ACS Partner Journal

CHEMICAL & BIOMEDICAL IMAGING

Open Access

pubs.acs.org/ChemBioImaging

Review

# Sequencing and Optical Genome Mapping for the Adventurous Chemist

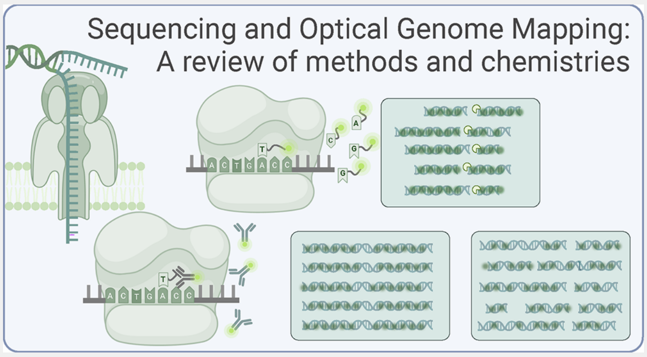
Elizabete Ruppeka Rupeika, Laurens D'Huys, Volker Leen, and Johan Hofkens*

Cite This: *Chem. Biomed. Imaging* 2024, 2, 784−807

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** This review provides a comprehensive overview of the chemistries and workflows of the sequencing methods that have been or are currently commercially available, providing a very brief historical introduction to each method. The main optical genome mapping approaches are introduced in the same manner, although only a subset of these are or have ever been commercially available. The review comes with a deck of slides containing all of the figures for ease of access and consultation.


Sequencing and Optical Genome Mapping:
A review of methods and chemistries

**KEYWORDS:** *sequencing methods, genome mapping, optical mapping, DNA, genomic sequences*

## INTRODUCTION

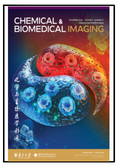
For billions of years, nature has encoded information within deoxyribonucleic acid (DNA), a biopolymer serving as the guide for the genetic architecture, operations, and upkeep of all living beings. Environmental influences together with the genetic code produce the remarkable range of observable phenotypic diversity of living entities.

The four letters of the DNA code are highly conserved organic molecules called nucleotides. Each nucleotide consists of a deoxyribose, a phosphate, and one of four nucleobases. There are two types of nucleobases—purines and pyrimidines. Adenine (A) and guanine (G) are purines, while cytosine (C) and thymine (T) are pyrimidines. In a single DNA strand, adjacent nucleotides are connected by a phosphodiester bond between the ribose and the phosphate moieties, forming the hydrophilic sugar−phosphate backbone. DNA and nucleobase structures are given in Figure 1. As discovered by Watson and Crick, using X-ray diffraction images of DNA made by Franklin and Wilkins, two single-strand polymers intertwine to form the well-known double helix.[1,2]

The double helix is stabilized by the sugar−phosphate backbone of each single strand and the hydrogen bonding between complementary hydrophobic nucleobases internally (Figure 1).[3] The complementary pairs are G and C and A and T. There is directionality to the DNA strands, and the strands in the double helix are antiparallel. Each strand has a five-prime (5′) end and a three-prime (3′) end, where the 3′ end of one strand matches with the 5′ end of the complementary strand (Figure 1).[3] DNA synthesis in vivo is always in the direction 5′ to 3′, with the leading strand being elongated continuously in the 5′ to 3′ (Figure 2) direction and the lagging 3′−5′ strand being synthesized discontinuously in the 5′−3′ direction via so-called Okazaki fragments—short rows of nucleotides.[3−5] DNA naturally exists in at least three different conformations, referred to as A-, Z-, and B-DNA, differing in the helical twist per base pair, the helical orientation, and the sectional diameter of the helix.[6] For example, B-DNA, which is the most naturally abundant, has a base-to-base distance (the rise per basepair) of 0.34 nm and a diameter of around 2 nm.[6]

The sequence in which the four nucleotides appears is not random.[7,8] DNA is transcribed into ribonucleic acid sequences (RNA), which in turn is read in sets of trinucleotides called codons.[8] Each codon corresponds to an amino acid, and some amino acids are encoded by several codons; together, the codon sequences constitute genes.[8] Protein coding regions account for only about 1% of the human genome: the rest is noncoding DNA.[9] The noncoding sequences are not fully understood, but we know that they contain regulatory elements like promoter and enhancer regions, which have spatiotemporal control over gene expression.[9] Some of these regions can be transcribed into regulatory ribonucleic acid (RNA) molecules.[9,10] These molecules participate in processes like gene expression, protein building, and protein production.

Received: August 14, 2024
Revised: October 3, 2024
Accepted: October 4, 2024
Published: October 25, 2024



https://doi.org/10.1021/cbmi.4c00060

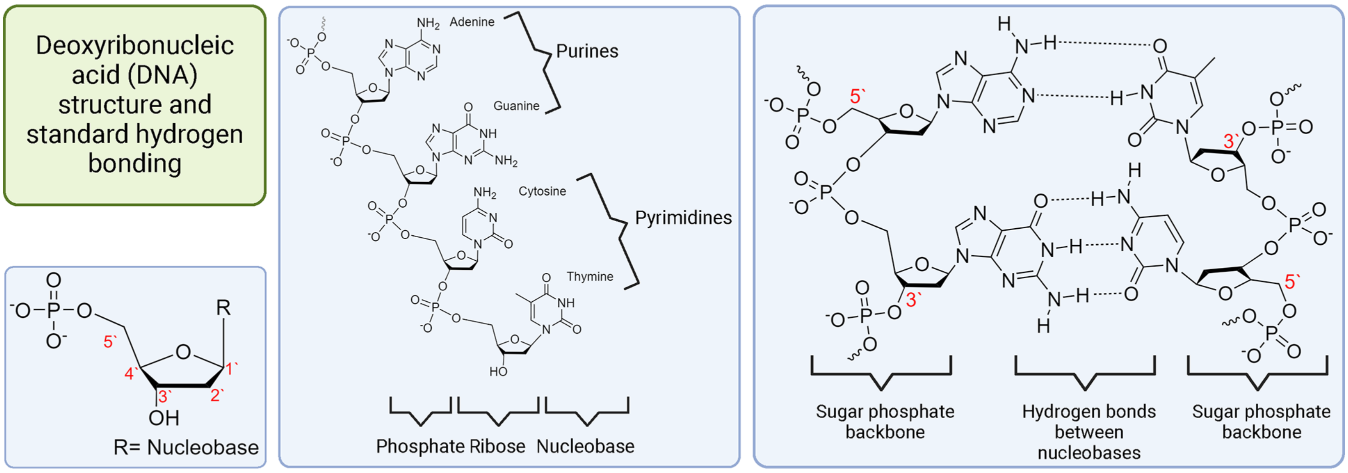


**Figure 1.** DNA structure. From left: nucleobase position numbering, purines and pyrimidines, and sugar phosphate backbone and hydrogen bonds between the two antiparallel strands.

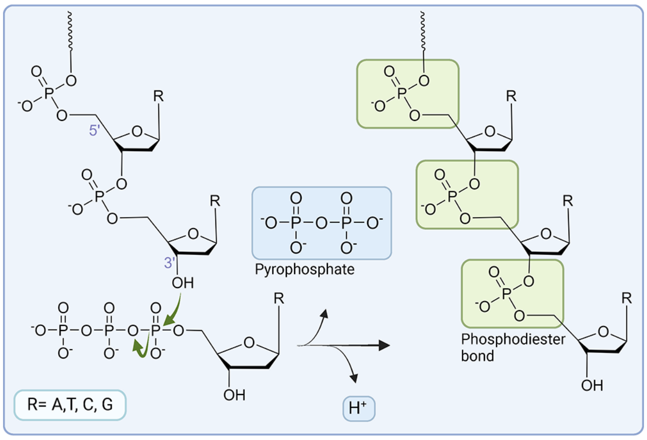


**Figure 2.** DNA elongation, triphosphate addition, release of a pyrophosphate, and formation of the phosphodiester bond of a DNA strand.

The central dogma of molecular biology posits that information flows from the level of genetic sequence to protein: DNA sequence is transcribed into RNA, which is subsequently translated into polymers of amino acids.[11] These polymers fold into proteins, which are the building blocks of every structure in the body (cell organelles, cells, tissues, enzymes etc.): the sequence of codons is paramount for successful protein production. An additional level of control is achieved via secondary modifications to the genetic sequence.[12] The field of investigation is called epigenetics, and the research covers modifications to the nucleotides and associated proteins that contribute to the control over what regions of the DNA are active and available for the enzymes responsible for transcription and translation of genomic information. These processes are reviewed at large elsewhere.[12]

The expansion of the field of genetics has fundamentally impacted other fields from medicine to agriculture. A notable milestone for the field was the publishing of the first full assembly of the human genome, which was the result of a race between the publicly funded, international collaboration called the Human Genome project (HGP) and the enterprise of Craig Venter, later known as Celera.[13,14] More about this unique and spectacular intersection between science and international politics that resulted in simultaneous *Nature* and *Science* publications in 2003 can be found in other dedicated sources.[13−15]

This review is formatted as a handbook of the chemistries employed by the main sequencing approaches that are or have been used by the global research community. In the coming sections the chemistry of the different reactions that contribute to the final base call are examined in detail. The methods are described in approximately chronological order. There are three generations of sequencing that are considered, starting with first attempts at reading the DNA sequence as such, followed by development of massively parallel sequencing approaches using polymerase chain reaction (PCR) amplification, and finally, arriving at PCR-free methods that read single molecules of DNA. These sections all consider base by base readouts of the genomic sequence. The generation of public repositories of genomic sequences from sequencing enabled development of methods that "read" genomic DNA in shorthand to identify known sequences or structural variants of sequences. These methods are generally referred to as genomic mapping and exhibit a range of techniques to obtain sequence-specific profiles of genomic sequences. The last sections of the review examine several strategies of genomic mapping, which all fall under the umbrella of optical mapping.

## ■ DNA SEQUENCING

### First Wave of DNA Sequencing

First-generation sequencing refers to two methods, both characterized by the use of polyacrylamide gel electrophoresis, that are named after their inventors: Maxam−Gilbert sequencing and Sanger sequencing. The Sanger and Coulson method reached the scientific community in 1975 (building up on previous publications, including 1973), while the Maxam and Gilbert method was developed around 1976 and first reached the scientific community in 1977.[16−18]

**Maxam−Gilbert Sequencing.** Maxam−Gilbert sequencing is based on cleavage of terminally radioactively ($^{32}$P) labeled single-stranded DNA (ssDNA) or double-stranded DNA (dsDNA) using chemical treatments that are base selective.[18] The radioactive nature of $^{32}$P allowed visualizing $^{32}$P-labeled molecules on an X-ray film, also known as an autoradiograph. The principle is the following: four combinations of reagents are used on four copies of the same DNA material to chemically damage and remove one base at a time

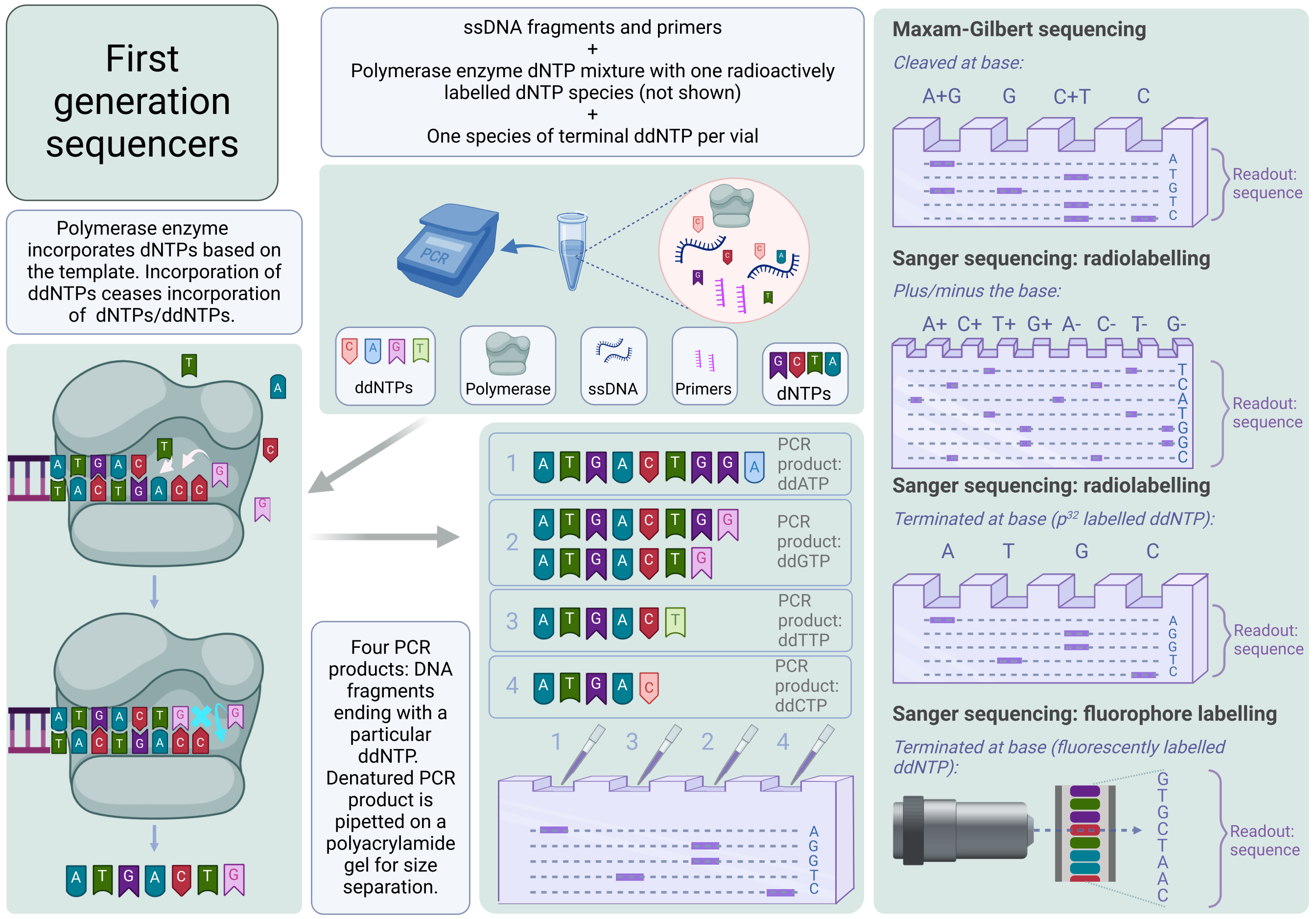

First generation sequencers
Polymerase enzyme incorporates dNTPs based on the template. Incorporation of ddNTPs ceases incorporation of dNTPs/ddNTPs.
ssDNA fragments and primers
+
Polymerase enzyme dNTP mixture with one radioactively labelled dNTP species (not shown)
+
One species of terminal ddNTP per vial
PCR
ddNTPs
Polymerase
ssDNA
Primers
dNTPs
1 A T G A C T G G A PCR product: ddATP
2 A T G A C T G G PCR product: ddGTP
A T G A C T G
3 A T G A C T PCR product: ddTTP
4 A T G A C PCR product: ddCTP
Four PCR products: DNA fragments ending with a particular ddNTP. Denatured PCR product is pipetted on a polyacrylamide gel for size separation.
1 3 2 4
A G G T C
A T G A C T G
Maxam-Gilbert sequencing
Cleaved at base:
A+G G C+T C
A T G T C
Readout: sequence
Sanger sequencing: radiolabelling
Plus/minus the base:
A+ C+ T+ G+ A- C- T- G-
T C A T G G C
Readout: sequence
Sanger sequencing: radiolabelling
Terminated at base (p32 labelled ddNTP):
A T G C
A G G T C
Readout: sequence
Sanger sequencing: fluorophore labelling
Terminated at base (fluorescently labelled ddNTP):
G T G C T A A C
Readout: sequence

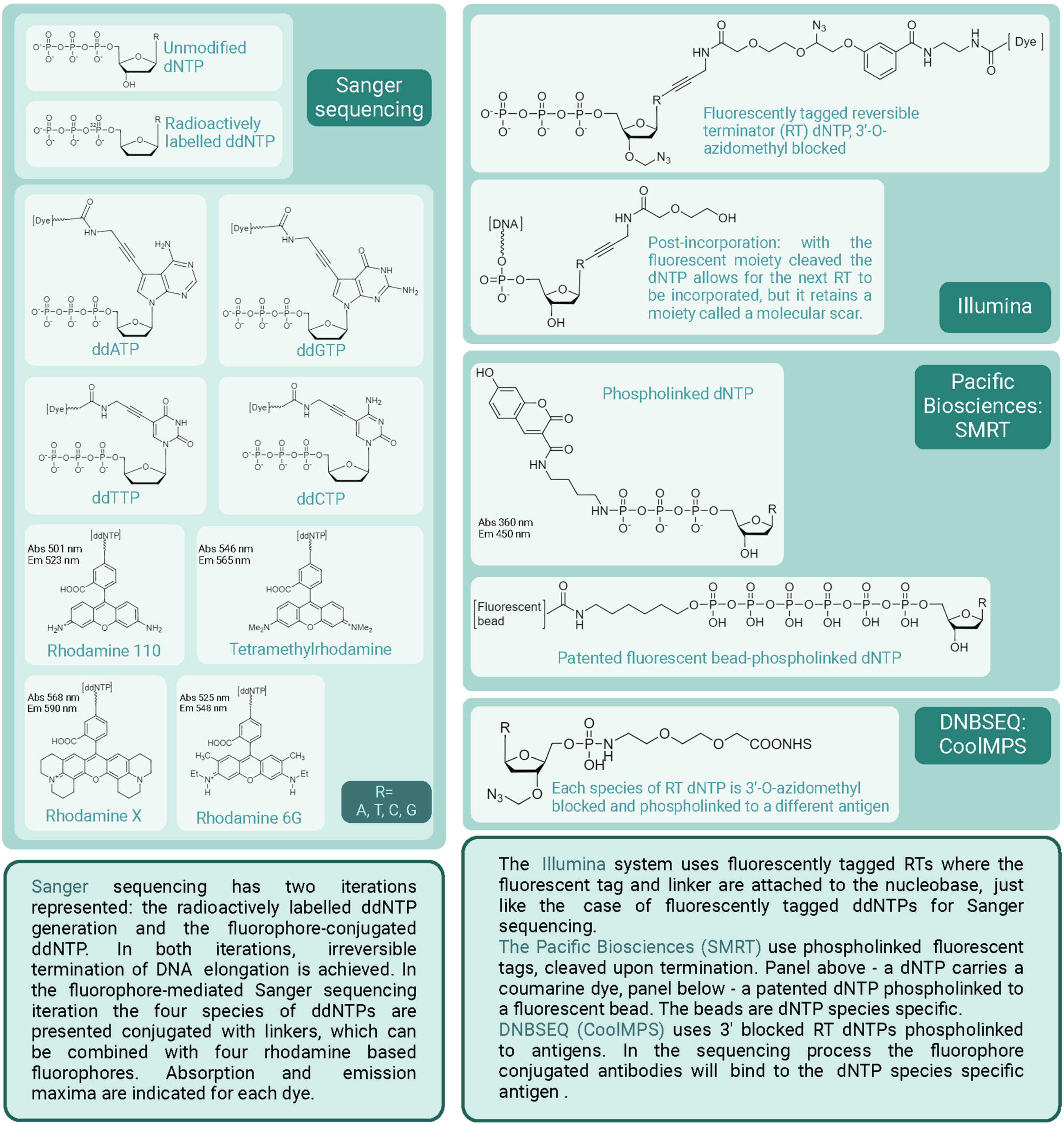
**Figure 3.** Overview of early sequencing methods: Radiolabeled Sanger sequencing (left and center) and (right) gel electrophoresis readout for (1) Maxam−Gilbert sequencing, (2) Sanger sequencing with radioactively labeled ddNTPs, and (3) Sanger sequencing with chain terminating radioactively labeled ddNTPs, and (4) Sanger sequencing with fluorescently labeled chain terminating ddNTPs.

