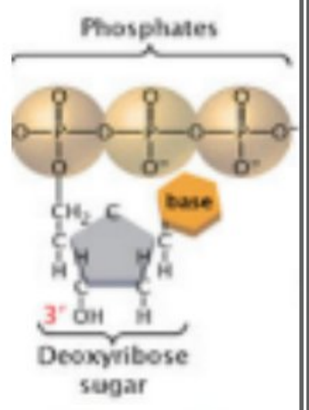
❖ DNA sequencing

- **DNA sequencing:** The process of determining the exact order of nucleotides in a genome or in a DNA fragment
- **<u>We use DNA sequencing in (the importance and purpose):</u>**
  - ➢ Identification of **genes** and their localization
  - ➢ Identification of **protein structure and function**
    - ✓ By knowing the <u>sequence of codons</u> which are translated to this protein → so we can know amino acids sequence → Knowing the structure and function of the protein
  - ➢ Identification of **DNA mutations**
    - ✓ By comparing the sequence of a DNA fragment with the sequence of the normal genome to know whether there are a mutation or a disease
  - ➢ Clarify (Elucidate) **Genetic variations** among individuals in health and disease
    - ✓ By sequencing the genome of individuals we can know (predict) the susceptibility for a certain disease → so giving advices how to take care and we can know the best treatment
  - ➢ **Evolutionary conservation** among organisms
    - ✓ To study the evolutionary similarities & differences between different species (organisms)

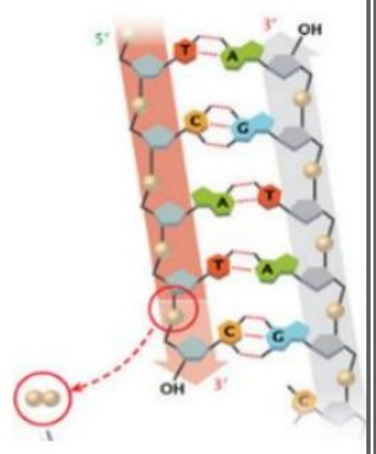
- **<u>DNA sequencing of organism genome (history):</u>**
  - The first genomes to be sequenced were **Viruses and prokaryotes** (simple & small) → Then Human **mitochondrial DNA**
  - The <u>first eukaryotic</u> genome sequenced was that of **yeast (Saccharomyces cerevisiae)**
  - The genome of a <u>multicellular organism</u>, the <u>nematode Caenorhabditis elegans</u>
  - Determination of the base sequence in the <u>human genome</u> was initiated in 1990
    - ✓ Not completed yet → some regions of the Y chromosome remains

- Different organisms differ in their genomes (sequence & size), usually more complex organisms have more complex (larger) genomes → but it is not necessary

❖ DNA synthesis/elongation (Replication)



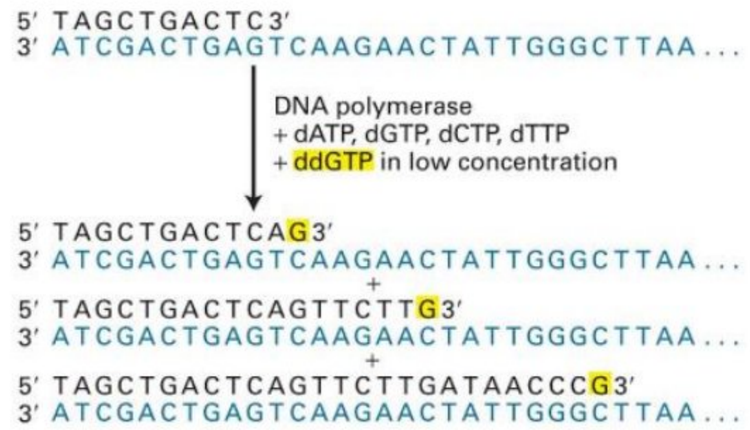
- A DNA fragment is replicated using DNA polymerase → using **Deoxyribonucleotides** as a <u>substrate</u>
  - ➢ (Deoxyribo) → the sugar is a pentose missing the **OH on the 2nd carbon**
  - ➢ **C-5** → is attached to **3 phosphate groups** (triphosphate)
- Replication done by forming **phosphodiester bond** between C-3 **(3' carbon)** which has OH group of a nucleotide with C-5 **(5' carbon)** of the next nucleotide
  - ➢ We get the energy for this reaction by the **cleavage and the release of 2 phosphates** (as a pyrophosphate molecule)



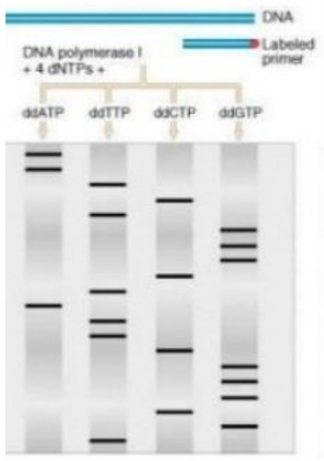
❖ The basic method of DNA sequencing

- It is the most popular & based on premature termination of DNA synthesis by **<u>di</u>deoxyribonucleotides**
  - ➢ **Dideoxyribonucleotides** ➔ Deoxy- on carbons number **2 & 3** (has 2 OH groups are reduced)
  - ➢ When a dideoxyribonucleotide is added to DNA → **we can't add any nucleotide after it** (because it don't have OH on carbon 3 → no phosphodiester bond with the next nucleotide)

