Swiss Institute of Bioinformatics

INTRODUCTION TO SEQUENCING DATA ANALYSIS

Sequencing Technologies

Deepak Tanwar
Frédéric Burdet

April 23-25, 2025

Adapted from previous year courses

elixir SWITZERLAND

Learning objectives

Understand the principles behind major DNA/RNA sequencing technologies

Identify applications of each sequencing technology in research

Evaluate technology limitations, especially those affecting read length and sequencing accuracy

Select appropriate sequencing methods for different genomic and transcriptomic analyses based on the experimental need

Quiz 1

What is the primary function of sequencing technologies?

A. Synthesizing DNA
B. Identifying proteins in cells
C. Determining the order of nucleotides in genetic material
D. Measuring gene expression levels

Sequencing technology refers to the various methods used to determine the order of nucleotides (the building blocks of DNA and RNA) in a strand of genetic material.

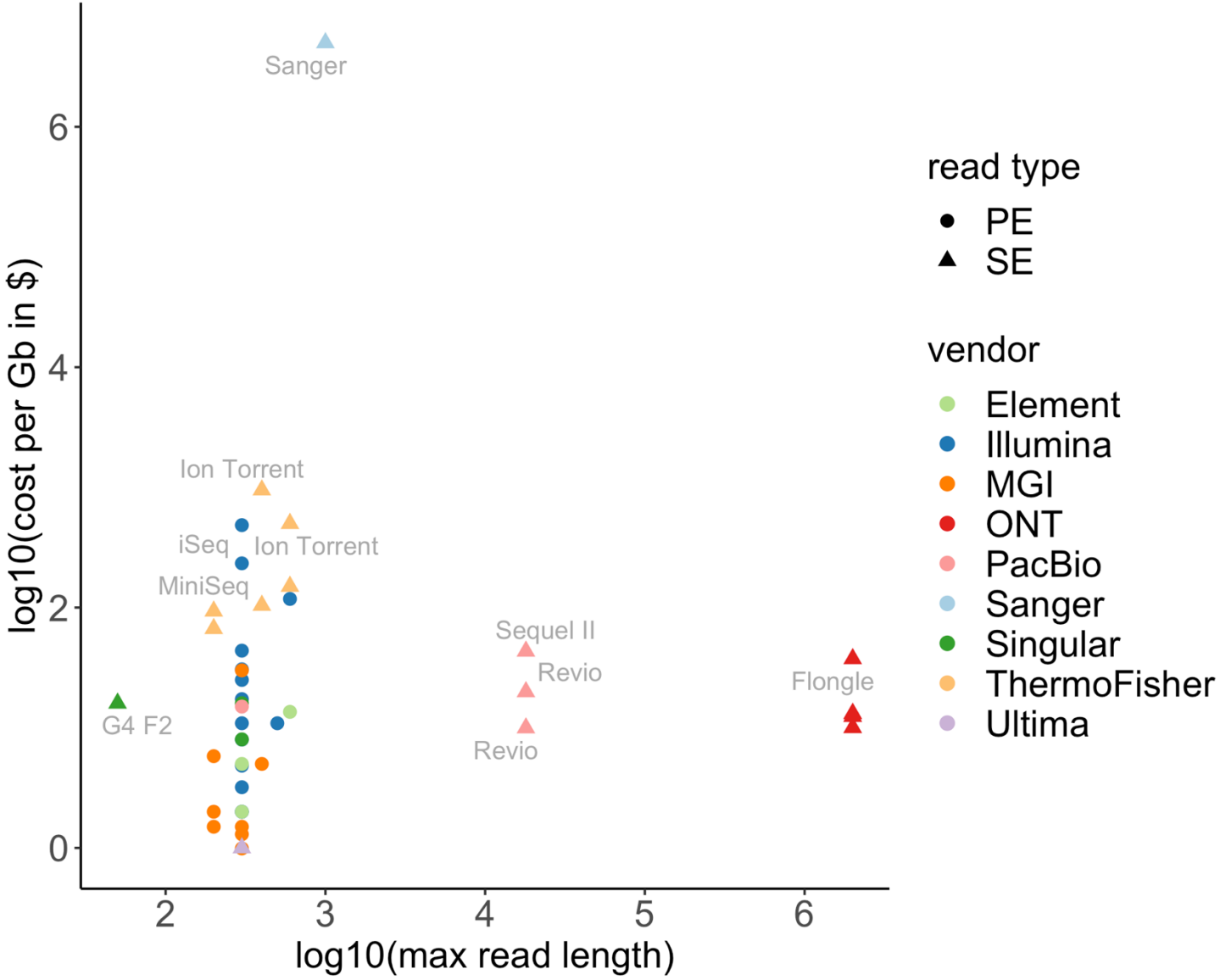
Sequencing technologies

Sanger sequencing

Second generation sequencing

Third generation sequencing

Comparison of Sequencing Technologies by Cost and Read Length

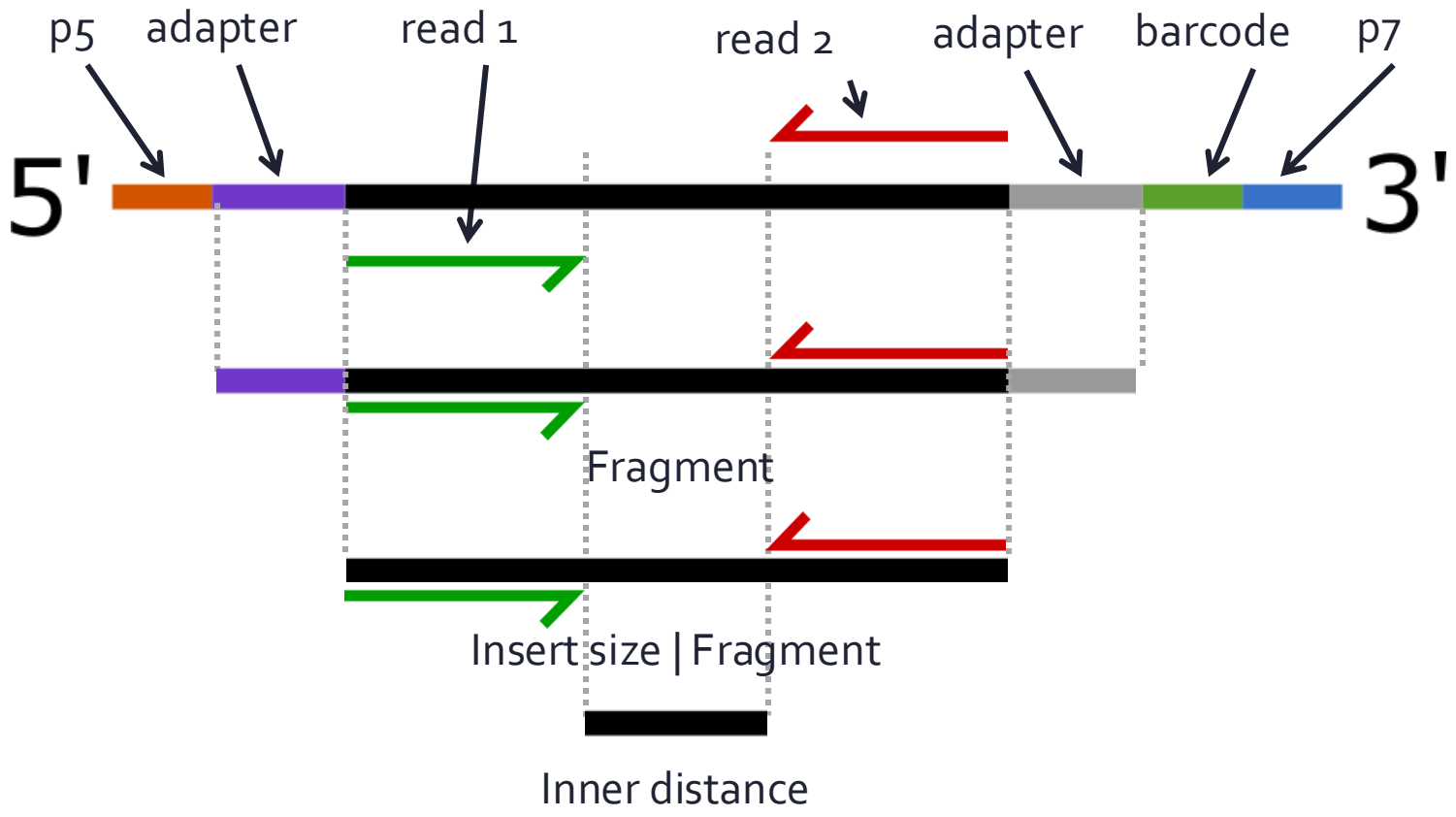


read type
- PE
- SE

vendor
- Element
- Illumina
- MGI
- ONT
- PacBio
- Sanger
- Singular
- ThermoFisher
- Ultima

drawn from: https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/ Albert Vilella

Some definitions



Some more definitions..

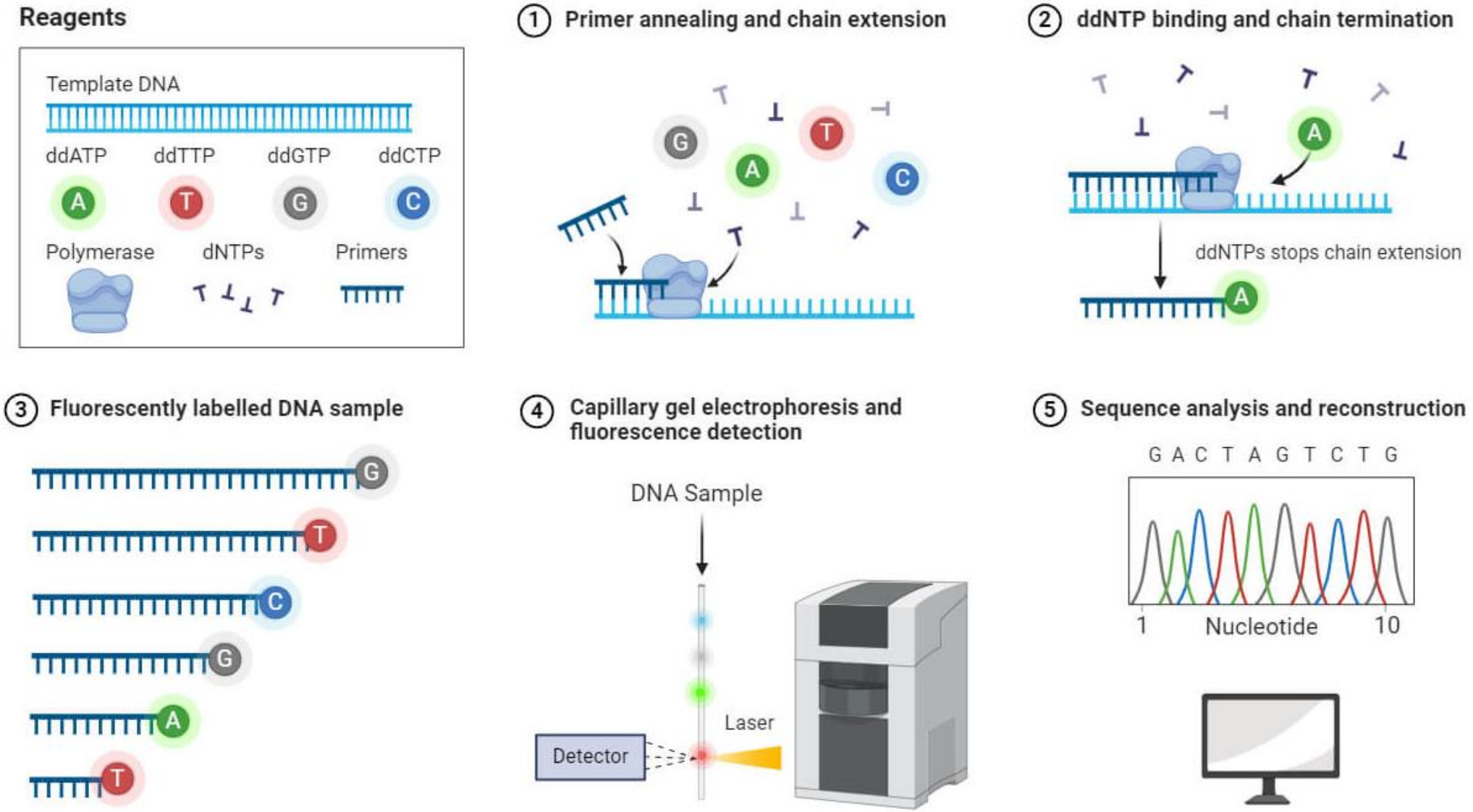
Library: fragments from one (c)DNA sample that share a barcode

Sequencing run: complete cycle of generating reads on a machine

Flow cell: physical platform where sequencing reactions take place. Used once in a sequencing run.

