# Third-generation-sequencing technologies: sequencing individual molecules

## The need to synchronize reactions across molecules limits the speed and read-length of second-generation sequencing technologies

The newest implementation of the sequencing-by-synthesis technology requires approximately 1 hour to add one additional base to each of the molecules in the flow cell. This is very slow[1] in comparison to the 1000-bases-per-second (=3'600'000 bp/h) speed, with which a DNA polymerase can incorporate nucleotides under optimal conditions.

This slowness of the sequencing-by-synthesis reaction cycle is ultimately owed to the need to maintain perfect synchrony between the reactions of the several thousand identical molecules in each of the clusters inside the flow cell. Without this synchrony, the fluorescent signal from the different molecules in a cluster would be out of register and become uninterpretable. The sequencing-by-synthesis reaction therefore had to be designed such that each reaction is prevented from progressing until all molecules in the flow cell have completed that reaction step. This requires a) a substantial waiting time at each reaction step to ensure all molecules have completed the reaction and b) time-consuming flushing-in and washing-out of the reagents that are needed for each individual step in the reaction cycle.

The requirement for synchrony between all DNA strands in a cluster also limits the maximal read-length that can be achieved with the sequencing-by-synthesis reaction. This is because, despite all efforts, there remains a finite chance that one of the molecules in the cluster will skip one of the reaction steps. This then puts that molecule permanently out of step with the other molecules in the cluster. Initially this is not a problem, because the signal from the other molecules in the cluster will overpower the signal from those few molecules that have fallen out-of lockstep. But, since there is no way for the molecules that have fallen out of lockstep to get back into lockstep, the fraction of out-of-lockstep molecules increases with every cycle. Eventually, after 200-300 reaction cycles so many molecules of the cluster will be out of lockstep that their random fluorescence signal drowns out the signal from those molecules that have remained in lockstep.

Third-generation sequencing technologies seek to sidestep these synchrony-based problems by performing the sequencing reaction on individual molecules.

## Pacific Biosciences (PacBio): Third-generation fluorescence-based single-molecule real-time (SMRT) sequencing

The conceptual advantages of performing a sequencing-by-synthesis-style reaction on individual DNA molecules are profound. Because there is no need to synchronize the reaction of multiple molecules in a cluster, the reaction no longer has to be stopped and reagents do not need to be exchanged at each step of the reaction cycle. Instead, all necessary reagents are added to the reaction mix and nucleotides are observed in real-time as they are added, one after another, to the growing DNA strand (figure 1). This simplifies the experimental protocol and has the potential to greatly speed up the sequencing reaction. The single-molecule approach also removes the synchronicity-imposed limits on read length and thus makes much longer reads possible.

[1] Remember, the tremendous overall performance of the sequencing-by-synthesis technology does not come from the speed of the individual reaction but from sequencing billions of DNA strands in parallel.

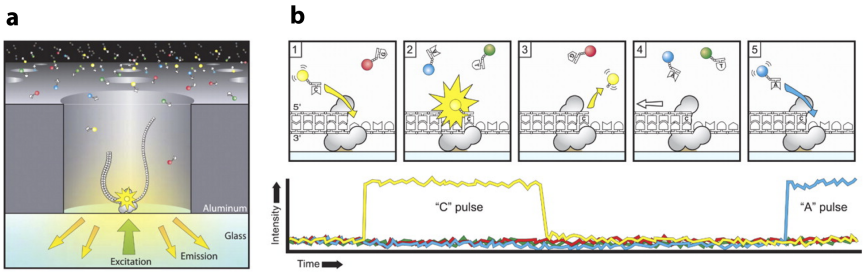


**Figure 1:** Principle of the PacBio single-molecule real-time sequencing technology. (a) Individual DNA polymerase molecules are immobilized at the bottom of 100 nm diameter nanowells etched into the bottom of the reaction chamber. (b) Individual fluorescently labeled nucleotides diffuse in and out of these nanowells and may be incorporated into the growing strand when they are complementary to the next unpaired nucleotide on the template strand. During the time between successful binding and the formation of the phosphodiester bond, the gamma phosphate-linked fluorophore is excited and emits an optical signal that is picked up by the instrument's zero-mode waveguide (ZMW) optics. The color of the fluorescence (trace bottom right) reveals which base is incorporated (Figure source: Eid et al., Science 2016)

## Optical monitoring of extremely small reaction volumes makes single-molecule sequencing possible

The challenge for such a single-molecule approach is that the signal that can be obtained from a single fluorophore molecule is limited[2]. Also, the fluorescence background in the sample is substantial, in particular given that the reaction mix contains a large amount of unincorporated fluorescently labeled nucleotides. The task for the scientists and engineers at PacBio, where the first viable single-molecule sequencing technique was developed, was clear: boost the signal obtained from the single fluorescently labeled nucleotide that is being incorporated into the chain while minimizing the fluorescence background from the other fluorescently labeled nucleotides diffusing in the solution.

