Sequencing Technologies - Part II

Sonja J. Prohaska

Computational EvoDevo, University of Leipzig
Santa Fe Institute, Santa Fe, NM, U.S.

WS 2026/27

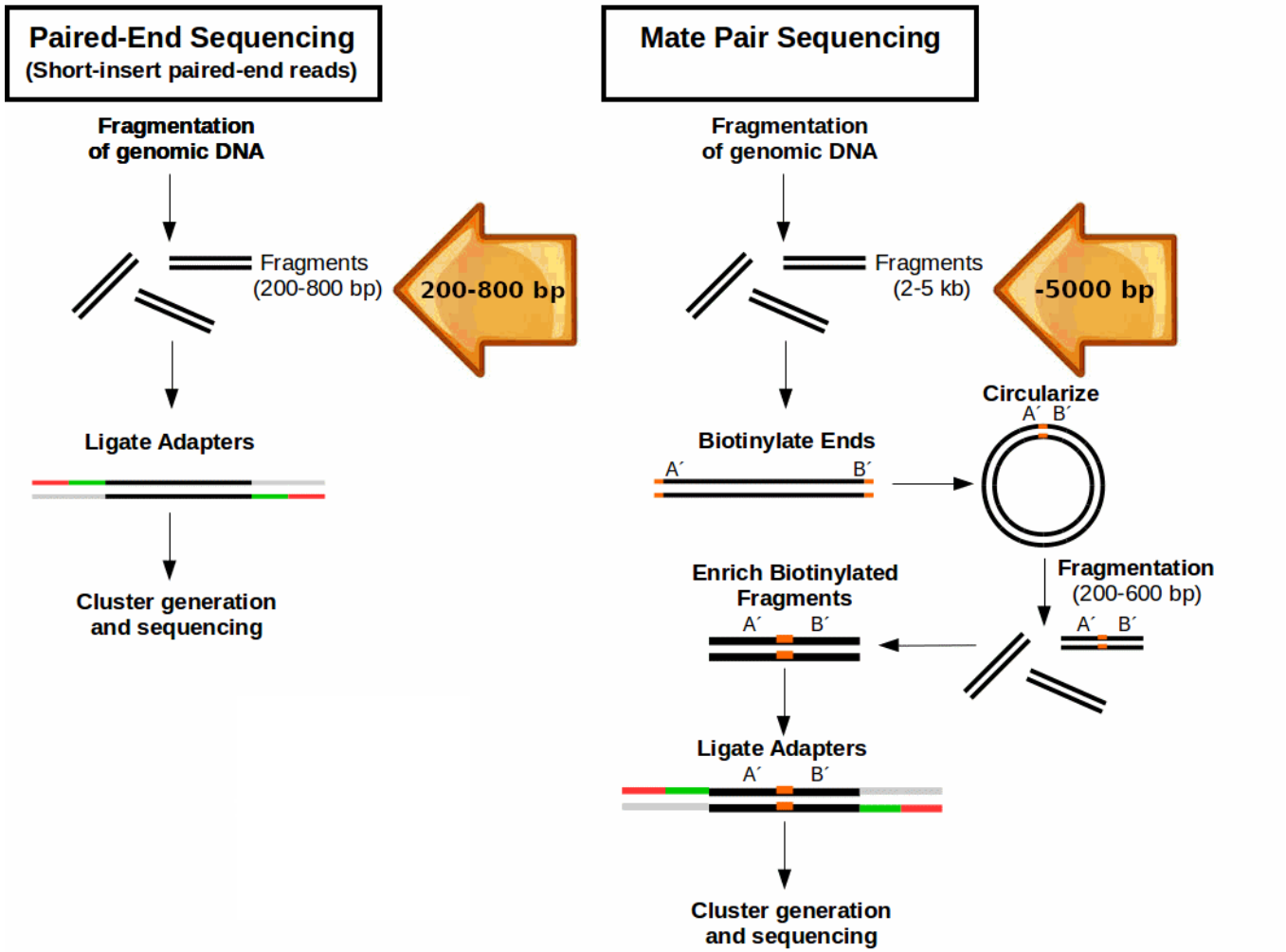
General High-Throughput Sequencing Workflow

- **sample preparation**: from organism to DNA
  - material collection, dissection
  - DNA extraction (careful,if nanopore)
- **library preparation**: from DNA to sequencing machine-specific input
  - DNA fragmentation and size selection (Illumina)
  - adapter ligation
  - (barcoding)
  - PCR amplification (Illumina)
- **sequencing**: from molecular input to digital output
  - signal capture, imaging (Illumina)
  - raw data generation
- **base calling**: from raw data to sequence
  - base calling
  - quality control
- **data analysis**

