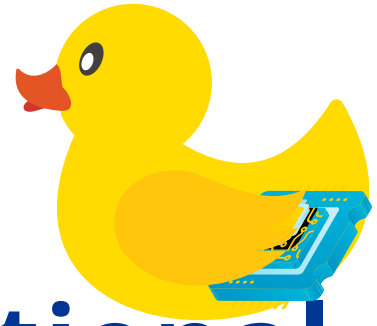


Computational Biology

(BIOSC 1540)

Lecture 02A

DNA sequencing

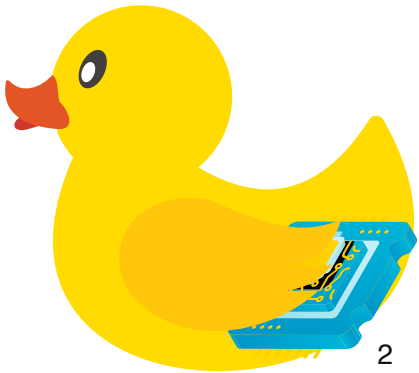
Foundations

Jan 14, 2025

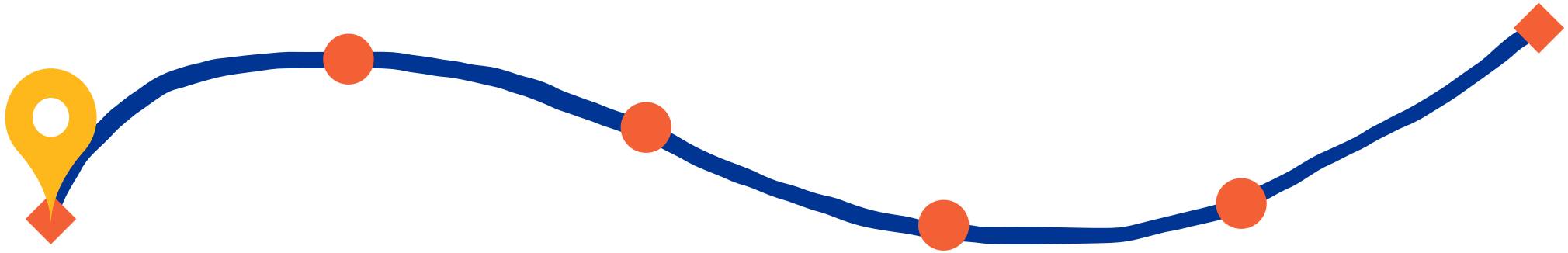
University of Pittsburgh

Announcements

- Assignment P01A is due Friday (Jan 17) by 11:59 pm
- Quiz 01 is in two weeks (Jan 28) and will cover from Lecture 02A to 03B



After today, you should have a better understanding of



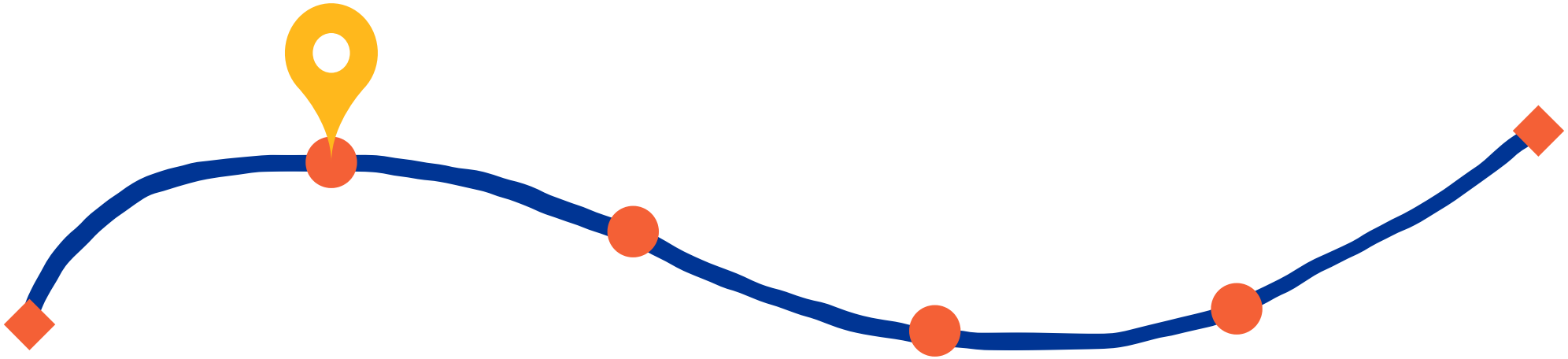
Importance and applications

of DNA sequencing

DNA sequencing revolutionizes biology and medicine through diverse applications

- **Medicine**: Enables precision medicine, genetic disease diagnosis, and cancer genomics.
- **Agriculture**: Enhances crop improvement, pest resistance, and livestock genetics.
- **Evolution**: Deciphers evolutionary relationships and molecular phylogenies.
- **Microbiology**: Identifies pathogens and studies microbial communities (e.g., metagenomics).
- **Ecology**: Monitors biodiversity and tracks species in ecosystems.

After today, you should have a better understanding of

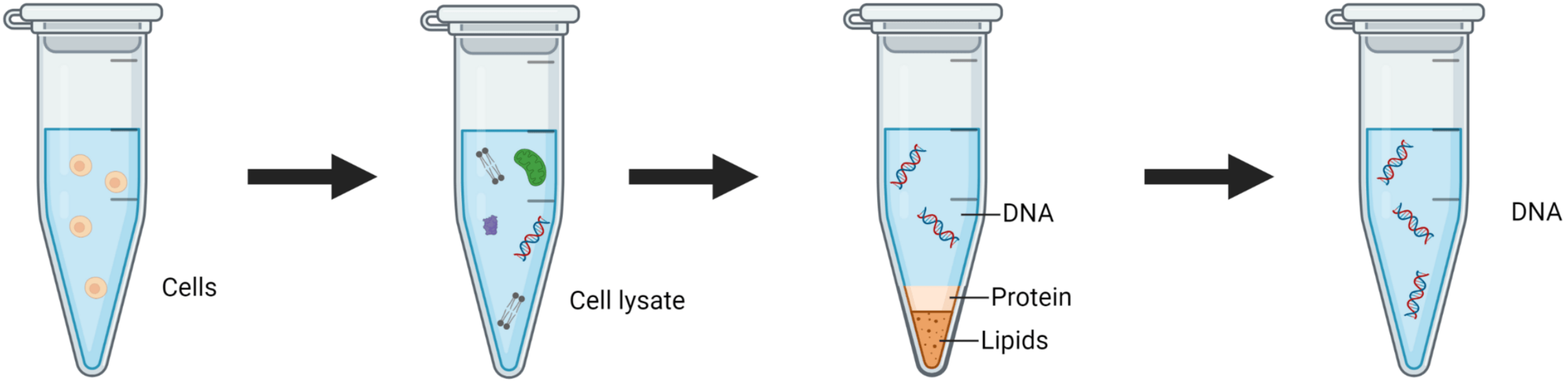


Techniques for extracting and purifying high-quality DNA

DNA extraction

How do we acquire our DNA sample?

Computationalists need to understand the underlying source of our data for quality control



Let's start with a bacterial culture

We let our bacterial culture produce our products of interest

Biotechnology frequently uses massive *E. coli* cultures to produce bioproducts

Fun fact: Pitt has a beer brewing class (ENGR 1933)

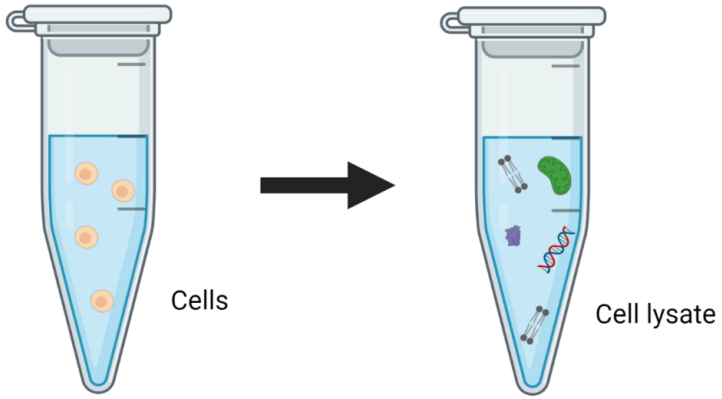
Separate cells from media

The first step is always to centrifuge and separate our cells and media

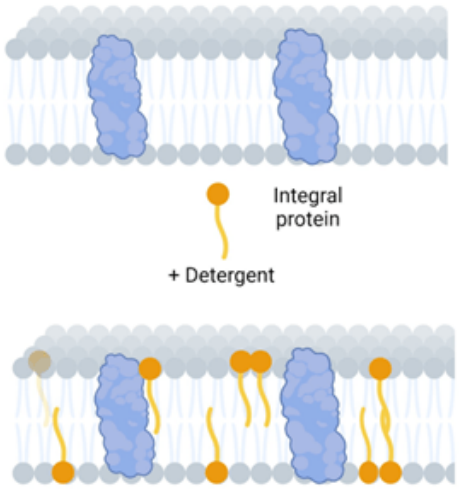
Keep the part that has our **component of interest** (DNA)

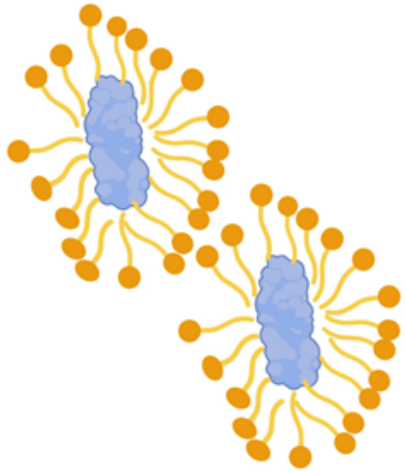
Great! We have our cells, but how can we get DNA out of our cells?

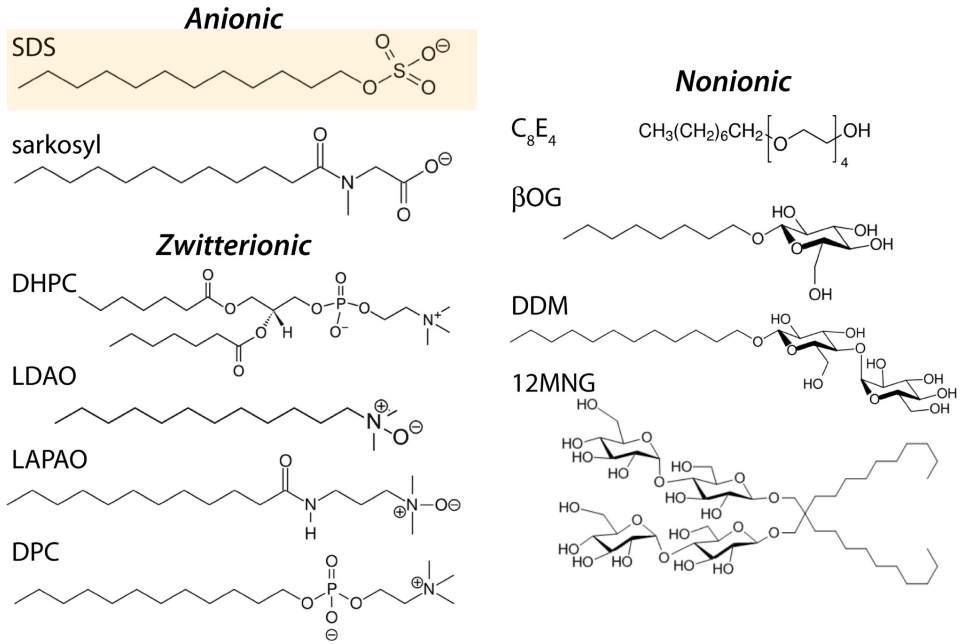
We break open our cells by lysing them



Chemical lysis destabilizes the lipid bilayer and denatures proteins



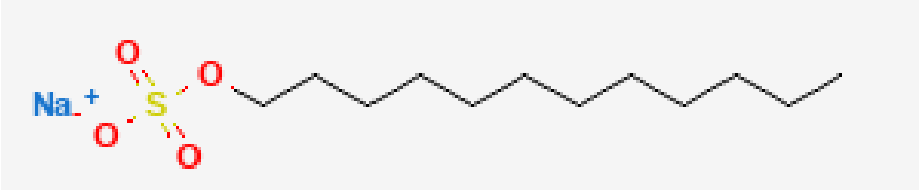




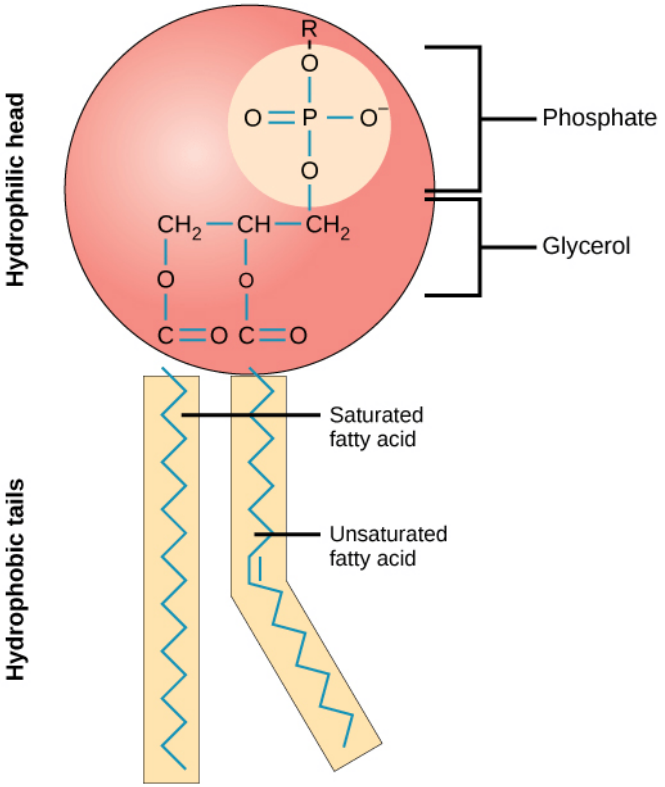
Surfactants have a hydrophilic head and hydrophobic tail

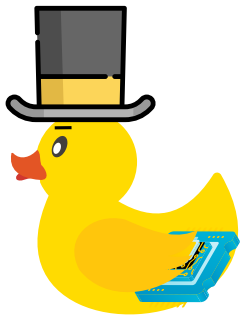
Wait, surfactants sound a lot like phospholipids?

