Brief Introduction to Next Generation Sequencing (NGS)

What is Next generation sequencing?

- Next generation sequencing (NGS) – parallel sequencing of a very large number of DNA molecules (millions or billions) present in the living system

- NGS is also known deep sequencing, massive parallel sequencing, high-throughput sequencing

- NGS can be used to establish a myriad of different structural and functional aspects of the cell

- NGS can be applied to both DNA and RNA, however RNA is still converted to DNA in library prep

- NGS methods can be categories into 2 large classes regarding application:

1. Structural methods that establish a primary structure of nucleotide sequences: *de-novo* whole genome sequencing (WGS); resequencing (whole genome or targeted) also known as variant calling

2. Functional or quantification methods that use numbers of sequenced fragments as a surrogate for the abundance of regulatory molecules: RNA-seq, small RNA-seq, ChIP-seq, ATAC-seq, Hi-C, metagenomics and many others

What is Next generation sequencing?

By read length NGS can be classified as short reads and long read

1. Short read – up to 300 bp (Illumina, Ion Torrent)
2. Long reads – up to 4 Mb (Oxford Nanopore, Pacific Bio)

By sequencing library configuration

1. Single-end → a DNA fragment is sequenced from one end
2. Paired-end → a DNA fragment is sequenced from both ends

By strand awareness at the library preparation step

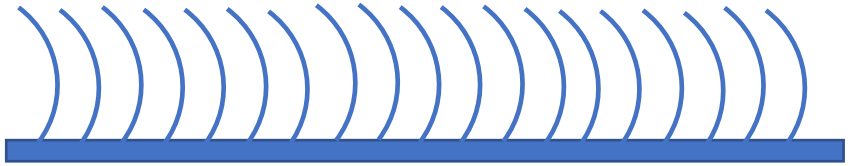
1. Non-stranded → we cannot tell which strand given sequence had originated from
2. Stranded → we can which strand (forward or reverse) the sequence was generated from

What are the steps of short reads sequencing

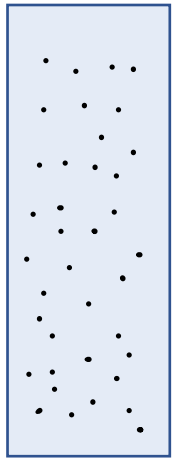
- DNA or RNA extraction, purification, and quantification (high quality is a must)
- Fragmentation of DNA to get it within a range of 200 – 800 bp (endonucleases, sonication, …) [bias introduced]
- Technology-specific adapters are ligated to both ends [bias introduced]
- The adapters may contain unique sample barcodes that allow to sequence multiple samples in one flow-cell lane and separate them later in a process called demultiplexing
- PCR amplification of DNA fragments with adapters attached [bias introduced]
- The libraries are loaded onto the flow-cells (Illumina) or their equivalent and sequenced

Illumina sequencing

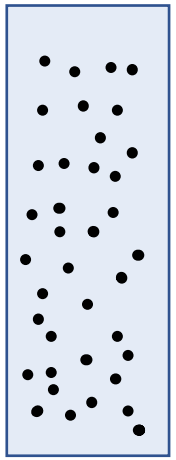
- Illumina uses sequencing by synthesis
- Illumina instrument combines in-situ PCR with microscopy
- The sequencing libraries are hybridized to a flow-cell – a glass slide covered with oligos complementary with adapters



Flow-cell side-view with "lawn" of complimentary oligos



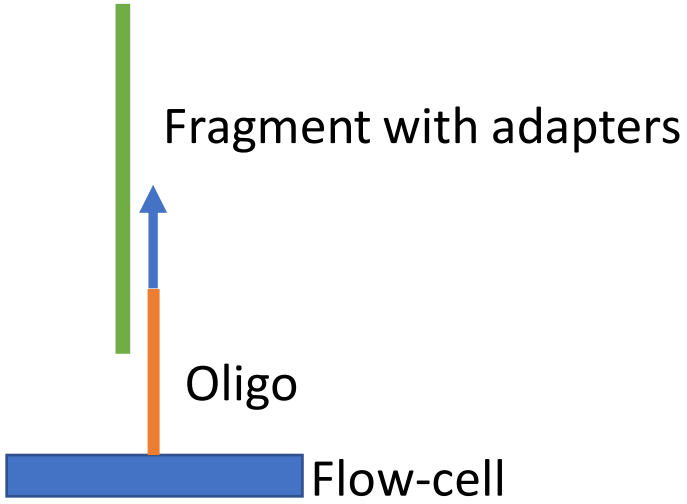
Top-down view of the flow cell with adapter ligated fragments (shown as dots) attached to oligo "lawn"



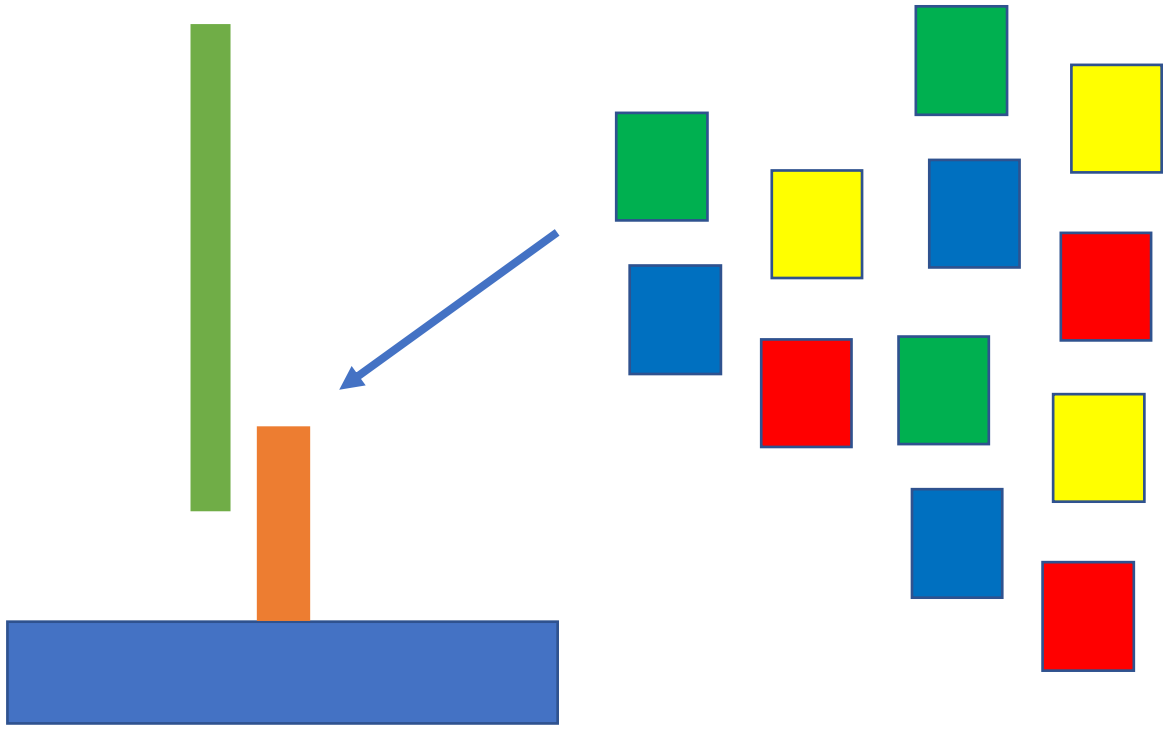
The dots corresponding to hybridized fragments have grown in size due to in-situ amplification. Now each fragment had was multiplied thousands of time and formed a **cluster**. In-situ PCR is needed to amplify the fluorescent signal once the sequencing starts

