

FINAL – Lecture 7

DNA Sequencing

﴿ وَإِن تَتَوَلَّوْا يَسْتَبَدِلْ قَوْمًا غَيْرَكُمْ ثُمَّ لَا يَكُونُوا أَمْثَلَكُمْ ﴾

اللهم استعملنا ولا تستبدلنا

Written by :

- Mohammed Alshwyiat
- Karam AL Qaisi

Reviewed by :

- Mo'awyah Alzghoul





DNA sequencing

Prof. Mamoun Ahram

School of Medicine

Second year, Second semester, 2024-2025

What is DNA sequencing?

- DNA sequencing is the process of determining the exact order of nucleotides in a genome. ➤ It can be human genome, particular gene, VNTRs, STRsetc.
- Importance:
 - Identification of genes and their localization.
 - How they are localized on the chromosome, on the short or long arms, near the telomere... etc.
 - Identification of protein structure and function.
 - Through sequential order of the gene, proteins 3D structure and function can be determined.
 - Identification of DNA mutations.
 - Genetic variations among individuals in health and disease.
 - Prediction of disease-susceptibility and treatment efficiency.
 - Genetic variations governs complex factors in health and disease states; these variations; such like RFLP, VNTRs, STRs, SNPs, make some people more susceptible to some diseases, more responsive to a particular drugs or certain drug dosages and even the severity of disease may vary according to which.
 - Evolutionary conservation among organisms.

DNA sequencing of organism genome

- Viruses and prokaryotes, first. ➤ **Because their genome are simple.**
- Human mitochondrial DNA.
- The first eukaryotic genome sequenced was that of yeast, *Saccharomyces cerevisiae*.
- The genome of a multicellular organism, the nematode *Caenorhabditis elegans*. **Then fruit fly.**
- Determination of the base sequence in the human genome was initiated in 1990.
- The initial draft was published in 2004.
- The complete sequence was published in August 2023.

Major findings

- The number of protein-coding genes is less than 20,000.
 - Many are common among other species like yeast, drosophila, and C. elegans, but others are unique.
 - The number of regulatory elements is significant (more than 30% of the genome). ➤ **This indicates how important regulation is.**
 - The non-coding genes (transcribed but not translated) such as microRNA and long noncoding RNA appear to be relevant (not mere noise)
 - **The portion of the human DNA that is transcribed is 75%, Most of the DNA creates non-coding RNA molecules (RNA that is transcribed but not translated).**
- **The Same number of genes as in mice.**
 - **Covers less than 2% of the human genome.**

Not to be memorized.

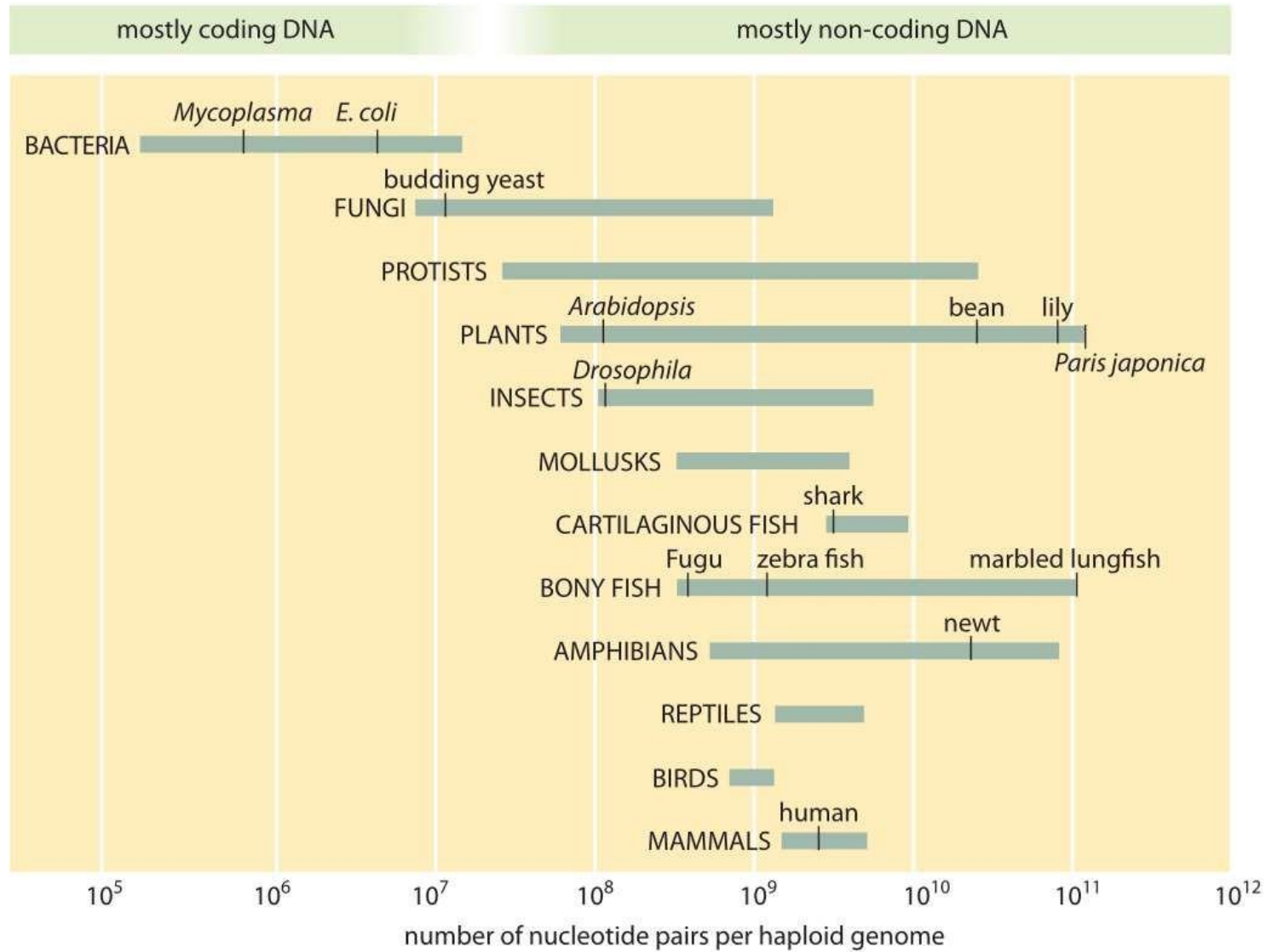
organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria <i>E. coli</i>	4.6 Mbp	4,300	1
budding yeast <i>S. cerevisiae</i>	12 Mbp	6,600	16
amoeba <i>D. discoideum</i>	34 Mbp	13,000	6
nematode <i>C. elegans</i>	100 Mbp	20,000	12 (2n)
fruit fly <i>D. 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)
viruses			
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
organelles			
mitochondria - <i>H. sapiens</i>	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
chloroplast - <i>A. thaliana</i>	150 kbp	100	1
eukaryotes - multicellular			
dog <i>C. familiaris</i>	2.4 Gbp	19,000	40
chimpanzee <i>P. troglodytes</i>	3.3 Gbp	19,000	48 (2n)

Protein-coding genes are almost the same between human and mice.

Even between human and chimpanzee.