Library Preparation

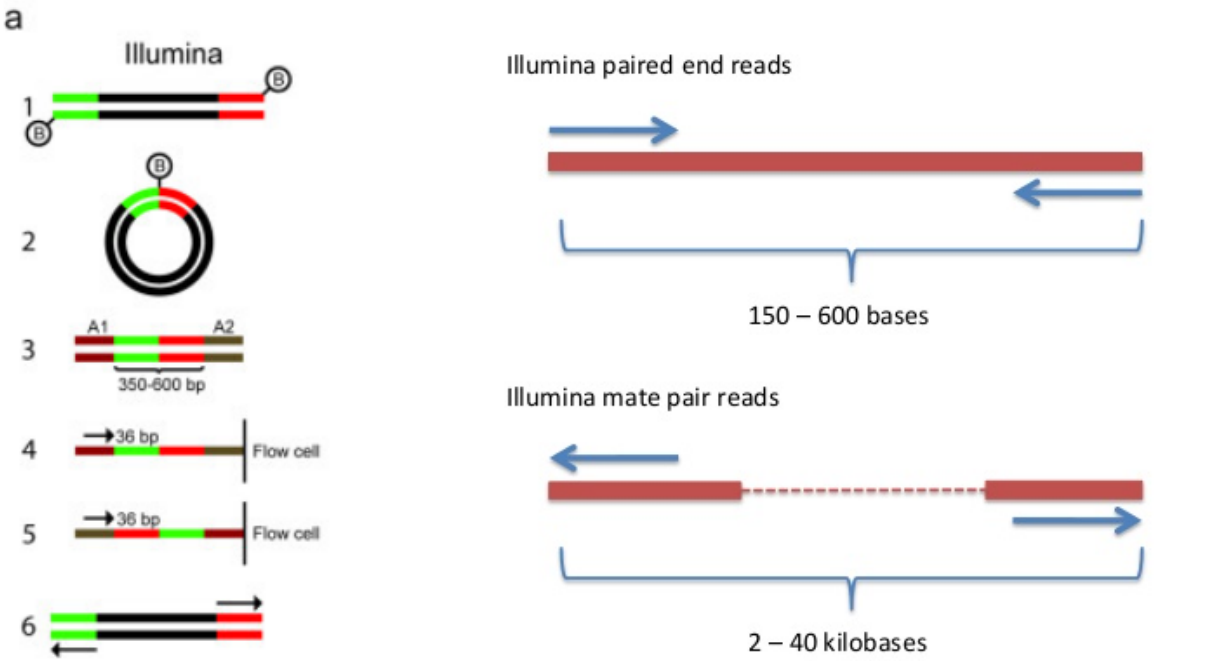
- Until now we have tried to sequence an DNA fragment from start to end.
- If it was longer than what our sequencing technology supported, we could only cover the first n nucleotides.
  - e.g. with Sanger: $n = 1 - 1000nt$
  - e.g. with Illumina: $n = 50 - 300nt$
- An alternative approach is to sequence the DNA fragment from **both ends**
- This way we get the first n and the last n nucleotides of the DNA fragment, which is longer than $2 \times n$

Library Preparation



https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for

Library Preparation



Principles for construction of mate-pair sequencing libraries. (a) Preparation of Illumina mate-pair libraries. Fragments are end-repaired using biotinylated nucleotides (1). After circularization, the two fragment ends (green and red) become located adjacent to each other (2). The circularized DNA is fragmented, and biotinylated fragments are purified by affinity capture. Sequencing adapters (A1 and A2) are ligated to the ends of the captured fragments (3), and the fragments are hybridized to a flow cell, in which they are bridge amplified. The first sequence read is obtained with adapter A2 bound to the flow cell (4). The complementary strand is synthesized and linearized with adapter A1 bound to the flow cell, and the second sequence read is obtained (5). The two sequence reads (arrows) will be directed outwards from the original fragment (6).

The approximate distance between read pairs can be used during read mapping to bridge repetitive regions.

Library Preparation

Possible applications for paired-end sequencing are listed there:
`https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for`

You can find a video explaining the concept there:
`https://www.youtube.com/watch?v=WneZp3fSJIk`

Sequencing Platforms in Comparison – Technical Data

	Sanger ABI 3730xl	**Illumina** MiSeq i100	**Illumina** NovaSeq X	**PacBio** Revio (HiFi)	**ONT** PromethION
read length typical (max.)	700–900 bp (∼1 kb)	2×150 bp (2×500 bp)	2×150 bp (2×300 bp)	?	?
accuracy per read	99.9–99.99 %	>99.9 %	>99.9 %	?	?
reads per flow cell[a]	96	10^8	$\sim 3 \times 10^{10}$	?	?
output per flow cell[a]	∼0.09 Mb	30 Gb	8–10.5 Tb	?	?
run time	20 min–3 h	4–24 h	17–48 h	?	?
cost per Gb[b]	$5–10 M	∼$60	$2–3	?	?
instrument price (list)	$290–400k	$49–109k	$985k	?	?

[a] Revio: per SMRT Cell (up to 4 per run); PromethION: per flow cell (up to 48 per run). [b] sequencing reagents or core-facility prices only, without library prep, labour and instrument. [c] estimated as output / mean read length. Maximum values are not all reached in the same run. As of Nov 2026.

The long-read platforms (PacBio, ONT) follow later in this lecture.

Second Generation Sequencing Technologies

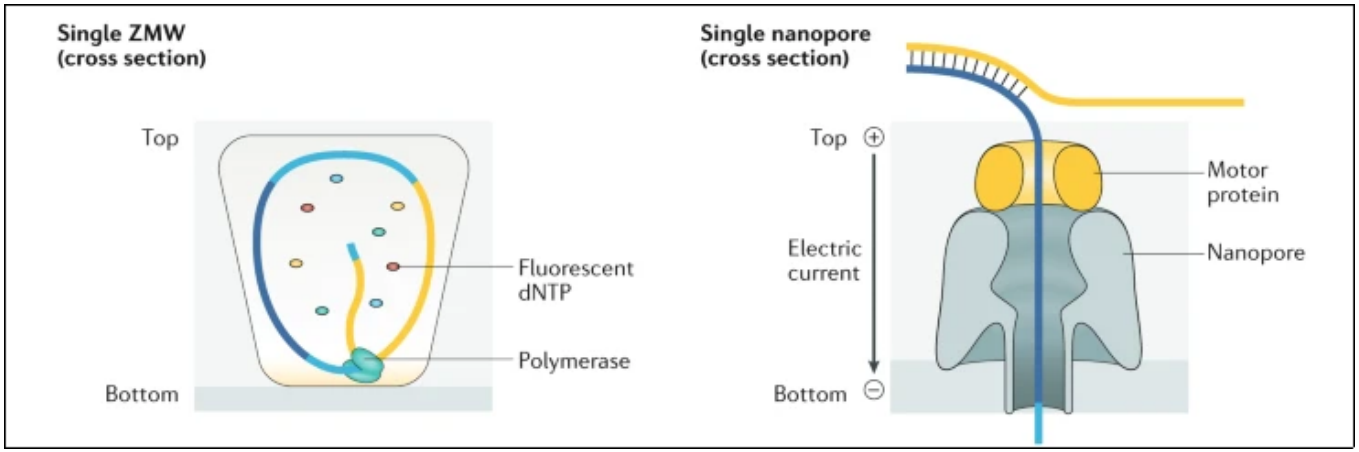
Several technologies using a variety of creative methodologies were on the market:

- automated Sanger sequencing
- Solid
- Helicos
- Roche 454 pyrosequencing

Illumina sequencing is practically the **only remaining second generation sequencing technology**. Conceptually it is not that different from the first generation Sanger sequencing.
It is a **short read** sequencing technology.

