

FINAL – Lecture 4

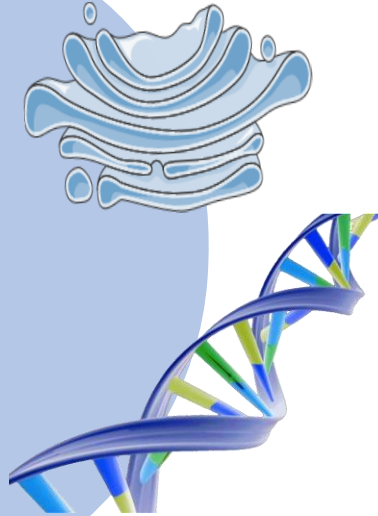
# The Human genome

﴿ وَإِن تَتَوَلَّوْا يَسْتَبَدِلْ قَوْمًا غَيْرَكُمْ ثُمَّ لَا يَكُونُوا أَمْثَلَكُمْ ﴾

اللهم استعملنا ولا تستبدلنا

Written by :

- Rama Al\_oweyrat
- Isra'a Mohammad



# Quiz for previous lecture

[Click here](#)

# Molecular Biology (2)

## The human genome

Prof. Mamoun Ahram

School of Medicine

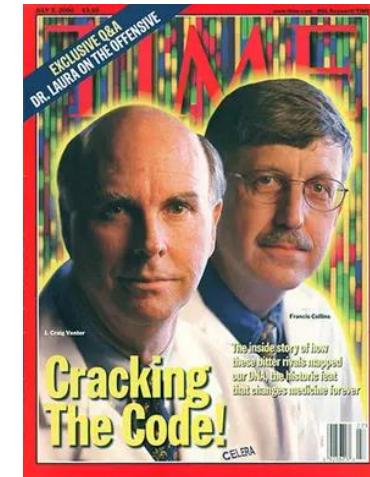
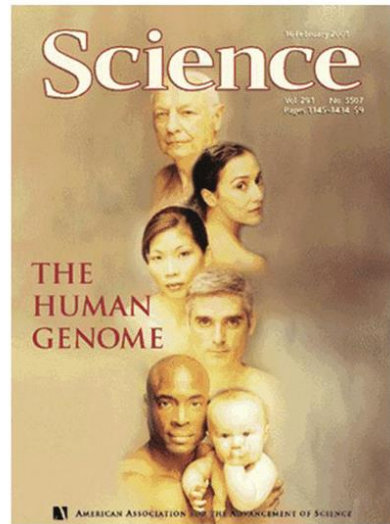
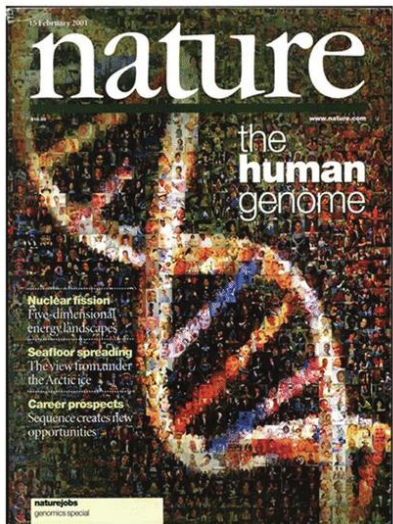
Second year, Second semester, 2024-2025

# The human genome project

→ Seven countries got together led by the United States to sequence the human genome & to know the order of nucleotides in the human DNA in the 24 chromosomes

- 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).

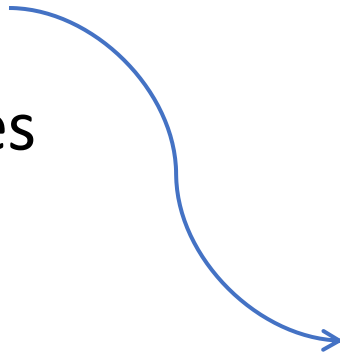
→ It was announced by Bill Clinton the president of the United States of America and the project was led mainly by Francis Collins and in the later stages there was competition by Greg venter who established a private company called Celera, they were able to also sequence the Human Genome using the preliminary information generated by Francis.



Nowadays we can sequence Human Genome in about \$100

# 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



Along with sequencing the human genome, scientists have been able to sequence the genomes of other organisms including advanced living beings like mice and chimpanzees and so on or even simpler ones like viruses & bacteria.



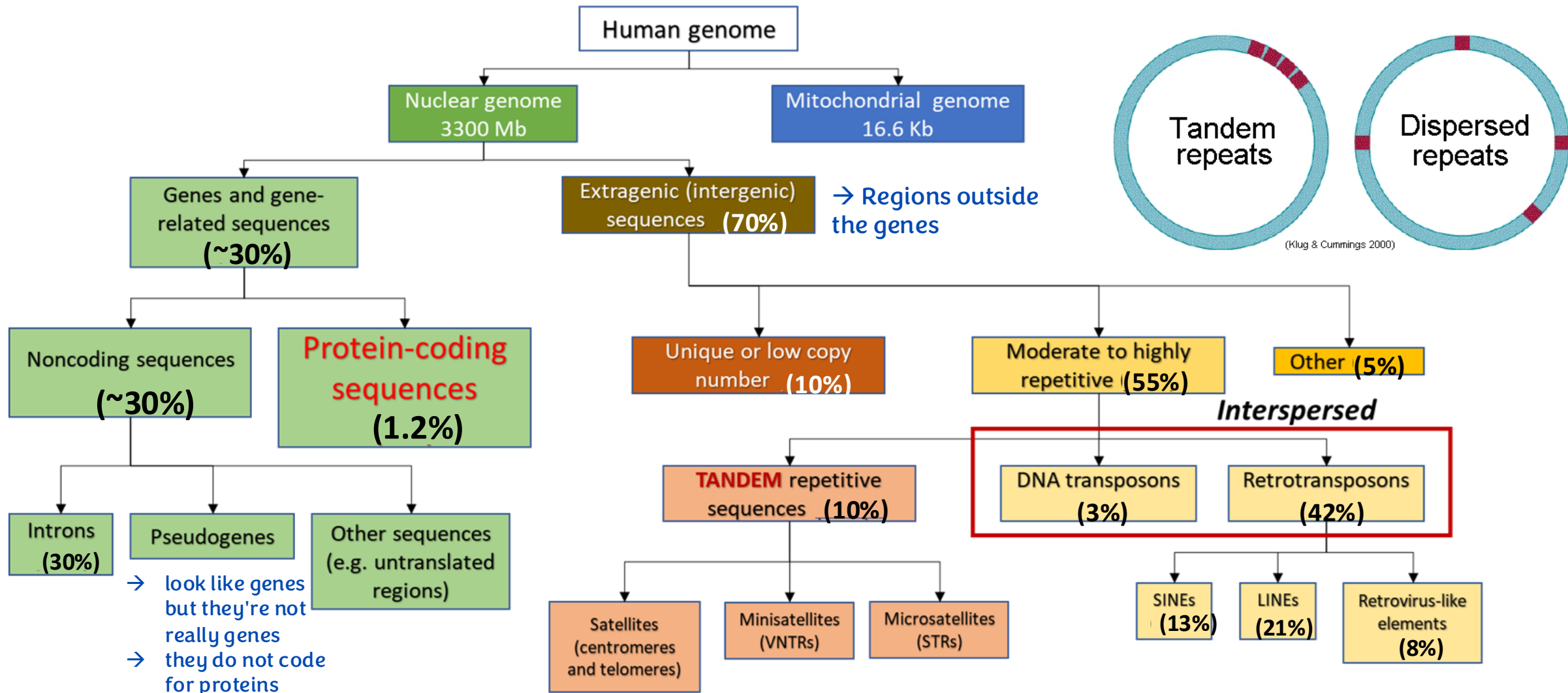
→ Comparative genomics is to look at & compare different genomes of different organisms

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

- Mouse has 40 chromosomes doublets and there are 20,000 protein coding genes; compare this to the human genome where there is 46 chromosomes doublets and there are about 20,000 genes also
- Some viral genomes are made of DNA, others are made of single stranded RNA and so on, there's a huge variety of genomes
- Mitochondrial chromosome genome is circular just like bacteria and it codes for 13 different proteins and these are needed the oxidative phosphorylation and the electron transport chain

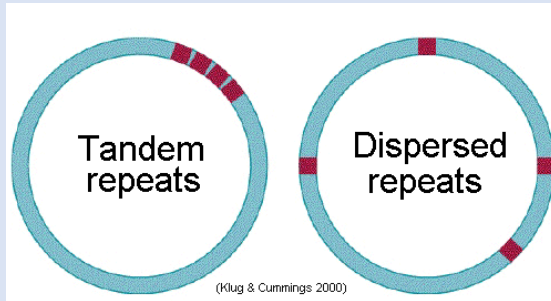
Similar number of genes in human and mouse genomes  
 -we all sort of have the same number of genes overall- we differ in very few genes, so what makes a human human and a mouse mouse? the noncoding parts regions of the genomes