The basic principle behind the solution implemented in the PacBio method is to monitor the fluorescence only in the extremely small volume of a few tenths of zeptoliters (1 zepto liter = $10^{-21}$ liters)[3] that immediately surrounds the polymerase.

Because the monitored reaction volume is so small, most of the time, it contains no nucleotide molecule at all - even at the micromolar nucleotide concentrations used in a typical DNA-synthesis reaction. The fluorescently labeled nucleotides (figure 2) will briefly diffuse into and out of the monitored volume on the microsecond time scale. A nucleotide might even sample the polymerase active site, but if it is not complementary to the template strand, it will detach and diffuse away again within a millisecond. Only when the nucleotide matches the base on the template strand will it bind tightly to the polymerase active site and remain in the monitored reaction volume long enough (approximately 100 milliseconds)[4] to generate a prolonged fluorescent signal. This signal can be used

[2]Boosting the fluorescent signal was the very reason why the sequencing-by-synthesis technology used clusters of identical molecules instead of individual molecules.

[3]The miniscule (<50x50x50 $nm^2$) optical-monitoring volumes required by this technique are far smaller than can be achieved with conventional diffraction limit optics. The combination of the well size, which is much smaller than the wavelength of light, with the special chip material causes a special optical configuration called a zero-mode waveguide (ZMW). Light impinging on a ZMW is not able to traverse the chip, so no light leaks into solution above the well. But, the electromagnetic field generated by the light can excite fluorescent molecules, which are located less then 50 nanometers from the bottom of the well. And, photons emitted by those fluorescent molecules can also "escape" through this ZMW to be detected by the instrument's optics.

[4]The DNA polymerase used in the PacBio system is substantially slower than the fastest natural polymerases. This gives the detection system time to pick up the fluorescent signal.

to determine the identity of the base that is being integrated. The fluorescence signal stops again when the phosphodiester-backbone bond is formed and the released fluorophore diffuses away.

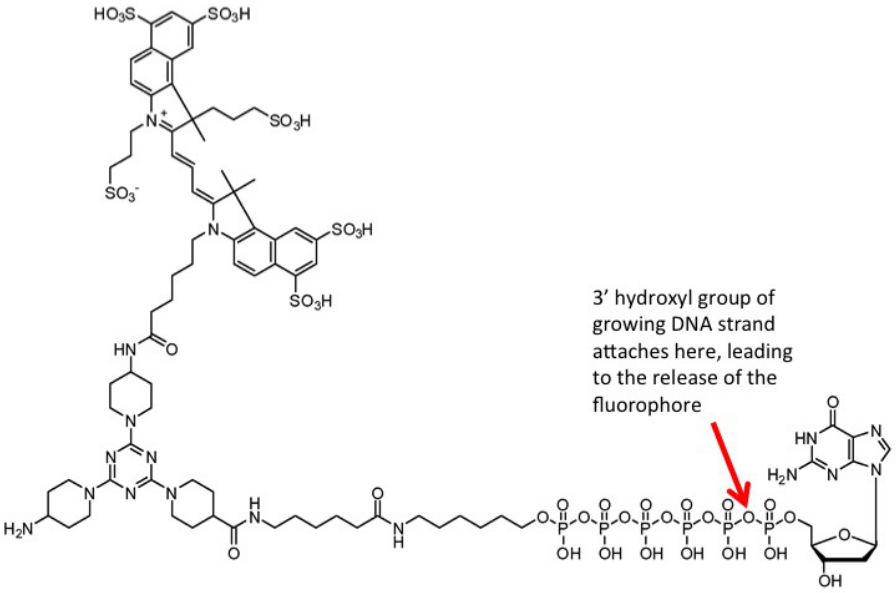


**Figure 2**: Molecular structure of a fluorescently labeled guanine nucleotide analog for PacBio single-molecule sequencing reactions. Note the attachment of the nucleotide to the terminal hexa(!)-phosphate group. As a result of this attachment method, the fluorophore is released when the nucleotide is incorporated into the growing DNA strand. Also note the long linker and large central triazin-based molecular scaffold (bottom left), which helps place the fluorescent moiety (top left) away from the polymerase surface. This placement reduces the chance for fluorescence quenching and photochemical reactions between the fluorophore and the polymerase that would destroy the fluorophore. The result is an increase in the strength of the fluorescent signal obtainable from a single fluorophore and a reduced chance to inflict chemical damage on the polymerase.