Illumina sequencing

Sequencing process



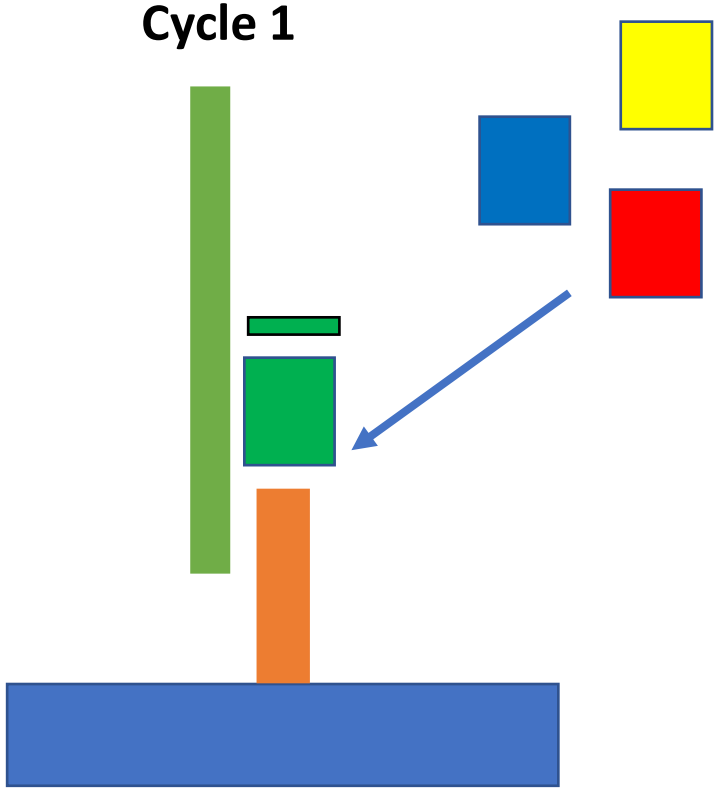
Nucleotides with fluorescent dye terminators