# Components of the human genome

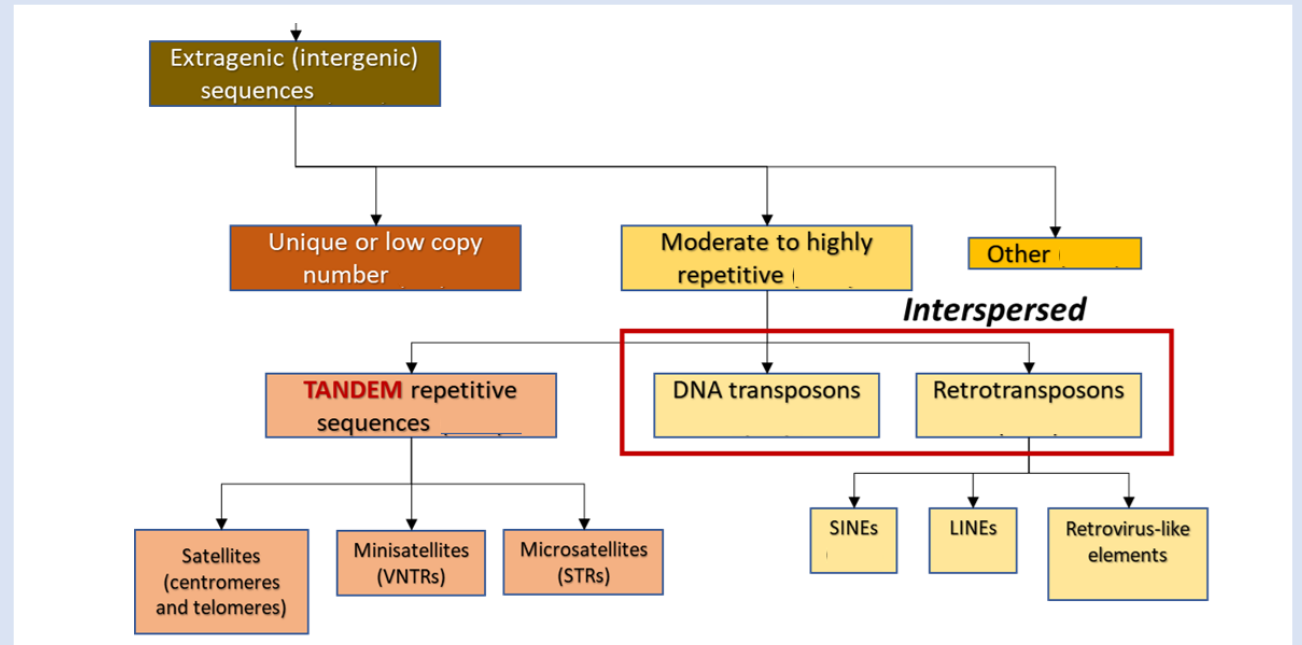


~5% of the genome contains sequences of noncoding DNA that are highly conserved (critical to survival).

**Note: All numbers are approximate**



- **Interspersed** means that they are dispersed repeats found in different regions of the human genome, distributed in the human genome all over the place in different chromosomes but are basically the same sequences
- **Tandem** repeats means that they come one after the other, so it's the same sequence but it's repeated one after the other. Tandem means that they are linked & associated with each other



→ Transposons can be short, some of them are long and others are viral based or viral-related sequence. Notice that they make up the majority of the human genome



# 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).

<b>Summary of ENCODE Results</b>	
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%

- 
- Near the end of the Human Genome Project in 2003, the US government started another project called The Encode Project.
  - They wanted to look at DNA and at the human genome more closely.
  - It was found that the percentage of the genome that is transcribed into RNA is 75% instead of the 2% that we were told about previously.
  - The protein coding genes –the relevant genes– 75% of the human genome can be transcribed into different RNA molecules.
  - The mission now is to know what these really do, their function, significance, are they mere noise or do they have a purpose?
  - There are other regions that are not transcribed but they are functional and can affect parts of the gene that are transcribed.
  - Many of the 75% of the human genome are actually RNA molecules that can be classified into two types: long and short.
-

# On March 31, 2022...

## 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.

→ It was hard finishing the 8% remaining part of the human genome because of the repeats, it took them 20 years to finish it and they were able to sequence the whole Human Genome except for the Y chromosome because it contains a lot of repetitive sequences.

*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

These 20,000 genes can make 86,000 different RNA messenger

# 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<sup>1,2,3</sup>. 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<sup>4,5</sup>. 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<sup>4</sup> and mapped available population variation, clinical variants and functional genomics data to produce a complete and comprehensive reference sequence for all 24 human chromosomes.



# Tandem repeats

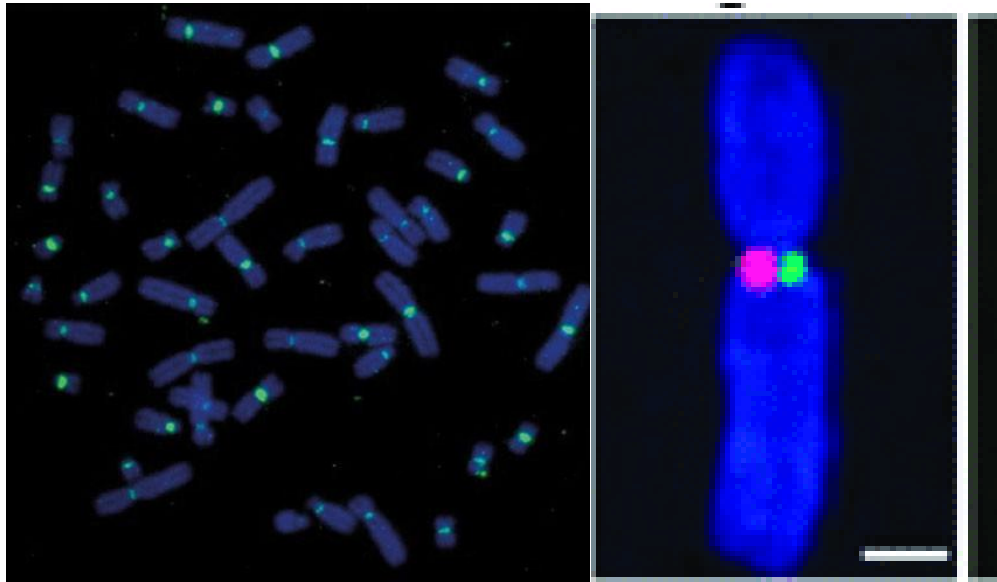
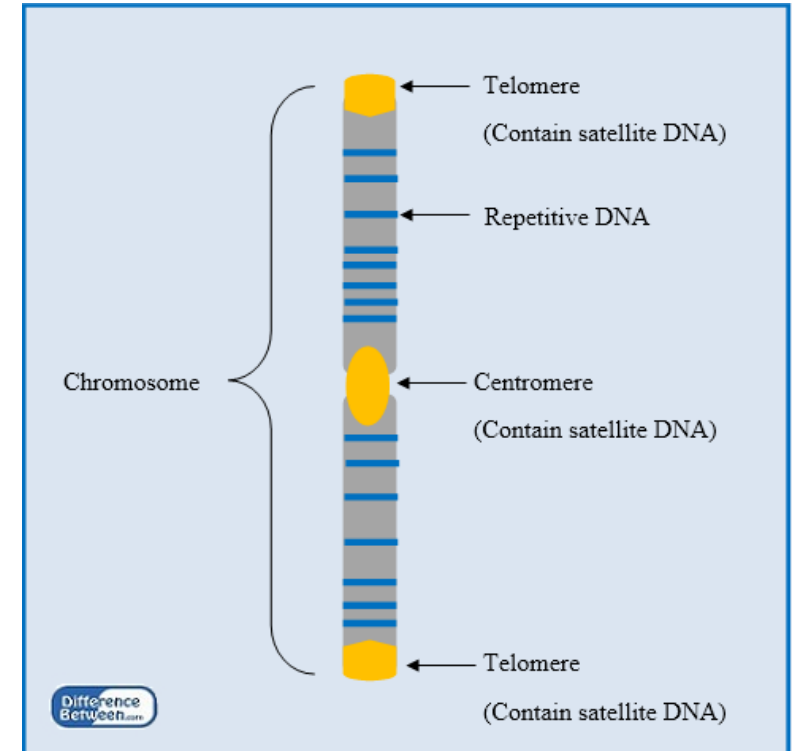
اللَّهُمَّ صَلِّ عَلَى مُحَمَّدٍ وَعَلَى آلِ مُحَمَّدٍ، كَمَا صَلَّيْتَ عَلَى إِبْرَاهِيمَ، وَعَلَى آلِ إِبْرَاهِيمَ، إِنَّكَ حَمِيدٌ مَجِيدٌ

