Second-generation sequencing technologies: tremendous performance gains through parallelization and miniaturization

A rapid succession of sequencing technologies has lead to a dramatic boost in sequencing performance and options

From the early 2000's, several new sequencing technologies became commercially available, each representing a substantial leap in performance over even the most advanced versions of Sanger sequencing. At the time, this group of technologies was referred to as "next-generation sequencing" (NGS). As these technologies are now about to be surpassed again by yet another generation of technologies, the term "next-generation" sequencing is becoming ambiguous.

To avoid any confusion, the various implementations of the Sanger method are now referred to as first-generation sequencing technologies. The technologies previously referred to as next-generation sequencing are now called second-generation technologies and the technologies, which are just starting to become available now, are called third-generation technologies.

Sometimes the assignment of a specific technology to the 1st, 2nd or 3rd generation is contentious, but one typically distinguishes between the three generations based on the following criteria.

First-generation technologies are marked by a clear separation of the biochemical sequencing reaction and the read-out process.

In **second-generation technologies**, this separation no longer exists. Instead, the biochemical reaction and the read-out are part of one integrated process that takes place in the same reaction vessel. Another feature shared by second-generation sequencing technologies is the need for an amplification step, in which individual molecules from the sequencing library are multiplied in a PCR-like process prior to the actual sequencing reaction.

Third-generation technologies have found ways to dispense with this amplification step and manage to sequence individual molecules from the sequencing libraries directly, thus avoiding potential pitfalls from this amplification step and making much longer sequence reads possible.

Sequencing by synthesis is the dominant second-generation sequencing technology

Among the second-generation technologies the sequencing-by-synthesis (SBS) method developed by the company Illumina is particularly popular and currently dominates the market for DNA sequencing. The introduction of this technology has increased the speed and reduced the cost of sequencing by a factor of 10'000!

Just as the Sanger method, the sequencing-by-synthesis method is based on a sequencing reaction that resembles the natural DNA-synthesis process and uses a primer, a polymerase, a template strand, and nucleotides for the synthesis of a new DNA strand. The main difference between the Sanger and SBS reactions is in the type of nucleotides that are used. In contrast to Sanger sequencing where a mix of natural and modified nucleotides is employed, SBS uses only modified nucleotides (see figure 1 for an example). Additionally, the nucleotide analogs used in Sanger and SBS sequencing are distinguished by the fact that the functional groups of the nucleotides (i.e., the fluorophore, which indicates the type of nucleotide and the terminator, which prevents the attachment of the subsequent nucleotide) are permanently attached for the Sanger nucleotides but can be chemically cleaved in the case of the SBS nucleotides.

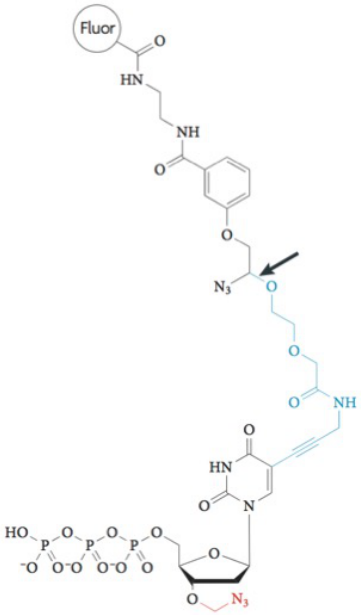

Figure 1: Chemical structure of a fluorescently labeled, chain-terminating dCTP analog used in the sequencing-by-synthesis method. During synthesis, the azide group shown in red blocks the 3'-hydroxyl group and prevents the incorporation of additional nucleotides into the growing strand. This group can be chemically removed to generate a 3'-hydroxyl group, which then permits incorporation of the next nucleotide. The moiety in blue is a chemically cleavable linker that attaches the fluorophore to the nucleotide. Each type of base (A, C, T, or G) carries a fluorophore of a different color.

The basic biochemical process of sequencing by synthesis is relatively simple (figure 2a). Template molecules are attached to a solid surface in a flow cell, a primer is annealed to the template strand, and a polymerase is bound. The individual cycle of the sequencing-by-synthesis reaction then proceeds as follows.

1. The flow cell is flooded with a solution containing the four different types of fluorescently labeled, chain-terminating nucleotides.
2. The polymerase uses the template strand as a guide to extend the primer with the complementary nucleotide. The terminating group on the newly incorporated nucleotide prevents the incorporation of additional nucleotides.
3. Any non-incorporated nucleotides are washed away.
4. The fluorescence signal of each template cluster of the sample is measured by fluorescence microscopy and reveals, which of the four nucleotides has just been incorporated.
5. Both the fluorophore and the terminating group are removed by chemical cleavage and washed away. This prepares the reaction for the next cycle.

The sequencing-by-synthesis (SBS) approach is amenable to miniaturization and massive parallelization

At first sight, the sequencing-by-synthesis reaction described above does not seem to represent much of a technological leap forward. The SBS reaction shares many commonalities with the Sanger reaction (use of a polymerase, fluorescently labeled nucleotides etc.). On the basis of an individual reaction, the SBS method is actually slightly inferior to the Sanger method on a number of parameters (speed of base incorporation, read length, error rate etc.).

