



# 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.
- Importance:
  - Identification of genes and their localization
  - Identification of protein structure and function
  - Identification of DNA mutations
  - Genetic variations among individuals in health and disease
  - Prediction of disease-susceptibility and treatment efficiency
  - Evolutionary conservation among organisms

# DNA sequencing of organism genome



- Viruses and prokaryotes, first
- Human mitochondrial DNA
- The first eukaryotic genome sequenced was that of yeast, *Saccharomyces cerevisiae*.
- The genome of a multicellular organism, the nematode *Caenorhabditis elegans*.
- 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.

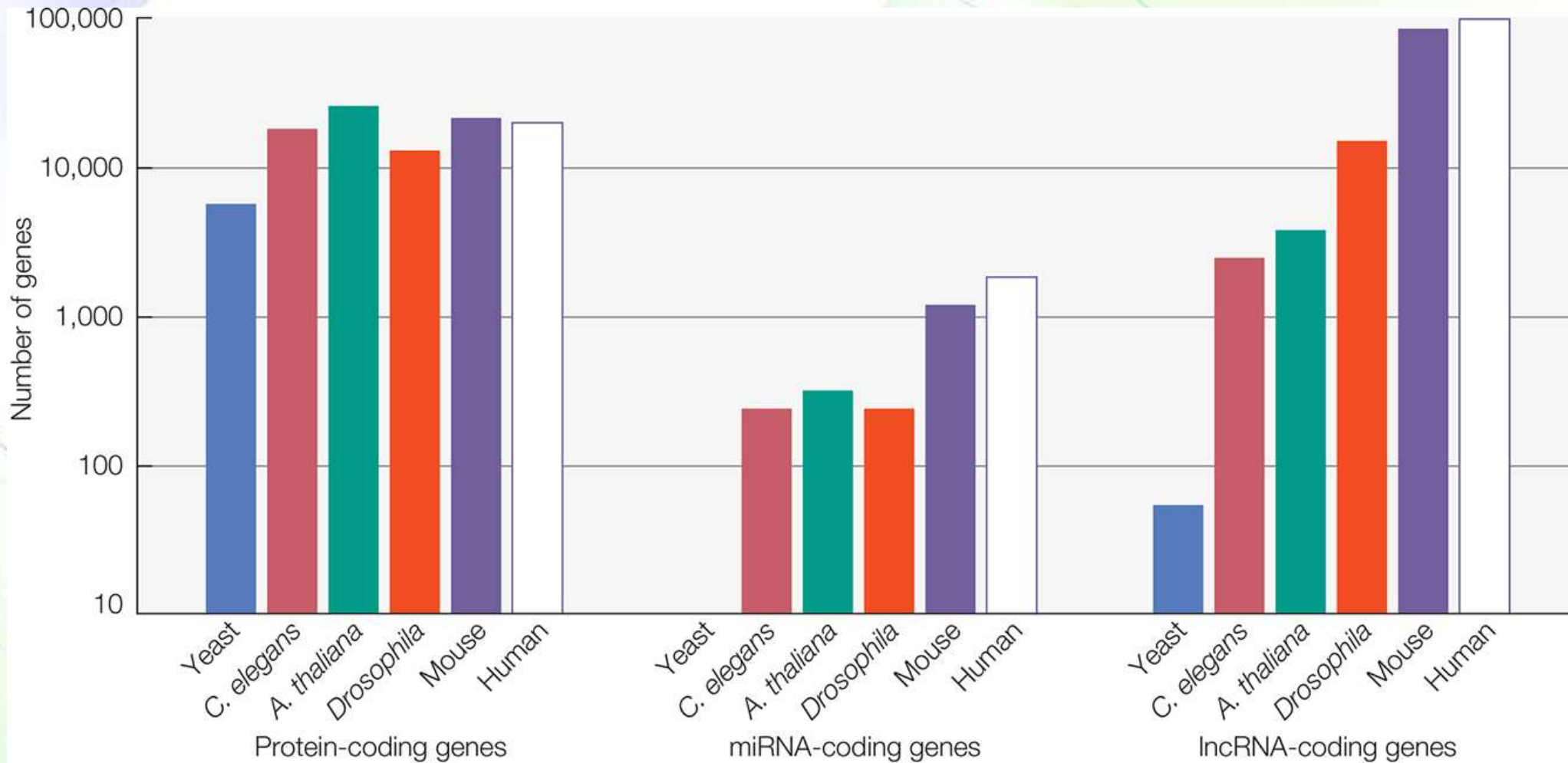


# Major findings



- 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).
- The non-coding genes (transcribed but not translated) such as microRNA and long noncoding RNA appear to be relevant (not mere noise)