# Satellite (macro-satellite) DNA

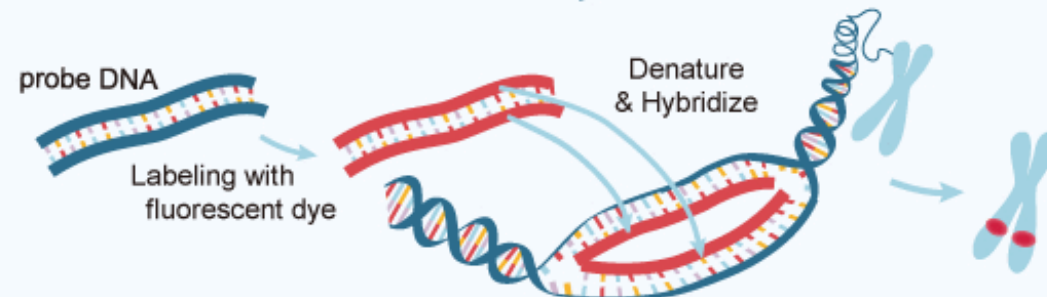
→ 5 to 300 base pair repeats repeated million to 10 million times

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

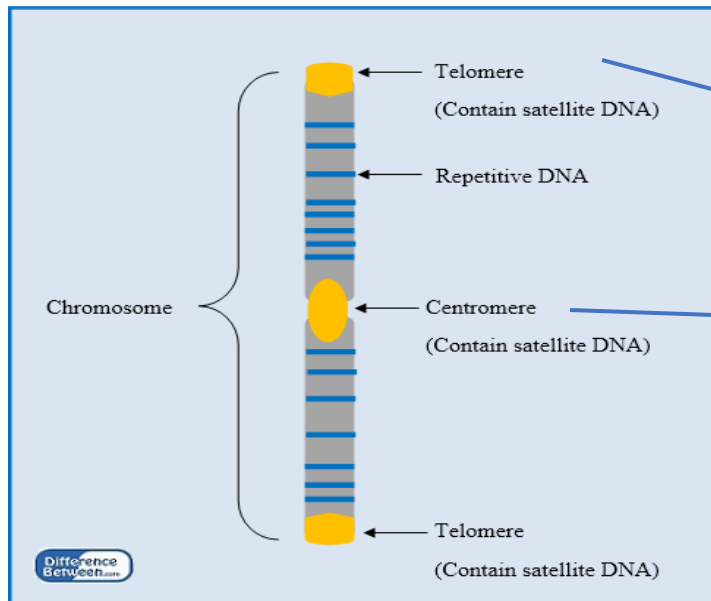
→ FISH is one of the most important techniques used to study chromosomes and chromosomal mutations



## Fluorescence In Situ Hybridization



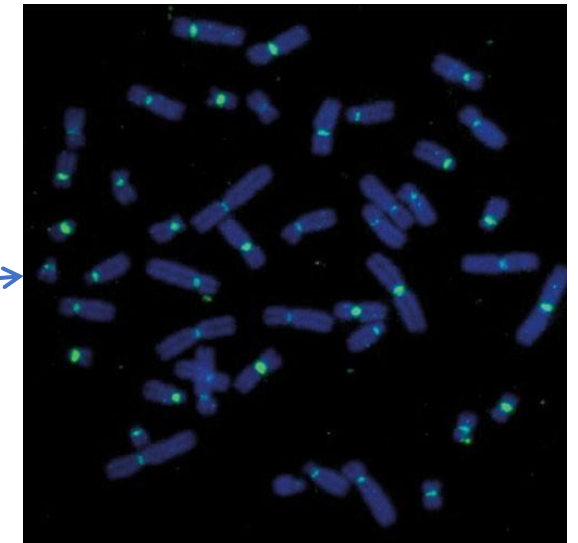




Telo means end; telomeres exist at the ends of chromosomes

- Centromeres are regions that exist in the middle of chromosomes -not necessarily in the center- but they are the constricted regions in the middle of chromosomes and they divide chromosomes into Q arm and P arm.
- P stands for petite -small & short- and the Q arm is the long arm.

- Using a technique known as fluorescence in situ hybridization, we can label different regions of the chromosome with probes and the probes are labeled with a fluorescent tag or radioactivity -mostly fluorescent tags- and the probes would bind to chromosomes at certain regions.
- You can label centromeres -as shown here- and you can label every single centromere since each chromosome has its own unique centromeric repeat we can use 24 different probes, each one is labeled with a different color so we can label each chromosome with a different color.
- We can also use FISH to label and target certain genes that exist on certain chromosomes like chromosome 5 or chromosome 10, etc. and to see where these regions or genes exist in what chromosomes.



# Fluorescence in situ hybridization (FISH)



Courtesy of Thomas Ried and Hased 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

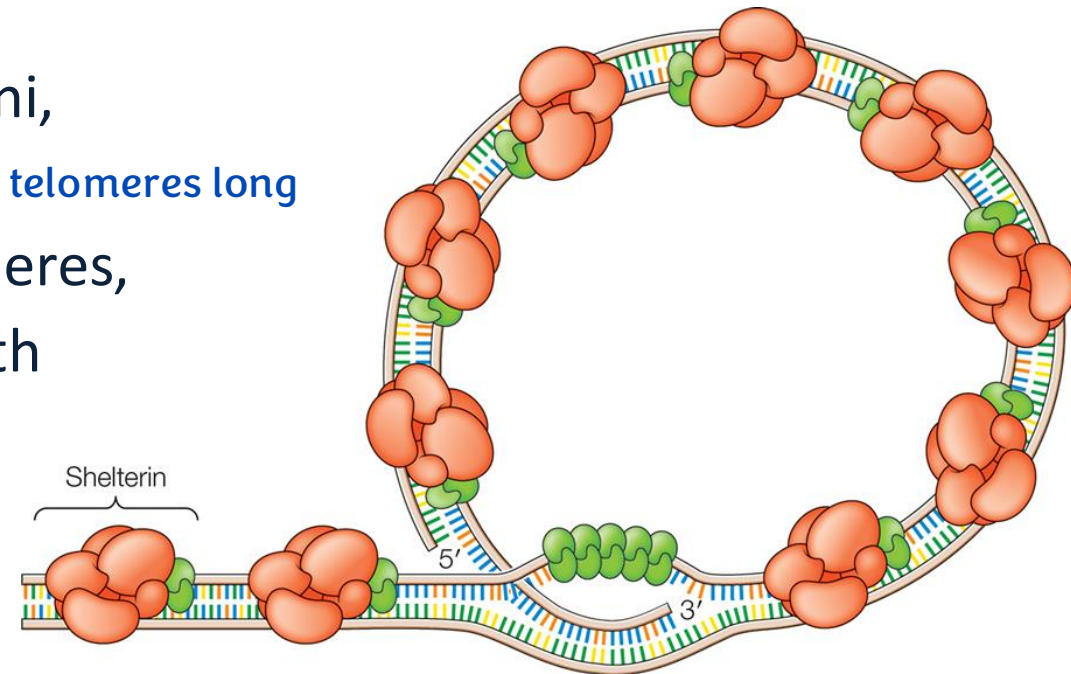
→ They exist at the end of chromosomes

- (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.

→ Telomeres are really important in stabilizing chromosomes & in keeping them intact

- **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, → This enzyme keeps telomeres long
  - maintaining the heterochromatic state of telomeres,
  - protecting DNA from deterioration or fusion with neighboring chromosomes

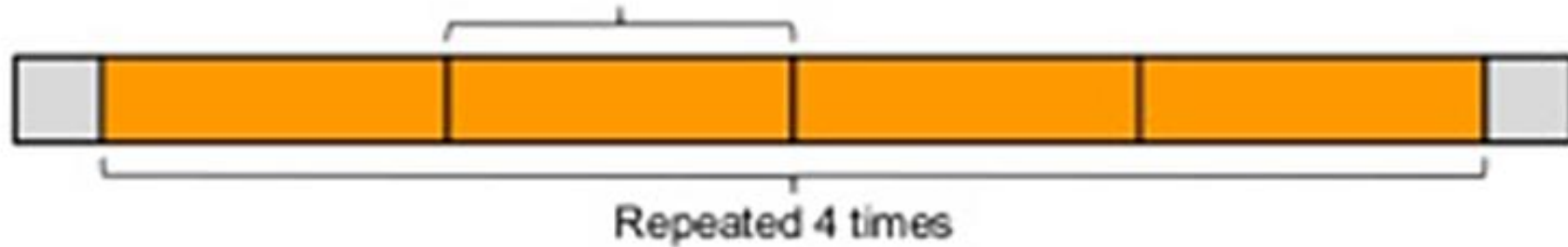
→ Remember, there are different types of RNA molecules that are transcribed but not translated, they're not used for making proteins rather they stay as single stranded RNA molecules and they can be classified into two parts either short or long.



