

# Forensic DNA analysis

Giulia Cereda

- Butler, J. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* . Elsevier Academic Press, New York. (chap.5)
- Buckleton, J., Triggs, C., and Walsh, S. (2005). *Forensic DNA evidence interpretation* . CRC Press, Boca Raton. (chap.1)

- DNA is the molecule that encodes the genetic instructions for the development and functioning of all known living organisms.



DNA has a double-strand structure, where each strand is made up of a sequence of four nucleobases: Adenine, Cytosine, Guanine and Thymine (usually denoted by four letters, A, C, G and T)

Each base is attached to a sugar molecule and a phosphate molecule to form **nucleotides**, arranged in the two long strands to form a spiral, with the shape of a double helix.

Bases of one strand pair up with bases of the other strand – A with T, and C with G – to form units called **base pairs**.

The instructions encoded in DNA are stored as a code made up of a double sequence of about 3 billions base pairs.



The order of the bases in the sequence determines the information available for building and maintaining an organism.

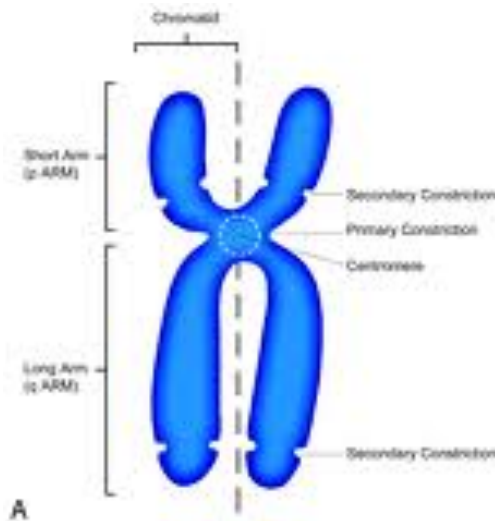
It is sufficient to consider one strand, such as

**AATTGCCTTTTAAAAA.**

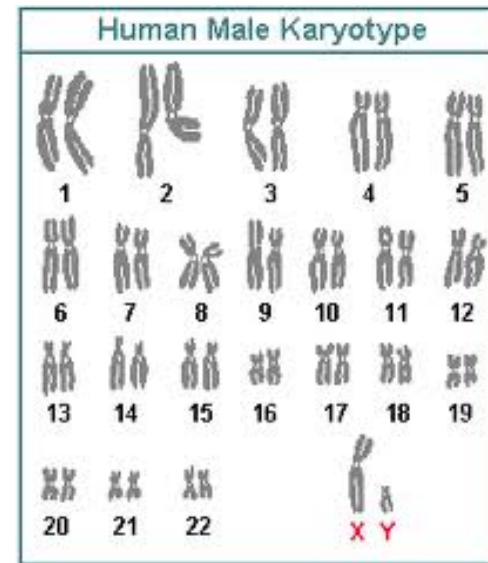
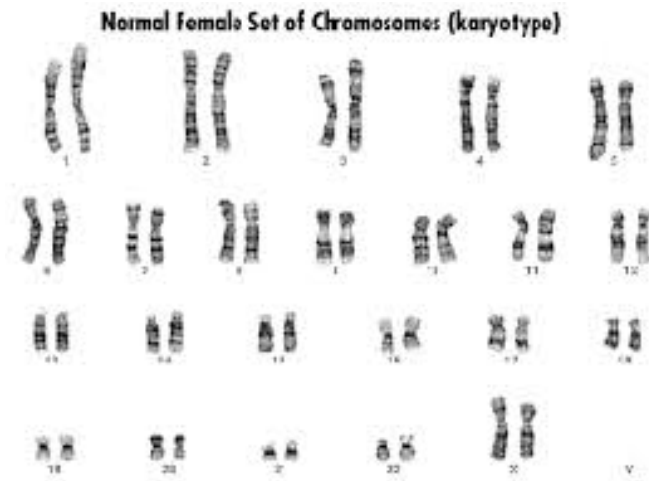
A distinct portion of DNA which codes instructions for a particular body's need (mostly the creation of proteins), is called **gene**.

There are about 32000 genes in the human DNA, which form the so-called **genome**.

All nucleotides are not aligned on a single chain: they are organised in thread-like structures called **chromosomes**.