Nucleotides per genomes

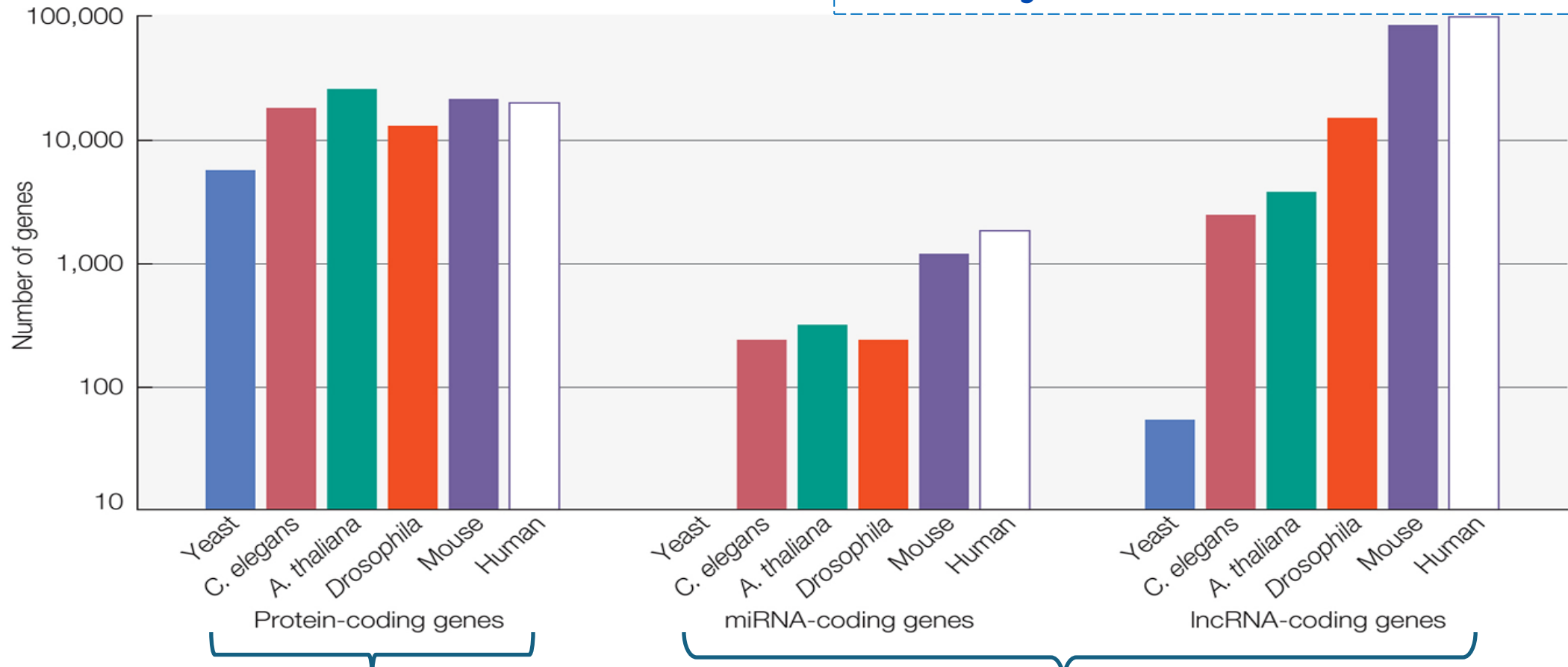
Not to be memorized indeed



Just pay attention how human genome is less than that of certain species of amphibians, bony fish and plants.

Noncoding RNAs and organismal complexity

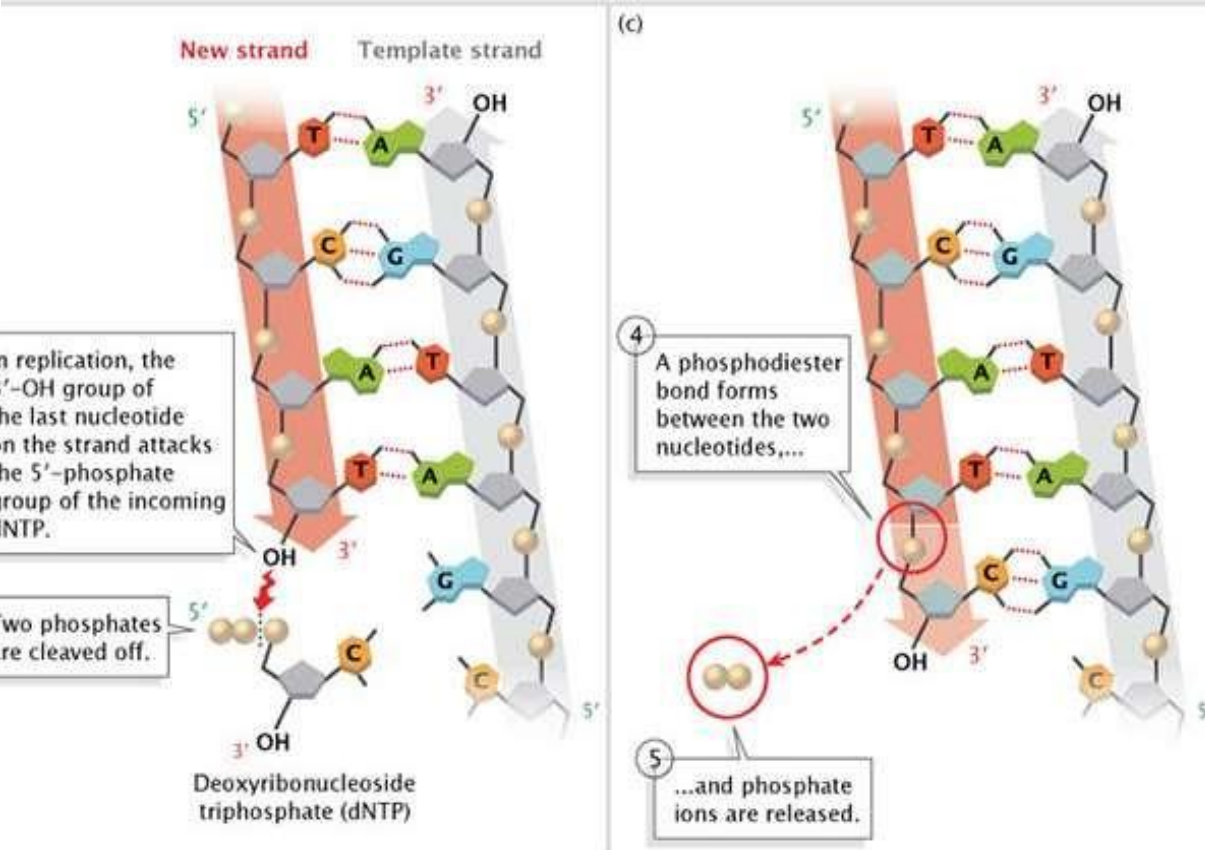
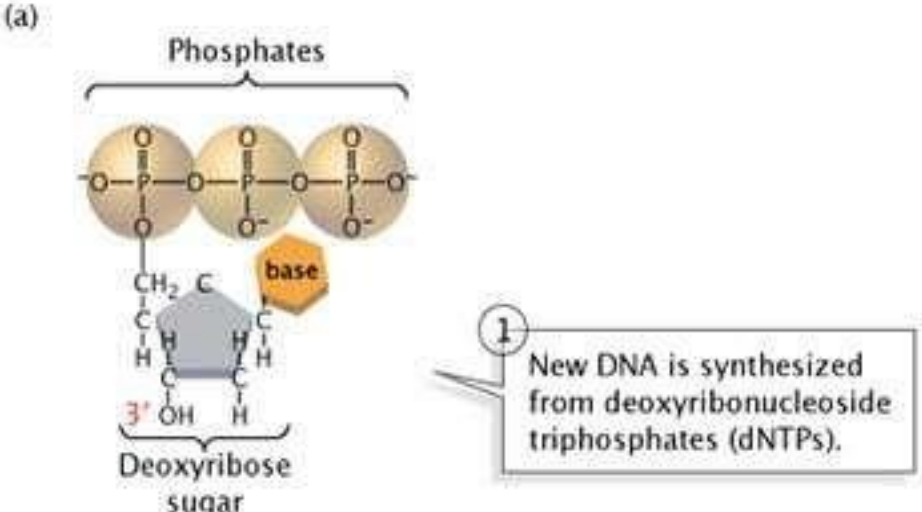
➤ Recall that the gene is defined as a sequence of DNA that is necessary for the synthesis of polypeptide, or functional RNA including, but not limited to, tRNA, rRNA, miRNA and lncRNA.



➤ One can see that there is, to some extent, similarity among different complex; like human and mice, and primitive; like fruit fly and yeast, organisms regarding protein-coding genes number.

➤ miRNA-coding or lncRNA-coding genes which indeed encode miRNA and lncRNA respectively, but not to be translated, the number of which reflects the complexity of the organisms. They are present in large number in human and mice compared to other primitive organisms. This gives us a sign that these molecules play a significant role in our genome, and that is why they seem to be relevant.

DNA synthesis/elongation



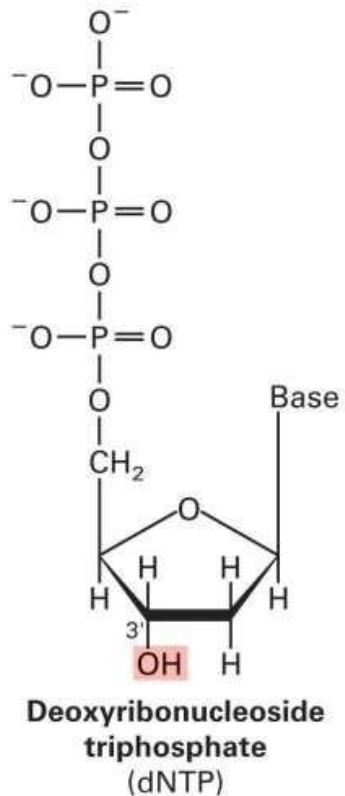
Please take a closer look at this figure for better comprehension.



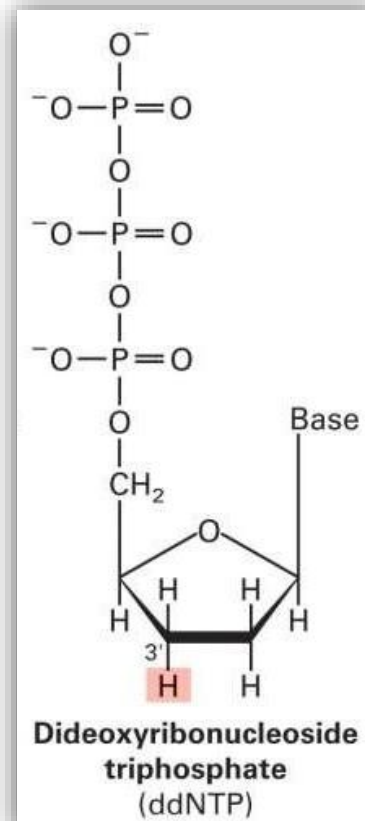
The basic method of DNA sequencing

- The most popular method is based on premature termination of DNA synthesis by dideoxynucleotides.