# Mini- and Micro-satellite DNA

Minisatellite: Variable Number Tandem Repeats (VNTR)

→ Tandem because one comes after the other; repeats because basically it's the same sequence repeated like several times.

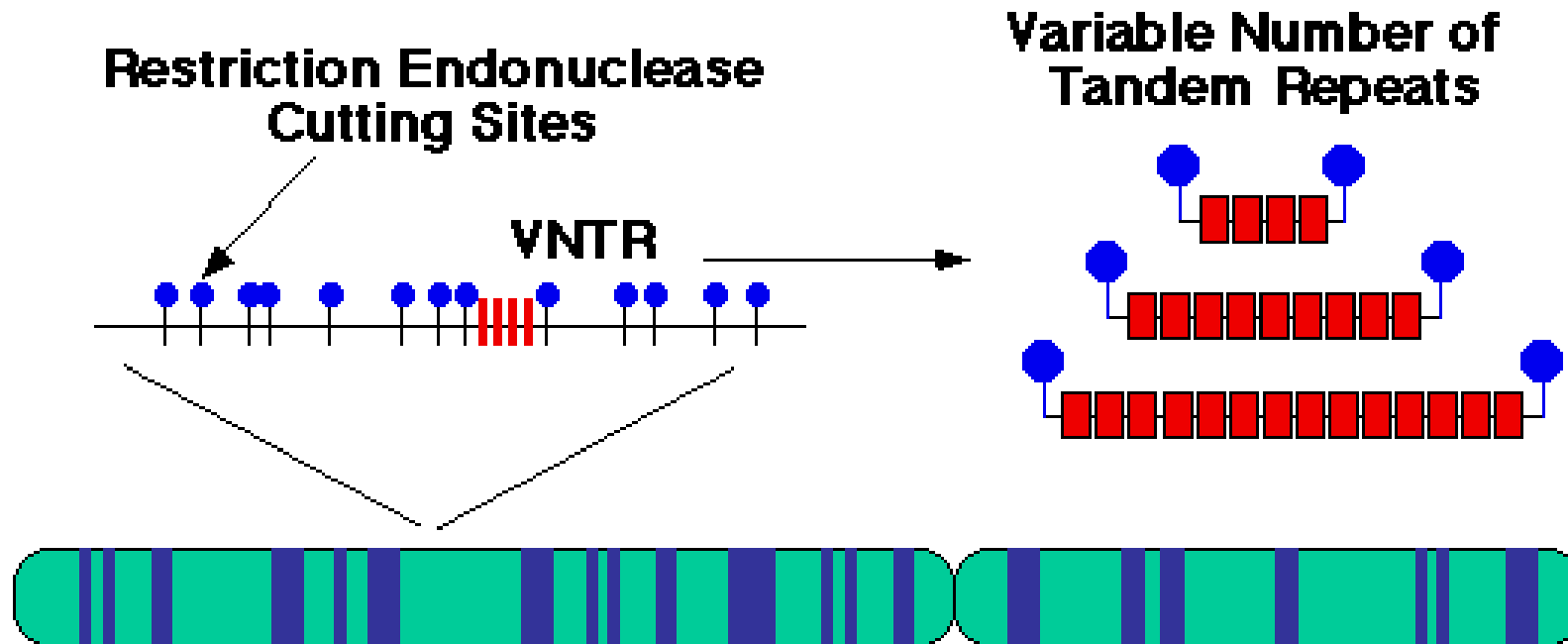


Microsatellite: Short Tandem Repeats (STR) – Simple Sequence Repeats (SSR)



# Mini-satellite DNA

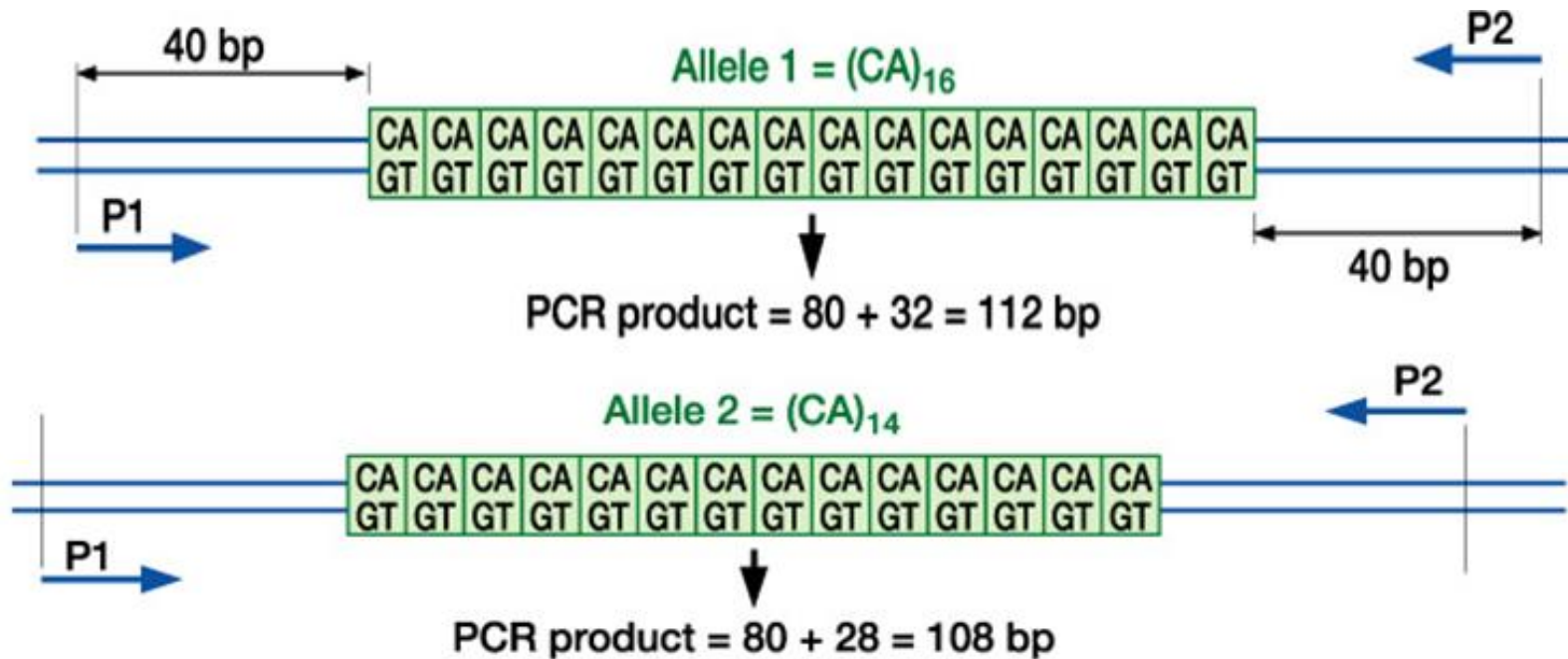
- Mini satellite sequences or VNTRs (variable number of tandem repeats) of 20 to 100 bp repeated 20-50 times



→ You can have VNTRs in different places and it can be the same VNTR or different sequences that are distributed throughout our genome