# Noncoding RNAs and organismal complexity

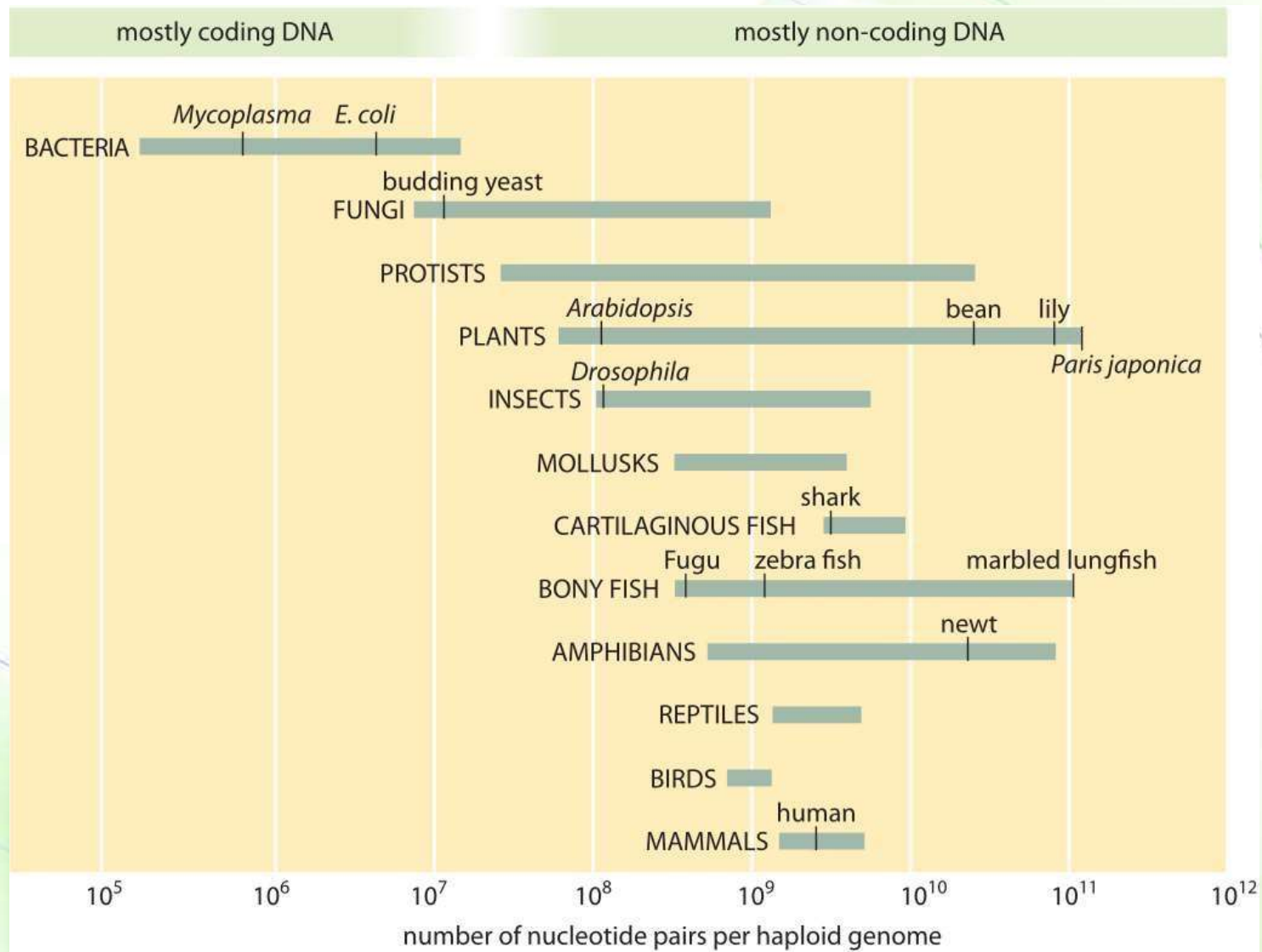




organism	genome size (base pairs)	protein coding genes	number of chromosomes
<b>model organisms</b>			
model bacteria <i>E. coli</i>	4.6 Mbp	4,300	1
budding yeast <i>S. cerevisiae</i>	12 Mbp	6,600	16
amoeba <i>D. discoideum</i>	34 Mbp	13,000	6
nematode <i>C. elegans</i>	100 Mbp	20,000	12 (2n)
fruit fly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



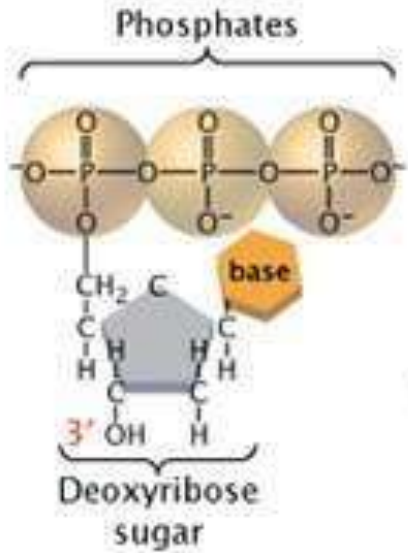
# Nucleotides per genomes



# DNA synthesis/elongation



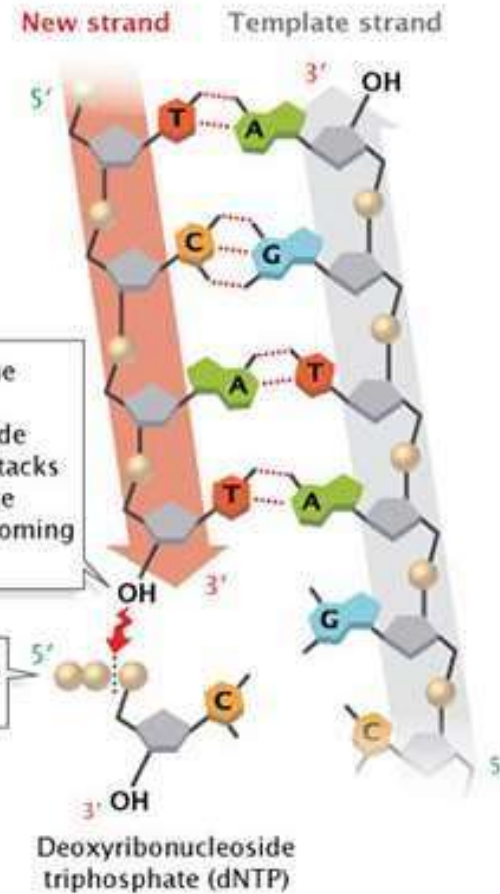
(a)



1 New DNA is synthesized from deoxyribonucleoside triphosphates (dNTPs).

2 In replication, the 3'-OH group of the last nucleotide on the strand attacks the 5'-phosphate group of the incoming dNTP.

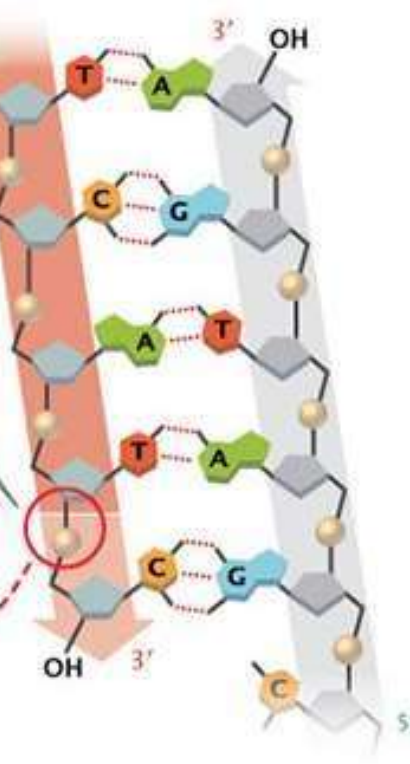
3 Two phosphates are cleaved off.



(c)

4 A phosphodiester bond forms between the two nucleotides,...

5 ...and phosphate ions are released.

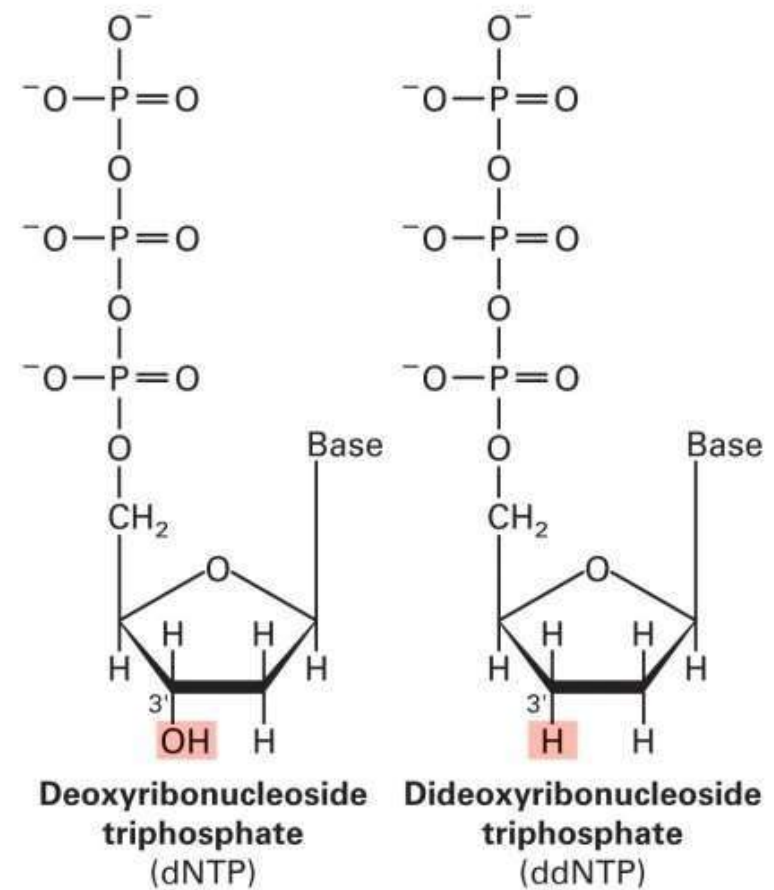




# The basic method of DNA sequencing



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



# 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



5' TAGCTGACTC 3'  
3' ATCGACTGAGTCAAGAACTATTGGGGCTTAA ...