This particular methodology is called Sanger method.



➤ **This is Normal substrate of DNA polymerase; deoxyribonucleoside triphosphate (dNTP), where it is deoxygenated at carbon (2).**



➤ **dideoxyribonucleoside triphosphate (ddNTP) on the other hand, it is additionally, as the name suggests, deoxygenated at carbon (3) besides carbon (2), this means that once added to the growing strand during DNA synthesis, it terminates the process, because then the growing terminus bears no hydroxyl group on carbon (3) to add the following subunit, ceasing DNA synthesis prematurely.**

The process...

- DNA synthesis is initiated from a primer that has been labeled with a radioisotope.
- Four separate reactions are run, each including deoxynucleotides plus one dideoxynucleotide (either A, C, G, or T)
- Incorporation of a dideoxynucleotide stops further DNA synthesis because no 3-hydroxyl group is available for addition of the next nucleotide

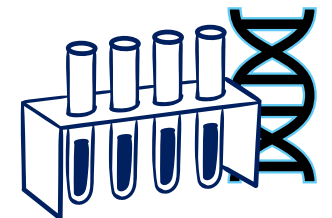
Generation of fragments

- A series of labeled DNA molecules are generated, each terminated by the dideoxynucleotide in each reaction
- These fragments of DNA are then separated according to size by gel electrophoresis and detected by exposure of the gel to X-ray film
- The size of each fragment is determined by its terminal dideoxynucleotide, so the DNA sequence corresponds to the order of fragments read from the gel

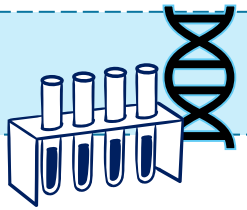
Please see the next few slides for more clarification.

Elucidation of the Sanger sequencing method

- Frederick Sanger sequencing method is based on the utilization of dideoxynucleoside triphosphate (ddNTP) for the premature termination of DNA synthesis.
- First of all, DNA sample is taken through saliva, buccal swap or blood sample, this DNA is indeed double-stranded, therefore it must be separated or denatured so that we have single-stranded DNA to be sequenced.
- Four tubes are brought, inside each, there are single-stranded DNA molecules (templates), DNA polymerase enzymes, certain radiolabeled primers, deoxyribonucleoside triphosphate (dNTPs; dGTP, dCTP, dTTP and dATP all together in each tube) with only little amount of specific type of dideoxynucleoside triphosphate (only one of the four ddNTPs in each of the four tubes) as well. The use of specific primer, **recall that DNA polymerase can not synthesize the complementary strand de novo unlike RNA polymerase**, indicates that some sequence on the DNA molecule is previously known so that we can chose the proper primer that is complementary to the DNA template.



Further Elucidation



- DNA polymerase, in the tube that contains **ddGTP** for instance, starts to synthesize the complementary strand on the template after the synthesis is initiated from the radiolabeled primers.
- as synthesis continues, DNA polymerase will encounter C nucleotide on DNA template, and it might thereupon add either **dGTP** (normal substrate) or **ddGTP** to the 3-prime terminus of the growing strand (complementary).
- the latter causes termination of the synthesis, however and for the most part, **dGTP** will be incorporated since **ddGTP** is present in low concentration in the tube as previously mentioned. one can conceive that it is actually a matter of probability.
- Imagine if we have **1000 DNA templates** with their hybridized primers in the tube before the reaction takes place, reaction starts, the polymerase, for instance, reads C nucleotide on the template and catalyzes the addition of either **dGTP** or **ddGTP**, consequently, **900 molecules** would have normal **dGTP** while the remaining **100 ones** would have **ddGTP**, the synthesis only proceeds for those **900 molecules**, the process continues until another C is encountered on the template, thereupon, out of the **900 molecules**, **800 ones** are incorporated by **dGTP**, and the remaining **100 molecules** with **ddGTP**, one can notice the pattern here. Eventually, there will be many radiolabeled prematurely-synthesized DNA fragments that vary in their length or size in each tube. Ratios mentioned in this text are arbitrary.

DNA synthesis of the complementary strand proceeds from 5' to 3'.

5' TAGCTGACTC3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...

DNA polymerase
+ dATP, dGTP, dCTP, dTTP
+ **ddGTP** in low concentration

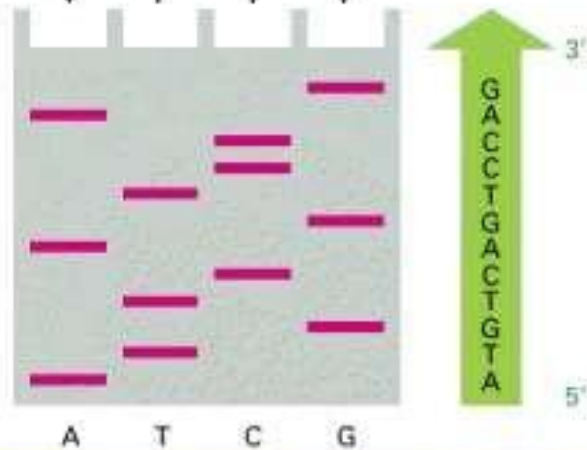
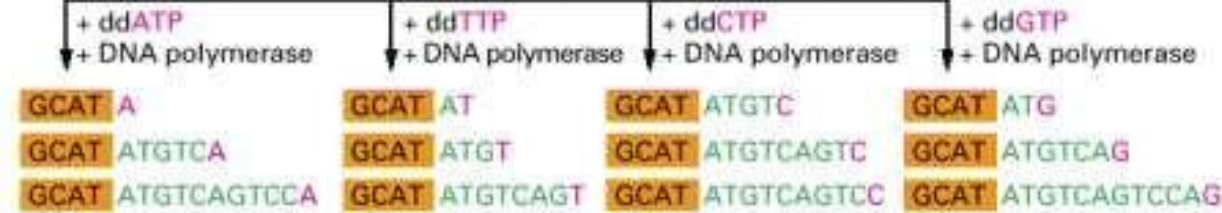
5' TAGCTGACTCA**G**3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...
+
5' TAGCTGACTCAGTTCTT**G**3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...
+
5' TAGCTGACTCAGTTCTTGATAACCC**G**3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...

And the process continues....

(C)



+ excess dATP
dTTP
dCTP
dGTP



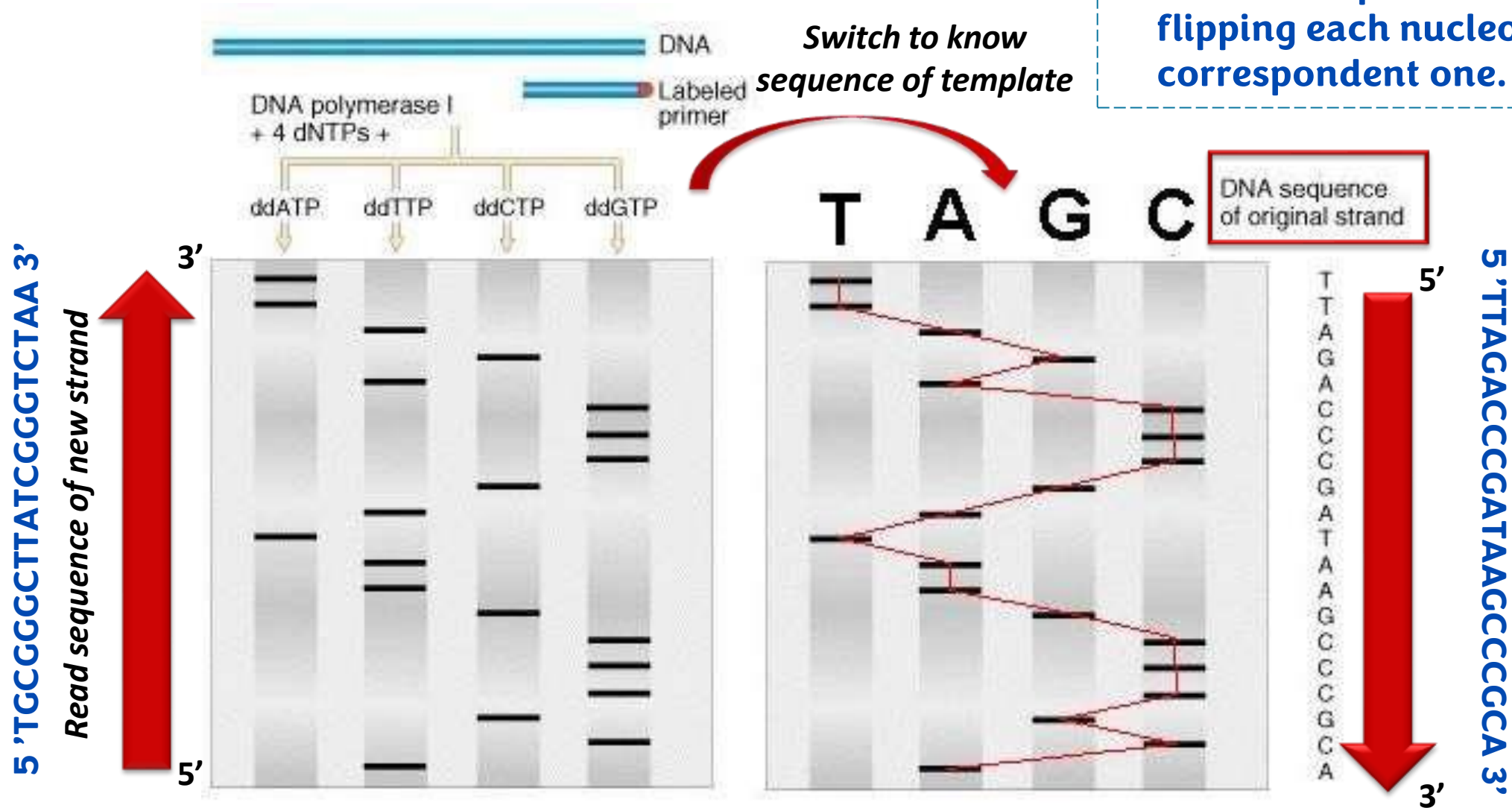
DNA sequence reading directly from the bottom of the gel upward, is
ATGTCAGTCCAG
1 12