to expose the corresponding sugar, which is also then cleaved from its 3′ and 5′ labels. The resulting fragments of DNA terminate at the different respective bases and are visualized on a polyacrylamide gel, where the distances that the fragments migrate during gel electrophoresis are directly correlated to the size of the fragment, which is visualized using a control lane with fragments of known sizes. The readout is a ladder of fragments where ordering the fragments by size will reveal the sequence of the bases.[18]

The chemistry of these four reactions is as follows.[18] T base moieties are cleaved from the backbone using hydrazine ($N_2H_4$), thereby exposing the weakened backbone and l. Hydrazine in combination with high salt concentrations (usually around 1.5−2 M) suppresses the hydrazine reaction with T, effectively targeting C. In both cases, the weakened backbone is cleaved using piperidine. For chemical treatment of A and G, dimethyl sulfate ($(CH_3)_2SO_4$) can be added to methylate G at the N-7 position and A at the N-3 position. As a result, A and G become susceptible to breakage at their glycosidic bond upon exposure to high temperatures and neutral pH. The DMS methylation reaction occurs five times faster for G, making this step more G specific. Additionally, the glycosidic bond of methylated A is less stable than that of methylated G. The use of formic acid after DMS methylation enhances the cleavage of A, bringing the A specificity closer to that of G. Finally, just like for T and C, the backbone is cleaved using piperidine. All four reactions start with fragmentation of the genomic DNA into fragments of random lengths. Once the reactions are complete, the reaction products are loaded on a polyacrylamide gel for readout (Figure 3, right). On this gel, a direct readout of T and A can be obtained, while G and C can be indirectly inferred from combining information from multiple lanes.[18]

**Sanger Sequencing.** In Sanger and Coulson's sequencing strategy, also called the plus−minus method, fragments of random size are generated through cyclic reactions with polymerase-mediated extension using a primer, deoxyribonucleotide triphosphates (dNTP), and a ssDNA template (Figure 3).[17]

After each cycle, the unincorporated dNTPs are removed from the reaction mixture. For the minus reaction, three dNTPs (all except the one in the plus reaction) are added. In contrast, only a single dNTP is added to the plus reaction. This is done for all dNTP combinations.[17] The resulting total of eight product samples is loaded in individual lanes of a polyacrylamide gel and visualized thanks to the radioactive label ($^{32}P$) that was incorporated in one of the dNTPs during the reaction (Figure 3, right). The readout is a ladder of fragments with the same initial sequence but terminating at different lengths with a known terminating element.[17] The sequence can be resolved by combining information from all of the lanes. Every pair of plus and minus samples yields fully


Unmodified dNTP
Radioactively labelled ddNTP
Sanger sequencing
ddATP
ddGTP
ddTTP
ddCTP
Abs 501 nm Em 523 nm
Rhodamine 110
Abs 546 nm Em 565 nm
Tetramethylrhodamine
Abs 568 nm Em 590 nm
Rhodamine X
Abs 525 nm Em 548 nm
Rhodamine 6G
R= A, T, C, G
Fluorescently tagged reversible terminator (RT) dNTP, 3'-O-azidomethyl blocked
Post-incorporation: with the fluorescent moiety cleaved the dNTP allows for the next RT to be incorporated, but it retains a moiety called a molecular scar.
Illumina
Phospholinked dNTP
Pacific Biosciences: SMRT
Abs 360 nm Em 450 nm
Fluorescent bead
Patented fluorescent bead-phospholinked dNTP
DNBSEQ: CoolMPS
Each species of RT dNTP is 3'-O-azidomethyl blocked and phospholinked to a different antigen

Sanger sequencing has two iterations represented: the radioactively labelled ddNTP generation and the fluorophore-conjugated ddNTP. In both iterations, irreversible termination of DNA elongation is achieved. In the fluorophore-mediated Sanger sequencing iteration the four species of ddNTPs are presented conjugated with linkers, which can be combined with four rhodamine based fluorophores. Absorption and emission maxima are indicated for each dye.

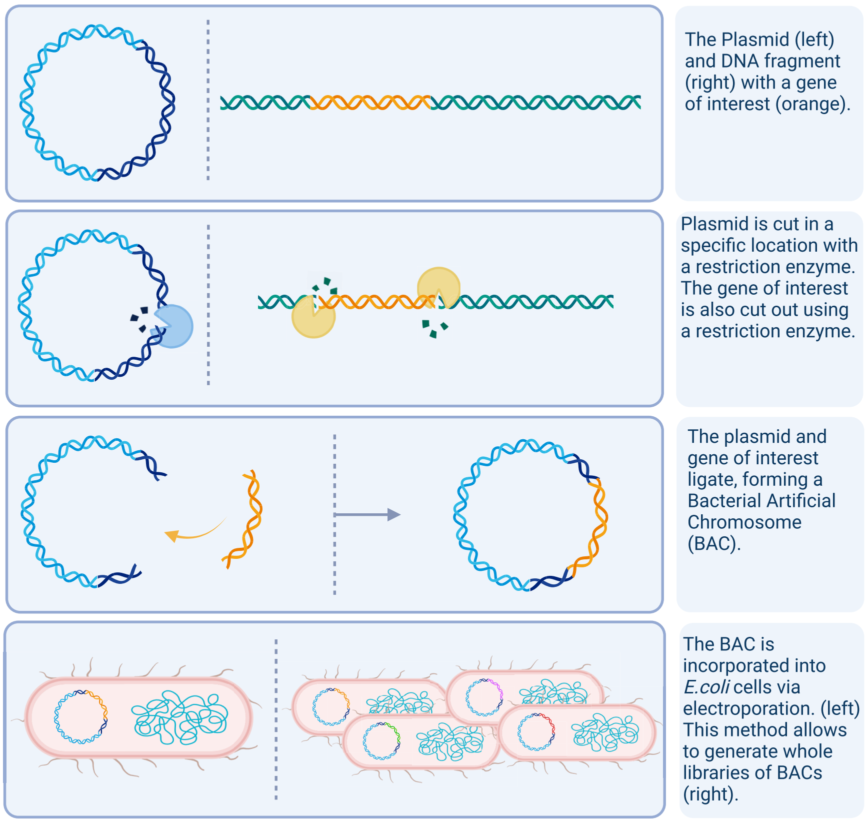
The Illumina system uses fluorescently tagged RTs where the fluorescent tag and linker are attached to the nucleobase, just like the case of fluorescently tagged ddNTPs for Sanger sequencing.
The Pacific Biosciences (SMRT) use phospholinked fluorescent tags, cleaved upon termination. Panel above - a dNTP carries a coumarine dye, panel below - a patented dNTP phospholinked to a fluorescent bead. The beads are dNTP species specific.
DNBSEQ (CoolMPS) uses 3' blocked RT dNTPs phospholinked to antigens. In the sequencing process the fluorophore conjugated antibodies will bind to the dNTP species specific antigen .

**Figure 4.** Structures of various dNTPs and ddNTPs used for various sequencing technology generations.

complementary information.[17,19] The labor-intensive plus−minus reaction scheme was later replaced by the chain-terminating Sanger sequencing, which became the method of choice before the rise of massively parallel sequencing.[19,20]

**Chain-Terminating Sanger Sequencing.** Chain-terminating Sanger sequencing was first showcased in 1975 and expanded in 1977.[17,20] It relies on the use of dideoxyribonucleotides (ddNTPs), which lack the 3′ OH group, disallowing phosphodiester bond formation, achieving chain termination by premature halting of the polymerase activity. Again, there are four reaction mixtures containing a primer, a template, the four dNTPs (one is always $^{32}$P labeled), and a single species of ddNTP.[20] After the reactions are complete, DNA fragments of varying sizes are generated thanks to random chain termination. The four samples are loaded on separate lanes on the polyacrylamide gel, similar to the Maxam−Gilbert and the plus-minus methods. Just like previous strategies, the sequence is inferred from the resulting ladder of fragments with known terminal ddNTP.[20]

A notable advancement was the replacement of radioactive labels with the newly developed fluorescent ddNTPs by DuPont.[21] Each of the ddNTP species carries a spectrally separated fluorophore (Figure 4, left); this example includes rhodamine fluorophores, while the original paper describes

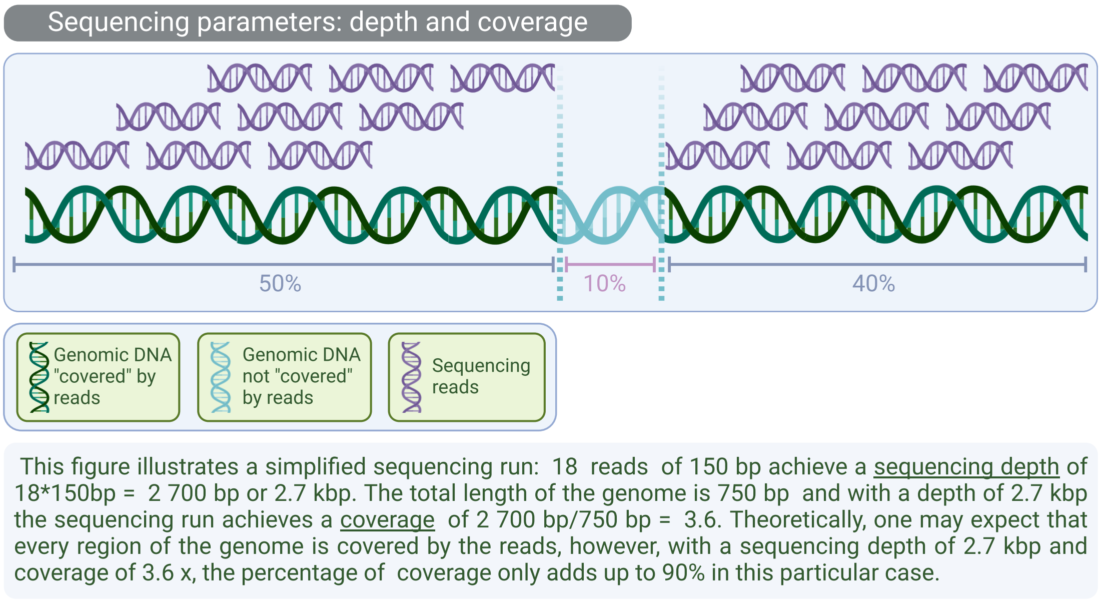

The Plasmid (left) and DNA fragment (right) with a gene of interest (orange).

Plasmid is cut in a specific location with a restriction enzyme. The gene of interest is also cut out using a restriction enzyme.

The plasmid and gene of interest ligate, forming a Bacterial Artificial Chromosome (BAC).

The BAC is incorporated into *E.coli* cells via electroporation. (left) This method allows to generate whole libraries of BACs (right).

**Figure 5.** Bacterial artificial chromosome generation process.


Sequencing parameters: depth and coverage
50%
10%
40%
Genomic DNA "covered" by reads
Genomic DNA not "covered" by reads
Sequencing reads

This figure illustrates a simplified sequencing run: 18 reads of 150 bp achieve a sequencing depth of 18*150bp = 2 700 bp or 2.7 kbp. The total length of the genome is 750 bp and with a depth of 2.7 kbp the sequencing run achieves a coverage of 2 700 bp/750 bp = 3.6. Theoretically, one may expect that every region of the genome is covered by the reads, however, with a sequencing depth of 2.7 kbp and coverage of 3.6 x, the percentage of coverage only adds up to 90% in this particular case.

**Figure 6.** Sequencing depth and coverage explained in an illustrated example. The missing 10% may look random in this example but in reality can be represented by regions that are prone to double-stranded breaks and fragmentation, homopolymeric regions or regions of high or low GC content impacting the local accuracy of enzymes like polymerase.[25,26] The missing 10% are shown as a single contiguous region for mere illustrative purposes as it is much more likely that problematic regions would be scattered across the genome. Even so, the cumulative amount of 10% in the example is very high for contemporary sequencing approaches.

succinylfluorescein dyes.[21,22] As a result, the sequencing reaction with chain termination could be performed in a single reaction with a single reaction product.[21] This in turn allowed use of capillary gel electrophoresis, where the reaction product, a mixture of DNA fragments of various lengths and different terminal ddNTPs, is loaded in a capillary tube instead of a gel slab (previous strategies) (Figure 3, right). The DNA fragments will descend the capillary tube in order of size and are imaged in that order, recording the florescent signal from the terminal ddNTP of each fragment.[21]

To this day, Sanger sequencing remains the gold standard of sequencing.[23,24] The limitations of the method dictate that typically Sanger sequencing is only used to check relatively short sequences of interest, measured in the tens of kilobases. The fragments of interest are amplified using polymerase chain reaction (PCR) and specially designed primers.[25] The PCR process will be discussed in more detail in the following sections. Originally, Sanger sequencing was commercialized by Applied Biosystems, currently available from Thermo Fisher. In fact, the principles of Sanger sequencing were used for the HGP. For this project, they used a DNA amplification system called bacterial artificial chromosome (BAC) (Figure 5) instead of the currently pervasive PCR systems.[14,26] BACs are plasmids employed as cloning vectors, which bacteria carry and copy upon multiplication. Genomic DNA of interest is fragmented, and the fragments ranging from 300 to 1000 kbp are ligated into BACs; in the case of the HGP, BACs carrying 100−500 kbp inserts were used.[26] The target bacteria (for example, *Escherichia coli*) are transfected with the BACs and cultured, thereby generating copies of the original genomic DNA fragments. Next, bacterial DNA is extracted and sequenced; the known bacterial genomic sequences are filtered out to generate assemblies of the original genomic DNA. This process is called hierarchical shotgun sequencing. BACs were seen as a way to circumvent issues with long repetitive or homopolymeric sequences.[14] Generally, contemporary sequencing methods do not employ cloning vectors for amplification.

**Sequencing Parameters.** The quality and capacity of a sequencer and sequencing runs are evaluated based on a set of parameters, such as run time, read length (a contiguous sequence "read" in one go in base pairs), accuracy, and error rate (typically assessed by mistaken base calls per every 100 000 bp). Other parameters include depth and coverage. Sometimes these terms are used interchangeably, referring to how many times the whole genome is sequenced in a sequencing run, though there are nuances. The depth of a sequencing run refers to the total read length obtained at the end of a sequencing run.[27] Meanwhile, the depth of coverage can be expressed using the formula $LN/G$, calculating to the ratio between the total number of bases read (total read length ($L$) × number of reads ($N$)) and the genome length (haploid genome length, $G$).[27] The depth of coverage can also be considered as coverage in terms of redundancy. There is also breadth of coverage or percentage of coverage, which refers to how much of the reference genome is recovered by the sequencing run at the given depth.[28] A simplified example to illustrate the terminology is given in Figure 6.

Chain termination with fluorescent ddNTPs brought the field of sequencing a step closer to automated DNA sequencing.[29,30] With an accuracy of 99.99% and read lengths between 400 and 1000 bp, Sanger sequencing in this format still has a relatively low throughput 84 kbp read per run of 3 h.[31] Due to the throughput limitation, analysis of mixed-genomic DNA samples would become extremely cumbersome and time consuming; each sequenced fragment has to fit in a particular location in the genome, which means fitting fragments of a maximum length at 1000 bp (kilobase pair, kbp) in, for example, a two million base pair (megabase pair, Mbp) bacterial genome. Given a sample that contains a genome or several genomes gigabase pairs (Gbp) in length, the task would become prohibitively complex. Hence, this method is mainly suited for pure samples, such as laboratory cultured bacteria (genomes typically in the Mbp range). To put some numbers on Sanger sequencing, in 1990 a project was launched to sequence the human genome, which is 3.2 billion bp.[15] The project took 13 years to complete, cost about $2.7 billion, and employed thousands of people. Currently, one may seek a commercial provider to sequence their own genome for less than a $1000 and expect a sequence within 1−2 months. The technological advancements described in the coming sections will reveal how this leap was possible.

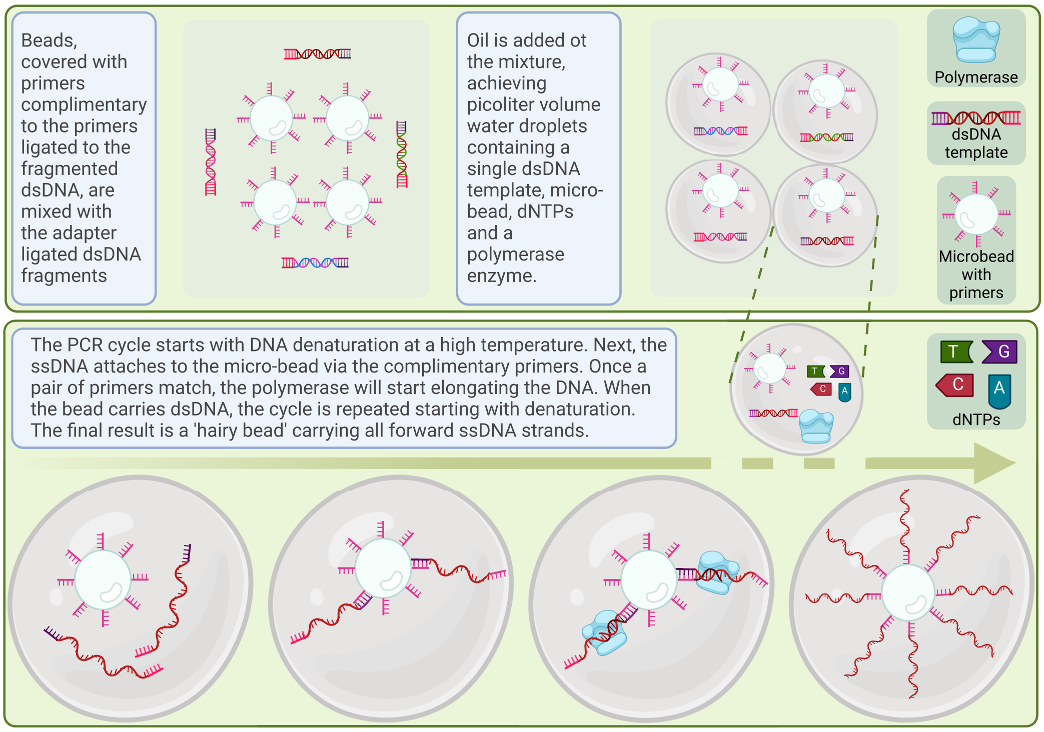
Generally, the first platforms that entered the market after Sanger sequencing were referred to as next-generation sequencing (NGS); these have effectively become second-generation sequencers (2$^{nd}$GS). Since the late 2000s, the literature also contained reports of third- and fourth-generation sequencing (3$^{rd}$GS and 4$^{th}$GS, respectively), referring to single-molecule approaches.[32,33]

**NGS Data Analysis.** The rise of NGS was concomitant with the rise of analysis algorithms designed to deal with the throughputs exceeding the outputs from the 1970s by orders of magnitude. A whole new field called bioinformatics erupted, starting with algorithms that compare sequences one element at a time with algorithms like Needleman−Wunsch.[34] From the moment that databases of known sequences were established, one may consider two types of algorithms: those that are used for assembling de novo and those that are used to align freshly sequenced material to the existing databases. For de novo sequencing, the earlier algorithms include Phrap and Newbler; a later example compatible with both 2$^{nd}$GS and 3$^{rd}$GS is SPAdes. For aligning, depending on the alignment specification (pairwise, local, global, ...), different algorithms are available; the best known is the Basic Local Alignment Tool (BLAST), but there are others such as Clustal, T-Coffee, and more. There are a number of parameters to consider and adjust for every approach (error rates, similarity thresholds, scoring of gaps in alignments, ...). One should be aware that any assembly and alignment is the result of a choice of sequencing technology, algorithm, and a number of parameters, which influence the final readout and subsequent conclusions about a sequence. Details on the bioinformatics pipelines exceed the scope of this review and are available elsewhere.[35]

## Second-Generation Sequencing

For 2$^{nd}$GS, the keywords are massively parallel sequencing, enabled by a paradigm called cyclic array sequencing.[36] Typically, 2$^{nd}$GS proceeds via four major steps: (1) collection and isolation of genomic material, (2) enrichment of the sequence(s) of interest, (3) sequencing, and (4) bioinformatic analysis of the data acquired.[37] Steps 1−3 will be discussed in detail in the coming paragraphs, while the methods for step 4 have been reviewed at large elsewhere.[35,38]

In step 1, the genomic material of DNA is extracted and isolated from a source, for example, a human cancer cell sample. Next, the extracted DNA, typically few kilobase pairs in length, is fragmented into short molecules between 100 and 500 bp, depending on the method used for fragmentation (i.e., vortexing, enzymatic digestion) and the requirements of the downstream sequencing platform.[37,39−41] DNA isolation methods range from more "old-school" approaches using phenol−chloroform−isoamyl alcohol to any number of commercially available kits. The most suitable methods differ depending on the source material, ranging from pure bacterial cultures, mammalian cell lines, and biopsies to complex environmental samples, such as stool and soil. In practice, local adaptations by the user are often required to achieve the


Beads, covered with primers complimentary to the primers ligated to the fragmented dsDNA, are mixed with the adapter ligated dsDNA fragments

Oil is added ot the mixture, achieving picoliter volume water droplets containing a single dsDNA template, micro-bead, dNTPs and a polymerase enzyme.