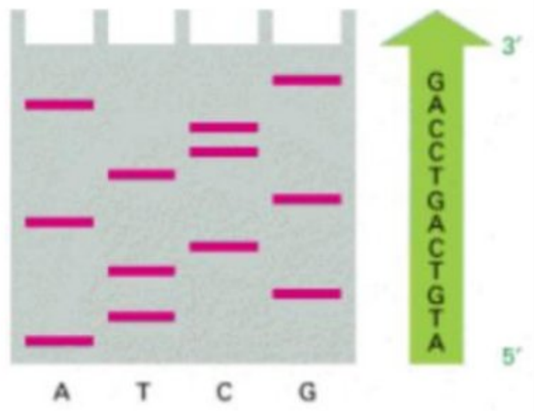
- **The Mechanism of this method:**
  - DNA polymerase starts synthesis (initiated) by a **primer** (can't start de novo) that has been **labeled with a radioisotope** (such as radioactive phosphorus) → to sequence 1 strand of the DNA fragment
  - We put the DNA fragment of interest in 4 tubes and we put the labeled primer, DNA polymerase, the **4 deoxynucleotides** (dA, dC, dG, dT) & a **dideoxynucleotide** in **a low concentration**
    - ➢ The first tube contains ddA, the second one ddC, the 3rd ddG, the 4th ddT having 4 reactions



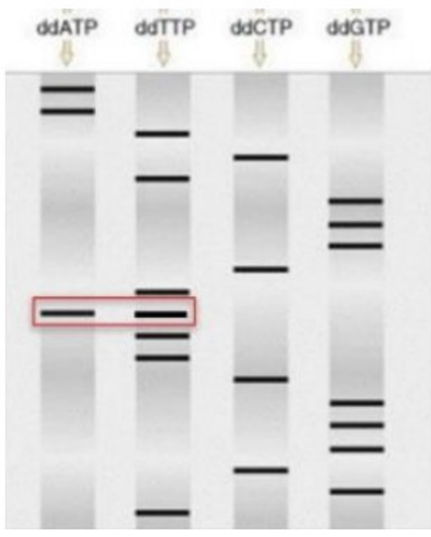
  - DNA polymerase starts synthesizing using deoxynucleotides mainly and on some fragments the dideoxynucleotide can be used terminating the synthesis of these strands forming many newly synthesized strands with **different lengths**
    - ➢ So, labeled DNA molecules are generated, each terminated by the dideoxynucleotide



  - We put the resulted fragments in **gel electrophoresis separating them according to size** and then detected by **X-ray** film → so now we can know the sequence of the newly synthesized fragment
    - ➢ The sequence of the newly synthesized strand → corresponds the **order of the resulted fragments** (from the smallest to the largest)
    - ➢ The sequence of the original strand is → **complementary and antiparallel** to the new strand



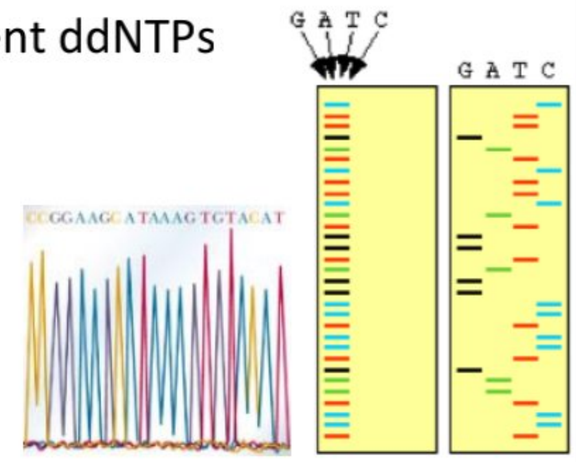
- **Note:**
- Some time there could some strands on the same level (the same length), why?
  - ➢ we have 2 homologous chromosomes → the individual has 2 forms of the gene → **heterozygous,** it can be:
    - ✓ **Polymorphism** (occurs **more than 1%** of the population)
    - ✓ **Mutation** (occurs in **less than 1%** of the population)



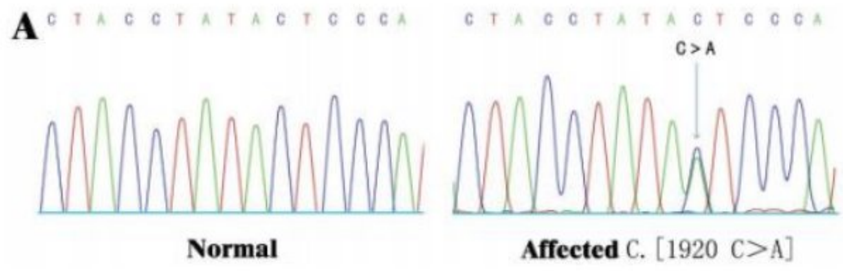
- We can know if this variation causes a disease by:
  - ➢ **Comparing** the sequence **with** the sequence of an individual that has the **disease** → if they both has this band → Carrier of the disease
  - ➢ If an individual is homozygous we can compare his sequence with the disease sequence if they have the same band → this individual is homozygous for the disease → affected by the disease

❖ Fluorescence-based DNA sequencing

- In this method → the 4 deoxynucleotides + the 4 dideoxynucleotides are included in the **same reaction**
- No radioactive labeling → but ddNTP are labeled with **fluorescent tag**
  - ➢ We don't use Radioactivity because it could produce mutations, so it is harmful
  - ➢ Using fluorescence makes it less laborious
- So the tube will have a template, primer, DNA polymerase, dNTPs & fluorescent ddNTPs
  - ➢ Each ddNTP has a tag with certain color (signal)
- So instead of having them separated in 4 lanes, we will have them in one lane and as they migrate, we have a sensor that reads the fluorescence and transforms the results in the shape of peaks, then it **reads the peak** and translates it into a letter