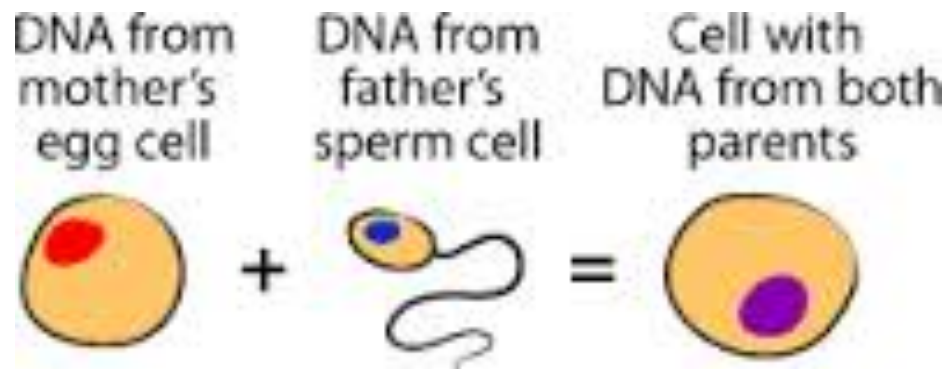
Humans have 46 chromosomes which form 23 couples of homologous chromosomes, identical to one another in shape and size, one inherited from the mother and one inherited from the father



One of these pairs is composed by the **sex chromosomes** which, among other functions, determine the sex of the individual.

The other 22 pairs of chromosomes are called **autosomal** chromosomes and determine the rest of the body makeup and functions.

The DNA code, or genetic code, is passed to the offspring through the paternal sperm and the maternal egg: the mother passes 23 chromosomes through her egg, the father passes 23 chromosomes through his sperm.

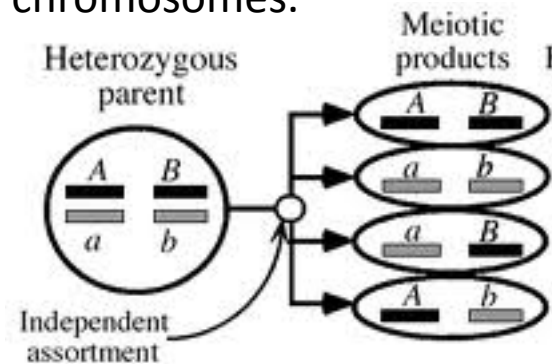


# DNA as identification tool

The entire sequence of the DNA is unique to each individual.

The reason for this variability is due to *recombination*, the process by which a combination of genes different from that of either parents is inherited by progeny.

Meiosis (or gamete cell production) is the process during which each reproductive cell receives randomly one representative of each pair of chromosomes.



Since all chromosomes are conveyed independently one another, from the same couple of parents at least  $2^{46}$  ( about 70000 billions) possible different offsprings can be generated.

Other sources of variability among individuals are **mutations**, which are changes in the nucleotide sequence of the genome, due to deletions, insertions or metamorphosis of some nucleotides.

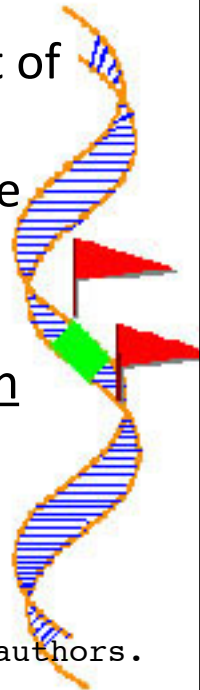


The unicity of each person's entire DNA sequence makes of DNA traces one of the most useful type of evidence for forensic identification

This is why, from its first use in the UK in 1986 (R vs Colin Pitchfork, Wambaugh (1989)), its use in forensic applications has become widespread.

Although the entire DNA material, is unique to each person, the greatest part of it is similar in all humans and only the 0.1% characterizes the different individuals: the variations which are present in this minute portion of DNA are called **polymorphisms**.

For forensic purposes, only those (some) portions of the DNA sequence which are known to display a polymorphism are analyzed: these zones are called **genetic markers (or loci)**



# Genetic markers

Portions of the DNA sequence which are known to display a polymorphism.