The PCR cycle starts with DNA denaturation at a high temperature. Next, the ssDNA attaches to the micro-bead via the complimentary primers. Once a pair of primers match, the polymerase will start elongating the DNA. When the bead carries dsDNA, the cycle is repeated starting with denaturation. The final result is a 'hairy bead' carrying all forward ssDNA strands.

**Figure 7.** Emulsion PCR workflow.

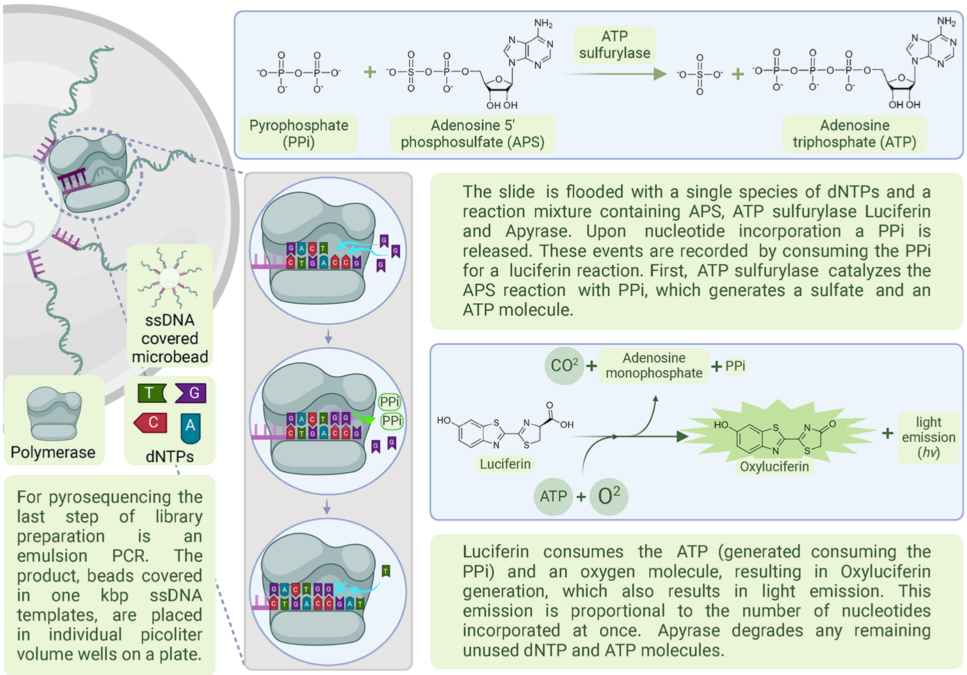
desirable purity and/or length of the extracted genomic material.

In step 2, the enrichment of the target sequence(s) is achieved using clonal amplification.[37] Most commonly used PCR strategies include solid-phase emulsion PCR and solid-phase bridge PCR.[42,43] The goal of clonal amplification is to generate thousands of copies of these short fragments, which would in subsequent sequencing steps provide a higher confidence due to more reads covering the same genomic location. In step 2, the individual DNA fragments from step 1 are processed further by ligating (attaching to the existing DNA in a seamless manner) special adapters (double-stranded DNA fragments) to both ends of each fragment.[37] The processed DNA fragments are referred to as the library of the sequencing run, and consequently, step 2 is the library preparation step. The ligated adapters are complementary to the primers used in the PCR-based amplification.[44] First, one needs to design primers that match the beginning of the region of interest in the DNA fragment or that will match universally.[44,45] The fragmented DNA is incubated with the PCR mix: the primers, a heat-tolerant DNA polymerase, and dNTPs. The mixture is subjected to a series of heating and cooling cycles to induce denaturation, primer annealing, and polymerase elongation of the DNA starting from the primer-annealed sites.[45]

Development of clonal amplification enabled sequencing of millions of sequence fragments in parallel (hence, massively parallel sequencing) with an increase in total output length ranging from Gbp to terabase pairs (Tbp), opening a new world for genetic research in medical, industrial, and, of course, academic settings.[31,46] Sequencing runs from sample prep to preanalyzed readout are automated and take hours to 2−3 days. These types of developments have also allowed for sequencing of mixed samples that contain many different genomes, enabling metagenomic research (genomes of communities of organisms, for example, microbes from soils, stool, water reservoirs).

**Emulsion PCR-Based Sequencing Platforms.** In emulsion PCR (ePCR), the adapters ligated to the ends of each fragment in the library are complementary to the primers attached to a magnetic microbead.[47] The emulsion is obtained by mixing oil and water, where the goal is for every water droplet in oil to contain a single library item from step 2, a single magnetic bead, and a PCR mix comprised of primers complementary to the adapter not attached to the bead, polymerase enzyme, and dNTPs.[47] The ePCR droplets function as compartmentalized microreactors, where no exchange of contents or even contact between two such systems takes place. The PCR reaction is typically repeated between 30 and 60 cycles (this number is highly subject to variation).[48,49] The reaction output is a bead carrying thousands of single-strand copies of the original DNA template, referred to as polymerase colonies (polonies). The ePCR process is illustrated in Figure 7. The downstream processing of the beads depends on the NGS platform as will be illustrated in the following sections; generally, the emulsion is broken, and the beads are deposited in individual chambers on a picotiter plate.[50,51] Ultimately, achieving a single ePCR mix along with a single library template per droplet is not trivial; the intricacies thereof are beyond the scope of this paper.[52]

**Sequencing by Synthesis.** Two examples of SBS sequencing platforms are Roche 454 and Thermo Fisher's Ion Torrent.[53,54] Roche 454 is a pyrosequencing platform, currently discontinued.[55] Here, after the ePCR step, the beads are deposited in individual wells on a picotiter plate. The wells are flooded with a reaction mixture containing a single species of nucleotide. It must be noted that in nature upon a nucleotide incorporation event, a pyrophosphate ($PP_i$) and a hydrogen ion ($H^+$) are released. The Roche 454 approach

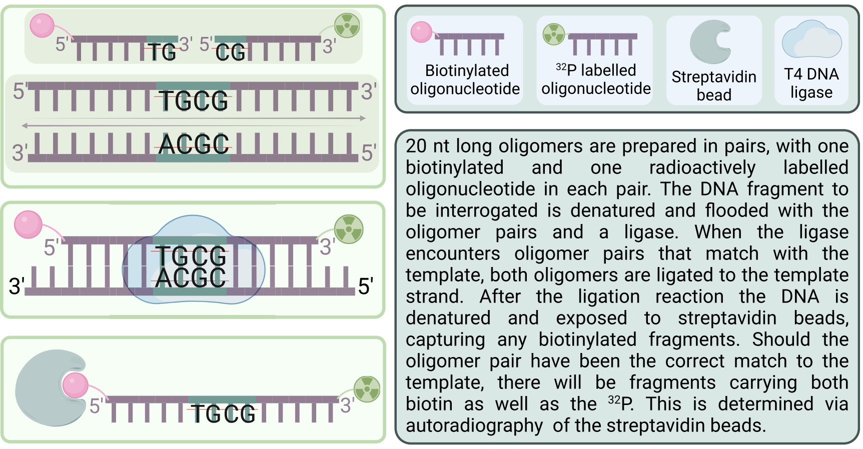


For pyrosequencing the last step of library preparation is an emulsion PCR. The product, beads covered in one kbp ssDNA templates, are placed in individual picoliter volume wells on a plate.

The slide is flooded with a single species of dNTPs and a reaction mixture containing APS, ATP sulfurylase Luciferin and Apyrase. Upon nucleotide incorporation a PPi is released. These events are recorded by consuming the PPi for a luciferin reaction. First, ATP sulfurylase catalyzes the APS reaction with PPi, which generates a sulfate and an ATP molecule.

Luciferin consumes the ATP (generated consuming the PPi) and an oxygen molecule, resulting in Oxyluciferin generation, which also results in light emission. This emission is proportional to the number of nucleotides incorporated at once. Apyrase degrades any remaining unused dNTP and ATP molecules.

**Figure 8.** Roche 454 sequencing workflow.



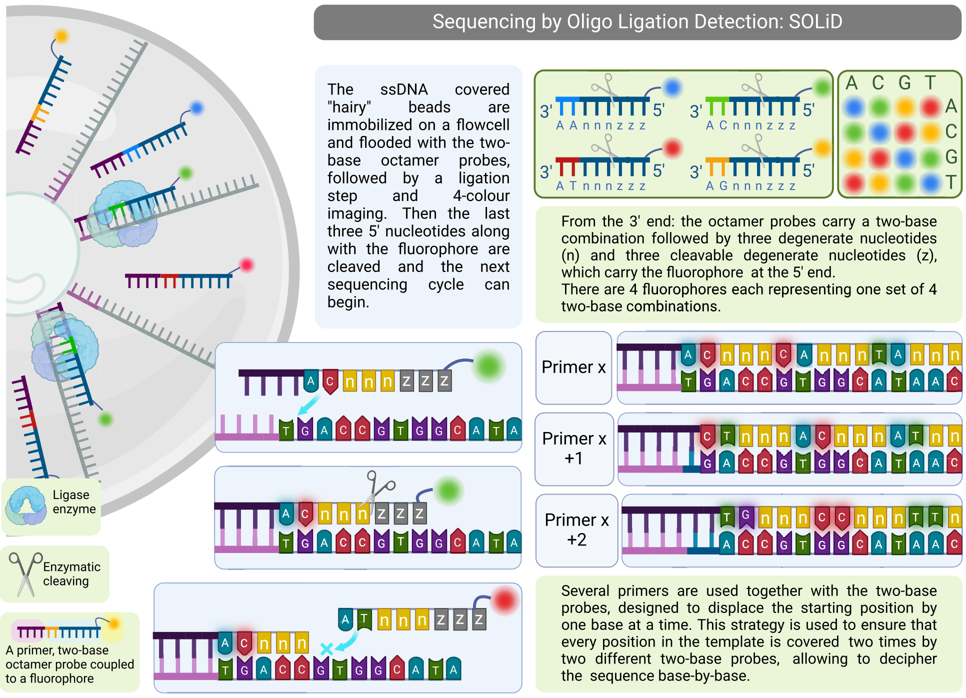
20 nt long oligomers are prepared in pairs, with one biotinylated and one radioactively labelled oligonucleotide in each pair. The DNA fragment to be interrogated is denatured and flooded with the oligomer pairs and a ligase. When the ligase encounters oligomer pairs that match with the template, both oligomers are ligated to the template strand. After the ligation reaction the DNA is denatured and exposed to streptavidin beads, capturing any biotinylated fragments. Should the oligomer pair have been the correct match to the template, there will be fragments carrying both biotin as well as the $^{32}$P. This is determined via autoradiography of the streptavidin beads.

**Figure 9.** An early approach to the Sequencing by Oligo Ligation Detection method.

records the release of the $PP_i$, and the reaction mixture contains a polymerase, adenosine triphosphate (ATP)−sulfurylase, adenosine 5′-phosphosulfate (APS), adenosine diphosphatase (apyrase), the luciferase enzyme, and luciferin.[56] Upon nucleotide incorporation, the ATP−sulfurylase and APS convert the $PP_i$ to ATP. Subsequently, luciferase catalyzes the conversion of luciferin to oxyluciferin by consuming the newly generated ATP, which emits visible light. Before the next cycle, apyrase degrades the remaining ATP and dNTPs.[56] The workflow of Roche 454 can be found in Figure 8. In the case of homopolymeric regions, the signal strength is proportional to the number of nucleotides incorporated. During a pyrosequencing routine, the four species of dNTPs are added sequentially and the sequence for each individual polymerase colony can be read by combining all signals per well on the picotiter plate.[56,57] The Roche 454 read length is ∼400 bp.

The Ion Torrent technology records the release of a proton (hydrogen ion, $H^+$) as the signal marking the event of nucleotide addition.[54] The hairy beads are deposited in wells on a semiconductor plate, and each well containing its single bead is flooded with a single species of dNTPs every 15 s. The well plate is effectively a semiconductor chip functioning as an ultrasensitive pH meter, detecting the change in pH as a $H^+$ is released.[54] In the case of homopolymeric regions, the signal is proportional to the amount of nucleotides incorporated.[54]

**Sequencing by Ligation.** Sequencing by Oligo Ligation Detection (SOLiD) relies on a strategy referred to as sequencing by ligation (SBL).[58,59] In its inception, the method relied on DNA ligase sensitivity for mismatches between a template strand and a candidate oligomer (Figure 9).[60,61] As a proof of concept, the sequence of a gene was interrogated for a single nucleotide polymorphism (SNP) using pairs of synthetic oligomers 20 nucleotides in length, where one of the oligomers is biotinylated and the other is radioactively labeled with a $^{32}$P.[60] When the oligomer pair is complementary to the sequence being interrogated, the ligase will ligate the two oligomers to the ssDNA of the template DNA. After another denaturation step, the reaction product is used for a

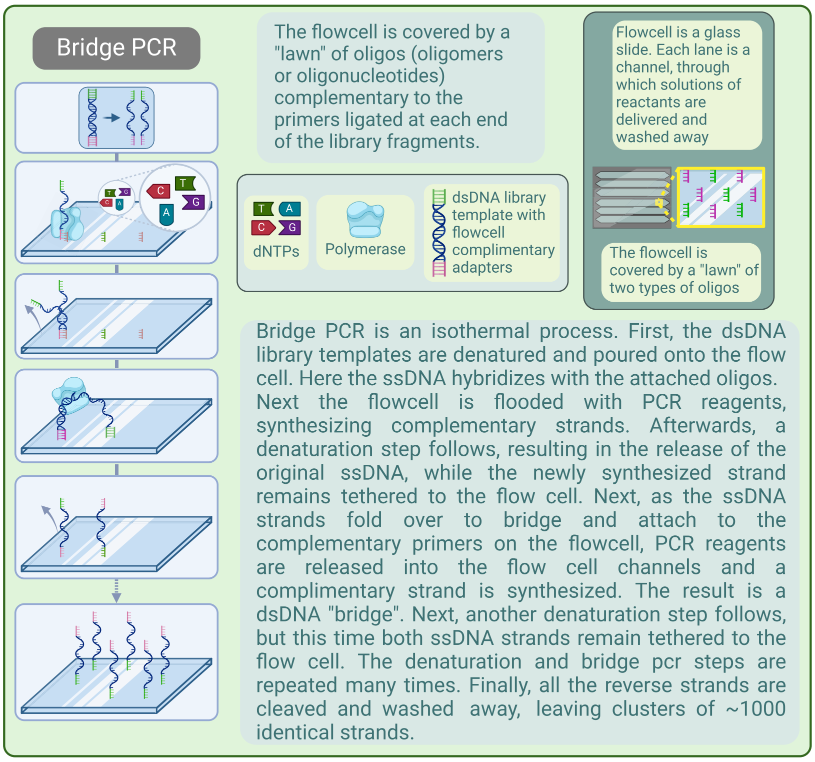


The ssDNA covered "hairy" beads are immobilized on a flowcell and flooded with the two-base octamer probes, followed by a ligation step and 4-colour imaging. Then the last three 5' nucleotides along with the fluorophore are cleaved and the next sequencing cycle can begin.

From the 3' end: the octamer probes carry a two-base combination followed by three degenerate nucleotides (n) and three cleavable degenerate nucleotides (z), which carry the fluorophore at the 5' end.
There are 4 fluorophores each representing one set of 4 two-base combinations.

Several primers are used together with the two-base probes, designed to displace the starting position by one base at a time. This strategy is used to ensure that every position in the template is covered two times by two different two-base probes, allowing to decipher the sequence base-by-base.

**Figure 10.** Sequencing by Oligo Ligation Detection (SOLiD) workflow.



The flowcell is covered by a "lawn" of oligos (oligomers or oligonucleotides) complementary to the primers ligated at each end of the library fragments.

Flowcell is a glass slide. Each lane is a channel, through which solutions of reactants are delivered and washed away

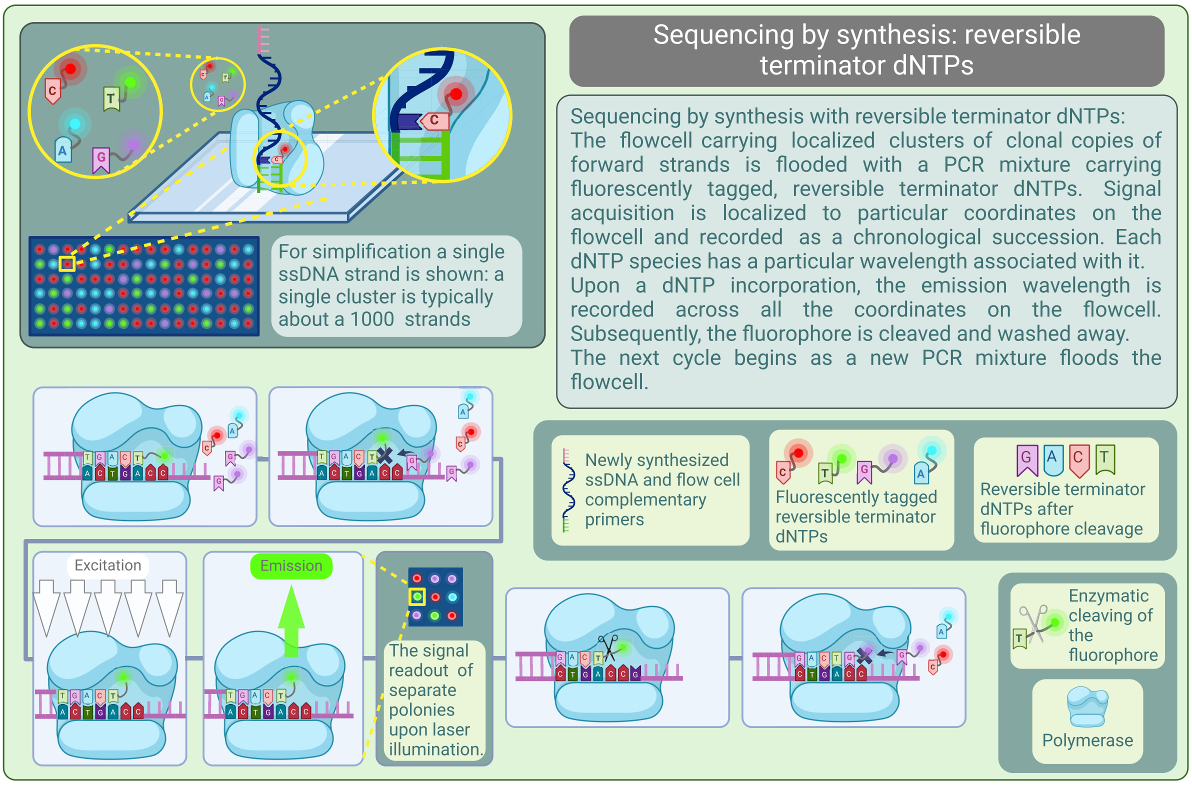
The flowcell is covered by a "lawn" of two types of oligos

Bridge PCR is an isothermal process. First, the dsDNA library templates are denatured and poured onto the flow cell. Here the ssDNA hybridizes with the attached oligos. Next the flowcell is flooded with PCR reagents, synthesizing complementary strands. Afterwards, a denaturation step follows, resulting in the release of the original ssDNA, while the newly synthesized strand remains tethered to the flow cell. Next, as the ssDNA strands fold over to bridge and attach to the complementary primers on the flowcell, PCR reagents are released into the flow cell channels and a complimentary strand is synthesized. The result is a dsDNA "bridge". Next, another denaturation step follows, but this time both ssDNA strands remain tethered to the flow cell. The denaturation and bridge pcr steps are repeated many times. Finally, all the reverse strands are cleaved and washed away, leaving clusters of ~1000 identical strands.