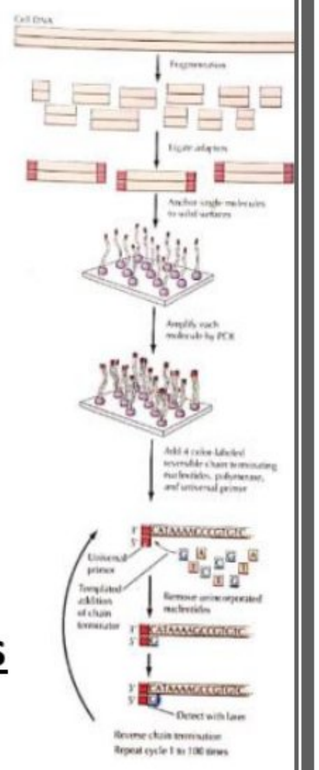


- **<u>What if we had 2 peaks in the same locations?</u>**
  - ➢ **Heterozygous** person ➔ Polymorphism or mutation
    - ✓ If the 2 peaks have different lengths → the longer wave is more present

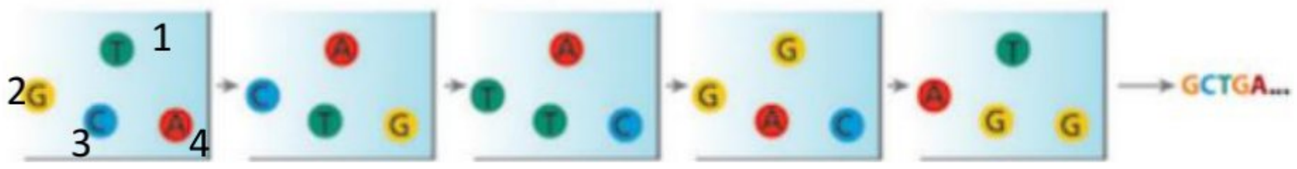


❖ Next-generation sequencing

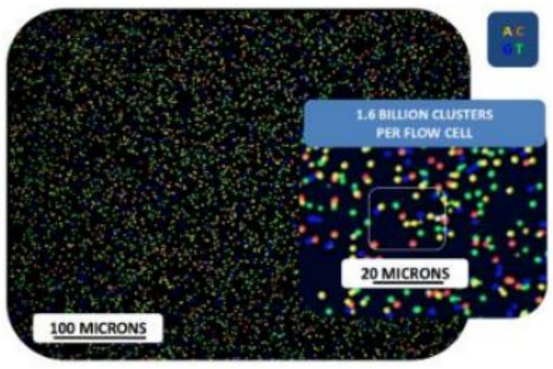
- The fastest method
- It is Done by:
  - ➢ **Fragmenting the cellular** DNA randomly and adding a DNA adapter (with a known sequence) to the end of each fragment
    - ✓ The formed fragment can have <u>overlap</u> between their sequences
  - ➢ We attach these fragments to a to a solid surface and amplified like PCR forming millions of clusters
  - ➢ **Adding primers that anneal to the adapter** → We add the same primer to all clusters because it is complementary to the adapter not the fragment itself
  - ➢ Then DNA polymerase start synthesizing the new strand using **<u>modified nucleotides</u>** (with terminating ends) as a substrate:
    - ✓ Each nucleotide have a **certain color** → detected by a <u>special camera</u>
    - ✓ When a nucleotide is added (incorporated) → no new nucleotides can be added until the incorporated on is chemically **modified** (**Activated**) ➔ now a new nucleotide can then be added to it
  - ➢ The cycle is repeated until sequencing the whole fragment



- For example the sequence of the newly synthesized of the:
  - ➢ 1st cluster: TAAGT
  - ➢ 2nd cluster: GCTGA
  - ➢ 3rd cluster: CTTAG
  - ➢ 4th cluster: AGCCG



- The solid surface contain millions of clusters → so after sequencing each cluster we use bioinformatics and a computer program to combine all information we have
  - ➢ The fragments overlap → so this program can give us the sequence of the original strand after analyzing the result (information)



❖ Polymerase Chain Reaction

- **Challenges in study DNA in research and medicine:**
  - Its hard to study **genetic variation** (such as STR, VNTR, SNPs and mutations)
  - Hard to deal with **minute amounts of genetic material** (such as that of dinosaurs and early human)
  - **Identification of organisms** (such as infectious agents)

- **Polymerase chain reaction (PCR):** A reactions allows the DNA from a selected region of a genome to be **amplified a billionfold** and effectively **purifying** this DNA away from the remainder of the genome
  - It is extremely sensitive; it can detect a single DNA molecule in a sample
  - It is really fast it takes only a few hours, it is also specific so it selects a specific region and amplify it
  - It's just like the cloning but it is faster, more selective and it is a **biochemical enzymatic reaction**
  - It is a chain (consecutive) reaction → repeated several times

- **Components of PCR reaction**
  - **The DNA template** (can be **circular or linear** but the resulted DNA molecules will be only linear)
  - **A pair of primers**
    - ✓ The point where DNA polymerase start amplifying (initiating polymerase activity)
    - ✓ **15-25 nucleotides**-long primers → to surround (limit) the target sequence
    - ✓ They are called primer 1 (**forward primer**) and primer 2 (**reverse primer**)
  - **All four deoxyribonucleoside triphosphates** (enzyme substrates dATP, dGTP, dCTP, dTTP)
  - **A heat-stable DNA polymerase**

