Genome sequencing

Introduction

Forward genetics is centered on phenotype-driven gene discovery. The approach consists of identifying a phenotype of interest in mutants derived by mutagenesis, and then determining the gene that is mutated and responsible for the phenotype. Prior to the introduction of technologies for manipulating the genome of an organism, this was the only genetic strategy available.

Reverse genetics, on the contrary, is a gene-driven approach. It starts with a cloned segment of DNA, or a sequence, which is used to introduce programmed mutations back into the genome to investigate gene product and function. This strategy relies on engineering the genome through adding, modifying, and replacing genes. A prerequisite for reverse genetics is the knowledge of the DNA sequence of an organism. Here, we will discuss different sequencing techniques and their applications.

Obtaining sequence information

In order to alter a genome and induce directed mutations, we need to know the relevant sequence information. The genome sequence of many organisms is now available, since the costs for sequencing continuously drop due to the development of more and more efficient sequencing techniques. Complete sequencing of a genome can nowadays be achieved within a few days. However, a big challenge still remains "reading" the genome, e.g., understanding which genome sequences contain biologically relevant information and to decipher what is the role of the other "junk" DNA to which no function has been assigned so far (functional genomics).

Here, we will discuss different sequencing techniques, starting with Sanger sequencing as the base and the golden standard. Then, we will introduce next generation sequencing (NGS) approaches, e.g., the Illumina technique that allows high-throughput sequencing, and finally discuss two examples of third-generation sequencing techniques that do not require DNA amplification any longer and capture the sequencing signal in real time. In this lesson, we will present the basic principles and differences between the methods. In the lecture, you will learn how these techniques are applied to answer specific research questions and discuss why one technique may be better suited to answer one question than the other.

The independence of information and physical representation in DNA

Before jumping into the details of the experimental methods of how to read the information stored in DNA, let us step back and consider the relationship between DNA as a physical object and the information contained in it. We know that DNA stores information in the sequence of the bases along its backbone. Remarkably, this base sequence has almost no influence on the physical properties of a DNA molecule or on the speed, efficiency, and accuracy, with which DNA enzymes such as DNA polymerase process it. This uncoupling of information from the physical properties of the storage medium has two important consequences:

1. The same DNA-sequencing methods can be applied to virtually any piece of DNA, regardless of the specific information stored in it. A DNA strand encoding the amino-acid sequence of a protein can be read just as a DNA strand encoding a promoter region of a gene or even a completely random sequence.
2. Determining the base sequence of a piece of DNA captures the most biological information about this piece of DNA.

The second, somewhat subtle point becomes clear when we contrast DNA with proteins. In the latter case, the sequence of a protein molecule is of course also relevant, but what really matters for understanding the biological function of a protein is not the amino-acid sequence, but the physical and chemical properties of its folded three-dimensional structure. By contrast, in the case of DNA, the information stored in the base sequence is of primary interest while the physical properties and structure of this piece of DNA convey little additional information.

Sanger sequencing: the first reliable and universal method for sequencing DNA

Sanger sequencing is named after its inventor Frederick Sanger, who not only invented the first reliable and universal method for DNA sequencing in 1977, but prior to that also developed a method to determine the amino-acid sequence of proteins. Each of these discoveries earned him a Nobel prize.

Sanger himself called his DNA-sequencing technique the "dideoxynucleotide chain-termination method" - a name, which nicely captures the central idea of his method. For his sequencing method, Sanger assembled all the components that are needed for a DNA-synthesis reaction in a test tube: a single-stranded template DNA, a primer, a polymerase, and triphosphate deoxynucleotides (dATP, dGTP, dTTP, and dCTP). When mixed together, the primer anneals to its complementary site on the template, the polymerase binds to the formed duplex and begins to extend the primer by incorporating complementary nucleotides according to the sequence of the template strand. In the synthesis reaction, the α-phosphate group of the incoming nucleotide gets attached to the 3'-hydroxyl group of the previously incorporated nucleotide (figure 1-1).

Sanger added one additional ingredient to the synthesis reaction: A dideoxynucleotide (ddNTP, figure 1-1 right). This is a chemically synthesized nucleotide analog, in which the 3'-hydroxyl group is replaced by a hydrogen atom. This nucleotide is incorporated by the polymerase just like any other nucleotide. But the lack of a 3'-hydroxyl group prevents the addition of any further nucleotides to this growing strand. Hence, as the original name of the method indicates, the dideoxynucleotide terminates the synthesis of the growing DNA chain.

