An Introduction to Sequencing Technologies

Image created using Microsoft copilot

A M Mahedi Hasan, PhD | CBiol MRSB | FHEA
Senior Research Associate (Computational Biology)
Treatment Resistance Group
UCL Cancer Institute

What will we learn today?

- Short read sequencing (next generation sequencing) techniques
- Long read sequencing (next to next generation sequencing) techniques
- Basis of library preparations
- Different DNA (and RNA) sequencing-based omics technology
- Single-cell (RNA) sequencing techniques
- Application of these techniques

(Bulk) Sequencing technologies

Short-read NGS (next generation sequencing):

- Sequencing by ligation (SBL): SOLiD and Complete Genomics (BGI)
- Sequencing by synthesis (SBS):

    Cycle Reversible Termination (CRT): Illumina, Qiagen

    Single Nucleotide Addition (SNA): 454, Ion Torrent

- Long-read sequencing (next to next generation sequencing): PacBio, ONT

Library preparation

Rationale:

- Sequencing by ligation:

  a probe sequence that is bound to a fluorophore hybridises to a DNA fragment and is ligated to an adjacent oligonucleotide for imaging. The emission spectrum of the fluorophore indicates the identity of the base or bases complementary to specific positions within the probe.

- Sequencing by synthesis:

  a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into an elongating strand.

- Two common steps in library preparation:

  Fragmentation

  Ligation to a common adaptor set for clonal amplification

Library preparation (cont..)

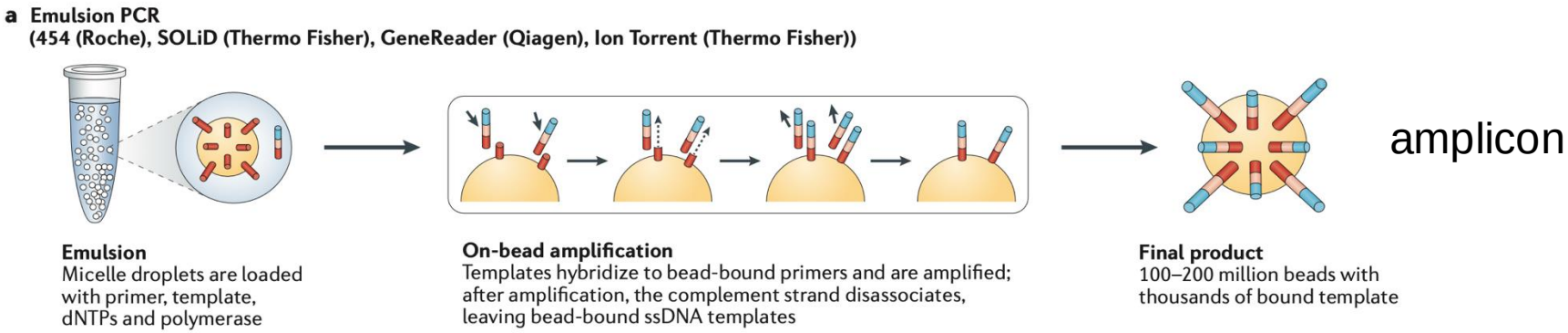
- The most important goal for most of SBL and SBS library preparation techniques is
- achieving clonal template DNA, sufficient enough (by forming a cluster of clonal template DNA) to get a clear signal.
- Each of these clusters can perform as reaction centres.
- A sequencing platform can collect information from many millions of reaction centres simultaneously, thus sequencing many millions of DNA molecules in parallel. It ensures “high throughput”.

Library preparation (cont..)

Different strategies used to generate clonal template populations:

1. **Bead-based** (SOLiD, Ion Torrent)
2. **Solid-state based** (Illumina)
3. **DNA nanoball generation** (Complete Genomics BGI)

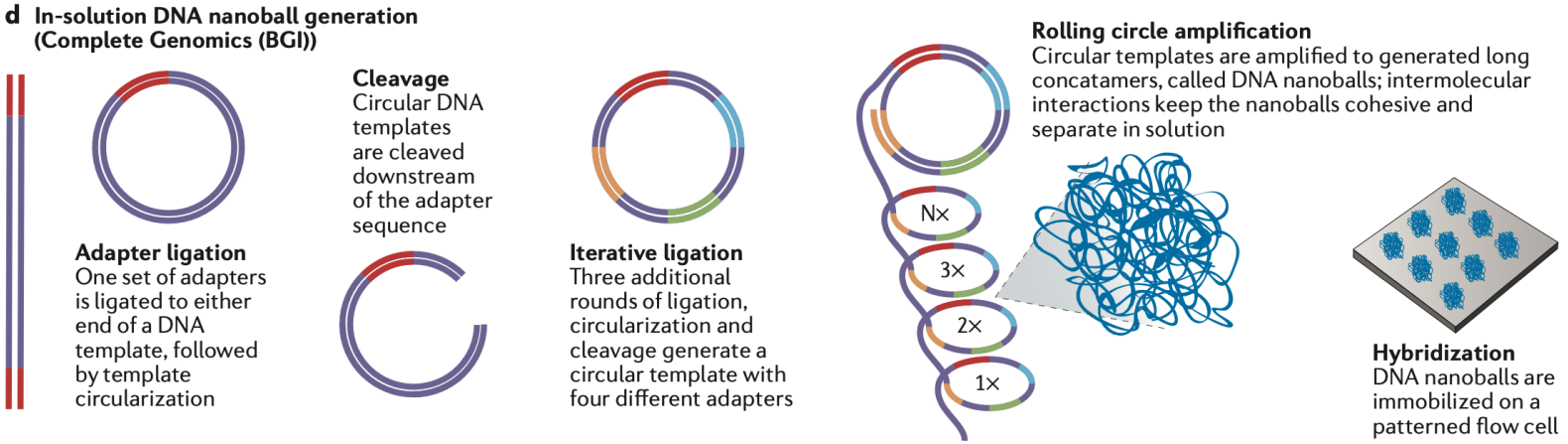
Library preparation (cont..)



bead-based (SOLiD, Ion Torrent):

- one adaptor is complementary to an oligo nucleotide fragment that is immobilized on a bead.
- Using emulsion PCR (emPCR), the DNA template is amplified such that as many as one million clonal DNA fragments are immobilised on a single bead.
- These beads can be distributed onto a glass surface or arrayed on a PicoTiterPlate (Roche Diagnostics).