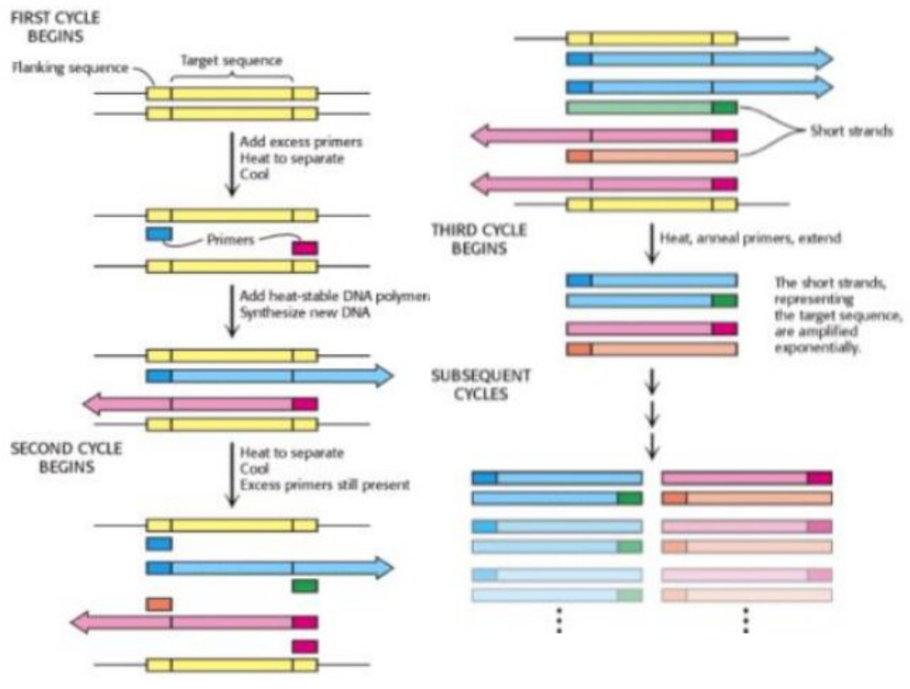
- **Notes:**
  - We use **DNA primers** instead of RNA primer → because they are **more stable**
  - The 2 primers to surround the region we need → each on at each end of the region

- **The PCR steps**
  - **Denaturation (at 95°C):** DNA is denatured into single- stranded molecules
  - **Reannealing (50°C to 70°C):** the primers anneal to the DNA
  - **Polymerization or DNA synthesis (at 72°C):** optimal for the polymerase

- These steps are repeated many times (about 25-30 times) → cycles
- Annealing temperature **depends on the primer itself** (its length , GC content ,... etc)

- **The DNA polymerase**
- Ordinary human Polymerases can't revive in these high temperatures they would be denatured being nonfunctional → so a **suitably heat-stable DNA polymerases** have been obtained from microorganisms whose natural habitat is hot springs
  - For example, the widely used **Taq DNA polymerase** is obtained from a **thermophilic bacterium**, Thermus aquaticus, and is thermostable up to 95°C

- In the first 2 cycles the process of synthesizing DNA contains long extended fragments
- In the **third cycle** → we have the **DNA size (fragment) that we want** → these fragments are now amplified to many copies



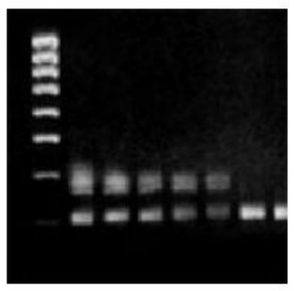
❖ PCR cycles

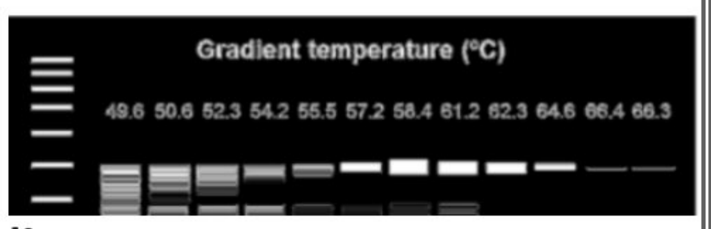
- **20-30 cycles** of reaction are required for DNA amplification
  - ➢ The products of each cycle serving as the DNA templates for the next-hence the term polymerase **chain reaction**
- Every cycle **doubles the amount of DNA**
- After 30 cycles, there will be over 250 million short products derived from each starting molecule

- These DNA fragments can be easily visualized as a discrete band of a specific size by **agarose gel electrophoresis** (1 band containing a huge number of identical DNA fragments)

- **What does determine the specificity of DNA amplification?**
  - ➢ The specificity of amplification depends on the specificity of the primers to not recognize and bind to sequences other than the intended target DNA sequences

- **How can you prevent the non-specific annealing?**
  - ➢ By using the **optimal temperature** for the reaction





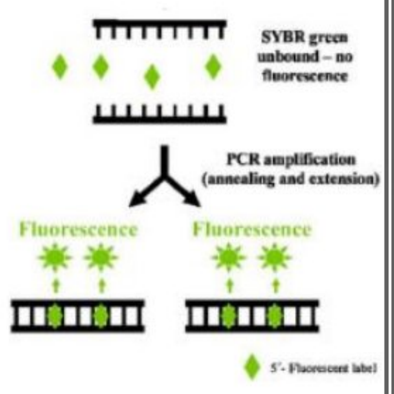
  - ✓ At low temperature → **less specific** annealing because there could be an imperfect hybridization
  - ✓ **The higher the temperature the more specific the annealing**
  - ✓ At Very high temperature → no H-bonding between strands & **no annealing**