Illumina sequencing

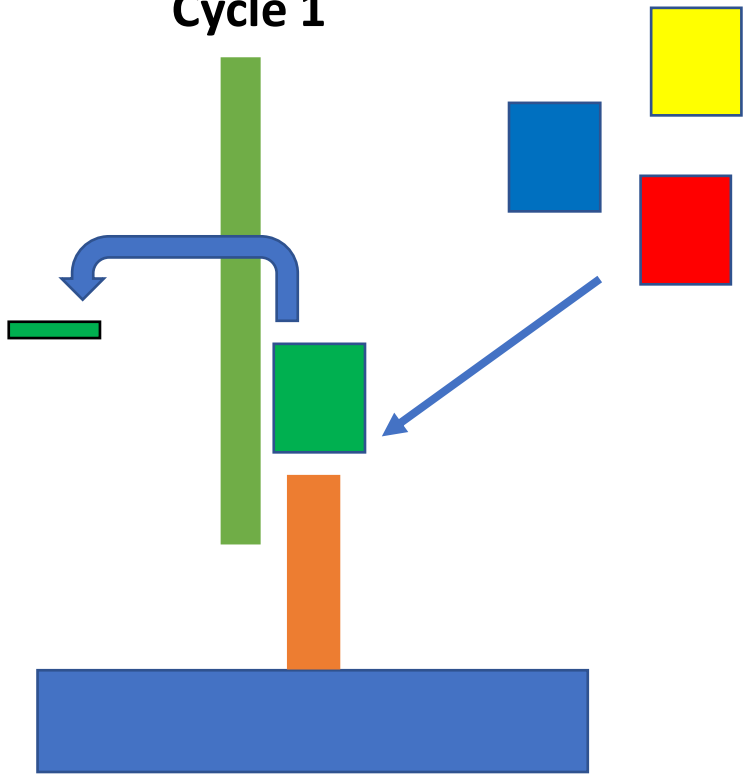
Sequencing process

Cycle 1



Attach nucleotide with fluorescent dye terminator, take a picture of clusters

Cycle 1

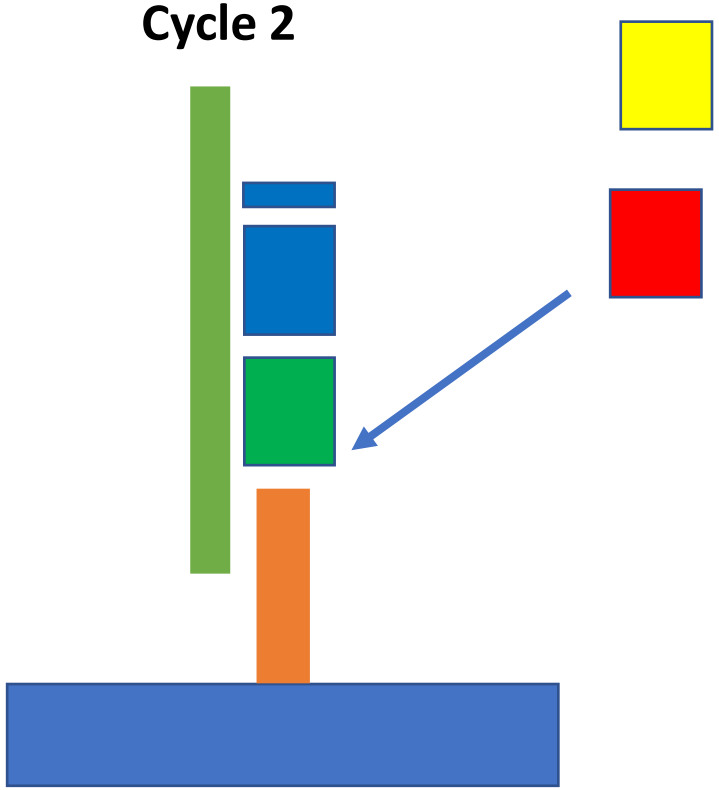


Remove terminator and wash it away

Illumina sequencing

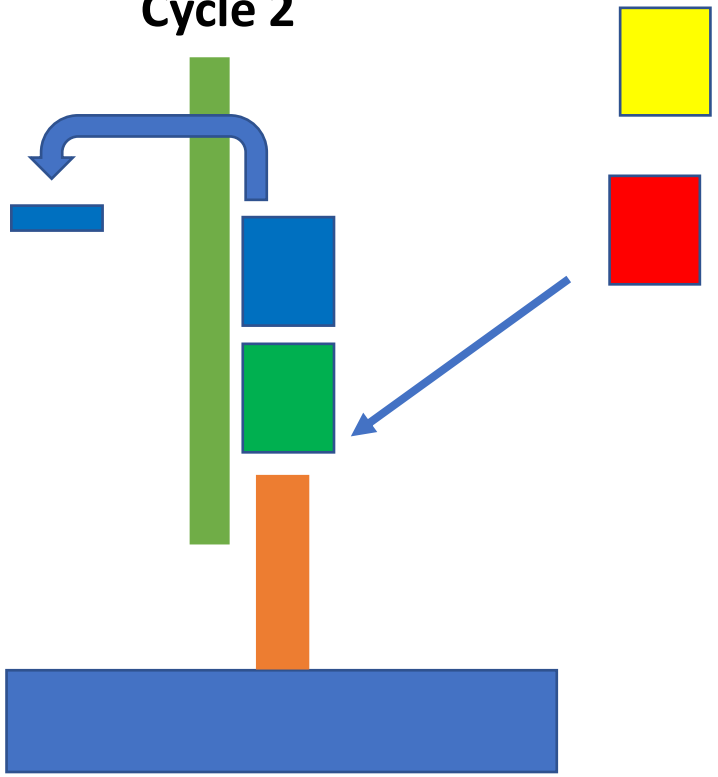
Sequencing process

Cycle 2



Attach next nucleotide, take a picture again

Cycle 2

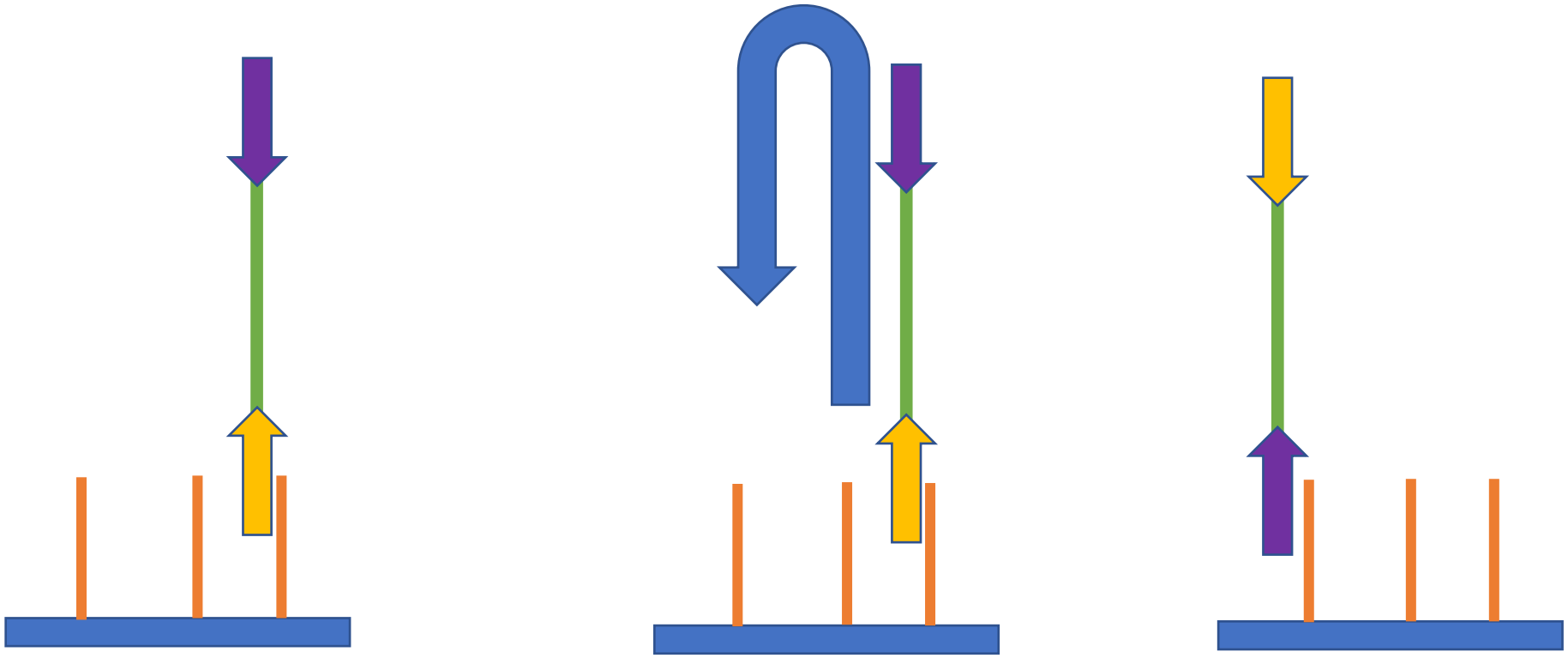


Remove terminator and wash it away, now we can attach next nucleotide

Illumina sequencing

Sequencing process

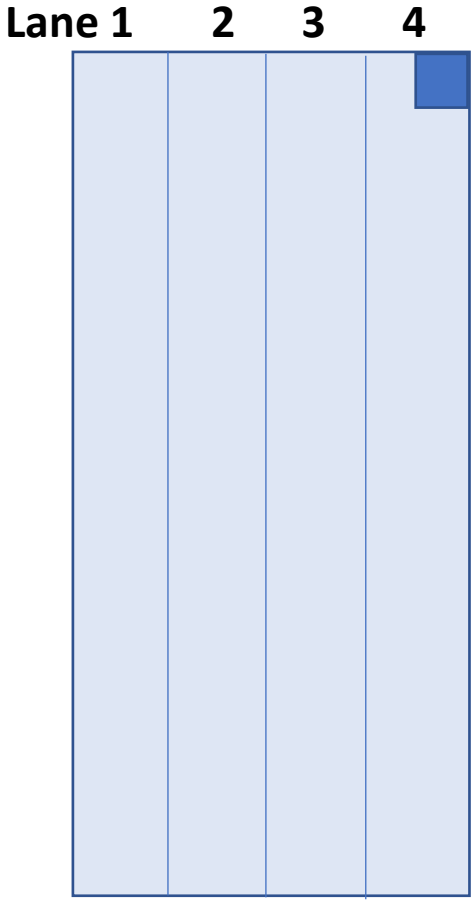