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

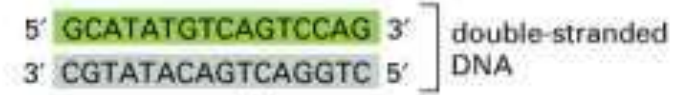
5' TAGCTGACTCA**G** 3'  
3' ATCGACTGAGTCAAGAACTATTGGGGCTTAA ...

+  
5' TAGCTGACTCAGTTCTT**G** 3'  
3' ATCGACTGAGTCAAGAACTATTGGGGCTTAA ...

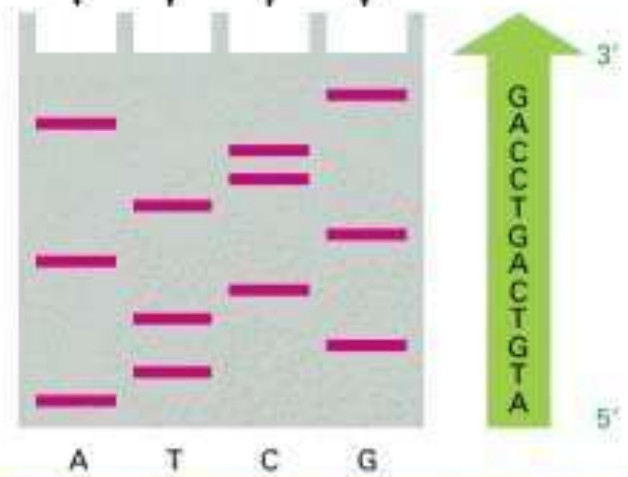
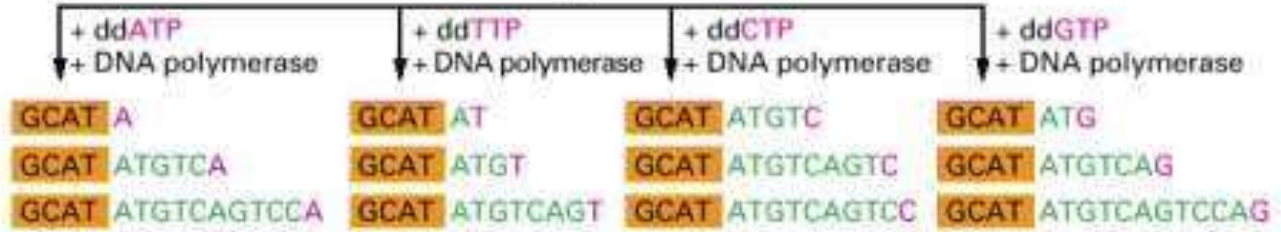
+  
5' TAGCTGACTCAGTTCTTGATAACCC**G** 3'  
3' ATCGACTGAGTCAAGAACTATTGGGGCTTAA ...



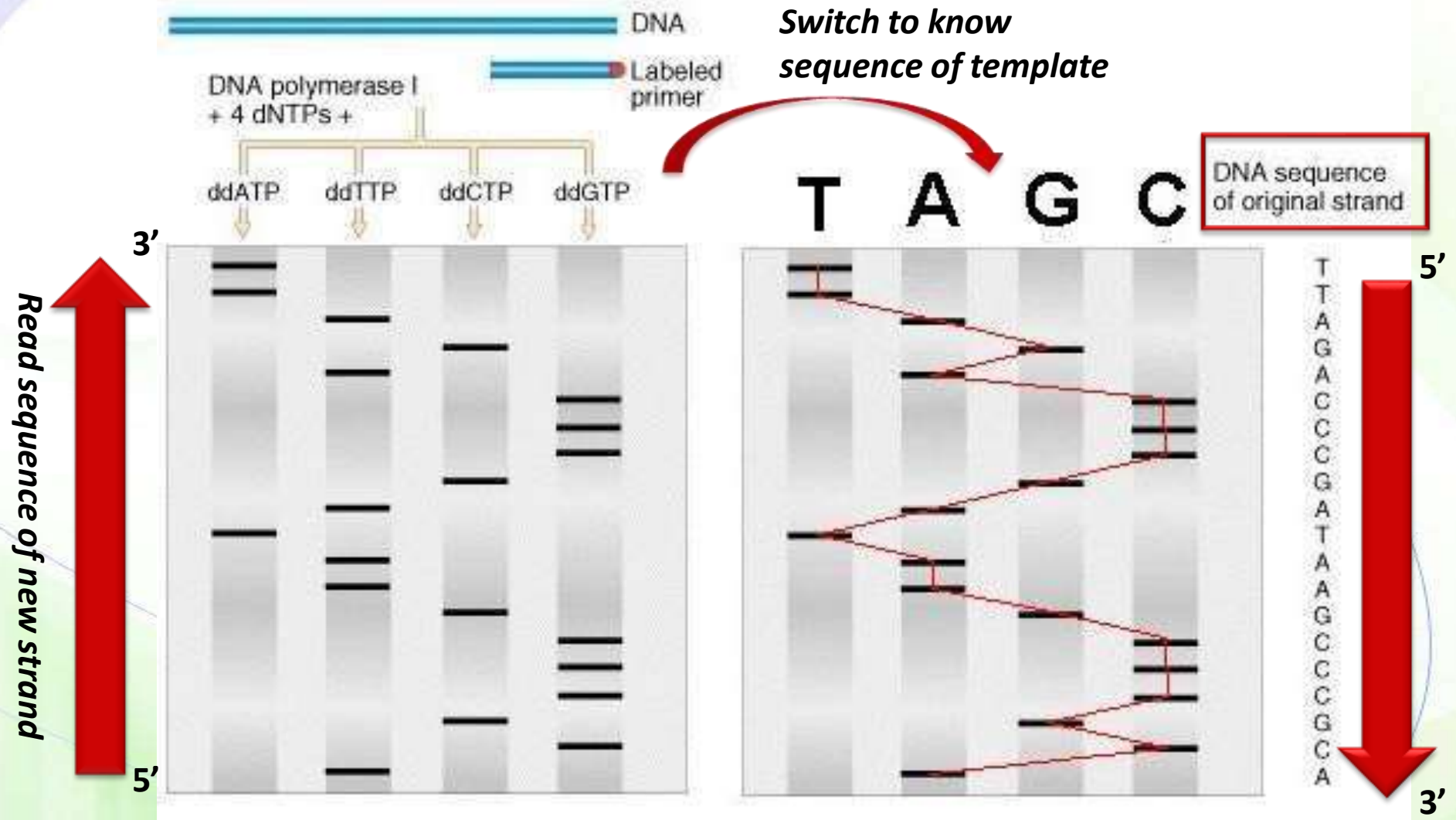
(C)