- We can use a mouse primers to amplify human genes → due to homology (similarities) in the sequence

❖ Uses of PCR

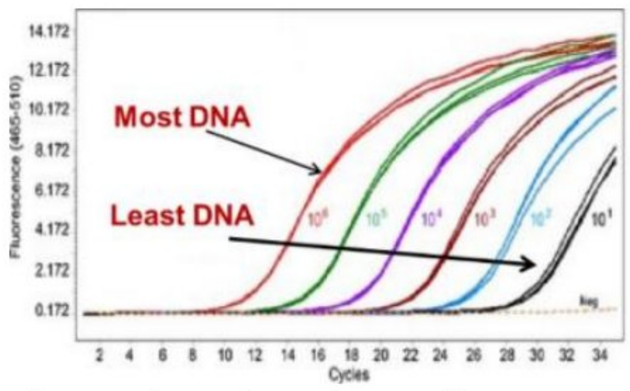
- ➢ **Molecular Fingerprint**
- ➢ **Genotyping**
- ➢ **Genetic Matching**
- ➢ **Detection of mutations**
- ➢ **Prenatal diagnosis**
- ➢ **Cloning**
- ➢ **Mutagenesis**
- ➢ **Molecular archeology**
- ➢ **Detection & classification of organisms**

❖ Real-time Quantitative PCR (qPCR)



- Here we utilize **SYBR green** which binds to double-stranded DNA and fluoresces only when bound (**No binding → no fluorescence**)
- **qPCR:** A way of relative quantitation of amount of DNA in a sample is by amplifying it in the presence of SYBR green

- So we use a specific primer for a specific sequence (such as viral DNA)
  - ➢ If this sequence is not present → no binding → no signal

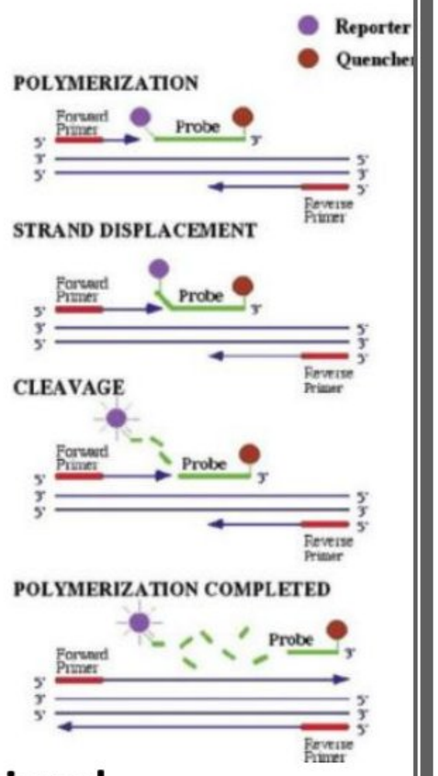


- At early cycles there is no signal detected because the amount of DNA is **very low for the instrument to detect** the fluorescence that is emitted from SYBR green
  - ➢ The greater the initial concentration → the greater amplification over cycles → we will reach the limit at which the instrument starts detecting signals earlier
  - ➢ The more the DNA (initial concentration) the **earlier a signal will be detected**

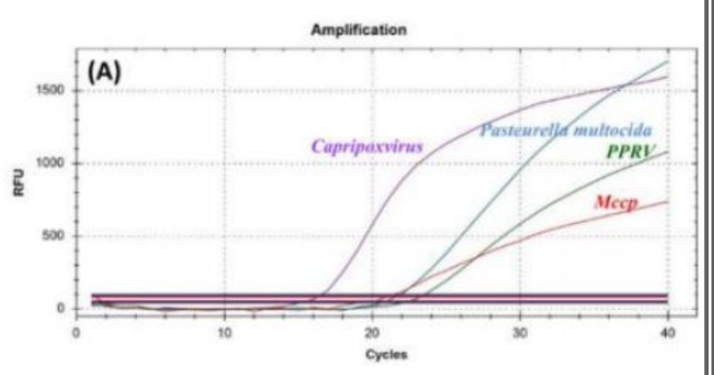
- **Threshold cycle (Ct):** The number of the cycle in which a signal is detected, measuring the **DNA amount**
- So we can use qPCR to know the **viral and bacterial load**
- **How can we be sure that this process is specific (Amplifying the needed fragment)?**
  - ➢ **By melting curve analysis**

- **Melting curve analysis:** Depends on the melting point of the fragments (where 50% of the DNA is denatured) → and there will be a peak of fluorescence observed
  - ➢ So we raise the temperature after the end of the reaction (qPCR)
    - ✓ If we have a **single peak** → there are a single type of fragments → so only the needed region (sequence) is amplified → **specific**
    - ✓ If we have a **more than 1 peak** → there are a many types of fragments → **not specific**
- Many peaks indicates different melting points → different fragments (they differ mainly in the **AT and CG pairs** amount)