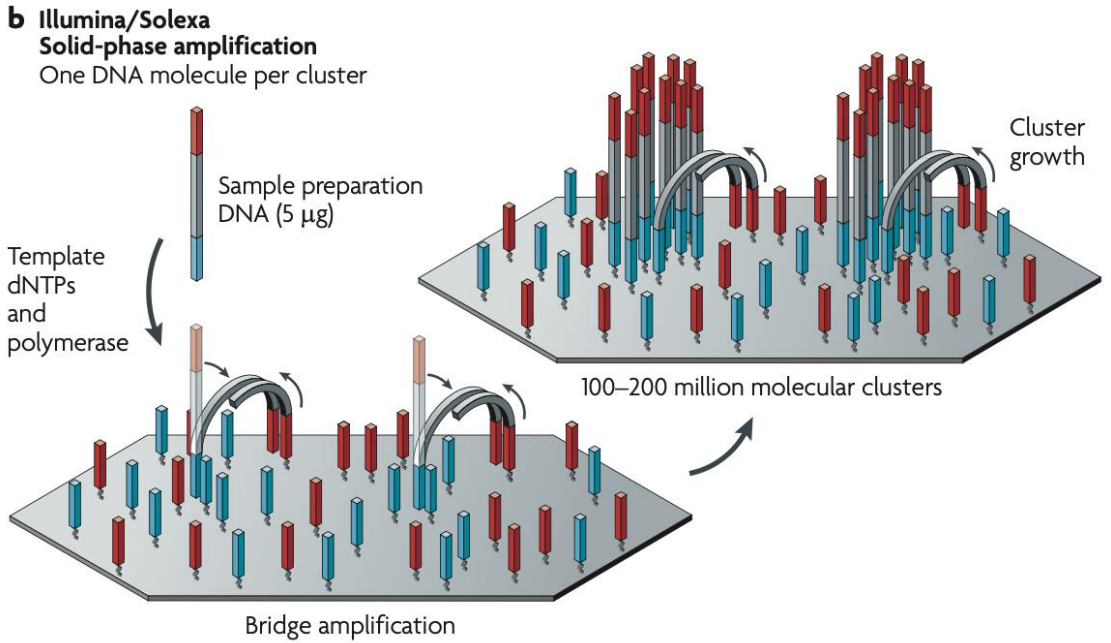
Library preparation (cont..)



DNA nanoball generation (Complete Genomics BGI):

- Currently only technology for template enrichment in solution.
- DNA undergoes an iterative ligation, circularization and cleavage process to create a circular template, with four distinct adaptor regions.
- Through the process of rolling circle amplification (RCA), up to 20 billion discrete DNA nanoballs are generated.
- The nanoball mixture is then distributed onto a patterned slide surface containing features that allow a single nanoball to associate with each location

Library preparation (cont..)



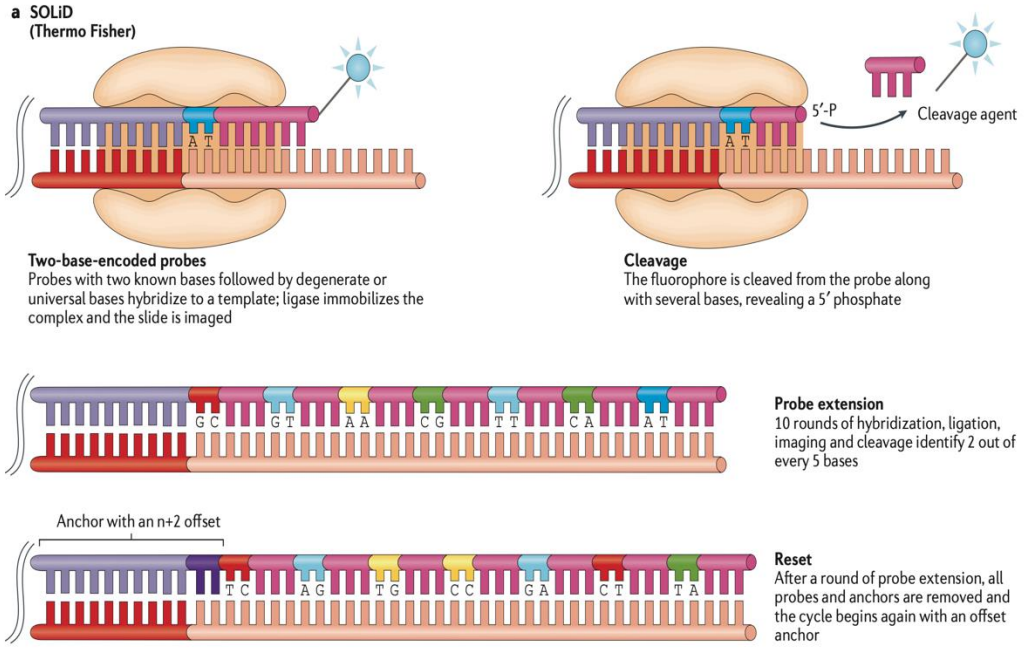
Solid-state based (Illumina):

Composed of two basic steps:

1. Initial priming and extending of the single-stranded, single-molecule template, and
2. Bridge amplification of the immobilised template with immediately adjacent primers to form clusters.

