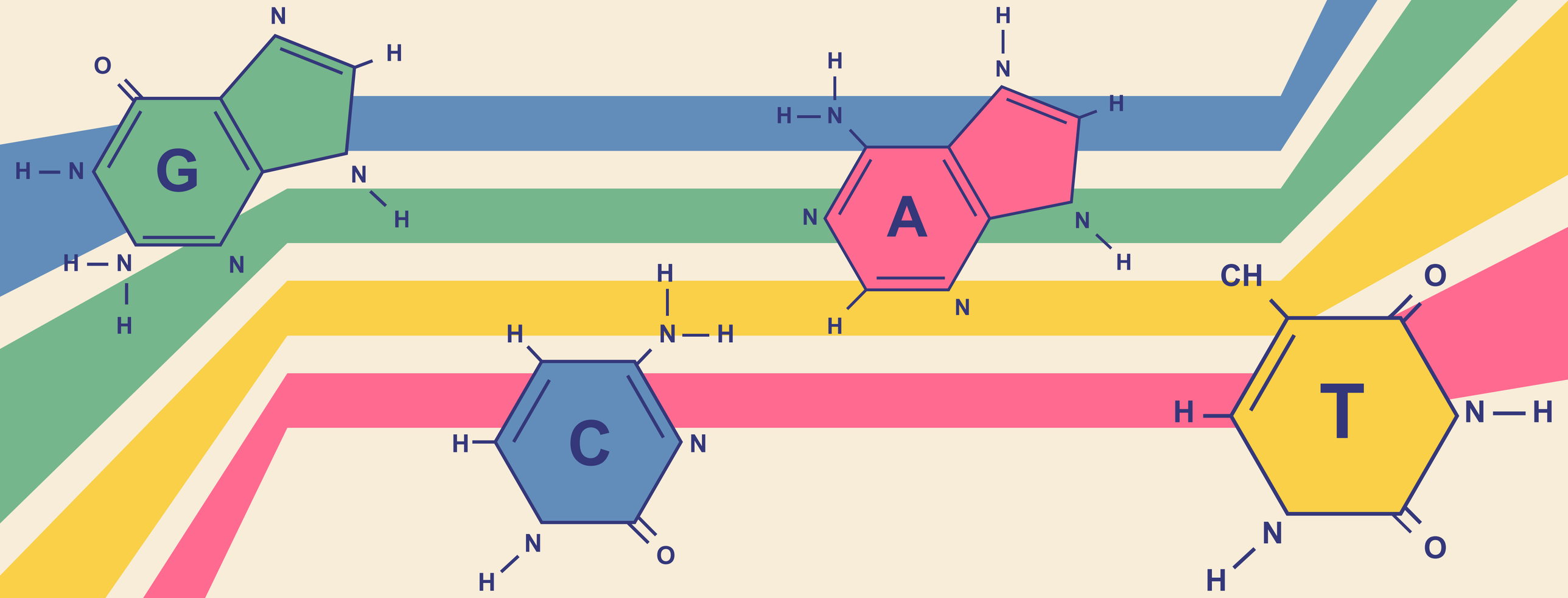
From Code to Genomes

: A Bioinformatics Adventure



By the end of today

- Terminiology
- How to navigate Sonic HPC
- Basic linux commands
- How to run a simple program

High Throughput Squencing

‘Capable of sequencing multiple DNA molecules in parallel, enabling hundreds of millions of DNA molecules to be sequenced at a time’

High Throughput Squencing

Also known as

- Second generation sequencing (IonTorrent, Illumina)
- Third generation sequencing (PacBio, Oxford Nanopore Technologies)
- Next generation sequencing

Terminology

16S rRNA or amplicon Sequencing

- Targeted to a specific gene or organism or both
- Typically involves PCR for amplification of target region
- Cheap and cheerful

Shotgun metagenomics

- Non targeted -agnostic
- Process is platform dependent
- Typically, involves fragmentation of sample to a specific size range

Whole genome sequencing

- Process is platform dependent
- Isolation of species of interest through traditional cultural based methods

Advantages

16S rRNA or amplicon Sequencing

- Can be performed on complex samples
- Amplification increases probability of capturing low-copy number events
- High confidence in presence/absence detection

Shotgun metagenomics

- Can be performed on complex samples
- Non-targeted: captures everything within a sample
- Lack of PCR results in a stronger correlation between read number and biological reality

Whole genome sequencing

- Can be run on all sequencing platforms
- Full length, high quality genomes can be obtained from most platforms

Disadvantages

16S rRNA or amplicon Sequencing

- You only capture the region you target - prone to primer bias and amplification errors
- Low correlation between read numbers and true biological presence
- Limited to categorical data outputs i.e. taxonomy

Shotgun metagenomics

- You may miss your desired target due to the presence of RNA/DNA from other organisms
- Expensive
- Incomplete genome assemblies
- Overwhelming data analysis

Whole genome sequencing

- Requires high input concentration prior to sequencing
- You need to be able to culture your organisms
- Expensive