What's the primary difference, and how does this change its behavior?



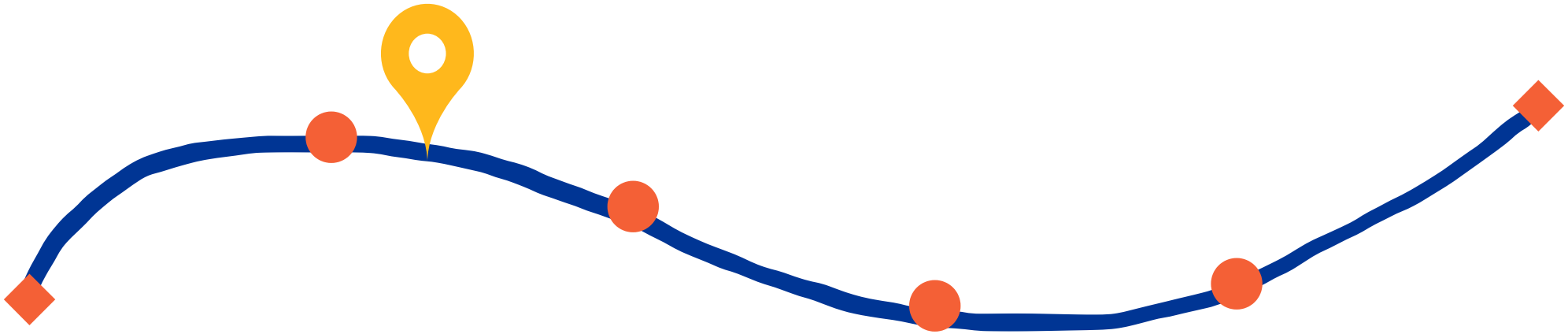
Surfactants possess a single hydrophobic tail. Why does the incorporation of these surfactants destabilize the phospholipid membrane?





Please note: TopHat questions are ungraded. Engaging honestly with the question will benefit you far more than any shortcuts.

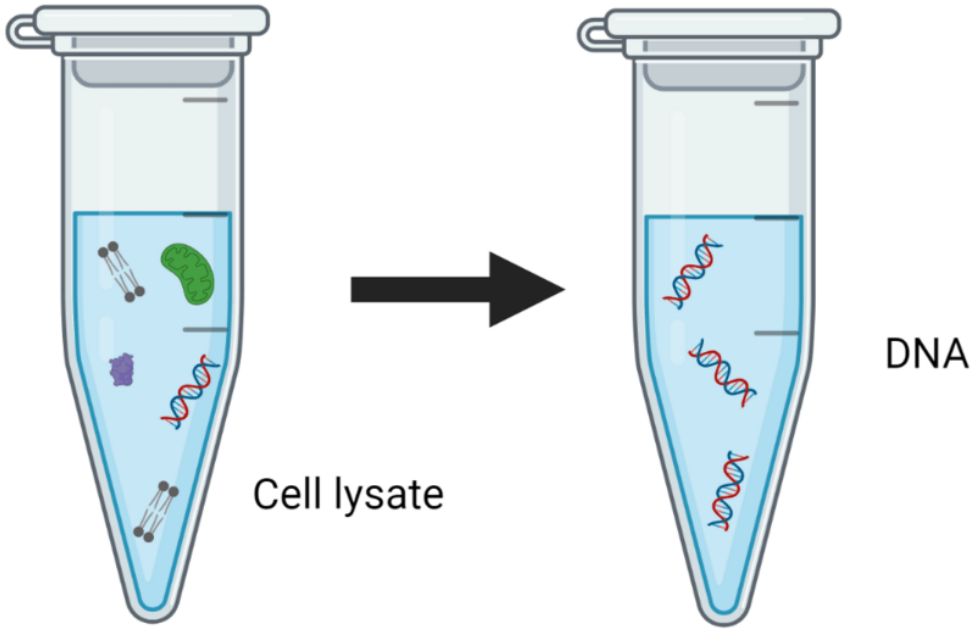
After today, you should have a better understanding of



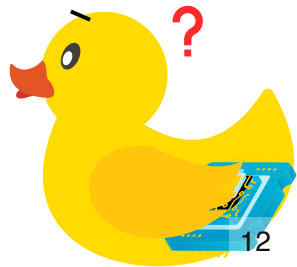
Techniques for extracting and purifying high-quality DNA

DNA purification

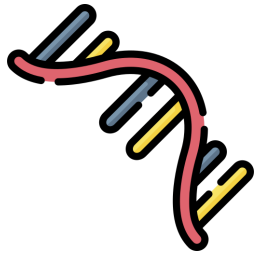
At this stage, we need to separate DNA from other biomolecules … how?



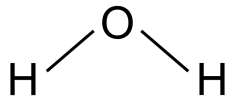
We need to **exploit physicochemical property differences** (such as solubility, charge, and hydrophobicity) to separate DNA from other biomolecules



Phenol-chloroform extraction exploits solubility and density differences



DNA | **RNA** | **Protein** | **Lipids**

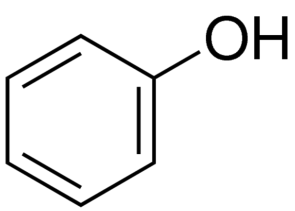


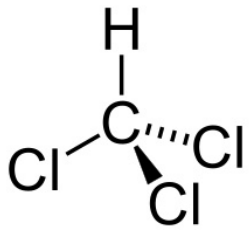
Water

Phosphate backbone (negative charged)

Denatures and aggregates at interface

Nonpolar





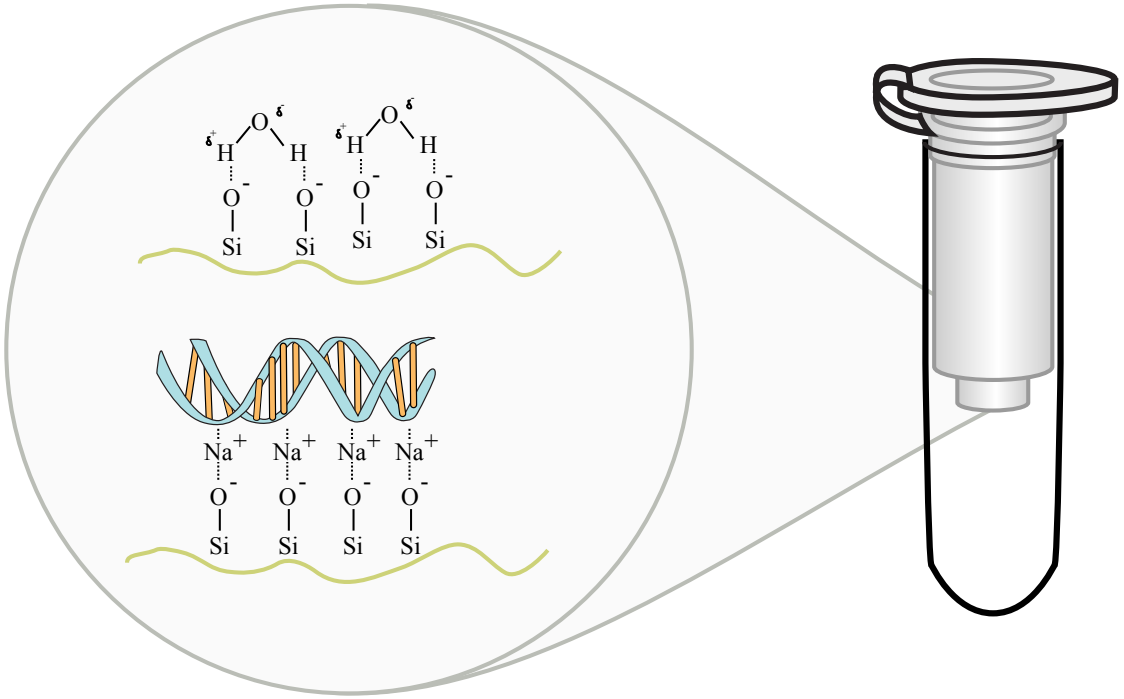
Phenol **Chloroform**

Collecting our aqueous phase selects only DNA and RNA

Silica column-based purification relies on ionic interactions

Under **high-salt conditions**, negatively charged **DNA binds to the positively charged silica membrane** via electrostatic interactions

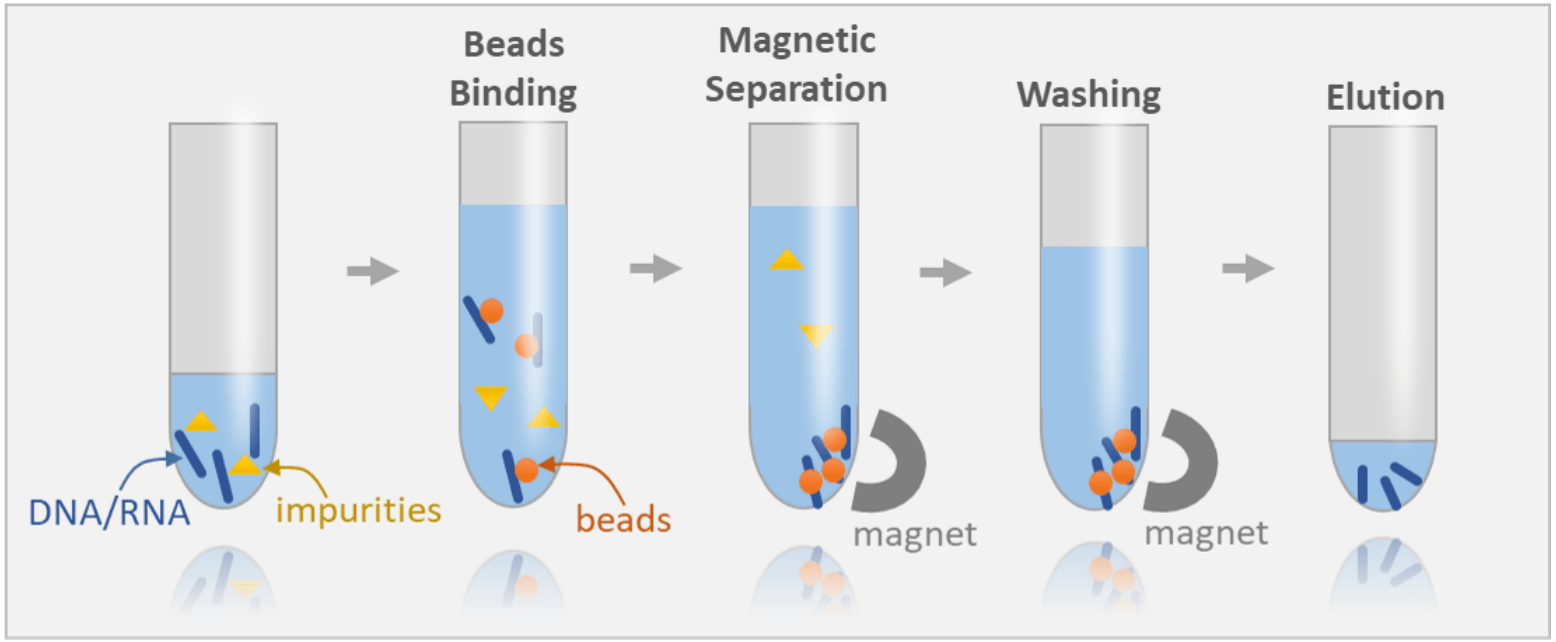
Contaminants like proteins and salts do not bind or are washed away



DNA is then eluted with a low-salt buffer or water

Magnetic beads rely on selective adsorption and surface chemistry

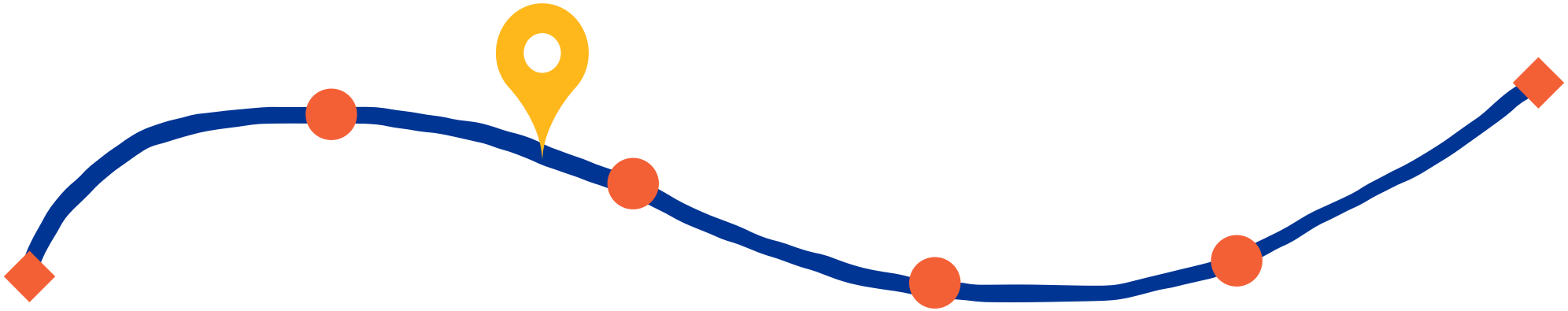
Magnetic beads coated with DNA-binding agents (e.g., silica or polymer) selectively adsorb DNA in the presence of binding buffers