Please take a closer look at this figure for better comprehension.

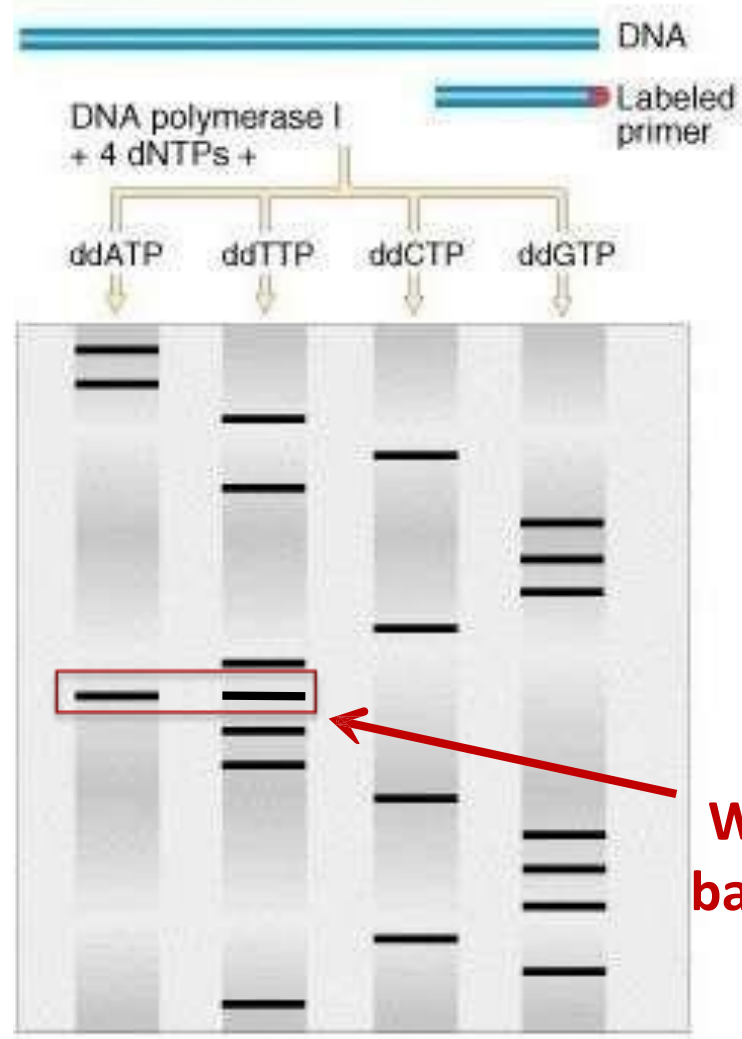
- This is the last step, DNA fragments can be separated according to their size by gel electrophoresis that has special gel with high resolution, meaning that, it can separate DNA fragments even with one nucleotide difference. Each tube is added separately to the wells, so each column corresponds to its nucleotide.
- the order of the radiolabeled fragments (or bands), as they migrate through the gel, reflects the actual DNA sequence (from 5' to 3') of the complementary strand reading from the bottom upward.

Please see next slide for more clarification.

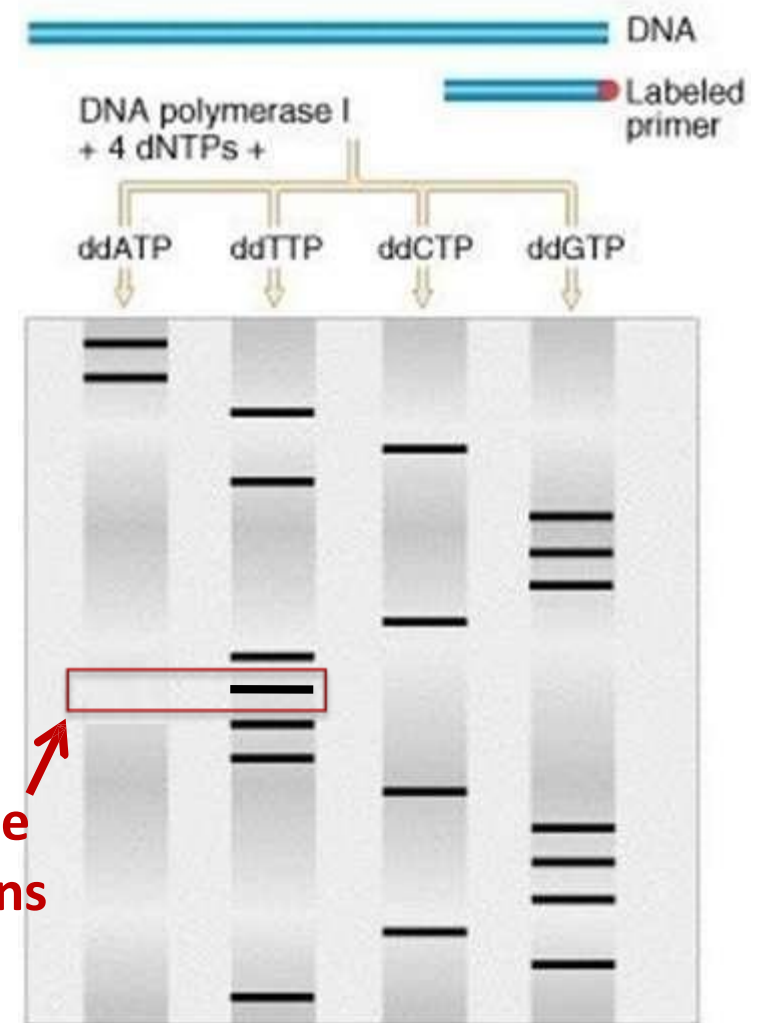
➤ We can read the complementary strand from the gel, and since DNA is antiparallel, template is read from the top downward after only flipping each nucleotide to its correspondent one.



Case 1



Case 2



What do the
band patterns
mean?

Please see next slide for
explanation.

Further Elucidation

➤ Case 1:

- ✓ Whenever we have 2 fragments (bands) line up in the same level, this indicates the person is heterozygous at this particular sequence or gene.
- ✓ Since we are diploid, that is, we have homologous chromosomes, one is maternally inherited while the other is paternally inherited in each pair of chromosomes.
- ✓ Therefore it could be single nucleotide polymorphism (SNP) that would account for this phenomenon, consequently, the maternal allele might bear (T) nucleotide at some position along the sequence, meanwhile the paternal one bears (A) nucleotide at the same position.