**Figure 11.** Solid-state (bridge) PCR workflow.

streptavidin bead assay, where only biotinylated DNA oligomers will be captured.[60] Next, the product of the streptavidin assay is exposed to an X-ray film. Thereby, it is confirmed whether there are any fragments carrying both biotin and streptavidin and, by extension, whether the original template matched with the oligomer pair.[60] Just like with Sanger sequencing, this method eventually arrived at the use of fluorescently labeled dNTPs.

Currently, the commercialized method, SOLiD, consists of generating sets of octamers and "fitting" them on a denatured template strand (Figure 10).[62,63] Each set of octamers contains 1 out of 16 possible 3′ two-base combinations (AT, AC, AG, AA, ...) followed by 6 degenerate nucleotides and a fluorescent


Sequencing by synthesis: reversible terminator dNTPs
C
T
A
G
Excitation
Emission
Newly synthesized ssDNA and flow cell complementary primers
Fluorescently tagged reversible terminator dNTPs
Reversible terminator dNTPs after fluorophore cleavage
Enzymatic cleaving of the fluorophore
Polymerase

Sequencing by synthesis with reversible terminator dNTPs: The flowcell carrying localized clusters of clonal copies of forward strands is flooded with a PCR mixture carrying fluorescently tagged, reversible terminator dNTPs. Signal acquisition is localized to particular coordinates on the flowcell and recorded as a chronological succession. Each dNTP species has a particular wavelength associated with it. Upon a dNTP incorporation, the emission wavelength is recorded across all the coordinates on the flowcell. Subsequently, the fluorophore is cleaved and washed away. The next cycle begins as a new PCR mixture floods the flowcell.

For simplification a single ssDNA strand is shown: a single cluster is typically about a 1000 strands

The signal readout of separate polonies upon laser illumination.

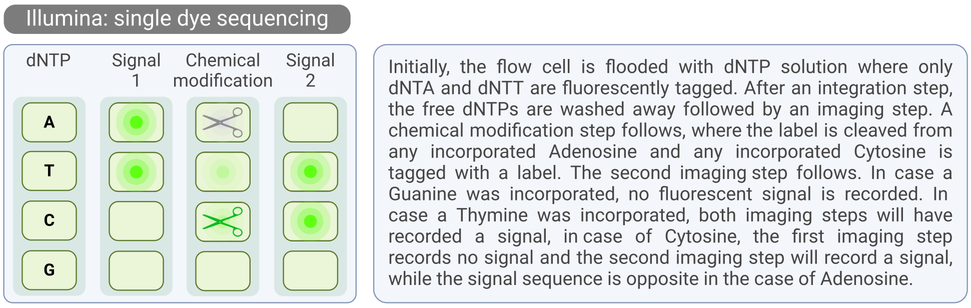
**Figure 12.** Sequencing by synthesis: Illumina workflow following the bridge PCR step.

label at the 5′ end. Each fluorescent label matches four two-base combinations. A variety of nonoverlapping sets of four fluorescent dyes/beads can be used.[61] A primer is used as a starting point for two-base probe hybridization. In each cycle, after the ligation step, any remaining free-floating octamers are washed away and the plate is imaged, thereby recording the fluorescent signal. Next, three degenerate nucleotides along with the fluorophores are cleaved off from the 5′ end of the hybridized oligomer.[61] The cycles are repeated until no more two-base probes can be incorporated, signaling the end of the round. The number of cycles depends on the read length. In the next round, a new primer with an offset of a single base is used and the procedure is repeated.[61] Typically, five rounds are used, after which the sequence can be deduced from the obtained color codes. Note that every base is interrogated twice, which leads to higher base-calling accuracy.[62] The ssDNA primers used at the start of each round are complementary to the adaptor sequences used for emulsion PCR clonal amplification. The read length is 50 + 35 bp.[62,64]

**Bridge Amplification.** *Sequencing by Synthesis.* Solid-phase bridge amplification, or bridge PCR, is an isothermal amplification reaction.[65] The final step of the library preparation for this workflow consists of ligating adapters on both ends of each fragment; these adapters are complementary to oligonucleotide primers covering the flow cell on which the bridge PCR reaction will be carried out.[65] The process starts by denaturation of the library templates into ssDNA, which then hybridize to the lawn of oligonucleotides on the flow cell (Figure 11). Next, the flow cell is flooded with PCR reagents, resulting in synthesis of the complementary strand to the flow cell-attached ssDNA.[65] The synthesis or elongation step is followed by denaturation and a wash step, resulting in only the newly built complementary strand remaining attached to the flow cell. This strand then bridges and hybridizes to a nearby oligonucleotide, complementary to its free adapter.[65] Another PCR cycle is initiated, resulting in a dsDNA bridge attached to the flow cell at each end. Another denaturation step follows, where the bridge formation separates into two flow cell-attached strands.[65] At this point, the reaction is complete and the next PCR cycle can start. After multiple PCR cycles, the ssDNA strands of both the forward and the reverse strand of the original dsDNA template form a localized cluster: a polymerase colony, also called polony. Once the PCR cycling is complete, the reverse strands are cleaved and washed away. Polony generation occurs simultaneously in localized clusters for all library fragments on the flow cell.[65]

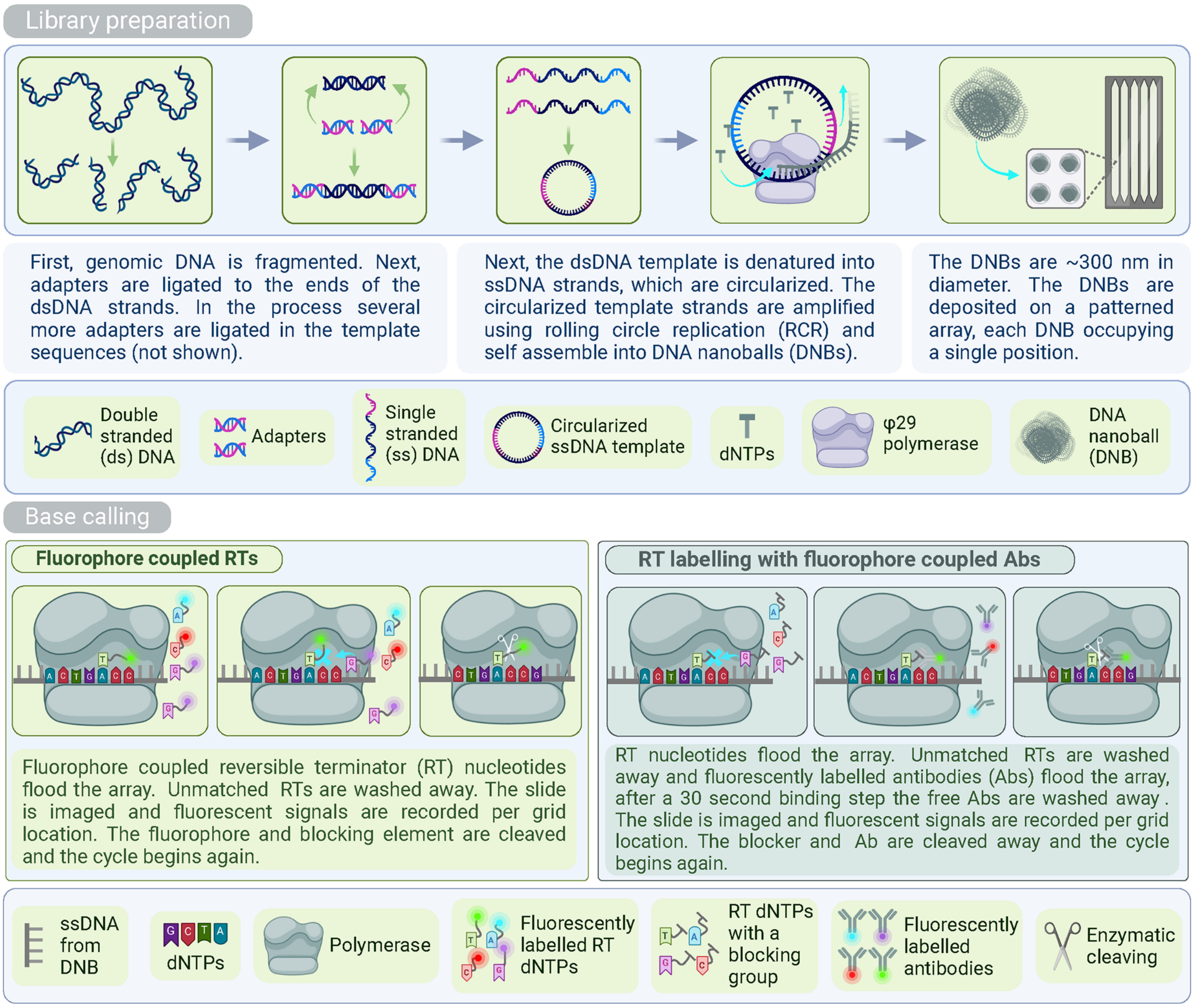
To briefly compare the two main PCR methods, solid-phase PCR and emulsion PCR, in general, ePCR is more efficient and allows for a much higher number of template copies. The downside of ePCR is that any error introduced in the early template copies will be amplified in all subsequent copies. In the case of bridge PCR, the templates may sometimes hybridize between themselves instead of primer on the flow cell, and the length of the DNA fragments that may be amplified on the flow cell used to be around 35 bp with current reads at 50, 150, and even 300 bp.[63,66]

Similarly to Sanger sequencing, Illumina sequencing technology uses an SBS strategy with chain-terminating nucleotides. However, the Illumina chain termination is reversible. Custom dNTPs contain a fluorescent group attached to the nucleobase and a blocking group at the 3′ position, where cleavage of the blocking group (along with the fluorophore) renders the 3′ end accessible to the polymerase.[67,68] The principles of reversible terminators as a category exceed the scope of this review.[67] In brief, the dNTP is 3′-*O*-azidomethyl blocked, meaning it carries an $N_3$, blocking its 3′



Initially, the flow cell is flooded with dNTP solution where only dNTA and dNTT are fluorescently tagged. After an integration step, the free dNTPs are washed away followed by an imaging step. A chemical modification step follows, where the label is cleaved from any incorporated Adenosine and any incorporated Cytosine is tagged with a label. The second imaging step follows. In case a Guanine was incorporated, no fluorescent signal is recorded. In case a Thymine was incorporated, both imaging steps will have recorded a signal, in case of Cytosine, the first imaging step records no signal and the second imaging step will record a signal, while the signal sequence is opposite in the case of Adenosine.

**Figure 13.** Illumina single-dye sequencing modification workflow.



First, genomic DNA is fragmented. Next, adapters are ligated to the ends of the dsDNA strands. In the process several more adapters are ligated in the template sequences (not shown).

Next, the dsDNA template is denatured into ssDNA strands, which are circularized. The circularized template strands are amplified using rolling circle replication (RCR) and self assemble into DNA nanoballs (DNBs).

The DNBs are ~300 nm in diameter. The DNBs are deposited on a patterned array, each DNB occupying a single position.

Fluorophore coupled reversible terminator (RT) nucleotides flood the array. Unmatched RTs are washed away. The slide is imaged and fluorescent signals are recorded per grid location. The fluorophore and blocking element are cleaved and the cycle begins again.

RT nucleotides flood the array. Unmatched RTs are washed away and fluorescently labelled antibodies (Abs) flood the array, after a 30 second binding step the free Abs are washed away . The slide is imaged and fluorescent signals are recorded per grid location. The blocker and Ab are cleaved away and the cycle begins again.

**Figure 14.** DNA nanoball (DNB) sequencing workflow: library preparation (top), base calling in DNBseq (lower left), and CoolMPS workflow (lower right).

OH group (Figure 4).[68,69] The nucleobase of the dNTP is conjugated to the fluorophore by a cleavable linker, and after the dNTP incorporation and imaging steps, the 3′ OH moiety is restored and the fluorophore is cleaved, leaving behind a slightly modified part of the original linker.[68,69] This is called a molecular scar, and it does not impede further nucleotide incorporation.[68]

The SBS is achieved in the following way (Figure 12). Initially, a solution containing four dNTPs coupled to four spectrally distinct fluorophores is added to the flow cell, which is already carrying the polonies of forward strands of the original library templates.[68] The sequencing cycle starts with a polymerase-mediated dNTP incorporation followed by washing away of unbound dNTPs. Next, the flow cell is imaged to capture fluorescent signals at four different wavelengths. Once the fluorescent group and blocking moiety are removed and washed away, the sequencing cycle can start over.[68] The readout is the temporal sequence of fluorescent signals on each grid location. The sequencing readout is the camera readout, a succession of nucleobase-signifying color signals, effectively yielding the sequence for each polony (i.e., library fragment). While the principle of Solexa/Illumina sequencing was first introduced in 1998, the localized clusters of bridge PCR became part of the workflow in 2004, bought from Manteia.[70]

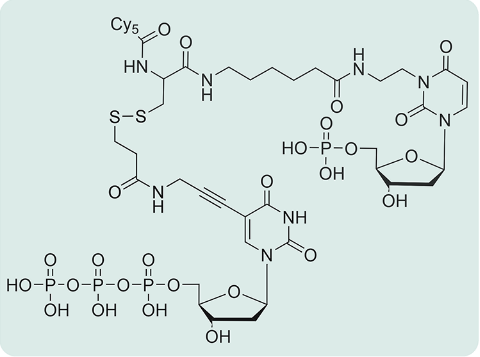
In later iterations of Illumina sequencing technology, two dyes and a single dye at a time are used, respectively. In the first case, two dNTPs are labeled with two different fluorophores, the third is labeled with both fluorophores, while the fourth dNTP carries no label.[71] The dNTP incorporation events are recorded with two imaging steps, one at each excitation wavelength, overall reducing the time and reagent cost.[71] In the case of a single dye, Illumina uses a flow cell of a complementary metal−oxide semiconductor (CMOS), where three dNTPs carry a single dye (Figure 13).[72] There are two imaging steps flanking a chemical modification step, which represents two types of chemical modifications for adenosine and cytosine and no modifications to thymine and guanine.[72] Further details of these reactions are proprietary.

**DNA Nanoball Sequencing.** Another big SBS player is DNA nanoball sequencing, showcasing a PCR-free library preparation. This is an MGI sequencing technology (subsidiary of Beijing Genomics Institute, BGI), emerging to the global market as recently as the late 2010s.

DNA nanoball sequencing, available commercially from the Chinese company MGI Tech (MGI), is also an SBS technology (Figure 14). The library prep, once again, starts with fragmentation followed by adapter ligation to the fragment ends and denaturation thereof.[73] The denatured ssDNA fragments circularize thanks to the adapters. A primer oligohybridizes with the adapters, generating a launch pad for the polymerase enzyme, and the $\varphi$29 polymerase is added.[73,74] As the polymerase enzyme generates a complementary strand, it immediately denatures, an example of rolling circle replication, a similar concept to circular consensus sequencing.[73,74] In this case, however, the denatured newly synthesized strand is collected into a structure resembling a "ball of yarn" where the "yarn" is the ssDNA consisting of tandem repeats.[73] This structure is referred to as a DNA nanoball (DNB). The fragment of interest has thereby been amplified, circumventing a PCR step. Together this whole process is referred to as combinatorial probe−anchor synthesis (cPAS).[74] The DNBs are ∼220−240 nm in diameter and are deposited on a patterned array flow cell.[73,74] The array consist of 300 nm wide seats called spots with one DNB per seat, spaced ∼500 nm apart. The seats are aminated to achieve a positive charge and attract the negatively charged DNB, while the rest of the flow cell is treated with the organosilicon bis(trimethylsilyl)amine (HDMS) to achieve hydrophobicity, disallowing any free DNBs to attach in between the designated seats.[73,74] At this step, phospholinked dNTPs and DNA polymerases are added to commence synthesis of the complementary strand (Figure 4 for dNTP structure). Every dNTP incorporation event will generate a signal as the dye-carrying moiety is naturally cleaved upon incorporation; the signal is recorded from above the chip.[73,74] The resulting string of signals associated with a particular location on the flow cell is interpreted in a similar fashion to the Illumina approach.

