6

Molecular Biology

Doctor 2019 | Medicine | JU

◉ Sheet ○ Slides

Done By

Ahmad Abuljadayel

Contributed in the Scientific Correction

Waheed Sharif

Contributed in the Grammatical Correction

Waheed Sharif

Doctor

Walhan Al-Shaer

Introductory information:

- **What is a genome?**

A genome is the genetic material of the organism which consists of DNA, and this DNA is divided into genes [coding region] and non coding DNA.

- The science that studies the genome is called **genomics.**
- **What is a gene?**

  It is the basic unit of inheritance and it consists of the DNA that is responsible for the expression of a functional protein..as we mentioned in the first lecture, the gene is the part of DNA that is transcribed to RNA which is then translated to a protein.

****Reminder:** Central dogma of molecular biology: Transcription of DNA to RNA then translation of that RNA to proteins with different functions..

**DNA→RNA→Protein

Now let's get started with our lecture..

To help us understand the genome, we have this table which is a comparison between different species in terms of total base pairs, genes, and chromosomes..

SPECIES	BASE PAIRS (estimated)	GENES (estimated)	CHROMOSOMES
Human (*Homo sapiens*)	3.2 billion	~ 25,000	46
Mouse (*Mus musculus*)	2.6 billion	~ 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

*Now before we understand this table, **what is a species?**

**It's a group of organisms that have the same main characteristics and they're able to interbreed among each other, producing fertile offspring .

Now, let's take a closer look at the table..

BASE PAIRS (estimated)	GENES (estimated)

*Now..as we can see here.. the total number of base pairs and genes in each species is **estimated**…but why? Why don't they have a fixed number of base pairs and genes?

Because they are a topic of major debate between scientists..For example, two decades ago, we [humans] were estimated to have **40000 genes, but now, scientists have estimated the number of genes in humans to be around **20000-25000** genes.

**VERY IMPORTANT:

*The main reason for studying this table is so we can see that there is no correlation between base pairs, genes and chromosomes..

For example: Look at the mouse..it has **2.6 billion** base pairs, **40** chromosomes and **25000** genes.

In humans: **3.2 billion** base pairs, **20000-25000** genes and **46** chromosomes.

Humans have many more base pairs than mice but the number of genes is almost the same, which shows that there is no correlation between base pairs and genes.

*Now..let's take a look at the next figure, which is a comparison between different organisms and the range for the number of base pairs per haploid genome.

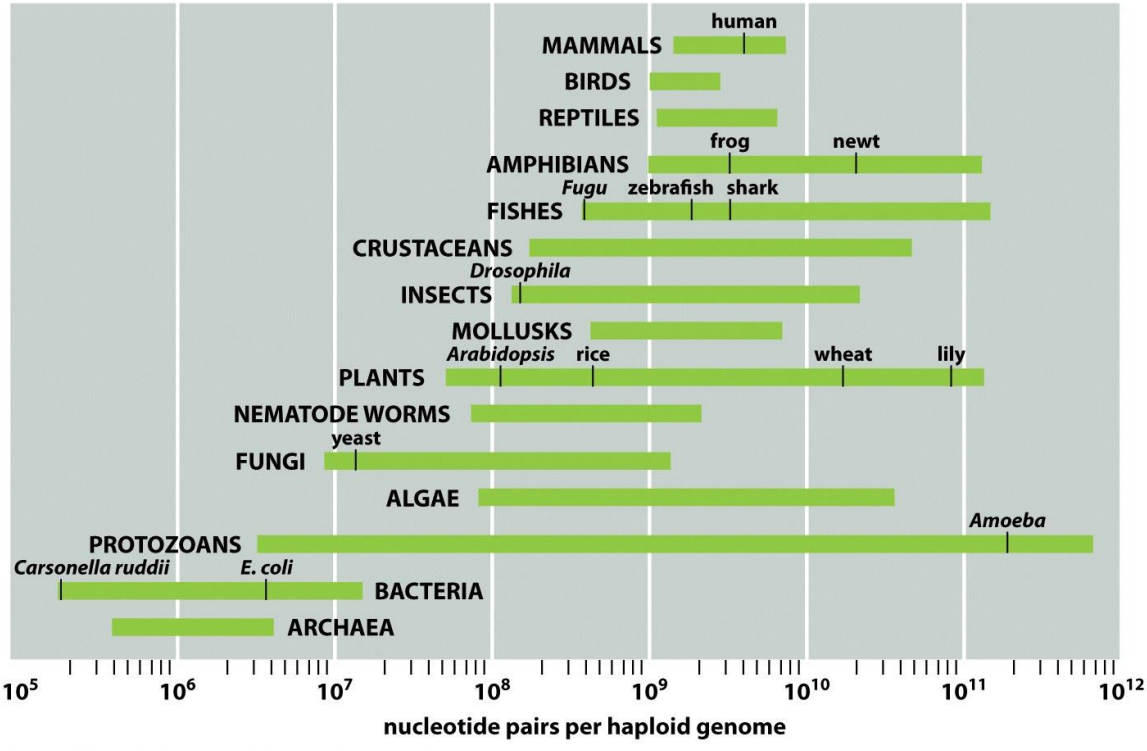


Figure 1-41 Essential Cell Biology 3/e (© Garland Science 2010)

****Reminder:** *Haploid: one set of chromosomes [n]..found in human germ cells

*In humans, somatic cells have 46 chromosomes, 23 from the mother [maternal] and 23 from the father [paternal], so two sets of chromosomes, we call it diploid [2n]..

so..