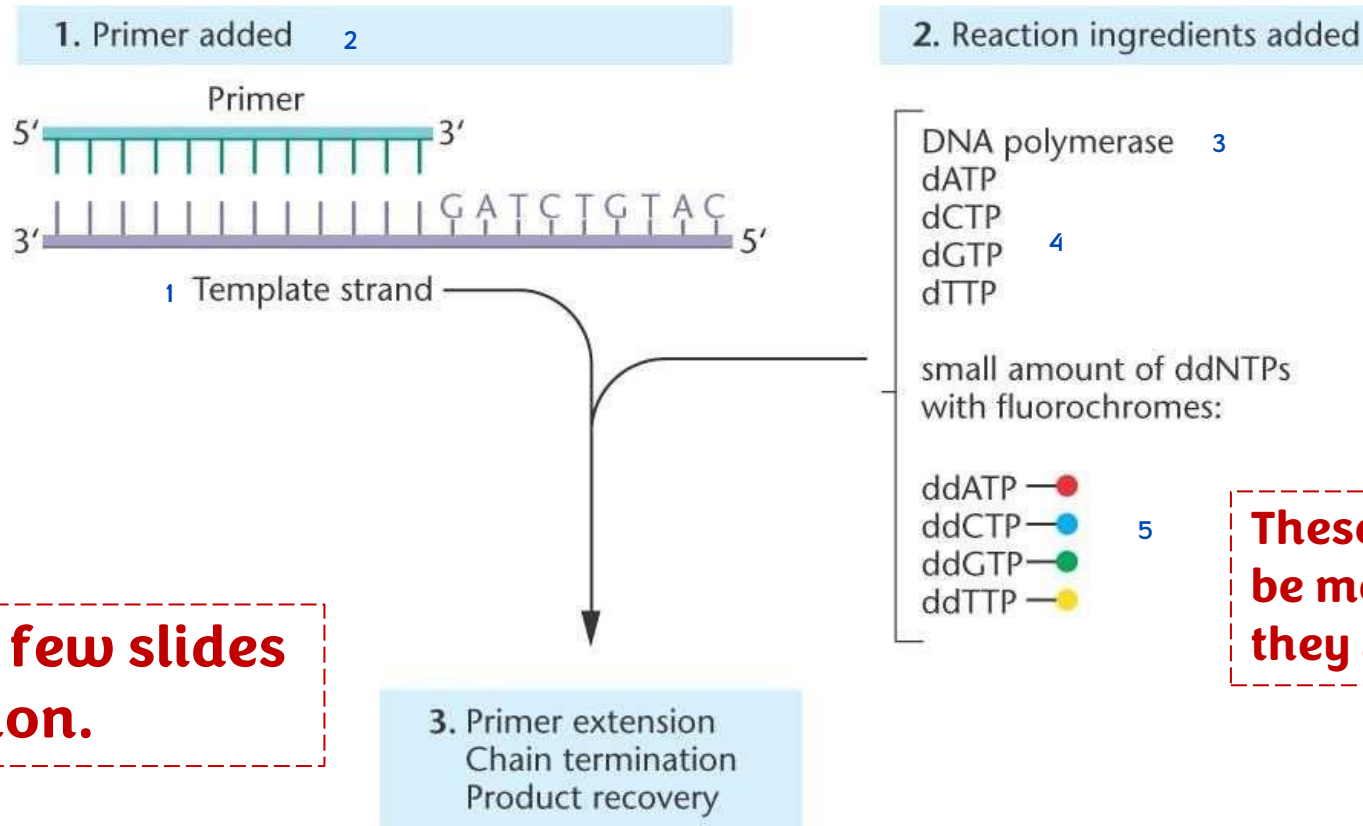
➤ Case 2:

- ✓ comparing to the previous case results, we would expect that there is a band in column (1) lining up with the other one, however this is not the case here, This person has (T) nucleotide in both alleles, and that means that the person is Homozygous due to polymorphism or mutation.



Fluorescence-based DNA sequencing

- Reactions include the four deoxynucleotides plus **the four dideoxynucleotides in the same reaction** with each ddNTP labeled with a unique fluorescent tag.



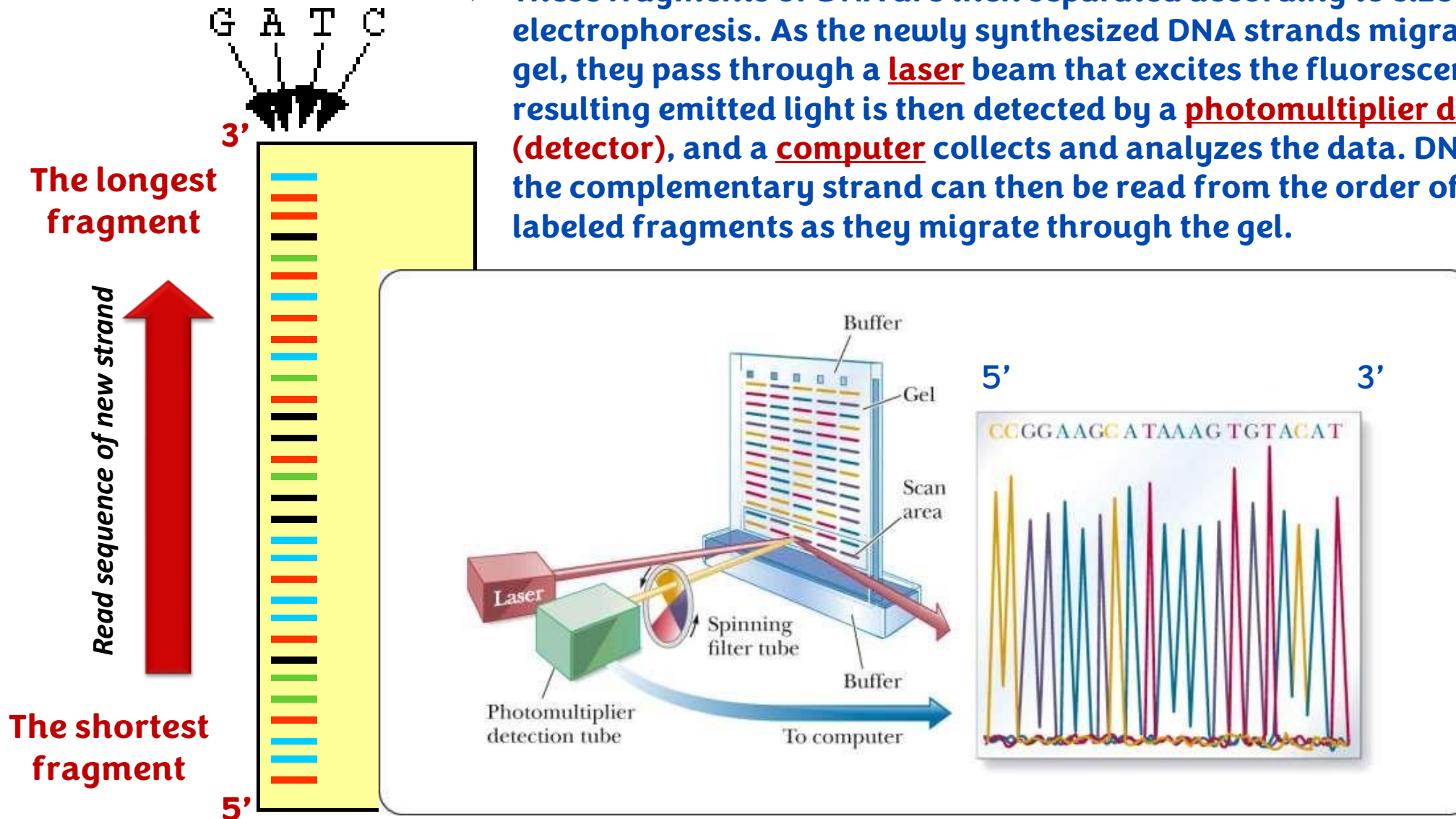
These colors are not to be memorized, because they are arbitrary.

Please see the next few slides for more clarification.

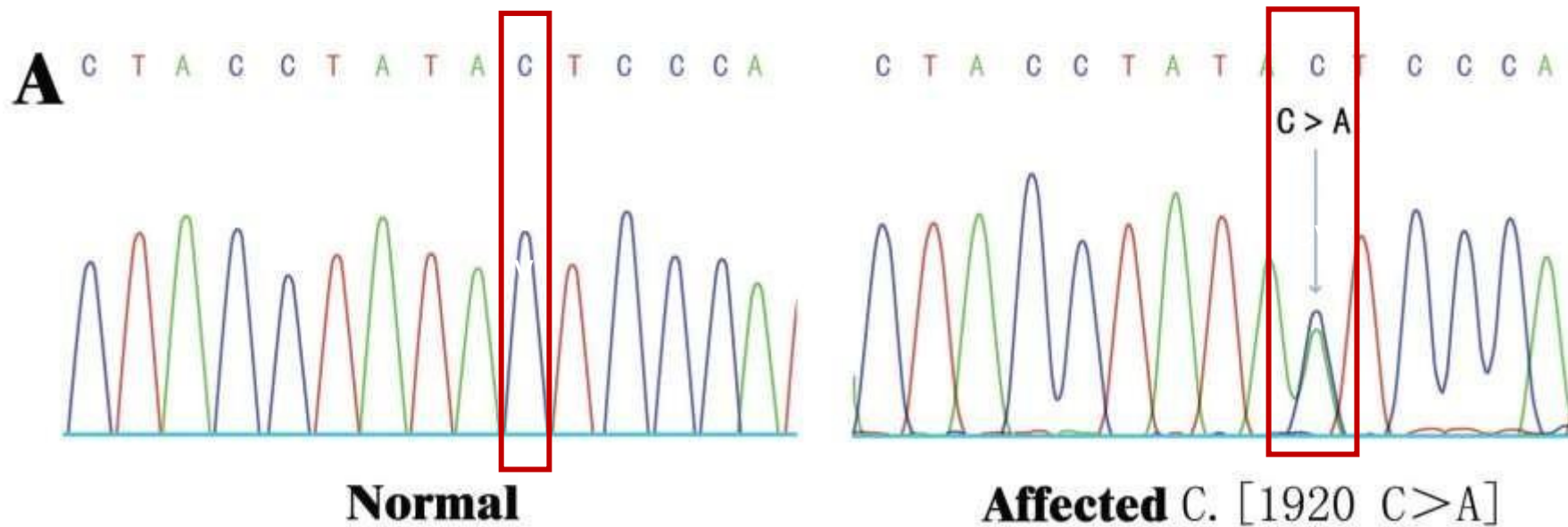
Elucidation of the Fluorescence-based sanger DNA sequencing