Lane: compartment within the flow cell. An Illumina flow cell often has multiple lanes (2 or 4)

Sanger sequencing



https://microbenotes.com/sanger-sequencing/

Sanger sequencing applications

Clinical Applications:

Genetic disease diagnosis (identifying mutations in specific genes)

Research Applications:

Validation of next-generation sequencing results

Forensic Applications:

DNA profiling for identification

Sanger sequencing applications

Clinical Applications:

Genetic disease diagnosis (identifying mutations in specific genes)

Research Applications:

Validation of next-generation sequencing results

Forensic Applications:

DNA profiling for identification

We are not covering Sanger sequencing in this course

Second generation sequencing

454 Pyrosequencing

- Discontinued due to technological advancements

Ion Torrent (semiconductor sequencing)

- This technology is faster and can be more cost-effective, but it generally has shorter read lengths and slightly lower accuracy compared to Illumina
- Up to ± 400 bp read length
- Homopolymers, such as TTTTT are impossible to sequence

Illumina (sequencing by synthesis)

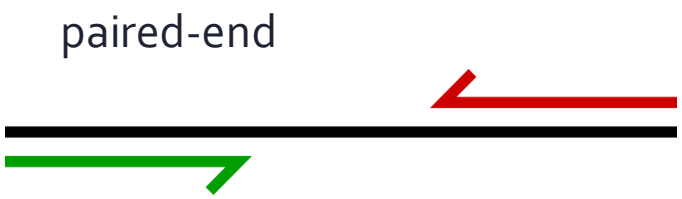
Illumina sequencing

Massive throughput: up to 16×10^{12} bases/run (NovaSeq X) = ~9,000 whole exomes

50 – 300 bp

Paired-end (or single-end)