They are of different types, depending of the type of polymorphism:

RFLP ([Restriction Fragment Length Polymorphism](#))

VNTR ([Variable Number of Tandem Repeat](#))

SNP ([Single Nucleotide Polymorphism](#))

**STR** ([Short Tandem Repeat](#))

SSLP ([Simple sequence length polymorphism](#))

AFLP ([Amplified fragment length polymorphism](#))

RAPD ([Random amplification of polymorphic DNA](#))

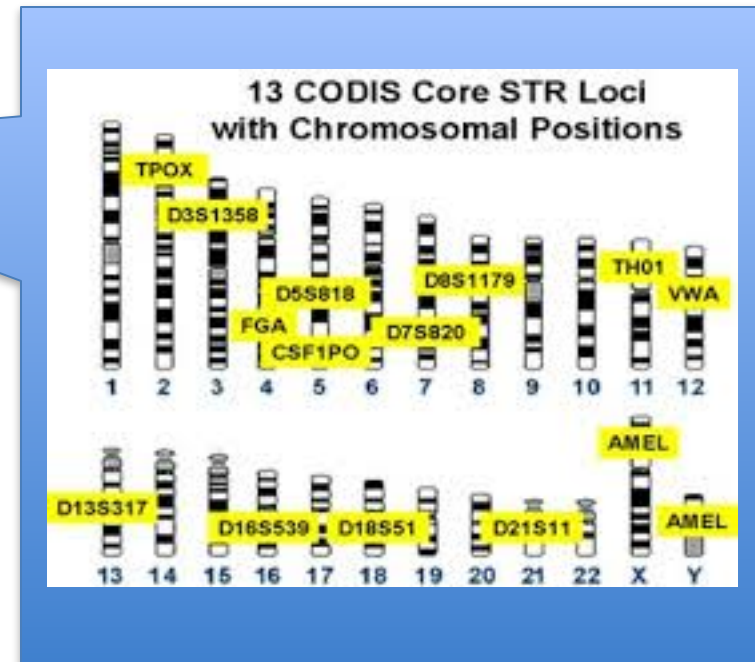
[Microsatelliti \(polimorfismo\),](#)

[SSR \(Simple sequence repeat\)](#)

SFP ([Single feature polymorphism](#))

DArT ([Diversity Arrays Technology](#))

RAD ([Restriction site associated DNA markers](#))



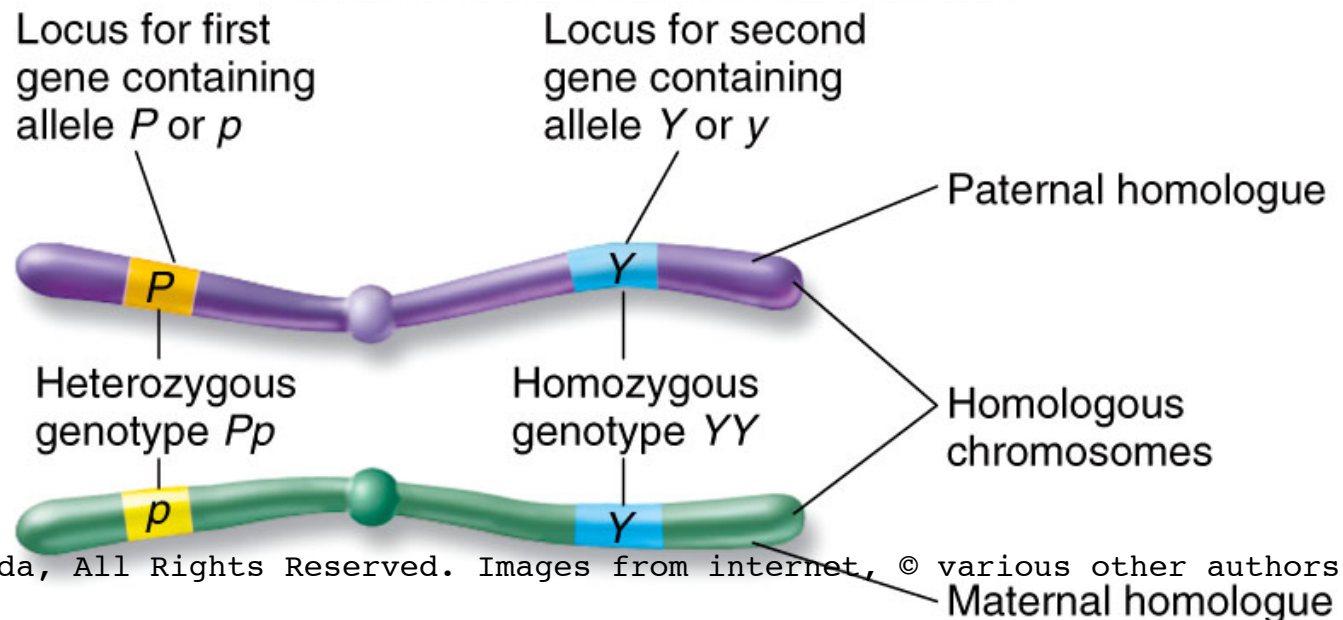
The alternative possible variants which the DNA sequence displays in these zones are called **alleles**.