## The PacBio technology outperforms sequencing-by-synthesis in terms of reaction speed but its lower degree of parallelization limits the throughput

Following formation of the phosphodiester bond, the polymerase moves the DNA strand to the next nucleotide position and is immediately ready to integrate the next nucleotide. As a result, the PacBio technology adds several nucleotides per second to the growing strand. This is a 10'000-fold improvement in the speed of DNA synthesis compared to the sequencing-by-synthesis method.

However, the highly sophisticated optical setup necessary for the PacBio technology allows parallelization of "only" a few thousand sequencing reactions (compared to the billions of parallel reactions possible with the sequencing-by-synthesis technology). As a result, the PacBio technology still lags behind the sequencing-by-synthesis technology in terms of overall throughput (i.e., bases sequenced per day per instrument).

## PacBio can generate continuous reads of >60'000 bases, but weak signals from single-molecule detection increase error rates

The advantage of the PacBio system then does not lie in the absolute throughput of the technology, but in the other advantage offered by single-molecule sequencing: the achievable read length. Because PacBio sequencing is a single-molecule technique, it completely sidesteps the synchronization problems that limit the read length of the sequencing by synthesis to 300 bases.

In principle, the PacBio sequencing reaction can continue indefinitely and the signal quality will remain identical throughout the entire read. In practical terms, read-length appears to be limited by very rare (1 per 15'000 bases) photochemical reactions, in which a photoexcited fluorophore reacts with and destroys the polymerase.

Because the photochemical damage events occur randomly, the distribution of read length from a PacBio experiment will be centered around 10'000 bases, but some reads will extend for more than 60'000 bases. These very long reads are particularly valuable for the assembly of difficult regions of a sequenced genome (e.g., regions containing repeats or gene duplications).

The downside of single-molecule techniques such as PacBio's is that they are inherently noisy. Figure 3 shows a short section of a fluorescence-intensity time trace recorded from a PacBio nanowell. Because phosphodiester-bond formation during DNA synthesis is a stochastic process, it will sometimes be fast (short pulse) or slow (long pulse). The interpretation of these time-traces is substantially more error-prone than for the sequencing-by-synthesis method. As a result, the error rate currently lies at around 10% with false insertions and deletions of individual nucleotides being the most common error type.

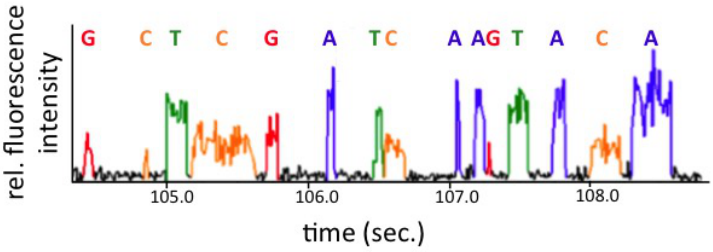


**Figure 3:** Fluorescence time trace from a single nanowell of a PacBio sequencer. The individual peaks correspond to the fluorescent signal from a single nucleotide being integrated into the growing DNA strand. Note the large variation in the duration from peak to peak. The section shown here corresponds to ~4 seconds. A full trace for a 30'000 nucleotide read has a duration of ~2.5 hours.

## PacBio sequencing is most often employed to provide a scaffold for the assembly of short-read sequencing data generated by other techniques

Pacific Biosciences is working on various strategies for reducing the error rate of the PacBio sequencing technologies, but it is unlikely that it will be able to reach the error rates of Sanger or SBS sequencing any time soon. Currently, the main use of PacBio sequencing data therefore lies in providing a long and continuous (albeit faulty) sequence template onto which shorter, lower-error reads from other techniques such as sequencing by synthesis can be assembled.

## The case for moving beyond fluorophores and polymerases to sequence DNA

The sequencing methods discussed so far are all based on the same core technologies: DNA polymerases and fluorescently labeled nucleotides. Evidently, this combination has been very successful so far. What could be the motivation for changing this "winning team"?

As it turns out, both fluorescent nucleotides and DNA polymerases face fundamental limitations and the sequencing technologies based on them are rapidly approaching these limits.