# Micro-satellite DNA

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

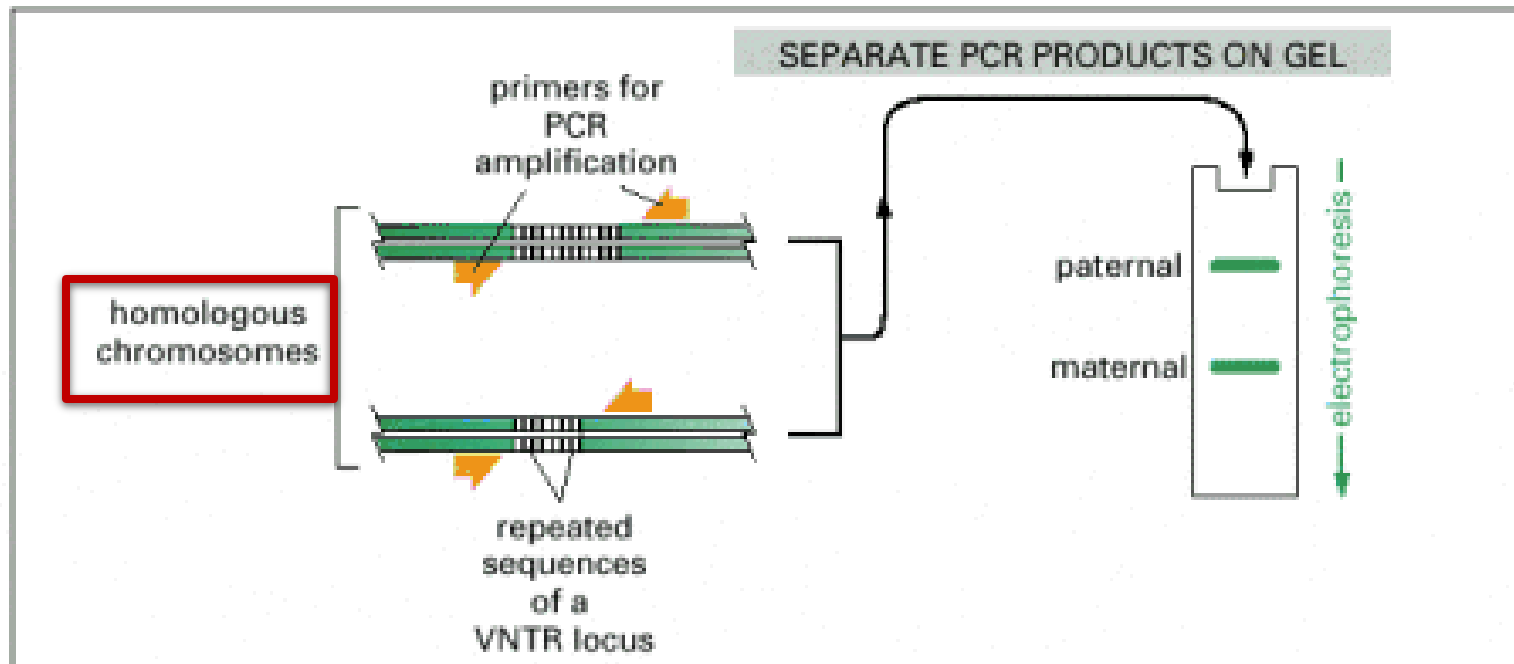




# Polymorphisms of VNTR and STR

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

we as individuals can have different repeats. On one chromosome the paternal chromosome we can have a repeat STR or VNTR with a certain number like 10, 20 times and on the maternal chromosome we can have the same repeat and the same sequence on the same chromosome we can have it repeated let's say 30 times.



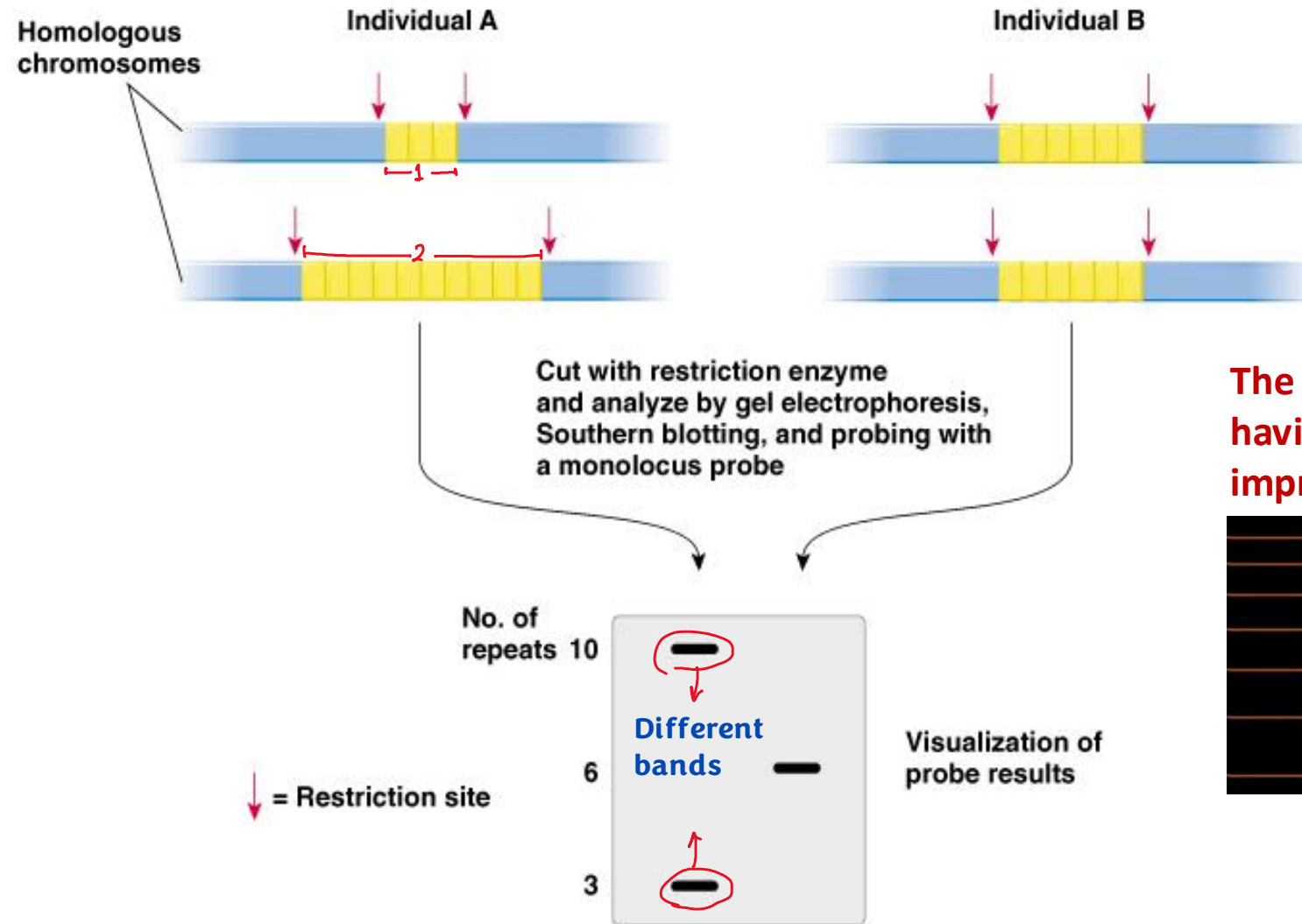
**Homologous chromosome (or homologs) are the set of one maternal and one paternal chromosome somatic diploid cells.**

**Homologous chromosome: the same chromosome except that we get it one from mother and one from father .**

we can have the repeats in different numbers on different chromosomes so the length and the size of DNA fragment that contains the DNA can be different and that is important because we can use it in forensic medicine , forensic testing , paternity testing and so on .

# STRs and VNTRs as DNA Markers

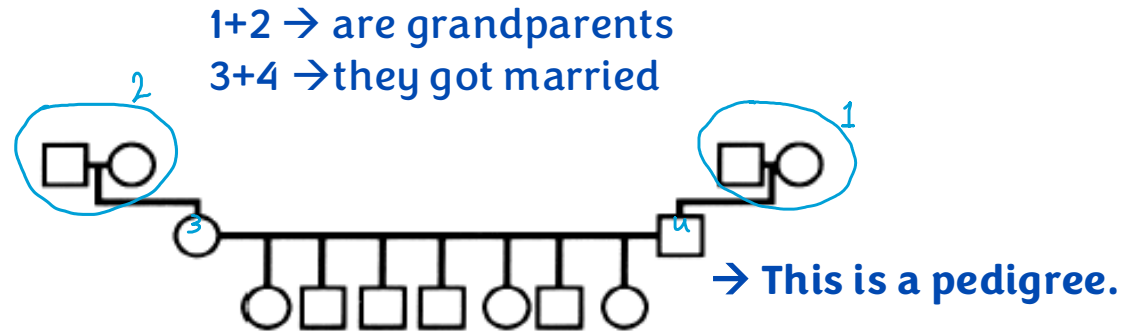
In **individual A** is heterozygous for a certain Str or untr so basically if we cut the region 1 and 2 we can have two different Bands . And in **individual B** a person can be homozygous for the untr or the Str



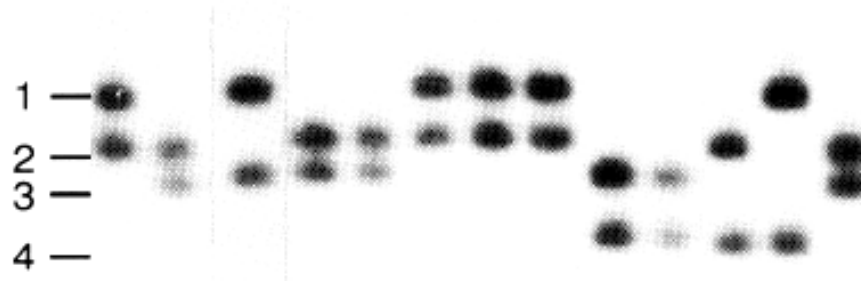
The likelihood of 2 unrelated individuals having same allelic pattern is extremely improbable.