- One cycle or sequencing corresponds to one base
- 100 cycles == 100 bp sequence



Paired-end sequencing → we flip the fragment attach it to a neighboring oligo and sequence from the other end

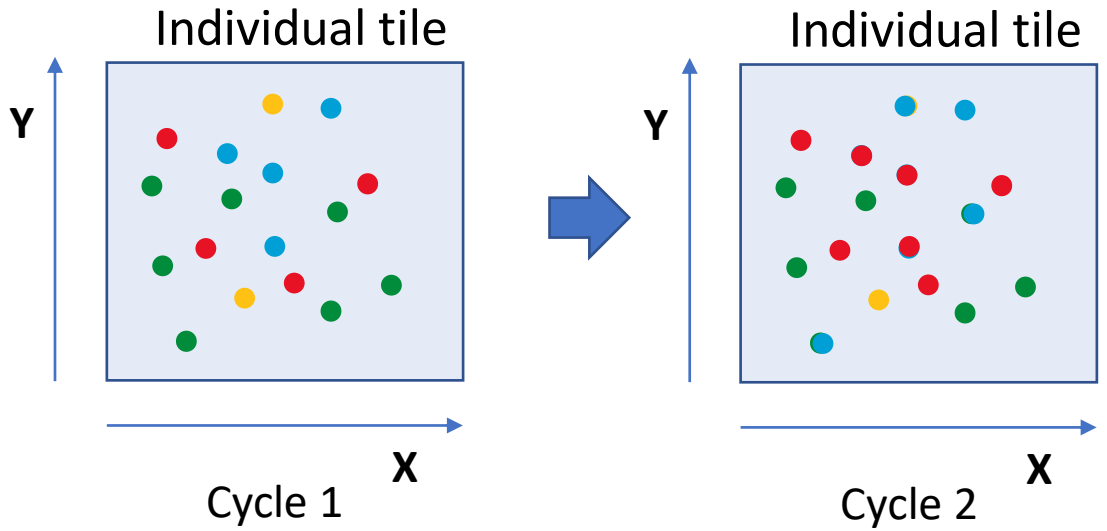
Illumina sequencing

Basecalling – converting image data into primary sequence



Flowcell

Tile: every lane is divided into tiles



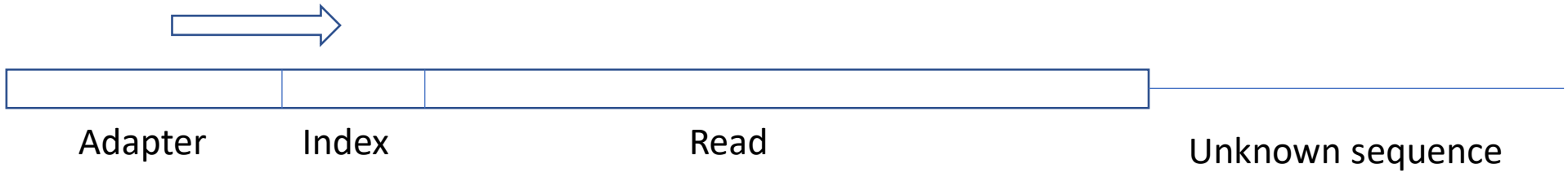
Every cluster has X Y coordinates withing a tile and every tile has a number within a lane

The read corresponds to a physical cluster that can be differentiated from other clusters by Lane # | Tile # | XY coordinate