Use Case

16S rRNA or amplicon Sequencing

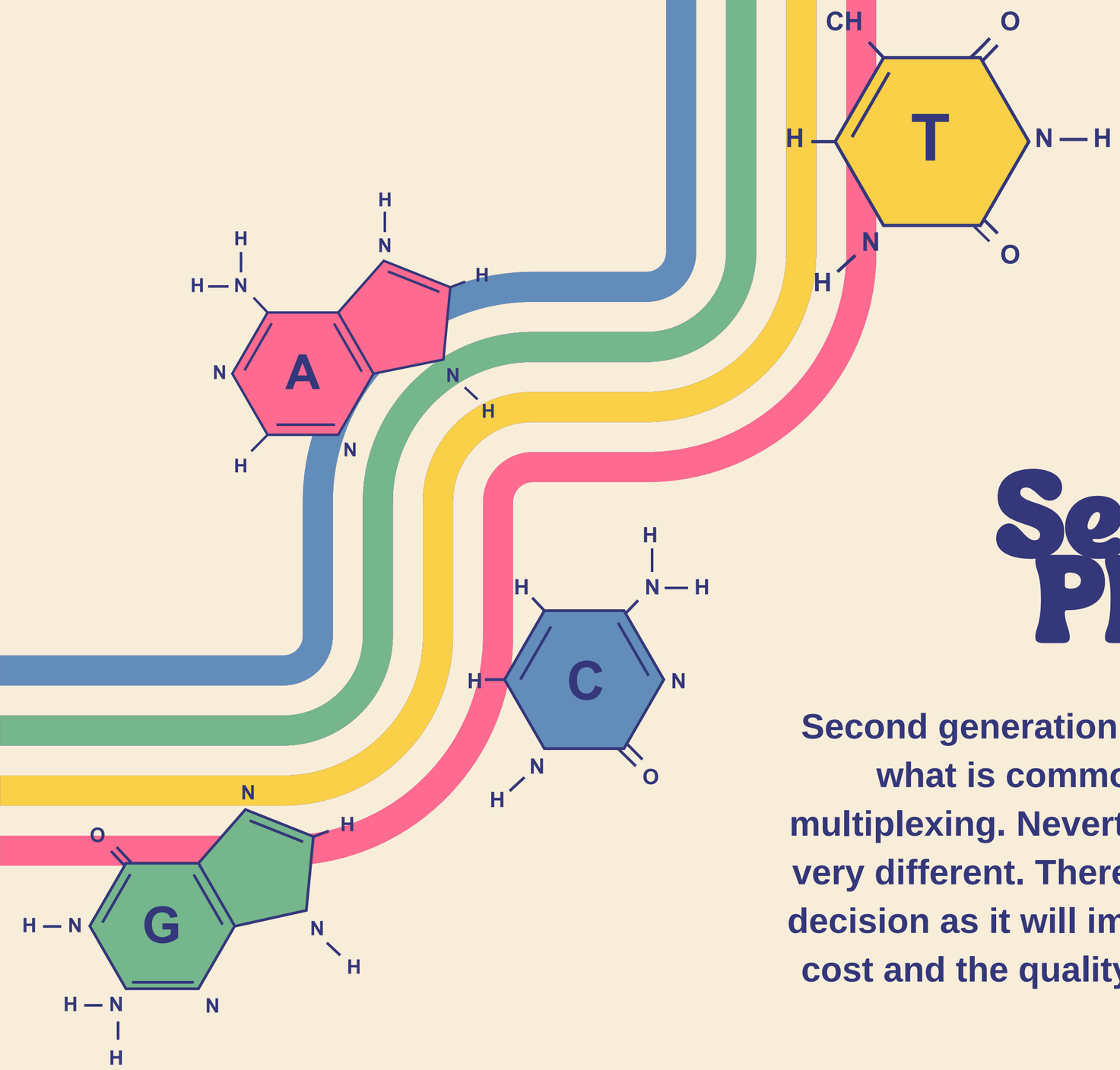
- Taxonomic classification to species level for a well characterized organism
- Presence/Absence of Bacteri/aArchaae at Genus level

Shotgun metagenomics

- Microbial Ecology studies: relationship between microorganisms in different environments
- Discovery projects - great way to find new species

Whole genome sequencing

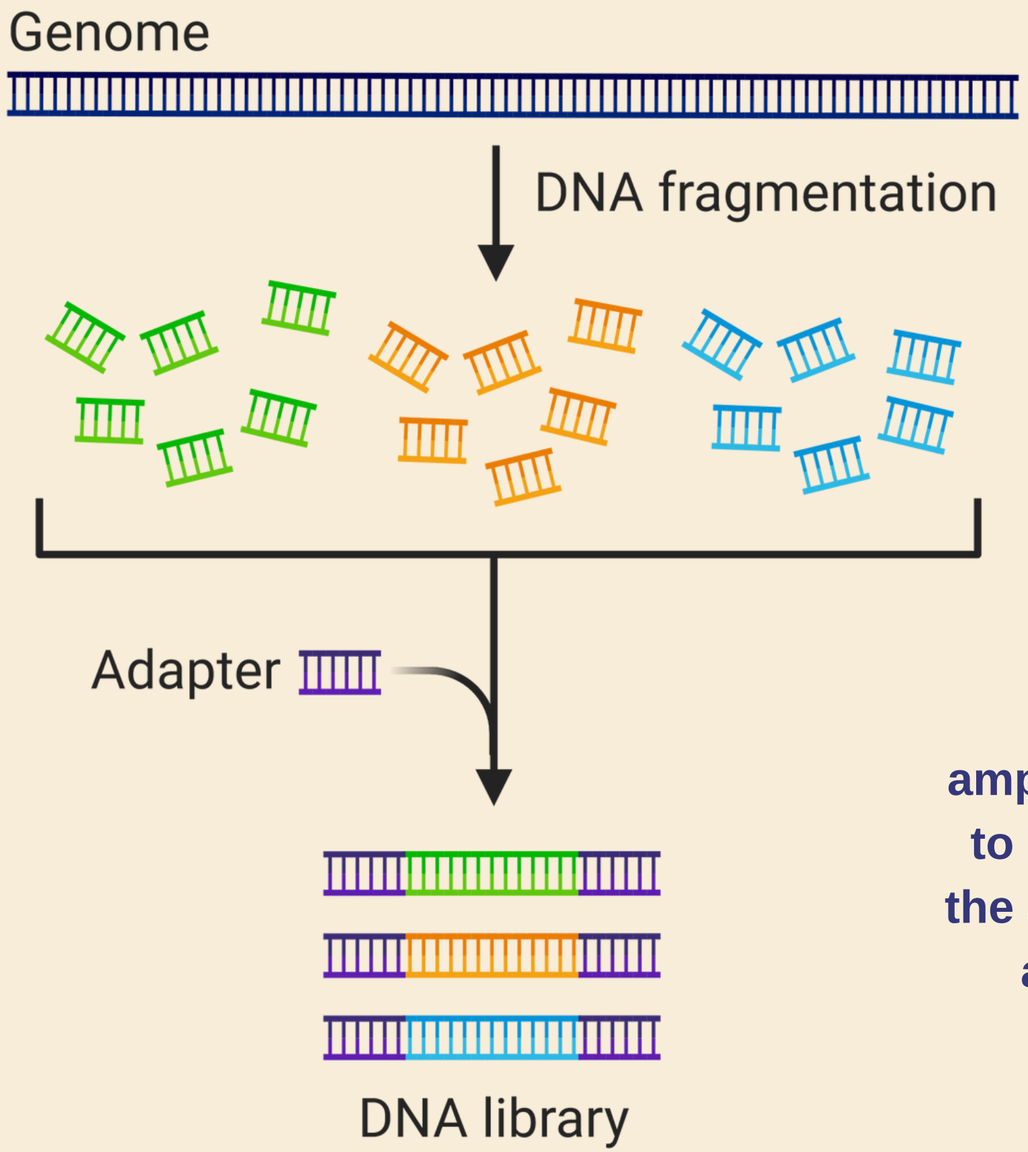
- Characterisation of a new species
- Generation of reference genomes for under studied organims

