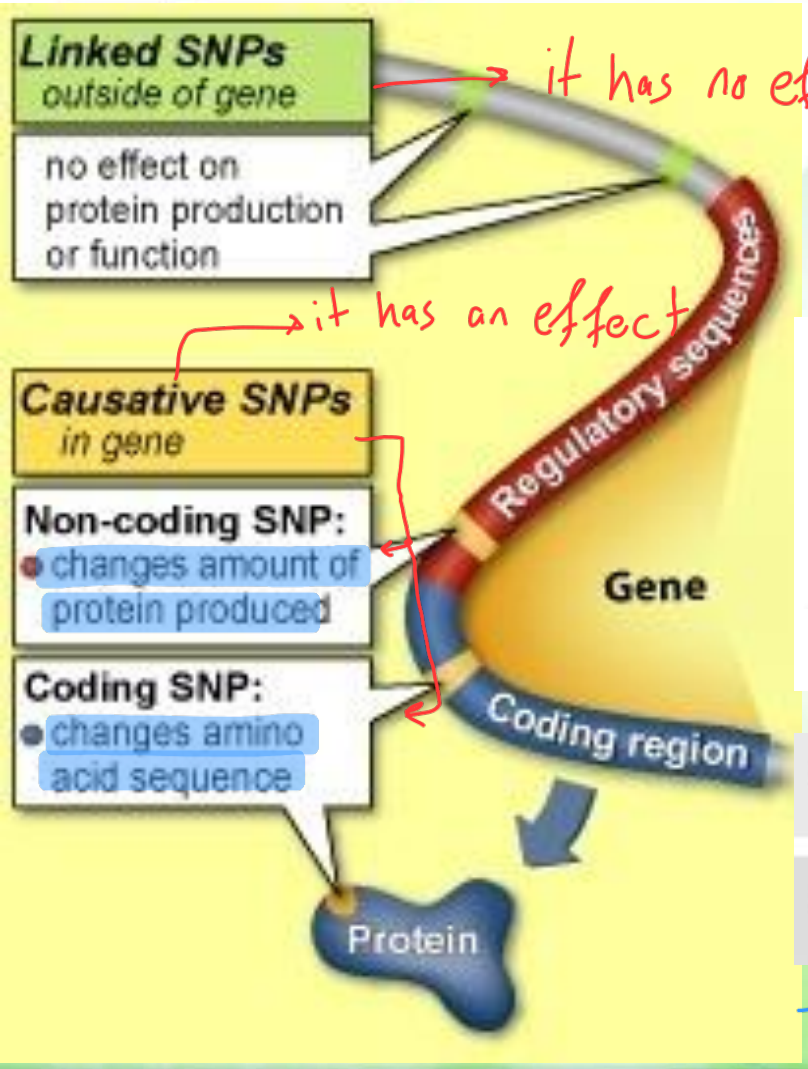
Both strands for Both chromsome



Categories of SNPs

do not forget we talk about 500k SNPs



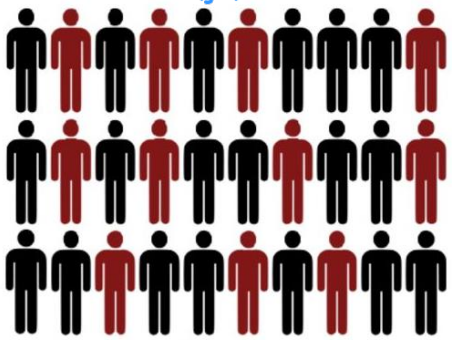


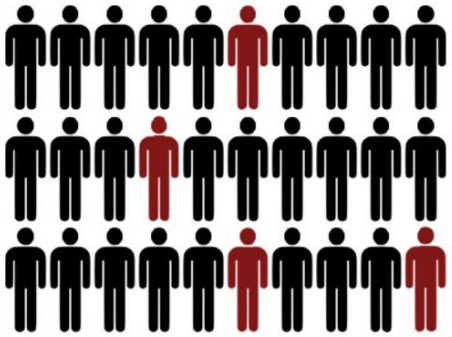
but if I removed this gene the phenotype will