The THIRD Generation of High-Throughput Sequencing

In the last decade alternative approaches have been brought to a stable level and are widely used nowadays. The most prominent examples are **PacBio** and **Oxford Nanopore (ONT)** sequencing. They are **single-molecule**[1], **real-time**[2] sequencing methods which make **long read** sequencing possible.



[1] **single-molecule**: no prior DNA amplification required
[2] **real-time**: lets scientists watch DNA sequencing happen as it occurs

Third Generation Sequencing Technologies

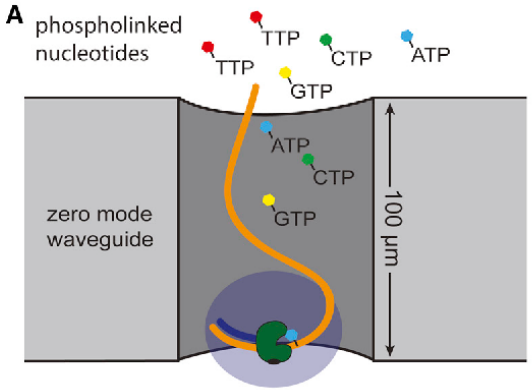
There are nicely animated videos for both sequencing technologies.
Nanopore sequencing:

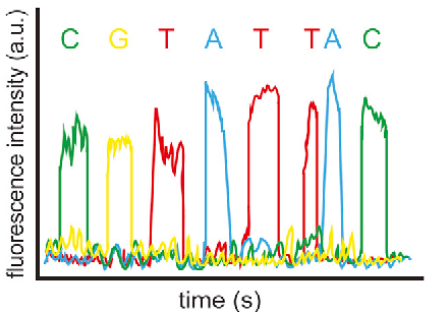
https://www.youtube.com/watch?v=qzusVw4Dp8w
https://www.youtube.com/watch?v=sv9fFeSd3kE

PacBio sequencing:
https://www.youtube.com/watch?v=lT3NqhKD840
https://www.youtube.com/watch?v=_lD8JyAbwEo
Advantages of Long-read sequencing:
https://www.youtube.com/watch?v=DDbeyf1FEEU

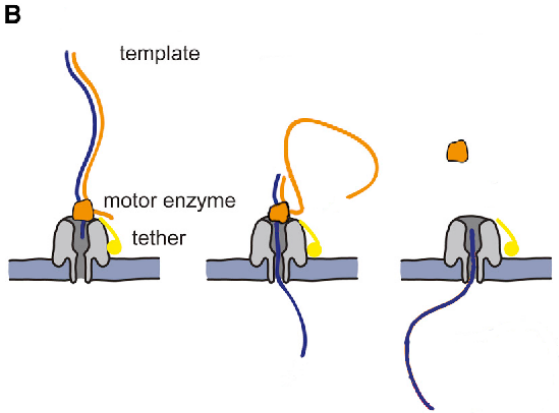
Third Generation Sequencing Technologies

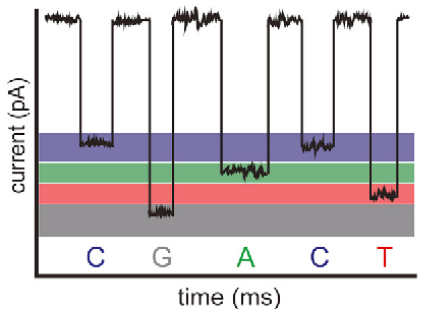
PacBio
single-molecule
real-time
no
sequencing by synthesis
fluorescence measurement





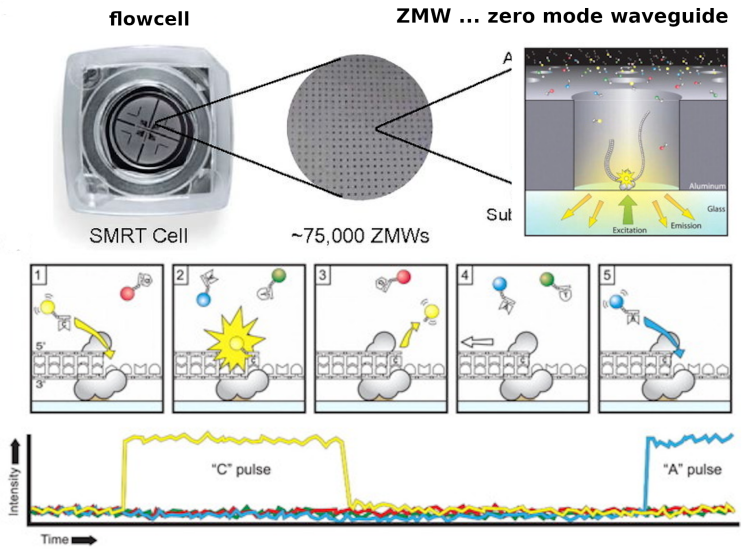
Nanopore
single-molecule
real-time
direct DNA/RNA sequencing
no
current measurement