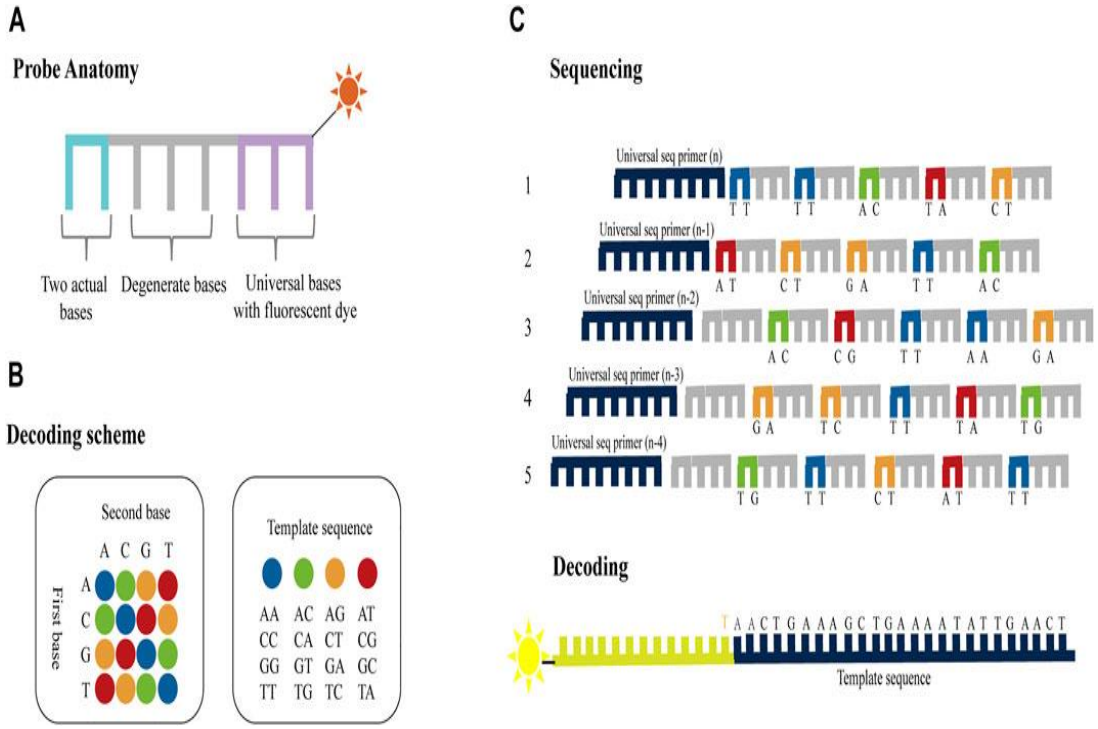
Sequencing by ligation

- Involve the hybridisation and ligation of labelled probe and anchor sequences to a DNA strand
- The anchor fragment encodes a known sequence that is complementary to an adapter sequence and provides a site to initiate ligation.
- The probes encode one or two known bases (one-base-encoded probes or two-base-encoded probes) and a series of degenerate or universal bases, driving complementary binding between the probe and template
- After ligation, the template is imaged (for the colour-coded fluorophore) and the known base (one-base-encoded) or bases (two-base-encoded) in the probe are identified. A new cycle begins after complete removal of the anchor–probe complex or through cleavage to remove the fluorophore and to regenerate the ligation site.

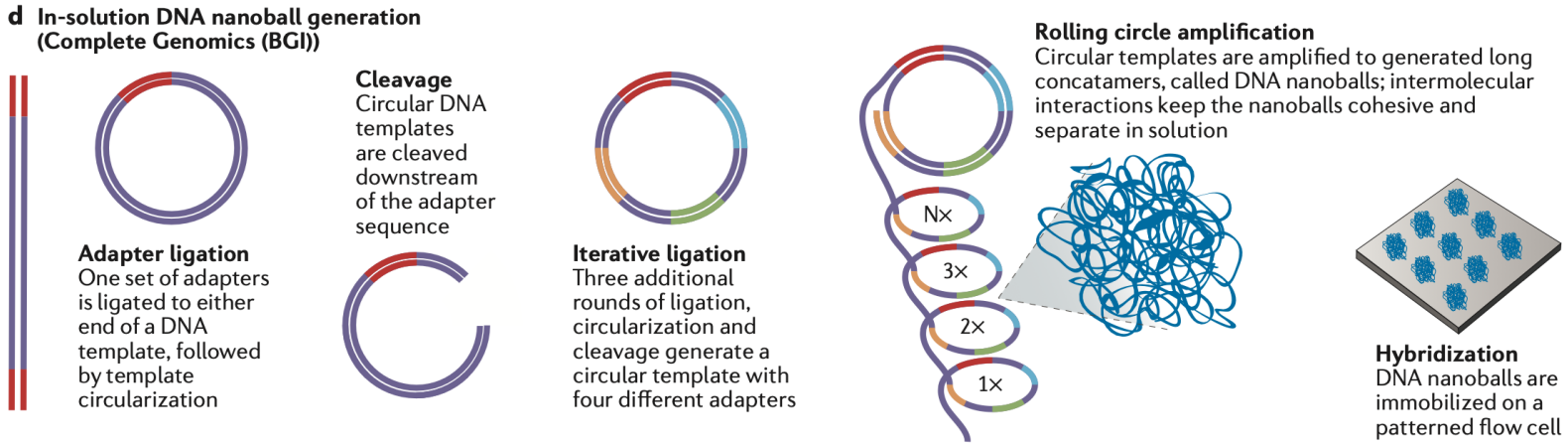


Sequencing by ligation (cont)

- The SOLiD platform utilizes two-base-encoded probes, in which each fluorometric signal represents a dinucleotide. Consequently, the raw output is not directly associated with the incorporation of a known nucleotide. Because the 16 possible dinucleotide combinations cannot be individually associated with spectrally resolvable fluorophores, four fluorescent signals are used, each representing a subset of four dinucleotide combinations (for dinucleotides by each colour).
- Thus, each ligation signal represents one of several possible dinucleotides, leading to the term colour-space (rather than base-space), which must be deconvoluted during data analysis.
- The SOLiD sequencing procedure is composed of a series of probe–anchor binding, ligation, imaging and cleavage cycles to elongate the complementary strand. Over the course of the cycles, single-nucleotide offsets (either n+1 or n-1) are introduced to ensure every base in the template strand is sequenced.