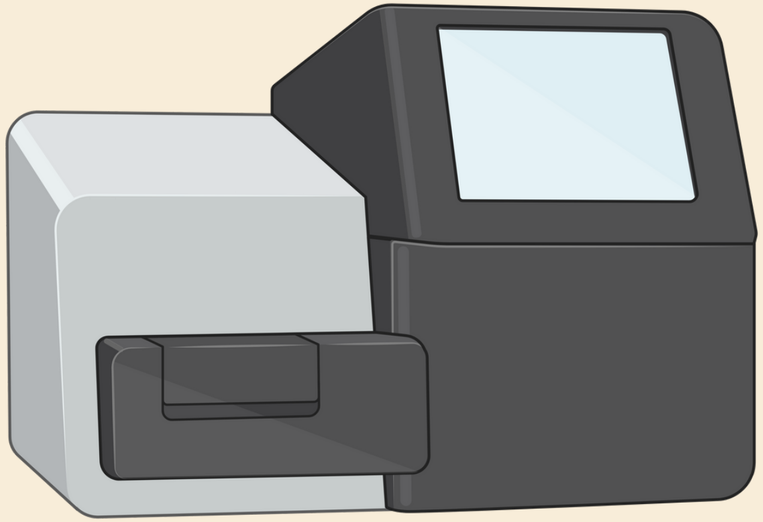


Sequencing Platforms

Second generation sequencing technologies can all perform what is commonly referred to as metabarcoding, i.e multiplexing. Nevertheless, how they sequence nucleotides is very different. Therefore, platform choice is a really important decision as it will impact the expected biases in your data, the cost and the quality of the subsequent genomic information.

① Library preparation

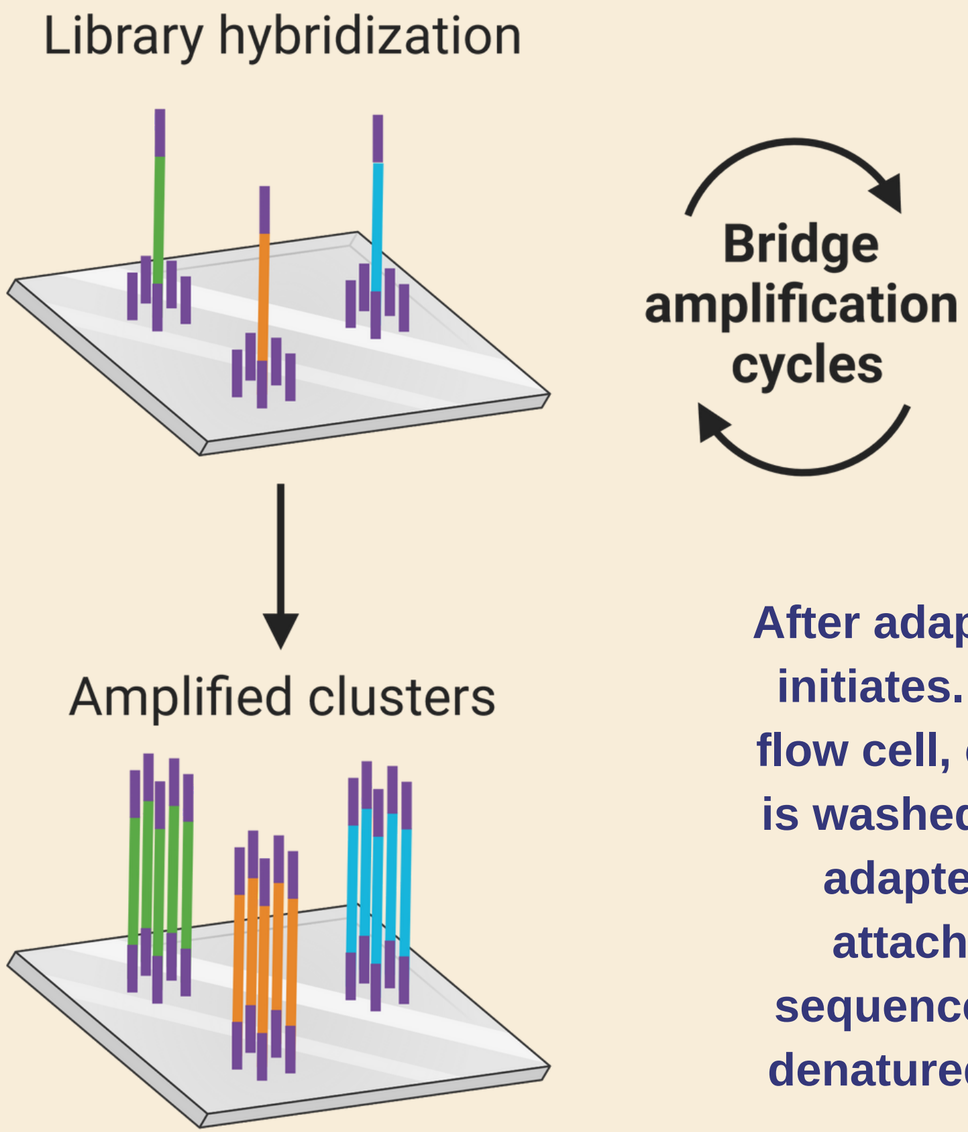




Illumina

Illumina relies on sequencing by synthesis with bridge amplification in order to generate DNA. In order for this process to happen, adapters (DNA sequences) are added to the end of the amplicons/fragements added during library preparation. The adapter sequences will be immobilized onto the flow cell.