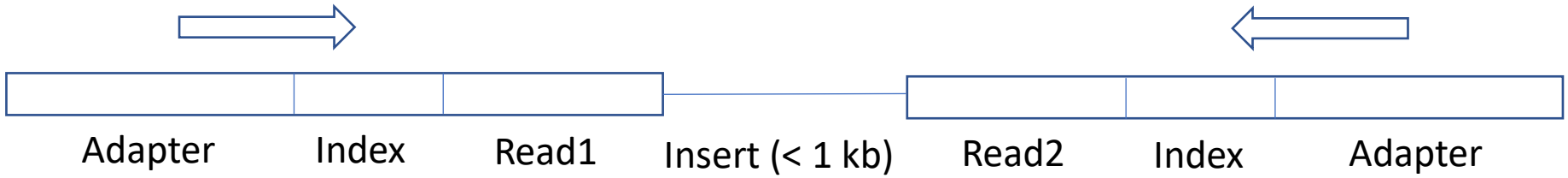
Illumina sequencing

Single-end vs paired-end reads

Single-end sequencing

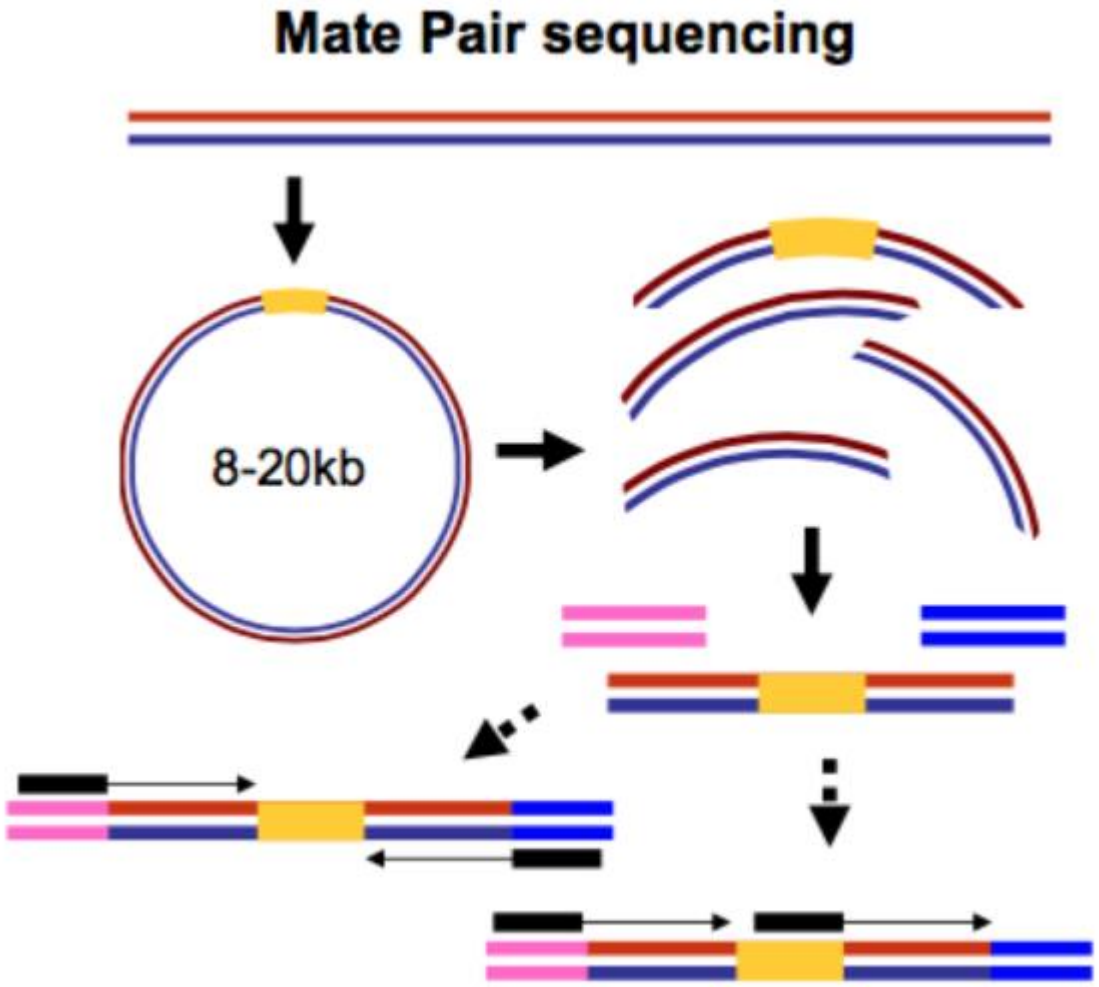


Paired-end sequencing



Illumina sequencing

Mate paired libraries – a special case of paired end sequencing



- In mate-pair libraries the inserts are much longer than in paired-end and can reach up to 20 kb
- When mapping mate-paired libraries the software must expect 1 – 20 kb distance between read pairs
- Mate-paired libraries are useful for de-novo assembly, genome finishing, and detecting large structural variants

Galaxy Project: https://tinyurl.com/bdh4e7a6

The output of Illumina sequencing

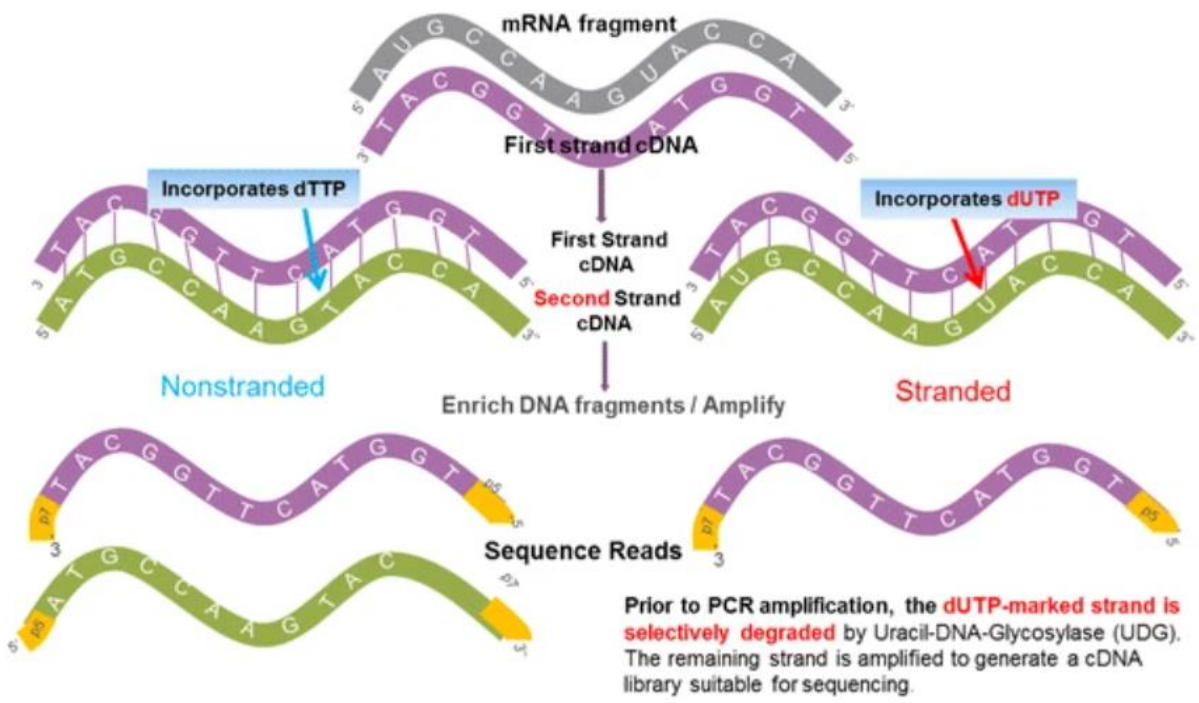
- The data output by Illumina sequencer comes as millions of DNA sequences in **fastq** format
- Typically, this files will be compressed and have *.fastq.gz* extensions
- Sometimes genome centers can deliver the reads as unaligned bam files *ubam*
- Paired-end sequencing will result in 2 fastq files – one for read 1 and read 2
- The order of reads in Read 1 and Read 2 must be the same
- Occasionally you might receive interleaved fastq file where both reads are in the same file and read 1 is directly followed by read2