In the case of DNA polymerase, the most obvious limitation is the enzymatic rate, with which individual polymerases can incorporate nucleotides into the growing strand. The DNA polymerases used in sequencing reactions are derived from natural polymerases that have undergone billions of years of evolutionary optimization to achieve a maximal reaction speed of 1000 bases per second. It therefore seems unlikely that further optimization by protein engineering will yield order-of-magnitude performance gains.

Likewise, fluorescence-based technologies pose fundamental limitations such as the spatial resolution that can be achieved with optical techniques. The PacBio technology is already pushing the spatial resolution of optical detection technology to its very limit. Another factor limiting the fluorescent signal is the number of excitation-emission cycles a fluorophore molecule can complete in a given time period without being destroyed by photochemical damage. This limits the speed and accuracy with which individual bases can be identified.

However, the greatest motivation for moving away from fluorescence-based detection systems appears to be the very high cost of the fluorescently labeled nucleotides. These molecules are extremely expensive to produce and in fluorescence-based sequencing, there is a fundamental limit of one fluorescently-labeled nucleotide that is consumed for each sequenced base.

## Nanopore technology: reading DNA sequences directly

The company Oxford Nanopore Technologies has developed the first commercially available DNA-sequencing technology that requires neither DNA synthesis nor fluorescently labeled nucleotides. Instead, Nanopore's technology reads the sequence of a DNA strand directly by passing it through a narrow protein pore in a membrane and measuring how the sequence of the passing DNA molecule influences the ionic current through this pore (figure 4).

The company guards many of the technical details of the sequencing technology as a trade secret, but the pore being used appears to be a genetically modified version of a porin protein. Porins are proteins that span the outer membrane of many bacteria where they permit the rapid, non-specific diffusion of hydrophilic small molecules across the membrane. The natural function of porins has nothing to do with DNA translocation. Porins simply happen to have a central pore that has just the right diameter (approx. 1nm) to allow a single strand of DNA to squeeze through.

In the Nanopore measurement cell, a single porin molecule is embedded into a polymer-based membrane that separates two chambers containing electrolyte solutions. By applying a small electrical voltage between these two chambers, the electrolyte ions flow through the pore and generate an electrical current that can be measured. Because the DNA molecule is also charged, it is pulled through the pore from one chamber to the other. By doing so, it restricts the flow of the electrolyte ions passing through the pore, which is reflected in the electrical current measured between the two chambers.

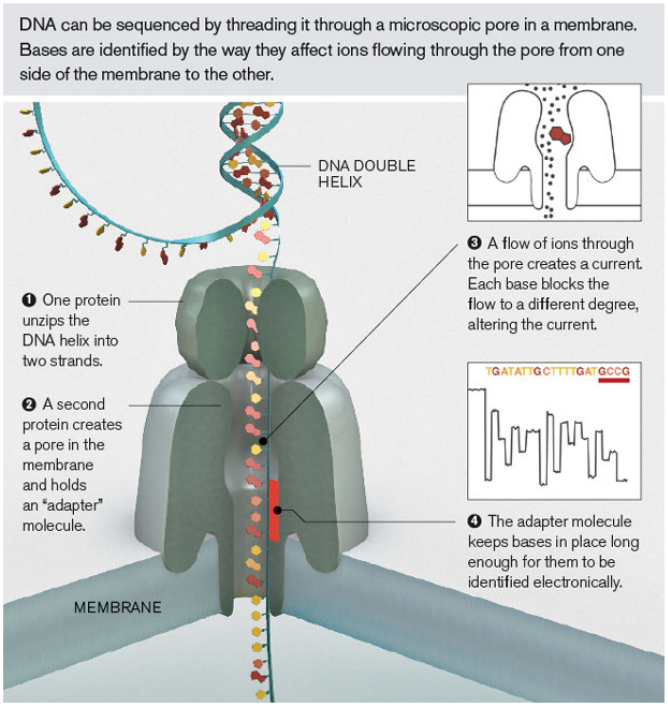


**Figure 4:** Operating principle of the Nanopore sequencing technology. A protein nanopore is embedded in a membrane that separates two chambers containing electrolyte solutions. A voltage applied across this membrane pulls ions and the DNA strand through the pore. The presence of the DNA in the pore restricts the flow of electrolyte ions in a way that depends on the base sequence of the DNA section that is currently traversing the pore. By measuring and analyzing the electrolyte current over time, the sequence of the DNA passing through the pore can be determined. (Figure source A. Schaffer, 2012, MIT Technology Review)

Due to their slightly different size, shape and polarity, the different bases have a slightly different effect on the electrolyte current. This change in the current can be measured and gives information about the base sequence of the DNA strand passing through the pore. The challenge in interpreting this signal is that this current not only depends on the base itself, but also on the identity of the neighboring bases. For example, a G in the context of the sequence CT**G**AC will give a different signal than a G in the context of the sequence TC**G**CC. This requires a rather sophisticated base-calling algorithm and makes it difficult to assess the quality of an individual base call.