# Real example



Large DNA fragment

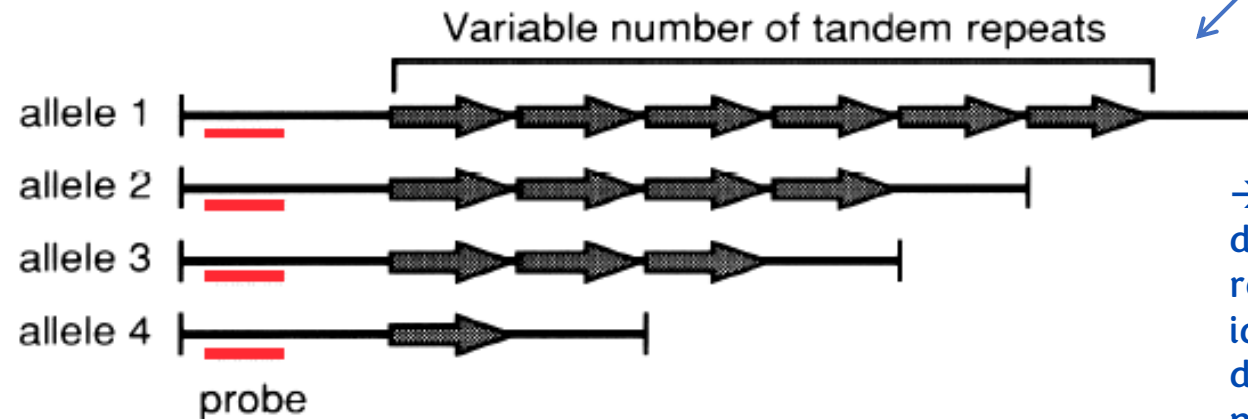


Small DNA fragment

single-locus probe but multiple alleles

We cut the DNA and get and form different fragments like these over here

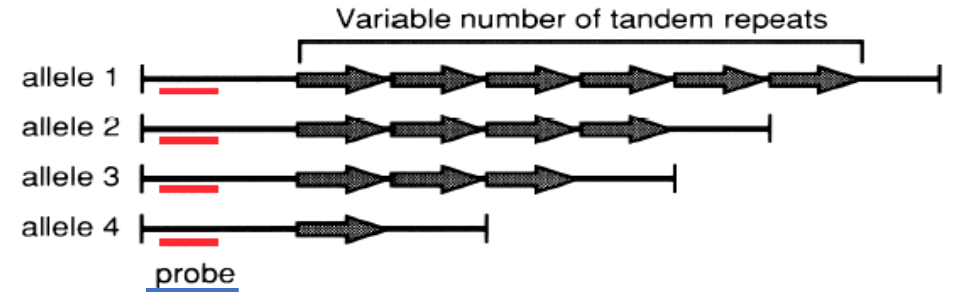
→ We separate DNA on a gel, then transfer the DNA from the gel to a membrane and add a probe.



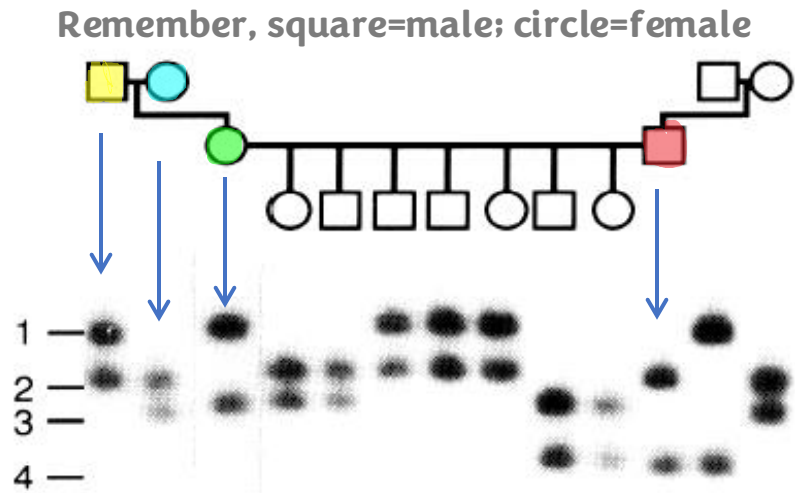
→ Variable length of fragments depending on number of repeats, other parts are identical in all fragments; main difference between them being number of repeats.

- Now, let's look at an example. Consider a pedigree of a family: male, female, grandparents, their daughter, and their son. The couple has seven children. Within this family, there is a specific VNTR present on for example chromosome 6. There are four alleles (Al=alleles)(Al1, Al2, Al3, Al4) with varying repeat numbers of VNTR : Al1 has six repeats, Al2 has four, Al3 has three, and Al4 has one repeat.
- If we use Southern blotting with a probe specific to this region ~the red line~ we can distinguish the alleles by their fragment lengths.
- We cannot use probe for VNTR because it would bind to all of them & we will not be able to distinguish different alleles, so we use probe that would bind outside VNTR region.

single-locus probe but multiple alleles



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



We are concerned with the size of fragments not necessarily intensity of the signal.

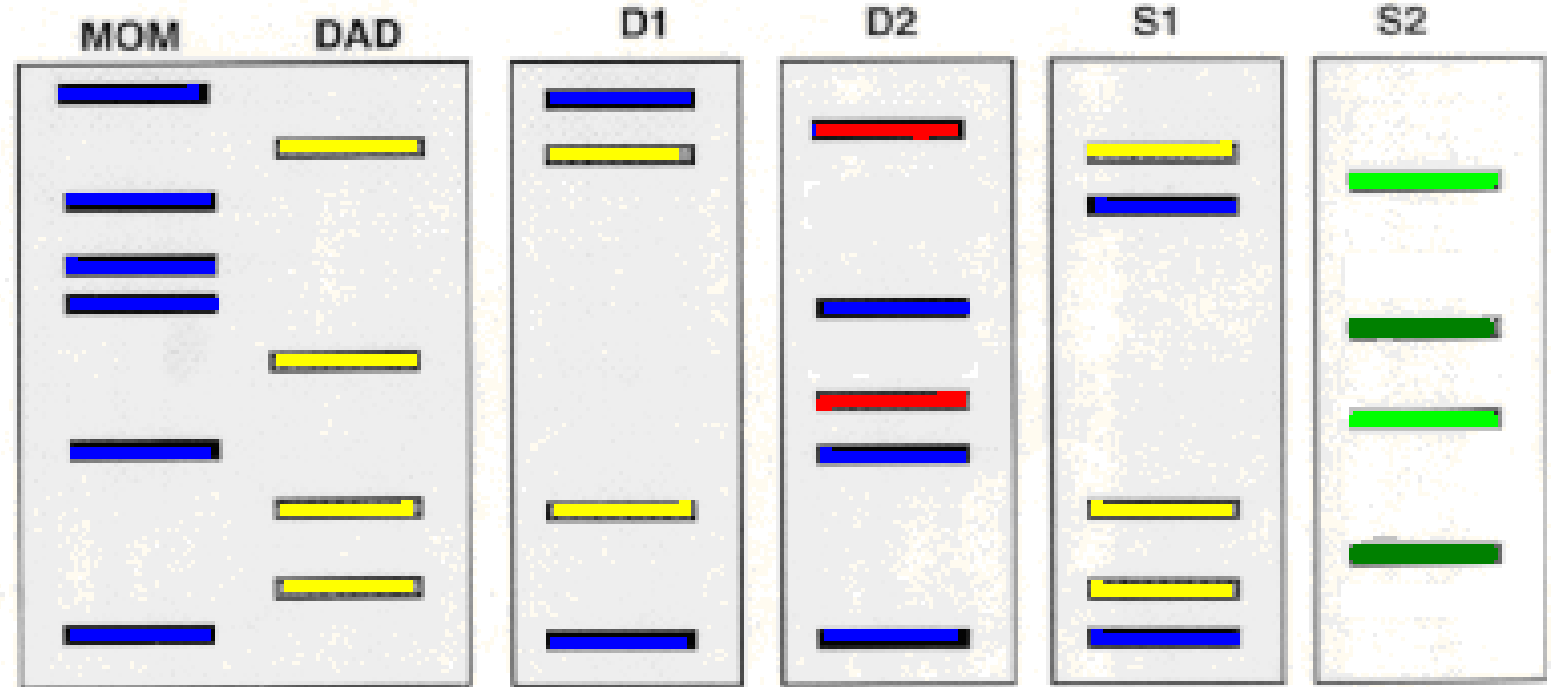
→ For example, the grandfather has Al1 on the chromosome and Al2 on the other chromosome, while the grandmother has Al2 and Al3. Their daughter inherits Al1 from her father and Al3 from her mother. The son of the other family inherits Al2 from his mother and Al4 from his father. When these two marry and have children, each child's genotype reflects a combination of alleles from both parents (all of the kids here are heterozygous).