Advantages of paired-end sequencing

- Twice the reads for the same sequencing run and library prep
- Better mapping results in repeated regions
- Better detection of genomic rearrangements
- Better resolution of multi-mappers
- When facing a choice whether to go with paired-end or single-end libraries always go with paired-end
- The exception are small RNA libraries and bisulfite sequencing

Advantages of strand-specific libraries

- Stranded libraries can differentiate between sense and anti-sense transcripts in RNA sequencing, improve detection of gene fusions
- Better estimates of gene expression levels



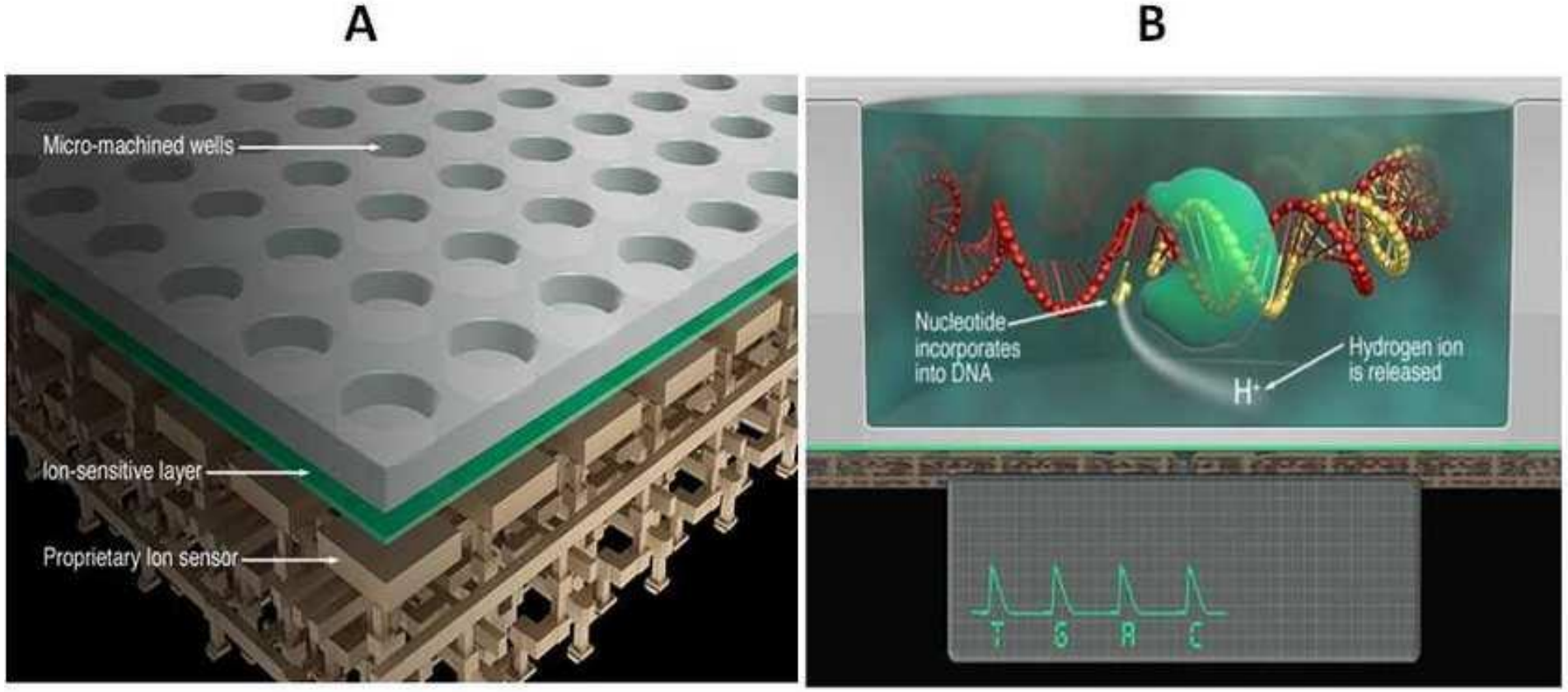
dUTP method for strand labeling Zhao et al, 2015

In the case of variant calling strandedness does does not appear to make a difference

Ion torrent sequencing

- Unlike Illumina Ion Torrent does not produce optical signal
- It relies on the expulsion of H+ ion when DNA polymerase adds dNTP to the growing DNA sequence
- The libraries are fragmented to about 200 bp and placed onto a bead
- The fragments are amplified on beads by emulsion PCR
- Beads carrying DNA fragments are then placed into the wells on the metal oxide semiconductor chip
- The slide is flooded with a single type of dNTP and pH is measured at every well
- Decrease in pH at a well signals successful incorporation of the nucleotide

Ion torrent sequencing



7: Design of Ion Torrent semiconductor sequencing chip technology. (A) Chip cross section semiconductor design viewing layer by layer. The top layer comprises the individual DNA polymerization reaction wells containing a DNA polymerase and the bottom two layers contain the FETs ion sensor. Each microwell has a matching FET detector that recognises a change in pH. (B) A side view of an individual reaction microwell illustrating DNA polymerase incorporation of nucleotides on a sequencing fragment. During this process, the hydrogen ion is released and detected by the FET. Taken from Niedringhaus et al. (2011).