- The previous basic sanger method is not actually safe due to the radioactivity. Instead, safer and better approach is carried out. Both depend on the incorporation of **ddNTPs** and the premature termination of DNA synthesis, however the primer in this technique is not radiolabeled, but rather the **ddNTPs** are fluorescent-labeled, each emits a particular color or signal to be detected
- unlike the former method where we have 4 separated reactions, here we have only one; involving single-stranded DNA molecules (templates) to be sequenced, DNA polymerase enzymes, certain primers, mixtures of deoxyribonucleoside triphosphate molecules (**dNTPs**) with only little amount of dideoxyribonucleoside triphosphate molecules as well (**ddNTPs**).
- The same principle applies here, DNA synthesis is initiated from the primers, If the enzyme, for instance, reads (A) nucleotide on the template, it will insert either **dTTP** or **ddTTP**, however, **dTTP** is most likely to be added. Once **ddTTP** is incorporated, synthesis ceases, and this process continues as such whenever the enzyme read A, T, C, or G nucleotides.
- Eventually, we will obtain prematurely-terminated fluorescent-labelled DNA fragments that differ from one another by only one nucleotide, ready to be electrophoresed, detected and visualized.

- Instead of utilizing **x-ray film** to visualize radioactive signals in the previous approach, here its automated (computerized).
- These fragments of DNA are then separated according to size by gel electrophoresis. As the newly synthesized DNA strands migrate through the gel, they pass through a **laser** beam that excites the fluorescent tags. The resulting emitted light is then detected by a **photomultiplier detection tube (detector)**, and a **computer** collects and analyzes the data. DNA sequence of the complementary strand can then be read from the order of fluorescent-labeled fragments as they migrate through the gel.

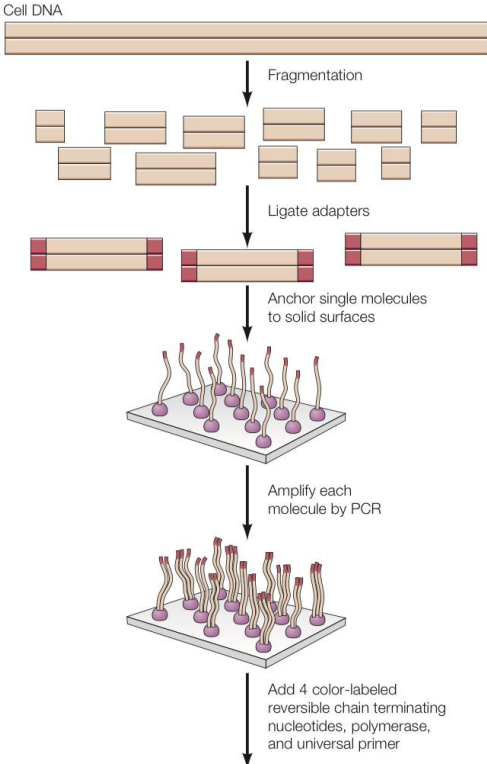


- Notice below to the right how we have two overlapping peaks corresponding to (A) and (C) nucleotides instead of the normal variant **CC**, this indicates that this person is **heterozygous AC** and has a mutation on one of the 2 alleles.
- If both alleles are mutated, we would obtain a single peak corresponding to (A) nucleotide (not shown in the figure) , indicating that the person is **homozygous AA**.
- Although Sanger sequencing approach is still used until now, it is unfortunately time-consuming, and not cost-effective.

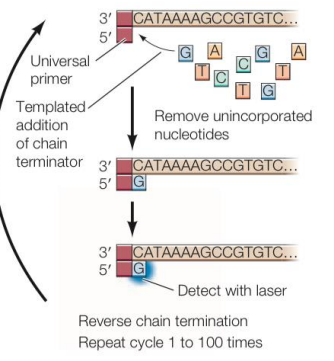


Next-generation sequencing

two main stages involved in this approach; cluster formation by PCR-like technique, and sequencing by synthesis using polymerase and special nucleotides. This is the general theme.



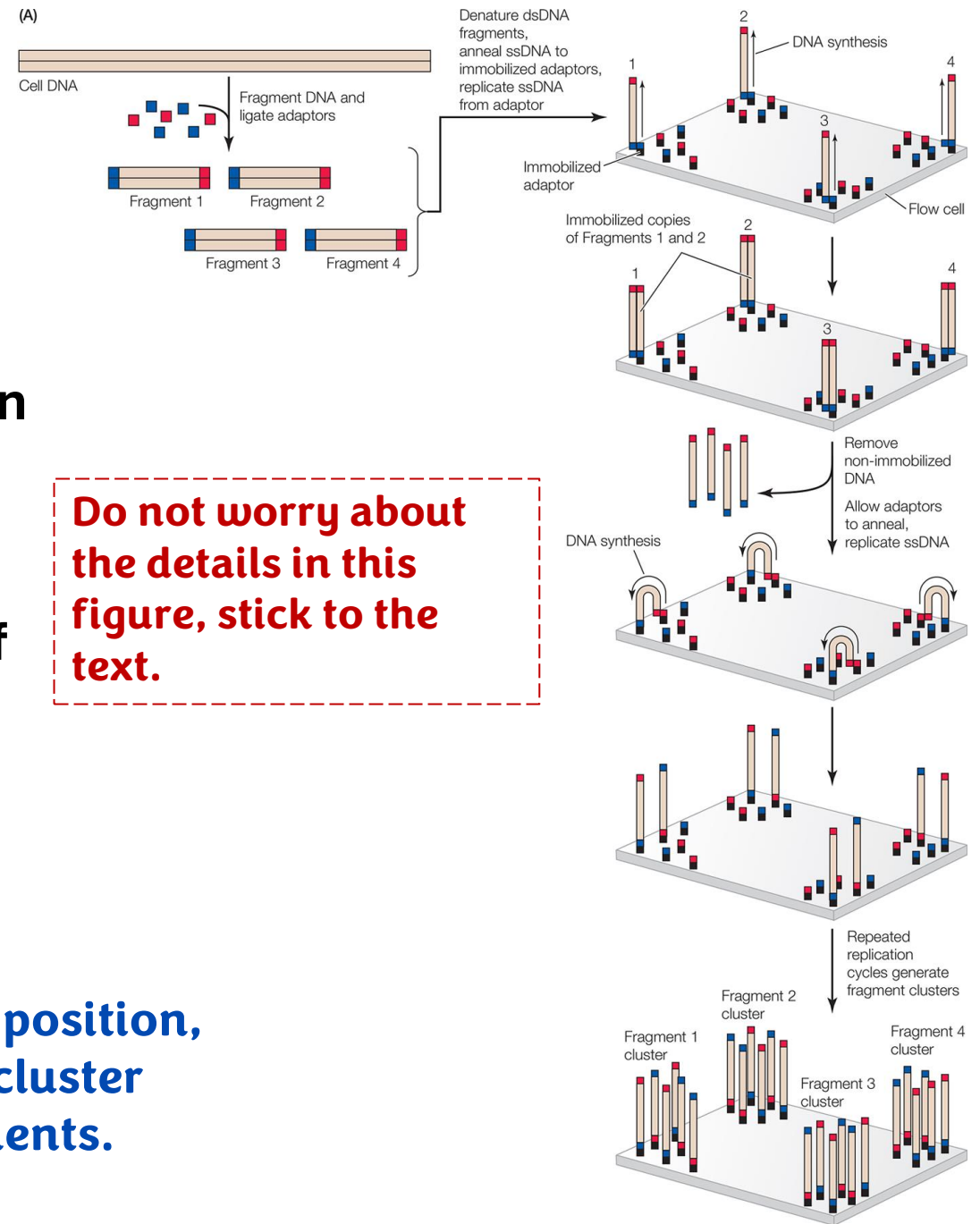
- Cellular DNA is fragmented. ➤ Random fragmentation leads to double-stranded fragments that are overlapping.
- DNA adapters are added to the ends of each DNA fragment. ➤ And then denatured to single-stranded fragments.
- Each DNA fragment is attached to a solid surface and amplified like PCR using primers that anneal to the adapter sequences. ➤ This solid surface contains small fragments that are complementary to the adaptors that are bound to the ends of each fragment (to be sequenced) attaching it to different position.
 - The adaptors can also contain sequences that can identify samples (like individual A, individual B, etc.) like bar codes
- Four-color special nucleotides are added and a single nucleotide is incorporated. ➤ these nucleotides are not dideoxynucleotides nor deoxynucleotides.
 - The nucleotides have to be chemically modified to add the following one.
- The color of the incorporated nucleotide is detected by a special camera and it is activated to allow for the addition of the subsequent nucleotide.
- A new nucleotide can then be added to it.
- The cycle is repeated.



cluster formation

Cellular DNA is fragmented, adapters are ligated to the ends of each fragment, and the double-stranded fragments are denatured to single strands. Single-strand fragments are then annealed to complementary adaptors (known as oligos; surface-bound adaptors complementary to the adaptors bound to the ends of the DNA fragments to be sequenced) that are immobilized on the surface of a flow cell and amplified by PCR, forming fragment clusters (many clusters each contains many identical DNA fragments), each originating from a distinct DNA fragment.