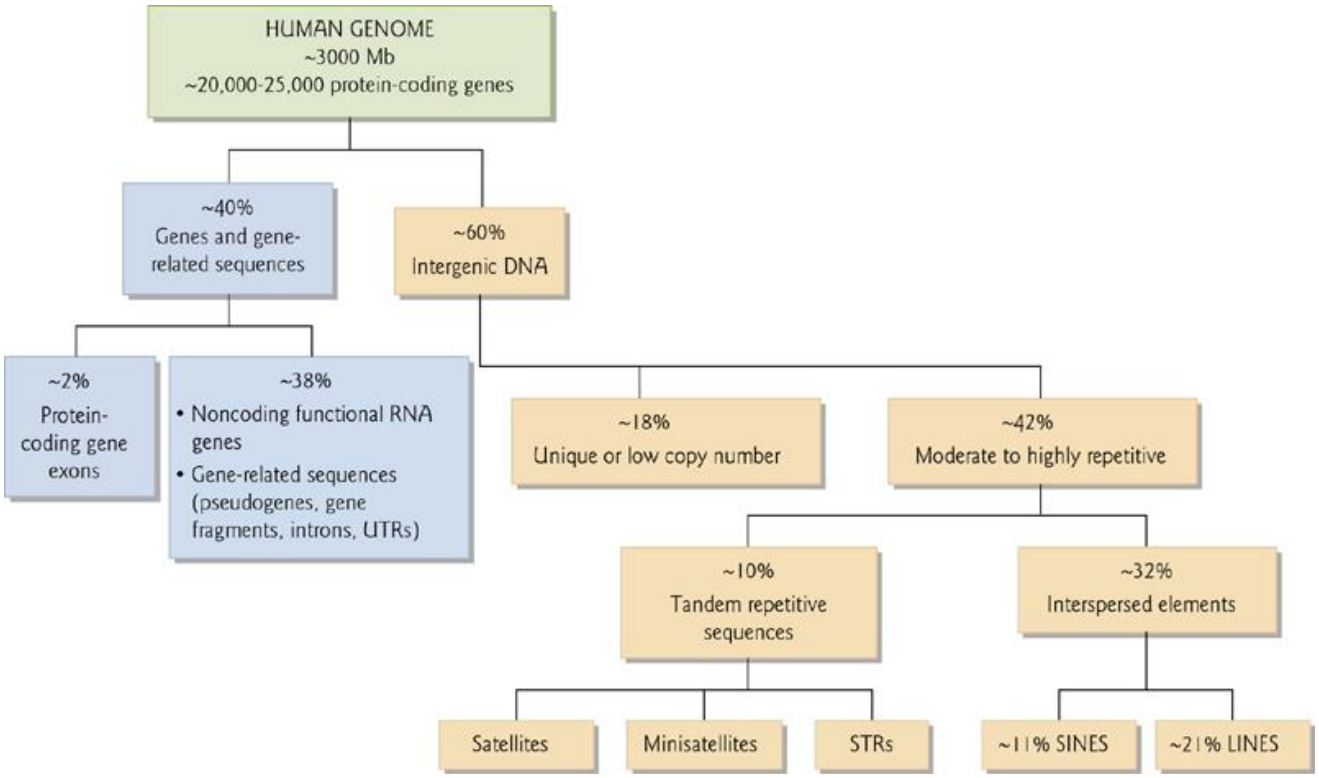
*Haploid genomes are found in human germ cells for example [the egg and sperm].

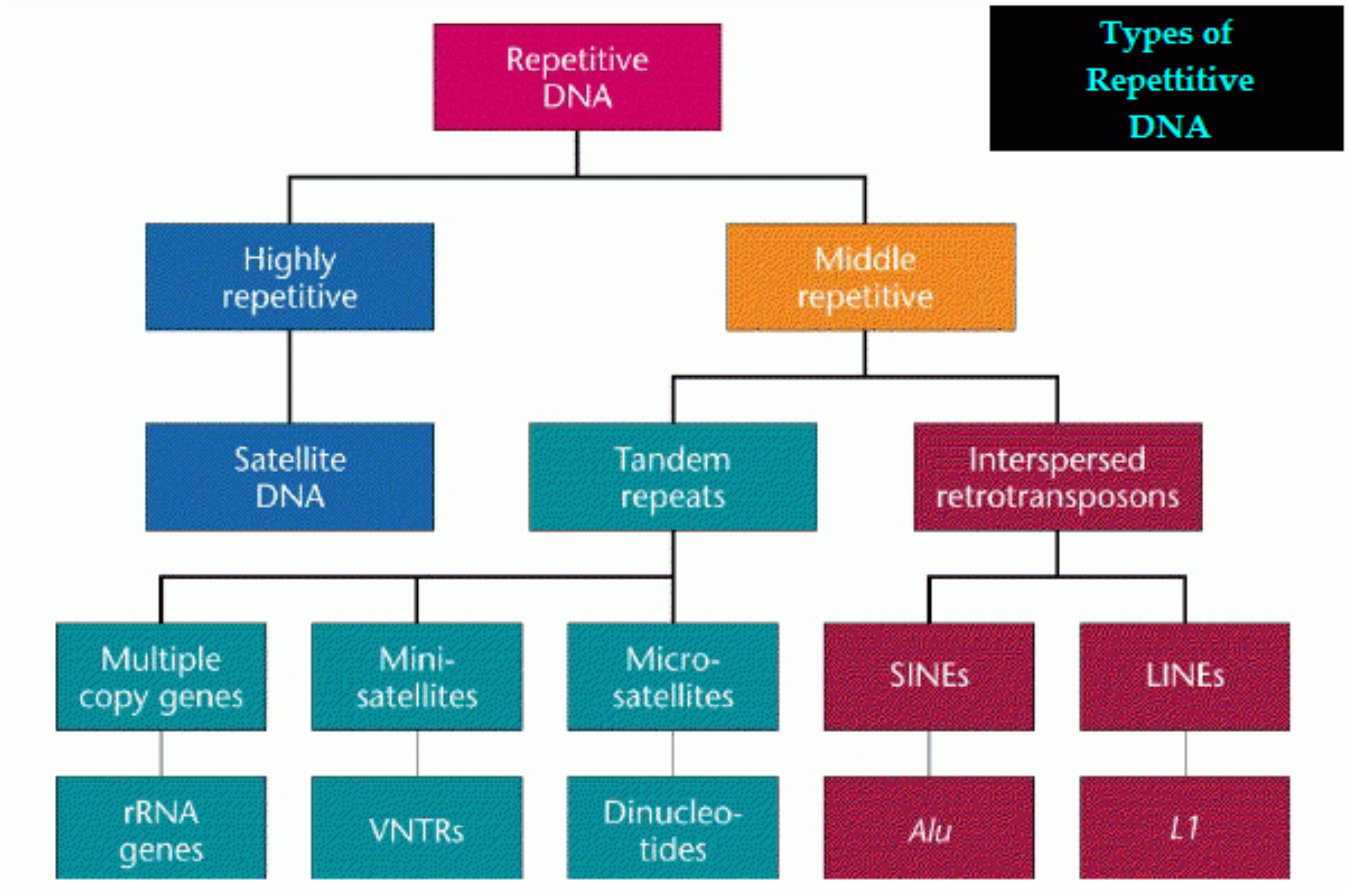
*Diploid genomes are found in somatic cells and they have **twice** the DNA content that haploid genomes have.