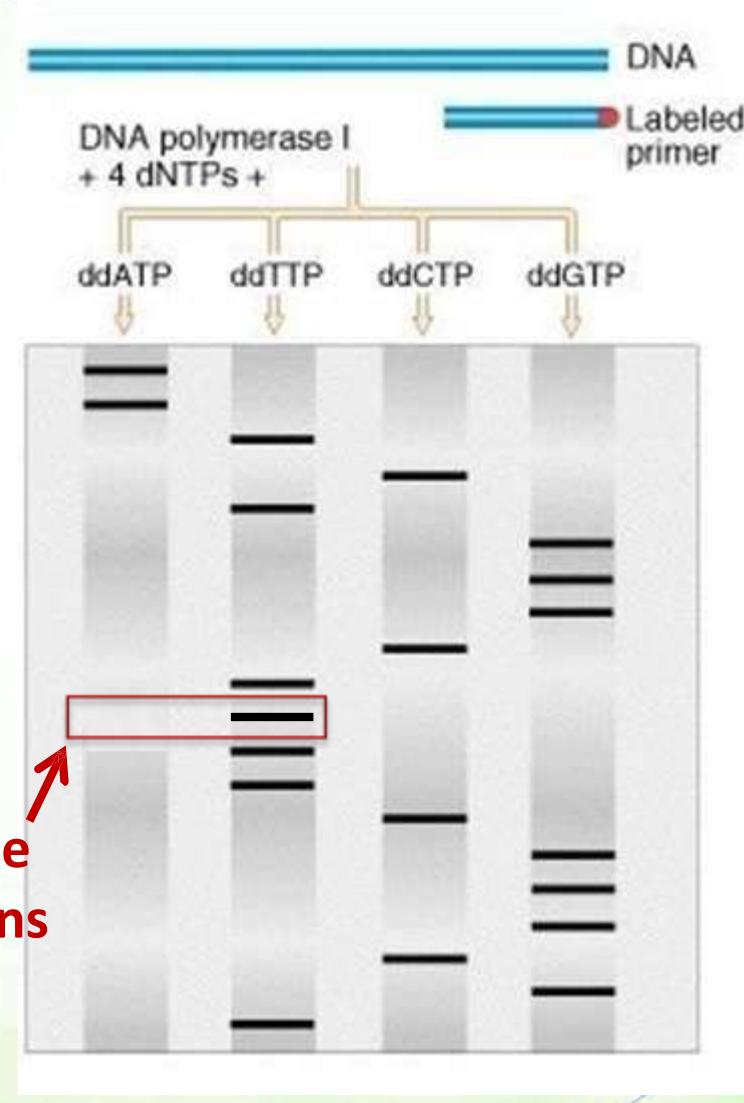
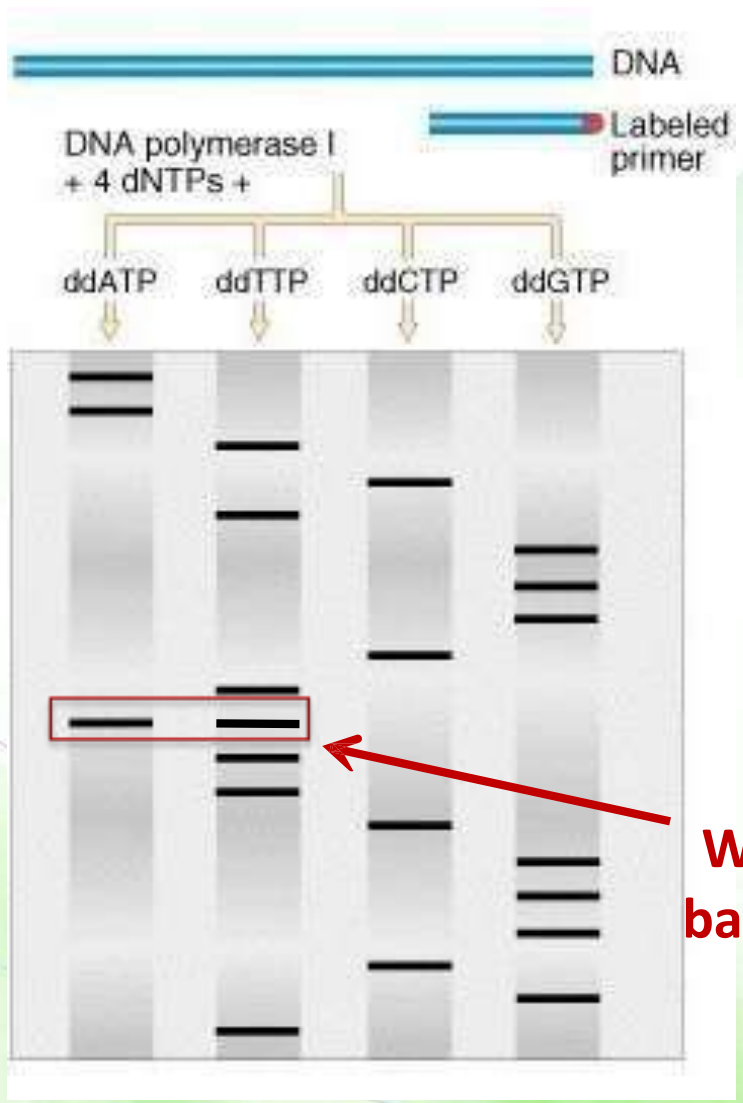