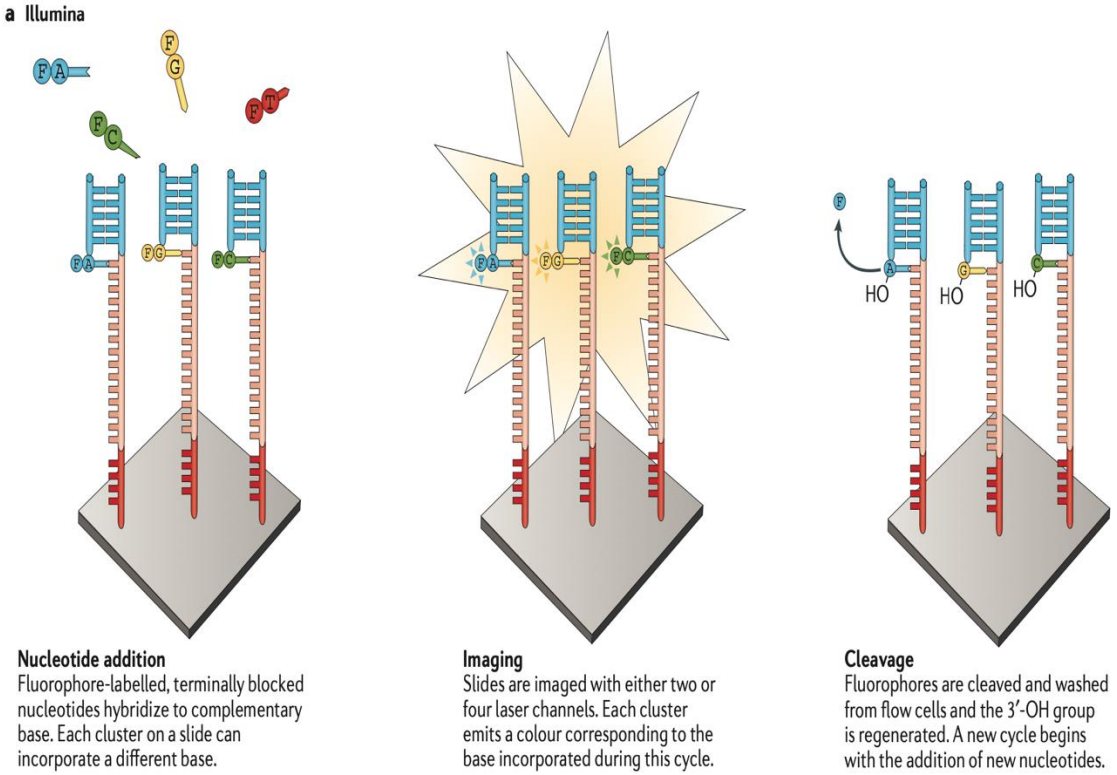


Complete Genomics Platform



- DNA sequencing using combinatorial probe–anchor ligation (**cPAL**) or combinatorial probe–anchor synthesis (**cPAS**).
- In cPAL, an anchor sequence (complementary to one of the four adaptor sequences) and a probe hybridize to a DNA nanoball at several locations. In each cycle, the hybridizing probe is a member of a pool of **one-base-encoded probes**, in which each probe contains a known base in a constant position and a corresponding fluorophore. After imaging, the entire probe–anchor complex is removed
- A new probe–anchor combination is hybridized. Each subsequent cycle utilizes a probe set with the known base in the **n + 1** position. Further cycles in the process also use adaptors of variable lengths and chemistries, allowing sequencing to occur upstream and downstream of the adaptor sequence.
- The cPAS approach is a modification of cPAL intended to increase read lengths of Complete Genomics' chemistry; however, at present, details about the approach are limited.

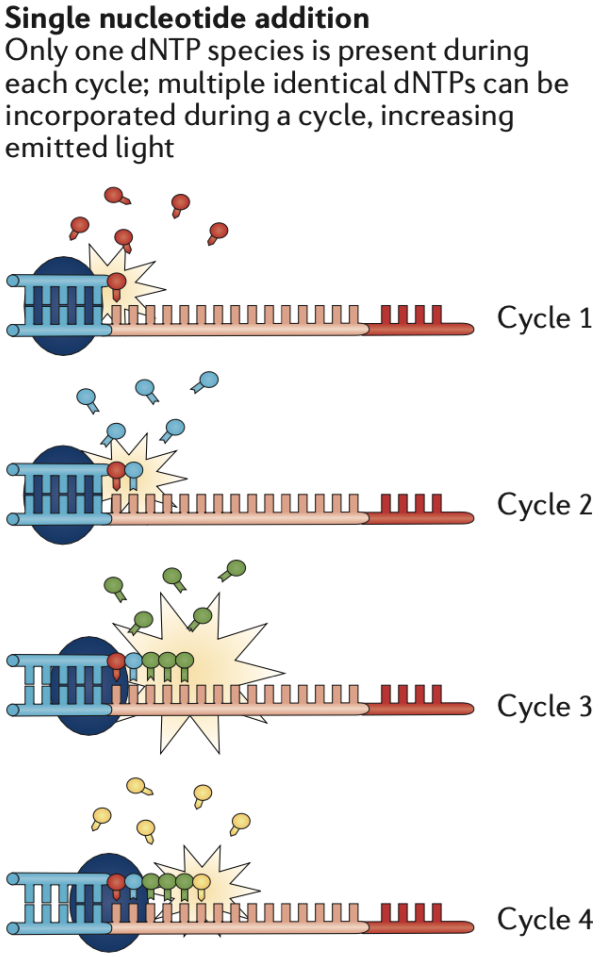
Sequencing by synthesis: CRT (Illumina)