***As we can see in the figure:**

- Base pairs [nucleotides] per haploid genome range from **10^9 to 10^10** in mammals, birds and reptiles.
- Protozoa [single celled eukaryotes] which are usually parasites have a very wide range of nucleotide pairs per haploid genome [from about 10^6 to 10^12].
- In bacteria there is only one single chromosome so there is a lower number of base pairs.
- **There is no relationship between the number of nucleotides and the level of development..** Example: amphibians and plants sometimes have more nucleotide pairs in their genome than humans.

Now..study these two figures carefully and memorize them:





***Notes on the first figure from the slides:**

- Introns make **26%** of the human genome while regulatory sequences make **8-20%** of it.
- 80% of the entire human genome is relevant (either transcribed, binds to regulatory proteins, or is associated with some other biochemical activity).

*So we can see in the first figure that only 2% of the human genome is responsible for coding of proteins [through exons]..so what about the remaining 98%??

****In general:** The human genome has 40% genes and 60% intergenic DNA

*The human genome has 38% noncoding gene related sequences, such as:

A. **Long non coding RNA and small non coding RNA.**
B. **Pseudogenes:** Non-functional segments of DNA that are similar to the functional gene, so a pseudogene resembles the functional gene. However, it isn't functional because it's imperfect: Its coding sequences are defective due to mutations. Therefore it doesn't function like a gene [it doesn't code for proteins].

C. **Introns**: Nucleotide sequences in premature RNA that are spliced during its maturation to the final RNA product. [Try to keep this in mind as we will talk about it in the upcoming lectures].
D. **UTRs [Untranslated regions]:** They are usually found in the non coding areas of mRNA at the 5' and 3' ends of the first and last exons

Note: anything I mention in purple is simply extra information to help you understand thoroughly. It's **NOT something the doctor mentioned in the lecture.

EXTRA INFO FOR FUTURE UNDERSTANDING:

***What is the difference between introns and UTRs?**

UTRs are found in modified mRNA.

Introns are spliced out during mRNA processing.

The 5' UTR is before the start codon AUG.

The 3' UTR is after the stop codon [UGA, UAG, UAA].

- **What is intergenic DNA?**

It can be defined as a kind of non coding DNA found between coding DNA regions [genes].

*Now we can see from the figure that 42% of the human genome consists of moderately to highly repetitive intergenic DNA. They are explained better in the second figure.

***Interspersed elements [Interspersed retrotransposons] have two types:**

1. **SINES:** Short interspersed nuclear elements such as **Alu** which is believed to play a role in gene expression.
2. **LINES:** Long interspersed nuclear elements. Example: **L1** which is a transposable element.

***Note from the slides:** ~5% of the genome contains sequences of noncoding DNA that are highly conserved indicating that they are critical for survival.

**Scientists believe that those sequences of noncoding DNA are involved in evolution.

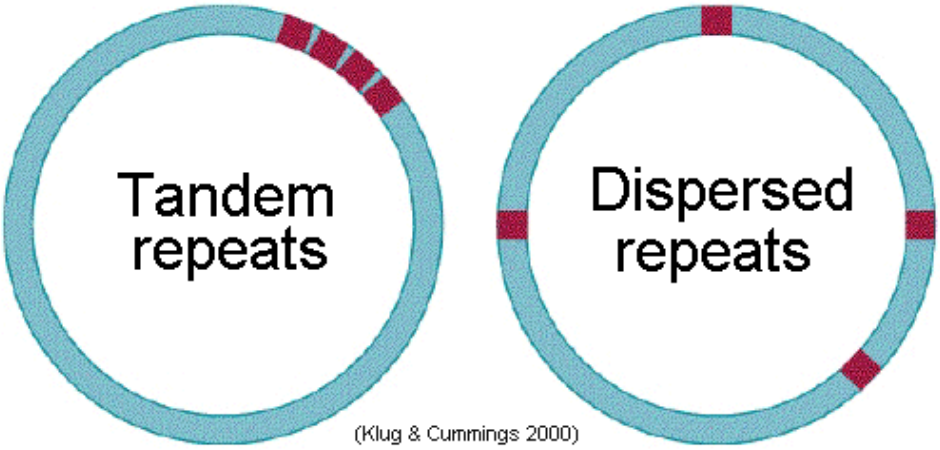
- **What is a repeat?**

It's basically a repeated sequence of nucleotides in DNA

***There are two types of repeats we need to know about: <u>Tandem repeats and dispersed repeats.</u>**

What is the difference between tandem repeats and dispersed repeats?!

*Let's take a look at this figure:

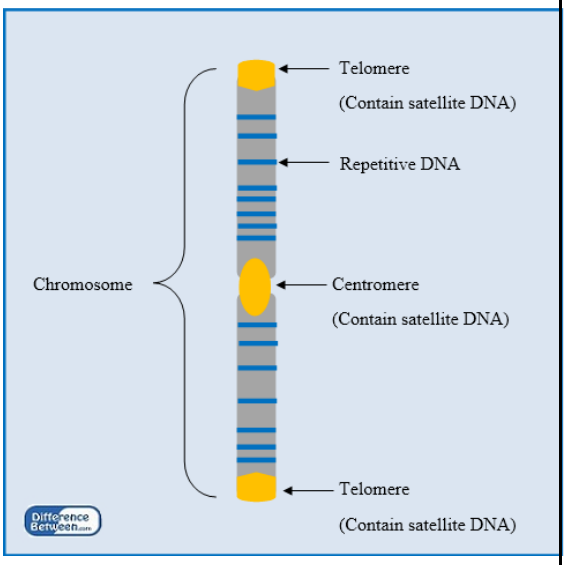


(Klug & Cummings 2000)

As we can see in the figure, **tandem repeats** are repetitive DNA sequences **<u>adjacent to each other</u>**, so these repeated DNA sequences can organize from end to end as a kind of array at specific locations in a chromosome.