Magnetic fields are used to separate beads with bound DNA from the solution, allowing for washing away impurities like proteins, RNA, and salts

Note: Nowadays, most labs use highly effective kits

After today, you should have a better understanding of

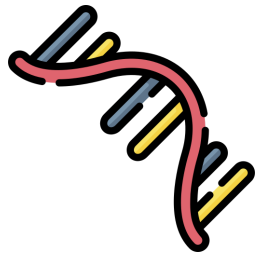


Techniques for extracting and purifying high-quality DNA

DNA quality quantification

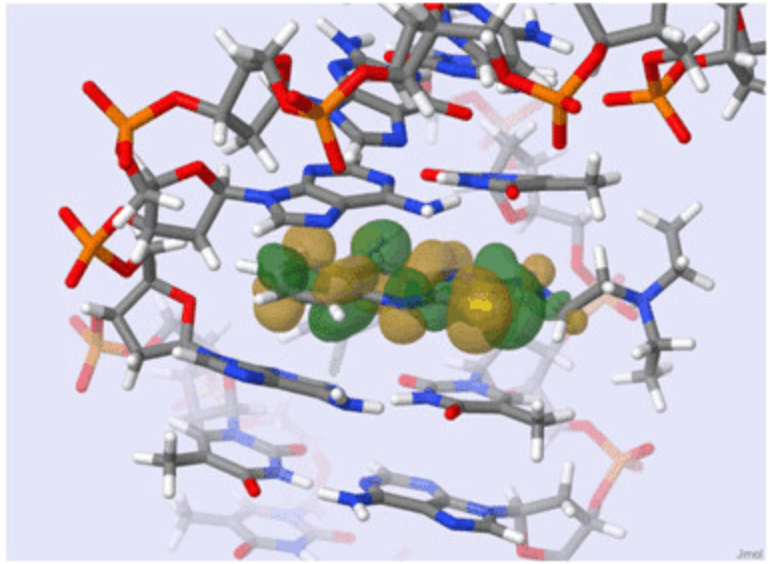
Before sequencing our sample, we should check the quality

DNA

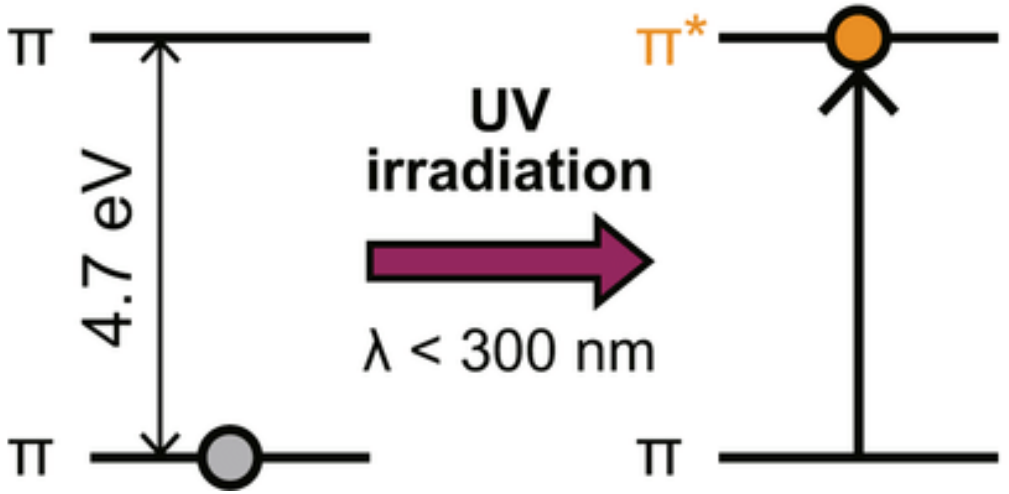
Likely contaminants	Why it's a problem
 RNA	RNA contamination can inflate DNA quantification readings due to similar properties
 Protein	Proteins can inhibit enzymatic reactions in library preparation and distort DNA quantification

UV radiation is selectively absorbed based on molecular structure

Molecules with **aromatic rings absorb UV light** strongly due to their conjugated π-electron systems

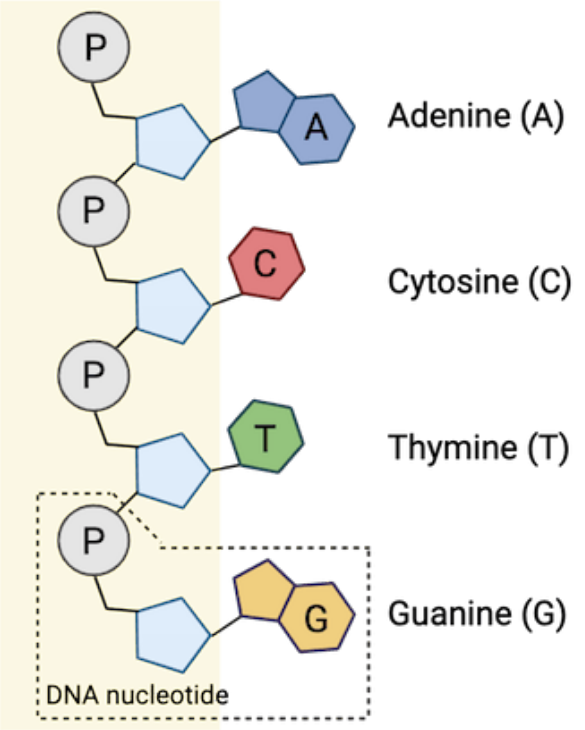


UV light excites electrons in the π-bonds of aromatic systems to higher energy states

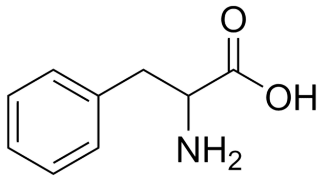


UV radiation is selectively absorbed based on molecular structure

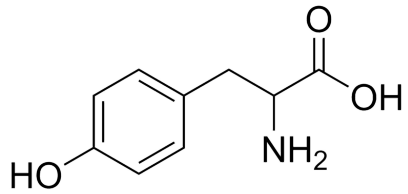
DNA and RNA absorb UV light at **260 nm** because their bases contain highly conjugated double bonds



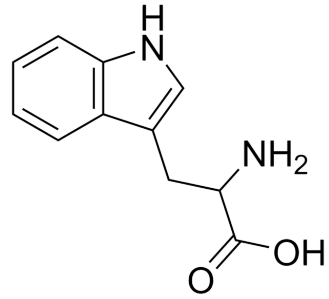
Proteins absorb UV light primarily at **280 nm**, mainly due to aromatic amino acids



phenylalanine



tyrosine

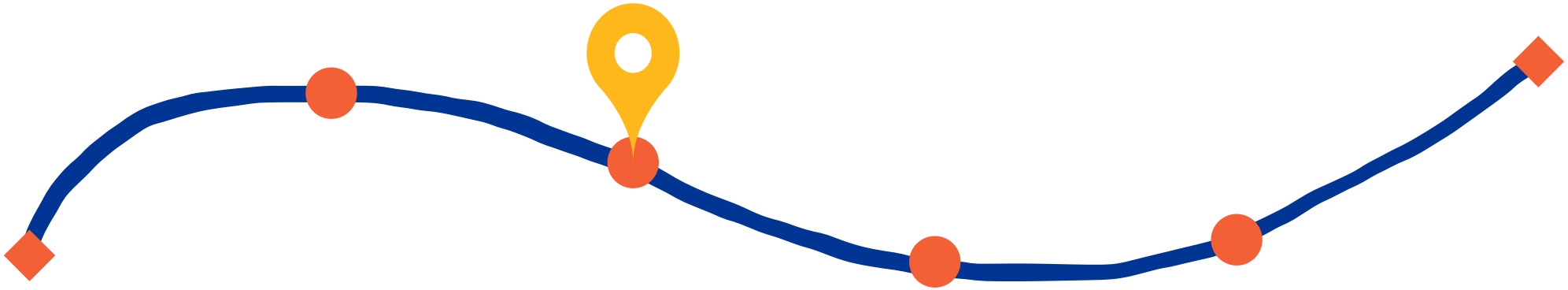


tryptophan

Substances	Ratio A260/A280
Optimum from literature	*1.7-2.0*
DNA	1.80
DNA + 0.4M guanidine HCL	1.74
DNA + 0.4M sodium acetate	1.76
DNA + 0.2% BSA	1.44
DNA + 0.05% phenol	1.59
DNA + 0.1% starch	1.67

A260/A280 ratio relates to sample purity

After today, you should have a better understanding of



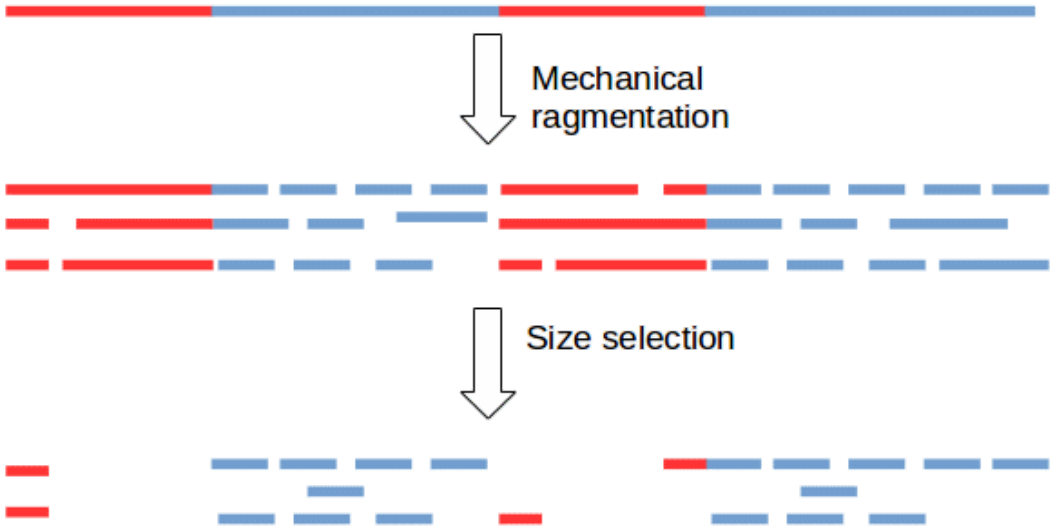
Steps in preparing DNA

libraries for sequencing

A DNA library is a collection of DNA fragments ready for sequencing

Fragmentation breaks DNA into smaller, manageable pieces

Long DNA molecules cannot be sequenced by most platforms due to size constraints



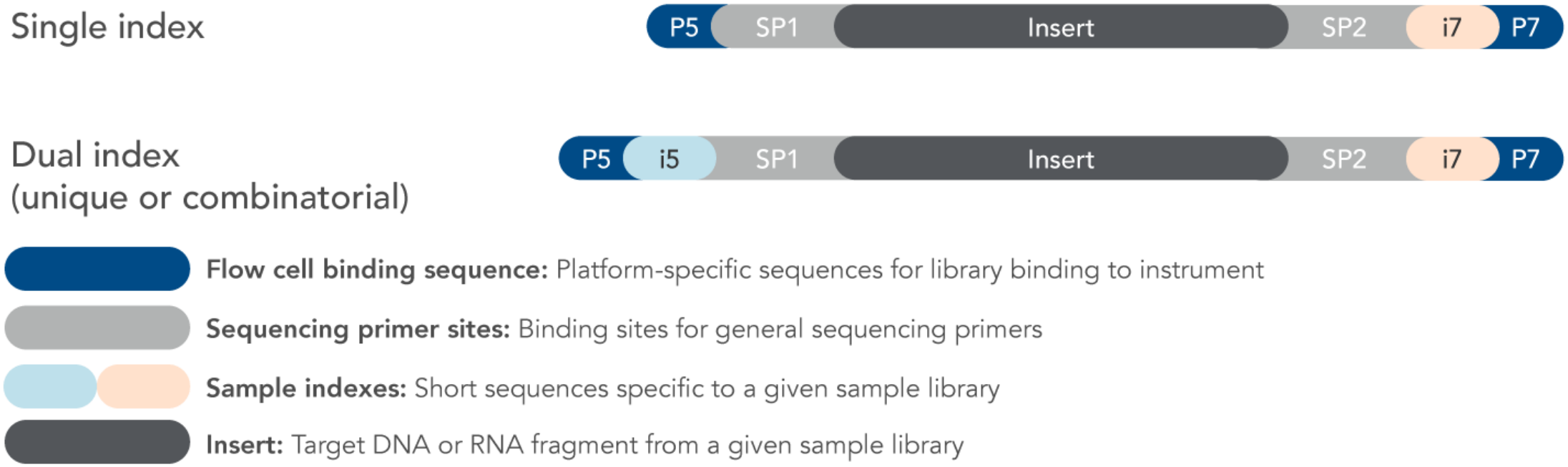
Methods include

- Mechanical shearing (e.g., sonication)
- Enzymatic digestion using restriction enzymes