❖ Taqman qPCR



- It is more specific, more sensitive, more reproducible & multiplexing
- It uses a primer, template, DNA polymerase, Deoxyribonucleotides and **probes**
  - ➢ This probe is specific to a **specific region to be amplified** → making this reaction **more specific**
  - ➢ The probe is bound to a **reporter** and a **quencher**
    - ✓ The reporter emits signal only **if it is far from** the quencher
- DNA polymerase used here has an **exonuclease activity**
- As the reaction proceeds → DNA polymerase is synthesizing → when it reaches the probe, it will start cleaving it → so the reporter is far from the quencher → **emitting signal**
  - ➢ No signal → no cleavage → the needed fragment is **not amplified**
- If we repeated the reaction several times → it will give the **same results** → **more reproducible**
- We can **amplify many regions of the same sample at the same time** by using the different probes specific to each region & attached to different reporters giving different signals → **multiplexing**



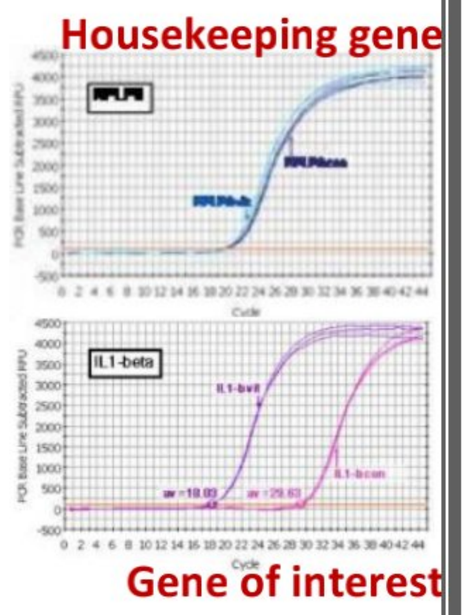
❖ Analysis of gene expression and RNA levels

- **Basic methods:** Northern blotting, in situ hybridization
- **Advanced methods:** real-time PCR, DNA microarray → these deal with DNA
- **Very advanced method:** RNA-seq

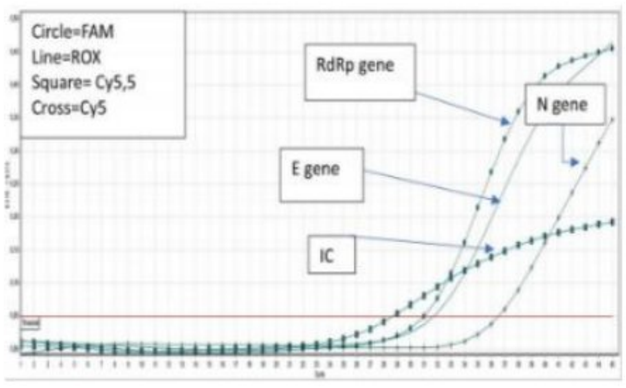
- To use the advanced methods on mRNA we must convert **mRNA into cDNA** by the enzyme **reverse transcriptase**:
  - ➢ We use a poly-T primer that binds to the Poly-A tail of the mRNA
  - ➢ After the synthesis of the first strand of cDNA → the mRNA is degraded and the 2nd strand is synthesized

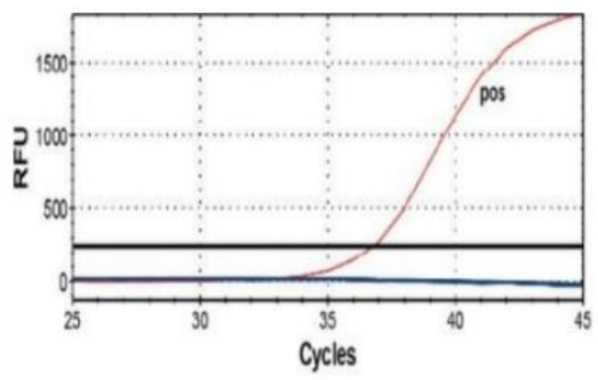
- **Now we can use real-time qPCR & Taqman PCR:**
  - ➢ So, amount of mRNA is reflected by the amount of cDNA
    - ✓ The more cDNA detected = the more mRNA ➔ so 1 mRNA → 1 cDNA …
- We can use qPCR with utilizing SYPR green on cDNA ➔ so the **more cDNA** (in the initial sample) the **earlier signal** is detected → **more mRNA** → this **gene is highly expressed**
- To be sure that the sample is properly collected we need a control → and this control is a **housekeeping gene**

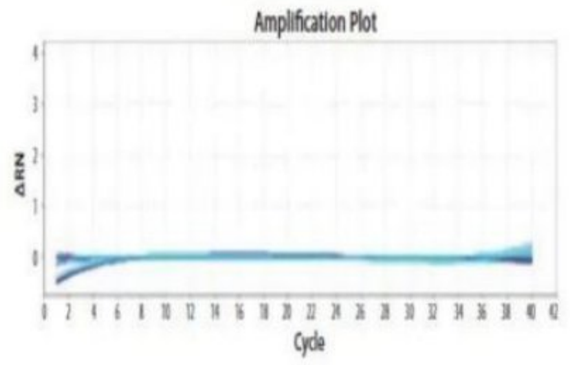
- **Housekeeping genes:** Genes that are expressed in a **constant rate** (**unaltered expression**) in all tissues & cells (such as actin & tubulin)
- When taking 2 samples of a certain gene we also examine the housekeeping gene
  - ➢ If both samples have the same expression for the house keeping gene so now we can study the gene of interest
  - ➢ If they don't have the same expression → the sample is not correctly taken
- In the next picture ➔ the gene in the purple sample is expressed more than it in the pink sample