An individual can have at most two alleles for the same polymorphism: the one carried by the paternal chromosome and the one carried by the maternal chromosome.

If these two alleles are equal, the individual is said to be **homozygous** at the specific marker, otherwise he is said to be **heterozygous**.

The couple of alleles present at an individual's genetic marker is called **genotype** and a combination of alleles at adjacent locations on a chromosome, inherited together, is called **haplotype**.

A **DNA profile** is the combination of genotypes of multiple markers.



# Different kind of polymorphisms

They are of different types of polymorphism:

RFLP ([Restriction Fragment Length Polymorphism](#))

VNTR ([Variable Number of Tandem Repeat](#))

SNP ([Single Nucleotide Polymorphism](#))

**STR** ([Short Tandem Repeat](#))

SSLP ([Simple sequence length polymorphism](#))

AFLP ([Amplified fragment length polymorphism](#))

RAPD ([Random amplification of polymorphic DNA](#))

[Microsatelliti \(polimorfismo\)](#), [SSR \(Simple sequence repeat\)](#)

SFP ([Single feature polymorphism](#))

DArT ([Diversity Arrays Technology](#))

RAD ([Restriction site associated DNA markers](#))

STR belong to the class of *length polymorphisms*: particular portions of the DNA molecule where a particular sequence of nucleotides repeats itself many times.

The different alleles are represented by the difference in the number of repetitions.

# STR polymorphisms

A short tandem repeat (**STR**) polymorphism is a length polymorphism made of a pattern of two or more bases, which are repeated directly adjacent to each other.

These patterns are called **words**, typically 4 base pairs long, repeated between 3 and 51 times: the polymorphism is represented by the difference in the number of repetitions of the same word, between the different individuals.

**STR markers** are specific regions of the DNA in which an STR polymorphism is known to exist.

STR alleles will be denoted as numbers corresponding to the number of repetitions of the same sequence.

Chromosome  $A_1$

..CGGGTATTGATTGATTGATTGATTGATTGATTGATTGGAAAGGT.....

Chromosome  $A_2$

..CGGGTATTGATTGATTGATTGATTGATTGAGTTGTATGAAAGGTC...

A repeating pattern, namely the word ATTG, can be seen on both chromosomes:

Chromosome  $A_1$

..CGGGT ATTGATTGATTGATTGATTGATTGATTGATTGATTGGAAAGGT..  
 $8 \times \text{ATTG}$

Chromosome  $A_2$

..CGGGT ATTGATTGATTGATTGATTGATTGTGTTGTATGAAAGGT..  
 $6 \times \text{ATTG}$

The genotype of this person at this locus, would be either (6,8) or (8,6). Conventionally, STR genotypes are sorted, thus we use  $(a, b)$  with  $a \leq b$ . Repeat numbers could be integers or decimals: for instance, if the repeat number is 9.3, this means that there are 9 repetitions and an incomplete repeat consisting of 3 more letters.

# Technical steps of DNA genotyping

## DNA extraction

DNA can be found on different biological materials, such as blood, sperm, saliva, and hairs, but also on objects which have been in contact with human cells.

After a trace is collected, its cells are broken down with chemical reagents, in order to reveal the DNA inside them.

After this, the DNA is purified, with the aim of separating it from other molecules that can potentially interfere or inhibit the process of analysis.

# Technical steps of DNA genotyping

## DNA amplification

In order to obtain a genetic profile from a DNA trace, it is necessary to amplify it across several orders of magnitude.