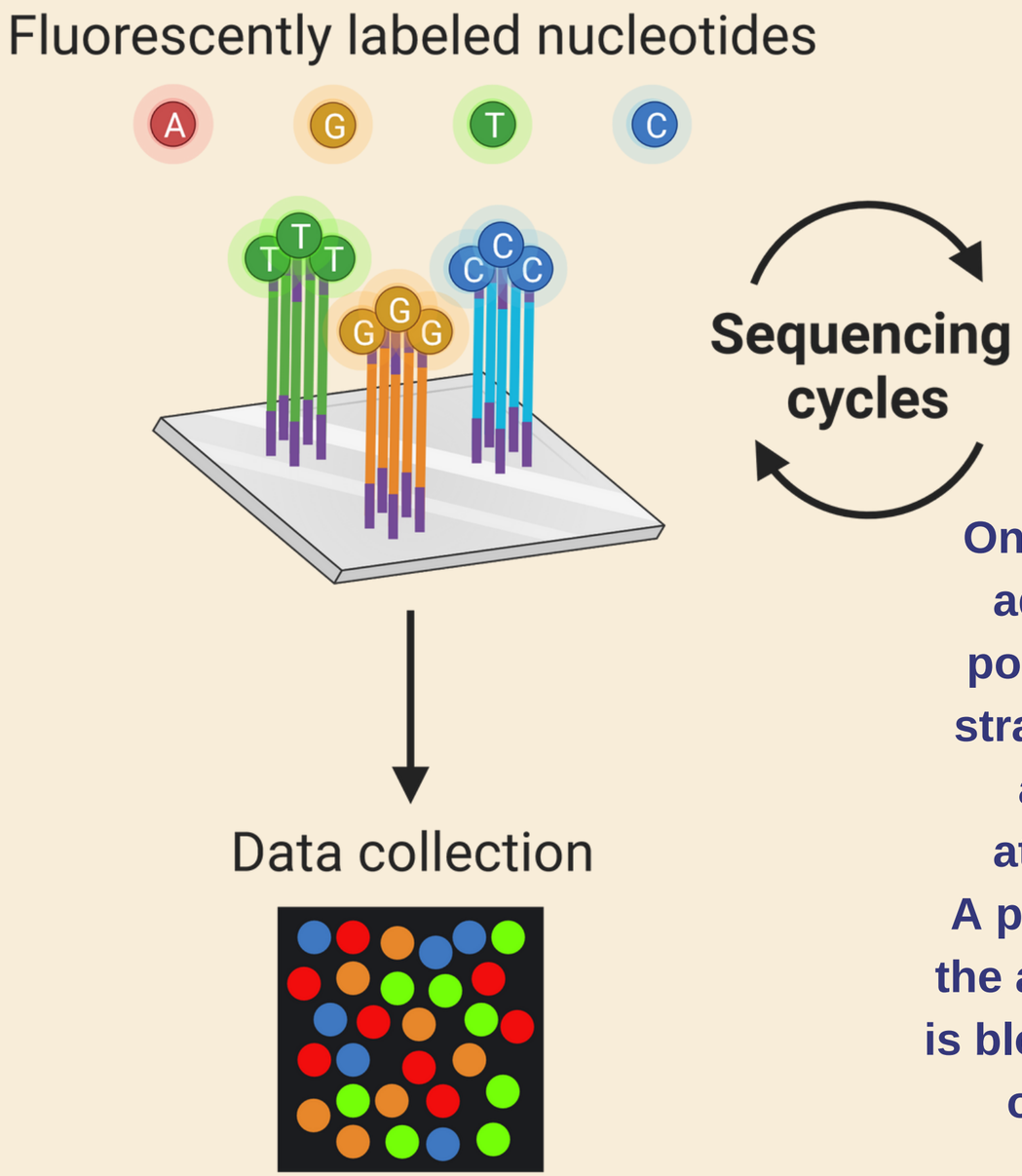
② DNA library bridge amplification



Illumina

After adapter immobilization to the flow cell, cluster generation initiates. A DNA polymerase ascends the strand linked to the flow cell, creating a complementary strand. The original strand is washed away, leaving the reverse strand. At its top, another adapter sequence is present. The DNA strand bends and attaches to the complementary oligo for the top adapter sequence, resulting in a dsDNA strand. This dsDNA strand is denatured by polymerases. This process repeats many times.