## Nanopore sequencing technology provides fast, ultra-long sequence reads in a small portable instrument

Just as the PacBio technology, the Nanopore techology has no fundamental read-length limit and the quality of the reads also stays constant over the entire read length. Because the technology is still very new, reliable performance indicators are not yet available. Read lengths of up to 200'000 bases have been reported and the error rate appears to be comparable to (or slightly higher) than the 10% achieved by the PacBio technology.

The area in which the Nanopore technology really excels is the speed with which individual DNA molecules can be read. As of fall 2016, this speed is at 500 bases per second - already very close to the fundamental speed limit of DNA-synthesis-based sequencing technologies.

The other advantage of the Nanopore technology lies in the way the sequence signal is read out. In the Nanopore technology, the signal is the electrical current generated by the ions passing through the pore alongside the DNA. Thanks to decades of research and development by the computer industry, semi-conductor-based technologies for measuring electrical currents have become incredibly precise, cheap and miniaturized. Using this technology, the whole readout process in the Nanopore instrument takes place on a cheap-to-manufacture semi-conductor chip - none of the lasers, fiberoptics, optical grating etc. of the fluorescence based instruments are needed. This means that the technology for a Nanopore sequencer can be packed into a device the size of a candy-bar (figure 5a) - much smaller than the kitchen-sized enclosure needed to house the newest PacBio instrument (figure 5b).

## Outlook on DNA sequencing

In less than 40 years, DNA sequencing has moved from an experiment that any molecular biologist could perform him/herself with readily available reagents and self-built equipment into a highly sophisticated, specialized, and lucrative industry. The performance gains generated in this process are astounding. For example, the speed of DNA-sequencing technologies dramatically outpaces other rapidly developing technologies such as computing.

For the moment, no end of this trend is in sight. New technologies are being introduced every few years and existing technologies continue to be improved. It is foreseeable that the cost for generating the raw sequencing data will become negligible compared to the cost of sample collection.

The next frontier for DNA sequencing will then lie in developing rapid, cheap, and reliable techniques for converting the raw sequencing data into continuous sequences that can be read from one end of each chromosome to the other. It may turn out that this will be a harder nut to crack than generating the sequencing data.

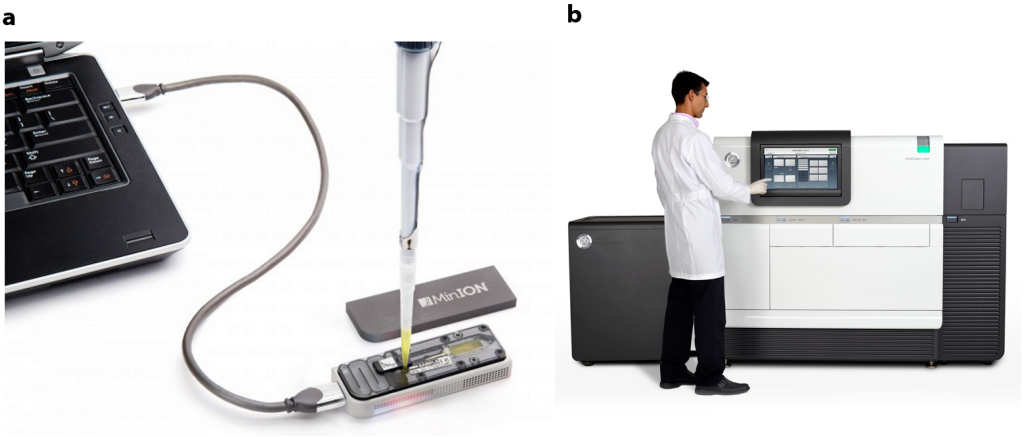


**Figure 5:** Size comparison between a Nanopore and a PacBio sequencing instrument. While the comparison is not entirely fair (the PacBio instrument performs many more sequencing reactions in parallel), it nevertheless demonstrates the miniaturization gain that can be achieved by moving beyond the traditional DNA-synthesis and fluorescence-based technologies.