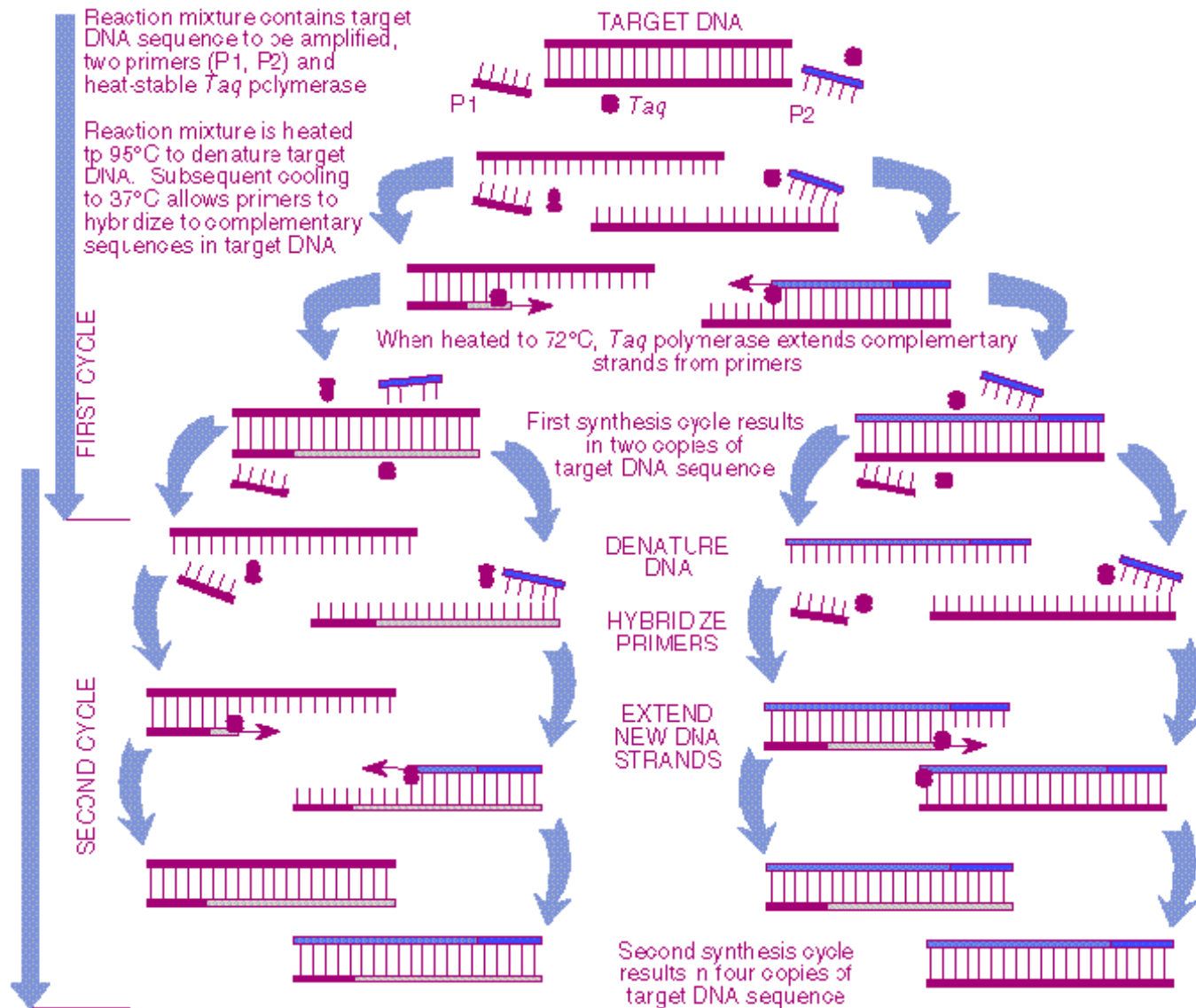
This is done through a biochemical technology, **called Polymerase Chain Reaction or PCR** (Mullis, Nobel prize 1993), which generates thousands of millions of copies of a chosen zone of DNA.

The method relies on about 30 cycles of repeated heating and cooling.

The target DNA sequence to be amplified is pinpointed through the use of primers, which are short DNA fragments containing sequences of nucleotides complementary (according to the rule A-T, C-G) to the target region, together with an enzyme, the DNA polymerase, that adds the building blocks in the proper order based on the template DNA sequence.



# DNA Amplification Using Polymerase Chain Reaction



# Technical steps of DNA genotyping

## DNA separation

In order to resolve the different amplified alleles one to another, a separation step is required to pull the different sized fragments apart, and to allow each allele to be distinguished from other alleles

This separation process is usually performed through *electrophoresis*, a technique which is used to distinguish molecules on the basis of their weight.