- **<u>The detection of SARS-CO-2:</u>**
  - ➢ We collect a sample → by a **nasopharyngeal swab** → and put it in a buffer solution
  - ➢ We extract the RNA of 3 genes specific for the virus (**E-gene**, RdRp-gene, N-gene)
  - ➢ RNA is converted into cDNA → which will be **amplified**
  - ➢ To be sure the sample is properly collected we also amplify a **human gene** (IC = internal control)







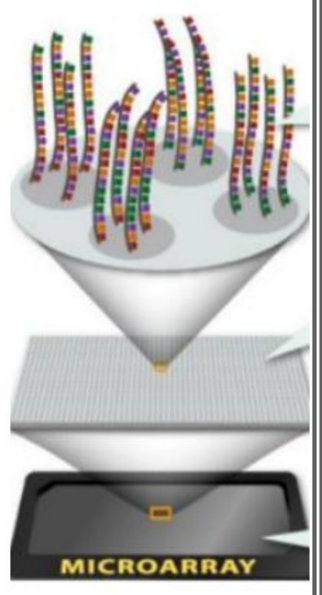
- The sample on the left → correctly collected → Positive to corona virus
- On the middle → correctly collected → Negative
- On the right → Not-correctly collected

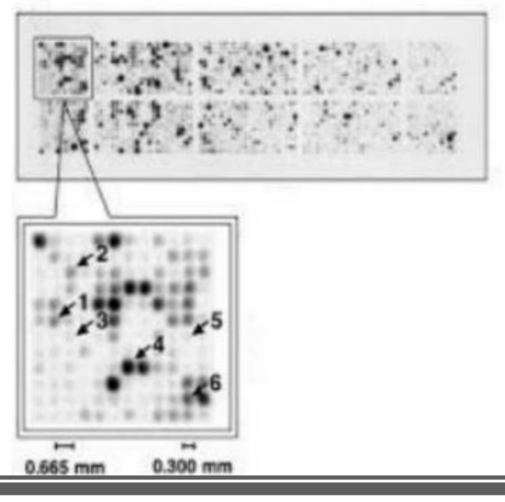
❖ The science of -omics

- **Genomics** → the science studying the whole genome in the cell
- **Transcriptomics** → the science studying the whole transcripts (RNAs) in the cell
- **Proteomics** → the science studying the whole proteins in the cell
- **Metabolomics** → the science studying the whole metabolites in the cell

❖ DNA microarrays

- DNA microarrays are solid surfaces (glass microscope slides or chips) spotted with up to tens of thousands of DNA fragments in an area the size of a fingernail
  - ➢ The exact sequence and position of every DNA fragment on the array is known
- We use it to study the transcriptome → allowing us to study thousands of gene at the same time
- Each array or spot contains a multiple **identical strands of DNA (probes)** that are unique for each spot → Each spot represents a gene that we know
- To study the expression of the genes in a cell:
  - ➢ All mRNA from the cells are extracted (isolated) and converted to **cDNA**
  - ➢ The cDNA is labeled with a **radioactive phosphorus**
  - ➢ We add the cDNA sample to the microarray:
    - ✓ Each cDNA will bind to the complementary probe (specific to it) → **<u>emitting a signal</u>**
    - ✓ More binding → stronger signal
      - **No signal → no expression of this gene**
      - **Strong signal → high expression**
      - **Little signal → low expression**
- So, Microarray tells us about the expression of all genes at the same time

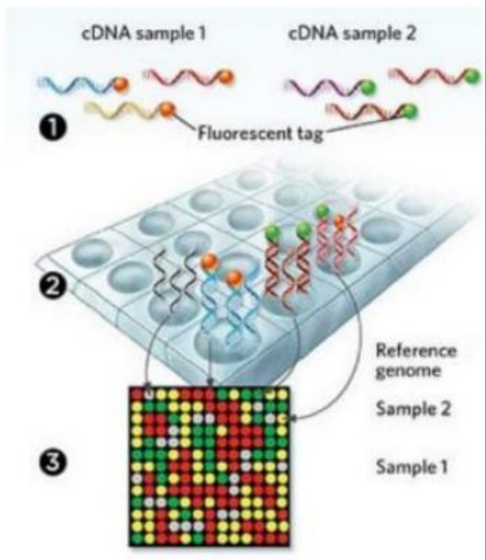




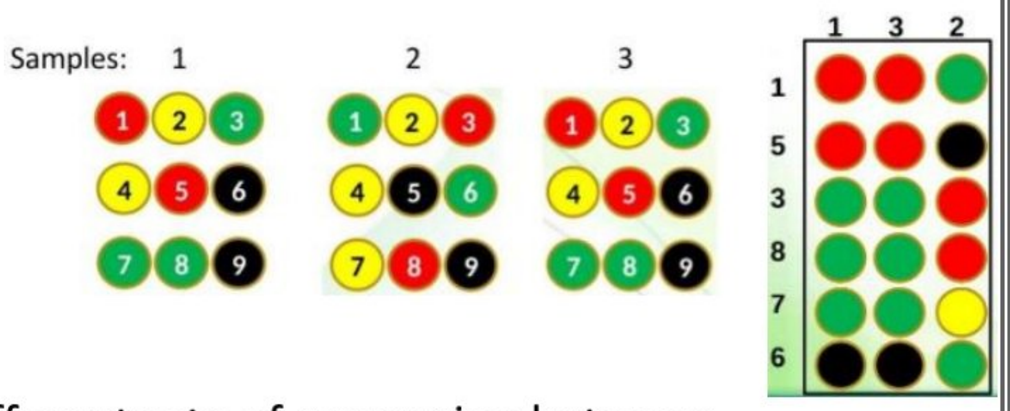
❖ Comparative expression

- We can use the microarray to **compare gene expression** between different samples (such as a normal cell with a cancer cell)
- It has the same mechanism of ordinary microarray but we **don't use radioactive** label, instead we **use a fluorescent tag** with a certain color for each sample