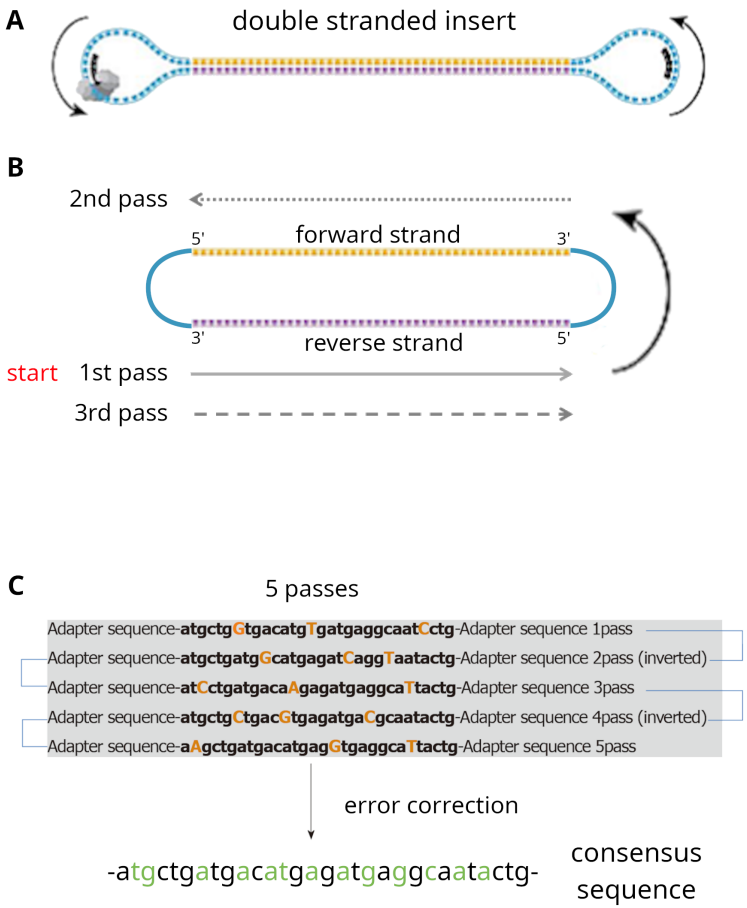


Pacific Biosciences (PacBio)

... is a **sequencing by synthesis** method.







https://www.youtube.com/watch?v=v8p4ph2MAvI

Pacific Biosciences (PacBio)

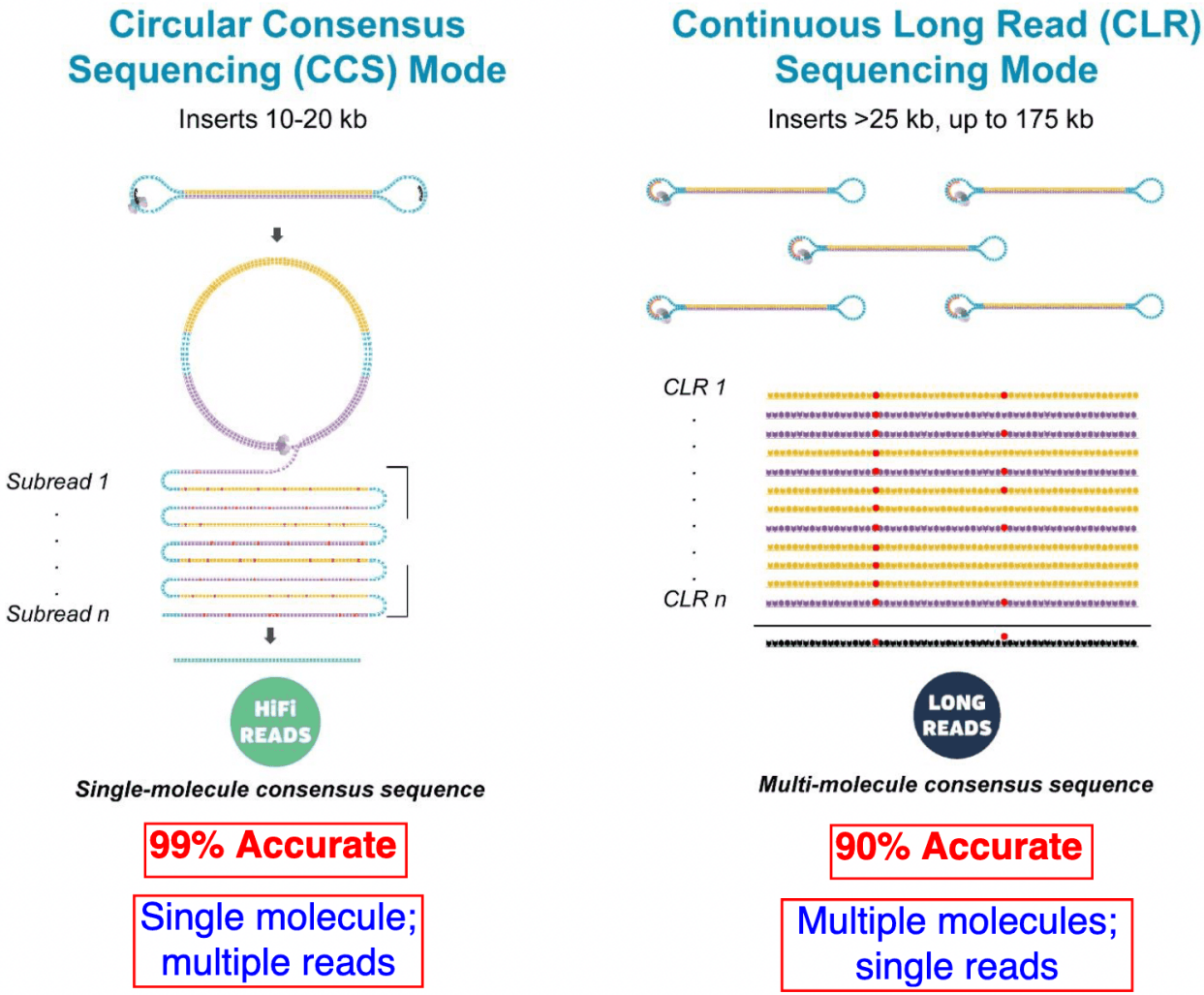
... is a **sequencing by synthesis** method.

Workflow in four steps

1. large DNA framents (with blunt ends)
2. SMRTbell adapter (hairpin shaped) ligation
   creates a **circular**, **single stranded** DNA
3. in one pass, one strand is sequenced
   in two passes, both strands are sequenced
4. alignment of multiple passes and error correction
   accuracy plateaus after approx. 15-20 passes
   polymerase processivity can limit the number of passes