Multiplexing

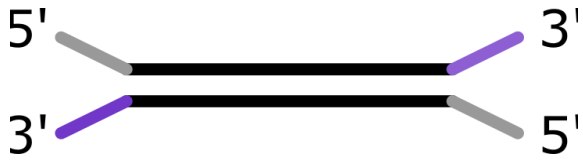


Illumina libray prep

shear + size select DNA

↓

Ligate adapters



↓

Barcode + p5/p7 sites

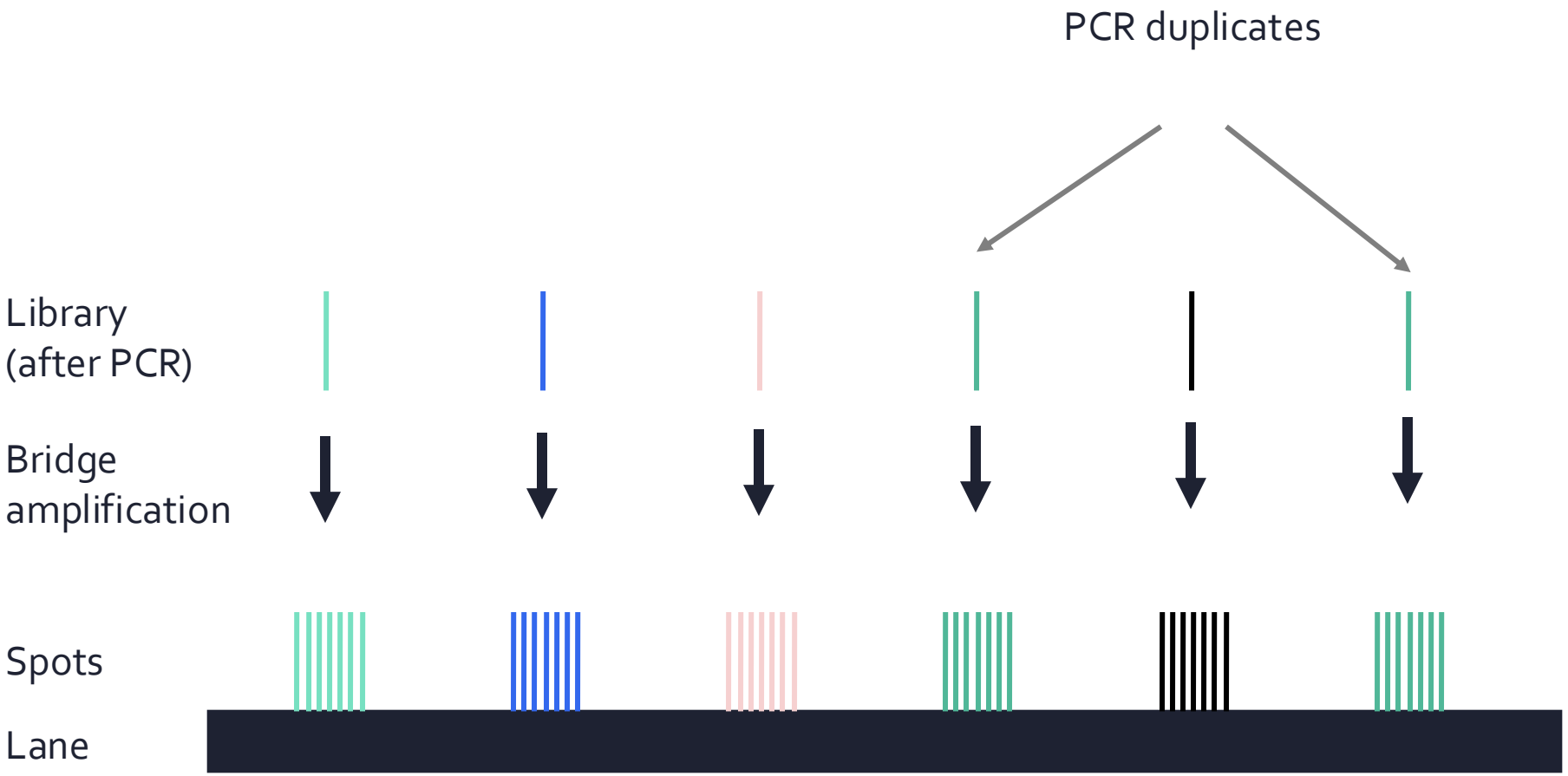
↓

PCR: 8-16 cycles

↓

Sequencing

Illumina Sequencing



Each spot represents one read pair

Illumina Sequencing by synthesis

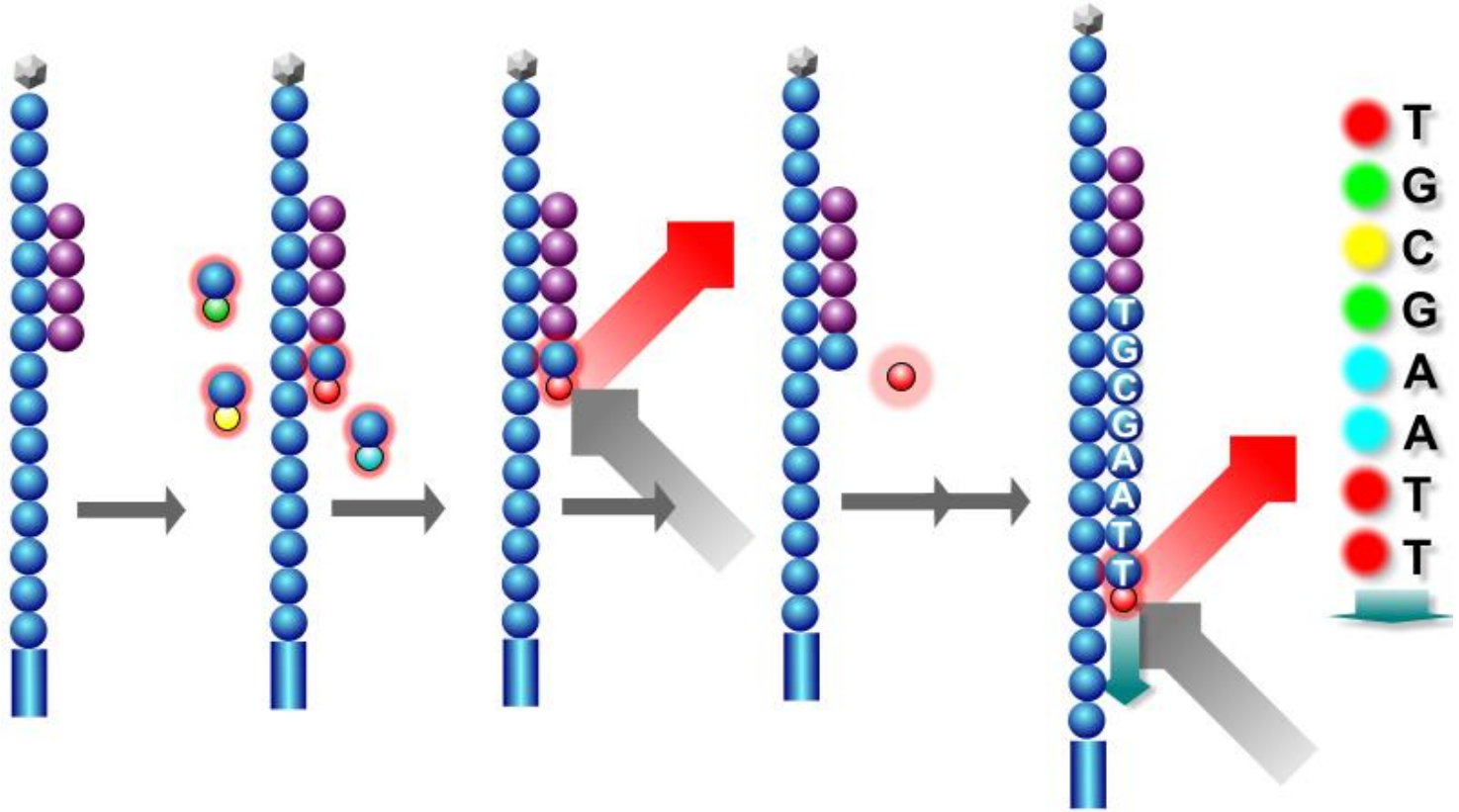


Image: Abizar Lakdawalla CC BY-SA 3.0

Second generation sequencing applications

Transcriptome characterization
e.g. RNA-seq – Gene expression, splicing, isoform detection

Epigenome characterization
e.g. ATAC-seq, Bisulfite-seq – Chromatin accessibility, DNA methylation

DNA-protein interactions
e.g. ChIP-seq – Transcription factor binding, histone modifications

Whole genome sequencing & assembly
e.g. short- and long-read WGS – De novo genome assembly, reference genome improvement

Second generation sequencing applications

Variant detection
e.g. Exome-seq, WGS – SNPs, indels, CNVs for disease association and diagnosis

Metagenome characterization
e.g. 16S rRNA sequencing, shotgun metagenomics – Microbiome studies, environmental genomics

Targeted sequencing
e.g. Amplicon-seq, hybrid capture panels – Focused gene panels for diagnostics

Second generation sequencing applications

Single-cell sequencing
e.g. scRNA-seq, scATAC-seq – Cell heterogeneity, developmental lineages, immune profiling

Spatial transcriptomics
e.g. 10x Visium, Slide-seq – Gene expression with spatial resolution in tissues

Single-cell epigenomics

e.g. scATAC-seq, scChIP-seq, scMethyl-seq – Chromatin accessibility, histone marks, methylation at single-cell level

Second generation sequencing applications

Multi-omics at single-cell level

e.g. SHARE-seq, 10x Multiome (RNA + ATAC), CITE-seq (RNA + protein)

Applications:

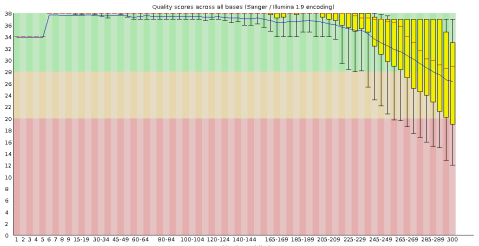
Linking transcriptome with epigenome or proteome