In the meantime, MGI has come forth with another variation of this sequencing method (Figure 14, lower right). The cPAS process remains as previously described, while the dNTPs used are nonfluorescently labeled and carry a 3′-*O*-azidomethyl blocking group (Figure 4).[75] Next, the sample is flooded with four nucleobase-specific antibodies conjugated with spectrally nonoverlapping fluorescent dyes.[75] Then, the unattached antibodies are washed away, and the sample is imaged. Finally, the antibody and the azidomethyl group are cleaved and washed away. According to Drmanac and colleagues, there is no molecular scarring upon removal of the blocking group.[75]

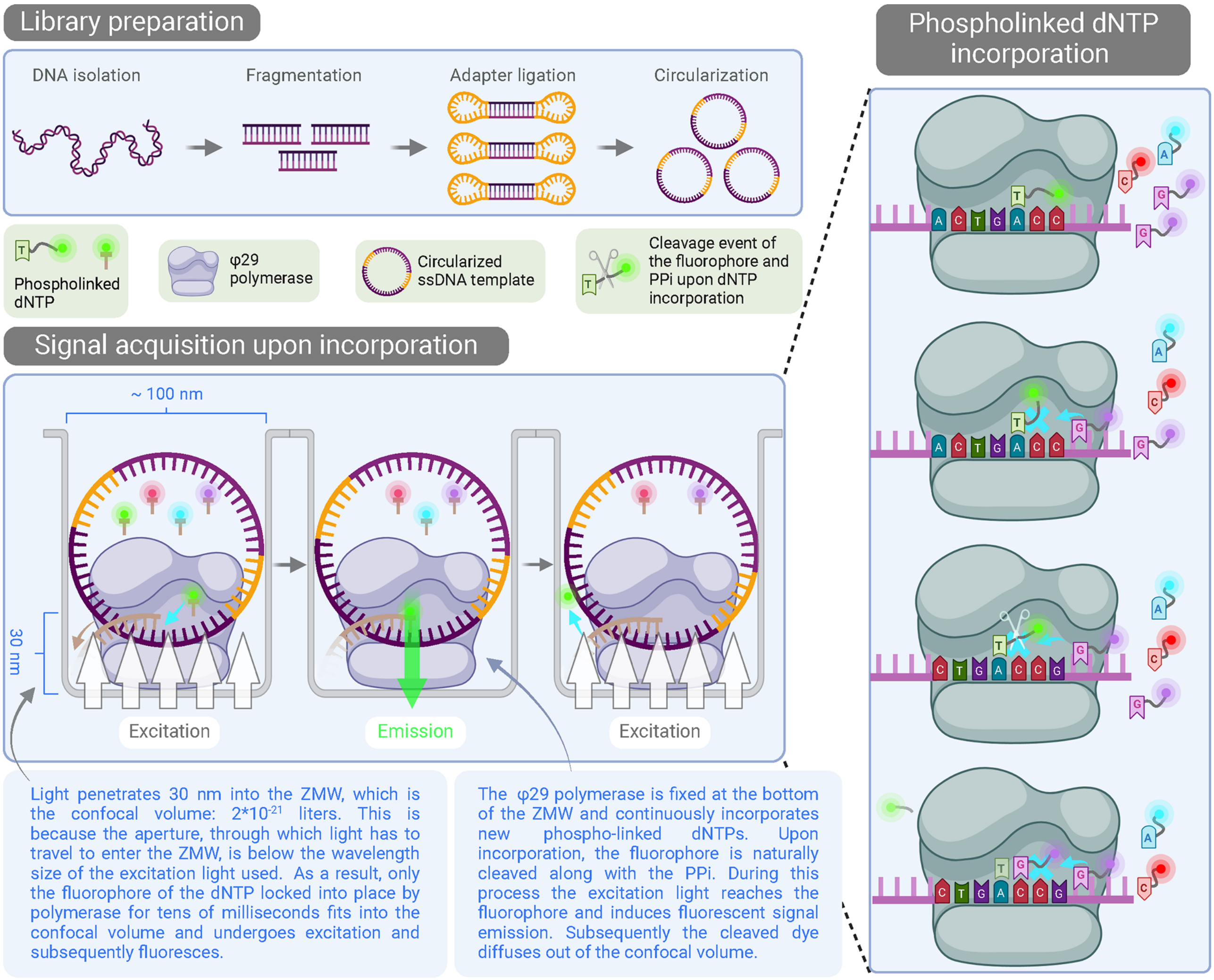
**Single-Molecule Sequencing: HeliScope.** It is worth mentioning that among the second-generation sequencers there is a discontinued single-molecule, PCR-amplification-free approach. The technology was spear headed by Stephen Quake in a 2009 publication, sequencing his own genome for \$50 000 at a time when the cost of sequencing a human genome was estimated at \$250 000−500 000 or even higher.[76,77] The HeliScope from Helicos Biosciences is an SBS machine processing single molecules of DNA, directly accepting genomic DNA molecules with lengths ranging from a few nucleotides to 100−200 nucleotides (Figure 15).[78]



An example of Helicos Biosciences Virtual Terminator dNTP: a VT-dNT. The blocking moiety (top right) is linked to a fluorophore, Cy5. This construct links to the nucleobase via a disulfide bond, which is also the cleavage site. Effectively, VT-dNTPs are another type of reversible terminators.

**Figure 15.** Reversible terminators used by Helicos Biosciences.

The process starts with preparing the DNA library; if the genomic DNA in question consists of fragments > 1 kb pair in size, the first step is sonication. Next, the DNA is denatured and incubated with an excess of dATPs and terminal transferases to generate polyadenine tails that stretch 90−200 nucleotides in length, which are eventually blocked using a ddNTP. Next, the library of ssDNA is loaded onto a glass flow cell, which is functionalized with single-stranded polythymine oligomers about 50 nucleotides in length, followed by a hybridization step. The polyadenine-carrying DNA molecules are hybridized to the flow cell-attached oligomers followed by addition of dNTTs, virtual terminator (VT) dNTPs for A, C, and G, along with a polymerase. The addition of dNTTs is necessary since the polyadenine tail typically extends longer than the polythymine oligomer. Once the polymerase encounters the first non-A residue, a VT dNTP is incorporated, blocking further polymerization reaction. The blocking is achieved due to the VT-dNTP structure, where a Cy5 fluorophore and a blocking moiety are connected to the dNTP by a disulfide bond at the nucleobase moiety.[78,79] The first imaging step merely dictates which positions on the flow cell carry attached molecules since they can be any of the three VT-dNTP species. The disulfide linkage of the VT-dNTPs is cleaved, and the fluorophores are washed away; the sequencing can begin. All of the VT-dNTP species carry the same fluorophore, so sequencing proceeds in cycles with single VT-



Light penetrates 30 nm into the ZMW, which is the confocal volume: $2*10^{-21}$ liters. This is because the aperture, through which light has to travel to enter the ZMW, is below the wavelength size of the excitation light used. As a result, only the fluorophore of the dNTP locked into place by polymerase for tens of milliseconds fits into the confocal volume and undergoes excitation and subsequently fluoresces.

The φ29 polymerase is fixed at the bottom of the ZMW and continuously incorporates new phospho-linked dNTPs. Upon incorporation, the fluorophore is naturally cleaved along with the PPi. During this process the excitation light reaches the fluorophore and induces fluorescent signal emission. Subsequently the cleaved dye diffuses out of the confocal volume.

**Figure 16.** Single-molecule real-time (SMRT) sequencing workflow.

dNTP species flooding the flow cell at a time, followed by an imaging, cleaving, and washing steps.

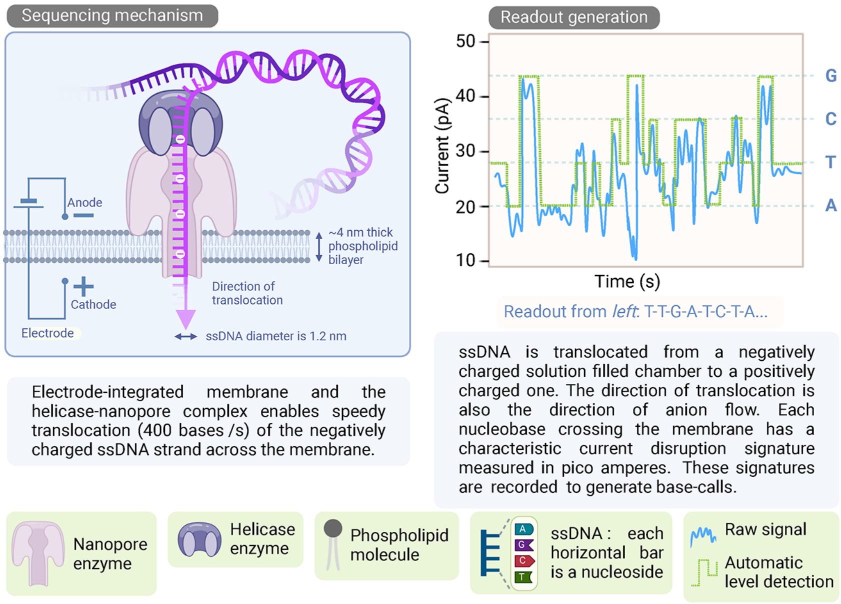
Helicos Biosciences went out of business in 2012; one of the reasons may have been the high price of the instrument, which sold for \$1.35 million in 2007 when the next cheapest machine was SOLiD, which sold for \$600 000.[77] Nevertheless, the strategy is notable, as will become clear in the next sections, avoiding amplification bias, being able to process long DNA molecules, and directly processing DNA molecules of a highly varied length, which are all advantageous attributes of a sequencing method.

## Challenges in NGS

While exceptionally powerful and often surprisingly error free, the NGS technologies do face some issues. For example, NGS platforms rely on the use of short reads, ranging between 75 and 400 bp, conducive to high read quality. The read lengths are limited in size to reduce polymerase-related errors and out-of-phase elongation within a single polony, since both of these problems are cumulative.[80,81] Consequently, the longer the read length, the higher the likelihood of read errors. The latter effect, called polony dephasing, arises from loss of synchronization during the sequencing reaction, for example, a skipped nucleotide at different locations in different molecules of the polony.[65] The result is increased discrepancy between base calls toward the end of the read within the same cluster, effectively adding to the overall noise for that polony.[80,81]

The real limitation of short reads becomes apparent upon encountering repetitive regions or structural variants that exceed the length of the longest read such as inversions or translocations. In such cases, either a de novo approach becomes necessary to decipher the fragments or, alternatively, long read sequencing methods can be used.[82,83] For short read sequencers, there is a mitigation strategy referred to as "paired ends" or "mate pairs" when applied to shorter fragments (200−800 bp) or longer fragments (2−5 kbp), respectively.[84,85] For the paired end approach, the library fragments are subjected to adaptor ligation at each end. For the mate pair approach, the ends are first conjugated with biotin, which contributes to the formation of circular DNA. The noncircularized DNA is removed, and the circularized DNA is fragmented once more (400−600 bp). The biotin-containing fragments are affinity purified using the biotin tag, and, finally, the flow cell complementary adaptors are ligated to both ends of the fragment. These fragments are now the library for sequencing. The key aspect is that the biotinylated fragments contain the ends of the original long fragments.

The paired end approach is clever but not all powerful.[86,87] In addition, the PCR-based clonal amplification step is prone to introducing library bias in very GC-rich or GC-poor regions,



Electrode-integrated membrane and the helicase-nanopore complex enables speedy translocation (400 bases /s) of the negatively charged ssDNA strand across the membrane.

ssDNA is translocated from a negatively charged solution filled chamber to a positively charged one. The direction of translocation is also the direction of anion flow. Each nucleobase crossing the membrane has a characteristic current disruption signature measured in pico amperes. These signatures are recorded to generate base-calls.

**Figure 17.** Oxford Nanopore (ONP) sequencing workflow.

causing them to be underrepresented or absent due to low read quality.[81] The single-molecule approaches of 3rdGS offer an array of solutions to the problems encountered by sequencers relying on short reads.[88]

## Third-Generation Sequencers

Third-generation sequencing is a markedly different approach to the earlier massively parallel sequencers, characterized by amplification-free sequencing and long reads, spanning in the kbp range. This increase in read lengths is a result of a shift in strategies: development of proprietary polymerase enzymes and a different application thereof as strategies that do not rely on polymerases at all. The former strategy belongs to Pacific Bioscience Single-Molecule Real-Time (SMRT) sequencing with proof of concept in 2009 and the latter to Oxford Nanopore Technology sequencing with papers from 2001 and 2012 contributing to the proof of concept.[89−91]

**SMRT.** Pacific Biosciences Single Molecule Real-Time (SMRT) is an SBS technology (Figure 16). Where previously at step 2 target sequences were enriched, SMRT requires no amplification. Instead, the library of dsDNA fragments is capped with a hairpin adapter at both ends and denatured, achieving an ssDNA with a circular topology.[89,92,93] This circular DNA construct will be sequenced over and over, effectively amplifying the signal not only for the template but also for each individual strand of the template. The actual sequencing procedure is as follows: each species of dNTPs carries a fluorophore that is phospholinked and is naturally cleaved upon incorporation.[94,95] In contrast to previously described SBS methods, SMRT is able to detect single-nucleobase incorporation events in real time. This is achieved by immobilizing a $\varphi29$ high-fidelity DNA polymerase at the bottom of a nanophotonic chamber called Zero Mode Waveguide (ZMW).[96,97] The polymerase derived from the *Bacillus subtilis* phage $\varphi29$ naturally has several advantageous properties: $\varphi29$ polymerase does not require a primase to commence replication, instead using a terminal protein as a primer; this polymerase immediately displaces the newly synthesized strand from the template and is highly processive (in other words, it continues replication without dissociation); finally, this enzyme is highly efficient and continues the polymerization reaction to generate ssDNA chains upward of 70 kbp in length with an incorporation rate of 1.5−6 bases/s (when employed by SMRT) and an error rate of $\sim10^{-5}$.[89,95,98,99] The polymerase used by SMRT is a mutant $\varphi29^{N62D}$ characterized by a reduced so-called proofreading or 3′−5′ exonuclease activity.[95]

The shape, size, and material of the nanophotonic structure, the ZMW, dictate uniquely suitable physical properties of the chamber, allowing for precise temporal separation of dNTP incorporation events as well as avoiding interference by the background fluorescence of free-floating dNTPs.[92] The ZMWs are typically around 70 × 100 nm in size with the laser reaching about 30 nm into the chamber; together these conditions ensure high precision in accurate signal detection.[92,93,96] The ZMW is illuminated by a laser from below; the fluorescent signal emission is also detected from below. The confocal-type aperture at the bottom of the well allows it to reject any fluorescent signal coming from fluorophores diffusing in the ZMW but outside of the so-called confocal volume ("out-of-focus" signal), only capturing signal released in the region where the polymerase is immobilized.[89,100] This detail is important since the polymerase requires relatively high concentrations of labeled dNTPs (0.1−10 $\mu$M) to function in the aforementioned beneficial manner.

Continuous replication of a circular template used by SMRT is called circular consensus sequencing.[89] A SMRT cell typically holds 150 000 ZMWs with modern SMRT cells carrying millions of ZMWs, and read lengths are dependent on the longevity of the polymerase enzyme. Library templates range between 10 and 20 kbp and read lengths between 10 and 25 kbp, achieving an accuracy of 89−99%.[89,101,102]

**Nanopore Sequencing.** Since the first conception of investigating biomolecules by funneling these molecules through a nanopore protein (porin) embedded in a membrane, the porins as well as solid-state nanopores have been shown to successfully achieve this task.[103−106] The idea is that a voltage is applied across the membrane/surface in which the pore is embedded, allowing for an ionic current to flow.[91] When a nucleotide passes through the pore, a signature perturbation in

Table 1. Main Parameters of Interest between Currently Commercially Available Sequencing Methods[a]

| Commercialized sequencing technology | Read length | Data generated per run | Accuracy (%) | Run time | Sequencing cost (survey) | Machine cost (new) |
|---|---|---|---|---|---|---|
| ONT | 10–60+ (up to 200) kbp | 2.8 Gb to 10 Tb | >99 | 72 h[b] | ∼€600 for 20–30 Gb | $2000–67 000 |
| PacBio SMRT | 10–25 kbp | 24–360 Gb | >99.5 | 24–30 h | €750–800 for 4 Gb | $60 000–779 000 |
| Sanger | 500–900 bp | 84 kbp | 99.999 | 8 h | €60–290 for 20–2000 bp | $100 000–300 000 |
| Ion Torrent | 200–400 bp | 30 Mb to 25 Gb | 98.22 | 14–24 h | | $64 000–104 000 |
| MGI | 50–300 bp | 7.5 Gb to 76.8 Tb | 94.8–98.98 | 5–106 h | €180 for 3 Gb | $230 000–1 000 000[c] [150 Gb to 7 Tb] |
| Illumina | 50–250 bp | 140 Mb to 16 Tb | 99.2–99.7 | 4–48 h | €120–270 for 3 Gb | $100 000–1 000 000 |
| Roche 454 | 700 bp | 0.7 Gb | 99.9 | 24 h | | $500 000[d] |
| SOLiD | 50 bp | 120 Gb | 99.94 | 7–14 days | | $495 000[d] |
| Helicos Biosciences | 30–35 bp | 20–30 Gb | 96 | 7–8 days | | $1 350 000[d] |

[a] SOLiD and Roche 454 sequencing is largely phased out and not available as a commercial service. The other technologies are relatively widespread, with Illumina, PacBio and ONT being very widespread in Europe, and MGI also available in the United States. The pricing is very tricky to determine, so a small survey of quotation was completed. Despite the advertised quick run times, a sequencing service on average takes between 3 and 6 weeks (15–30 weekdays); this is prior to any bioinformatic analysis. Additionally, the price estimate survey was conducted for the same type of samples (except Sanger sequencing), which is 5 metagenomic DNA samples, with 20 M reads for approaches like Illumina and MGI and the equivalent by other technologies. Consequently, the table is filled in with approximate price and data amount generated per sample. The long-read sequencers like ONT and PacBio do suffer from error rates higher than the high-accuracy rates reported by the manufacturers, which is why, especially in the case of ONT, a much larger sequenced data volume was required. Finally, no data was available for Ion Torrent as a commercial service. [b] The ONT does not have a set run time; rather, data is generated as needed, and the maximum readouts are provided by the manufacturer with a run time of 72 h. [c] The pricing of new MGI machines is not freely available with few exceptions. The system that produces 76.8 Tb of data per a 3 day run is also the one that can deliver a $99 genome. However, it is not offered directly for sale. The biggest unit for sale is the 7 Tb machine at $1 000 000. [d] There is no up to date information about the cost of new units of these machines, so these are prices from the mid-2000s to mid-2010s. There is, however, an assortment of variously priced secondhand and refurbished machines available, corroborating the idea that while production is discontinued the sequencers are still in use.

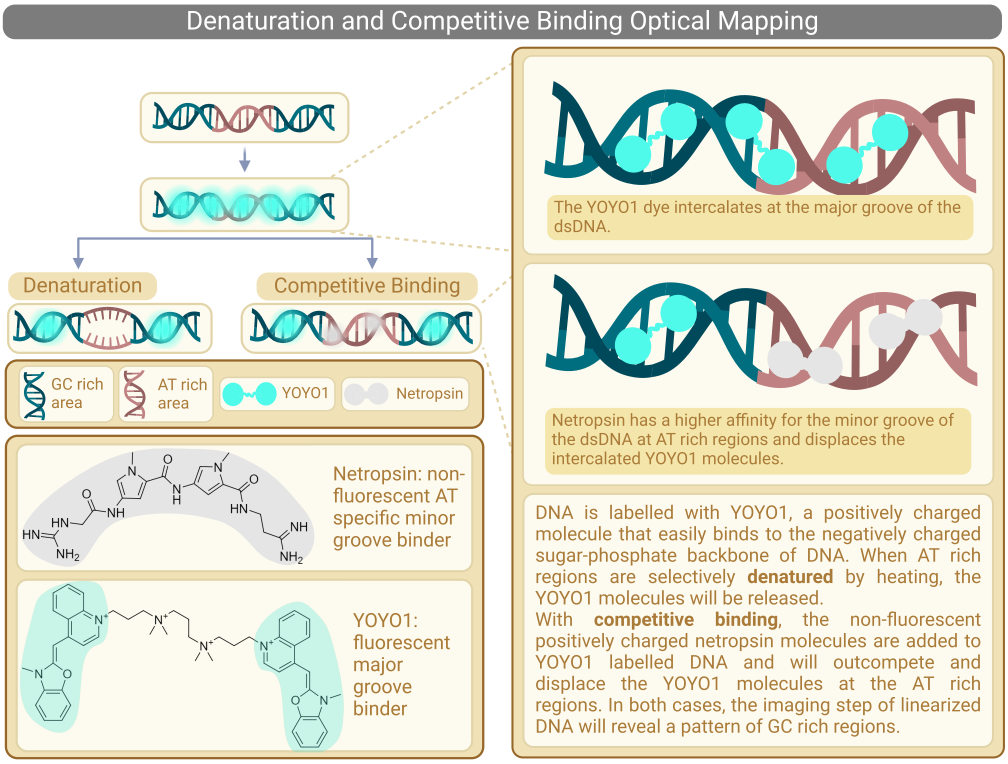
the current can be observed.[91] The more known porins include *Staphylococcus aureus* α-hemolysin and *Mycobacterium smegmatis* porin A (MspA), while solid-state nanopores include Si nanopores in $SiN_X$ membrane.[91,107,108] Helicases, transportases, and translocases are some of the additional biomolecules used to facilitate the funneling process, for example, by unwinding DNA or slowing down the speed of DNA translocation to allow recording of individual translocation events. Commercial sequencers are available from Oxford Nanopore Technology (ONT).