③ DNA library sequencing

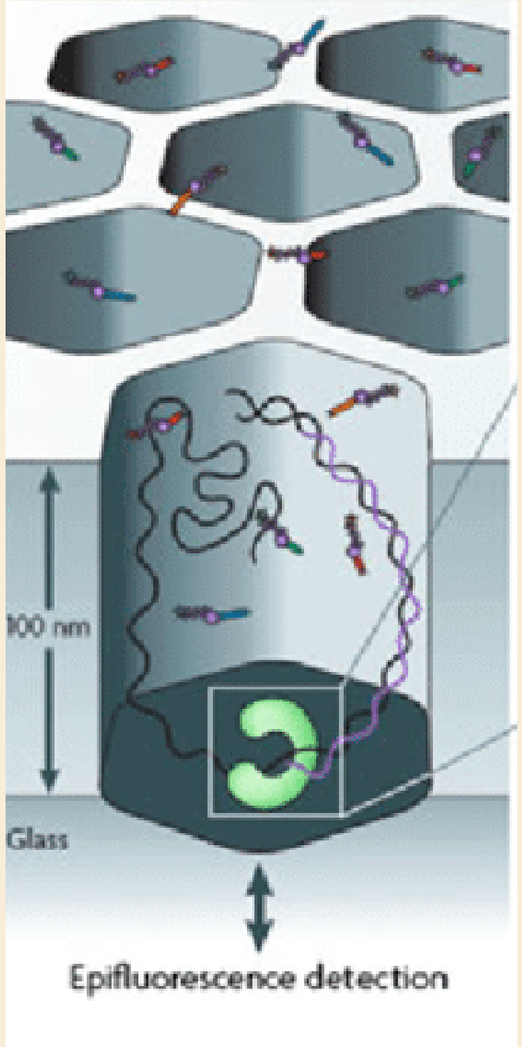


Illumina

Once the DNA strand has been read, the strand that was just added is washed away. Then, the index 1 primer attaches, polymerizes the index 1 sequence, and is washed away. The strand forms a bridge again, and the 3' end of the DNA strand attaches to an oligo on the flow cell. The index 2 primer attaches, polymerizes the sequence, and is washed away. A polymerase sequences the complementary strand on top of the arched strand. They separate, and the 3' end of each strand is blocked. The forward strand is washed away, and the process of sequence by synthesis repeats for the reverse strand.

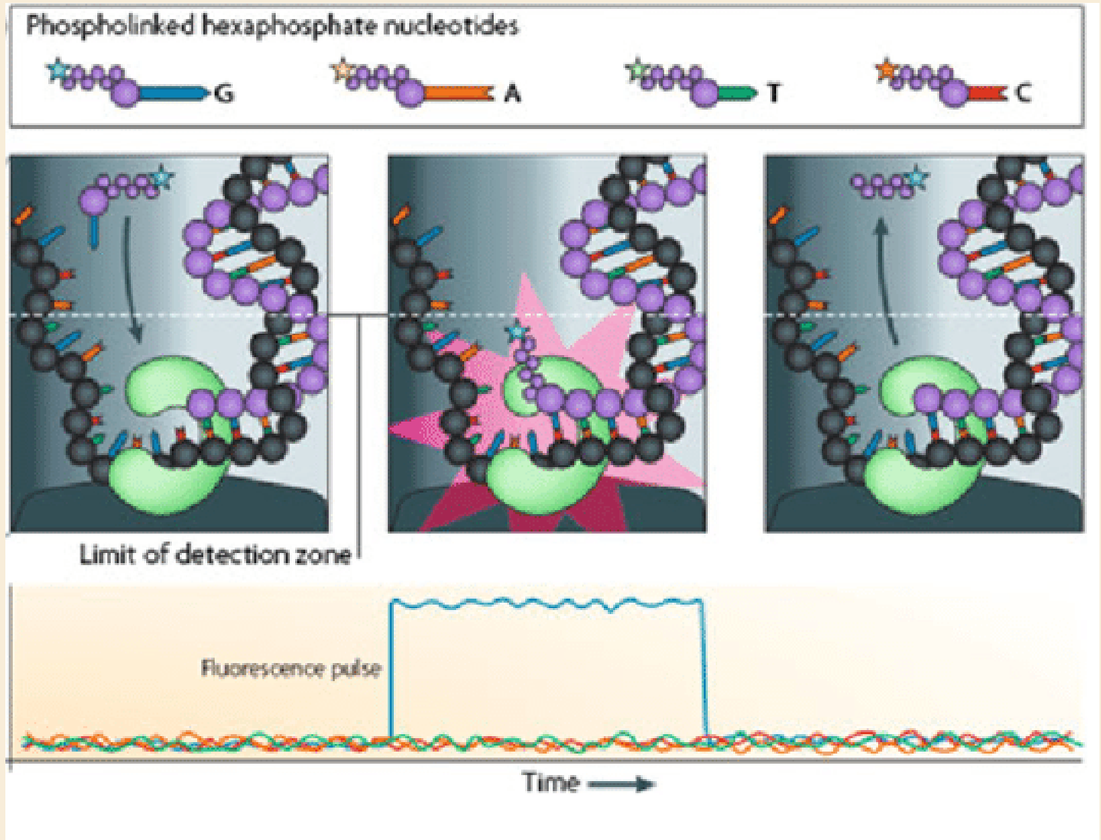
PacBio (3rd gen)

- ZMWs are subwavelength optical nanostructures in a thin metallic films. They are very powerful and capable of confining excitation volume to attoliter range.
- ZMWs can be used to isolate individual molecules for optical analysis at physiologically relevant concentrations.