→ We can see that each child has their own unique genotype, and while some may share a genotype -some genetic makeup being similar or identical in different individuals which is why we do several VNTRs to link a sample with a person- the majority are distinct.

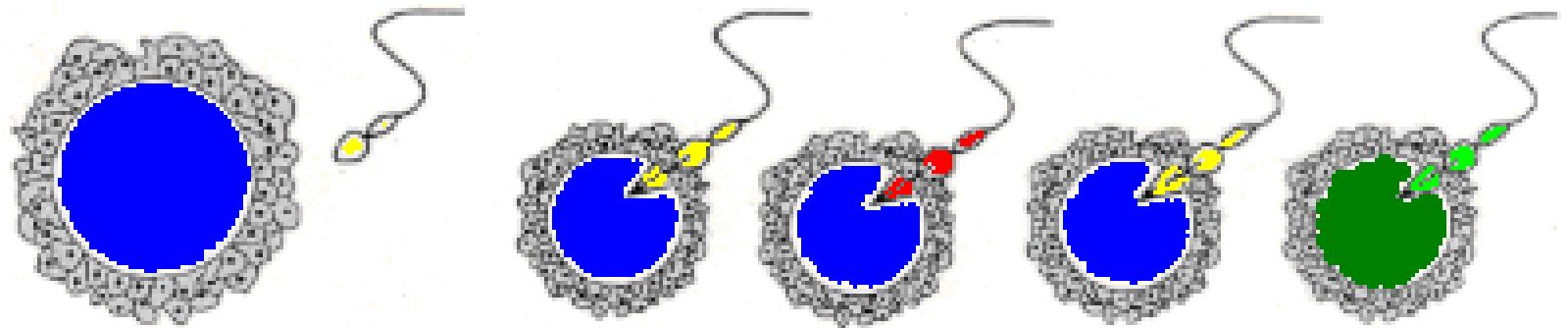
# Paternity testing

This is the genetic the molecular profile, also known as genetic profile, genetic fingerprinting & molecular fingerprinting.

- Each child has a fragment inherited from the mother and another from the father.
- D1 & S1 have fragments that can be linked to both parents genetic profile (actual children of both parents).
- D2 has fragments from the mother & no fragments from the dad.
- S2 is not the son of either parents.



If we observe a child whose genotype does not match either parent, it could indicate non-paternity (used in paternity testing).



# Single nucleotide polymorphism (SNPs)

- All of the previously mentioned tandem repeats are variables, different among individuals.
- Another source of genetic variation is single nucleotide polymorphisms (SNPs). SNPs are variations in a single nucleotide in the genome that occur at specific positions.

→ Polymorphism means different shapes for single nucleotides

- 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

→ This person is homozygous for a specific SNP (GG) & is heterozygous for another (AC).

Individual 1

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

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

Individual 2

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

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

Individual 3

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

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

Individual 4

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

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

Individual 5

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

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

Individual 6

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

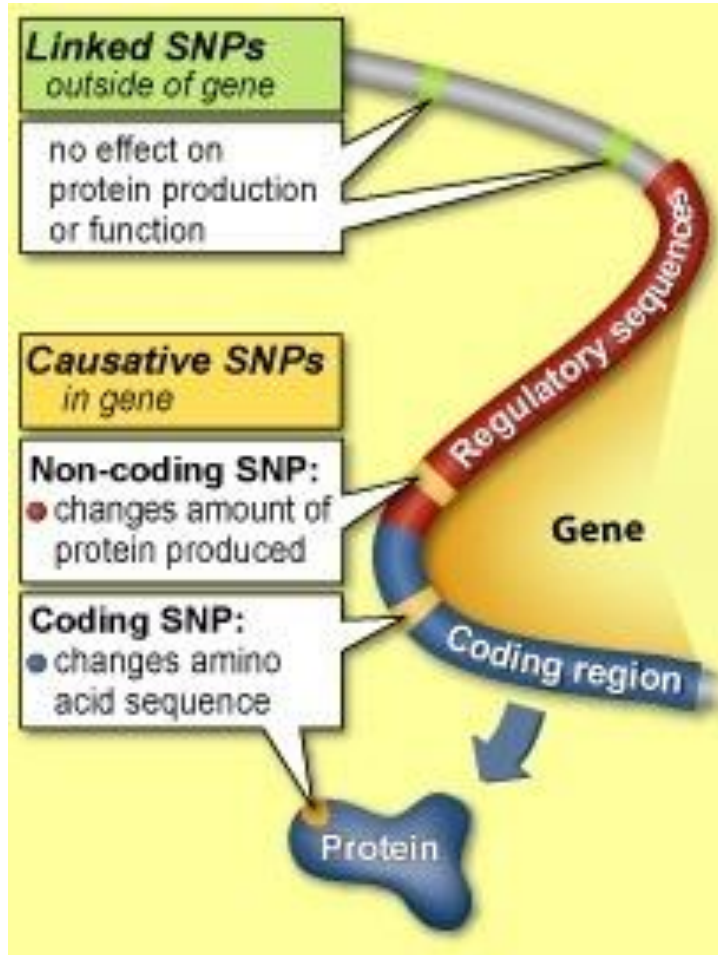
Chr 2 *copy2* ...CGATATTCC**T**ATCGAATGTC...  
 ...GCTATAAGG**A**TAGCTTACAG...

	Homozygous SNP		Heterozygous SNP	
Paternal allele	AACTGGACTT	<b>G</b>	AAGCATCTACGTT	<b>A</b> TCCATGAAG
Maternal allele	AACTGGACTT	<b>G</b>	AAGCATCTACGTT	<b>C</b> TCCATGAAG
Frequency in population:		G 51%	T 49% (minor allele)	A 90%
				C 10% (minor allele)

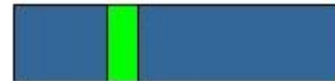
To be classified as a SNP, the variation must exist in more than 1% of the population; otherwise, it is considered a mutation (basically, a SNP is a type of mutation that exists in large proportion of a population).

SNPs can be homozygous or heterozygous, depending on whether the nucleotides at a specific position are identical or different on homologous chromosomes.

# Categories of SNPs

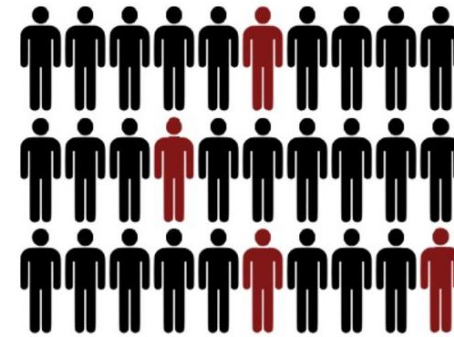


Linked SNPs



Cases

Causative SNPs



Controls

TTGGCCAGCTGGACCGAGGGGCGATGAC

TTGGCCAGCTGGATGAGGGGCGATGAC

- 
- The significance of SNPs depends on their location.
  - SNPs in coding regions can change the amino acid sequence of a protein, potentially affecting its function.
  - While those in regulatory regions can regulate gene expression of the coding region – doesn't code for protein but regulates– and may cause less protein coded from coding region for example, which would affect person's health.
  - SNPs in these regions are referred to as causative SNPs because they directly affect the phenotype → maybe less efficient or mutated.
  - However, most SNPs are linked SNPs that do not cause a phenotype –no functional significance– but are associated with certain traits or diseases that are linked but doesn't cause any phenotype. For example, person who has a disease has a G in this part of the genome, doesn't mean causation it just indicates linkage for some particular reason. G basically acts as a marker for higher probability of disease affecting an individual.
  - In one population, there would be variation between individuals based on different SNPs. For example, one person may require one pill of Panadol while another may require two, they all have the same enzyme & gene but the gene in one person may be expressed less efficiently.
-



# Interspersed repeats

اللَّهُمَّ صَلِّ عَلَى مُحَمَّدٍ وَعَلَى آلِ مُحَمَّدٍ، كَمَا صَلَّيْتَ عَلَى إِبْرَاهِيمَ، وَعَلَى آلِ إِبْرَاهِيمَ، إِنَّكَ حَمِيدٌ مَجِيدٌ