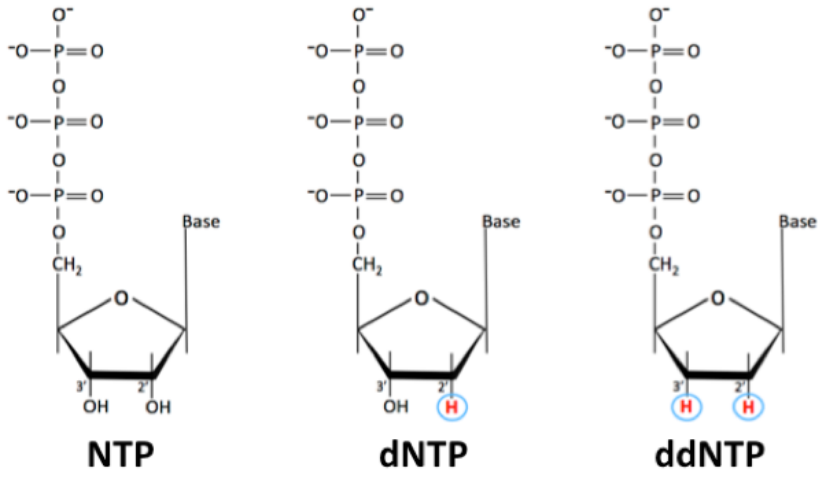

Figure 1-1 Comparison of the chemical structure of NTPs, dNTPs and ddNTPs. NTPs, such as ATP, serve as the energy currency of the cell. dNTPs are the building blocks of DNA. They lack the hydroxyl group on the 2' carbon of the sugar ring. This stabilizes DNA against hydrolysis. ddNTPs are synthetically generated compounds that are used as chain terminators in Sanger sequencing. In ddNTPs, the 3'-hydroxyl group of the sugar ring, which is needed for the attachment of the next nucleotide, has been replaced by a hydrogen atom.

Let us now consider what happens if we add a small amount of just one type of dideoxynucleotide, for example, ddATP, to the DNA-synthesis reaction. Synthesis proceeds normally until the polymerase encounters a thymidine on the template strand. At that time, the polymerase can either incorporate a natural dATP in the growing strand and synthesis will continue with the next nucleotide; or, an artificial ddATP is incorporated, in which case the synthesis of this particular strand is terminated. Which of the two happens is determined by chance and by the relative concentration of dATP and ddATP. Those strands where a "normal" dATP was incorporated will continue to grow until the polymerase encounters another thymidine, in which case there is again a chance that synthesis is terminated by addition of a ddATP, etc. (figure 1-2).