`https://www.youtube.com/watch?v=v8p4ph2MAvI`

PacBio: HiFi Reads vs. Long Reads

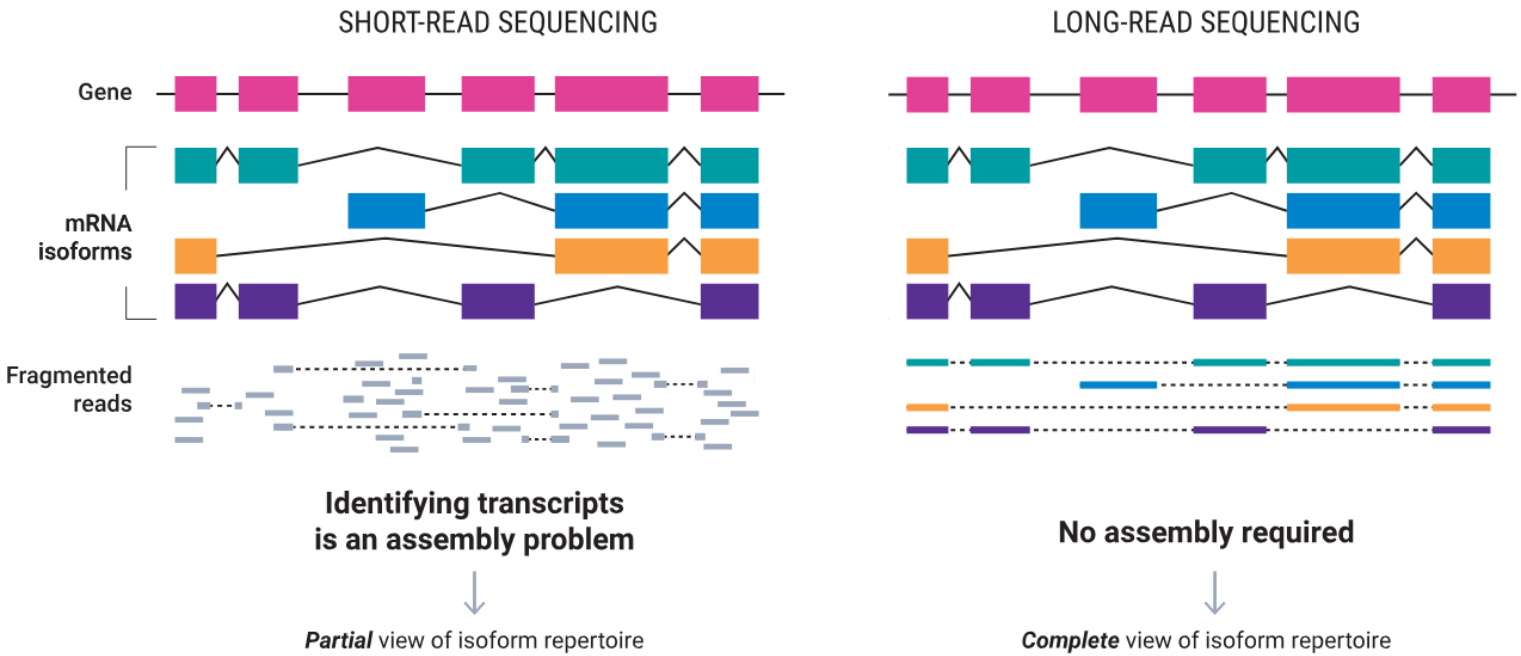


- Read length and accuracy trade off against each other.
- HiFi has largely replaced CLR.
- Both types also detect methylation (5mC and 6mA).

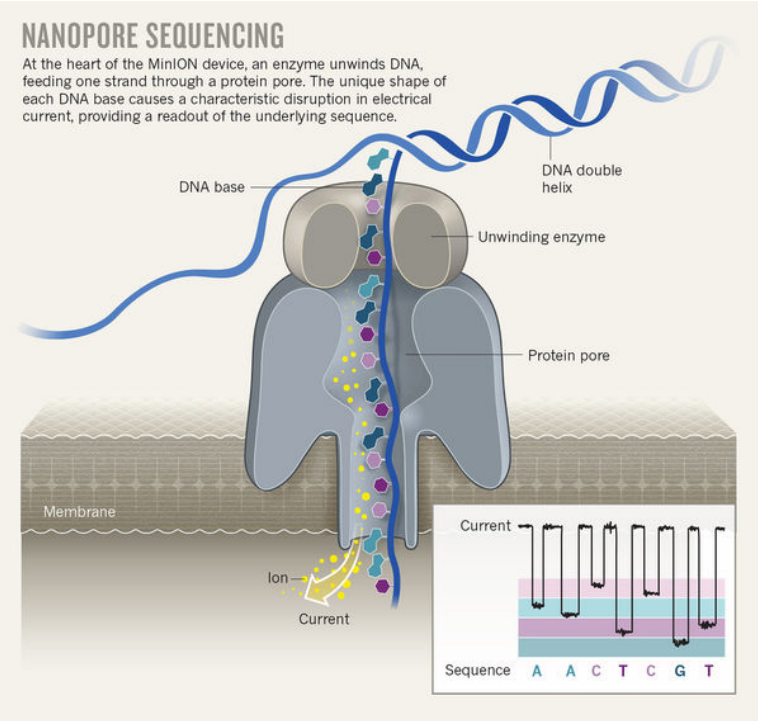
Iso-seq (PacBio): RNA Isoform Sequencing

Identifying transcript isoforms is challenging

- **standard RNA-seq** relies on short-read sequencing and computational assembly of overlapping reads spanning the splice-junction.
- **Iso-seq** (based on PacBio) allows isoform identification by sequencing of full length transcripts.



Oxford Nanopore Technology (ONT) – How it works

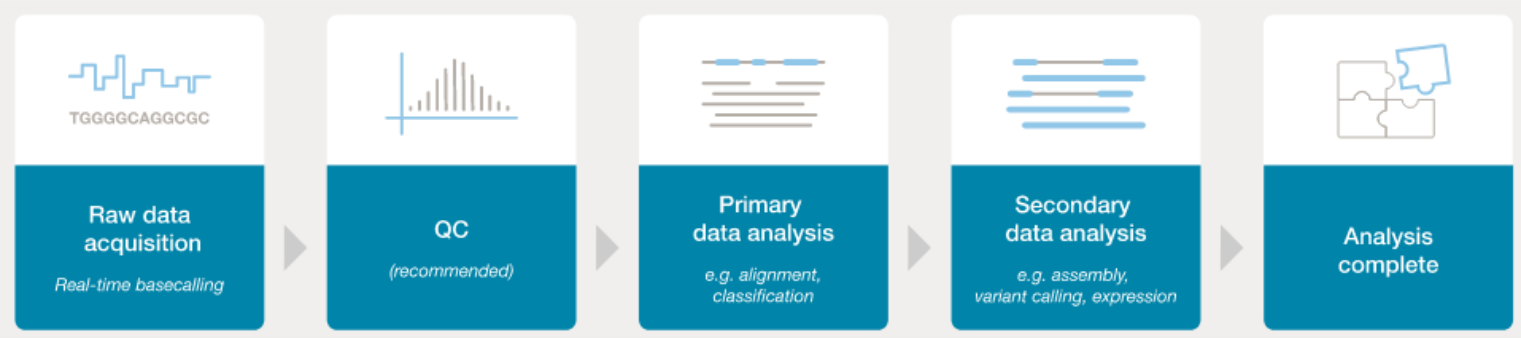




- flowcell contains 5×512 nanopores
- a single DNA strand is pushed through the pore
- about 5 bases are in the pore at a time
- the current change across the membrane reflects the change in base composition