- A DNA template is primed by a sequence that is complementary to an adapter region, which will initiate polymerase binding to this double-stranded DNA (dsDNA) region.
- During each cycle, a mixture of all four individually labelled and 3′-blocked deoxynucleotides (terminator bound dNTPs) are added. After the incorporation of a single dNTP to each elongating complementary strand, unbound dNTPs are removed and the surface is imaged to identify which dNTP was incorporated at each cluster.
- The fluorophore and blocking group can then be removed and a new cycle can begin.
- Illumina's suite of instruments for short-read sequencing range from small, low-throughput benchtop units to large ultra-high-throughput instruments dedicated to population-level whole-genome sequencing (WGS).
- dNTP identification is achieved through total internal reflection fluorescence (**TIRF**) microscopy using either two or four laser channels.
- In most Illumina platforms, each dNTP is bound to a single fluorophore that is specific to that base type and requires four different imaging channels, whereas the NextSeq and Mini-Seq (even the current top of the range NovaSeq) systems use a two-fluorophore system.. The Illumina CRT system accounts for the largest market share for sequencing instruments compared to other platforms

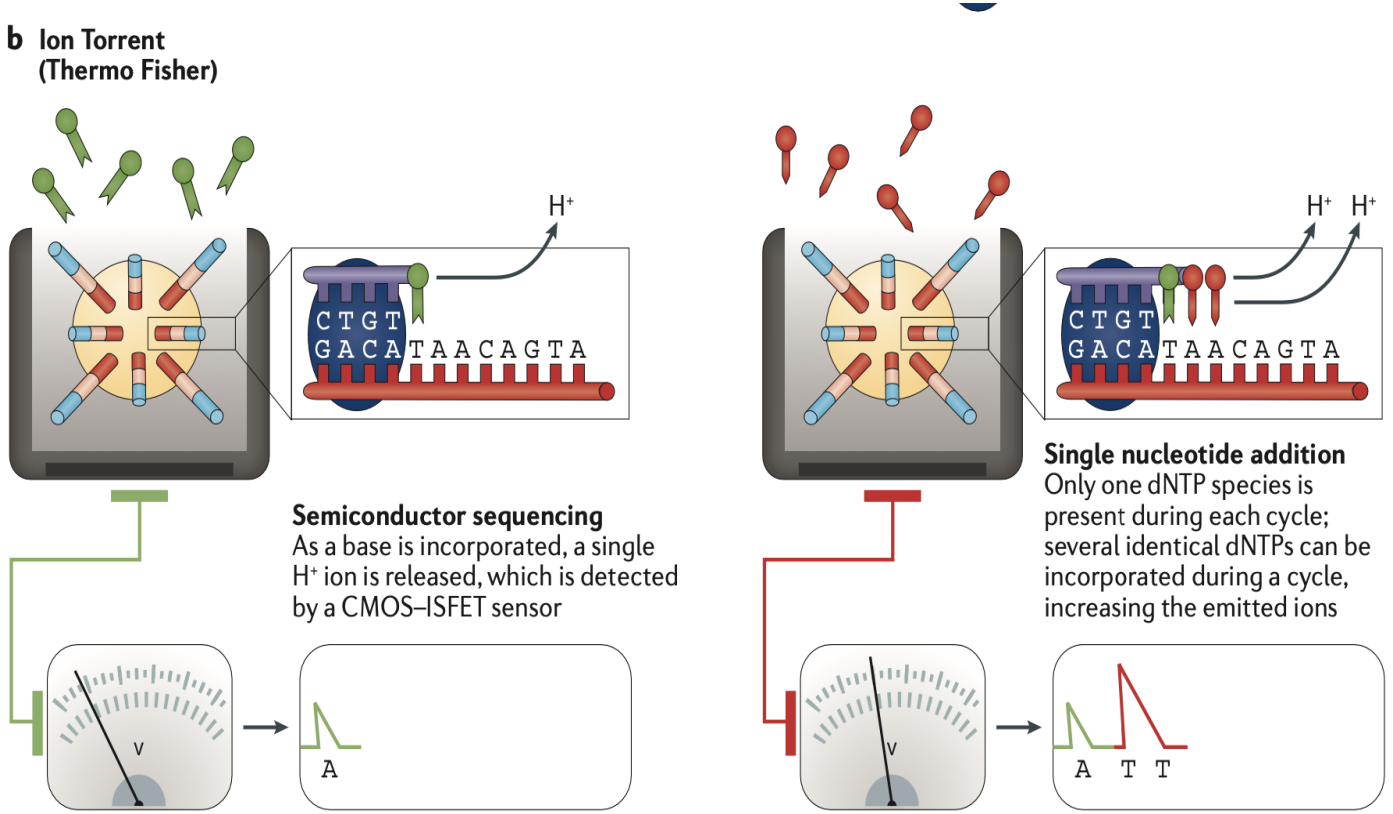
Sequencing by synthesis: SNA (454, Ion Torrent)

- Unlike CRT, SNA approaches rely on a single signal to mark the incorporation of a dNTP into an elongating strand. As a consequence, each of the four nucleotides must be added iteratively (cycle 1 –> cycle 2 –> cycle 3 –> cycle 4) to a sequencing reaction to ensure only one dNTP is responsible for the signal.
- Furthermore, this does not require the dNTPs to be blocked, as the absence of the next nucleotide in the sequencing reaction prevents elongation.
- The exception to this is homopolymer regions where identical dNTPs are added, with sequence identification relying on a proportional increase in the signal as multiple dNTPs are incorporated.
- The first NGS instrument developed was the 454 pyrosequencing device. It was discontinued in 2016.