+ excess dATP  
dTTP  
dCTP  
dGTP



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

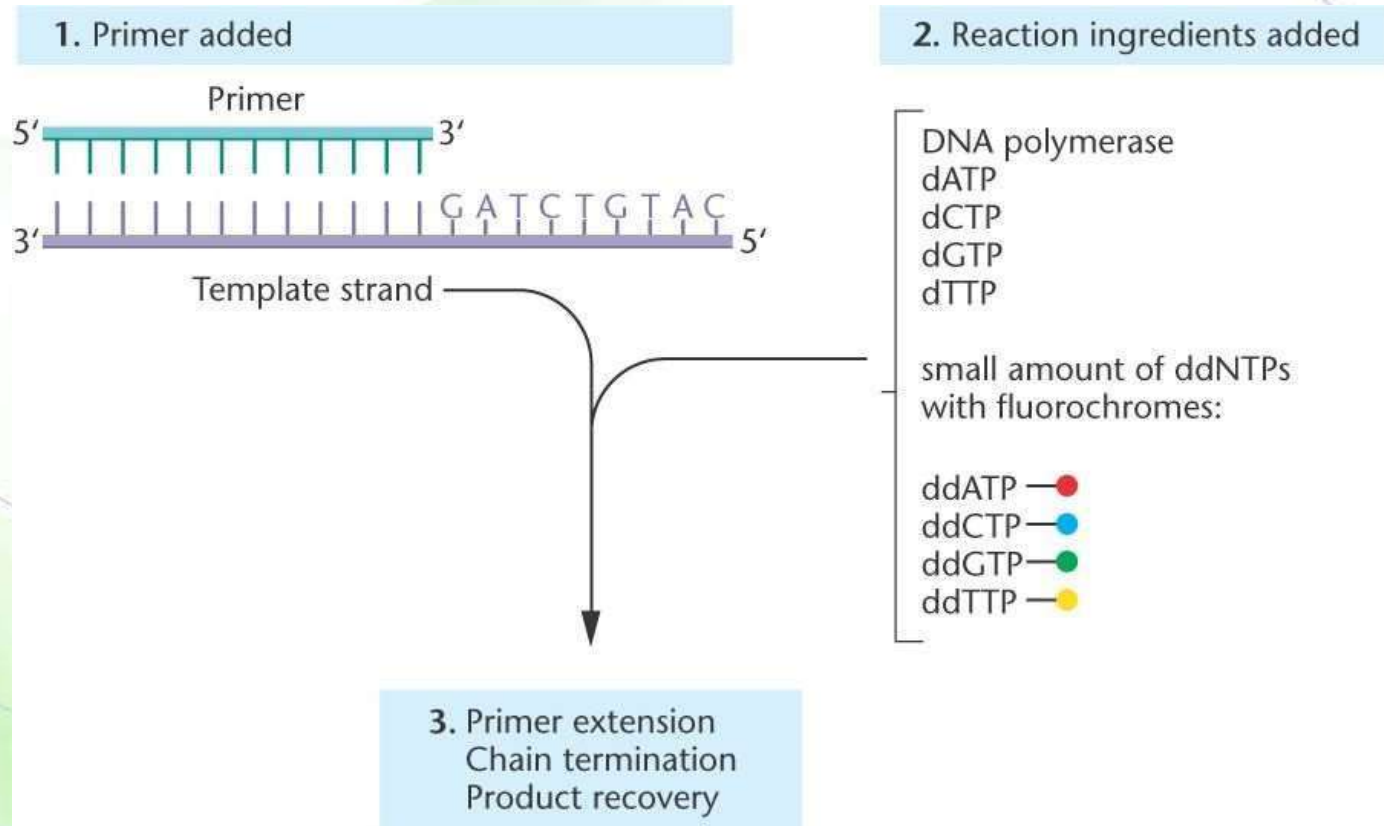




# Fluorescence-based DNA sequencing



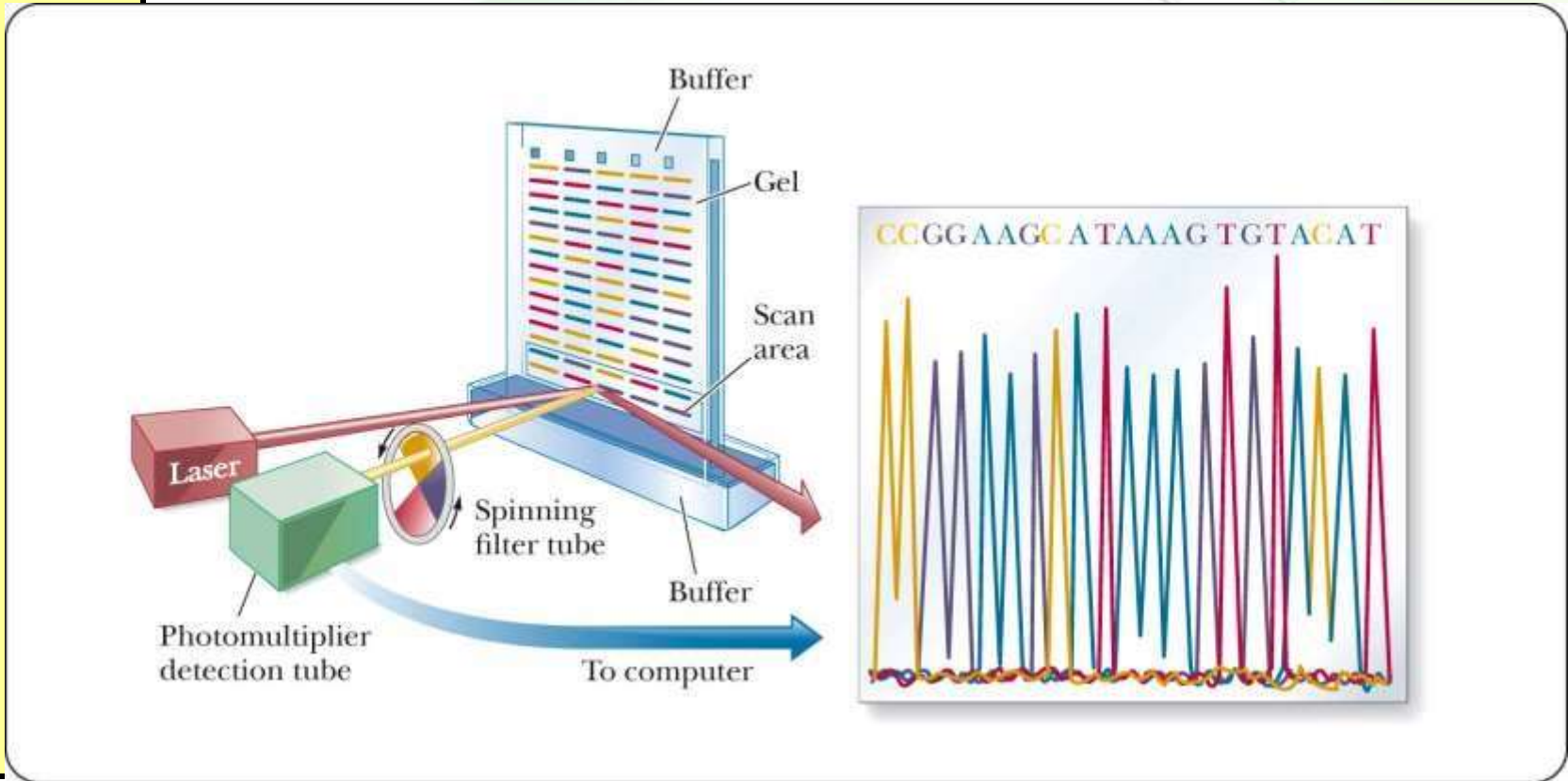
- Reactions include the four deoxynucleotides plus **the four dideoxynucleotides in the same reaction** with each ddNTP labeled with a unique fluorescent tag.