<u>Dispersed repeats</u> are found as single elements **<u>dispersed</u>** around the genome, so the repeated DNA sequences are dispersed from each other, they're not adjacent to each other.

***Satellite DNA [also called macro-satellite DNA]:** Regions of 5-300 bp repeated 10^6-10^7 times (10% of genome). So, this type of intergenic DNA makes 10% of the genome.



Satellite DNA can be found in different places in the chromosome, like **<u>telomeres</u> and **<u>centromeres.</u>**

**A centromere is a specialized kind of DNA that links sister chromatids in the chromosome.

***So, in the centromere, we have centromeric repeats [171 bp long] which are unique to each chromosome [each chromosome has a unique centromeric repeat].

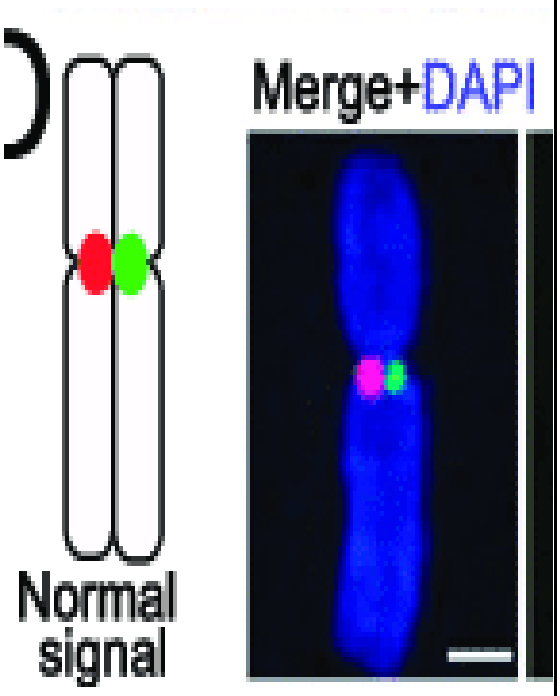
***How can we view centromeric repeats? By <u>fluorescent in situ hybridization</u> [FISH].**

*In this method, we use a fluorescent probe which binds to its complementary strand in the chromosome, then it produces fluorescence in the centromere.

* We have a fluorescent signal from the centromere as a result of the binding of the probe to its complementary strand.

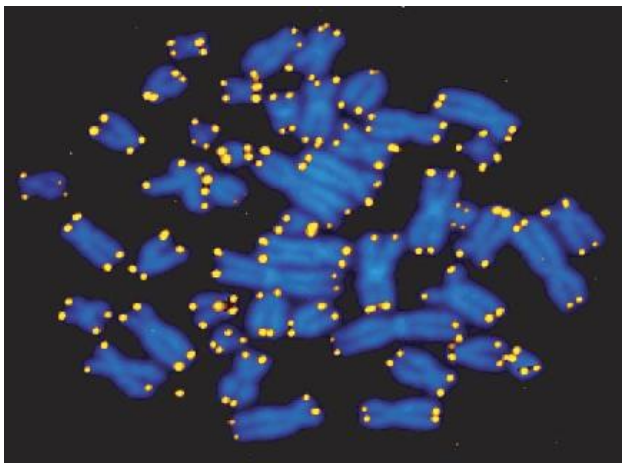
* So if we get this signal we know the chromosome is normal.

* If we don't get the signal then the chromosome has a disease.



*We also have <u>telomeric repeats..</u>

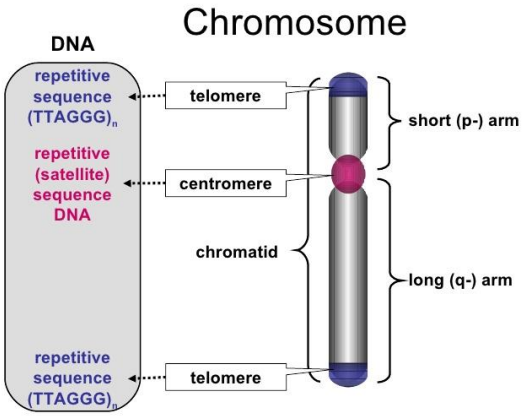
*Here we have once again used the FISH method to reveal the telomeres of the chromosomes. You can see them fluorescing in this photo due to the binding of fluorescent probes to their complementary telomeric DNA sequences at either ends of the chromosomes.



*Once again, here we can see a chromosome with its repetitive DNA sequences specifically at its telomeres and centromere.