Understanding gene regulation networks

Immune and tumor microenvironment studies

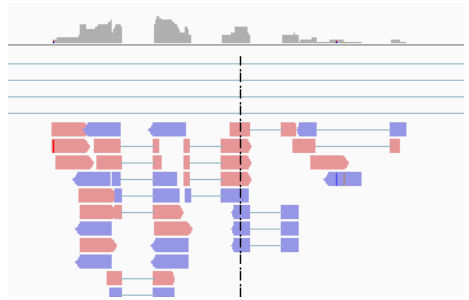
Simple workflow of data analysis

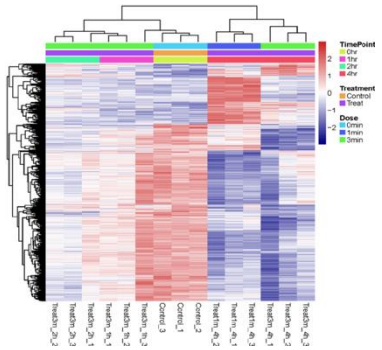
Sequencing



Quality control

Alignment



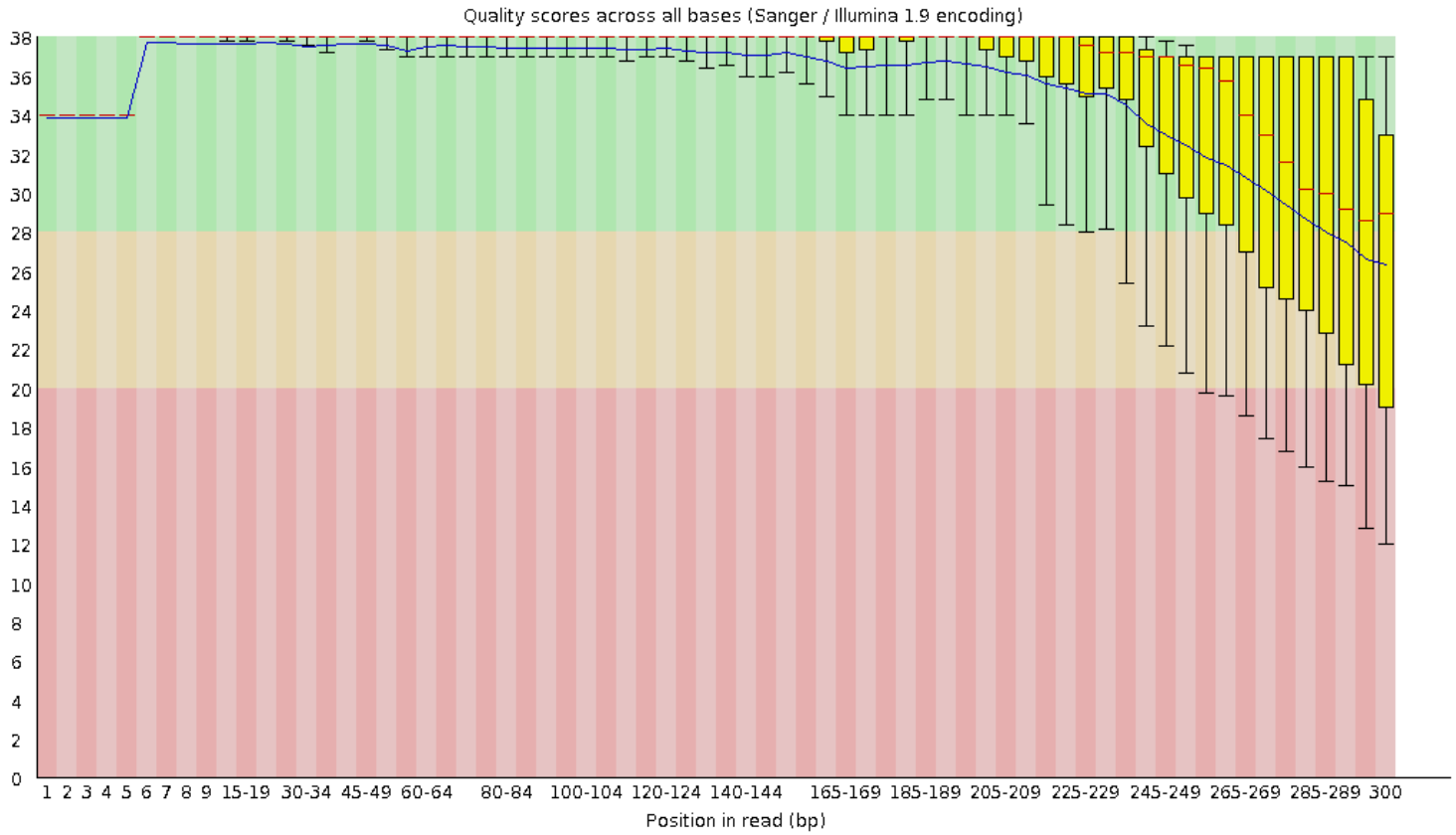


Down-stream analysis

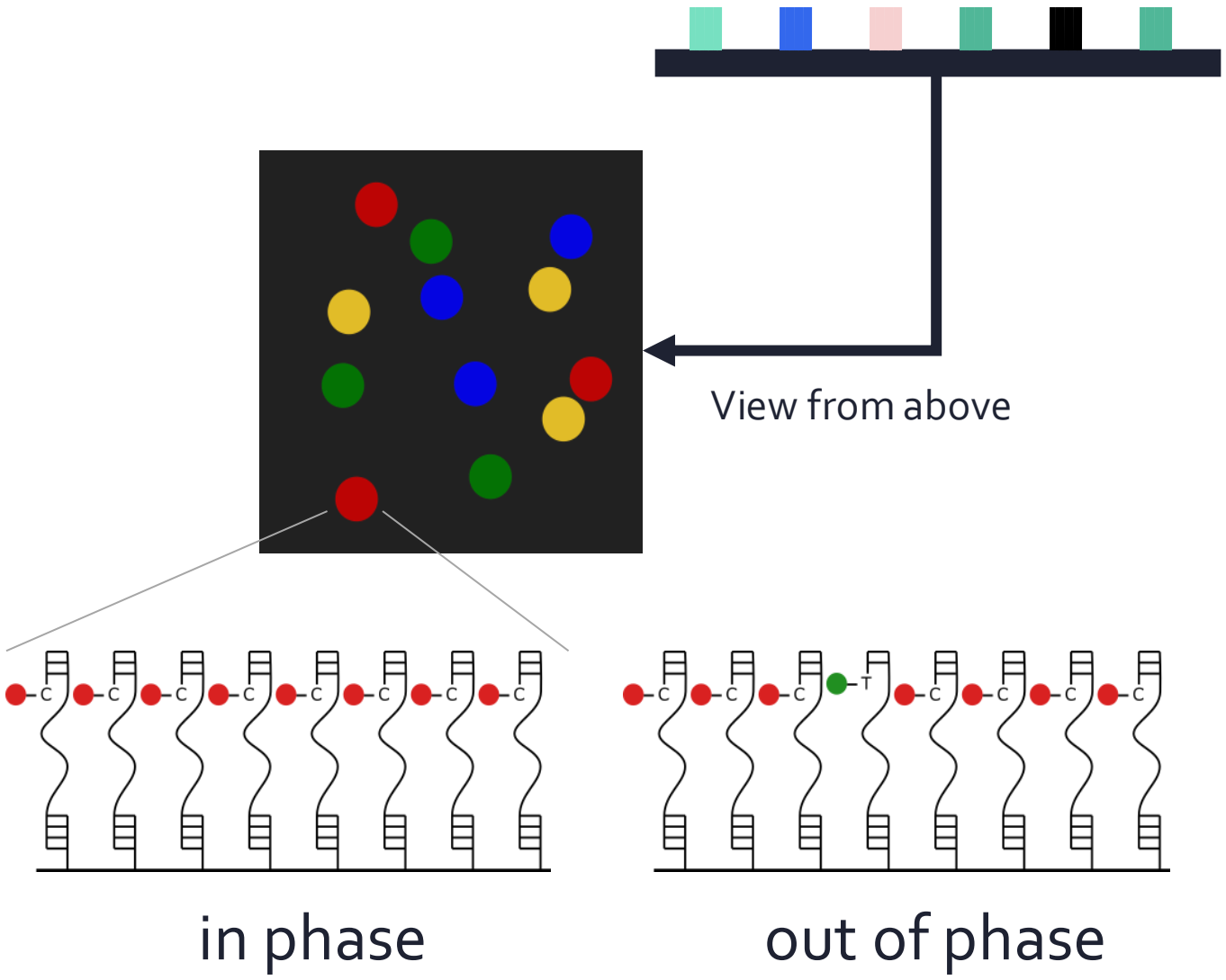
Visualisation

Illumina - limitations

Sequence quality declines towards the end



Phase Sequencing



Illumina - limitations

Maximum read length: 300 bp

Read length is limited by out-of-phase signal

How to reconstruct:

- Repeats?
- Isoforms?
- Structural variation?
- Haplotypes?
- Genomes?

Quiz 2

What is a common limitation of Illumina sequencing?

A. Very high error rates