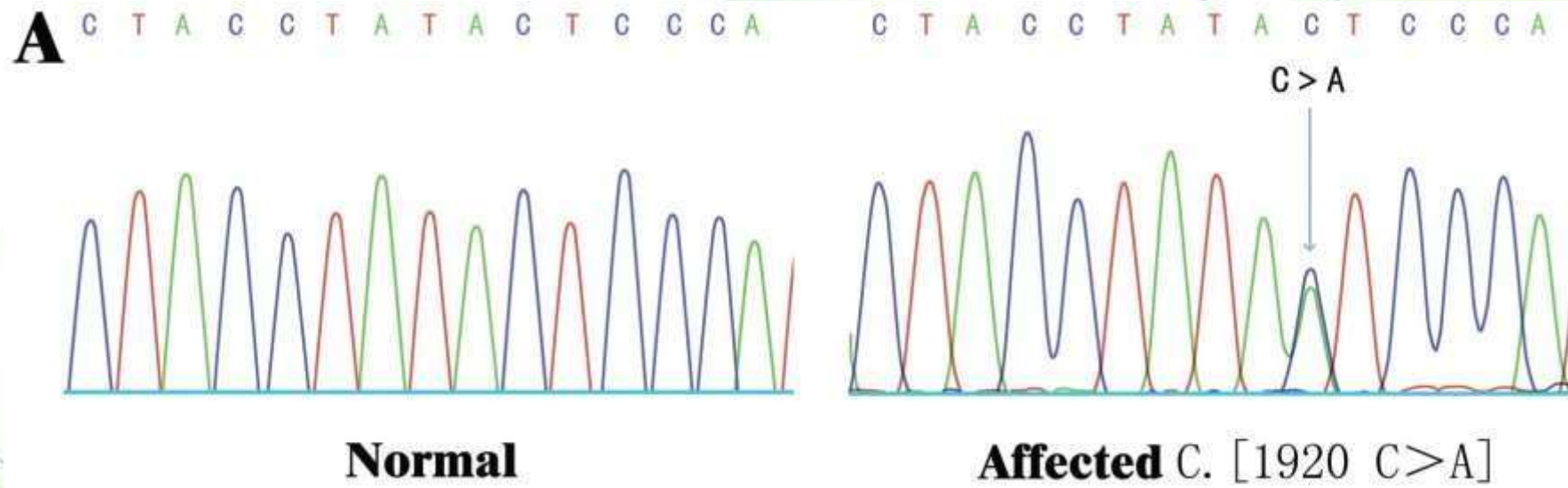






G A T C

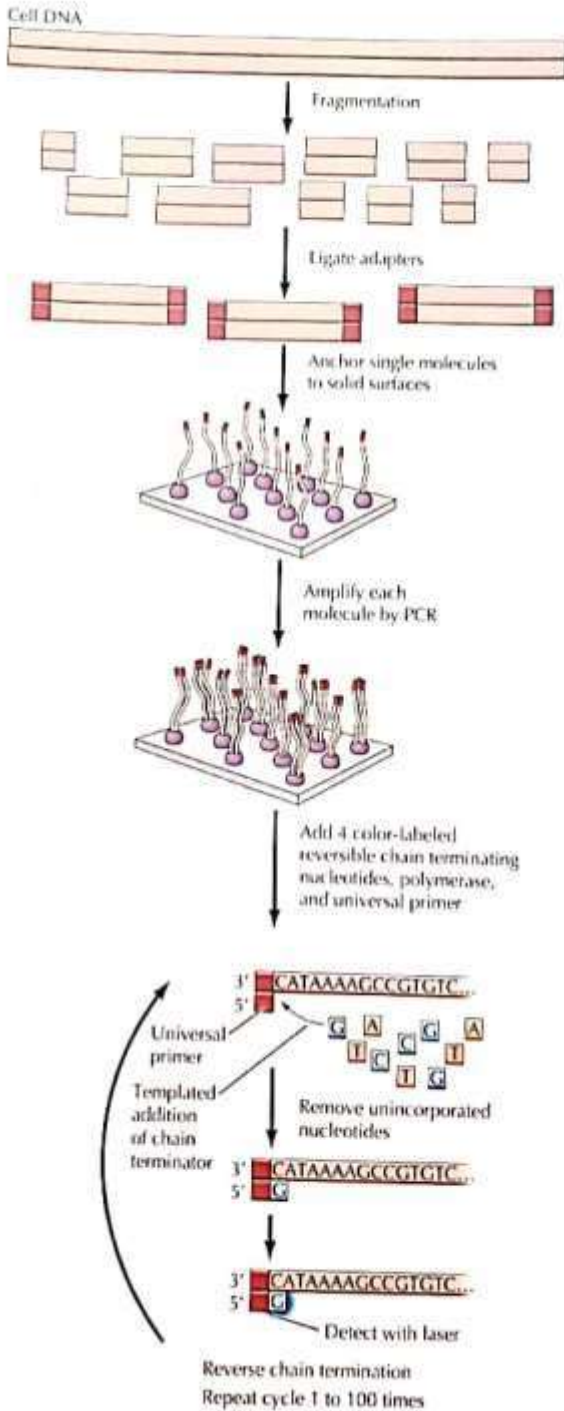




**What does it mean?**

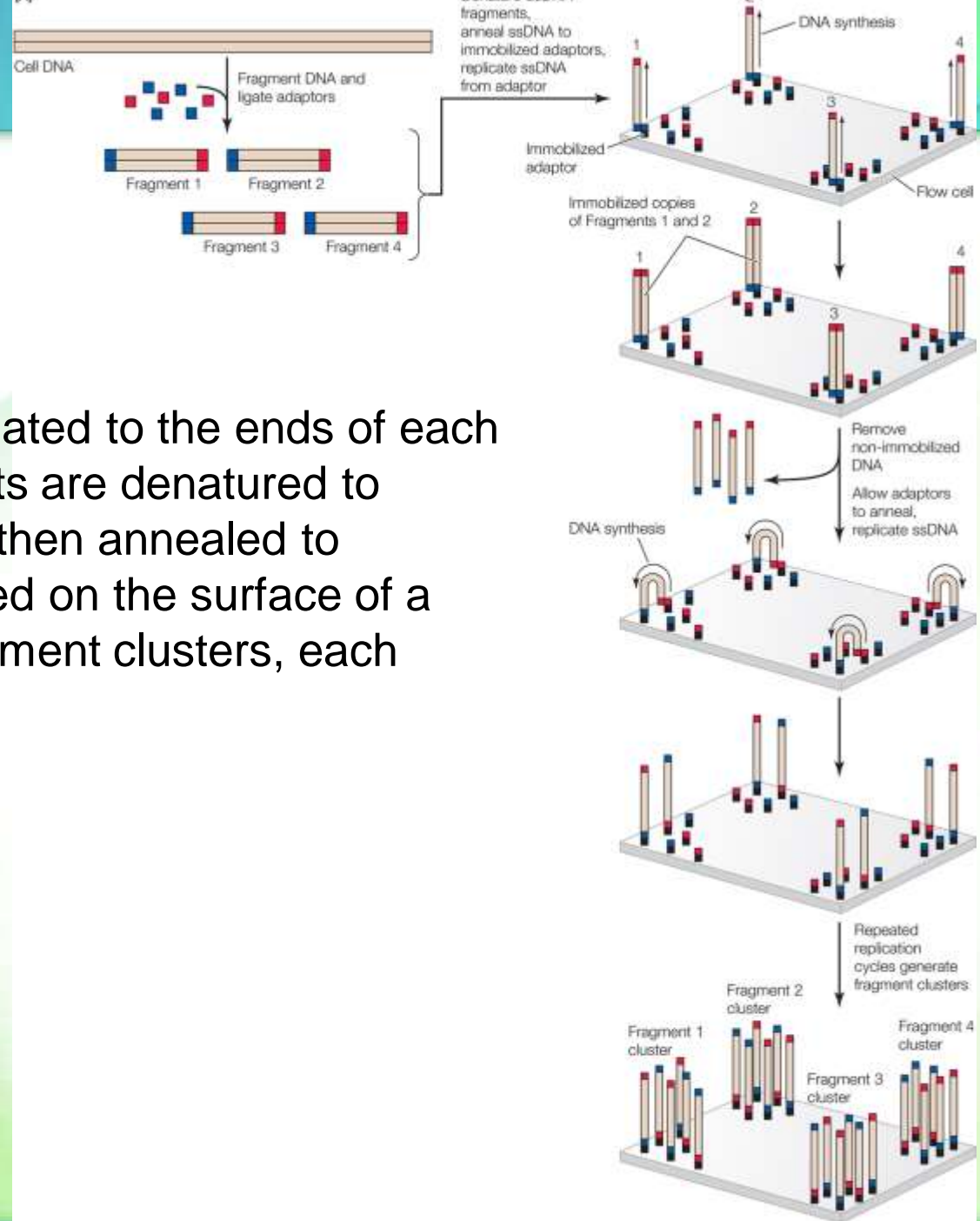


# Next-generation sequencing



- Cellular DNA is fragmented.
- DNA adapters are added to the ends of each DNA fragment.
- Each DNA fragment is attached to a solid surface and amplified like PCR using primers that anneal to the adapter sequences.
  - 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.
  - 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.