*We can see that the chromosome has a short [p-] arm and a long [q-] arm

* The centromere links between the short arm and the long arm



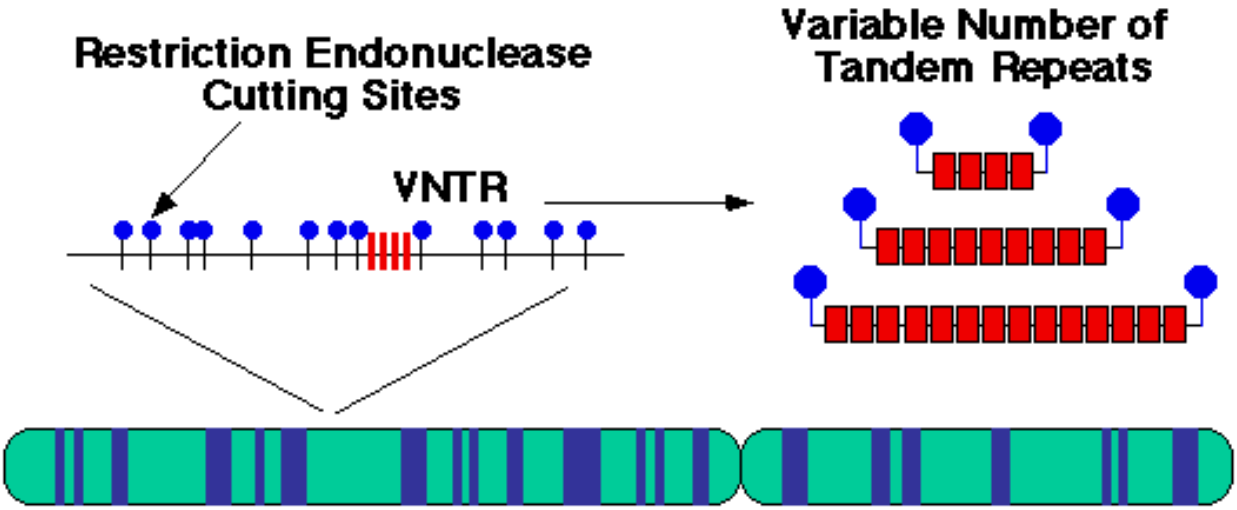
***Mini satellite DNA:**

*Another type of tandem repetitive DNA, with its sequences being called VNTRs [Variable number of tandem repeats].

*VNTRs consist of 20-100 bp repeated 20-50 times.

***Why are they considered VARIABLE?**

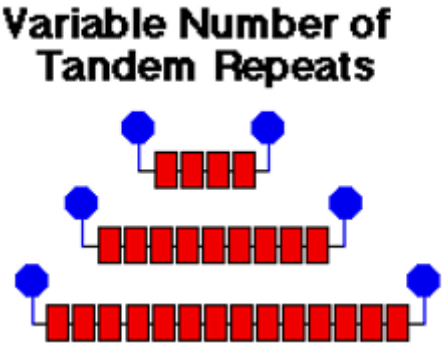
**Because there are different sizes of these repeats from site to site on the chromosome



*In this figure, we can see that when we use restriction endonuclease to cut at the restriction sites [marked in blue], we get different sizes of DNA sequences due to the different numbers of tandem repeats..[VNTRs are marked in **red**].

*To make it clearer:

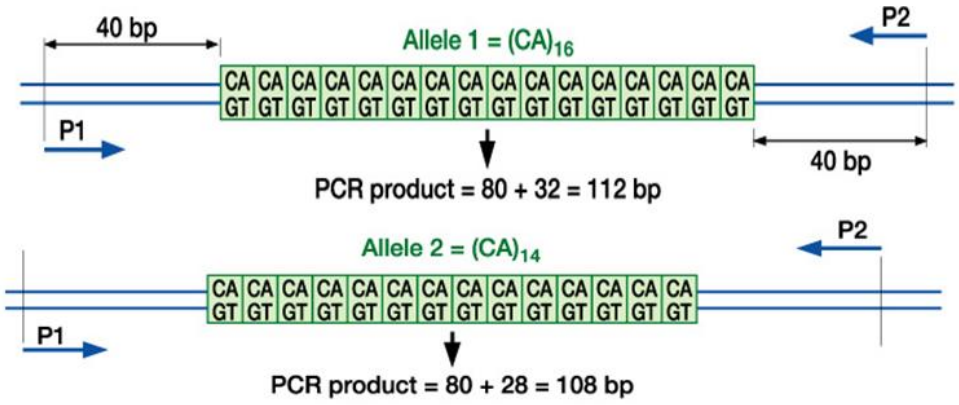
As we can see here, restriction endonucleases have cut the DNA at restriction sites. Now, the sizes of these fragments vary due to the variation in the number of tandem repeats. We have small, medium and large fragments. Each **red** rectangle marks a VNTR..



***Micro-satellite DNA:**

*In micro-satellite DNA, we have STRs (short tandem repeats), which are DNA sequences of **2 to 10** bp repeated **10-100** times.

*Take a look at this figure. We have two alleles. In each allele we have a DNA sequence that consists of two alleles from two different individuals.



*Each DNA sequence has **two primers** with each primer being 40 bp long, and between the two primers we have an **STR** with a 2 bp long DNA sequence [CA] repeated 16 times in allele 1 and 14 times in allele 2… So:

*In allele 1, we have two primers with each one being 40 bp long..40*2=80 bp. AND we have an STR [in this case CA] repeated 16 times, so 16*2=32 bp. Now we add them together and we find that there is a total of 112 bp.

*In allele 2: Two primers, 40 bp each..40*2=80 bp

STR repeated 14 times, 14*2=28 bp

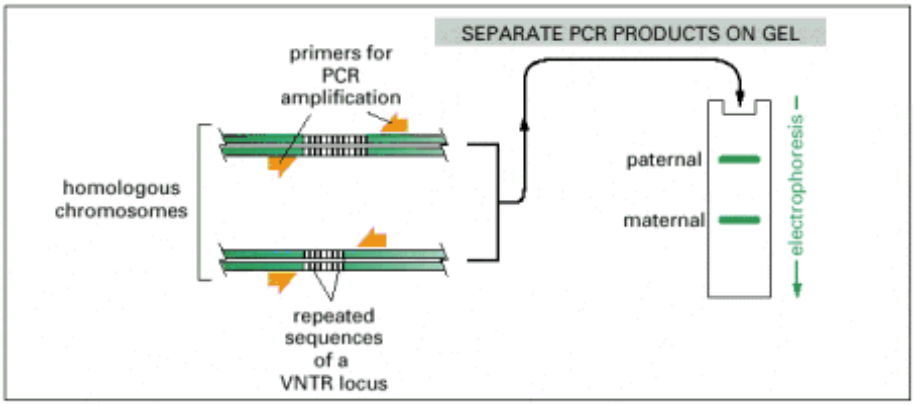
80+28=108 bp…

Now, you might be wondering: What's the benefit of all this? Well, we can use it to differentiate between individuals… By using **PCR [Polymerase Chain Reaction]** we can amplify these two alleles, and we will find that there is a 112 bp long DNA and a 108 bp long DNA… The benefit is that we will be able to distinguish different individuals by finding the STR length in the alleles of each individual as we will discuss shortly..