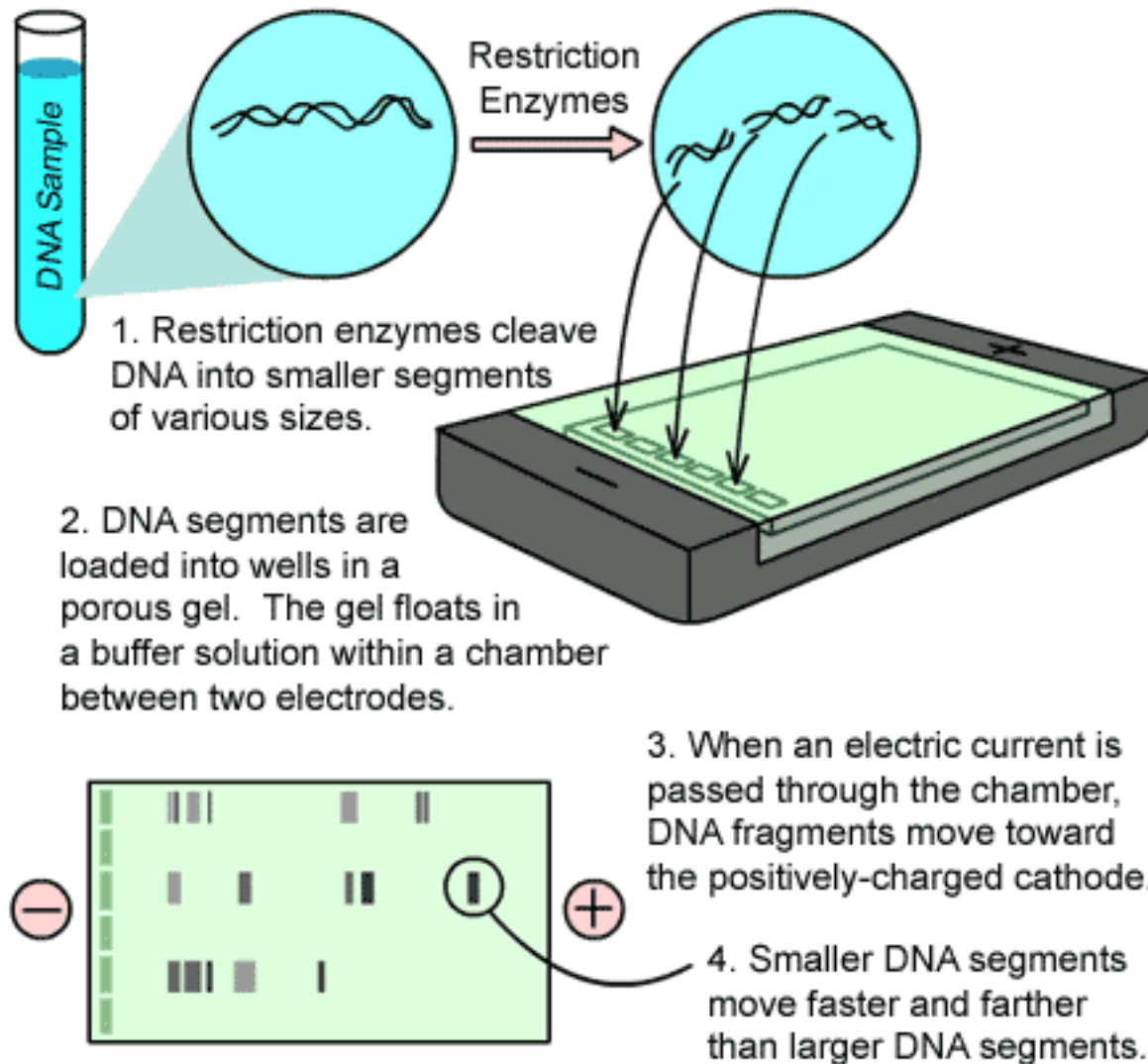
The process is based on the migration undertaken by charged molecules, when immersed in a liquid and exposed to the electrical field generated by a couple of electrodes of opposite charge: negatively charged molecules move to the positive electrode, and vice versa.

This movement has a different speed, on the basis of the size of the molecules: light molecules will move faster than heavier ones and this causes the alleles to sort themselves according to weight.

If applied to the amplified DNA fragments, the process consists of injecting the amplified sample into a gel, then to pass an electrical current through the gel, causing the alleles, which are all negatively charged, to move towards the positive pole.