B. Requires radioactive labeling
C. Decline in sequence quality toward read end
D. Cannot be used for RNA sequencing

Third generation sequencing: Long reads

Crux: maximizing signal from a single-molecule base read-out

Single molecule, so no out-of-phase signal

Two frequently used platforms:

- Oxford Nanopore Technology
- PacBio SMRT sequencing

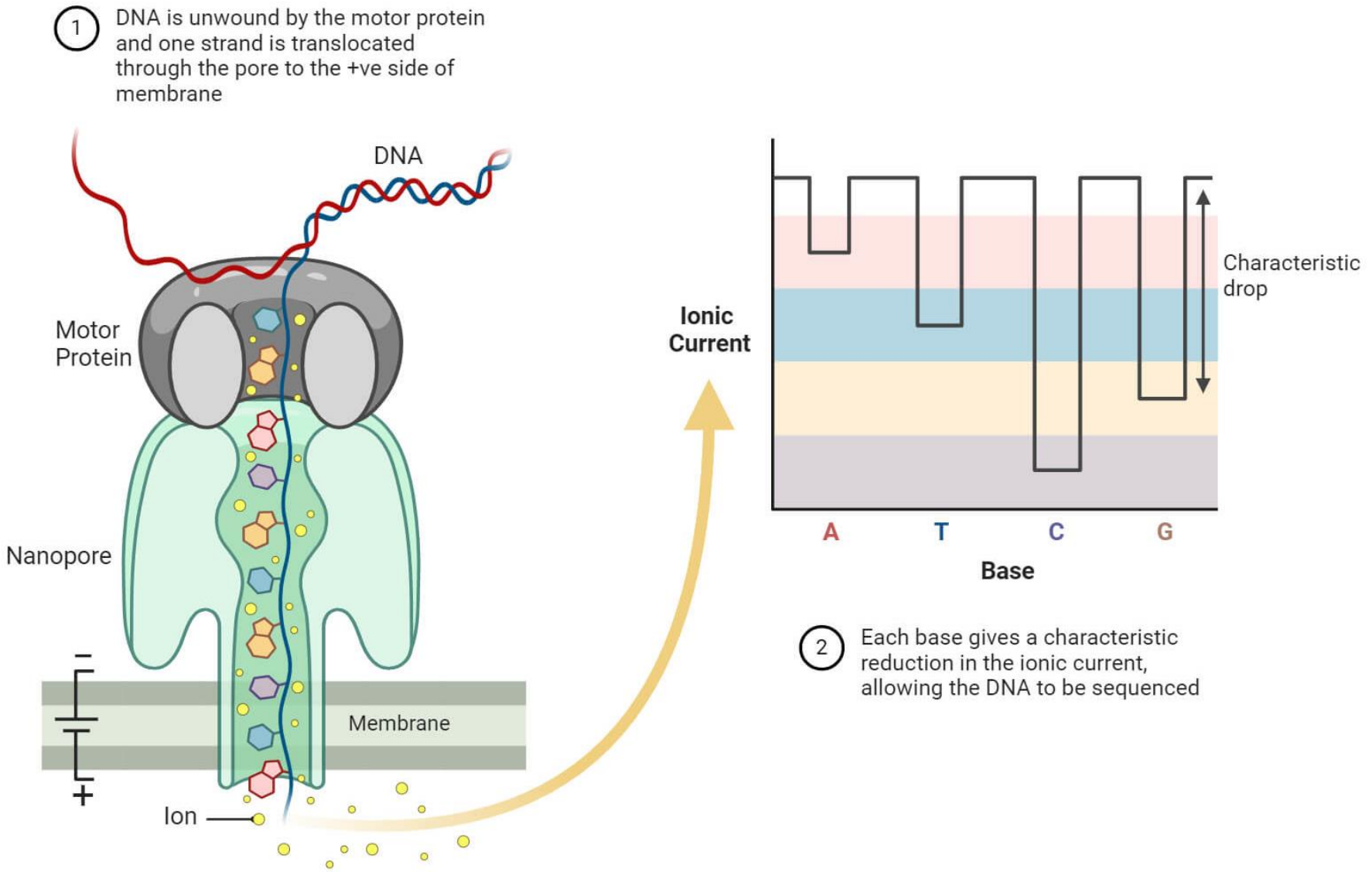
Oxford Nanopore technology

Based on changes in electrical current

Well-known for its scalability and portability

~95-97% accuracy

Oxford Nanopore technology principle



Oxford Nanopore technology sequencers

MinION

GridION

Flongle

PromethION

Quiz 3

Which sequencing method uses changes in electrical current to identify bases?