* STRs and VNTRs are highly variable among individuals **(polymorphic).**

* They are useful in DNA profiling for forensic testing [also for paternity tests].

*In this figure, we can see homologous chromosomes [one is paternal and the other is maternal], and in each one there are two primers. Between the primers, we have repeated DNA sequences [VNTRs in this case].

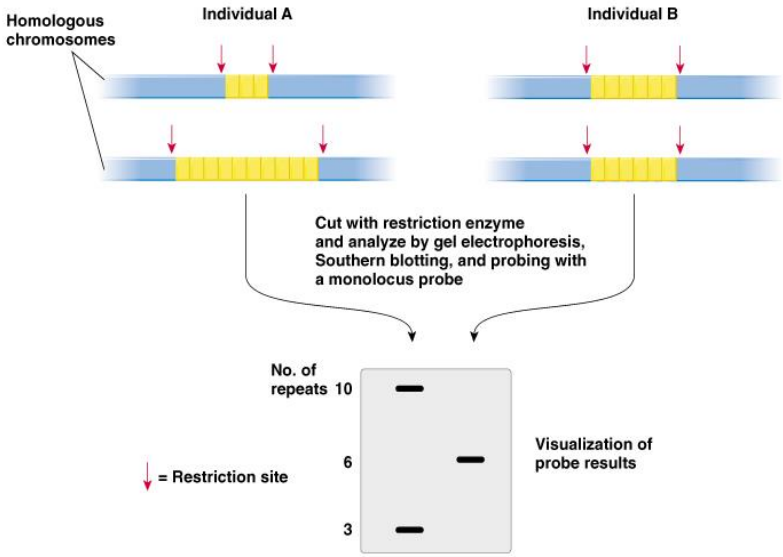


*By using PCR, we find that the DNA in one chromosome is longer than the DNA in the other chromosome due to the longer DNA having more VNTRs… That's why when we use gel electrophoresis, we get two different bands.

*Now that we know all of that information, let's utilize it!

** **How can we use microsatellites and VNTRs as DNA markers?** In other words, **how can we distinguish different individuals using microsatellites and VNTRs??**

*Concentrate on this figure...



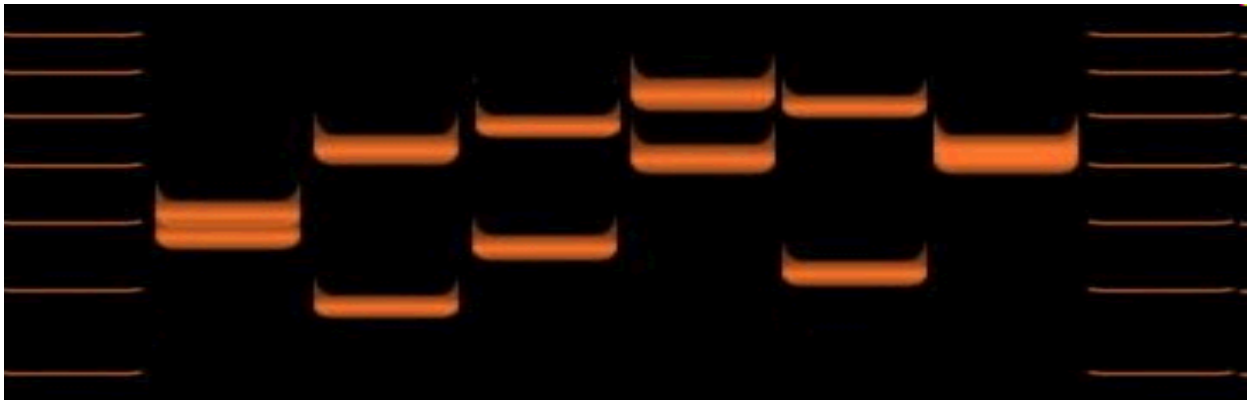
* We have homologous pairs of chromosomes from two individuals [Obviously, it's the same pair of chromosomes between the two individuals… So if we take pair number 16 from individual 1 for example, that means we will also take pair number 16 from individual 2], so we use restriction endonucleases which cut at restriction sites [here the ends of tandem repeats]..

Then, by using the southern blotting and gel electrophoresis techniques, we can see that individual A has a chromosome with 10 repeats and its homologous chromosome has 3 repeats.

Individual B has a chromosome with 6 repeats and its homologous chromosome has 6 repeats as well, that's why they only appear as one band in the agarose gel, they have the same length.

So, we could easily tell the two individuals apart after cutting the DNA with restriction endonucleases and binding of fluorescent probes to the DNA repeats [VNTR or STR].

**Now take a look at this figure, which shows the VNTR allelic length variation among 6 individuals:



**Here we have used the southern blotting and gel electrophoresis techniques after cutting the DNA at restriction sites to expose the VNTRs. As you can see, the bands vary in their positions along the gel, which means that the number of repeats is different for each individual. Also, as you can see, in each column we have two bands, indicating the length of the DNA is different between the homologous chromosomes of the same individual.

**** VNTR can be used to identify and differentiate individuals because it's almost impossible to have two different individuals with the same allelic pattern [same allelic length variation].**

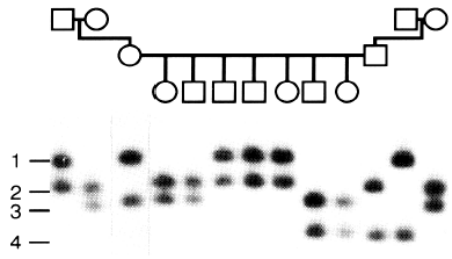
****Reminder:** We can use this for crime scenes and paternity tests.

Take a deep breath..

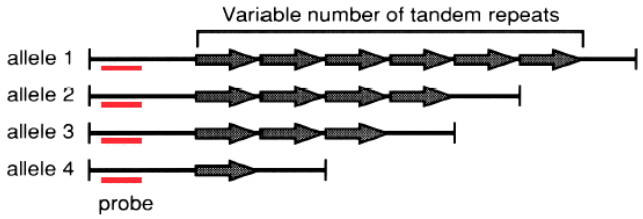
we're almost done.