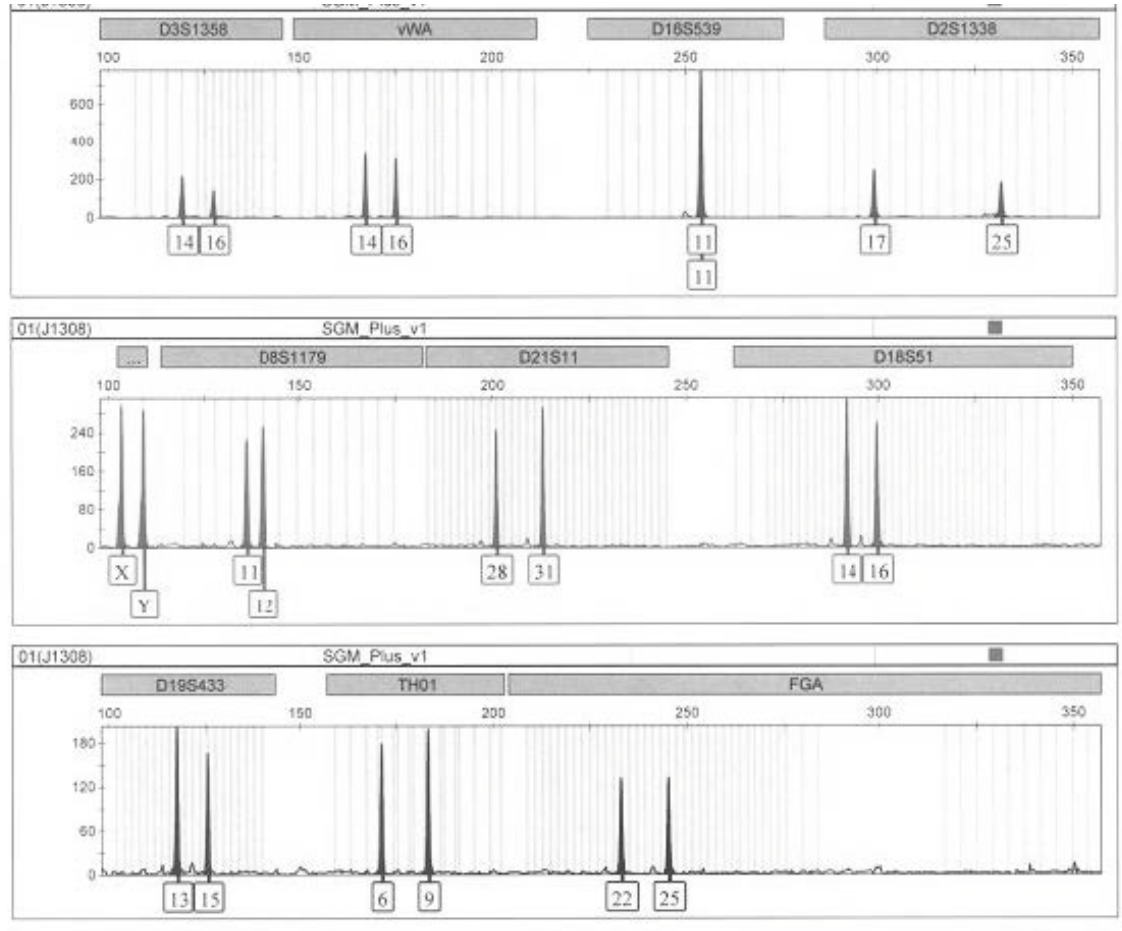
The speed depends only on the length of the fragment

The measurement and analysis are mostly done with a specialised gel analysis software. Capillary electrophoresis results are typically displayed in a trace view called an [electropherogram](#).



The output from the process is a graph, called **peak profile**, in which the horizontal axis gives the base pair measurement, and the vertical axis the light intensity.

Each peak indicates the presence of an allele, where the height is a measure of the amount of the allele in the amplified sample.



# DNA mixtures

DNA mixtures are stains that contain genetic material from more than one person.

This can be due to a contact between the different individuals' DNA material, anytime before the trace is collected

Mixtures are commonly found, after sex rapes, in vaginal swabs obtained from the victim, as the traces may contain material coming from the victim, the perpetrator, but also from other consensual partner

The main factor identifying a mixture is the presence of three or more alleles at at least one locus.

The criticality of mixtures is represented by the difficulty of discerning the particular genotype of each contributor: this happens every time the different contributors share some alleles in some locus

Alleles detected in locus 1	Alleles of the possible donors		
11,12,13	A=11,11 B=12,13	A=11,12 B=12,13	A=11,11 B=11,12 C=11,13

A way for correctly characterising the different contributor's allele, is to use information about the peak heights (or area).

***Identification and separation of DNA mixtures using peak area information***

Cowell, R. G. and Lauritzen, S. L. and Mortera, J.