Sequencing by synthesis: SNA (454, Ion Torrent)

- The Ion Torrent was the first NGS platform without optical sensing. The Ion Torrent platform detects the H^+ ions that are released as each dNTP is incorporated.
- The resulting change in pH is detected by an integrated complementary metal-oxide semiconductor (**CMOS**) and an ion-sensitive field-effect transistor (ISFET).
- The pH change detected by the sensor is imperfectly proportional to the number of nucleotides detected, allowing for limited accuracy in measuring homopolymer lengths.



Comparison of short-read platforms

SOLiD and Complete Genomics systems:

- A very high accuracy **(~99.99%)** as each base is probed multiple times.
- But true variants are missed while **few false variants are called**.
- There is also evidence that the platforms share some under-representation of AT-rich regions and the SOLiD platform displays some substitution errors and some GC-rich under-representation.
- The most limiting factor is the widespread adoption of these technologies is the **very short read lengths**. 75 bp for SOLiD and up to 100 bp for Complete Genomics, limiting their use for genome assembly and structural variant detection applications.
- The runtimes on the order of **several days**, the SOLiD system has been relegated to a small niche within the industry.

Illumina:

- **Dominates** the short-read sequencing industry owing, in part, to its maturity as a technology, a **high level of cross-platform compatibility** and its wide range of platforms.
- The suite of instruments available ranges from the **low-throughput MiniSeq** to the **ultra-high-throughput HiSeq X**, which is capable of sequencing ~1,800 human genomes to 30× coverage per year. Further diversification is derived from the many options available for runtime, read structure and read length (up to 300 bp).
- As the Illumina platform relies on a CRT approach, it **is much less susceptible to the homopolymer errors observed in SNA platforms**.
- Although it has an overall accuracy rate of >99.5%, the platform does display some **under-representation in AT-rich and GC-rich regions**, as well as a tendency **towards substitution errors**.

Comparison of short-read platforms

Both the 454 and the Ion Torrent systems -

- offer superior read lengths compared to other short-read sequencers with reads up to an **average of 700 bp and 400 bp, respectively**, providing some advantages for applications that focus on repetitive or complex DNA.
- Insertion and deletion (indel) errors dominate, although the overall error rate is similar to other NGS platforms in non-homopolymer regions.
- Homopolymer regions are problematic for these platforms, which lack single-base accuracy in measuring homopolymers larger than 6–8 bp.

Illumina suite allows for a wide range of applications

- Genome sequencing through WGS or exome sequencing
- Epigenomics applications, such as ChIP–Seq (chromatin immunoprecipitation followed by sequencing),
- ATAC–seq (assay for transposase-accessible chromatin using sequencing) or
- DNA methylation sequencing (methyl-seq); and
- Transcriptomics applications through RNA sequencing (RNA-seq)
- The two-colour labelling system used by the NextSeq and MiniSeq platforms increase speed and reduces costs by reducing scanning to two colour channels and reducing fluorophore usage. However, the two-channel system results in a slightly higher error profile and underperformance for low-diversity samples owing to more ambiguous base discrimination.

Long read sequencing

Currently, there are two main types of long-read technologies:

1. single-molecule real-time sequencing approaches

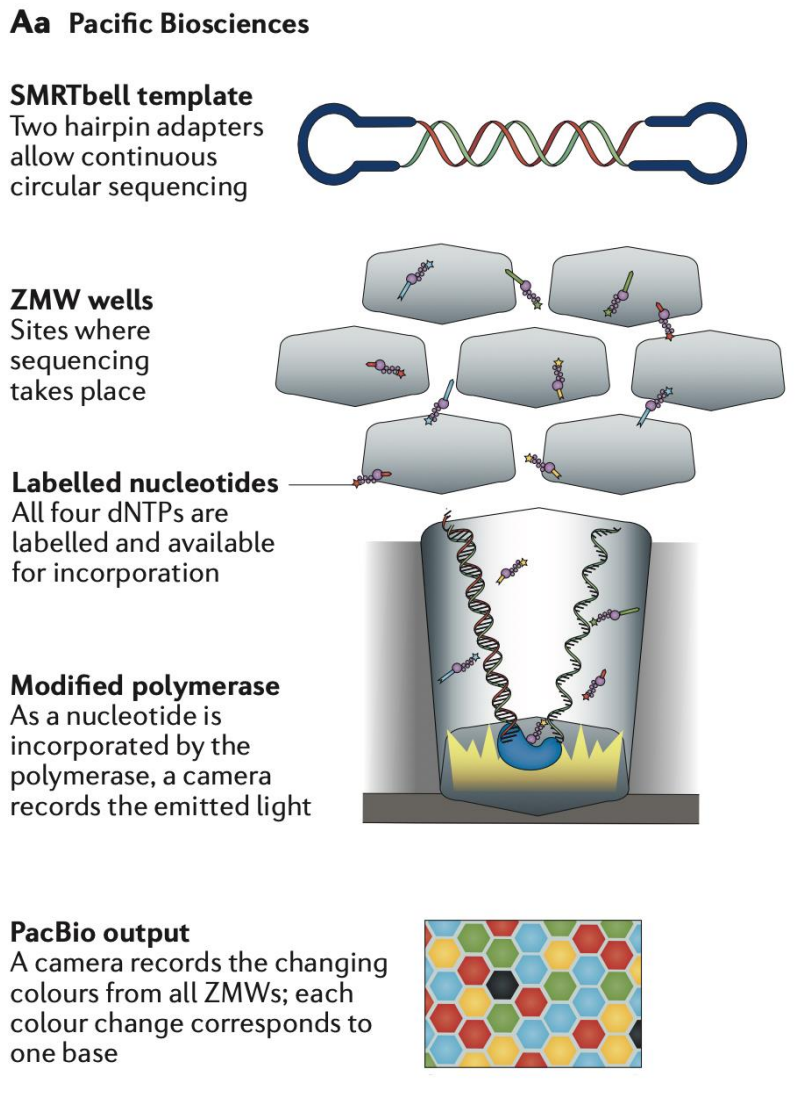
    - The single-molecule approaches differ from short-read approaches in that they do not rely on a clonal population of amplified DNA fragments to generate detectable signal, nor do they require chemical cycling for each dNTP added.