*This is an example of the usage of VNTR in paternity tests.

*Here we have a family tree [pedigree] with each member's alleles below them.



single-locus probe but multiple alleles



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

*For example: The grandparents on the left: The grandfather has allele 1 and allele 2, the grandmother has allele 2 and allele 3..

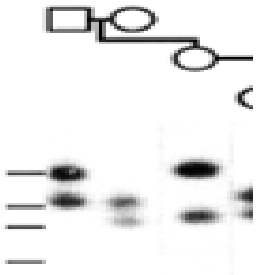
****Reminder:** Each square indicates a male while each circle indicates a female.

***But hold on…** We haven't mentioned what the alleles consist of…

*Each allele has a specific number of VNTRs… Once again, we can detect that by southern blotting and gel electrophoresis… Binding of a fluorescent probe to the DNA and so on…

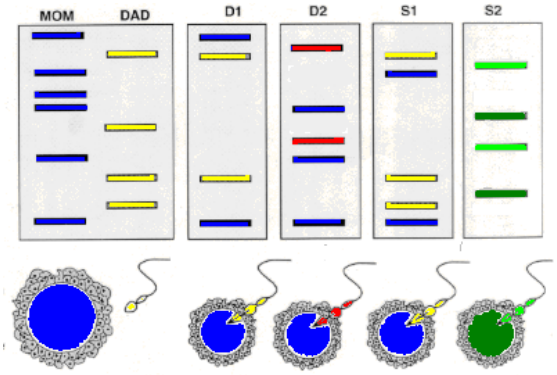
*Now, the point of this chart is to see that we can match every child to their parents…

*Take a look here for example… We can see that the daughter has alleles 1 and 3, she got allele 1 from her father and allele 3 from her mother… And once again, the alleles have a different number of tandem repeats, and as you can see, the VNTRs have benefitted us.. We can link generations through analysis of VNTRs…



Here we have another paternity test, where we can use the VNTRs to find the size of each DNA molecule, and by using gel electrophoresis, we get the DNA bands.

Obviously, the children have to inherit some of them from the mother and some of them from the father…



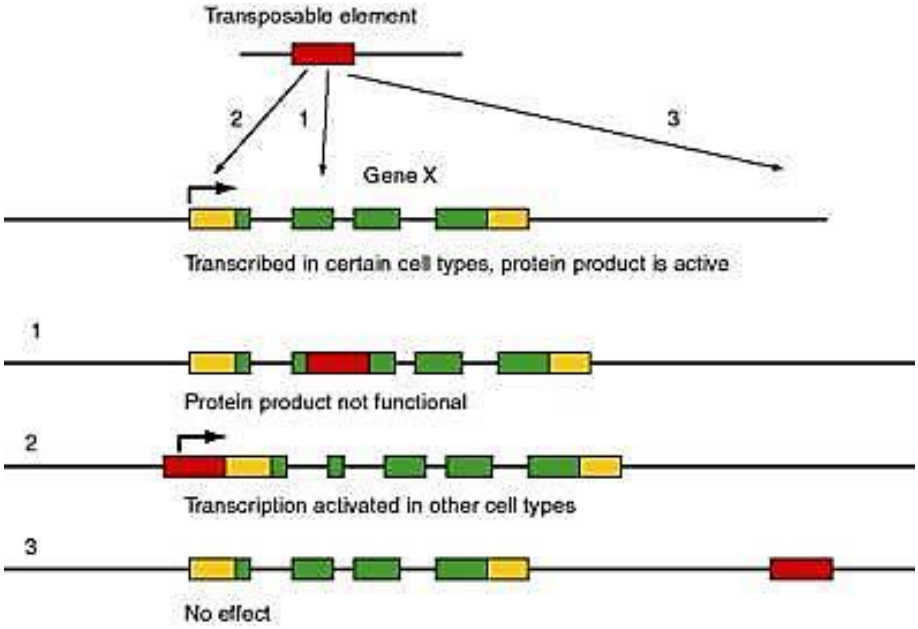
***Transposons [Jumping genes]:**

- They are segments of DNA that can move from their original position in the genome to a new location. [That's why they are called jumping genes, they can move from one position to another in the genome]
- We have two classes of transposons:
    - **A.** DNA transposons (2-3% of human genome)
    - **B.** RNA transposons or retrotransposons where transposable DNA sequence moves by RNA intermediate (40% of human genome) which have two types as we have mentioned earlier:
        - I. Long interspersed elements (LINEs) [Example: **L1**]
        - II. Short interspersed elements (SINEs) – An example is **Alu** (300 bp), which, as we have mentioned earlier, is believed to play a role in gene expression.

**Over 99% of the transposons in the human genome lose their ability to move, but we still have some active transposable elements [the remaining 1%] that can sometimes cause diseases such as Hemophilia A and B, severe combined immunodeficiency, porphyria, predisposition to cancer, and Duchenne muscular dystrophy.

**Take a look at this figure…

**Here we have a transposable element [a transposon capable of jumping]… It could jump into 3 different positions… Let's explain each one and its outcome…



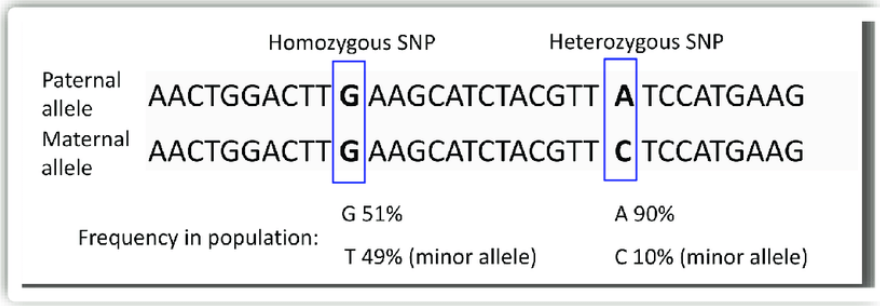
1. The transposable element inserts itself **inside** the gene...If that happens, we get a completely different code for the protein...So then we might end up with a dysfunctional protein that causes disease.
2. The transposable element inserts itself **in the beginning** of the gene. If it does that, it can switch on the expression of that gene in other cell

types…However, that is usually unnecessary… In other words, transcription of that gene in other cell types due to the insertion of the transposable element in the beginning of the gene is usually unnecessary.