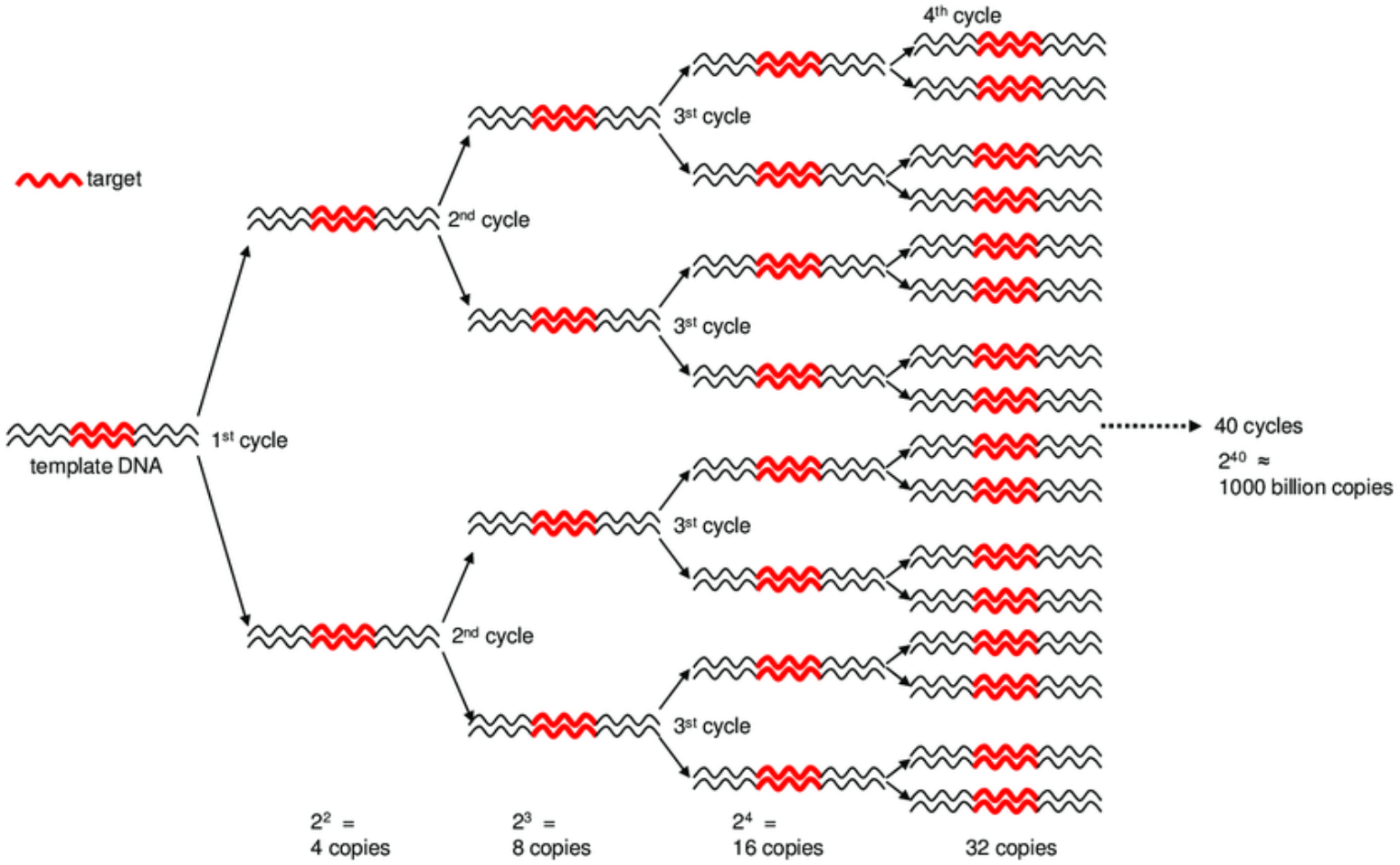
DNA is fragmented to an optimal size range (e.g., 200–500 bp) for efficient sequencing and alignment

Adapter ligation enables amplification and sequencing

Adapters are short, synthetic DNA sequences that are ligated to the ends of DNA fragments during library preparation



PCR amplification ensures sufficient DNA for sequencing



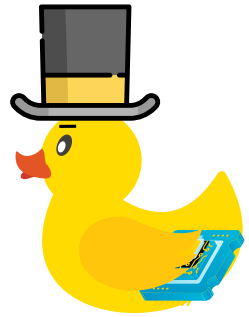
During next-generation sequencing library preparation, short "adapter" sequences are added to the ends of DNA fragments. Which of the following best describes the primary reason for adding these adapters?

A. To link multiple fragments into a single chain for more efficient sequencing.

B. To selectively remove unwanted DNA fragments before sequencing for a better distribution.

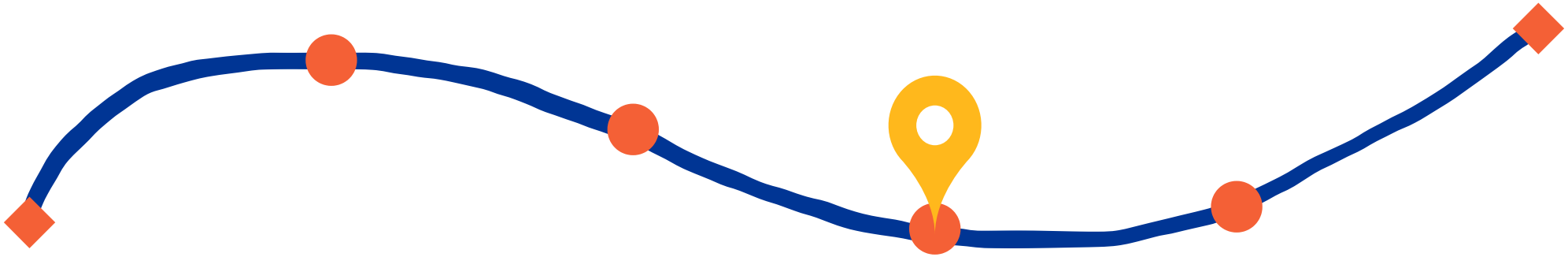
C. To incorporate chemical modifications that prevent secondary structure formation.

D. To provide binding sites for PCR and enable recognition by the sequencing instrument.



Please note: TopHat questions are ungraded. Engaging honestly with the question will benefit you far more than any shortcuts.

After today, you should have a better understanding of

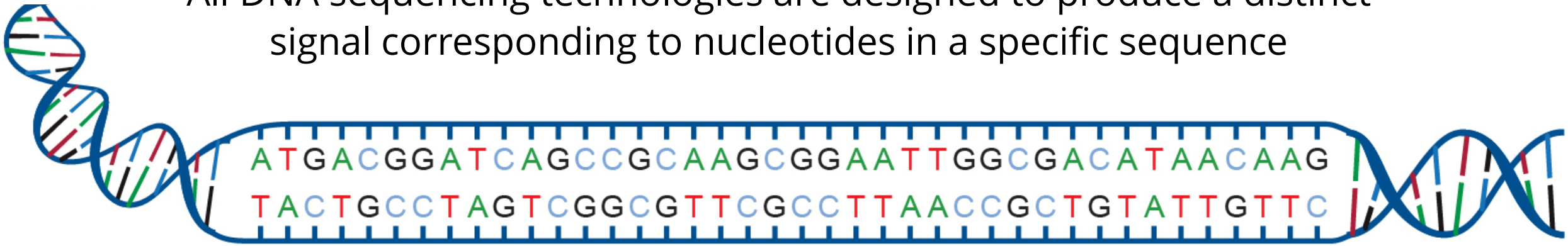


Principles and innovations of

DNA sequencing technologies

Our main problem: Determine the precise ordering of nucleotides

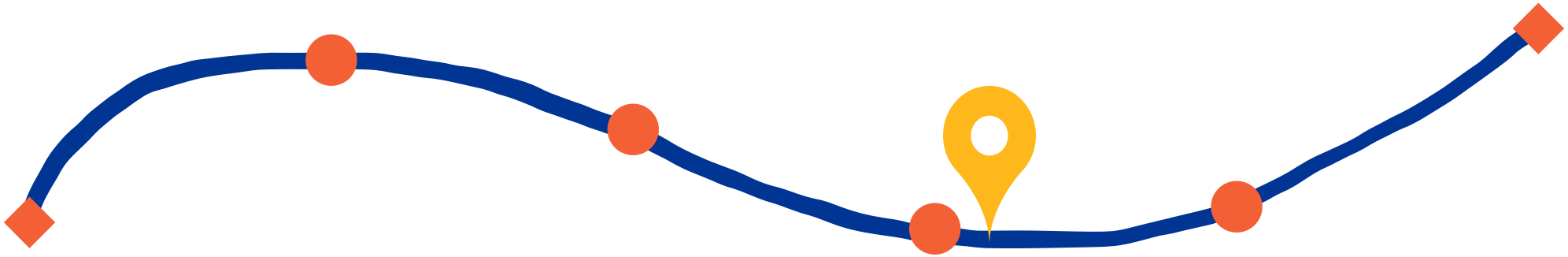
All DNA sequencing technologies are designed to produce a distinct signal corresponding to nucleotides in a specific sequence



Common signals

- **Optical:** Generated by the interaction of light with nucleotides, often through fluorescence or absorbance.
- **Electrical:** Variations in current or voltage as nucleotides interact with a sensing element.
- **Chemical:** Produced by enzymatic or chemical reactions.

After today, you should have a better understanding of



Principles and innovations of

DNA sequencing technologies

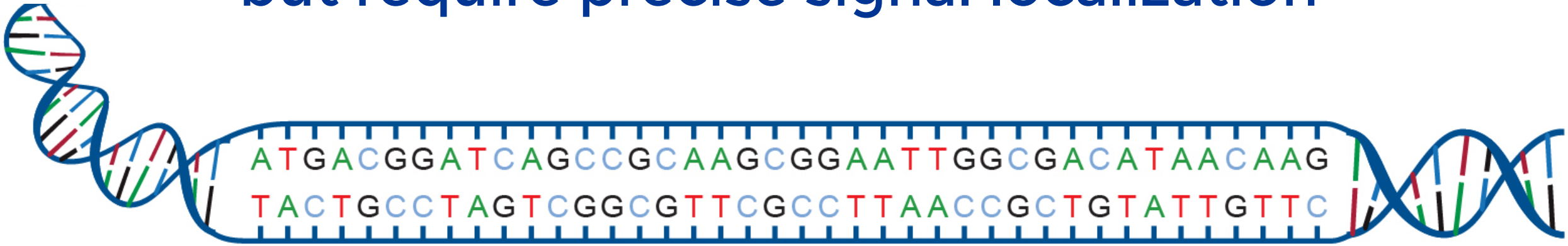
Chain termination (Sanger)

DNA elongation happens rapidly and continuously

We use DNA polymerase + excess nucleotides to make copies of DNA

https://omics.crumblearn.org/sequencing/dna/pcr/dna-elongation.html

Fluorescent tags enable nucleotide detection but require precise signal localization

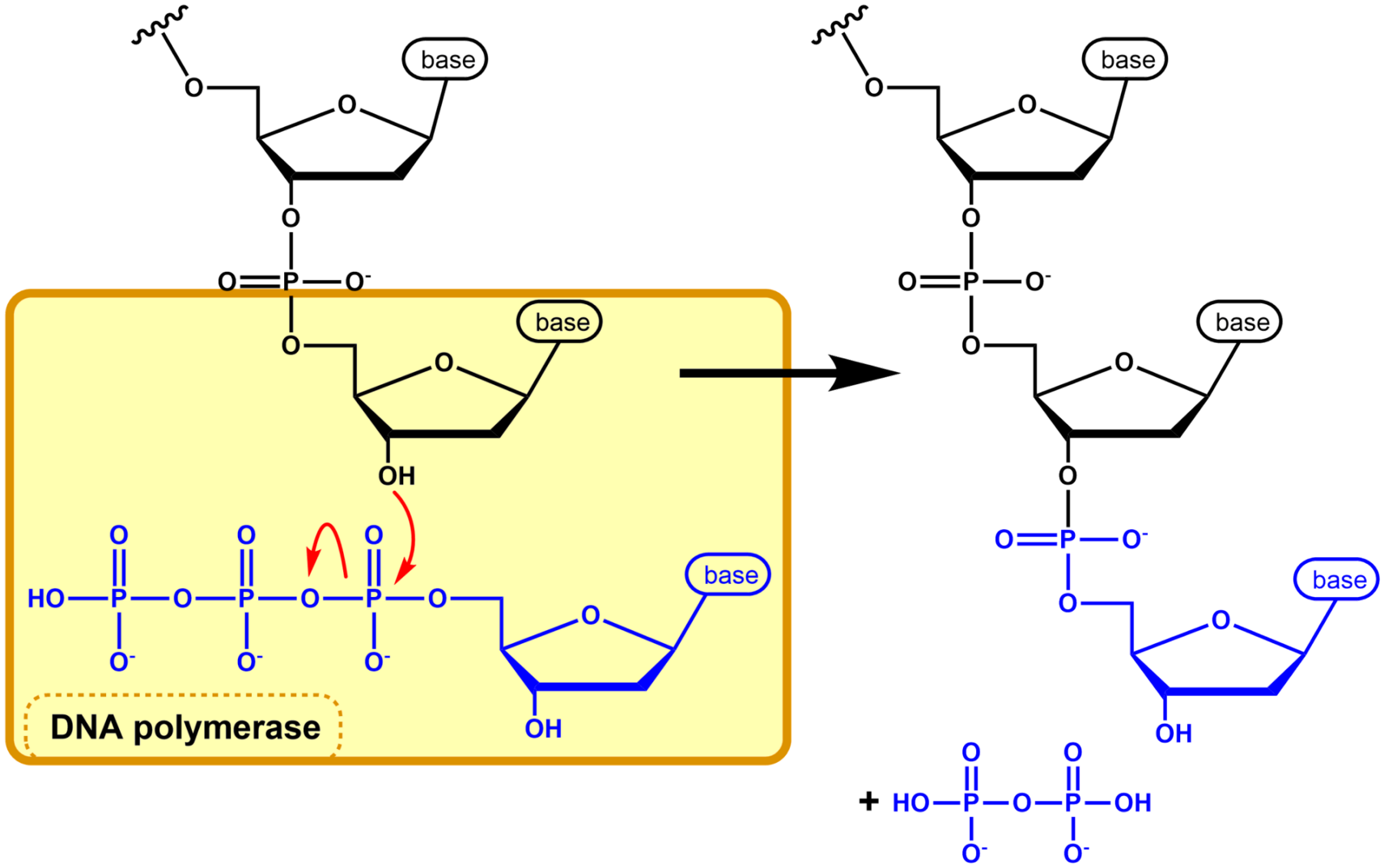


When excited by light, fluorescent tags emit distinct signals, providing a mechanism to detect nucleotide identity

Issue: How can we determine where the signal is coming from in the sequence?