ONT is considered as a third-generation sequencing method. The first published attempts at funneling DNA through a membrane using a porin protein date back to 1996 to a report by Deamer with a following publication in 2001 showing that nanopore–ssDNA conjugates are able to discriminate between matched and mismatched complementary strands to the conjugated ssDNA oligomer.[90,104] In 2005, the company Oxford Nanopore Technologies was founded in the United Kingdom with their first commercially available platform, MinION, entering the market in 2015.[109] Interestingly, as soon as 2016, a series of tests began to evaluate the MinION performance in space on the International Space Station.[110] The first successful polymerase enzyme–nanopore combinations yielded a DNA translocation speed of 30 bases/s.[106,111]

Currently, the technology works as follows (Figure 17). ssDNA strands are funneled through a nanopore enzyme, base-by-base, with ONT reporting speeds up to 400–450 bases/s.[106,111] The translocation speed is controlled by a helicase enzyme, also responsible for unwinding the dsDNA: the speed has to be controlled to ensure that every nucleotide translocation is properly detected.[112] A constant voltage is applied across the planar phospholipid bilayer membrane in which the nanopore is embedded. The membrane separates two chambers filled with salt solutions. With a voltage difference applied between the two chambers, an ion current flows through the pore. The voltage also drives the negatively charged DNA through the pore toward the positively charged compartment.[104] When a nucleobase passes through the nanopore channel the ion current is altered with a nucleobase-specific signature, which is recorded and serves as the proxy for base calling.[104] Currently, base calling is conducted on 5- or 6-mers using a hidden Markov model and the Viterbi algorithm to determine the succession of the 5- or 6-mers.[113] The nanopore is 1.2 nm in diameter, allowing the passage of a single DNA (or RNA) strand at a time.[104]

The library preparation step is amplification free: the DNA is fragmented, and one end of the fragment is capped by a hairpin adapter, effectively allowing for both strands to be read by the pore protein in succession.[113] Ultimately, there are a total of three different library preparation ways: one taking no more than 10 min, and another one geared toward 100+ kbp fragment preparation. A typical nanopore sequencing flow cell holds thousands (e.g., 2048 for the MinION platform) of such nanopores.[114] Read lengths, theoretically, only depend on the size limit of DNA extraction, with reported ultralong reads reaching 50 and 100 kbp.[114,115] Despite the tremendous progress made so far, base-calling error rates are reported to be anywhere between 15% and 5%. Widespread efforts directed at the development of error-correcting strategies are ongoing.[116−118]

**Summary: Commercial Sequencers.** Table 1[24,64,119−128] illustrates the main parameters of interest between currently commercially available sequencing methods, including the discontinued methods like SOLiD and Roche 454 Pyrosequencing (for these data goes back to 2012–2014). Entries that show data generated per run and run times in ranges refer to the minimum and maximum outputs and run times of the


Denaturation and Competitive Binding Optical Mapping
Denaturation
Competitive Binding
GC rich area
AT rich area
YOYO1
Netropsin
Netropsin: non-fluorescent AT specific minor groove binder
YOYO1: fluorescent major groove binder

The YOYO1 dye intercalates at the major groove of the dsDNA.

Netropsin has a higher affinity for the minor groove of the dsDNA at AT rich regions and displaces the intercalated YOYO1 molecules.

DNA is labelled with YOYO1, a positively charged molecule that easily binds to the negatively charged sugar-phosphate backbone of DNA. When AT rich regions are selectively **denatured** by heating, the YOYO1 molecules will be released. With **competitive binding**, the non-fluorescent positively charged netropsin molecules are added to YOYO1 labelled DNA and will outcompete and displace the YOYO1 molecules at the AT rich regions. In both cases, the imaging step of linearized DNA will reveal a pattern of GC rich regions.

**Figure 18.** Optical genome mapping: mechanisms of action for denaturation mapping and competitive binding (CB).

technologies. In most cases, the ranges also refer to the use of different types of machines, for example, benchtop sequencers as opposed to bulky standalone machines. It is obvious that price is an important factor for sequencing as well, but it is tricky to determine exactly the costs of a sequencing run. One reason is that the smaller, palm-sized units of ONT, the benchtop sequencers, and the large wardrobe-resembling units will to a certain extent have different target use. For example, the palm-sized ONT is geared toward field work and is designed for ease of transport and use outdoors, while the benchtop sequencers are more suitable for in-lab sequencing, and the large, bulky devices are most likely to be found in core facilities. The price ranges for machine costs reflect the various designs of devices available. Another reason that complicates pricing is that for very large sequencers the machine itself is extremely costly, reaching astonishing prices, for example, one million dollars for one of the newest Illumina machines.[129] However, it is with these types of hugely expensive machines where running the machine at capacity allows for a full human genome to be sequenced for \$100−300.[130] ONT, for example, offers much cheaper machines. However, the consumables remain relatively expensive at \$600−900 per run, so it would be difficult to achieve a \$100 genome; at the same time, a starting kit of one of their sequencers is currently available for ∼\$2000.[131,132] Having said this, the consumables for the large and expensive units are also expensive, but the cost is offset by how many sequencing runs take place in parallel. A brief survey of pricing was conducted for this paper, revealing that, on average, MGI and Illumina whole genome sequencing costs about the same. For example, when looking to sequence less than 10 samples to obtain 3 Gb per sample, it costs between ∼\$200 (MGI and Illumina in the United States) and ∼€500 (Illumina in Europe).

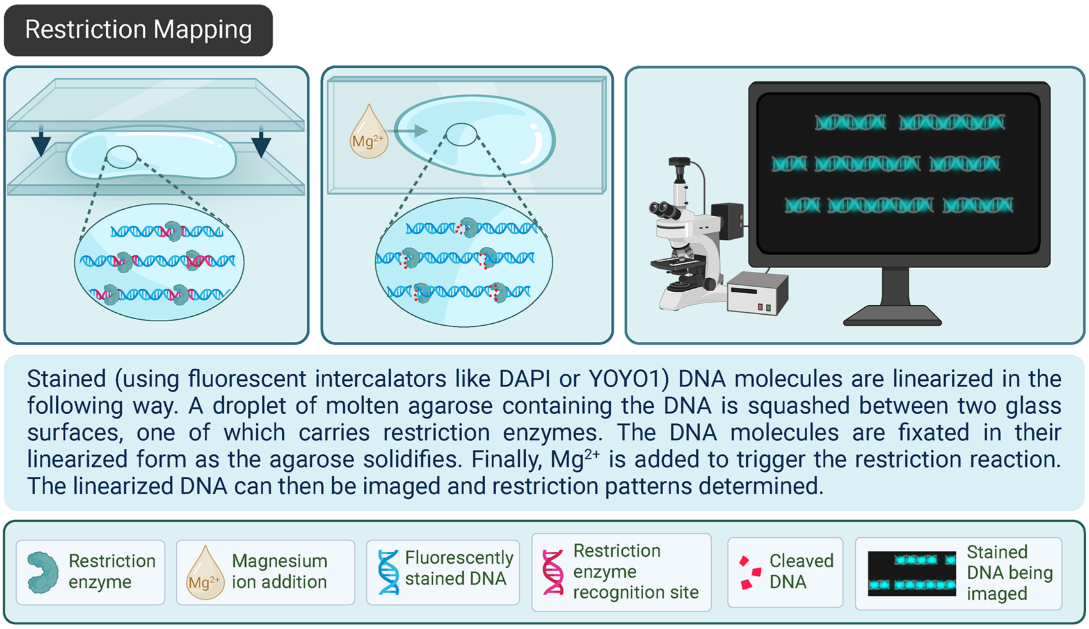
It is worth mentioning that the biggest market share in 2022 went to Illumina at a staggering 80% (reported the same in 2023) with the next contender, Ion Torrent at 7%, with MGI, ONT, and PacBio following at 6%, 4%, and 3% respectively.[133−135]

## ■ OPTICAL GENOME MAPPING

As indicated in the beginning, the following sections will describe methodologies of investigating shorthand of DNA in place of base-by-base readouts. In other words, the techniques described in these sections refer to visual examination of genomic DNA based on sequence-specific markers. Depending on the marker and marking strategy, the readouts allow one to map and identify the DNA, DNA regions, and/or structural variants of the regions of interest. Typically, these methods are characterized by their use of fluorescence microscopy and are referred to as optical mapping (OM). The availability of high-quality sequences generated using whole genome sequencing (WGS) methods described above enables OM technologies as means to either focus on the sequence variation or massively speed up identification using patterns or markers as shorthand for the base-by-base sequence. Currently, all of these methods rely on WGS sequence readouts and can, therefore, be considered as complementary analytical tools. One of the main advantages of the OM approaches is the reduction of data amount extracted from sequences to identify them, thereby alleviating the computational resources required to subsequently identify elements like structural variants.

### Nonenzymatic OM

**Competitive Binding/Affinity-Based OM.** Competitive binding (CB) OM is a well-known example of enzyme-free DNA mapping.[136] The competition to bind takes place on a dsDNA molecule between the fluorescent bis-intercalator YOYO-1 (a homodimer of Oxazole Yellow) and the natural



Stained (using fluorescent intercalators like DAPI or YOYO1) DNA molecules are linearized in the following way. A droplet of molten agarose containing the DNA is squashed between two glass surfaces, one of which carries restriction enzymes. The DNA molecules are fixated in their linearized form as the agarose solidifies. Finally, $Mg^{2+}$ is added to trigger the restriction reaction. The linearized DNA can then be imaged and restriction patterns determined.

**Figure 19.** Optical genome mapping: restriction mapping workflow.

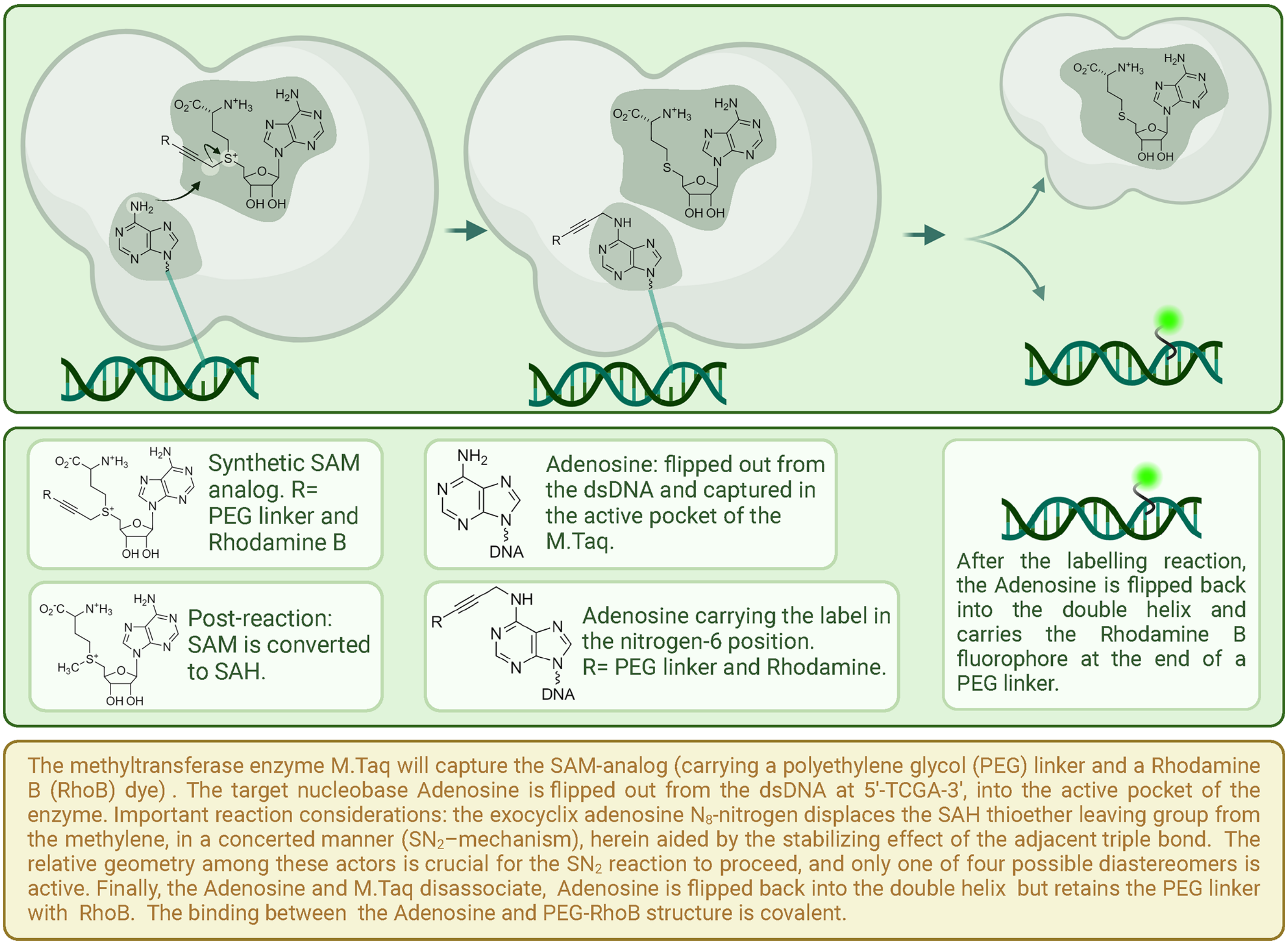
antibiotic, nonfluorescent minor groove binder netropsin.[136,137] The method was successfully demonstrated by the Westerlund lab in 2012 using bacteriophages lambda and T4, which are 48.5 and 166.5 kbp in size, respectively.[136] In 2014, the Westerlund lab demonstrated the method on *Escherichia coli* DNA fragments of 50−150 kbp, where the full genome is close to 5 Mbp in size.[138] The principle of CB is as follows: YOYO-1-stained DNA is subjected to netropsin treatment, whereby the netropsin molecules displace the fluorophores at AT-rich regions.[136,138] The reaction can be achieved in one or two steps. YOYO-1 has a high overall or sequence-independent binding constant ($K_b > 10^{10}$ $M^{-1}$), while netropsin has a higher, sequence-dependent affinity for AT-rich regions ($K_b = 10^8$ $M^{-1}$) compared to GC-rich regions ($K_b = 10^5$ $M^{-1}$).[139−141] Thanks to the high sequence specificity, netropsin when added in excess (150−8000 netropsin to YOYO-1) displaces or outcompetes the YOYO-1 already bound on the dsDNA.[136,138] DNA is linearized by molecular combing on a positively charged glass slide or stretched through nanofluidic channels on a chip.[142−145] Upon imaging, the YOYO-1-netropsin-stained molecules yield intensity variations directly revealing AT-rich regions by the absence of fluorescence. To reduce any signal distortion of the intensity trace rising from thermal fluctuations and diffusion of the DNA molecules in the nanochannels, time series (kymographs) are recorded, aligned, and averaged.[136,138] The map resolution achieved with CB is on the order of kilobase pairs. The CB method has been applied for profiling of clinical urine samples as well as partially mapping the human genome at a resolution on the order of hundreds of kilobase pairs from peripheral blood mononuclear cells (PBMCs).[139,146]

Notably, the pioneer of molecular combing, Bensimon, has participated in the commercialization (Genomic Vision) of an optical mapping method whereby DNA linearized by molecular combing is hybridized with fluorescent probes in a fashion resembling fluorescence in situ hybridization (FISH).[147,148] This method produces genomic maps of 100+ kbp and can be used for genomic rearrangement investigations and studies of DNA replication.[149]

**Denaturation OM.** Similar to competitive binding OM, denaturation OM unveils AT- and GC-rich locations using fluorescence intensity variations by making use of sequence-dependent denaturation of DNA (Figure 18).[150,151] AT-rich regions have lower melting (denaturing) temperatures compared to GC-rich regions, so when DNA is subjected to heating, that only denatures AT-rich regions and is only stained with YOYO-1 at GC-rich regions since this dye only forms a stable complex with nondenatured dsDNA. Excess YOYO-1 is not a concern as unbound YOYO-1 exhibits a 1000-fold lower intensity than its bound counterpart.[140] Denatured regions will appear nonfluorescent. DNA is linearized using nanofluidic devices to record kymographs using fluorescence microscopy.[150,151]

## Enzymatic OM

**Restriction OM.** Restriction OM can be considered as the on-surface alternative of pulsed field gel electrophoresis (PFGE) with restriction enzymes. In PFGE with restriction enzymes, DNA is subjected to restriction followed by separation of the fragments on an agarose gel applying pulsed current, effectively yielding genome-specific patterns.[152] Restriction typing with OM is similar insofar as visual maps of the restricted fragments are generated (Figure 19).[153] The process is different, however. PFGE as well as restriction mapping rely on the highly site-specific action of restriction enzymes. The restriction processing of DNA was developed by Schwartz and colleagues in the early 1990s and was initially applied to generate ordered restriction maps of *Saccharomyces cerevisiae* chromosomes.[153] Here, single DNA molecules are linearized by liquid flow: a mixture of DNA molecules in molten agarose is pipetted on a coverslip and covered with a microscopy slide treated to carry a restriction enzyme on the surface. Subsequent gelling of the agarose causes fixation of the linearized DNA molecules.[153] Typically, the DNA molecules would be stained with YOYO-1 or DAPI (4′,6-diamidino-2-phenylindole, fluorescent minor groove binder with a preference for AT-rich regions), allowing imaging with fluorescence microscopy. Before imaging, $Mg^{2+}$ is added to



Synthetic SAM analog. R= PEG linker and Rhodamine B

Adenosine: flipped out from the dsDNA and captured in the active pocket of the M.Taq.

Post-reaction: SAM is converted to SAH.

Adenosine carrying the label in the nitrogen-6 position. R= PEG linker and Rhodamine.

After the labelling reaction, the Adenosine is flipped back into the double helix and carries the Rhodamine B fluorophore at the end of a PEG linker.

The methyltransferase enzyme M.Taq will capture the SAM-analog (carrying a polyethylene glycol (PEG) linker and a Rhodamine B (RhoB) dye) . The target nucleobase Adenosine is flipped out from the dsDNA at 5'-TCGA-3', into the active pocket of the enzyme. Important reaction considerations: the exocyclix adenosine $N_8$-nitrogen displaces the SAH thioether leaving group from the methylene, in a concerted manner ($SN_2$–mechanism), herein aided by the stabilizing effect of the adjacent triple bond. The relative geometry among these actors is crucial for the $SN_2$ reaction to proceed, and only one of four possible diastereomers is active. Finally, the Adenosine and M.Taq disassociate, Adenosine is flipped back into the double helix but retains the PEG linker with RhoB. The binding between the Adenosine and PEG-RhoB structure is covalent.