Pacific bio sequencing

- Pacific Biosciences (PacBio) uses single-molecule real-time PCR (SMRT) technology
- PacBio sequences occurs in real-time, it does require a pause between the steps
- The template (SMRTbell) is a single stranded circular DNA molecule created by ligating hairpin adaptors both ends of the target sequence
- SMRTbell is loaded onto a chip called SMRTcell and diffuses into a tiny separated volume called Zero-mode wave guide (ZMW)
- ZMW is smallest available volume for light detection
- Each ZMV contains a single polymerase immobilized at the bottom
- Four fluorescent labeled nucleotides are added to ZMV, the light signal is emitted every time dNTP is incorporated by DNA polymerase

Pacific bio sequencing

- The light pulses emitted by ZMV are recorded a “movie” and interpreted as a sequence, called continuous long read (CLR)
- PacBio RS II instrument produces “movies” 0.5 – 4 h in length
- The template is circular, and once the DNA polymerase passes the adapter, it replicates a complementary strand of the target sequence thus completing a circle called a “pass”
- Multiple “passes” of the same template can be completed within the same run
- In this scenario a CLR can be split into multiple reads by cutting out adapter sequence
- The consensus sequence of multiple subreads produces circular consensus sequence (CCS) that has much higher accuracy then single reads
- Nucleotide modifications, such as DNA methylation can be detected directly in PacBio sequencing

Pacific bio sequencing

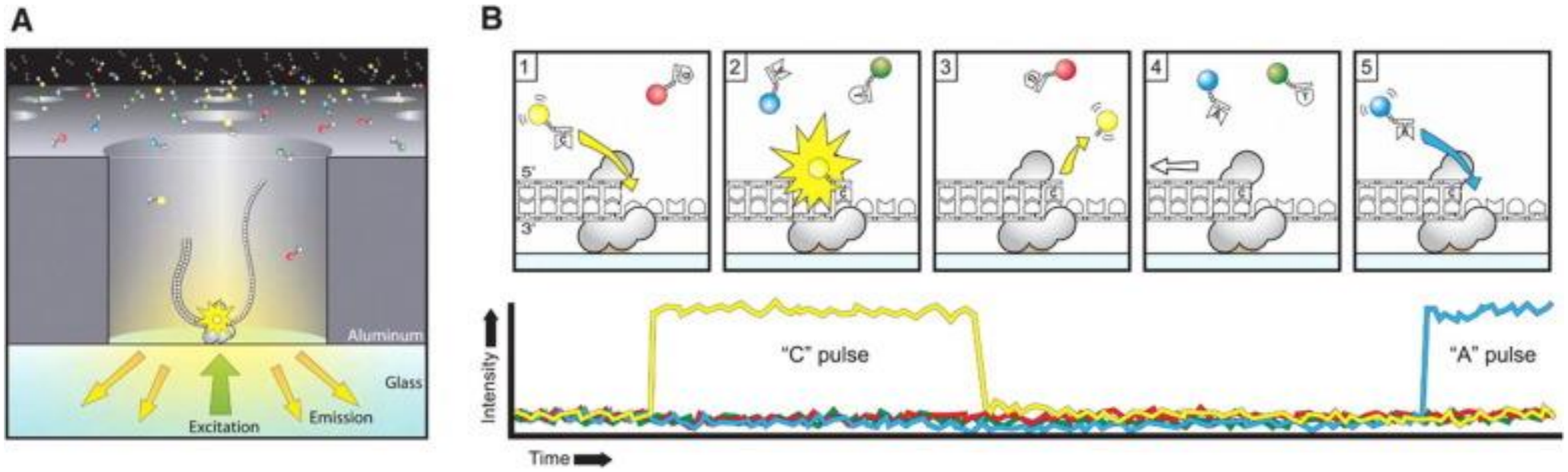
Hairpin adaptors (green) are ligated to the end of a double-stranded DNA molecule (yellow and purple), forming a closed circle. The polymerase (gray) is anchored to the bottom of a ZMW and incorporates bases into the read strand (orange).

https://www.sciencedirect.com/science/article/pii/S1672022915001345#f0010

Pacific bio sequencing

Each SMRT cell contains 150,000 ZMWs. Approximately 35,000–75,000 of these wells produce a read in a run lasting 0.5–4 h, resulting in 0.5–1 Gb of sequence.