2. synthetic approaches that rely on existing short-read technologies to construct long reads in silico.

    - the synthetic approaches do not generate actual long-reads; rather, they are an approach to library preparation that leverages barcodes to allow computational assembly of a larger fragment.

Single-molecule long-read sequencing: Pacific Biosciences (PacBio)

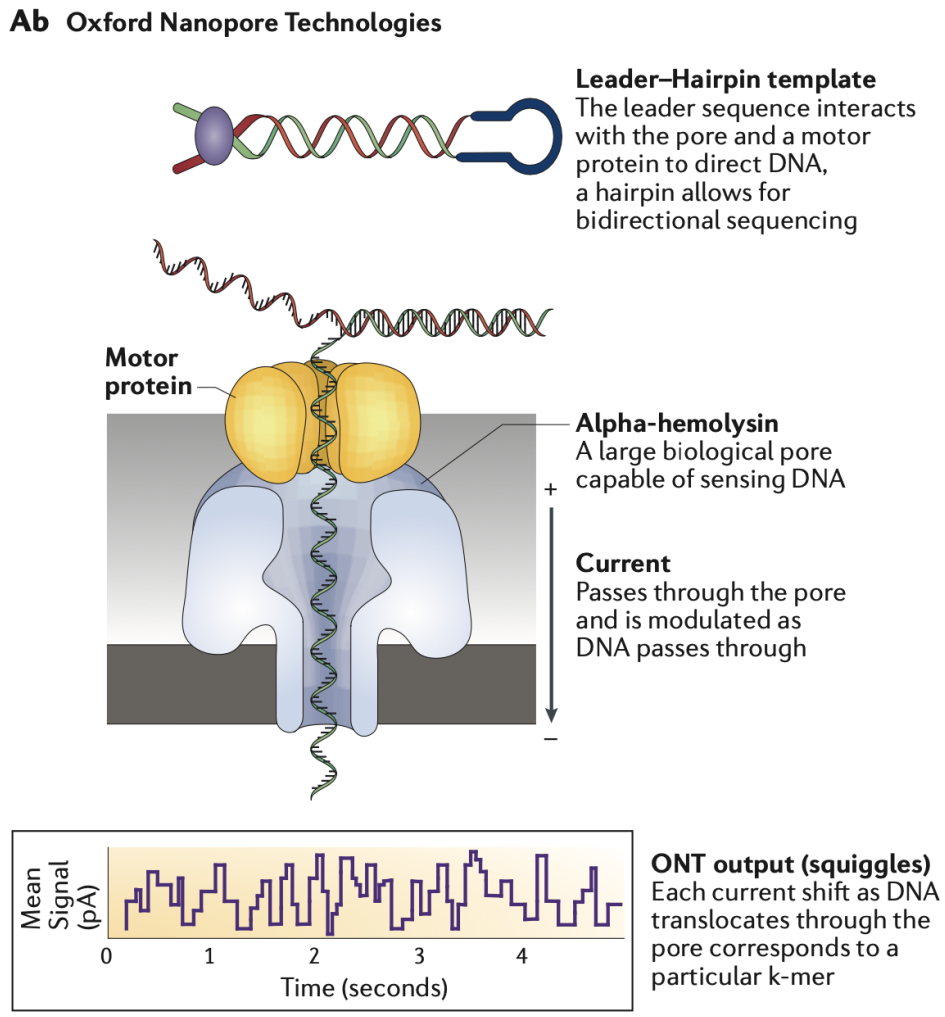
- The instrument uses a specialized flow cell with many thousands of individual picolitre wells with transparent bottoms — zero-mode waveguides (ZMW).
- PacBio fixes the polymerase to the bottom of the well and allows the DNA strand to progress through the ZMW.
- By having a constant location of incorporation owing to the stationary enzyme, the system can focus on a single molecule.
- dNTP incorporation on each single-molecule template per well is continuously visualised with a laser and camera system that records the colour and duration of emitted light as the labelled nucleotide momentarily pauses during incorporation at the bottom of the ZMW.
- The polymerase cleaves the dNTP-bound fluorophore during incorporation, allowing it to diffuse away from the sensor area before the next labelled dNTP is incorporated.
- SMRTbell template is formed by using Two hairpin adapters at both ends of the template DNA allow continuous circular sequencing (allows each template to be sequenced multiple times)
- Although it is difficult for DNA templates longer than ~3 kb to be sequenced multiple times, shorter DNA templates can be sequenced many times as a function of template length.