**Figure 20.** Optical genome mapping: Fluorocode mechanism of action, a doubly activated synthetic analogue of methyltransferase cofactor, labeling process. Reproduced with permission from ref 181. Copyright 2024 ACS Omega.

the gelled mix, diffusing through the agarose pores and triggering the enzymatic reaction.[153] Introduced cuts become visible as nonfluorescent gaps (∼1 $\mu$m) due to partial DNA coil relaxation surrounding the restriction site. Time-lapse recording is to account for local DNA molecule motion due to the mild fixation conditions.[154] For analysis of molecule length, two complementary approaches are used: relative intensity with an internal reference and relative length with the same set of reference traces.[154,155]

Improvements in restriction OM came with the development of more advanced molecular combing approaches, involving polylysine-treated or silanized glass surfaces, as well as more elaborate microfluidic and nanofluidic devices.[155−158] Due to the more stable fixation and deposition control of DNA molecules associated with these new combing methods compared to agarose embedding, data quality and throughput could be improved. Consequentially, more molecules could be visualized, internally aligned, and averaged, yielding higher quality, long-range restriction maps.[158,159]

Restriction OM has an inherently lower resolution due to the indirect readout of restriction sites, namely, the cleaved gap and measurement of integrated intensity across the visible trace instead of measuring single fluorescent emitters.[160] The most frequently used restriction enzymes recognize 6−8 bp sequences, resulting in an average recognition site density of one site every 4−65 kbp, highly depending on the recognition site and the particular genome.[159,161] While the reduced information load per molecule can be compensated by increasing the fragment size, obtaining high molecular weight DNA especially from complex environmental samples is not trivial with established kits typically aimed at short-read NGS applications. However, there certainly are both commercial and academic attempts at HMW DNA extraction.[162,163]

**Nick Repair OM.** While nick repair (NR) as a strategy has been around since the 1970s, the method has gained more traction since the 2000s, resulting in a commercial platform.[164−166] NR OM is characterized by a number of sequential enzymatic reactions. First, a site-specific nicking enzyme (e.g., Nt.BspQI) introduces a single-stranded break in the DNA.[166] For a typical nick enzyme, the recognition site runs between 4 and 8 bp.[167] Nick enzyme treatment is followed by a polymerase treatment, where the nicks are recognized and repaired by DNA polymerase I, catalyzing two simultaneous reactions. As the 5′− 3′ exonuclease or proofreading activity of the polymerase removes nucleotides at the nick site, it simultaneously incorporates new, fluorescently tagged nucleotides in the same 5′−3′ direction.[166,168] The original Rigby and Berg method from the 1970s is called nick translation, where instead of fluorescent labels radioactive labels were used.[164] Some of the more modern methods would also stain the nick-repaired DNA with intercalating dyes, such as DAPI or YOYO1, to visualize the

Table 2. Summary of the Main Parameters of Interest for Optical Mapping Technologies[a]

| Optical mapping technology | Read length (kbp) | Nonenzymatic/ enzymatic | Commercial provider |
|---|---|---|---|
| Competitive binding | 50−150 | nonenzymatic | |
| Denaturation OM | 100+ | nonenzymatic | |
| Molecular combing | 100+ | nonenzymatic | Genomic Vision[b] |
| Restriction OM | ∼10−26 | XhoI | |
| Methyltransferase-mediated OM | 30+ | M.*Taq*I | Perseus Biomics |
| Nick repair OM (nick translation) | 100+ | Nb. BbvCI, Vent(exo-) polymerase | Bionano Genomics |

[a]Perseus Biomics is specializing in human gut microbiome studies, while Bionano Genomics is specializing in genome structural variant monitoring in the context of clinical oncology. [b]Genomic Vision ceased commercial operations during the writing of this review.

whole DNA strand.[166] Labeled DNA molecules are stretched on specially treated coverslips or in nanochannel arrays and subsequently imaged with fluorescence microscopy.[165,166]

A major advantage of NR OM compared to other OM methods discussed is the covalent binding of the fluorescent tags. Thereby, potential loss or perturbation of this tag during combing is no longer a problem. NR OM may exhibit false positives resulting from accidently induced or naturally present nicks in the target strand.[169]

In 2016, McCaffrey and co-workers replaced the nicking enzyme with a Cas9 D10A protein carrying a mutation in one of its two nuclease domains.[170] Consequently, the Cas9 D10A catalyzes single-stranded DNA breaks, achieving the same enzymatic activity as nickases. An inherent part of the Cas9 system is a special RNA guide molecule, the sequence of which can be altered based on the target sequence.[171] The guide-targeted recognition sites can be as large as 23 bp, making this method ideally suited for tagging genes and assessing gene copy numbers, for example.

**Methyltransferase-Mediated OM.** There is another enzyme-based OM method: in this case, the method relies on the use of prokaryotic DNA methyltransferases (MTases) transferring a synthetic cofactor analogue or part of it to a target location on a DNA strand, first shown in 1998 by Pignot and colleagues.[172] In nature, these enzymes catalyze a highly site-specific methyl group transfer onto dsDNA, with *S*-adenosyl-L-methionine (SAM or AdoMet), serving as the methyl donor and reaction cofactor.[173] Methylation occurs on the base moiety of the nucleotide, and in the case of MTases, typically used for optical mapping, the reaction product is $N^6$-methyladenine.[174] MTase labeling works in the following way: the MTase slides along the dsDNA, and once it finds its target sequence, it will flip the target nucleotide out of the double helix.[173−177] The active pocket of the enzyme carries SAM, ensuring the required proximity for the transalkylation reaction to occur: the methyl group is transferred to the substrate through a second-order nucleophilic substitution ($S_N2$) reaction, converting SAM to *S*-adenosyl-L-homocysteine (SAH).[175−178] DNA MTases exhibiting tolerance to synthetic SAM analogues is not trivial; however, research indicates that a variety of functional groups can be site-specifically transferred to DNA.[177] One type of such synthetic analogues to natural SAM is doubly activated methyltransferase cofactors (Figure 20).[177,178]

DNA Fluorocode OM is an enzymatic DNA-labeling approach combining such a doubly activated cofactor with a DNA MTase.[161,178−181] Fluorocode OM, like previous OM techniques, works with single linearized molecules of DNA with molecule or read length ranging between 30 and 60 kbp. Fluorocode OM was first introduced by the Hofkens group in 2011 with a later iteration in 2020.[169,179] The enzyme used is the thermophilic methyltransferase *M.TaqI* with recognition sequence 5′-TCGA-3′.[172,179] The fluorescent label is a doubly activated synthetic SAM analogue conjugated to a fluorophore.[178,179] The DNA is labeled in a single-step reaction as the MTase delivers the linker with the fluorophore to the nitrogen-6 position of the adenosine, where they remain covalently bound. Next, the DNA is linearized using molecular combing and imaged using fluorescence microscopy.[143,179] The resulting images of signal sequences or barcodes allow cross-taxonomic identification and species-level taxonomic resolution, and depending on the sample, strain level investigation can be achieved.[179−181] The Fluorocode method is being commercialized by Perseus Biomics for human metagenomics applications.

Other enzyme-based OM methods are used for research of structural variants (SV) of mammalian genomes. For example, Bionano Genomics used an enzyme with the recognition site 5′-CTTAAG-3′. The approach is similar to Fluorocode OM, with the main application being identification of SVs, for example, in the context of oncology.[182,183]

**Summary Optical Mapping Technologies.** Optical mapping encompasses a number of technologies, only a few of which are breaking into the market at this time. Table 2 provides a succinct summary of the main parameters when comparing the different optical mapping methods. Notably, even the shortest read length for OM technologies is in the range of tens of kilobase pairs, while for sequencing it was ONT alone that could provide read lengths in the order of kilobase pairs. Optical mapping technologies occupy a research and market niche complementary but also competitive to sequencing. While sequencing is here to stay, optical mapping is in the process of proving itself as a reliable actor both in the academic and in the commercial sectors.

## ■ CONCLUSION

Genomic sequencing started off by eroding DNA fragments targeting one nucleotide species at a time and generating a ladder of fragments of different lengths. Almost simultaneously, a different method was proposed where ssDNA has its complementary strand generated, again with halted elongation at particular nucleotide species. In both cases, a ladder of DNA fragments is generated with known final nucleotide species. In both cases, the DNA was radioactively labeled and visualized on an agarose gel. The next step was moving away from using the toxic radioactive labels and using fluorescent markers instead, eventually switching to a polyacrylamide gel. These methods, while accurate, nevertheless, turned out to be too slow, and scaling up was not easily achievable. Alternative strategies were needed, where more DNA could be sequenced, retaining the accuracy of the final readout. Massively parallel sequencing allowed for both. The strategy for accuracy was simply copying many times over the original fragmented DNA template. Generally, the copying was achieved in two general ways: either preamplifying the DNA using a PCR reaction (eventually amplifying copies of copies of the original template) and subsequently sequencing the amplified copies all at once (Illumina) or continuously amplifying the original template by sequencing it (SMRT). A third, later version includes generating a single DNA strand of tandem repeats of a template (all of the repeats generated from the original template; DNBseq). These microreactions were all designed to take place in parallel in the hundreds of thousands to millions, which allow one to interrogate large genomes billions of base pairs in size.

Additionally, it turned out that the use of fluorophores, first introduced in later iterations of Sanger sequencing, is extremely conducive to DNA investigations, as almost all of the subsequently developed sequencing methods use fluorophores. There are exceptions, of course, where the electrochemical properties of the DNA elongation events (Ion Torrent) or DNA strand translocation across a nanopore (ONT) are recorded. Another exception making use of the electrochemical properties of a dNTP incorporation event is using the released $PP_i$ for a subsequent luficerin-to-oxyluciferin conversion (Roche 454). In those cases, the DNA polymer is considered in its natural form. In cases where DNA is sequenced using fluorophores, typically it is via dNTPs that are conjugated with fluorophores. There have been various strategies of the exact fluorophore attachment, phospholinked dNTPs, where the fluorophore is naturally cleaved upon incorporation, nucleobase-conjugated fluorophores, where the fluorophore may be cleaved, leaving behind a small linker, and antibody-conjugated fluorophores, where the antibody is specific to each species of reversible chain-terminating dNTP. The read lengths we may obtain of genomic DNA vary from 100−150 bp to 100−150 kbp for the single-molecule techniques like ONP and OM.

Effectively, this means that there are different ways of interrogating the same sequences with methods that have complementary strong points; for example, where methods with shorter reads will struggle at repetitive genomic domains, methods with long reads will allow for reliable scaffolds. Where a reference genome can be established with accurate short read base-by-base readouts, other methods with less precise readouts but cheaper and faster profiling capacity of large number of samples together can allow for a more efficient delivery of particular readouts.

This review has discussed the advancements in the chemistries of sequencing technologies from inception of an idea to commercial products and services across a spectrum of technologies in a broad sketch of the various technical possibilities exhibited by existing technologies. Together these stories illustrate the flight of scientific imagination, which remains constrained by the chemical and physical properties of the DNA and fluorophores and the biochemical processes of reactions like DNA synthesis, methylation, and ligation. Given the power and precision of existing technologies, one may only expect the future to bring technologies closer to the public by reduced costs and improved confidence. For example, genetic testing of family members now is a routine step when encountering a genetic disease in a patient. Commercial services like Ancestry and 23andMe offer the public insights into their own personal and family history: the kits are available via mail order and are financially accessible. The $1000 genome is already history, and we may only expect access to sequencing to be a more mundane affair, more integrated into various aspects of the modern life including healthcare and food production, animal husbandry, and more, even technologies like information storage.[184]

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/cbmi.4c00060.

> Discovery of DNA and development of sequencing methods; DNA structure; DNA synthesis; amplification by bacterial artificial chromosomes; PCR reaction applied in the early sequencing methods; early sequencing methods, premassively parallel sequencing; Sanger sequencing, modified ddNTPs; emulsion PCR, pyrosequencing; Sequencing by Oligo Ligation Detection, early concept; Sequencing by Oligo Ligation Detection, commercialized; Sequencing by Oligo Ligation Detection, commercialized; solid-state (bridge) PCR; Illumina, four-dye process; Illumina, single-dye process; Helicos Biosciences, single-molecule sequencing; SMRT library preparation; SMRT sequencing; nanopore sequencing; DNBSeq library preparation; MGI sequencing, DNBSeq and CoolMPS; parameters of a sequencing run; denaturation and competitive binding OGM; restriction mapping; methyltransferase-directed OGM; video sources of commercialized technologies (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**Johan Hofkens** − *Faculty of Science, Chemistry, KU Leuven, Leuven, Flanders 3001, Belgium; Max Planck Institute for Polymer Research, Mainz, Rheinland-Pfalz 55128, Germany;* orcid.org/0000-0002-9101-0567; Email: johan.hofkens@kuleuven.be

### Authors

**Elizabete Ruppeka Rupeika** − *Faculty of Science, Chemistry, KU Leuven, Leuven, Flanders 3001, Belgium;* orcid.org/0000-0002-7093-8147

**Laurens D'Huys** − *Faculty of Science, Chemistry, KU Leuven, Leuven, Flanders 3001, Belgium;* orcid.org/0000-0001-6325-9720

**Volker Leen** − *Perseus Biomics B.V., Tienen 3300, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/cbmi.4c00060

### Notes

All figures (except Figure 20) are originally created for this manuscript and were made using Biorender 2024, InkScape, and Chemdraw.

The authors declare the following competing financial interest(s): Johan Hofkens and Volker Leen are co-founders of Perseus Biomics.

## ■ ACKNOWLEDGMENTS

J.H. acknowledges financial support from the FWO (Fonds voor Wetenschappelijk Onderzoek, grant number

G0C1821N), the Flemish Government through long-term structural funding Methusalem (CASAS2, Meth/15/04), and the Max Planck Institute through an MPI fellowship.

## ■ REFERENCES

(1) Franklin, R. E.; Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **1953**, *171*, 740−741.
(2) Watson, J. D.; Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953**, *171*, 737−738.
(3) Alberts, B.; et al. DNA Replication Mechanisms. *Molecular Biology of the Cell*, 4th ed.; Garland Science, 2002.
(4) Burgers, P. M. Solution to the 50-year-old Okazaki-fragment problem. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3358−3360.
(5) Okazaki, R.; Okazaki, T.; Sakabe, K.; Sugimoto, K.; Sugino, A. Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proc. Natl. Acad. Sci. U. S. A.* **1968**, *59*, 598−605.
(6) Mandelkern, M.; Elias, J. G.; Eden, D.; Crothers, D. M. The dimensions of DNA in solution. *J. Mol. Biol.* **1981**, *152*, 153−161.
(7) Alberts, B.; et al. From DNA to RNA. *Molecular Biology of the Cell*, 4th ed.; Garland Science, 2002.
(8) Alberts, B.; et al. From RNA to Protein. *Molecular Biology of the Cell*, 4th ed.; Garland Science, 2002.
(9) Chi, K. R. The dark side of the human genome. *Nature* **2016**, *538*, 275−277.
(10) Statello, L.; Guo, C.-J.; Chen, L.-L.; Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 96−118.
(11) Crick, F. Central Dogma of Molecular Biology. *Nature* **1970**, *227*, 561−563.
(12) Armstrong, L. *Epigenetics*; Garland Science, New York, 2020; DOI: 10.1201/9780429258862.
(13) Venter, J. C.; et al. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304−1351.
(14) Lander, E. S.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860−921.
(15) Barranco, C. The Human Genome Project. *Nat. Res.* **2021**, DOI: 10.1038/d42859-020-00101-9.
(16) Sanger, F.; Donelson, J. E.; Coulson, A. R.; Kössel, H.; Fischer, D. Use of DNA Polymerase I Primed by a Synthetic Oligonucleotide to Determine a Nucleotide Sequence in Phage f1 DNA. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70*, 1209−1213.
(17) Sanger, F.; Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **1975**, *94*, 441−448.
(18) Maxam, A. M.; Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **1977**, *74*, 560−564.
(19) Sanger, F.; et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **1977**, *265*, 687−695.
(20) Sanger, F.; Nicklen, S.; Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **1977**, *74*, 5463−5467.
(21) Prober, J. M.; et al. A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides. *Science* **1987**, *238*, 336−341.
(22) Bergot, B. J. et al. Spectrally resolvable rhodamine dyes for nucleic acid sequence determination. US5366860, 1989.
(23) Malamon, J. S.; et al. A comparative study of structural variant calling in WGS from Alzheimer's disease families. *Life Sci. Alliance* **2024**, *7*, No. e202302181.
(24) Cheng, C.; Fei, Z.; Xiao, P. Methods to improve the accuracy of next-generation sequencing. *Front. Bioeng. Biotechnol.* **2023**, *11*, 982111.
(25) Crossley, B. M.; et al. Guidelines for Sanger sequencing and molecular assay monitoring. *J. Vet. Diagn. Investig. Off. Publ. Am. Assoc. Vet. Lab. Diagn. Inc* **2020**, *32*, 767−775.
(26) Alberts, B. et al. Isolating, Cloning, and Sequencing DNA. *Molecular Biology of the Cell*, 4th ed.; Garland Science, 2002.
(27) Sims, D.; Sudbery, I.; Ilott, N. E.; Heger, A.; Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **2014**, *15*, 121−132.
(28) How to calculate the coverage for a NGS experiment; ECSEQ Bioinformatics; https://www.ecseq.com/support/ngs/how-to-calculate-the-coverage-for-a-sequencing-experiment.
(29) Smith, L. M.; et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **1986**, *321*, 674−679.
(30) Swerdlow, H.; Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* **1990**, *18*, 1415−1419.
(31) Moorcraft, S. Y.; Gonzalez, D.; Walker, B. A. Understanding next generation sequencing in oncology: A guide for oncologists. *Crit. Rev. Oncol. Hematol.* **2015**, *96*, 463−474.
(32) Shendure, J.; Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **2008**, *26*, 1135−1145.
(33) Ke, R.; Mignardi, M.; Hauling, T.; Nilsson, M. Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Hum. Mutat.* **2016**, *37*, 1363−1367.
(34) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(35) Kim, J.; Ji, M.; Yi, G. A Review on Sequence Alignment Algorithms for Short Reads Based on Next-Generation Sequencing. *IEEE Access* **2020**, *8*, 189811−189822.
(36) Metzker, M. L. Sequencing technologies — the next generation. *Nat. Rev. Genet.* **2010**, *11*, 31−46.
(37) Valencia, C. A.; Pervaiz, M. A.; Husami, A.; Qian, Y.; Zhang, K. *Next Generation Sequencing Technologies in Medical Genetics*; Springer New York: New York, 2013; DOI: 10.1007/978-1-4614-9032-6.
(38) Mishra, P.; et al. Genome assembly and annotation. In *Bioinformatics*; Singh, D. B., Pathak, R. K., Eds.; Academic Press, 2022; Chapter 4, pp 49−66; DOI: 10.1016/B978-0-323-89775-4.00013-4.
(39) Poptsova, M. S.; et al. Non-random DNA fragmentation in next-generation sequencing. *Sci. Rep.* **2014**, *4*, 4532.
(40) Sambrook, J.; Russell, D. W. Fragmentation of DNA by nebulization. *CSH Protoc.* **2006**, *2006*; DOI: 10.1101/pdb.prot4539
(41) Ribarska, T.; Bjørnstad, P. M.; Sundaram, A. Y. M.; Gilfillan, G. D. Optimization of enzymatic fragmentation is crucial to maximize genome coverage: a comparison of library preparation methods for Illumina sequencing. *BMC Genomics* **2022**, *23*, 92.
(42) Heather, J. M.; Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **2016**, *107*, 1−8.
(43) Goodwin, S.; McPherson, J. D.; McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333−351.
(44) Guo, J.; Starr, D.; Guo, H. Classification and review of free PCR primer design software. *Bioinformatics* **2021**, *36*, 5263−5268.
(45) Khehra, N.; Padda, I. S.; Swift, C. J. Polymerase Chain Reaction (PCR). *StatPearls*; StatPearls Publishing: Treasure Island, FL, 2024.
(46) Sequencing Platforms | Illumina NGS platforms; https://www.illumina.com/systems/sequencing-platforms.html.
(47) Adessi, C.; et al. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* **2000**, *28*, No. e87.
(48) Omelina, E. S.; Ivankin, A. V.; Letiagina, A. E.; Pindyurin, A. V. Optimized PCR conditions minimizing the formation of chimeric DNA molecules from MPRA plasmid libraries. *BMC Genomics* **2019**, *20*, 536.
(49) Shao, K.; et al. Emulsion PCR: A High Efficient Way of PCR Amplification of Random DNA Libraries in Aptamer Selection. *PLoS One* **2011**, *6*, No. e24910.
(50) Tewhey, R.; et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat. Biotechnol.* **2009**, *27*, 1025−1031.

(51) Chai, C. Principle of Emulsion PCR and Its Applications in Biotechnology. *J. Anim. Reprod. Biotechnol.* **2019**, *34*, 259−266.

(52) Siu, R. H. P.; et al. Optimization of on-bead emulsion polymerase chain reaction based on single particle analysis. *Talanta* **2021**, *221*, No. 121593.

(53) Margulies, M.; et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005**, *437*, 376−380.

(54) Merriman, B.; Rothberg, J. M. Progress in Ion Torrent semiconductor chip based sequencing. *Electrophoresis* **2012**, *33*, 3397−3417.

(55) Slatko, B. E.; Gardner, A. F.; Ausubel, F. M. Overview of Next Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* **2018**, *122*, No. e59.

(56) Nyren, P.; Pettersson, B.; Uhlen, M. Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Anal. Biochem.* **1993**, *208*, 171−175.

(57) Ronaghi, M.; Elahi, E. Discovery of Single Nucleotide Polymorphisms and Mutations by Pyrosequencing. *Comp. Funct. Genomics* **2002**, *3*, 51−56.