Oxford Nanopore Technology (ONT) – Properties

- raw data is processed into base called data using **machine-learning**



Properties

+ portability (sequencing at remote places)
+ simple library preparation
+ long reads
+ direct detection of nucleotide modifications
– lower per-read accuracy than HiFi (simplex ∼99 %)
– different types of systematic errors

  e.g. problems with homopolymers (i.e. -TTTTTTTTT-)

Detecting Base Modifications

Oxford Nanopore enables detection of

- 4 DNA modifications (a.o. 5mC, 5hmC, 6mA)
- 5 RNA modifications (a.o. m^6A, m^5C, pseudouridine, inosine)

PacBio enables detection of

- 3 DNA modifications (5mC, 5hmC (since 2026), 6mA)

Illumina's 5-base solution (since 2025)

- does not truly detect a 5th base, but it uses
  – enzymatic conversion: 5mC $\rightarrow$ T (the reverse of bisulfite/EM-seq)
  – the opposite strand
  – and bioinformatics (DRAGEN) (similar to Bis-SNP)
  to predict 5mC in DNA

A community issue: 5mC (on DNA) and m^5C (on RNA) describe the same chemical change.

Sequencing Platforms in Comparison – Technical Data

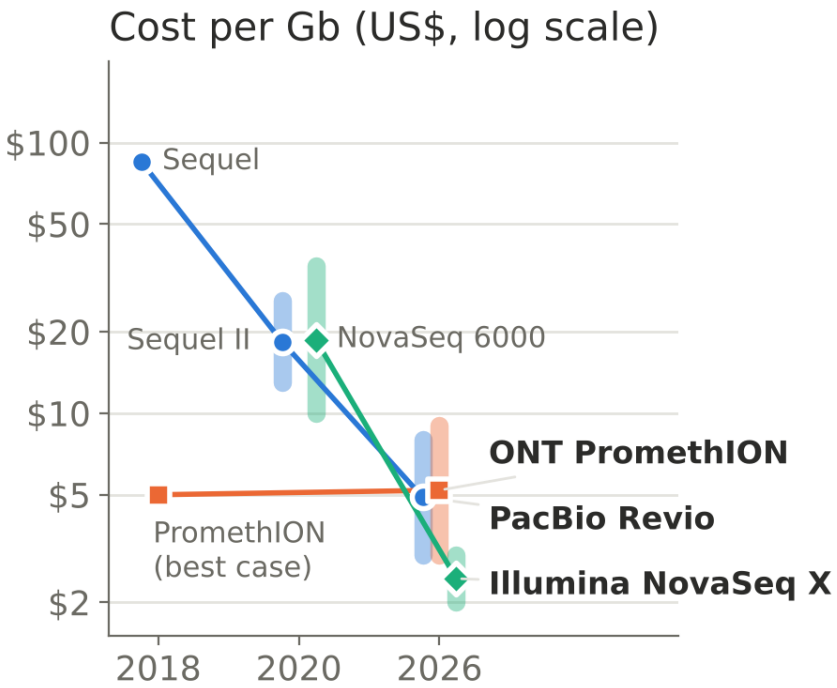
	Sanger ABI 3730xl	**Illumina** MiSeq i100	**Illumina** NovaSeq X	**PacBio** Revio (HiFi)	**ONT** PromethION
read length typical (max.)	700–900 bp (∼1 kb)	2×150 bp (2×500 bp)	2×150 bp (2×300 bp)	15–20 kb	10–50 kb (>1 Mb)
accuracy per read	99.9–99.99 %	>99.9 %	>99.9 %	99.9–99.95 %	∼99 % (duplex 99.9 %)
reads per flow cell[a]	96	10^8	$\sim 3 \times 10^{10}$	$\sim 7 \times 10^{6}$ [c]	$\sim 10^{7}$ [c]
output per flow cell[a]	∼0.09 Mb	30 Gb	8–10.5 Tb	90–120 Gb	100–200 Gb
run time	20 min–3 h	4–24 h	17–48 h	∼24 h	up to 72 h
cost per Gb[b]	$5–10 M	∼$60	$2–3	$3–8	$3–9
instrument price (list)	$290–400k	$49–109k	$985k	$599k	∼$450k

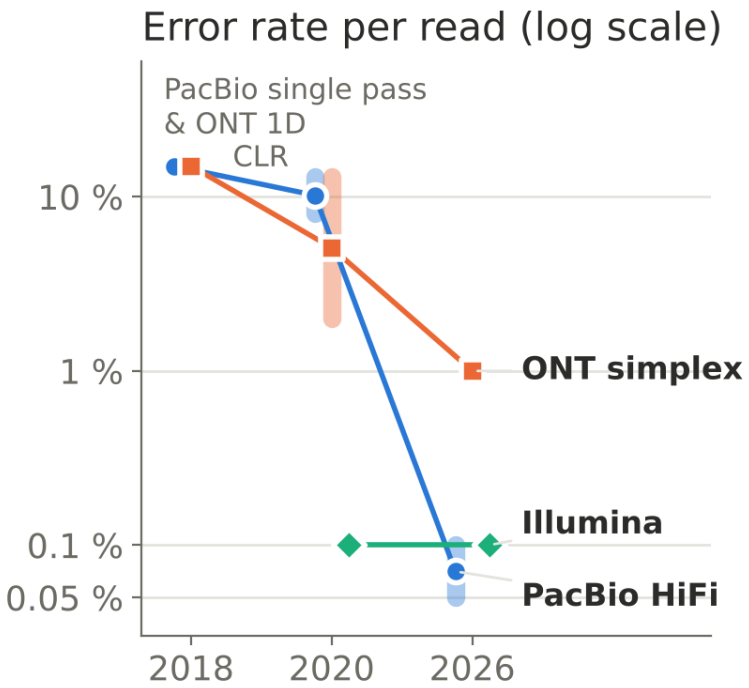
[a] Revio: per SMRT Cell (up to 4 per run); PromethION: per flow cell (up to 48 per run). [b] sequencing reagents or core-facility prices only, without library prep, labour and instrument. [c] estimated as output / mean read length. Maximum values are not all reached in the same run. As of Nov 2026.