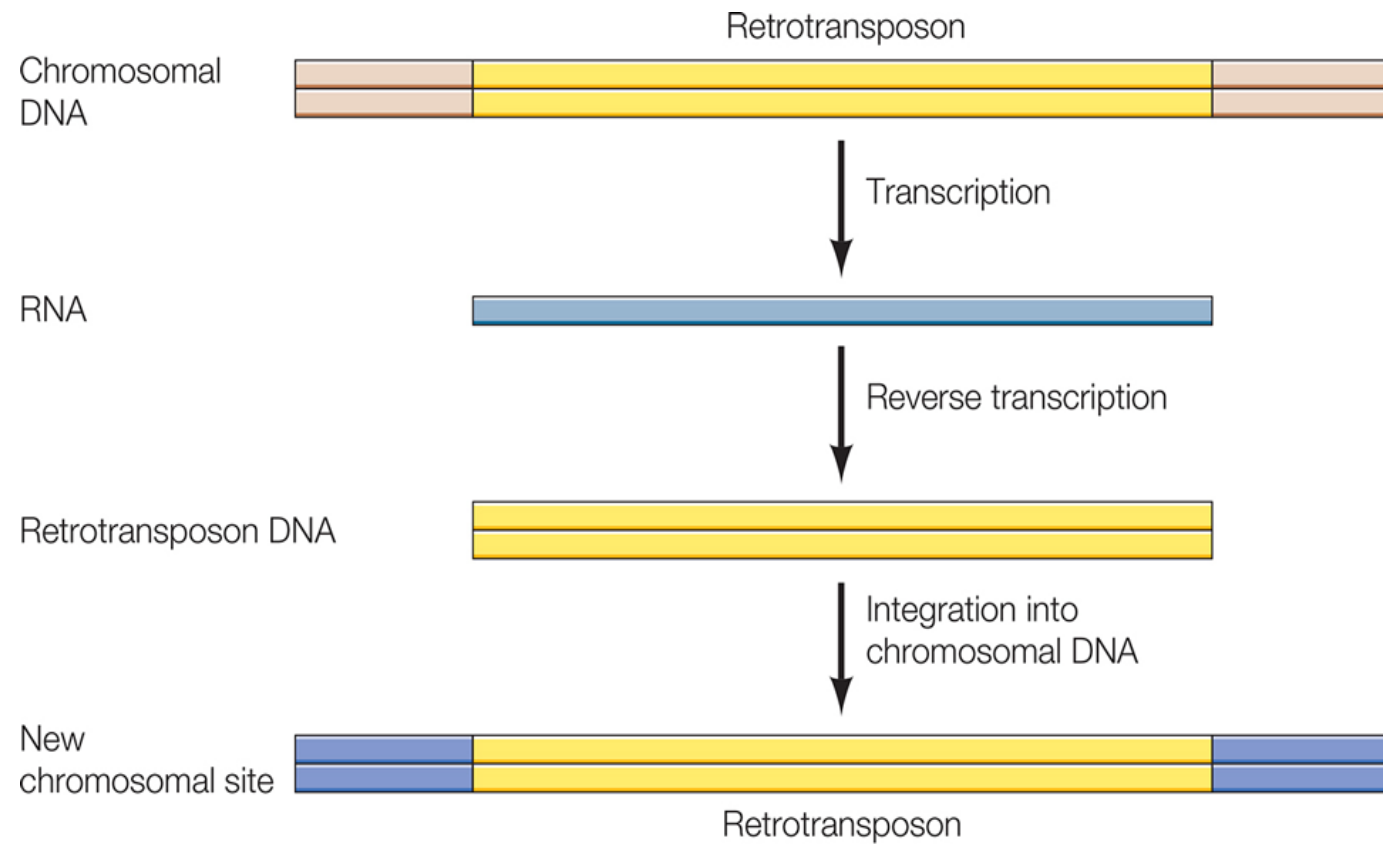
# Transposons (jumping genes)

→ Majority of our transposons have lost the ability to change places in our genome -can happen in some organisms like pigs-

- 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) → Alu exists all over human genome & can be integrated into protein coding genes.
    - Retrovirus-like elements (8%) → Genome is RNA not DNA, like HIV

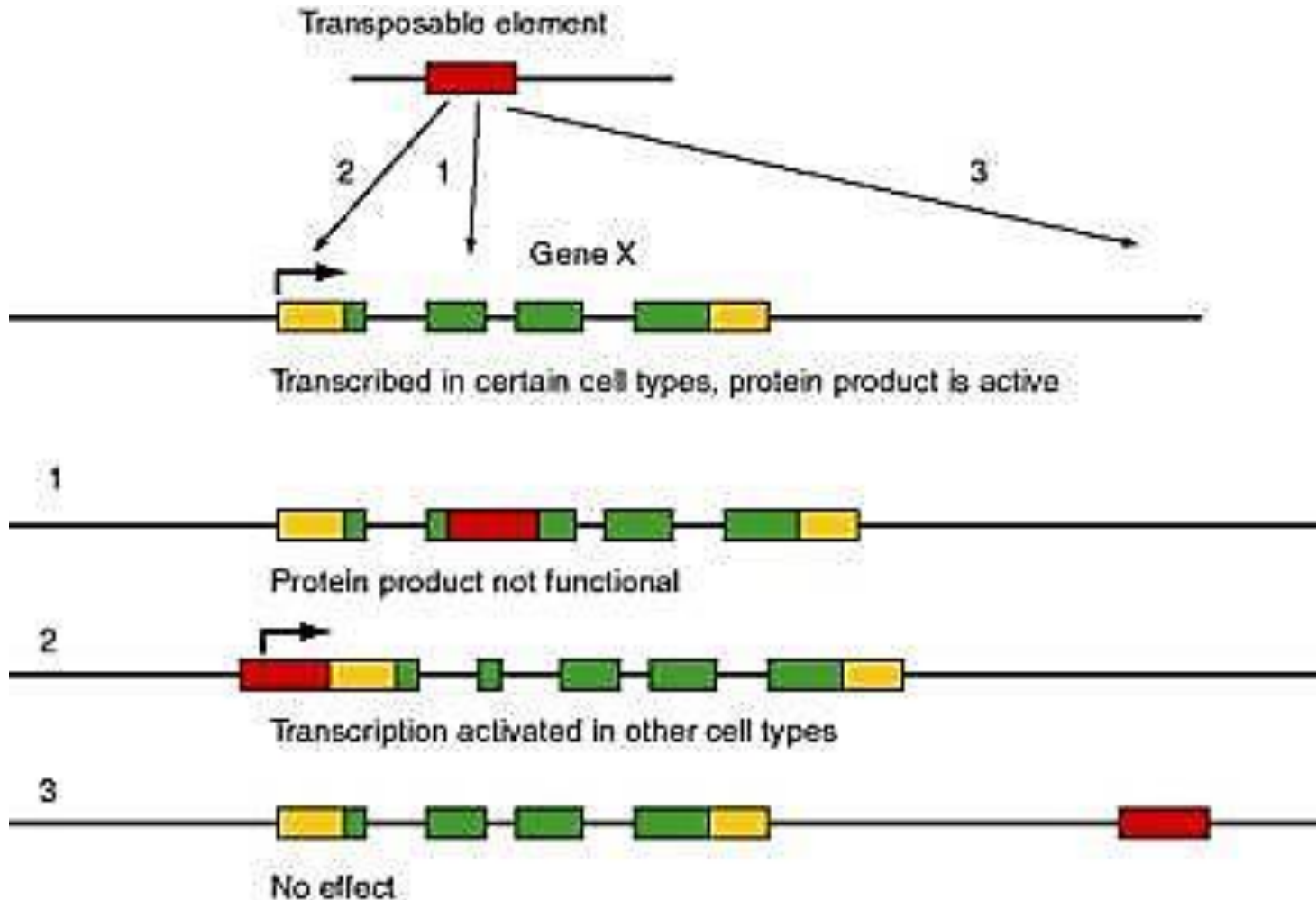
# How do retrotransposons move and integrate?

- A retrotransposon present at one site in chromosomal DNA is transcribed into RNA.
  - Reverse transcriptase is an enzyme coded by some retrotransposons
- 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.
  - Integrase can also be coded by certain retrotransposons
  - Majority of retrotransposons do not have reverse transcriptase & integrase gene, only some transposons can be transcribed & translated producing these enzymes that help in their movement.





# 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.

→ The significance of transposons changing their place depends on where they move & get integrated

→ They can integrate:

1. Within gene making protein defective
2. Or integrate within regulatory sequence, like promoter or enhancer region, and affect activity of that gene
3. Outside of genes & regulatory sequences with no effect

For any feedback, scan the code or click on it.



Corrections from previous versions:

Versions	Slide # and Place of Error	Before Correction	After Correction
V0 → V1			
V1 → V2			

Additional Resources:

رسالة من الفريق العلمي:

عَنْ أَنَسِ رَضِيَ اللَّهُ عَنْهُ عَنِ النَّبِيِّ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ قَالَ:  
«لَا يُؤْمِنُ أَحَدُكُمْ، حَتَّى يُحِبَّ لِأَخِيهِ مَا يُحِبُّ لِنَفْسِهِ».

[صحيح] - [متفق عليه] - [صحيح البخاري - 13]

دعواتكم للمسلمين المستضعفين بالفرج والنصر