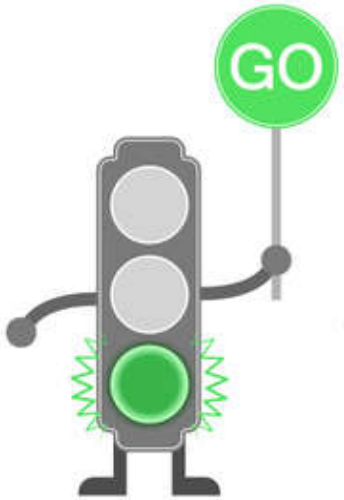
The length of a DNA fragment can be used to specify a nucleotide location (i.e., the last nucleotide)

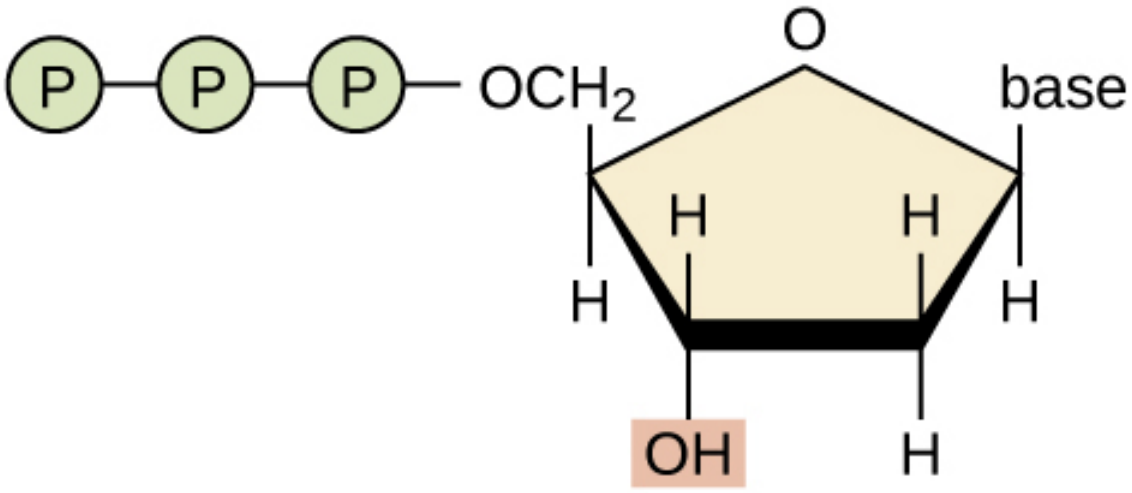
3' OH is required for DNA elongation



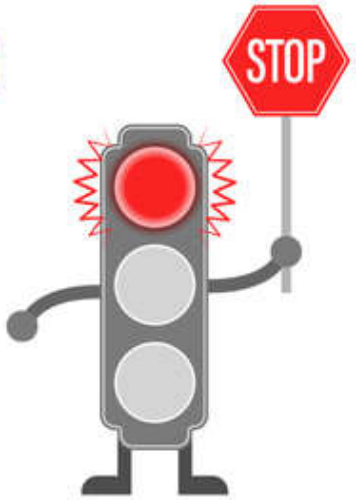
What happens if we don't have the 3' OH? We cannot add another nucleotide

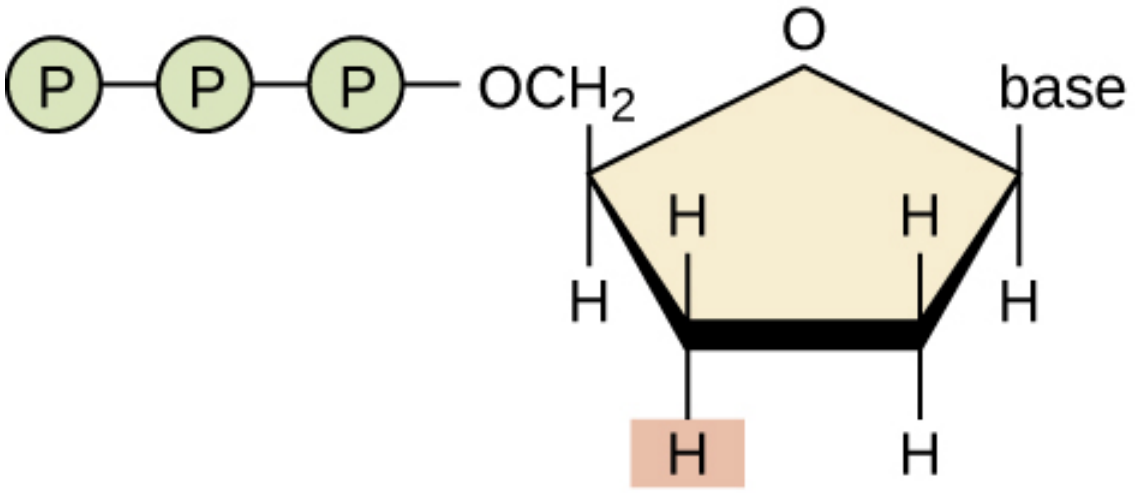
Di-deoxynucleotides stop replication





deoxynucleotide (dNTP)





dideoxynucleotide (ddNTP)

ddNTP will randomly stop DNA elongation

When DNA polymerase adds a **ddNTP**, it cannot add any other nucleotide

We will be left with DNA strands of variable length with an optical-based signal at the end

https://omics.crumblearn.org/sequencing/dna/first-gen/sanger/principles/chain-termination.html

Ratio is usually **1**:**100**

By sorting DNA fragments by length, we can identify the last nucleotide is

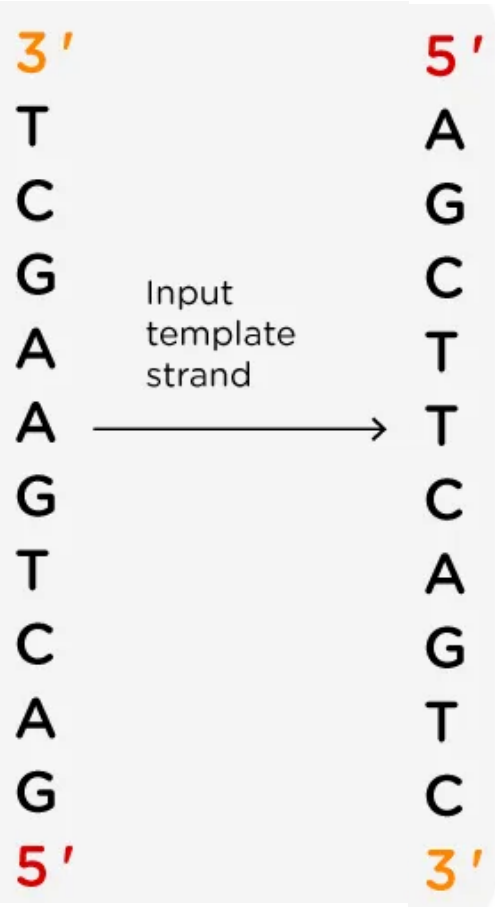
Variable-length fragments

C*GAAGTCAG 5′
A*G 5′
A*GTCAG 5′
G* 5′
T*CAG 5′
G*AAGTCAG 5′
C*AG 5′
A*AGTCAG 5′
G*TCAG 5′
T*CGAAGTCAG 5′

Fragments sorted by length

G* 5′
A*G 5′
C*AG 5′
T*CAG 5′
G*TCAG 5′
A*GTCAG 5′
A*AGTCAG 5′
G*AAGTCAG 5′
C*GAAGTCAG 5′
T*CGAAGTCAG 5′

Last nucleotide order



Original setup

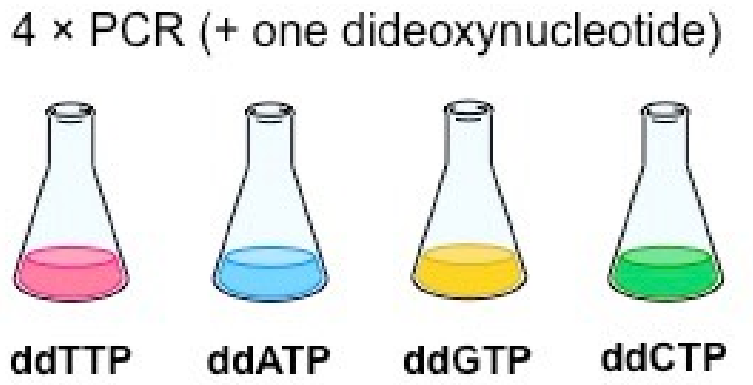
1. Split DNA sample into four beakers
2. Add all four dNTPs to each beaker
3. Add some amount of radioactive ddNTP in a single beaker
4. Add Taq polymerase and let PCR run

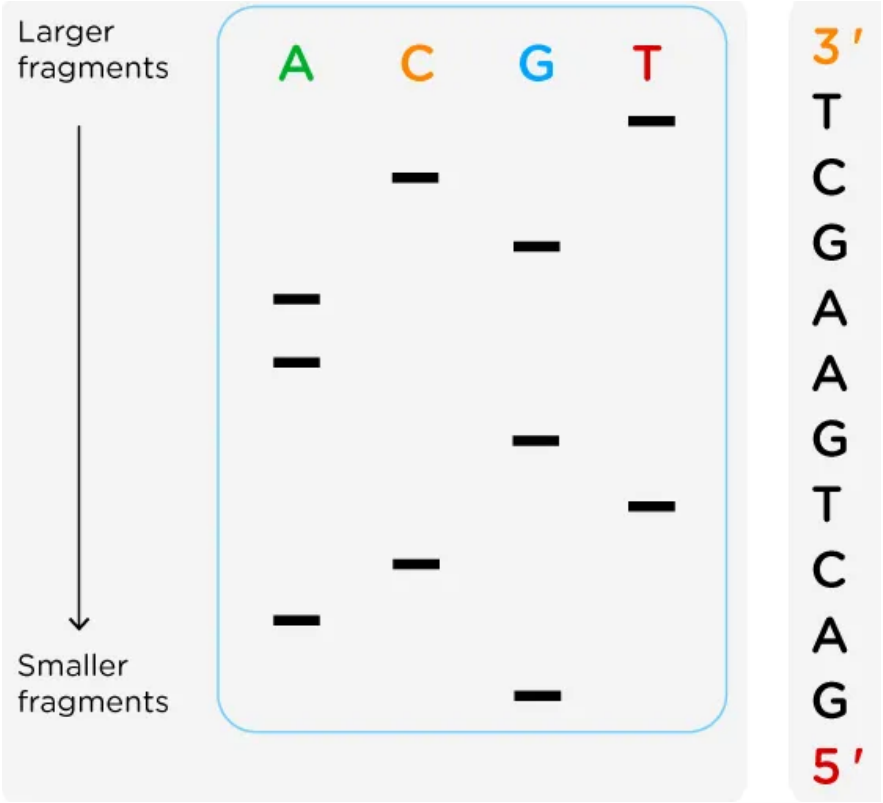
Once we have fragments, how can we separate them by length?

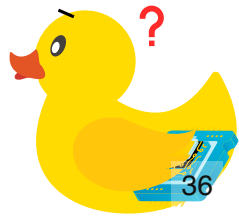
Gel electrophoresis!

Why would we need separate beakers?

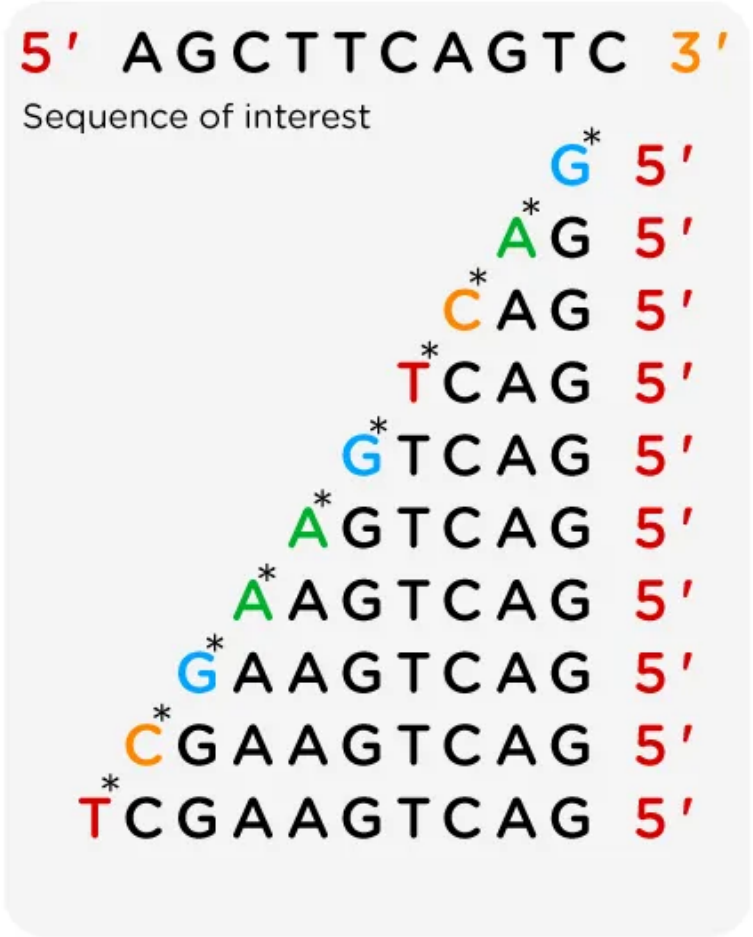
Cannot differentiate between radioactive nucleotides