Sequencing Platforms in Comparison – Properties

	Sanger ABI 3730xl	**Illumina** MiSeq i100	**Illumina** NovaSeq X	**PacBio** Revio (HiFi)	**ONT** PromethION
amplification	✓	✓	✓	✗	✗
single molecule, real-time	✗	✗	✗	✓	✓
native RNA	✗	✗	✗	✗ (cDNA)	✓
base modifications	✗	via conversion	via conversion	5mC, 5hmC, 6mA	4 DNA, 5 RNA
portable	✗	✗	✗	✗	✓ MinION
typical errors	read ends	substitu-tions	substitu-tions	homo-polymers	indels, homo-polymers
best for	single loci, clone checks	amplicons, small genomes	genomes, RNA-seq at scale	de novo assembly, isoforms	ultra-long reads, direct RNA, field work

The Numbers Age Fast: 2018 – 2020 – 2026





- Long reads have caught up with short reads in **cost per Gb** and **accuracy**.
- PacBio: from single-pass CLR (∼15 % errors) to HiFi consensus (<0.1 %).
- Always check the **date and source** of a technical comparison.

Bars: reported range; markers: geometric mean; time axis not to scale; labels name the instrument. 2018: Dijk *et al.* (2018) Trends in Genetics; 2020: Logsdon *et al.* (2020) Nature Reviews Genetics; 2026: manufacturer and core-facility data (reagents only).

Base (calling) quality – the FASTQ File

While there are different sequencing technologies fortunately a common file format for the base called sequencing data is used by all of them. It is an extension of the "fasta" format and is called "fastq". In the next slide is an example.

The 'Q' in 'fastq' stands for 'quality' and the **Q score (Phred quality score)** indicates the accuracy of the base call. It is the probability of a base call error.

$$Q = -10 \times log_{10}(p),$$

where p is the estimated probability that the base call is wrong.

- a logarithmic scale represents the probability of an incorrect base call

  Q10 – 1 in 10 chance of error (90% accuracy)

  Q20 – 1 in 100 chance of error (99% accuracy)

  Q30 – 1 in 1,000 chance of error (99.9% accuracy)

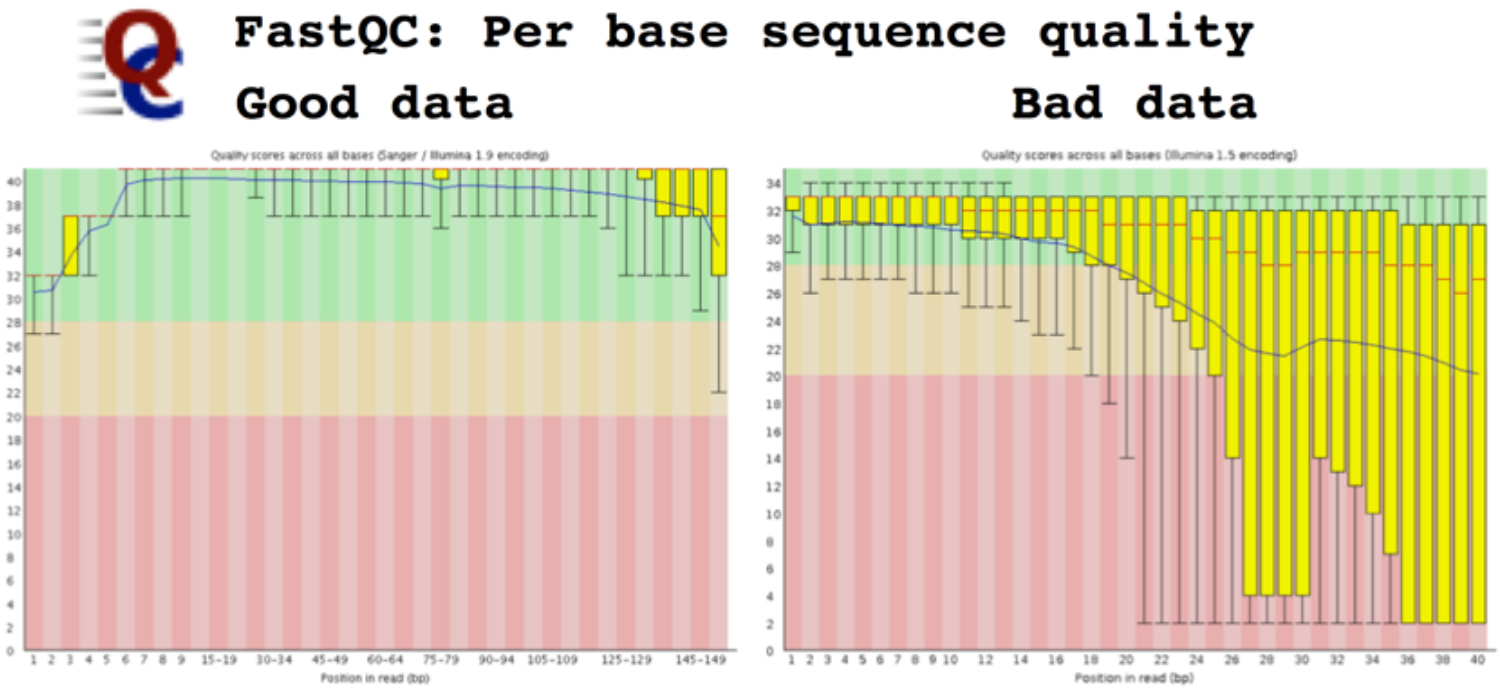
  Q40 – 1 in 10,000 chance of error (99.99% accuracy)

Base (calling) quality – the FASTQ File



- Line 1: label or sequence identifier – starts with @
- Line 2: raw sequence – string of A, C, G, T and N
- Line 3: separator line – starts with +
- Line 4: Phred quality score – string of ASCII characters
  Each character corresponds to a base position in line 2
  ASCII character codes specify the Q score

Base (calling) quality



Notice the different scaling in the left and right panel.