3. The transposable element inserts itself **away** from the gene… In that case, nothing will happen… It won't have any effect.

****SNPs [Single nucleotide polymorphisms]:**

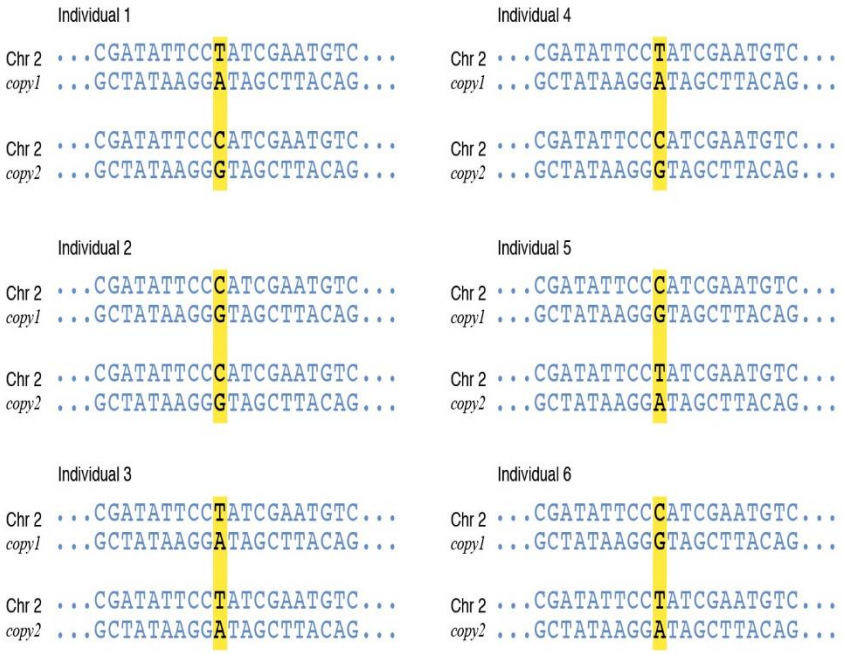
- They are another source of genetic variation
- Here we see single-nucleotide substitutions of one base for another [For example: I substitute A for T at a specific spot on the gene which could cause a change in its expression].
- Two or more versions of a sequence must each be present in at least one percent of the general population
- SNPs occur throughout the human genome - about one in every 300 nucleotide base pairs…Which means: There are about 10 million SNPs within the 3-billion-nucleotide human genome. However, only 500 thousand of those SNPs are thought to be relevant.



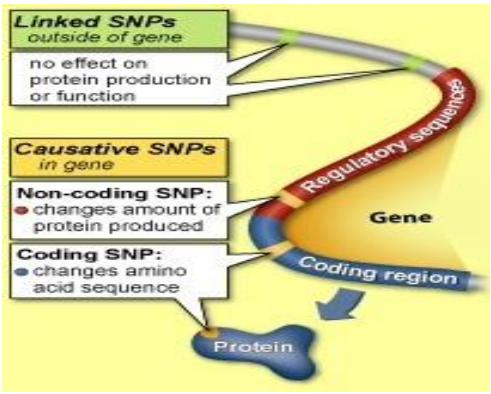
***A heterozygous SNP** is the one where there is a different nucleotide in that position in each homologous chromosome of the same pair [For example, in the paternal allele I have the nucleotide A in a certain position while in the maternal allele I have the nucleotide C in the position equivalent to that of A on the paternal allele…]

***A homozygous SNP** is the one where the same nucleotide is in that position in each homologous chromosome of the same pair [For example, I have the nucleotide C in a certain position in the paternal allele…Also, I have the nucleotide C on the position equivalent to it in the maternal allele]…

*Here we can see 6 different individuals with different single nucleotide polymorphisms…Some individuals have the base C while others have the base T in the same position…Also, that could differ between the homologous chromosomes of the same individual [And that would be called heterozygous SNP as we have mentioned earlier.]



***Types of SNPs:**



1. **Linked SNPs:** They occur outside the gene [Replacement of a nucleotide occurs outside the gene]…It has no effect.
2. **Causative SNPs:** They occur in the gene… We have non-coding SNPs which happen by replacement of a nucleotide at non-coding sequences in the gene such as promoters. **It could change the amount of protein produced.**
   We also have coding SNPs which occur in the coding regions of the gene. Therefore, the replacement of a single nucleotide could **change the codon and produce a different amino acid**, which would then **alter the amino acid sequence of the protein**, so we would end up with a dysfunctional protein.

*****Important notes:**

- Non-coding SNPs could change the speed of gene expression.
- Coding SNPs could increase or decrease the activity of the produced protein.
- An important example on causative SNPs is **personalized medicine**

**What is personalized medicine?

- It's a medicine that is given to patients based on their status.

*As humans, we can be classified to slow metabolizers, fast metabolizers and intermediate metabolizers… That depends on single nucleotide polymorphisms, as they could slow down metabolism.

The question is: What has personalized medicine got to do with this?

*Slow metabolizers have slower detoxificating enzymes in their livers, while **fast metabolizers** have faster detoxificating enzymes in their livers. So, in slow metabolizers, the medicine is detoxificated slowly…Therefore, it stays in their bodies for longer periods of time, and that's why they get small doses. However, in **fast metabolizers**, the medicine is detoxificated quickly, so it stays in their bodies for a short period of time. Therefore, we need to give them large doses in order for the medicine to have its effect.

So, basically, a small dose of a certain medicine given to a slow metabolizer will have the same effect as a large dose of the SAME medicine given to a fast metabolizer.

Good luck everyone, I wish you all the best…