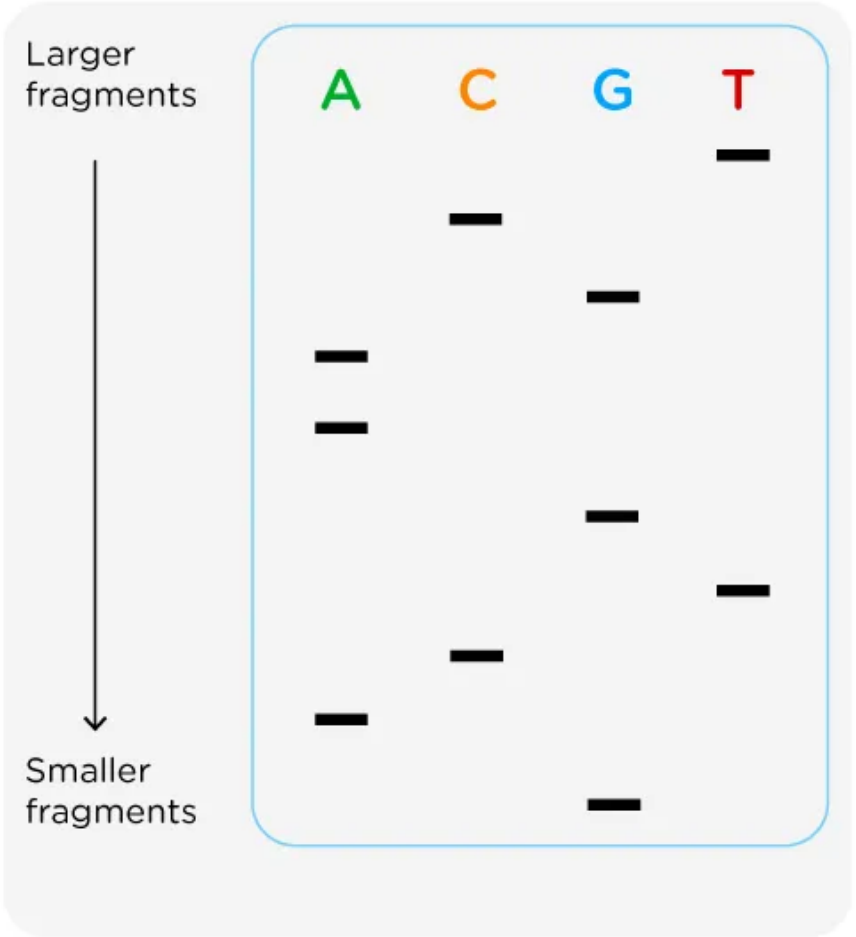


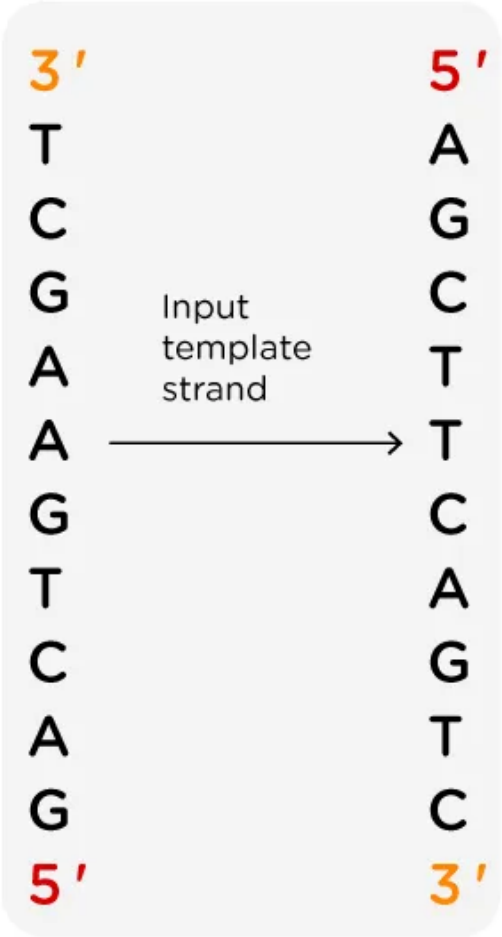




We can build our sequence based on what (radioactive) ddNTP is at that position

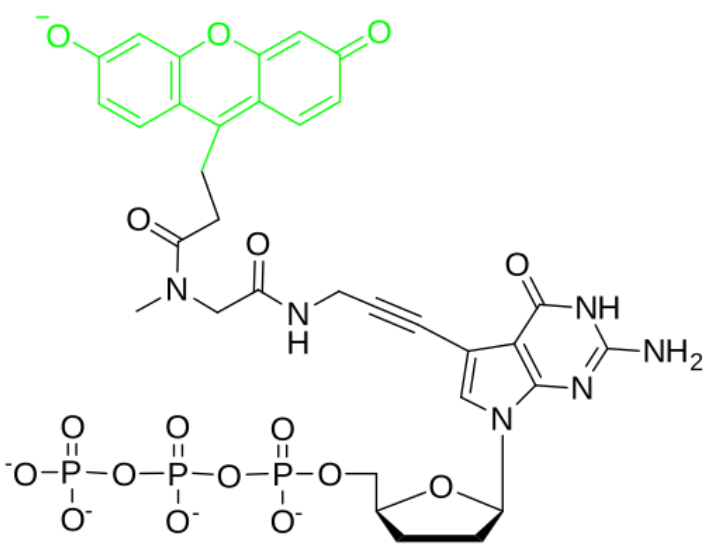




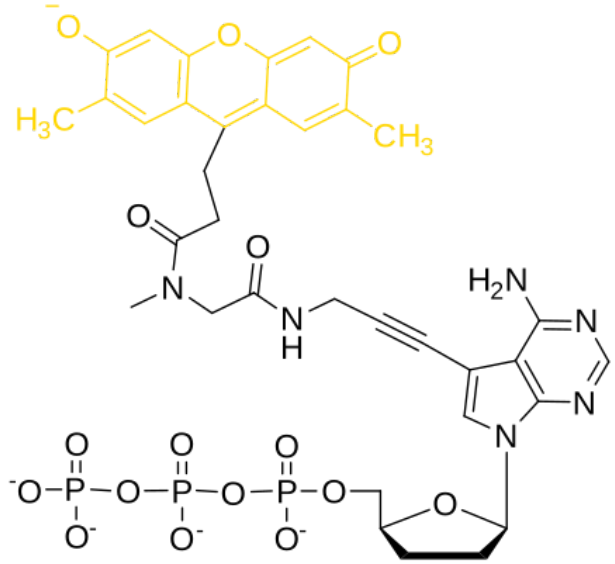


Now we use fluorescence to distinguish ddNTPs

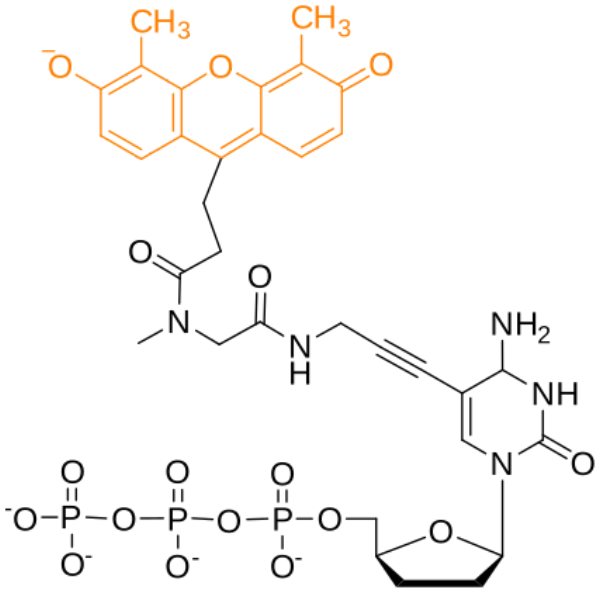
Only need one PCR!



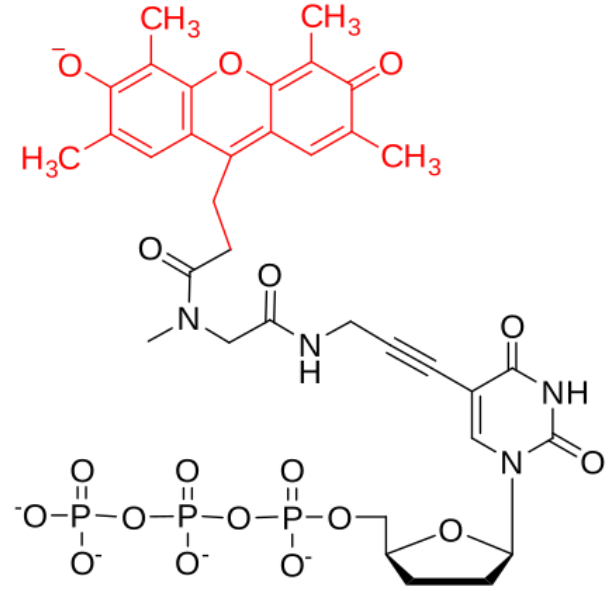
G-505



A-512



C-519

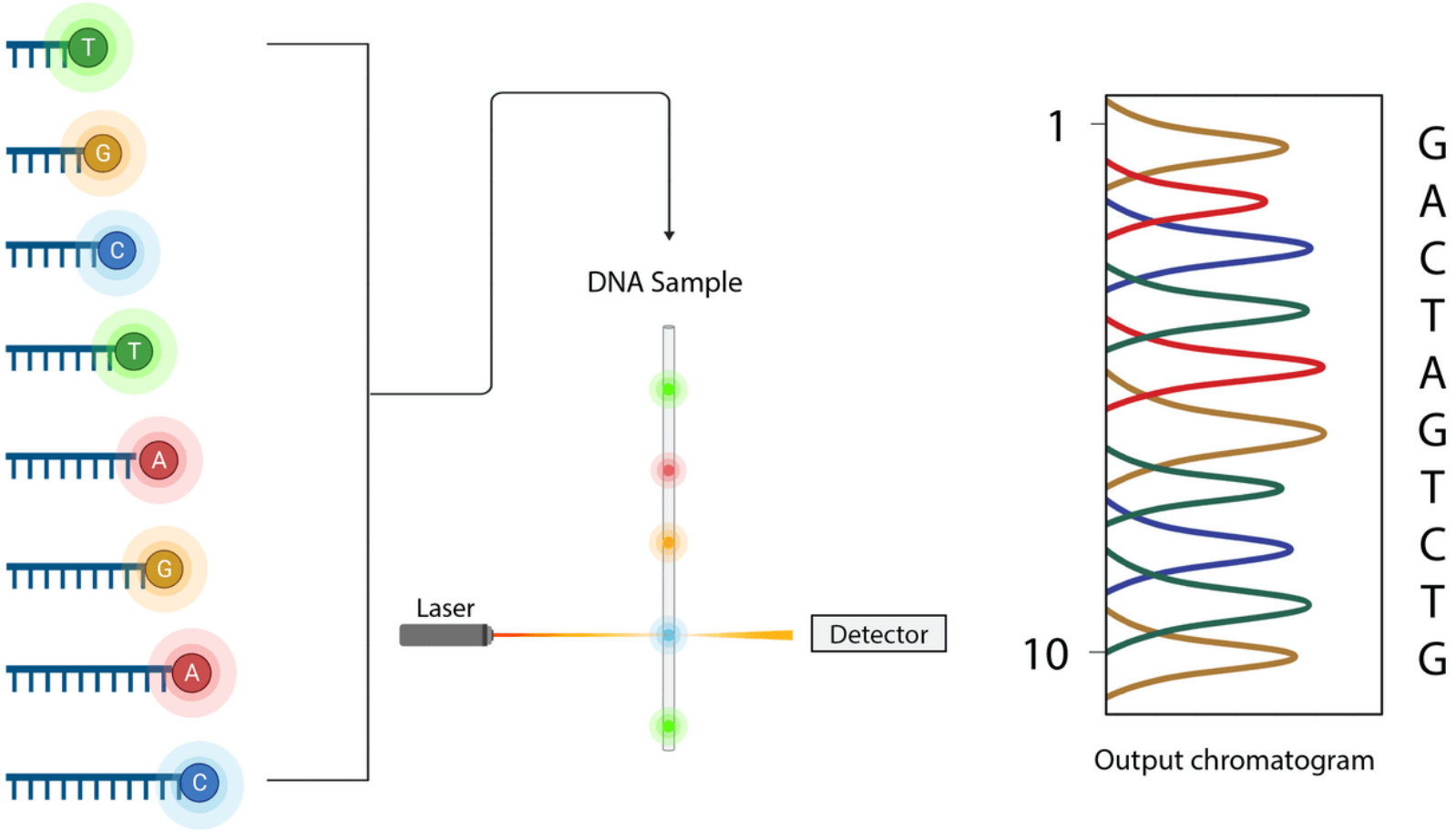


T-526

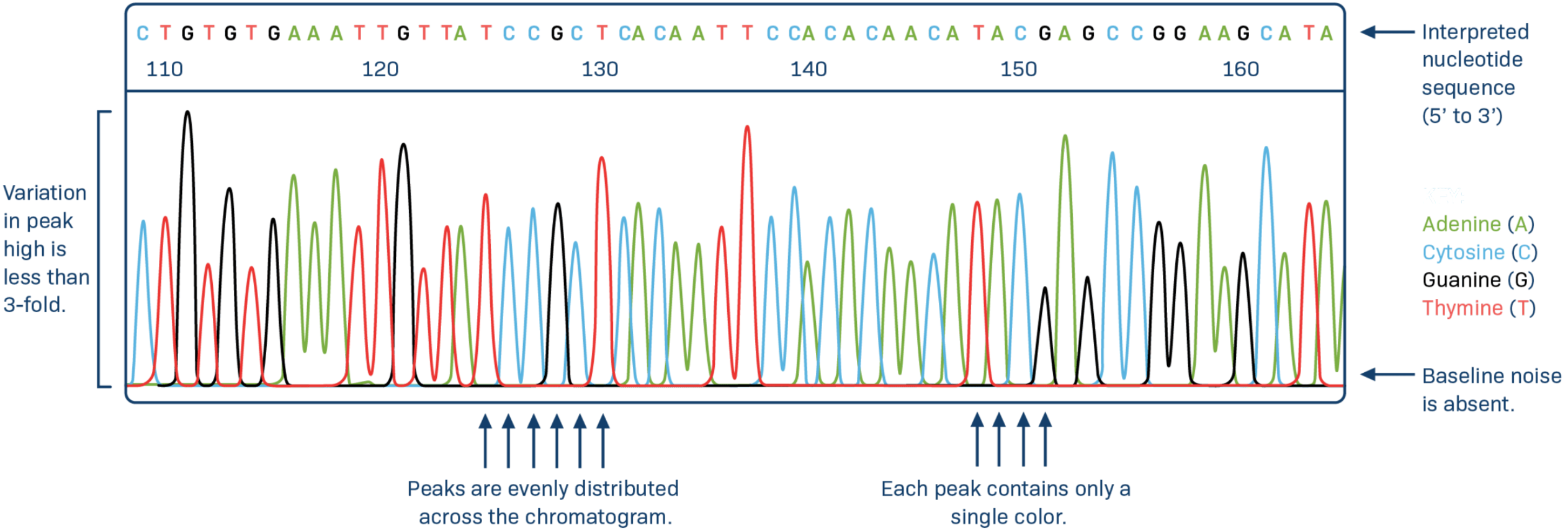
We also can automate fragment separation