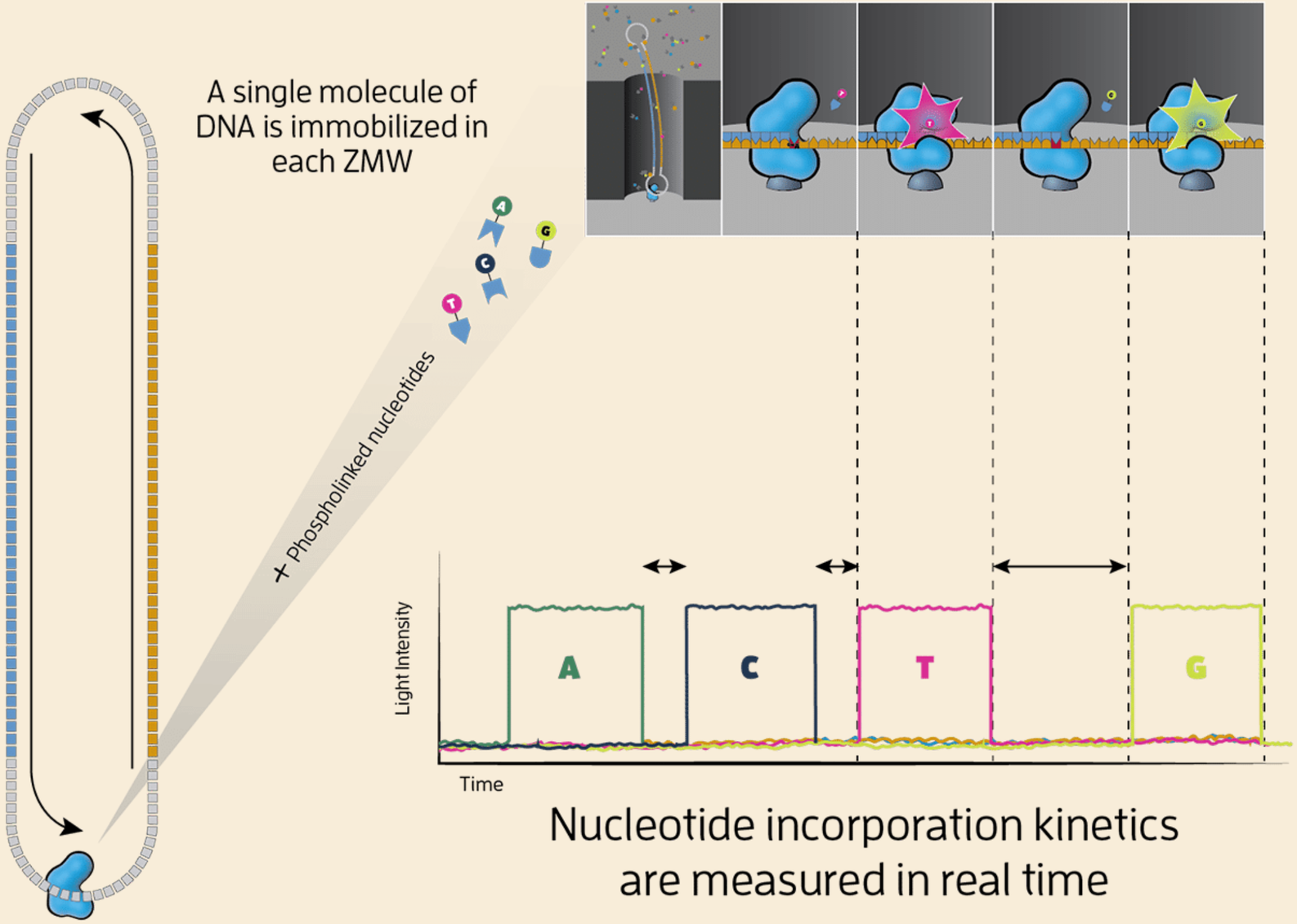


PacBio (3rd gen)

- PacBio uses arrays of ZMWs for real-time analysis of single-molecule reactions or binding events
- Circuarlised pieces of sample DNA are immobilized at the bottom of the glass surface of the ZMWs and once free floating nucleotides are added, the DNA polymerase attached during library preparation begins to copy the template.
- The light emitted through the ZMW indicates which bases have been generated.



PacBio (3rd gen)

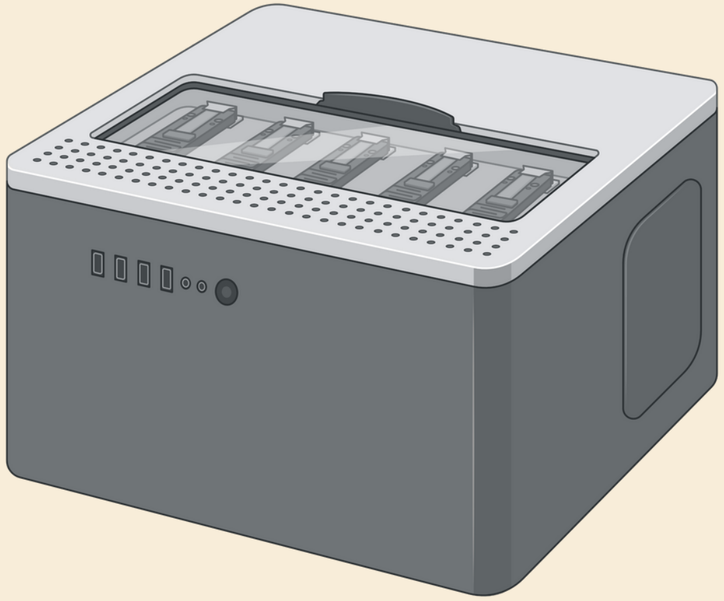


As anchored polymerases incorporate labeled bases, light is emitted

Directly detect DNA modifications during sequencing

Nucleotide incorporation kinetics are measured in real time

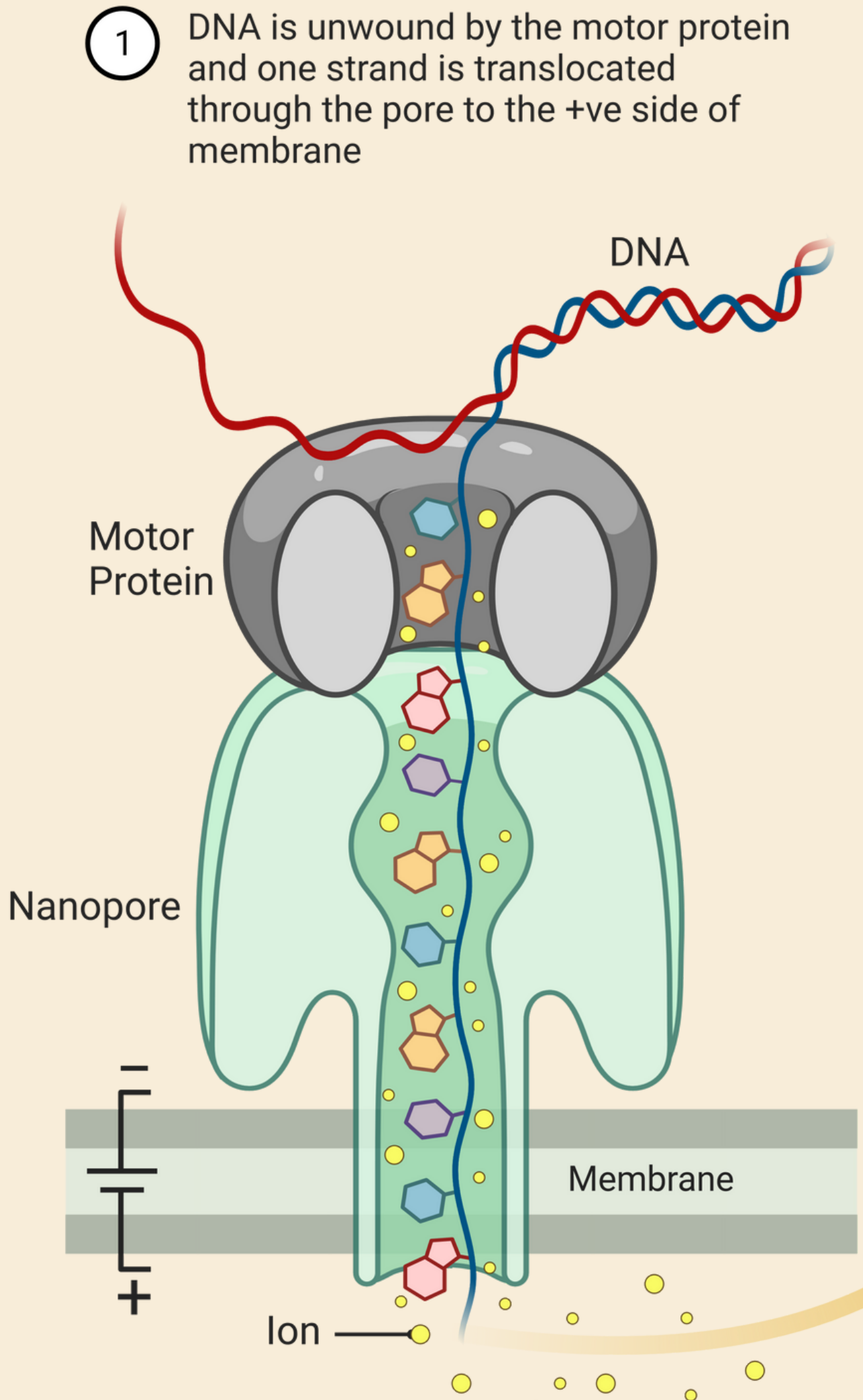
Oxford Nanopore Technology (3rd gen)