Not removed

Some of them has C but little

So some of them has T but little





healthy

TTGGCCAGCTGGA**C**GAGGGCGATGAC

TTGGCCAGCTGGA**T**GAGGGCGATGAC

that does NOT mean T is the reason of the desease

Interspersed repeats

Repeated Regions but with spaces between them

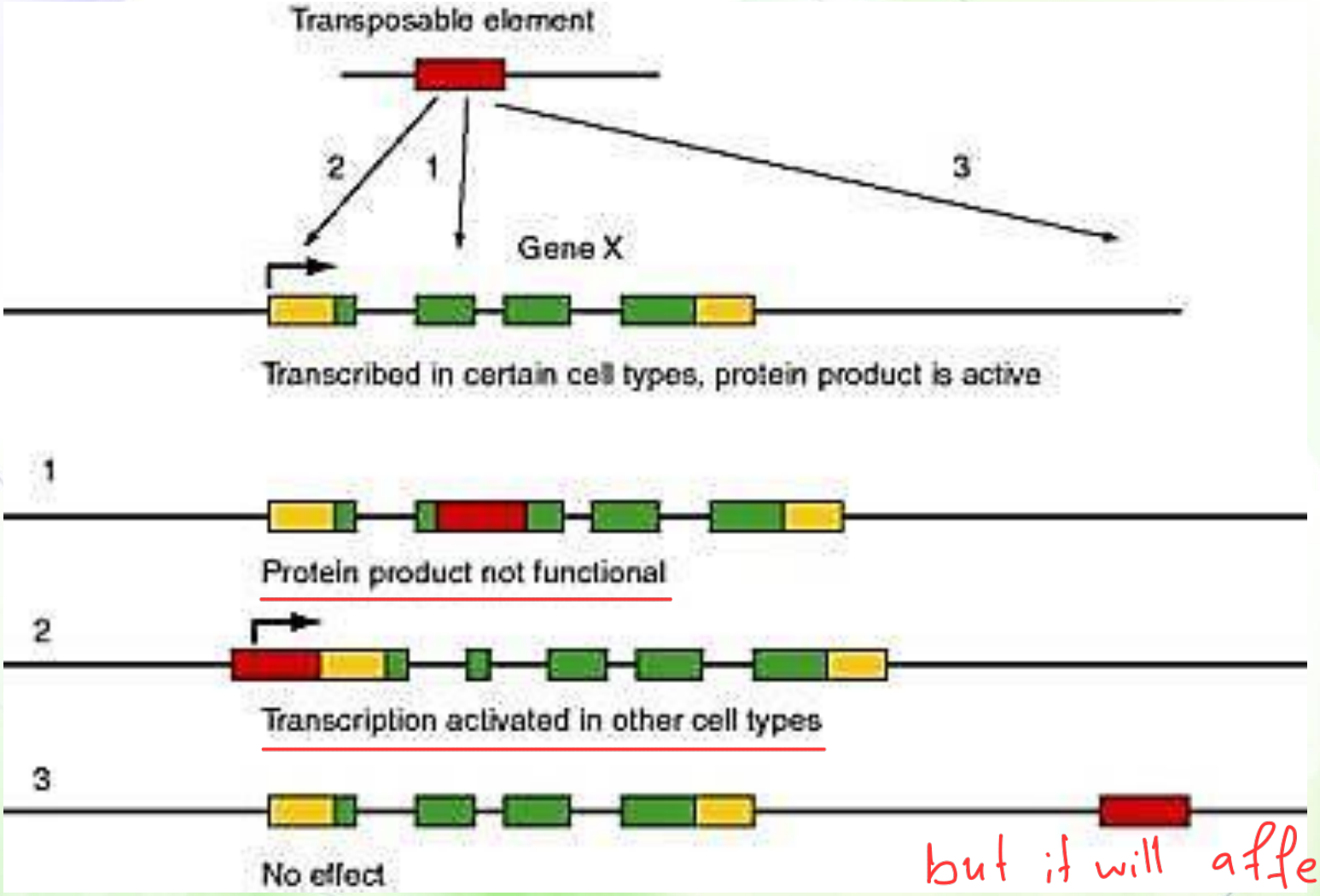
(متباعدة)

Transposons (jumping genes)

Just another name

↳ can move (but in human it mostly lost ability to move)

- 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) from DNA of virueses
  - RNA transposons or retrotransposons (42% of human genome).
    - Long interspersed elements (LINEs, 21%)
    - Short interspersed elements (SINEs, 13%)
      - An example is Alu (300 bp) → an restriction endo clease called Alu can cut this element
    - 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.



but it will affect other non coding elements