Capillary gel electrophoresis can accelerate fragment length sorting and detection

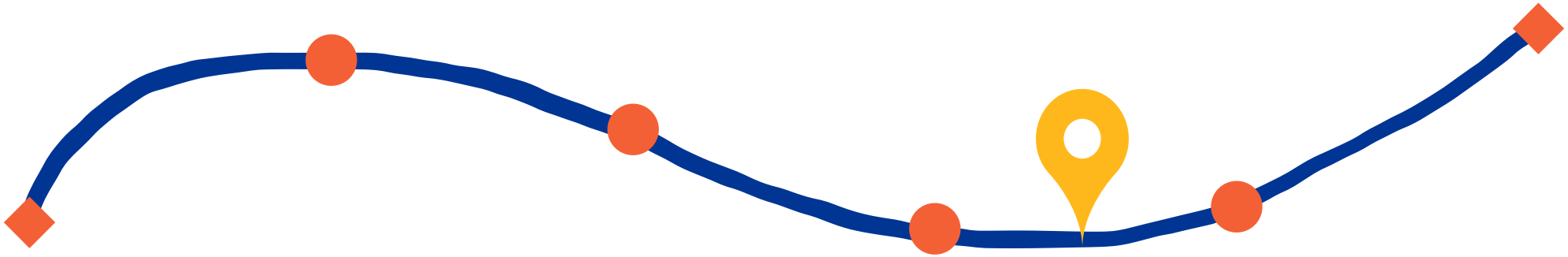
Unique fluorescence signal per ddNTP produces a **chromatogram**



Ideal chromatogram



After today, you should have a better understanding of



Principles and innovations of

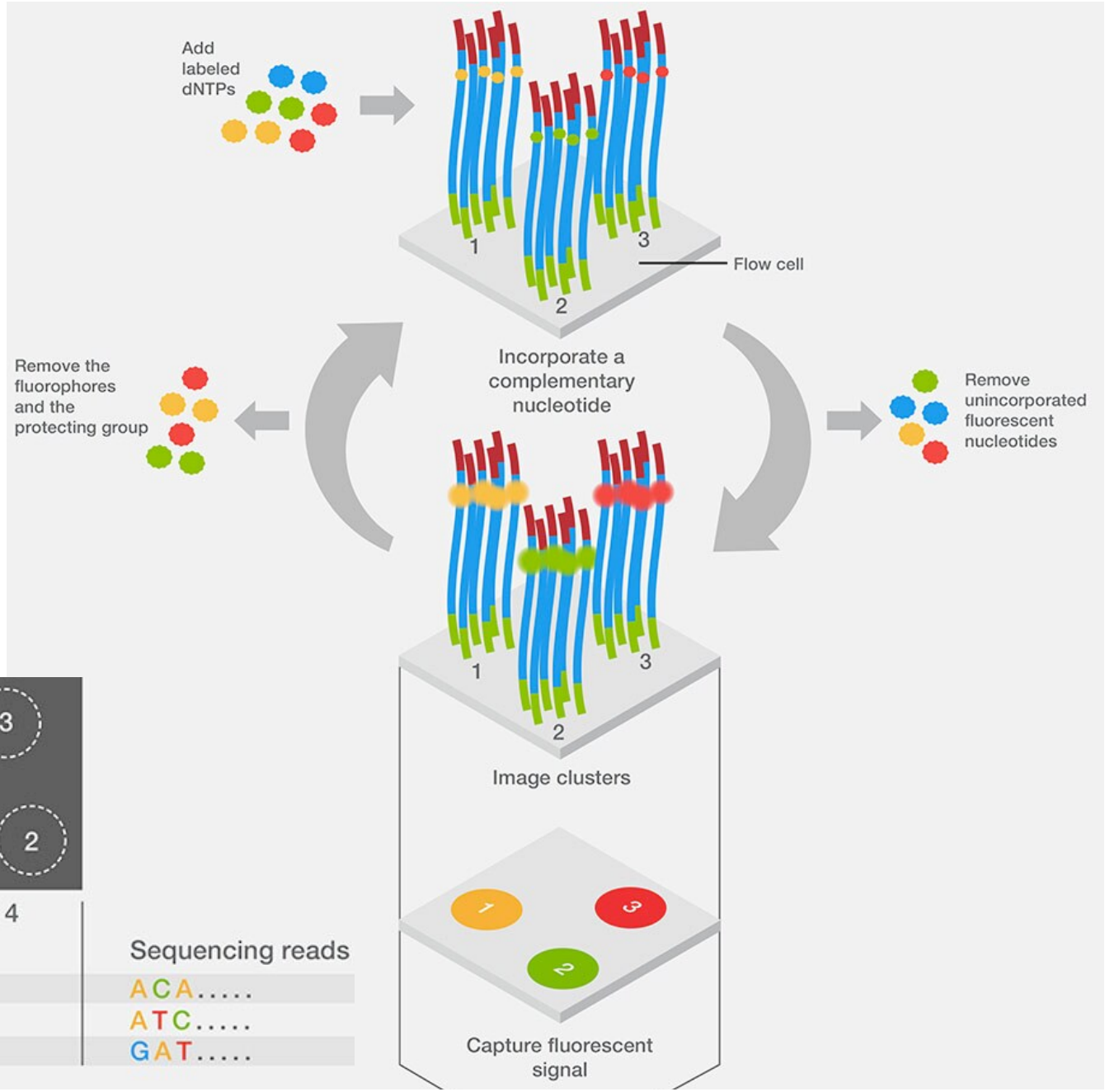
DNA sequencing technologies

Sequencing by synthesis (Illumina)

Sanger sequencing is highly accurate but lacks scalability and speed for large-scale sequencing

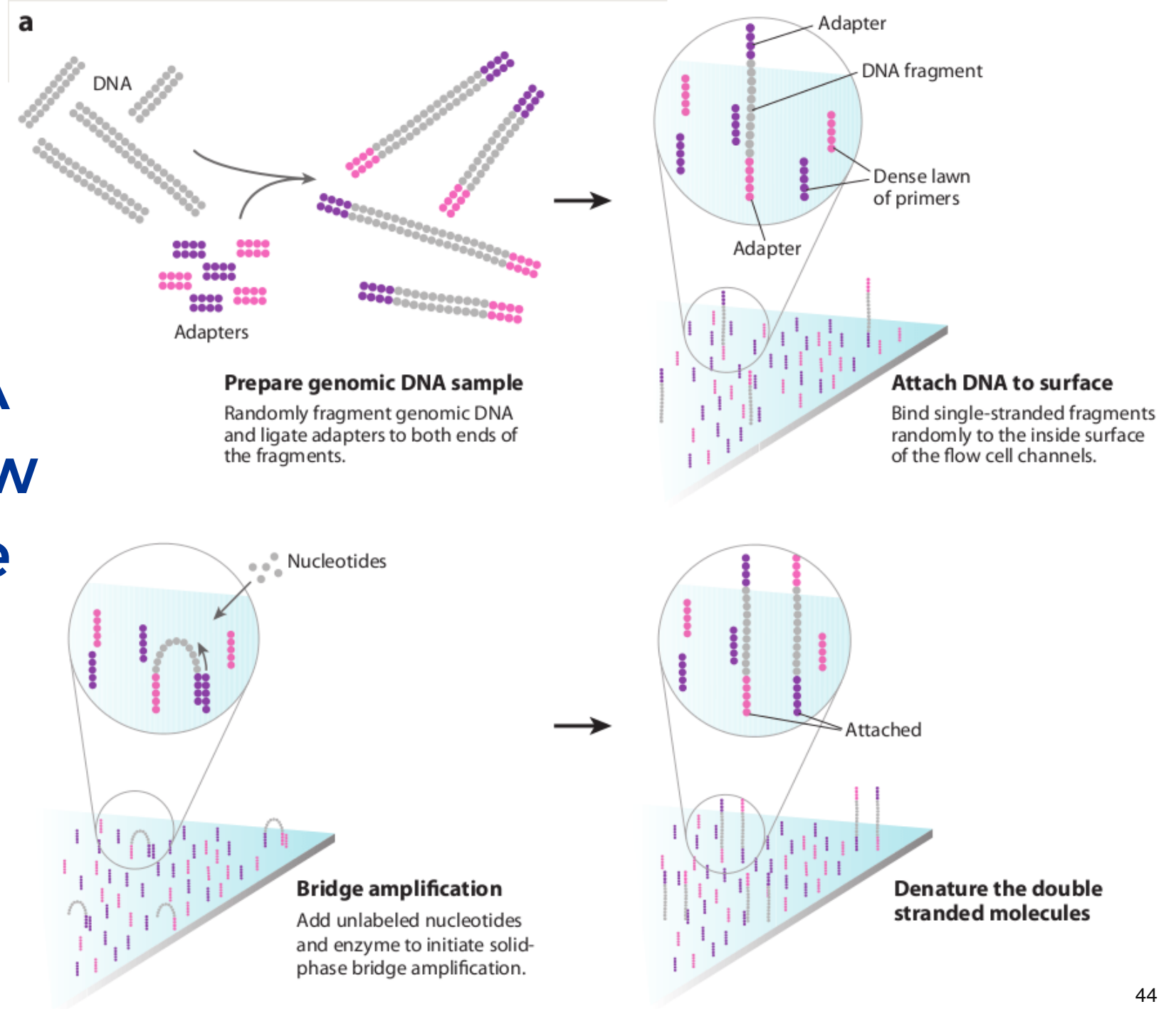
What if we could identify nucleotides as they are being added, allowing us to sequence faster and at a larger scale?

Sequencing by synthesis identifies nucleotides as DNA strands are being synthesized



Calls	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Sequencing reads
1	A	C	A	–	ACA.....
2	A	T	C	–	ATC.....
3	G	A	T	–	GAT.....

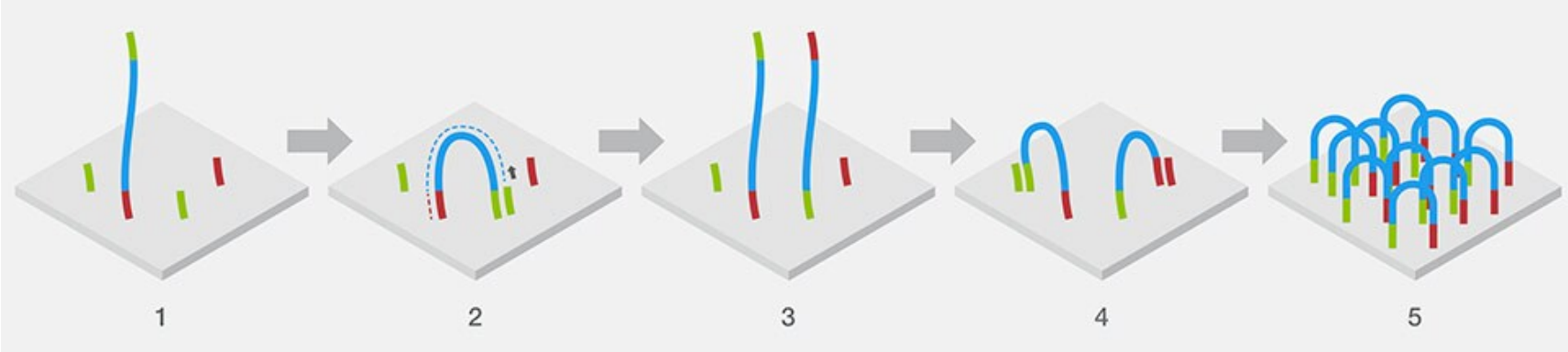
Immobilizing DNA fragments on a flow cell enables stable signal detection