- For example, if we compared a normal cell with a cancer cell:
  - ➢ We label the cDNA from the normal cell with a green color
  - ➢ The cDNA from the cancer cell with red color
  - ➢ We are both samples to the same microarray → the computer will analyze the colors
  - ➢ If the spot has:
    - ✓ **Red color** → this gene (of the spot) is expressed more in the cancer cell
      - ▪ So it can be a **cancerous gene**
    - ✓ **Green color** → this gene (of the spot) is expressed more in the normal cell
      - ▪ So it can be a tumor **suppressor gene**
    - ✓ **Yellow color** → expressed in both cell at the same rate
    - ✓ **White or black** → no expression on either cell

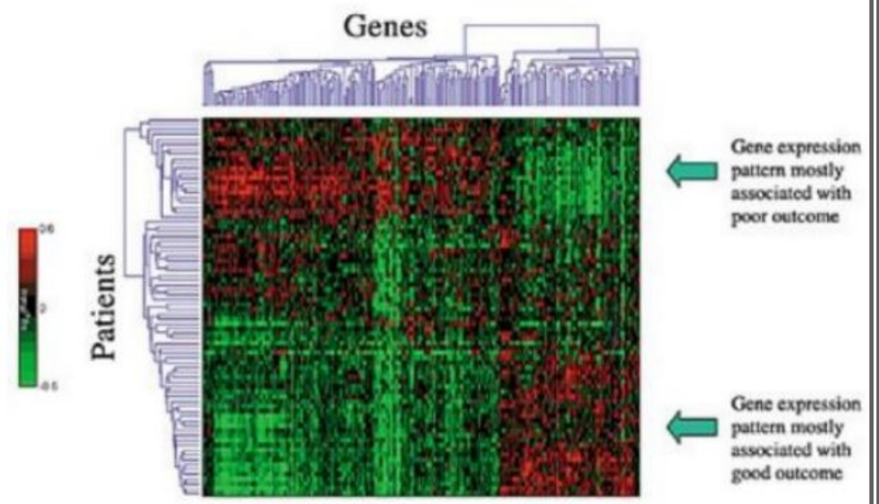


- Normally we compare between **normal with a large number of cancer samples, example:**
  Here we have 3 samples → each 1 of them is compared to the normal:
  - ➢ So, **green** means that it is more expressed in normal cell BUT **less expression (down regulation)** in this sample of cancer cells
  - ➢ **Red** means that this gene is **highly expressed (up regulation)** in this sample
  - ➢ Yellow → expressed as much as the normal cells
  - ➢ Black or white → not expressed in either cell



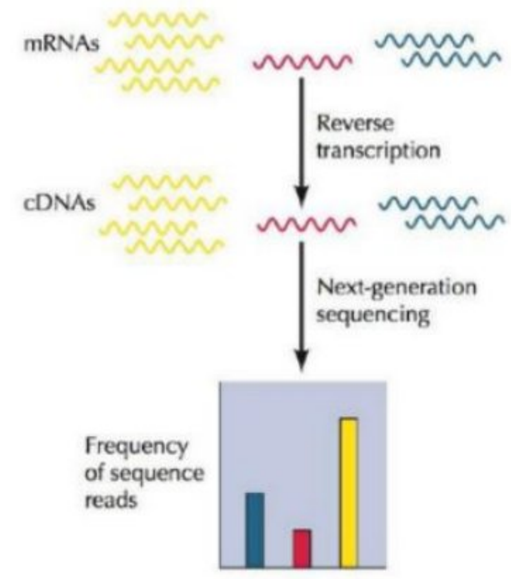
- These samples are combined using bioinformatics:
  - ➢ When we look to these samples → we notice that the expression of genes 2, 4 & 9 doesn't differ between the samples → so we **eliminate them**
  - ➢ We are interested in the other genes that have different rate of expression between samples (for example: gene 3 in sample 1 is down regulated but it is up regulated in sample 2)
  - ➢ By comparing the samples we notice that samples **1 & 3 are similar**

- We can benefit from that → by comparing large number of patient to know about the simulates between their conditions
  - ➢ Patients represented in region number 1 have:
    - ✓ Down regulation of B genes
    - ✓ Up regulation of A genes
  - o Doctors concluded that these patients → give **poor therapy outcomes**
  - ➢ Patients represented in region number 2 have:
    - ✓ Down regulation of A genes
    - ✓ Up regulation of B genes
  - o Doctors concluded that these patients → give **good therapy outcomes**



- So, we can use gene expression to predict if the patient have good or poor therapy outcomes → to know how to deal with him

❖ RNA-seq (RNA sequencing)



- Cellular RNA is reverse transcribed to cDNAs, which are subjected to next-generation sequencing → **Knowing the sequence of RNA**
- The relative **amount (level) of each cDNA (mRNA)** is indicated by the frequency at which its sequence is represented in the total number of sequences read

- RNA-seq can be used to:
  - ➢ Characterize **novel transcripts**
  - ➢ Identify **splicing variants**
  - ➢ Profile the expression levels of **all transcripts**
    - ✓ So this how we knew that 75% of human genome is transcribed

- **Microarrays are limited to detect transcripts corresponding to known genomic sequences** (because we use certain probes to identify them)