- Each single-stranded fragment is amplified at each position, forming a cluster using a PCR-like technique. Each cluster contains hundreds of the same identical DNA fragments.



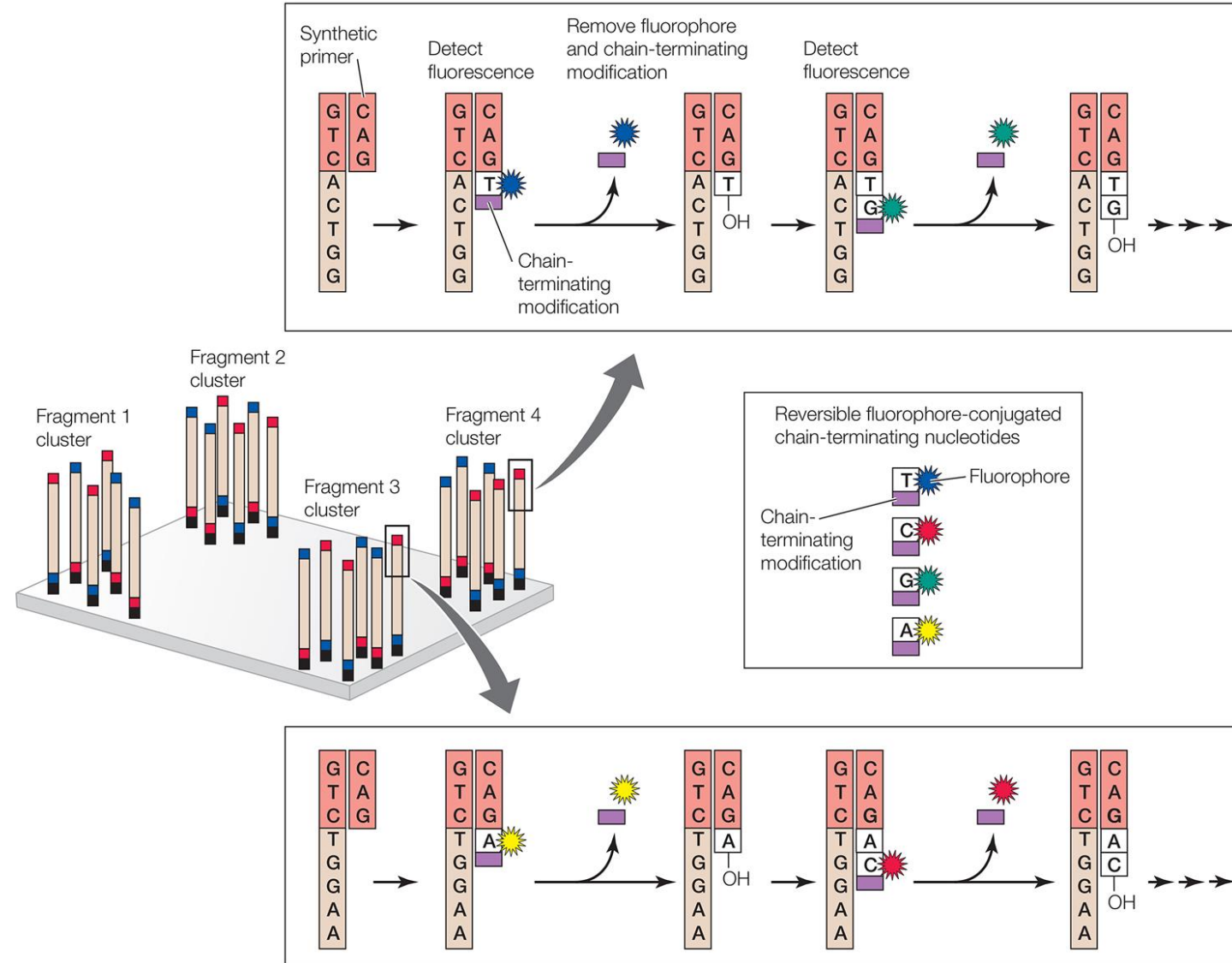
Do not worry about the details in this figure, stick to the text.

Sequencing by synthesis

Sequencing by synthesis then determines the nucleotide sequence within each cluster, which uses four reversible fluorophore-labeled, chain-terminating nucleotides that are added one at a time by DNA polymerase and a primer that recognizes the adapter sequence. After addition of each nucleotide, the fluorescent color within each cluster is detected by a laser. The fluorophore and chain-terminating modifications are then removed, and the next nucleotide is added for another cycle. Repeated cycles can determine sequences 50–300 nucleotides long in each cluster.

Please pay attention at every detail in this figure for better comprehension.

(B) Sequencing by synthesis

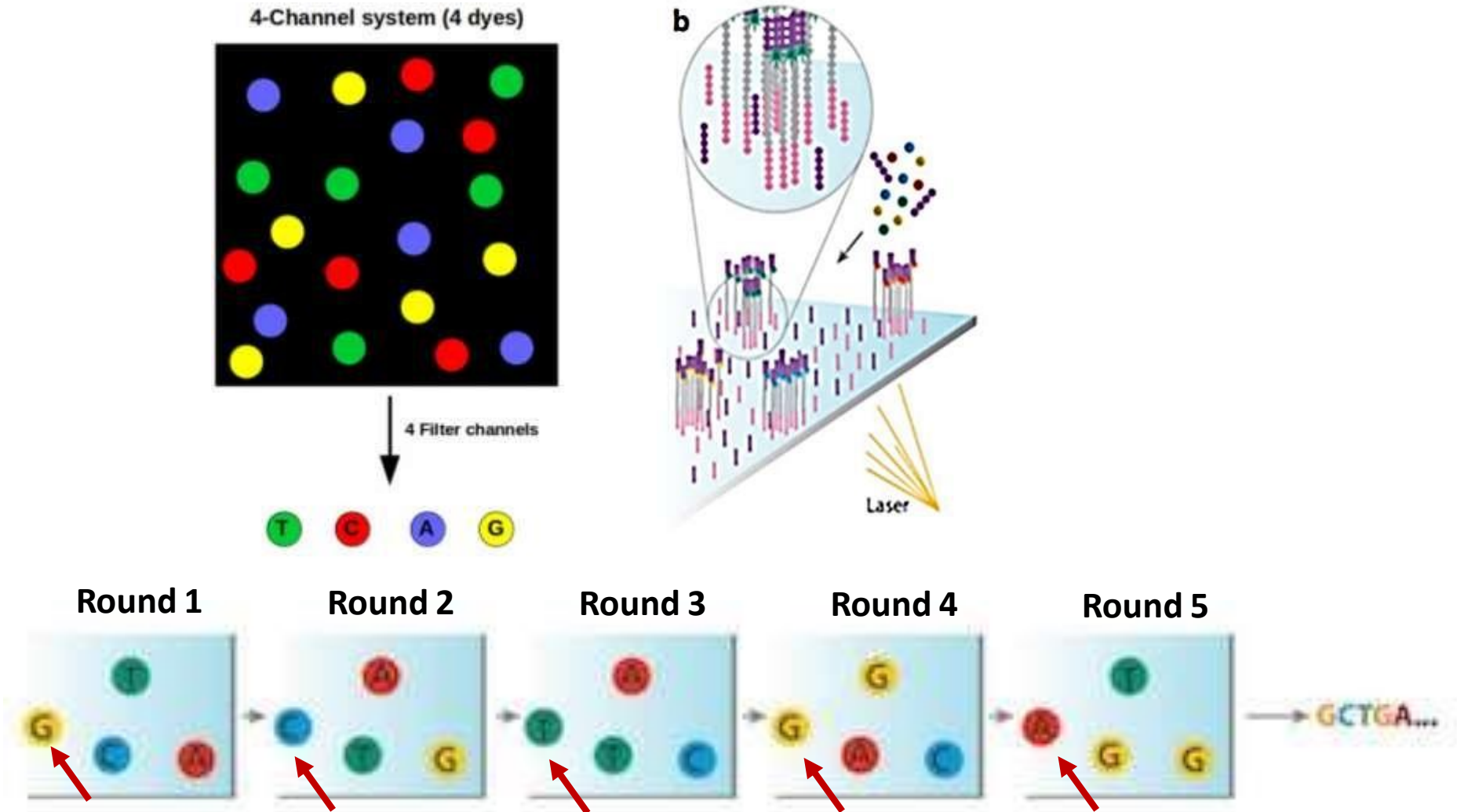


Worthy explanatory notes

- These nucleotides emit specific color once activated by laser. And subsequently have to be chemically-modified via removing **fluorophore** and **chain-terminating modification**, so that the subsequent subunit can be incorporated to the growing chain, and the cycle continues.
- Remember that these nucleotides are incorporated one by one to the DNA fragment each time.
- Each cluster, which contains many identical DNA fragments, emit distinct signal every time a nucleotide is simultaneously inserted to these fragments present in the cluster.
- Therefore, millions of clusters are simultaneously emitting enormous number of distinct signals each time the nucleotides are incorporated to the DNA fragments, all can be detected by a camera.