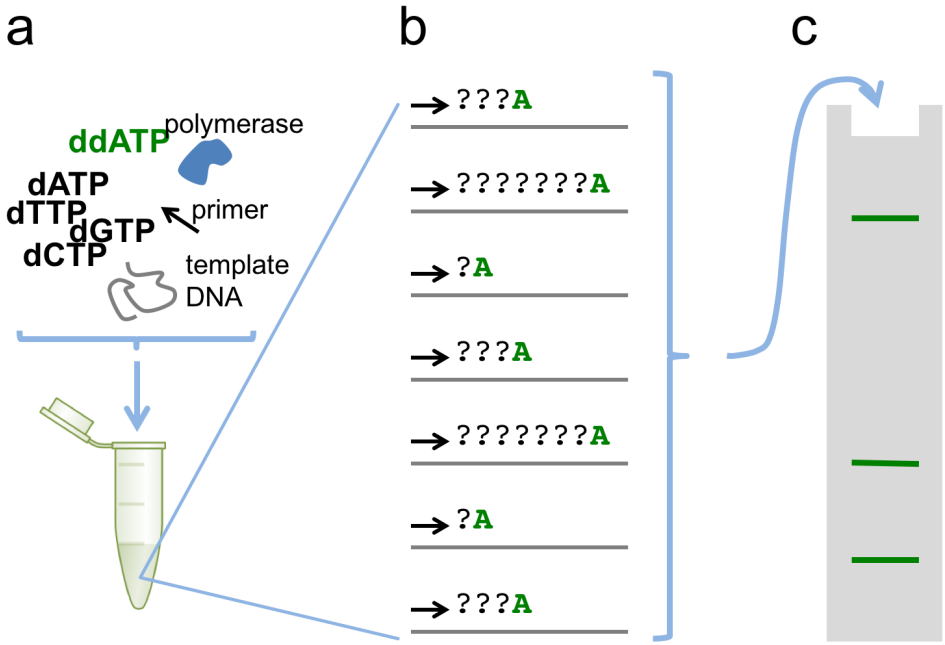

Figure 1-2 Sanger DNA sequencing uses random termination of DNA synthesis to determine the locations of a specific base in a DNA sequence. During the synthesis reaction **(a)**, a specific dideoxynucleotide (e.g., ddATP) included in the synthesis reaction may be inserted in the growing DNA chain wherever a T is present in the template strain, causing termination of the chain at this position. The result is a population of DNA strands **(b)** where each strand terminates with an adenine. Separating these DNA fragments according to length by gel electrophoresis **(c)** reveals the relative position of adenine nucleotides in the newly synthesized DNA strands.

The result of this synthesis reaction is a population of DNA strands of different lengths (fragments). The reason the Sanger methods works is that these fragments contain information about the positions where the template strand contained a base complementary to the dideoxynucleotide used in the reaction. If, for example, the template contained a T at position 2, 4 and 9, then a ddATP nucleotide may have been incorporated at these positions in one of the growing strands and would have terminated the synthesis of this strand. As a result, the products of the reaction will contain newly synthesized strands that are 2, 4 and 9 bases long. At positions where the leading strand contained any of the other three bases, synthesis will have simply continued. Thus, none of the newly synthesized strands will be 1, 3 ,5 ,7 or 8 bases long.

By performing the synthesis in four separate reactions, each containing a different dideoxynucleotide, one obtains four populations of newly synthesized strands. Together, the fragments of these strands encode the complete base sequence of the DNA strand. The second step of Sanger sequencing is to read out this information by separating the newly synthesized strands by gel electrophoresis. Shorter

DNA strands will move through the gel more quickly than longer strands. To visualize the positions of the DNA bands on the gel, early versions of the Sanger method used radioactively labeled ddNTP molecules and separated the products of the four sequencing reactions on adjacent lanes of the same gel. Exposing the gel to a photographic film would reveal a staircase pattern of bands corresponding to the different lengths of newly synthesized DNA strands (figure 1-3). By stepping along this staircase from the bottom of the gel (shortest strand) to the top, the sequence of the newly synthesized strand can be read in the 5'-to-3' direction. This sequence, which is read out from the gel, is complementary to the DNA strand that was sequenced.

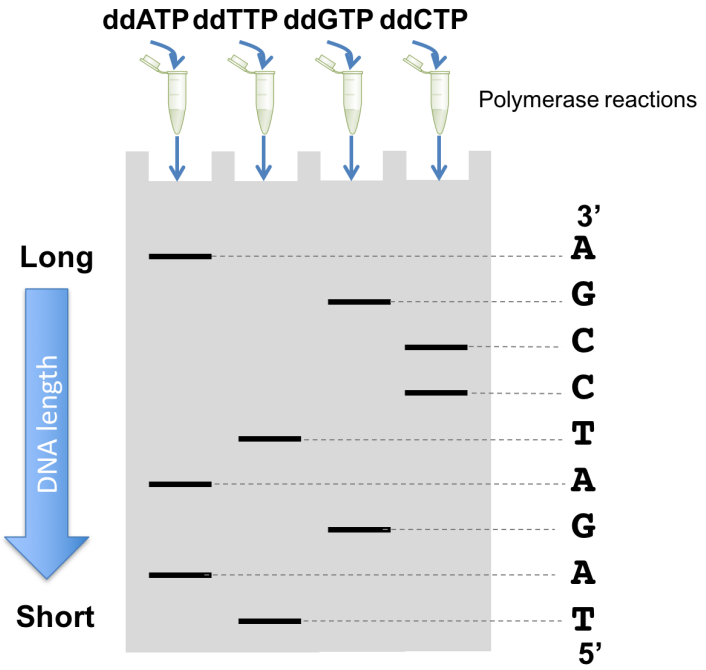

Figure 1-3 Reading out the DNA sequence from an autoradiograph generated by the Sanger method. The products of the four separate sequencing reactions are separated on a gel. The gel is then exposed to a photographic film and the blackened bands on the film reveal the relative length of the DNA strands generated in the four reactions. By "stepping" from the smallest to the largest strand the sequence of the entire strand can be read off in the 5'-to-3' direction.