FastQC, written by Simon Andrews of Babraham Bioinformatics, is a very popular tool used to provide an overview of basic quality control metrics for raw next generation sequencing data. It is optimized for Illumina data. For longer reads there is specialized (but less popular) software. In the left panel, you can see the common pattern for good data. i.e. the quality at the beginning and the end is sightly worse than in the middle.

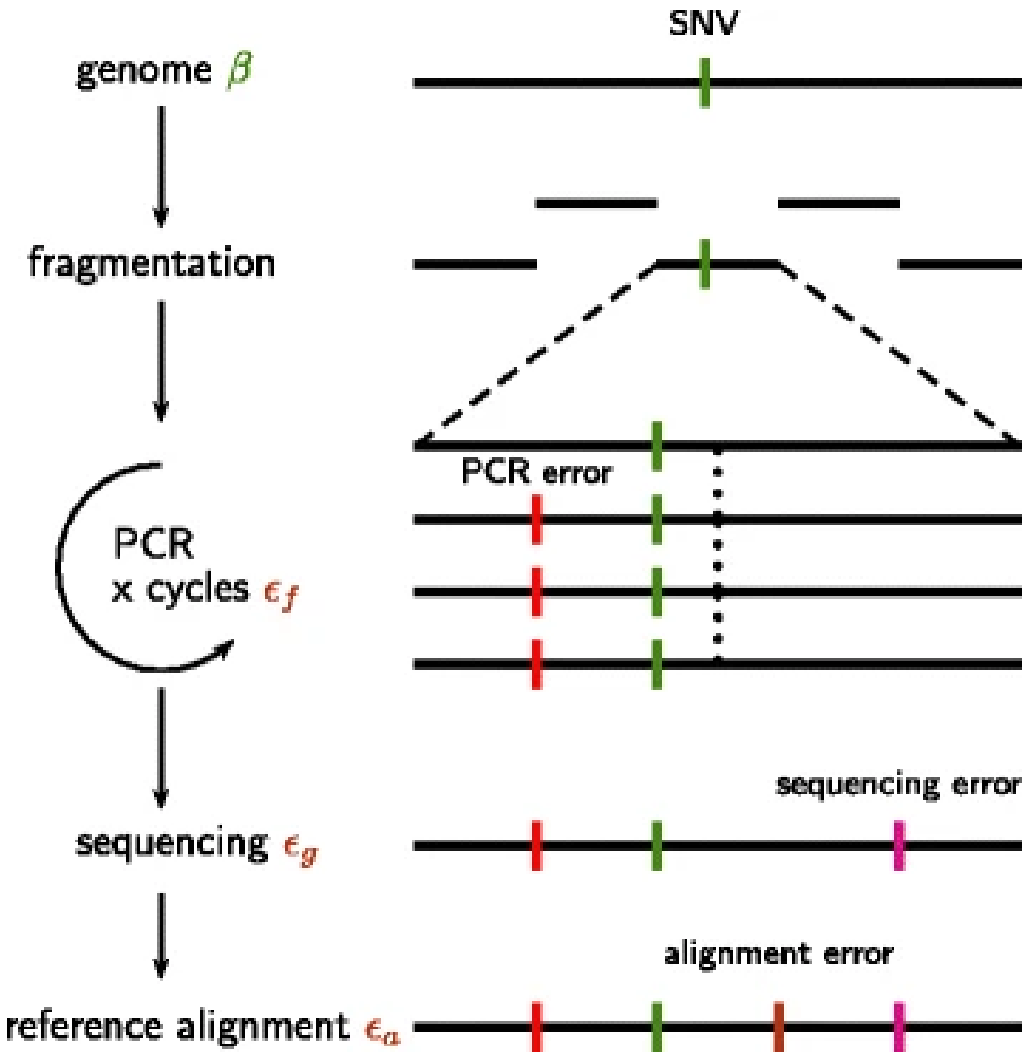
Base (calling) quality – the FASTQ File

It is important to keep in mind that the quality score only refers to the **quality of the base call**. Which means how confident is the base caller that it called the base correctly. If there were alterations to the sequence before the sequencing (e.g. during PCR) a very good quality score can still be achieved despite the exchange of a nucleotide relative to the original sequence.

The **base calling quality** is not to confuse with the **read mapping quality** which specifies how well a read matches the reference sequence.

Different Sources of Variation/Errors

Single nucleotides that do not match the reference sequence can have different sources.



Summary

In sequencing, the following key parameters matter

- how much input DNA is required?
- is the query sequence amplified before sequencing?
- what is the (average and maximum) read length?
- what is the averange base calling acurracy?
- what is the cost per Gb (gigabase, 10^9 bases)?
- how fast are the results available?
- can epigenetic modifications be directly detected?
- **which platform suits the sequencing task best?**

The goal is to understand which sequencing platform (e.g. Sanger, Illumina, PacBio, Nanopore) is the best choice for a specific sequencing task, and why.

Sequencing - Self-assesment questions

Which sequencing technology should be used for the following hypothetical projects and why? Assume that sequencing platforms are available.

- You want to sequence a single gene locus covering a few hundred nucleotides.
- You want to sequence a large number of Corona test probes per day in a cheap and reliable way.
- You suspect that a human genome underwent genomic rearangements (probably due to cancer) and want to sequence a few samples.
- You want to do the later in the middle of an undeveloped country. (You have to bring your sequencer with you.)
- You want to generate a new animal genome for which no reference genome is available. You have moderate funding.
- You want to sequence RNA and directly measure RNA modifications.

If you have trouble answering the questions, read the review by Logsdon *et al.* (2020): `https://doi.org/10.1038/s41576-020-0236-x`