Forensic Science International

(2007) 166, 28--34

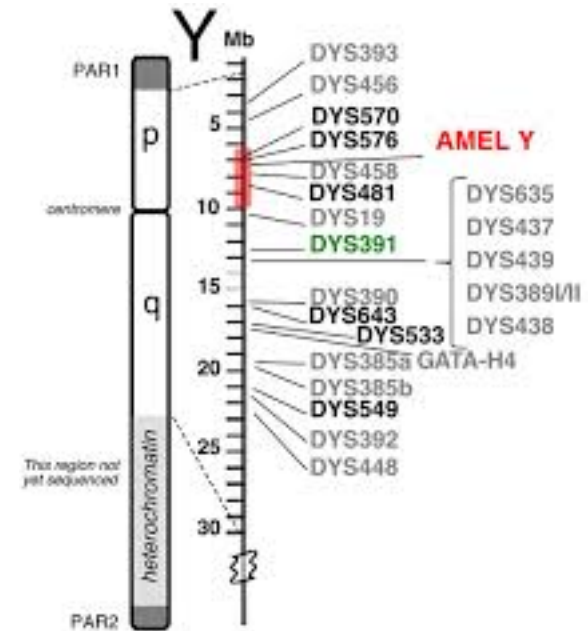
# Y-STR markers

Y-STR markers are STR markers situated on the Y-chromosome.

They are used in forensic caseworks, especially for their capacity to reveal male-specific Y-STR allele in female/male DNA mixtures, even if extremely unbalanced (when classical STR markers are not performing adequately).

Limitations:

- they are usable only for mixtures with a specific gender mismatch (Y-STRs only detect male component's DNA in a female background)
- Y-STR can be quite common in a population and patrilineal relatives of a suspect cannot be excluded as being the contributors of the stain, if no mutations occur





# The Likelihood ratio approach

Given a DNA stain and a matching suspect, whose genotype is known, two main hypotheses may typically be of interest for a correct evaluation of the available evidence:

$H_p$ : the DNA in the crime stain came from the suspect.

$H_d$ : the DNA in the crime stain came from an unknown donor

The largely accepted method for assigning probative value to evidence in order to discriminate between the two mentioned propositions, is the calculation of the Bayes factor (BF), regularly called in forensic context **likelihood ratio (LR)**

$$LR = \frac{P(E | H_p, I)}{P(E | H_d, I)}$$

The likelihood ratio is used to quantify the way in which new information can change the belief, or 'odds', that a particular hypothesis is true.

Prior odds and posterior odds are the odds before and after introducing information, such as a new piece of evidence.

As part of Bayes' theorem, the likelihood ratio connects prior odds to posterior odds in the following way:

$$\underbrace{\frac{P(H_p|E, I)}{P(H_d|E, I)}}_{\text{Posterior odds}} = \underbrace{\frac{P(E|H_p, I)}{P(E|H_d, I)}}_{\text{Likelihood ratio}} \underbrace{\frac{P(H_p | I)}{P(H_d | I)}}_{\text{Prior odds}}.$$

The likelihood ratio, which is a measure of the probative value of the findings E with respect to two alternative hypotheses, is to be distinguished from the conditional degree of belief on the same hypotheses (represented by posterior odds)

# Example

E: Profile of the suspect = profile on the stain.

D3S1	vWA	D16	D2S1	D8S1	D21	D18	D19	TH01	FGA	AMEL
<b>14</b>	<b>14</b>	<b>11</b>	<b>17</b>	<b>11</b>	<b>28</b>	<b>14</b>	<b>13</b>	<b>6</b>	<b>22</b>	<b>X</b>
<b>16</b>	<b>16</b>	<b>11</b>	<b>25</b>	<b>12</b>	<b>31</b>	<b>16</b>	<b>15</b>	<b>9</b>	<b>25</b>	<b>Y</b>

H<sub>p</sub>: The stain belongs to the suspect

H<sub>d</sub>: The stain belongs to someone else

$$LR = \frac{P(E|H_p)}{P(E|H_d)} = \frac{1}{f_{profile}}$$

$$LR = \frac{P(E|H_p)}{P(E|H_d)} = \frac{1}{f_{profile}}$$

Autosomal markers

$$f_{profile} = \prod_{m=1}^{mk} f_{profile}^m$$

$$LR = \prod_{m=1}^{mk} LR_m$$

This is due to the recombination:  
autosomal STR are chosen far enough  
not to pass together.

Y-STR markers

They do not recombine, they cannot be  
considered as independent, so the  
product rule does not apply.