How then does SBS achieve its tremendous boost in performance relative to Sanger sequencing? The answer is that the SBS method can be parallelized and miniaturized. Instead of improving the speed or accuracy of the individual reaction, SBS makes it possible to perform billions (!!!) of parallel sequencing reactions in a flow cell the size of a microscope slide (figure 2b).

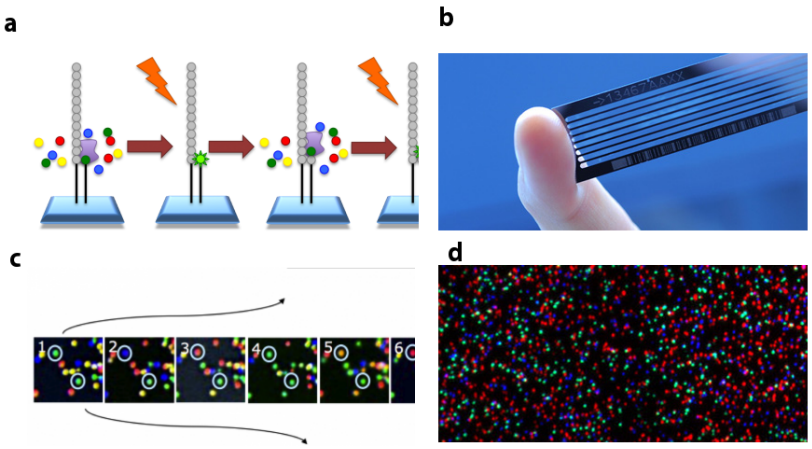

Figure 2: Sequencing by synthesis achieves its throughput through miniaturization and massive parallelization of a simple sequencing reaction. a) Three cycles of a sequencing-by-synthesis reaction. The polymerase (purple) extends the primer by a single nucleotide. Unincorporated nucleotides are washed away and the fluorescence signal is detected. The fluorophore and the chain-terminating group are cleaved off and washed away. Then the cycle is repeated. b) Flow cell used in sequencing-by-synthesis instruments. The DNA strands undergoing the sequencing-by-synthesis reaction are chemically attached to the interior surface of the flow cell's transparent channels. c) Reading a sequence from a series of fluorescence images. Each colored dot corresponds to one sequence cluster of DNA molecules from the sequencing library. The images are recorded after each cycle of the sequencing reaction. d) Zoomed-up view of a small portion of a flow cell surface. Each dot corresponds to a different DNA molecule undergoing sequencing. The highest-performing SBS instruments are able to sequence several billion molecules in a single flow cell.

The key change enabling this performance boost of the SBS technology is that the DNA fragments being sequenced are attached to a solid surface (figure 2a). This means that individual reactions can be identified by their xy-coordinates. This makes it possible to perform many individual sequencing reactions side-by-side. Further, integrating the biochemical reaction steps and the physical read-out of the sequence signal into one continuous process means that the sequence information can be read

out while the molecules stay attached to the surface. No physical transfer of reaction products from one instrument to another is necessary. The SBS instrument simply has to record a high-resolution image of the flow cell's surface after each reaction cycle and the series of color changes at a given xy-coordinate reveals the sequence of the corresponding DNA molecule.

Sequencing by synthesis employs bridge amplification to generate clusters of identical molecules to yield a detectable optical signal

The above description of SBS may have given the impression that each of the colored dots visible on the flow cell surface (figure 2c, d) corresponds to a single DNA molecule. In fact, this is not the case. In the optical setup used in SBS instruments, the signal from a single fluorophore molecule would be too weak to be detected reliably. Therefore, before the actual sequencing-by-synthesis reaction can be performed, the individual DNA molecules bound to the flow cell need to be amplified in a PCR-like process called **bridge amplification** (figure 3). During bridge amplification, one end of the DNA molecules always remains attached to the flow-cell surface. A few cycles of this reaction therefore generates a tight cluster of approximately one thousand copies of the initial DNA molecule. During the subsequent sequencing reaction the joint fluorescent signal from this cluster of molecules then has sufficient strength for reliable detection.

A big part of the current efforts to improve the SBS technology further is dedicated to optimizing this cluster generation process. The goal is to fit as many clusters on the surface as possible, but to still maintain the clear optical separation between neighboring clusters that is necessary for unambiguous signal read out.