Pacific bio sequencing

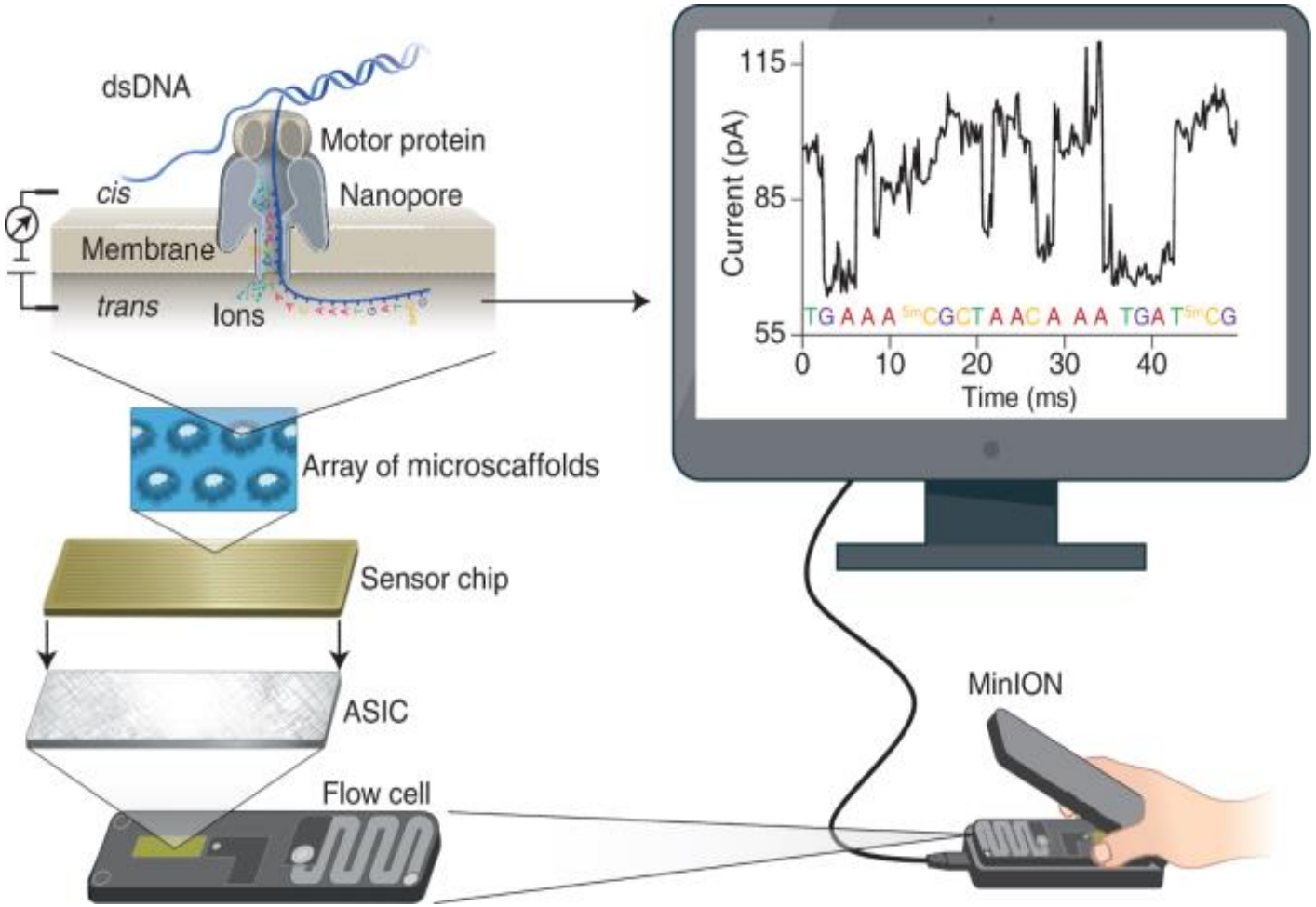


A. A SMRTbell (gray) diffuses into a ZMW, and the adaptor binds to a polymerase immobilized at the bottom. **B.** Each of the four nucleotides is labeled with a different fluorescent dye (indicated in red, yellow, green, and blue, respectively for G, C, T, and A) so that they have distinct emission spectrums. As a nucleotide is held in the detection volume by the polymerase, a light pulse is produced that identifies the base.

Oxford nanopore sequencing

- Technology relies on nanoscale protein pore – “nanopore”
- The nanopore is embedded in electrically resistant membrane
- The difference of potentials exists on both sides of a membrane
- Single stranded negatively charged DNA or RNA is driven from the negatively charged side of the membrane to the positive side through the nanopore
- Translocation speed is controlled by a motor protein that ratches nucleic acid through the nanopore in a stepwise fashion
- The motor protein is helicase that unwinds double-stranded and turns it single stranded while it is driven through the nanopore
- The ionic current through the nanopore changes during a nucleotide pass, and the configuration of this change depend on the nucleotide and its chemical modifications

Oxford nanopore sequencing



A MinION flow cell a total of 2,048 nanopores used to sequence DNA or RNA. The wells are inserted into an electrically resistant membrane supported by an array of microscaffolds connected to a sensor chip. Each channel associates with a separate electrode in the sensor chip and is controlled and measured individually by the application-specific integration circuit (ASIC). Ionic current passes through the nanopore because a constant voltage is applied across the membrane. Under the control of a motor protein, a double-stranded DNA (dsDNA) molecule (or an RNA–DNA hybrid duplex) is first unwound, then single-stranded DNA or RNA with negative charge is ratcheted through the nanopore, driven by the voltage. As nucleotides pass through the nanopore, a characteristic current change is measured and is used to determine the corresponding nucleotide type at ~450 bases per.

https://www.nature.com/articles/s41587-021-01108-x

Oxford nanopore sequencing

- The range of sequence length for Oxford Nanopore is between 500 – 2.3 Mb with typical length between 10 – 100 kb
- One MiniION flow cell generated about 25 Gb of sequencing data if good quality DNA is sequenced
- Oxford Nanopore is direct sequencing that does not require fragmentation, ligation, amplification or fluorescence
- Oxford Nanopore MiniION is portable and can be used in a field
- Nanopore sequencing is still quite error prone with sequencing accuracy at about 5%
- Nanopore can detect chemical modifications of the nucleotides (DNA methylation, hydroxylation, etc.)
- The accuracy of MiniIon sequencing can be improved with rolling circle amplification, where multiple copies of a DNA molecule are concatenated in a single strand. The contiguous strand is sequenced and used to generate a consensus sequence

Comparison of sequencing platforms

Illumina is leader in terms of output with about 1500 Gb of reads produced per sequencing run, Ion Torrent will output about 50 Gb

For PacBio – 500 Gb in CSS mode, however 10 – 16 Gb after forming consensus sequences and retaining highly accurate amplicons

For Oxford Nanopore MinIon – 50 Gb in ideal conditions, larger devices like PromethIon can produce up to 150 Gb

In terms of error rates:

→ Illumina: various instruments have different error rates, typically ~0.5%, for example HiSeqXTen – 0.087% and Miniseq 0.613%

Link to the paper about Illumina error rates:
https://academic.oup.com/nargab/article/3/1/lqab019/6193612

→ Ion torrent: error rate in range of 0.48% - 1.12%; sensitive to homopolymers that cause deletions

Link to the paper about Ion Torrent error rates: https://www.nature.com/articles/s41598-017-08139-y

Comparison of sequencing platforms

→ PacBio error rates: about 13% for raw datasets, with high (8%) level of insertions

→ Nanopore errors: about 12%, evenly distributed (4% each) between substitutions, deletions and insertions

Consensus correction improves error rates for both technologies decreasing the error rate to about 1%

Link to paper: https://academic.oup.com/nargab/article/2/2/lqaa037/5843804

Read length for different platforms:

→ Illumina – up to 2 by 300 bp for paired-end reads

→ Ion Torrent – 200 – 400 bp

→ PacBio – 10 – 25 kb

→ Oxford Nanopore – 10 – 100 kb for long read and 100 – 300 for ultra long read