This radioactivity-based version of the Sanger method was soon replaced by a fluorescence-based version (figure 1-4). In this fluorescence-based approach, each of the four dideoxynucleotides is labeled with a different fluorophore. Thus, the type of nucleotide (A, T, C, or G) that terminated the synthesis of a strand is encoded in the color of the fluorophore attached to this strand. This made it possible to combine the four, previously separate synthesis reactions into a single reaction and to separate all of the newly synthesized DNA strands on a single lane of a gel. Most importantly, reading out the base sequence could now be automated by placing a color-sensitive fluorescence detector at the bottom of the gel that detects each of the fluorescing bands as they pass the detector (figure 1-4).

Performance and application of Sanger sequencing

Through steady development and automation of all aspects of the Sanger method, the efficiency of the technique was improved continuously to the point where a dedicated technical staff member could sequence about 1 million bases per day at a cost of 2'000 - 3'000 US dollars.

Yet, the fundamental principle of sequencing, terminating the growth of DNA strands by random incorporation of modified nucleotides and then separating the resulting DNA molecules by gel electrophoresis, has remained the same. The maximal read length of the Sanger method is 1000 bases. It is limited by the relative size resolution obtainable by electrophoresis. The challenge can be understood by considering that to read the 100th base in a sequence, the electrophoresis step has to separate two strands of DNA that are 99 and 100 bases long. This corresponds to a relative size difference of 1/100. But, to read the 1000th base, the electrophoresis needs to separate two strands that are 999 and 1000 bases long. This brings the required relative resolution to 1/1000, which approaches the current technical limit. The error rate of Sanger sequencing typically lies in the range of one error per 10'000 - 100'000 base pairs, which is extremely low.

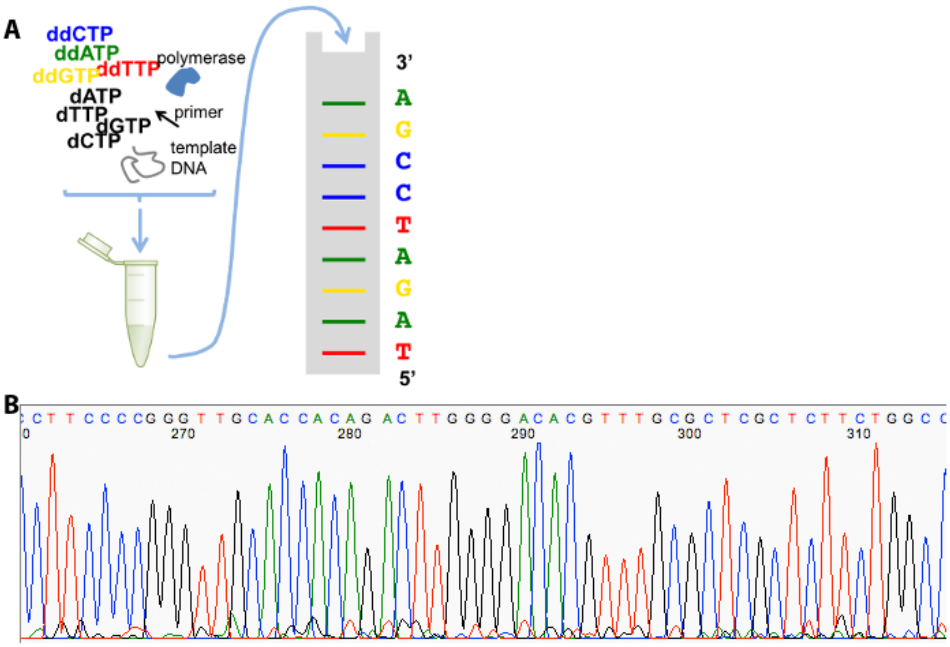

Figure 1-4 Fluorescence-based Sanger sequencing. (A) Fluorescently labeled ddNTP analogs, each fluorescing in a different color, are used to distinguish the different chain-termination products of the polymerase reaction. **(B)** A chromatogram recorded by a fluorescence detector placed at the bottom of a capillary gel. Each peak corresponds to a band on the gel. The sequence, which is complementary to the sequenced DNA strand, in 5'-to-3' order is read from left to right.