(58) Shendure, J.; et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* **2005**, *309*, 1728−1732.

(59) McKernan, K. J.; et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **2009**, *19*, 1527−1541.

(60) Landegren, U.; Kaiser, R.; Sanders, J.; Hood, L. A Ligase-Mediated Gene Detection Technique. *Science* **1988**, *241*, 1077.

(61) McKernan, K.; Blanchard, A.; Kotler, L.; Costa, G. Reagents, methods, and libraries for bead-based sequencing. EP20100158867, 2016.

(62) Mitra, S.; et al. Analysis of the intestinal microbiota using SOLiD 16S rRNA gene sequencing and SOLiD shotgun sequencing. *BMC Genomics* **2013**, *14*, S16.

(63) Pettersson, E.; Lundeberg, J.; Ahmadian, A. Generations of sequencing technologies. *Genomics* **2009**, *93*, 105−111.

(64) Liu, L.; et al. Comparison of Next-Generation Sequencing Systems. *J. Biomed. Biotechnol.* **2012**, *2012*, 1−11.

(65) Holt, R. A.; Jones, S. J. M. The new paradigm of flow cell sequencing. *Genome Res.* **2008**, *18*, 839−846.

(66) Sequencing Read Length | How to calculate NGS read length; https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html.

(67) Chen, F.; et al. The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. *Genomics Proteomics Bioinformatics* **2013**, *11*, 34−40.

(68) Bentley, D. R.; et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456*, 53−59.

(69) Milton, J.; Liu, X. Modified nucleosides and nucleotides and uses thereof. US9334328B2, 2007.

(70) *History of Illumina Sequencing*; Illumina, 2014; https://web.archive.org/web/20141012151855/http://technology.illumina.com/technology/next-generation-sequencing/solexa-technology.html.

(71) *2-Channel SBS Technology | Faster sequencing and data acquisition*; Illumina; https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html.

(72) *Illumina CMOS Chip and One-Channel SBS Chemistry*; Illumina, https://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/cmos-tech-note-770-2013-054.pdf.

(73) Drmanac, R.; et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **2010**, *327*, 78−81.

(74) Fehlmann, T.; et al. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenetics* **2016**, *8*, 123.

(75) Drmanac, S. et al. CoolMPS: Advanced massively parallel sequencing using antibodies specific to each natural nucleobase. *BioXRiv* **2020** DOI: 10.1101/2020.02.19.953307.

(76) Pushkarev, D.; Neff, N. F.; Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **2009**, *27*, 847−850.

(77) Yuzuki, D. Observations about Helicos, a single molecule sequencer from 2008; Silent Valley Consulting, 2023; https://silentvalleyconsulting.com/blog/observations-about-helicos-a-single-molecule-sequencer-from-2008/.

(78) Thompson, J. F.; Steinmann, K. E. Single Molecule Sequencing with a HeliScope Genetic Analysis System. *Curr. Protoc. Mol. Biol.* **2010**, *92*, 1−14.

(79) Bowers, J.; et al. Virtual Terminator nucleotides for next generation DNA sequencing. *Nat. Methods* **2009**, *6*, 593−595.

(80) Pfeiffer, F.; et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **2018**, *8*, 10950.

(81) Aird, D.; et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **2011**, *12*, R18.

(82) Reinert, K.; Langmead, B.; Weese, D.; Evers, D. J. Alignment of Next-Generation Sequencing Reads. *Annu. Rev. Genomics Hum. Genet.* **2015**, *16*, 133−151.

(83) Mantere, T.; Kersten, S.; Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. *Front. Genet.* **2019**, *10*, 426.

(84) Van Nieuwerburgh, F.; et al. Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic Acids Res.* **2012**, *40*, No. e24.

(85) What is mate pair sequencing for? ECSEQ Bioinformatics; https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for.

(86) Sahlin, K.; Chikhi, R.; Arvestad, L. Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics* **2016**, *32*, 1925−1932.

(87) Paterson, A. L.; et al. Mobile element insertions are frequent in oesophageal adenocarcinomas and can mislead paired-end sequencing analysis. *BMC Genomics* **2015**, *16*, 473.

(88) Amarasinghe, S. L.; et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020**, *21*, 30.

(89) Eid, J.; et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **2009**, *323*, 133−138.

(90) Howorka, S.; Cheley, S.; Bayley, H. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat. Biotechnol.* **2001**, *19*, 636−639.

(91) Manrao, E. A.; et al. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **2012**, *30*, 349−353.

(92) Levene, M. J.; et al. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science* **2003**, *299*, 682−686.

(93) Korlach, J.; et al. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 1176−1181.

(94) Korlach, J. Phospholink nucleotides for sequencing applications. US20130122490A1, 2013.

(95) Korlach, J.; et al. Long, Processive Enzymatic Dna Synthesis Using 100% Dye-Labeled Terminal Phosphate-Linked Nucleotides. *Nucleosides Nucleotides Nucleic Acids* **2008**, *27*, 1072−1083.

(96) Levene, M. J.; Korlach, J.; Turner, S. W.; Craighead, H. G.; Webb, W. W. Zero-mode clad waveguides for performing spectroscopy with confined effective observation volumes. US6917726B2, 2005.

(97) Blanco, L.; Bernad, A.; Salas, M. PH$\varphi$29 DNA polymerase. US5001050A, 1991.

(98) Blanco, L.; et al. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **1989**, *264*, 8935−8940.

(99) del Prado, A.; et al. New insights into the coordination between the polymerization and 3′-5′ exonuclease activities in $\phi$29 DNA polymerase. *Sci. Rep.* **2019**, *9*, 923.

(100) The term confocal aperture refers to confocal microscopy, where a pinhole is used to control the amount of light reaching the detector, effectively filtering out light emission beyond the confocal volume.[96,97] Notably, even the smallest confocal pinhole in a confocal microscope is several orders of magnitute larger than the diameter of a SMRT ZMW, ranging from tens to hundreds of micrometers.

(101) Ardui, S.; Ameur, A.; Vermeesch, J. R.; Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* **2018**, *46*, 2159−2168.

(102) Minoche, A. E.; et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biol.* **2015**, *16*, 184.

(103) Derrington, I. M.; et al. Nanopore DNA sequencing with MspA. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 16060−16065.

(104) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 13770−13773.

(105) Chien, C.-C.; Shekar, S.; Niedzwiecki, D. J.; Shepard, K. L.; Drndic, M. Single-Stranded DNA Translocation Recordings Through Solid-State Nanopores on Glass Chips at 10-MHz Measurement Bandwidth. *ACS Nano* **2019**, *13*, 10545−10554.

(106) Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **2021**, *39*, 1348−1365.

(107) Laszlo, A. H.; Derrington, I. M.; Gundlach, J. H. MspA nanopore as a single-molecule tool: From sequencing to SPRNT. *Methods San Diego Calif* **2016**, *105*, 75−89.

(108) Cherf, G. M.; et al. Automated Forward and Reverse Ratcheting of DNA in a Nanopore at Five Angstrom Precision. *Nat. Biotechnol.* **2012**, *30*, 344−348.

(109) *Company history*. Oxford Nanopore Technologies, 2021; https://nanoporetech.com/about-us/history.

(110) Castro-Wallace, S. L.; et al. Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Sci. Rep.* **2017**, *7*, 18022.

(111) *Continuous development and improvement*. Oxford Nanopore Technologies, 2022; https://nanoporetech.com/about-us/continuous-development-and-improvement.

(112) Caldwell, C. C.; Spies, M. Helicase SPRNTing through the nanopore. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 11809−11811.

(113) Lu, H.; Giordano, F.; Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics* **2016**, *14*, 265−279.

(114) Deamer, D.; Akeson, M.; Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **2016**, *34*, 518−524.

(115) Jain, M.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338−345.

(116) Rang, F. J.; Kloosterman, W. P.; de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **2018**, *19*, 90.

(117) Wang, L.; Qu, L.; Yang, L.; Wang, Y.; Zhu, H. NanoReviser: An Error-Correction Tool for Nanopore Sequencing Based on a Deep Learning Algorithm. *Front. Genet.* **2020**, *11*, 900.

(118) Sahlin, K.; Medvedev, P. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat. Commun.* **2021**, *12*, 2.

(119) Cost per Gigabase. *41J Blog*, 2022; https://41j.com/blog/2022/09/cost-per-gigabase/.

(120) NovaSeq 6000 System|Powerful high-throughput sequencing system; https://www.illumina.com/systems/sequencing-platforms/novaseq.html.

(121) Hon, T.; et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci. Data* **2020**, *7*, 399.

(122) *MGI sequencing platforms: High-throughput gene sequencers, DNBSEQ sequencing technology*; MGI Product-MGI Tech website-Leading Life Science Innovation; https://en.mgi-tech.com/products/.

(123) Guide to NGS Platforms; https://www.biocompare.com/Editorial-Articles/590720-Guide-to-NGS-Platforms/.

(124) Cuber, P.; et al. Comparing the accuracy and efficiency of third generation sequencing technologies, Oxford Nanopore Technologies, and Pacific Biosciences, for DNA barcode sequencing applications. *Ecol. Genet. Genomics* **2023**, *28*, No. 100181.

(125) *NGS, qPCR, or Sanger Sequencing: Selection Guide*; Azenta Life Sciences, https://www.azenta.com/blog/ngs-qpcr-dpcr-or-sanger-sequencing-assay-selection-guide.

(126) Applied Biosystems. *Applied Biosystems SOLiD® 4 System Instrument Operation Guide*, 2012; https://www.yumpu.com/en/document/read/8844193/applied-biosystems-solidr-4-system-instrument-operation-guide-.

(127) Garrido-Cardenas, J. A.; Garcia-Maroto, F.; Alvarez-Bermejo, J. A.; Manzano-Agugliaro, F. DNA Sequencing Sensors: An Overview. *Sensors* **2017**, *17*, 588.

(128) Marine, R. L.; et al. Comparison of Illumina MiSeq and the Ion Torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses. *J. Virol. Methods* **2020**, *280*, No. 113865.

(129) Mullin, E. The Era of Fast, Cheap Genome Sequencing Is Here. *Wired* **2022**; https://www.wired.com/story/the-era-of-fast-cheap-genome-sequencing-is-here/.

(130) MGI Breaks the $100 genome barrier, barely. *SanDiegOmics*; https://sandiegomics.com/mgi-breaks-the-100-genome-barrier-barely/ (2023).

(131) MinION portable nanopore sequencing device. Oxford Nanopore Technologies, https://nanoporetech.com/products/sequence/minion.

(132) Pricing, https://rtsf.natsci.msu.edu/genomics/pricing.aspx.

(133) 2022 Sequencing Market Share − Same As It Ever Was (For Now). *SanDiegOmics*, 2023; https://sandiegomics.com/2022-sequencing-market-share-same-as-it-ever-was-for-now/.

(134) Illumina (ILMN) Leads the Market With 80% Share. *Yahoo Finance*, 2024; https://finance.yahoo.com/news/illumina-ilmn-leads-market-80-123127888.html.

(135) Arslan, S.; et al. Sequencing by avidity enables high accuracy with low reagent consumption. *Nat. Biotechnol.* **2024**, *42*, 132−138.

(136) Nyberg, L. K.; et al. A single-step competitive binding assay for mapping of single DNA molecules. *Biochem. Biophys. Res. Commun.* **2012**, *417*, 404−408.

(137) Rye, H. S.; et al. Stable fluorescent complexes of double-stranded DNA with bis-intercalating asymmetric cyanine dyes: properties and applications. *Nucleic Acids Res.* **1992**, *20*, 2803−2812.

(138) Nilsson, A. N.; et al. Competitive binding-based optical DNA mapping for fast identification of bacteria - multi-ligand transfer matrix theory and experimental applications on Escherichia coli. *Nucleic Acids Res.* **2014**, *42*, No. e118.

(139) Müller, V.; et al. Enzyme-free optical DNA mapping of the human genome using competitive binding. *Nucleic Acids Res.* **2019**, *47*, e89−e89.

(140) Glazer, A. N.; Rye, H. S. Stable dye−DNA intercalation complexes as reagents for high-sensitivity fluorescence detection. *Nature* **1992**, *359*, 859−861.

(141) Zimmer, C.; Marck, C.; Schneider, C.; Guschlbauer, W. Influence of nucleotide sequence on dA.dT-specific binding of Netropsin to double stranded DNA. *Nucleic Acids Res.* **1979**, *6*, 2831−2837.

(142) Bensimon, A.; et al. Alignment and Sensitive Detection of DNA by a Moving Interface. *Science* **1994**, *265*, 2096−2098.

(143) Deen, J.; et al. Combing of Genomic DNA from Droplets Containing Picograms of Material. *ACS Nano* **2015**, *9*, 809−816.

(144) Tegenfeldt, J. O.; et al. The dynamics of genomic-length DNA molecules in 100-nm channels. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 10979−10983.

(145) Müller, V.; Westerlund, F. Optical DNA mapping in nanofluidic devices: principles and applications. *Lab. Chip* **2017**, *17*, 579−590.

(146) Müller, V.; et al. Cultivation-Free Typing of Bacteria Using Optical DNA Mapping. *ACS Infect. Dis.* **2020**, *6*, 1076−1084.

(147) Michalet, X.; et al. Dynamic Molecular Combing: Stretching the Whole Human Genome for High-Resolution Studies. *Science* **1997**, *277*, 1518−1523.

(148) Herrick, J.; Bensimon, A. Introduction to molecular combing: genomics, DNA replication, and cancer. *Methods Mol. Biol. Clifton NJ.* **2009**, *521*, 71−101.

(149) Lengronne, A.; Pasero, P.; Bensimon, A.; Schwob, E. Monitoring S phase progression globally and locally using BrdU incorporation in TK+ yeast strains. *Nucleic Acids Res.* **2001**, *29*, 1433−1442.

(150) Reisner, W.; et al. Single-molecule denaturation mapping of DNA in nanofluidic channels. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 13294−13299.

(151) Welch, R. L.; Sladek, R.; Dewar, K.; Reisner, W. W. Denaturation mapping of Saccharomyces cerevisiae. *Lab. Chip* **2012**, *12*, 3314−3321.

(152) Wang, X.; Jordan, I. K.; Mayer, L. W. A Phylogenetic Perspective on Molecular Epidemiology. In *Molecular Medical Microbiology*, 2nd ed.; Tang, Y.-W., Sussman, M., Liu, D., Poxton, I., Schwartzman, J., Eds.; Academic Press: Boston, MA, 2015; Chapter 29, pp 517−536; DOI: 10.1016/B978-0-12-397169-2.00029-9.

(153) Schwartz, D. C.; et al. Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping. *Science* **1993**, *262*, 110−114.

(154) Samad, A.; Huff, E. F.; Cai, W.; Schwartz, D. C. Optical mapping: a novel, single-molecule approach to genomic analysis. *Genome Res.* **1995**, *5*, 1−4.

(155) Yokota, H.; et al. A new method for straightening DNA molecules for optical restriction mapping. *Nucleic Acids Res.* **1997**, *25*, 1064−1070.

(156) Dimalanta, E. T.; et al. A Microfluidic System for Large DNA Molecule Arrays. *Anal. Chem.* **2004**, *76*, 5293−5301.

(157) Riehn, R.; et al. Restriction mapping in nanofluidic devices. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10012−10016.

(158) Gupta, A.; Kounovsky-Shafer, K. L.; Ravindran, P.; Schwartz, D. C. Optical mapping and nanocoding approaches to whole-genome analysis. *Microfluid. Nanofluidics* **2016**, *20*, 44.

(159) Zhou, S.; et al. A Whole-Genome Shotgun Optical Map of Yersinia pestis Strain KIM. *Appl. Environ. Microbiol.* **2002**, *68*, 6321−6331.

(160) Valouev, A.; Schwartz, D. C.; Zhou, S.; Waterman, M. S. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15770−15775.

(161) Neely, R. K.; et al. DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. *Chem. Sci.* **2010**, *1*, 453.

(162) Maghini, D. G.; Moss, E. L.; Vance, S. E.; Bhatt, A. S. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nat. Protoc.* **2021**, *16*, 458−471.

(163) Zhang, Y.; et al. DNA Extraction: A Simple Thermoplastic Substrate Containing Hierarchical Silica Lamellae for High-Molecular-Weight DNA Extraction (Adv. Mater. 48/2016). *Adv. Mater.* **2016**, *28*, 10810−10810.

(164) Rigby, P. W.; Dieckmann, M.; Rhodes, C.; Berg, P. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *J. Mol. Biol.* **1977**, *113*, 237−251.

(165) Xiao, M.; et al. Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Res.* **2007**, *35*, No. e16.

(166) Lam, E. T. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **2012**, *30*, 771.

(167) Wei, H.; et al. Mapping the nicking efficiencies of nickase R.BbvCI for side-specific LNA-substituted substrates using rolling circle amplification. *Sci. Rep.* **2016**, *6*, 32560.

(168) Das, S. K.; et al. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res.* **2010**, *38*, No. e177.

(169) Neely, R. K.; Deen, J.; Hofkens, J. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers* **2011**, *95*, 298−311.

(170) McCaffrey, J.; et al. CRISPR-CAS9 D10A nickase target-specific fluorescent labeling of double strand DNA for whole genome mapping and structural variation analysis. *Nucleic Acids Res.* **2016**, *44*, No. e11.

(171) Ran, F. A.; et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **2013**, *8*, 2281−2308.

(172) Pignot, M.; Siethoff, C.; Linscheid, M.; Weinhold, E. Coupling of a Nucleoside with DNA by a Methyltransferase. *Angew. Chem., Int. Ed.* **1998**, *37*, 2888−2891.

(173) Buryanov, Y.; Shevchuk, T. The use of prokaryotic DNA methyltransferases as experimental and analytical tools in modern biology. *Anal. Biochem.* **2005**, *338*, 1−11.

(174) Sánchez-Romero, M. A.; Cota, I.; Casadesús, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr. Opin. Microbiol.* **2015**, *25*, 9−16.

(175) Goedecke, K.; Pignot, M.; Goody, R. S.; Scheidig, A. J.; Weinhold, E. Structure of the N6-adenine DNA methyltransferase M•TaqI in complex with DNA and a cofactor analog. *Nat. Struct. Biol.* **2001**, *8*, 121−125.

(176) Cheng, X. AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res.* **2001**, *29*, 3784−3795.

(177) Dalhoff, C.; Lukinavicius, G.; Klimasauskas, S.; Weinhold, E. Direct transfer of extended groups from synthetic cofactors by DNA methyltransferases. *Nat. Chem. Biol.* **2006**, *2*, 31−32.

(178) Goyvaerts, V.; et al. Fluorescent SAM analogues for methyltransferase based DNA labeling. *Chem. Commun.* **2020**, *56*, 3317−3320.

(179) Bouwens, A.; et al. Identifying microbial species by single-molecule DNA optical mapping and resampling statistics. *NAR Genomics Bioinforma.* **2020**, *2*, No. lqz007.

(180) D'Huys, L.; et al. Assessing the Resolution of Methyltransferase-Mediated DNA Optical Mapping. *ACS Omega* **2021**, *6*, 21276−21283.

(181) Ruppeka-Rupeika, E.; et al. Optical Mapping: Detecting Genomic Resistance Cassettes in MRSA. *ACS Omega* **2024**, *9*, 8862.

(182) Pang, A. W. C.; et al. Analytic Validation of Optical Genome Mapping in Hematological Malignancies. *Biomedicines* **2023**, *11*, 3263.

(183) Wight, D. J.; et al. Unbiased optical mapping of telomere-integrated endogenous human herpesvirus 6. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 31410−31416.

(184) Buko, T.; Tuczko, N.; Ishikawa, T. DNA Data Storage. *BioTech* **2023**, *12*, 44.

## ■ NOTE ADDED AFTER ASAP PUBLICATION

This review published ASAP on October 25, 2024. The graphics in the paper have been updated and the corrected version reposted on October 30, 2024.