ONT sequencing utilizes nanopores, which are small holes at the nanometer scale, embedded in a synthetic membrane.The nanopore serves as a microscopic tunnel through which a single-stranded DNA molecule passes.

Oxford Nanopore Technology (3rd gen)

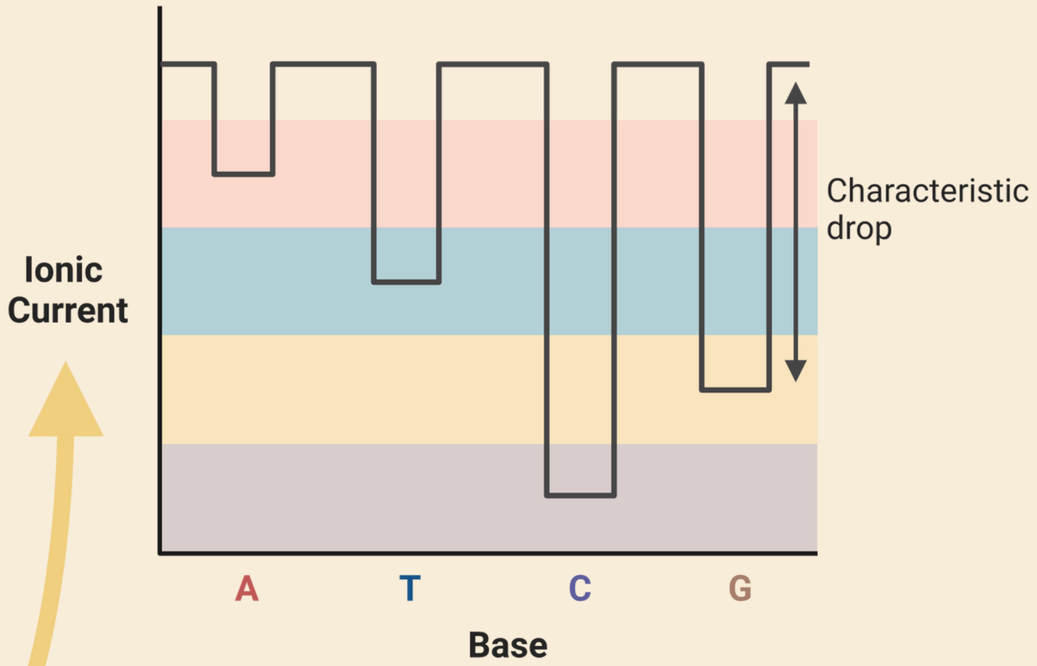
A single DNA strand is introduced to the nanopore. An electric current is applied across the nanopore, creating an electric field. As the DNA strand threads through the nanopore, the electric field causes the individual nucleotides to disrupt the current in a manner that is characteristic of each nucleotide.



Oxford Nanopore Technology (3rd gen)

As the DNA nucleotides pass through the nanopore, they cause characteristic disruptions or modulations in the electric current. These disruptions are detected and recorded as electrical signals by specialized sensors.

The recorded electrical signals are then translated into DNA base sequences. Each of the four DNA bases (adenine, thymine, cytosine, and guanine) generates a unique signal pattern, allowing for real-time base identificatio