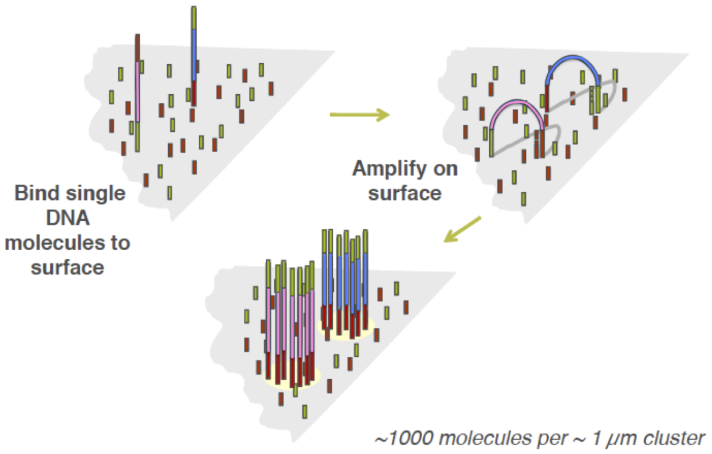

Figure 3: Bridge amplification (a PCR-like process) is used to amplify individual surface-bound DNA molecules into clusters of sequence-identical molecules. Adapter sequences (red and green), that were attached to the ends of the initial DNA molecules (pink and blue) during library generation, bind complementary primers that are attached to the flow cell's surface. The primers are extended using a DNA-polymerase reaction. Heat denaturation separates the newly formed double-stranded DNA molecules and the amplification reaction can be repeated. The process converts each of the initial DNA molecules into clusters of ~1000 identical, single-stranded DNA molecules ready for the sequencing-by-synthesis reaction.

Adapter sequences enable surface attachment and ID-tagging of DNA fragments

In the discussions above, we have seen that the sequencing-by-synthesis methodology uses primers to initiate the sequencing reaction and selectively attach and release the ends of the DNA pieces to the surface of the flow cell via complementary short DNA molecules (oligos) etc. All of these processes are based on sequence complementarity between the oligos and primers on the one hand and the DNA pieces to be sequenced on the other hand. Yet, all these DNA pieces have different sequences and these sequences are unknown. How does this work?

The solution is synthetic adapter oligos (figure 4), which are enzymatically attached to the ends of the unknown DNA sequences. These adapter oligos provide stretches of sequences that are identical (with the exception of the ID tags) for each of the DNA molecules in the sample so that the same sequencing primers and surface-attachment oligos can be used for each DNA molecule.

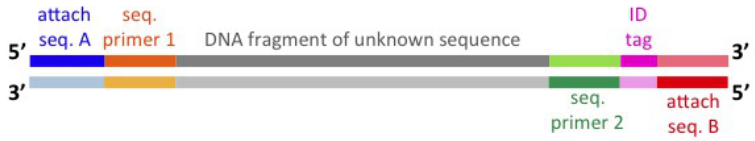

Figure 4: Schematic illustration of a DNA molecule targeted for sequencing (gray) after attachment of adapter oligos (colored regions). The resulting DNA molecule is now ready to be loaded into the flow cell for cluster amplification and sequencing. The attachment sequences A and B are used to attach the molecule to complementary oligos immobilized on the flow-cell surface and to facilitate bridge amplification. The sequencing-primer regions 1 and 2 are complementary to the sequencing primers used to sequence the two ends of the unknown DNA fragment. The ID-tag facilitates the identification of the sample, from which the unknown DNA fragment was derives. This makes it possible to sequence DNA from different samples (e.g., from different patients) in the same flow cell. The different regions are not drawn to scale. The unknown DNA sequence is typically 500-1000 bp in length while the sequencing primer, surface attachment, and ID-tag sequences have a length of ~25, ~30 and 8 bp, respectively.

A step-by-step outline of the workflow for sequencing by synthesis

The order, in which the various aspects of the sequencing-by-synthesis method were presented above, was chosen to reflect the internal logic of this method. This mode of presentation may make it difficult to understand the actual workflow for performing sequencing-by-synthesis in the laboratory.

Here is a list of the individual steps involved in DNA sequencing in the chronological order, in which they are performed in the lab.

- Tens of thousands of copies of the complete DNA to be sequenced are isolated from a specimen. For a human genome this corresponds to 5 micrograms of DNA extracted from a tissue sample of roughly one cubic millimeter in size.
- The DNA is mechanically sheared to yield DNA fragments with a typical length of 500-1000 bp.

- The DNA fragments are ligated to adapter oligos containing ID-tag, surface-attachment, and primer-complementary regions.
- Library molecules are bound to the flow-cell surface so that individual molecules are physically well separated from one another.
- Individual library molecules are amplified via bridge amplification to generate clusters of identical molecules.
- 150-300 sequencing-by-synthesis cycles are performed to determine the sequence of the DNA fragments and any included ID-tag sequences.

Optional steps for paired-end sequencing:

- The newly synthesized strand is released and the opposite end of the template is attached to the surface.
- Sequencing-by-synthesis is repeated to obtain the sequence of the opposite end of the fragments.

The sequencing-by-synthesis is constantly improved and adapted to new applications

As was the case with the Sanger technology, the sequencing-by-synthesis technology is undergoing continuous improvements and is being adapted to an ever-expanding range of applications. The details of the individual steps in the various implementations of the technique may therefore vary considerably from the ones outlined above, but the underlying principles are conserved.

The performance characteristics of the newest Illumina sequencing-by-synthesis instruments are very impressive. Several billion individual DNA fragments can be sequenced in parallel with read lengths of up to 300 bases and an error rate of approximately 0.1%.