The detection

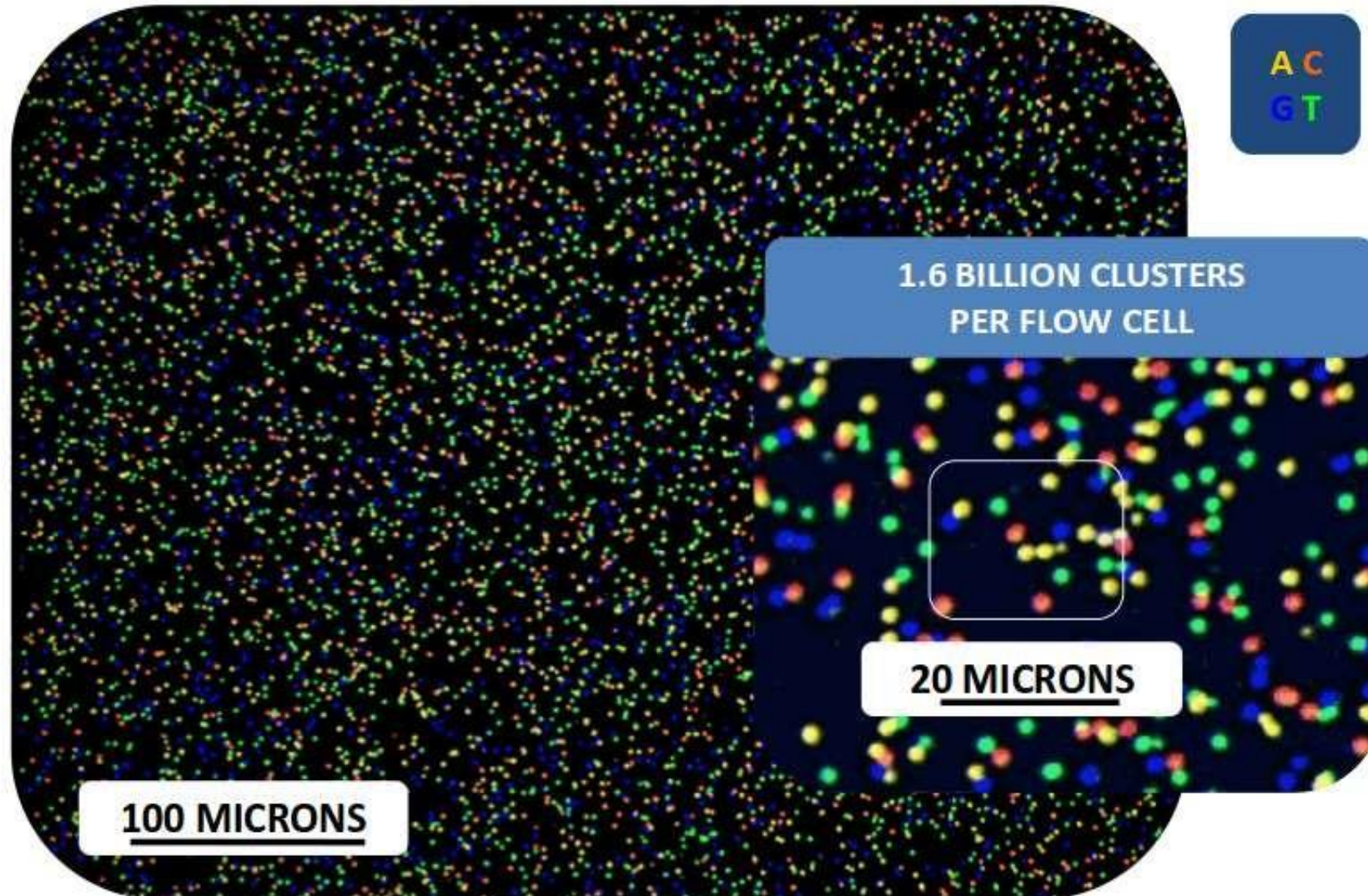
- here we have 4 clusters, you can read DNA fragment sequence in each one through 5 rounds.



- For the cluster shown by the arrow; **5-GCTGA.....-3** and so on.

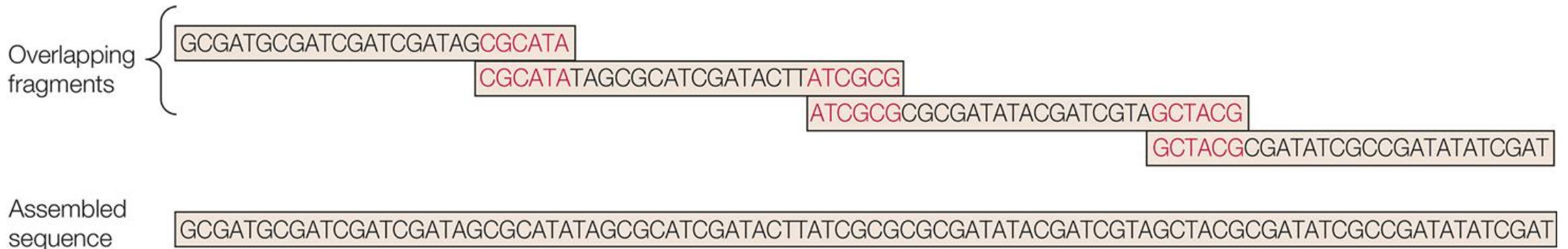
A real look

- **Fireworks!!, millions of clusters, camera can distinguish different signals being emitted from which.**



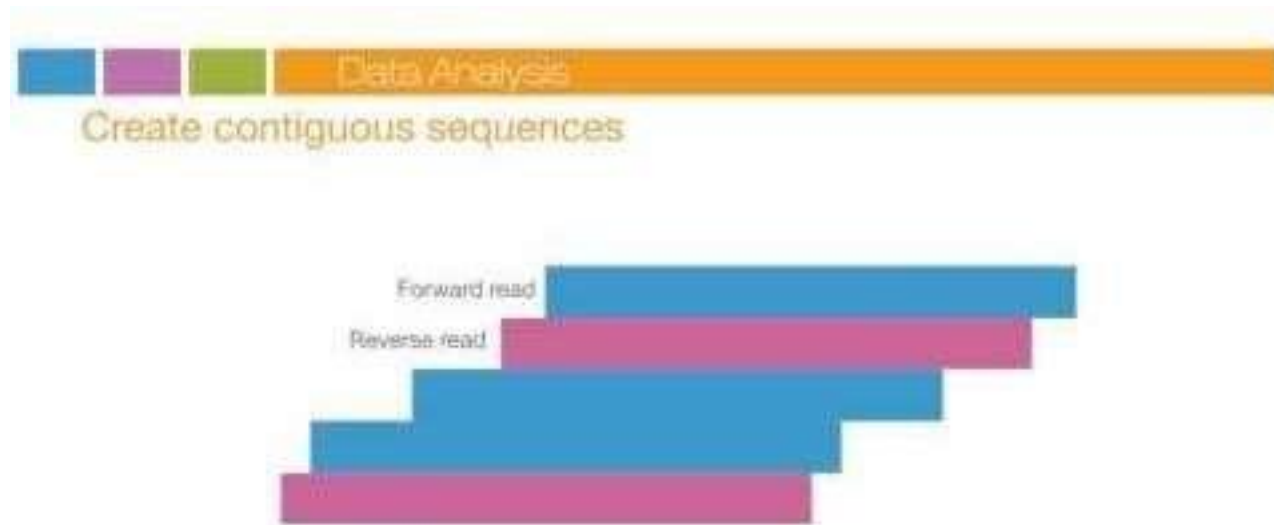
Sequence assembly

- The sequences of millions of fragments are generated.
- They can be assembled into a contiguous sequence by identifying fragments with overlapping sequences.



➤ **Remember that these sequences are overlapping.**

<https://www.youtube.com/watch?v=womKfikWlxM>



**It really does not hurt to watch some extra references!.
Very helpful to comprehend this method.
Disclaimer; not all the details in this video are required.**

For any feedback, scan the code or click on



Corrections from previous versions:

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

Additional Resources:

رسالة من الفريق العلمي:

M.Z.

حفظ الله الأردن قيادةً و وطنًا و شعبًا، حفظ الله
سوريّة و أحرارها و فلسطين و مقاوميهها و لبنان و
أراضيها.
الشام شامنا لو الزمن ضامنا.