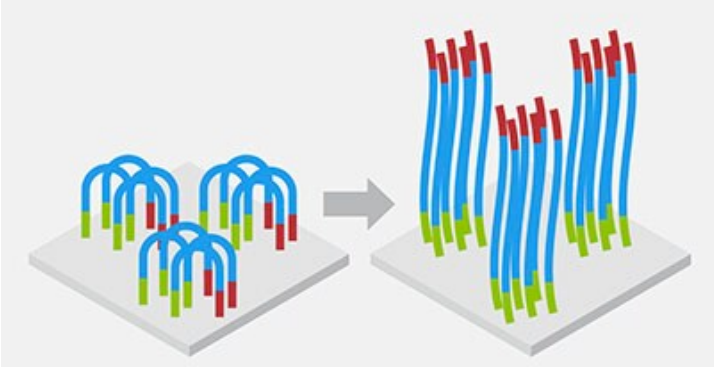
Bridge amplification generates clusters of identical DNA fragments, amplifying the signal for detection

Even with immobilization, the signal from a single fragment is often too weak to detect

Bridge amplification creates double-stranded bridges



Double-stranded clonal bridges are denatured with cleaved reverse strands



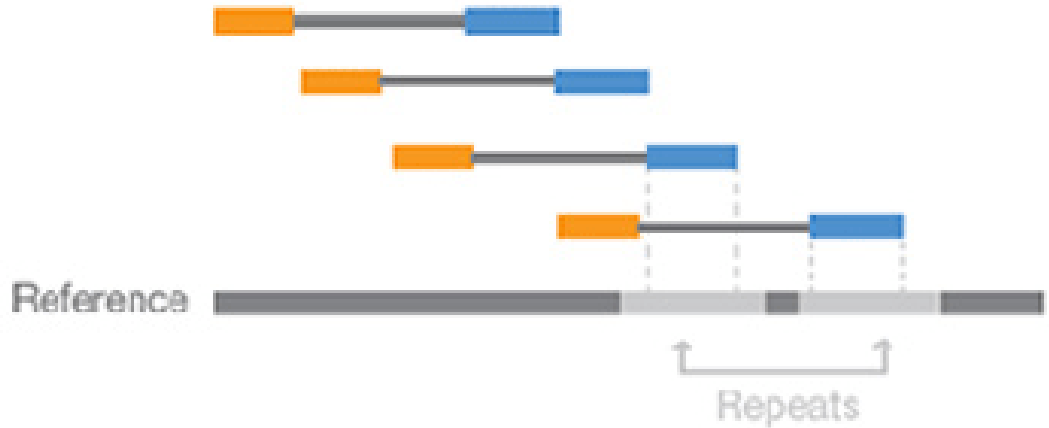
Clusters will give off a stronger signal compared to a single fragment

Paired-End Reads

Alignment to the Reference Sequence

Forward

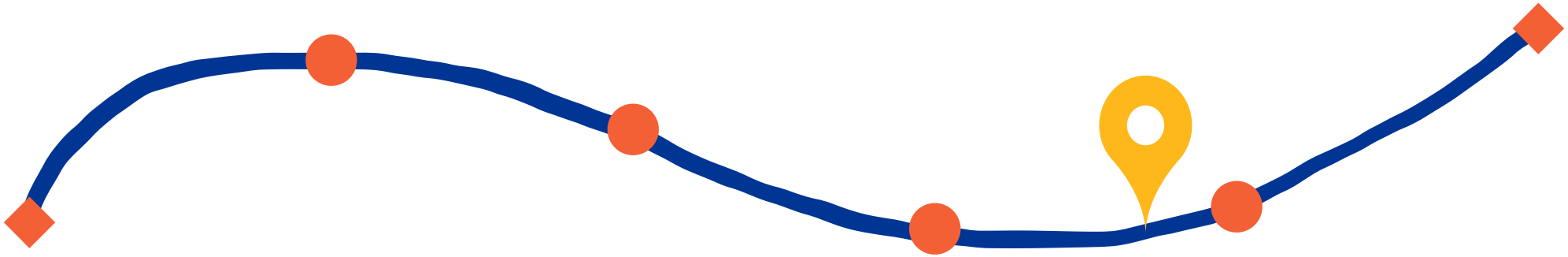
Reverse



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

More on this in later lectures

After today, you should have a better understanding of



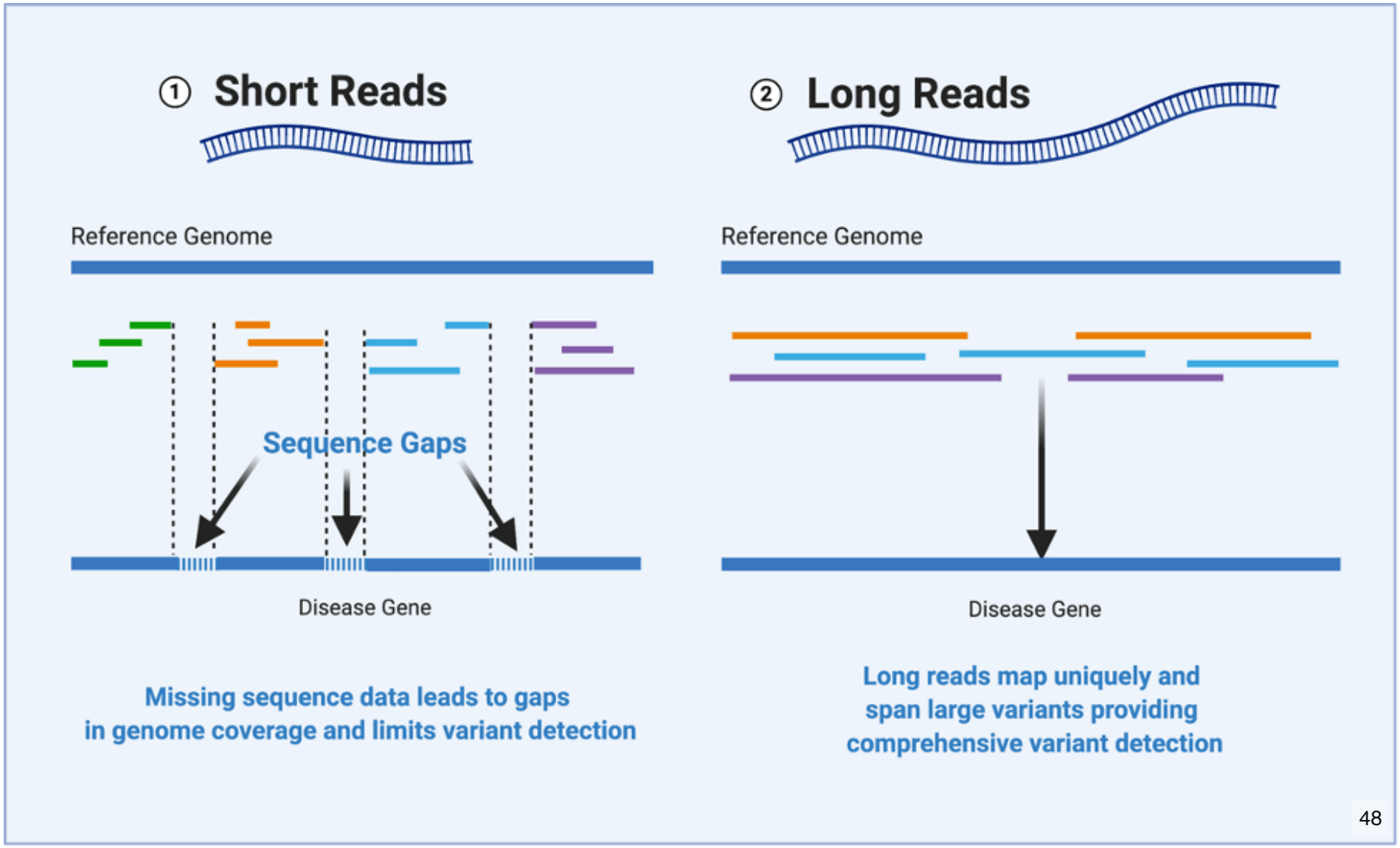
Principles and innovations of

DNA sequencing technologies

Single molecule sequencing (Nanopore)

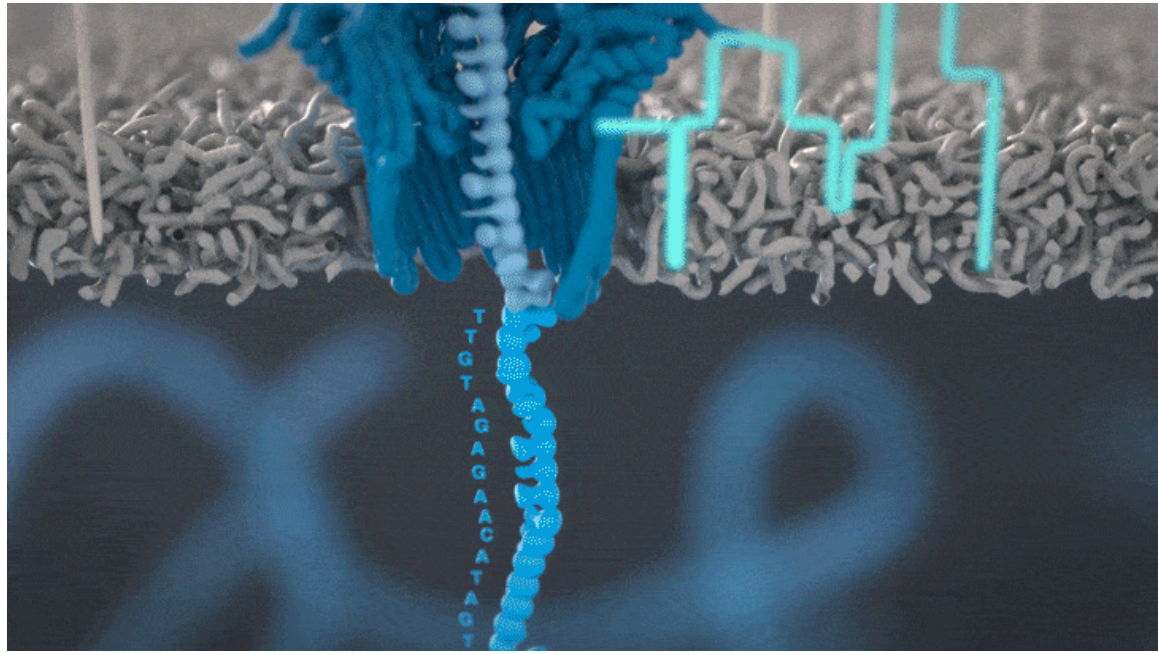
Illumina sequencing is cost-effective, scalable, and highly parallel, but limited by short read lengths

Short DNA reads make genome assembly difficult, especially in repetitive regions

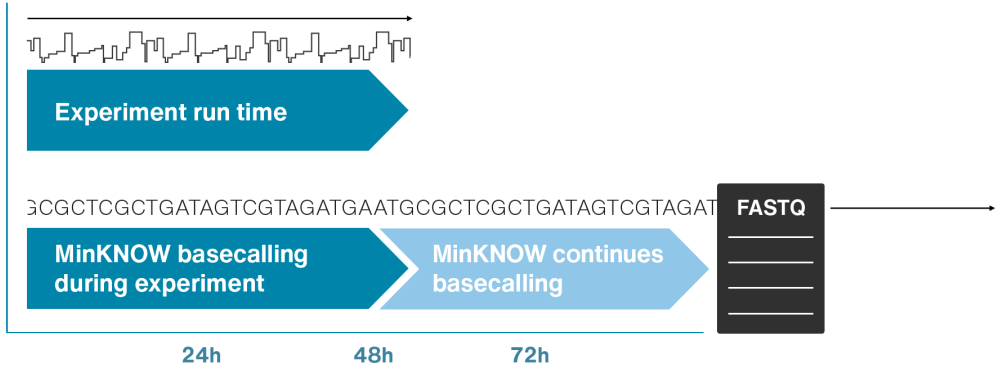


Single-molecule sequencing enables long-read sequencing by reading DNA molecules directly

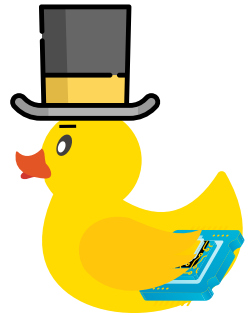
Nanopore sequencing detects nucleotide sequences by measuring changes in ionic current as DNA passes through a pore



- DNA passes through a nanopore driven by an electric field.
- Each nucleotide disrupts ionic current in a unique, measurable way.
- Real-time signal capture translates into nucleotide sequence.



Match each modern sequencing technology with the correct combination of features or characteristics.

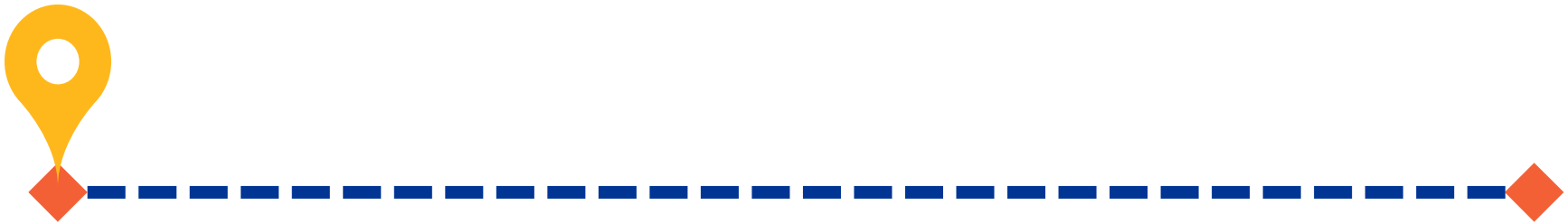


Please note: TopHat questions are ungraded. Engaging honestly with the question will benefit you far more than any shortcuts.

Before the next class, you should

Lecture 02A:
DNA sequencing - Foundations

Lecture 02B:
DNA sequencing - Methodology



Today

Thursday

- P01A is due Friday, Jan 17th
- P01B will be released Friday, Jan 17th
- CByte 01 will be released Friday, Jan 17th