Single-molecule long-read sequencing:

Oxford Nanopore Technologies (ONT):

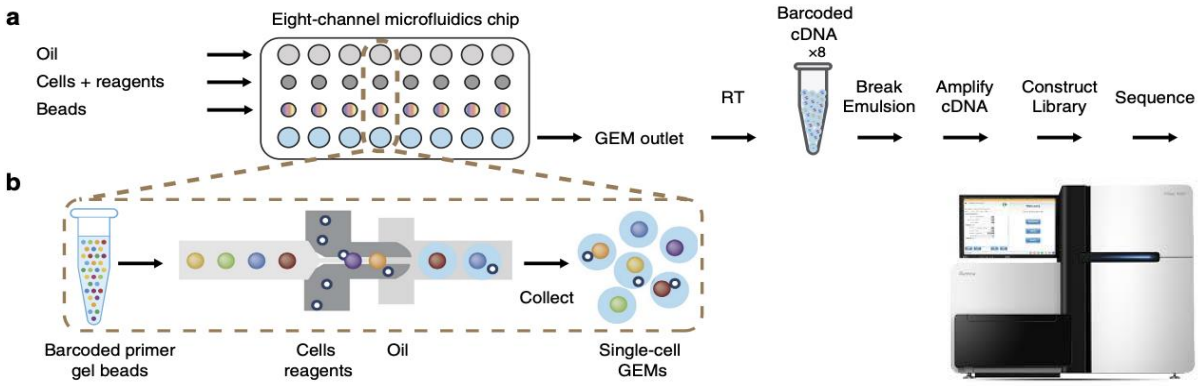
- In 2014, the first consumer prototype of a nanopore sequencer — the MinION.
- Nanopore sequencers directly detect the DNA composition of a native ssDNA molecule.
- To carry out sequencing, DNA is passed through a **protein pore** as current is passed through the pore. As the DNA translocates through the action of a secondary motor protein, a voltage blockade occurs that modulates the current passing through the pore.
- The temporal tracing of these charges is called **squiggle space**, and shifts in voltage are characteristic of the particular DNA sequence in the pore, which can then be interpreted as a k-mer.
- Rather than having 1–4 possible signals, the instrument has more than 1,000 — one for each possible k-mer, especially when modified bases present on native DNA are taken into account.
- ONT MinION uses a leader– hairpin library structure. This allows the forward DNA strand to pass through the pore, followed by a hairpin that links the two strands, and finally the reverse strand. This generates 1D and 2D reads in which
- both '1D' strands can be aligned to create a consensus sequence '2D' read.



Benefits of long read sequencing

- Genomes are highly complex with many long repetitive elements, copy number alterations and structural variations that are relevant to evolution, adaptation and disease.
- However, many of these complex elements are so long that short-read paired-end technologies are insufficient to resolve them.
- Long-read sequencing delivers reads in excess of several kilobases, allowing for the resolution of these large structural features.
- Such long reads can span complex or repetitive regions with a single continuous read, thus eliminating ambiguity in the positions or size of genomic elements.
- Long reads can also be useful for transcriptomic research, as they are capable of spanning entire mRNA transcripts, allowing researchers to identify the precise connectivity of exons and discern gene isoforms.

Single cell sequencing



- Cells were combined with reagents in one channel of a microfluidic chip, and gel beads from another channel to form GEMs.
- Gel beads loaded with primers and barcoded oligonucleotides are first mixed with cells and reagents, and subsequently mixed with oil-surfactant solution at a microfluidic junction.
- Single-cell GEMs are collected in the GEM outlet.
- RT takes place inside each GEM, after which cDNAs are pooled for amplification and library construction in bulk.
- During data processing, reads from a cell is clustered based on UMI.

Applications

- WGS has become one of the most widely used applications in NGS. Through this technology, researchers can obtain –
    1. the most comprehensive view of genomic information and
    2. associated biological implications
- 1000 Genomes Project
- Whole-exome and targeted sequencing
- By limiting the size of the genomic material used, more individual samples can be sequenced within a sequencing run which can increase both the breadth and the depth of a genomic study.
- coverage as high as 10,000X can be required to validate rare variants.
- Protein–DNA interactions:
- Can be probed by enriching for protein-interacting DNA fragments, often through immunoprecipitation as in the case of ChIP–Seq.
- Conversely, ATAC–Seq uses a hyperactive transposase to generate short-read NGS compatible DNA fragments from regions unprotected by proteins or nucleosomes.

Applications (cont..)

- Methylation patterns:
- Methyl-Seq involves the capture and enrichment of methylated DNA, selective
- digestion of methylated or unmethylated regions and/or modification of a methylated base such that it introduces a SNP into the DNA sequence.

- Long read sequence:
- A recent paradigm shift in NGS is the ability to sequence very long stretches of DNA.
- Repetitive and complex regions have historically been difficult to assemble and resolve using short-read sequencing approaches. Recently, using long-read sequencing technology, Chaisson *et al* were able to add more than 1 Mb of novel sequence to the human GRCh37 reference genome through gap closure and extension, and they identified >26,000 indels that were ≥50 bp in length, providing one of the most comprehensive reference genomes available.
- Transcriptomic research has also benefited from greater accessibility to NGS. Today, researchers are leveraging the power of NGS to deeply sequence down to single-transcript sensitivity.
- Recently, researchers at the Stanley Royd Hospital in the United Kingdom used MinION sequencing to monitor an outbreak of Salmonella enterica.
- MinION sequencing was used to monitor the transmission history and evolution of the Ebola virus as the outbreak unfolded at the European Mobile Laboratories in Guinea.

References

- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49
- Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews. Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x
- Bagger, F. O., Borgwardt, L., Jespersen, A. S., Hansen, A. R., Bertelsen, B., Kodama, M., & Nielsen, F. C. (2024). Whole genome sequencing in clinical practice. *BMC Medical Genomics*, *17*(1), 39–16. https://doi.org/10.1186/s12920-024-01795-w
- Singh, V. (2022). Advances, challenges, and opportunities in DNA sequencing technology. In *New Frontiers and Applications of Synthetic Biology*. Elsevier Science & Technology. https://doi.org/10.1016/B978-0-12-824469-2.00022-1
- Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002
- Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049–14049. https://doi.org/10.1038/ncomms14049
- https://www.youtube.com/watch?v=mI0Fo9kaWqo&t=16s
- https://www.youtube.com/watch?v=PFwSe09dJX0&t=290s