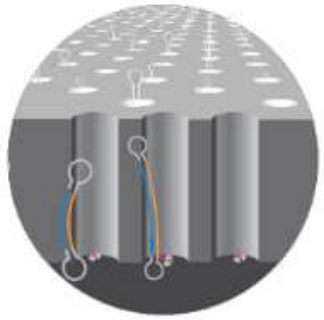
A. Illumina
B. Sanger
C. Oxford Nanopore
D. PacBio SMRT

PacBio sequencing

SMRT Sequencing

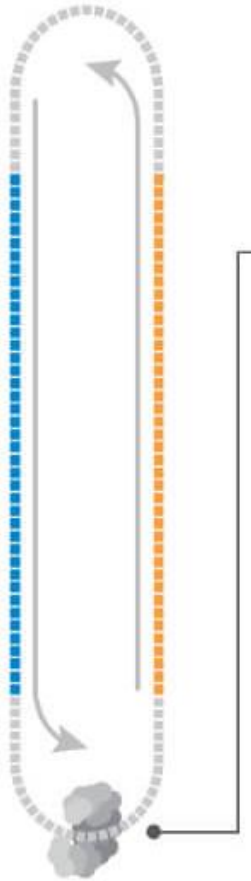
- Technology: Single Molecule, Real-Time (SMRT) sequencing.
- Process: Uses zero-mode waveguides (ZMWs) to observe single DNA molecules in real-time.
- Accuracy: High accuracy due to real-time detection of nucleotide incorporation.
- De Novo Genome Assembly: Ideal for assembling complex genomes, including those with repetitive regions
- Epigenetic Studies: DNAme

PacBio sequencing: SMRT

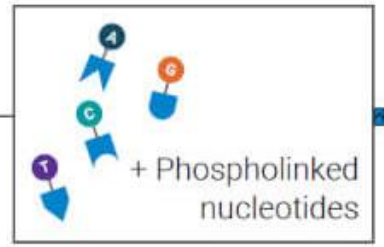


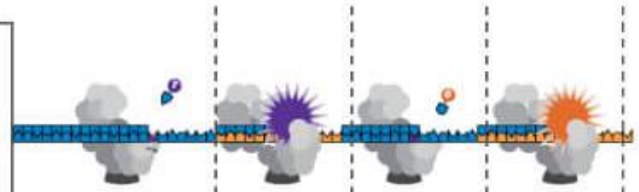
SMRT® Cells contain millions of zero-mode waveguides (ZMWs)

SMRTbell templates enable repeated sequencing of circular template with real-time detection of base incorporation



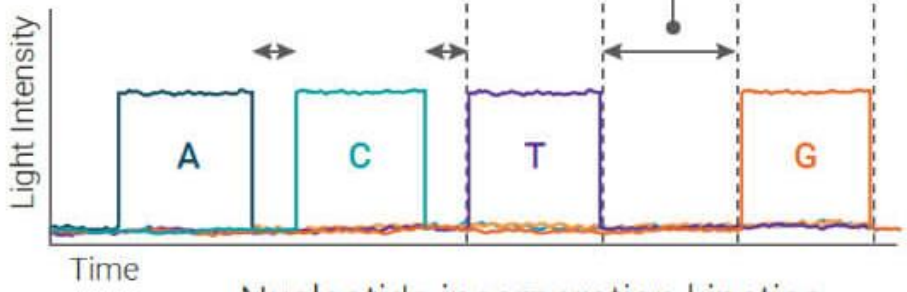
A single molecule of DNA is immobilized in each ZMW





As anchored polymerases incorporate labeled bases, light is emitted

Directly detect DNA modifications during sequencing



Nucleotide incorporation kinetics are measured in real time

PacBio sequencing

Circular Consensus Sequencing (CCS)

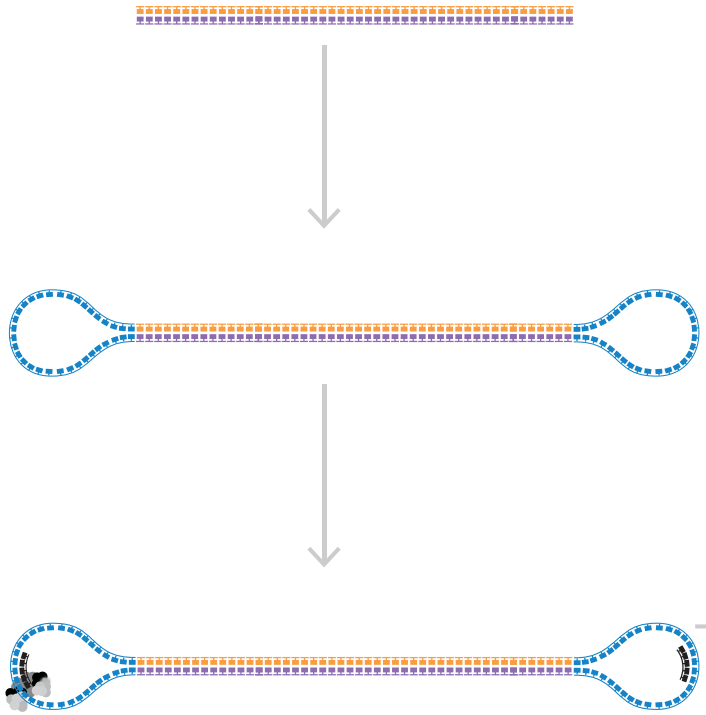
- Technology: Uses SMRTbell libraries for circularized DNA.
- Process: DNA is sequenced in repeated passes, generating long reads.
- Output: HiFi reads with 99.9% accuracy, ideal for detailed genomic studies.
- High-Accuracy Genome Assembly: Produces highly accurate long reads (HiFi reads) for assembling complex genomes
- Rare Disease Research: Helps in identifying genetic variants associated with rare diseases