2 Each base gives a characteristic reduction in the ionic current, allowing the DNA to be sequenced

Use Case

Illumina

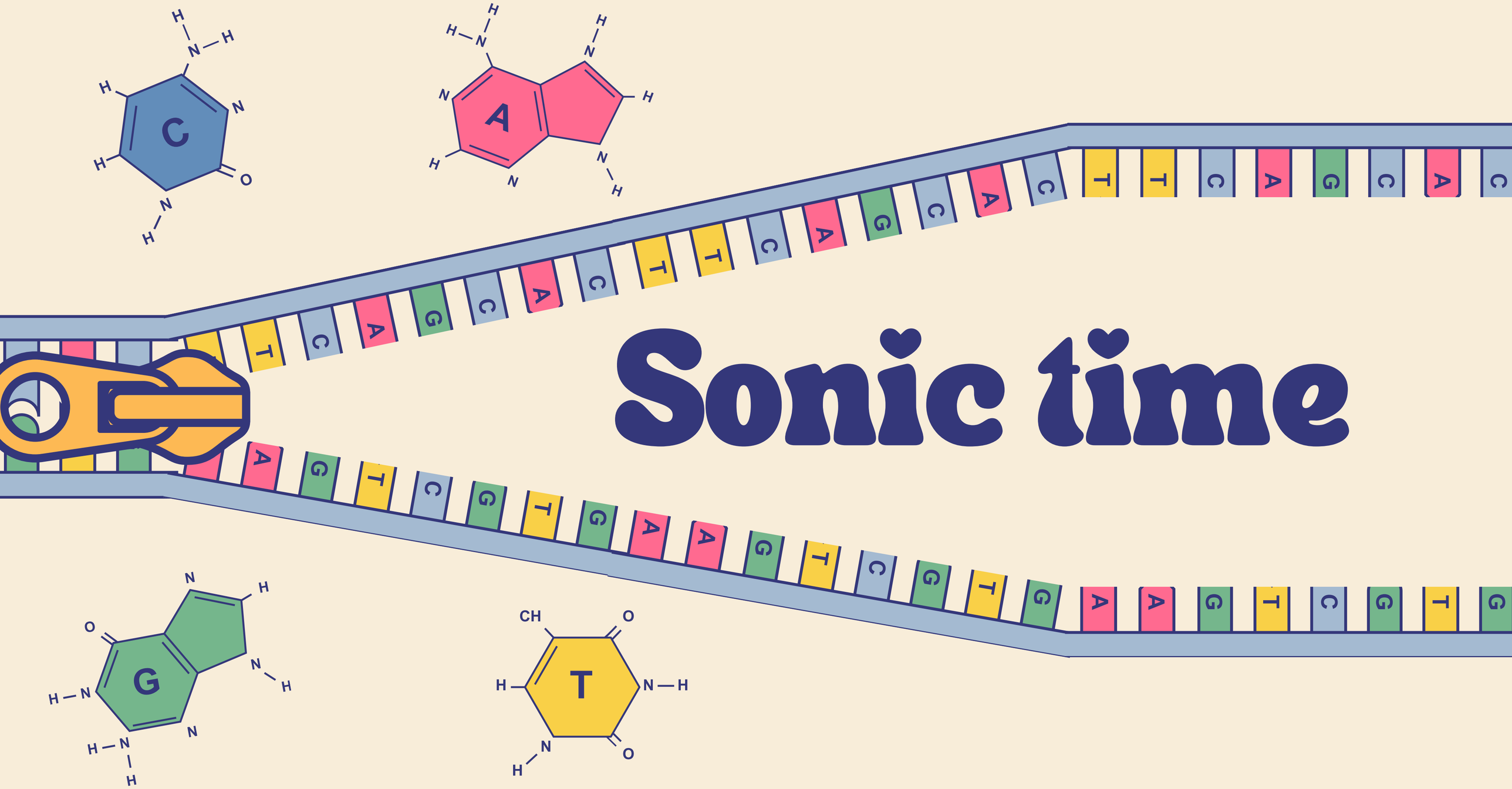
- Need a method that can be approved within accredited systems
- Working with organisms without G/C or A/T rich regions
- High sample volumes
- SNP identification

PacBio

- Money-rich, time-poor
- Working with large genomes
- Reference genomes generation
- Novel species genomes
- SNP identification

ONT

- Just poor
- Working with large genomes
- Novel species genomes
- Working with organisms with G/C or A/T rich regions
- Need real-time results



Sonic time

Step 1:

Install VS Code on your computer

Step 3:

Sign up for GitHub account

Next steps

Link your VS Code to your GitHub account

Do this via the accounts and manage tab in the bottom left hand corner of VS Code

Add Extensions to VS Code

- Remote-SHH
- Remote-SSH Editing Configuration Files
- Remove Explorer

Set up Config File

Instructions coming ....

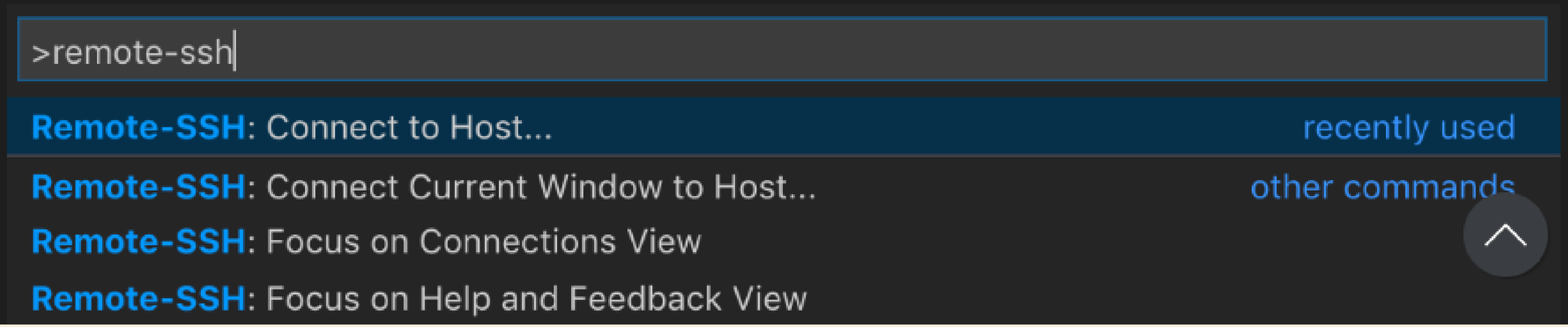
Logon to the Sonic HPC

ssh username@login.ucd.ie

Follow the step-by-step tutorial or if you have a simple SSH host setup, connect to it as follows:

1. Press `F1` and run the **Remote-SSH: Open SSH Host...** command.
2. Enter your user and host/IP in the following format in the input box that appears and press enter: `user@host-or-ip` or `user@domain@host-or-ip`
3. If prompted, enter your password (but we suggest setting up key based authentication).
4. After you are connected, use **File > Open Folder** to open a folder on the host.

You can press `F1` to bring up the Command Palette and type in `Remote-SSH` for a full list of available commands.



Set up Config File

- **Head to the command bar and write >remote-SSH: you should see a list of commands pop up, you should pick: remote-SSH:Set up Config File**