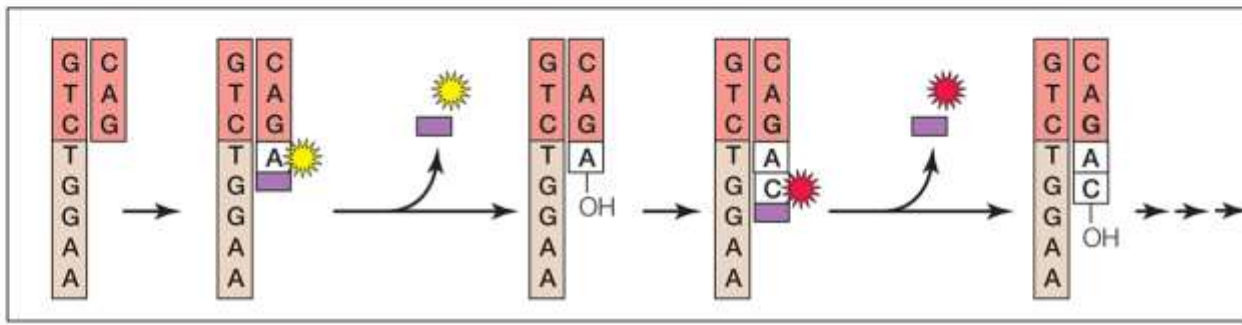
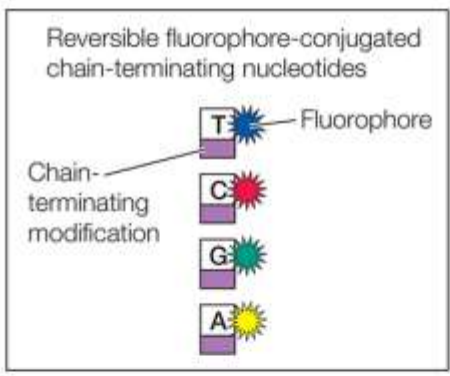
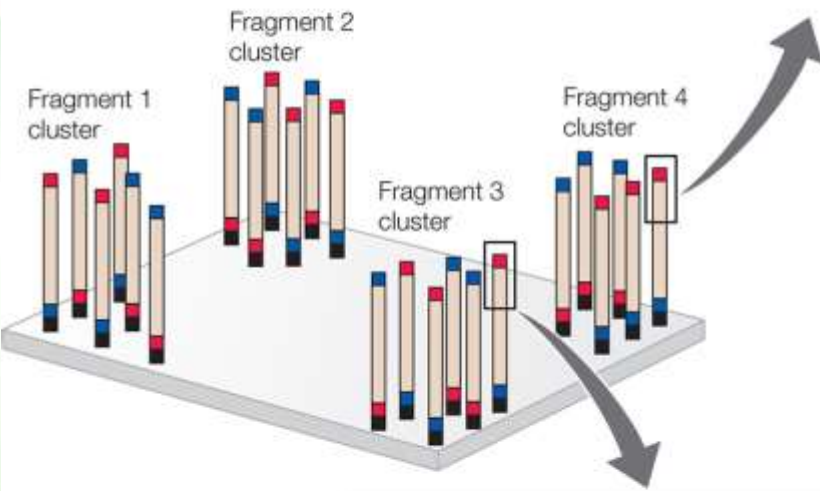
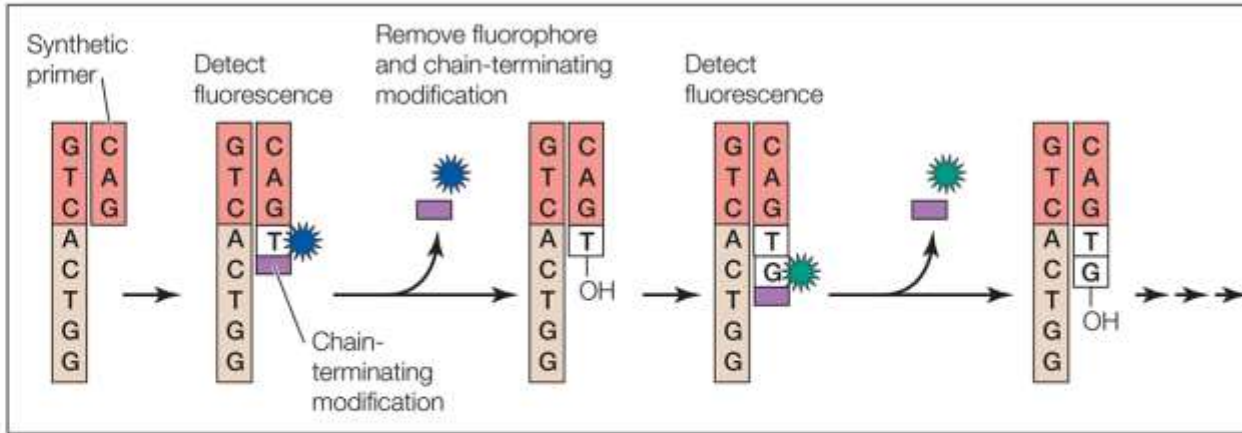
Cellular DNA is fragmented, adapters are ligated to the ends of each fragment, and the double-stranded fragments are denatured to single strands. Single-strand fragments are then annealed to complementary adaptors that are immobilized on the surface of a flow cell and amplified by PCR, forming fragment clusters, each originating from a distinct DNA fragment.