PacBio sequencing: CCS

Start with high-quality double stranded DNA

Prepare SMRTbell libraries

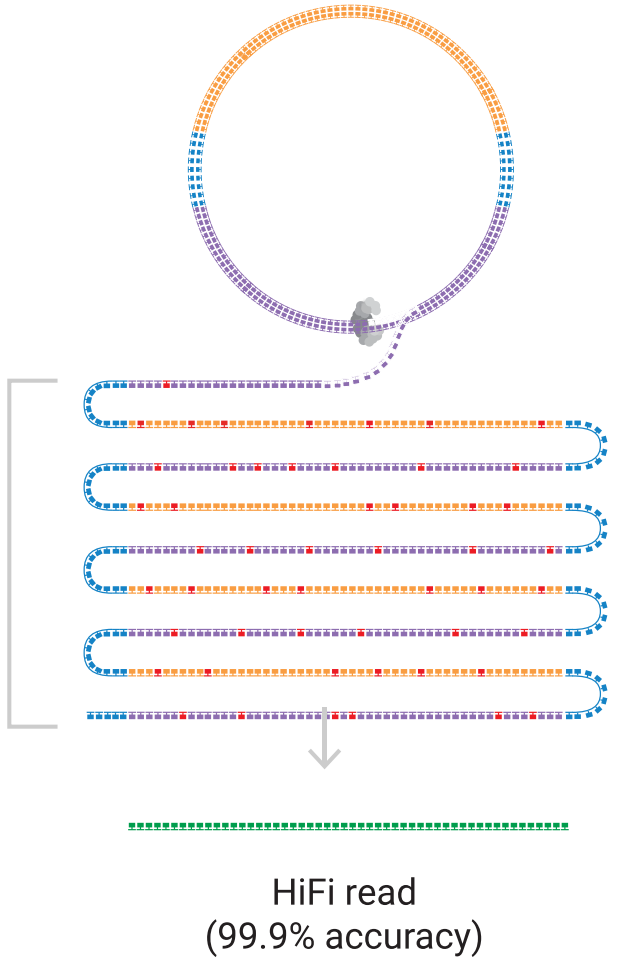
Anneal primers and bind DNA polymerase



Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus and methylation status are called from subreads



HiFi read
(99.9% accuracy)

Quiz 4

What feature makes PacBio's HiFi reads highly accurate?

A. Realtime single molecule synthesis
B. Use of nanopores
C. Repeated sequencing passes through circular DNA
D. Short-read fragment stitching

How to choose your sequencing method?

Read length

Accuracy

Availability

Costs

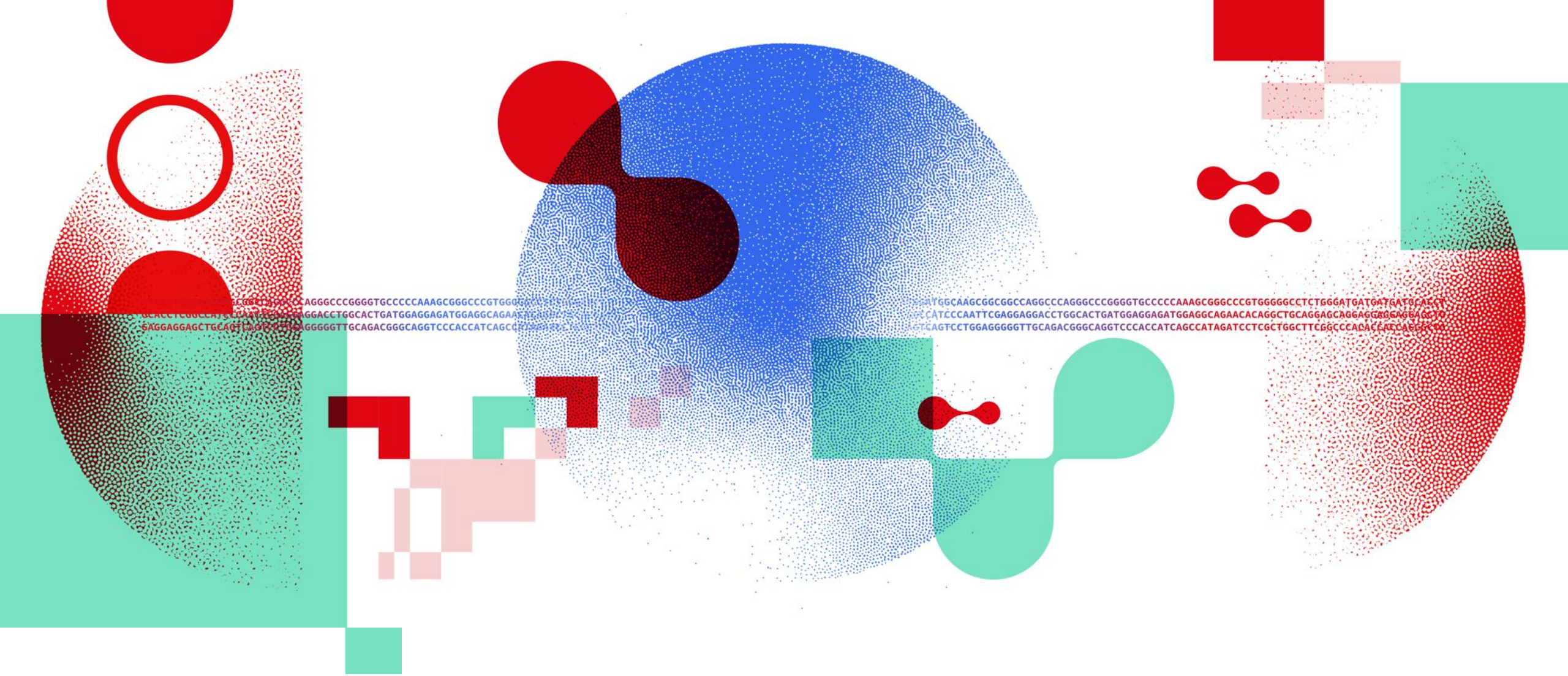
Throughput

Summary

Sanger sequencing: A legacy method used for mutation identification and validation, especially in clinical and forensic applications.

Second generation (Next-Gen) sequencing: Focused on **Illumina** (most widely used due to high throughput and cost-efficiency).

Third generation sequencing: Includes **Oxford Nanopore** (portable, moderate accuracy) and **PacBio SMRT** (very high accuracy with HiFi reads).



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss

elixir SWITZERLAND

SIB