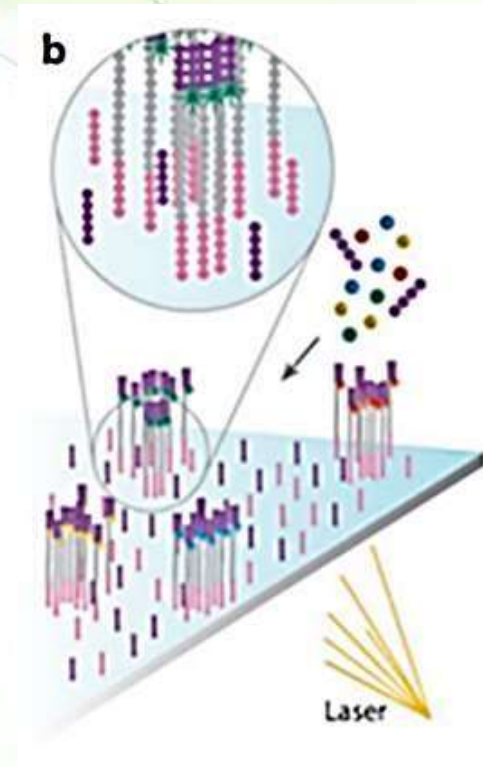
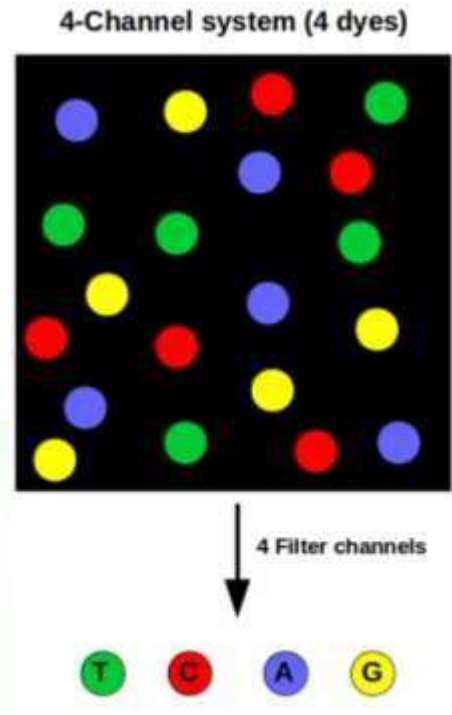


Sequencing by synthesis then determines the nucleotide sequence within each cluster, which uses four reversible fluorophore-labeled, chain-terminating nucleotides that are added one at a time by DNA polymerase and a primer that recognizes the adapter sequence. After addition of each nucleotide, the fluorescent 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.

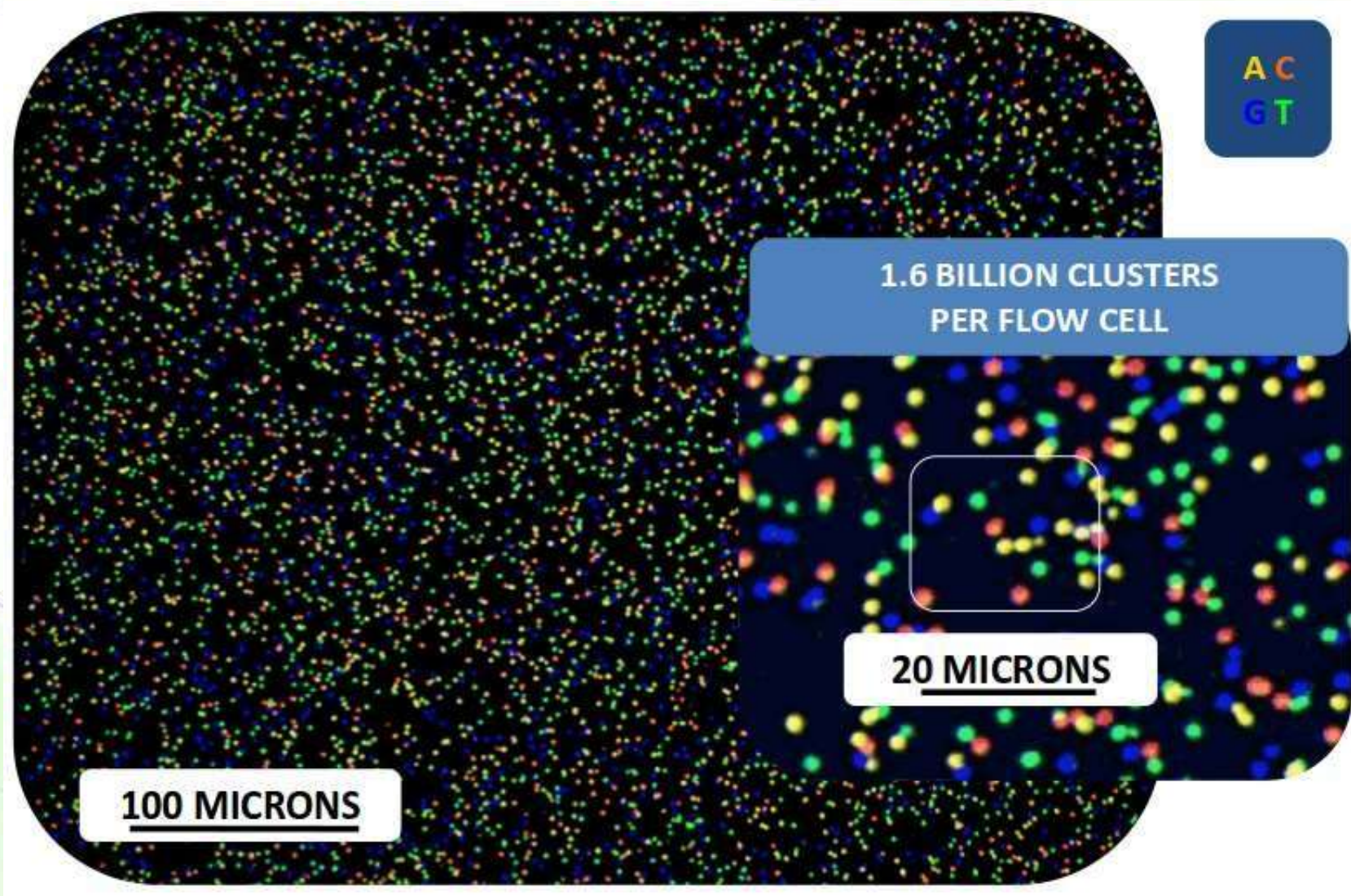
(B) Sequencing by synthesis



# The detection



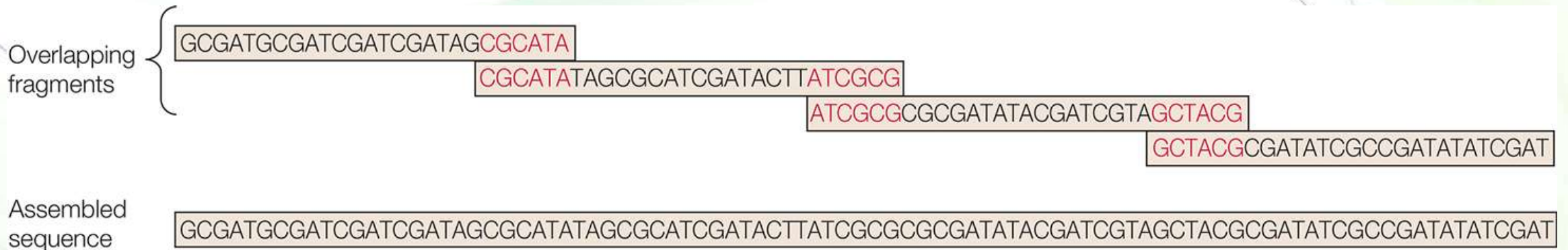
# A real look



# Sequence assembly



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







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

**Data Analysis**  
Create contiguous sequences