Most large-scale DNA sequencing projects (e.g., whole-genome analysis) now employ the second- and third-generation methods that will be discussed in the next section. Still, Sanger sequencing remains popular for many day-to-day applications in molecular-biology laboratories and for medical diagnostics. There are two reasons for this continued popularity. First, the Sanger method employs sequencing primers that can be chosen by the experimenter. This allows the targeted sequencing of a specific portion of a larger DNA molecule while the 2nd and 3rd generation workflows will sequence the entire DNA in the sample. Second, the Sanger method can read long stretches of DNA in a single reaction with very low error rates. In medical diagnostics it is therefore still common to verify candidate mutations found by 2nd and 3rd generation sequencing methods via the Sanger method.

Development of sequencing technologies has led 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 readout 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 sequencing.

Third-generation technologies have found ways to circumvent this amplification step and 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.

Just as the Sanger method, the sequencing-by-synthesis method is based on a sequencing reaction that resembles the natural DNA-synthesis process. 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 1-5 for an example). Additionally, while in Sanger sequencing the functional groups of the nucleotides (i.e., the fluorophore, which indicates the type of nucleotide and the terminator that prevents the attachment of the subsequent nucleotide) are permanently attached, the groups can be chemically cleaved in the case of the SBS nucleotides.

The basic biochemical process of SBS is relatively simple (figure 1-6a). Template molecules are attached to a solid surface (called 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 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 the sample is measured 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.

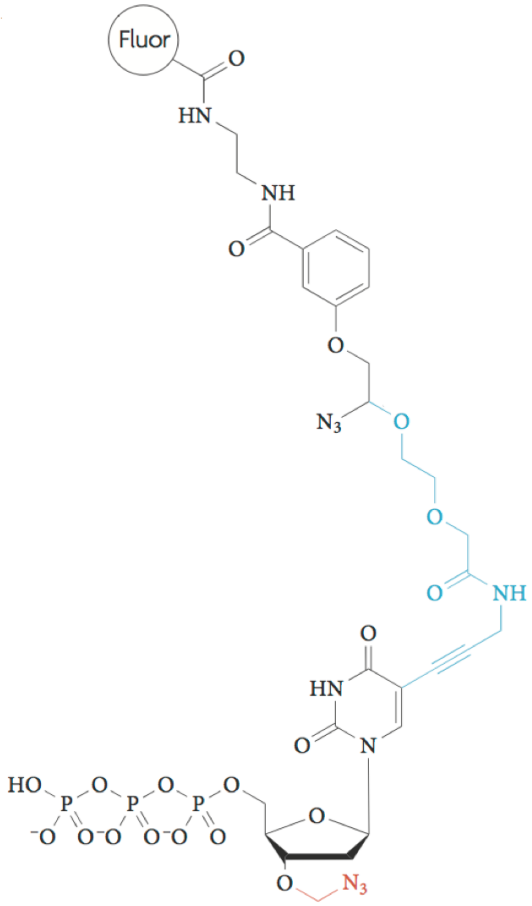

Figure 1-5 Chemical structure of a fluorescently labeled, chain-terminating dCTP analog used in the sequencing-by-synthesis method. During synthesis, the azide group shown in red blocks protects the 3'-hydroxyl group and prevents the incorporation of additional nucleotides into the growing strand. This 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 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 step forward. The SBS reaction shares many components with the Sanger reaction (use of a polymerase, fluorescently labeled nucleotides 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 1-6b).

The key to this is that in SBS, the DNA fragments to be sequenced are attached to a solid surface. Thus, individual reactions can be identified by their xy-coordinates (figure 1-6c). This makes it possible to perform many individual sequencing reactions side-by-side. Further, integrating the biochemical reaction steps and the physical readout of the sequence signal into one continuous process allows reading out the sequence information 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 (figure 1-6d).

The colored dots visible on the flow cell surface (figure 1-6c/d) might look like they correspond to a single DNA molecule. This is actually 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. 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. This PCR takes place directly on the flow cell. Therefore, primers have been attached to the flow cell's surface before adding the DNA to be sequenced. These primers bind the DNA and allow amplification of each fragment, generating 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.

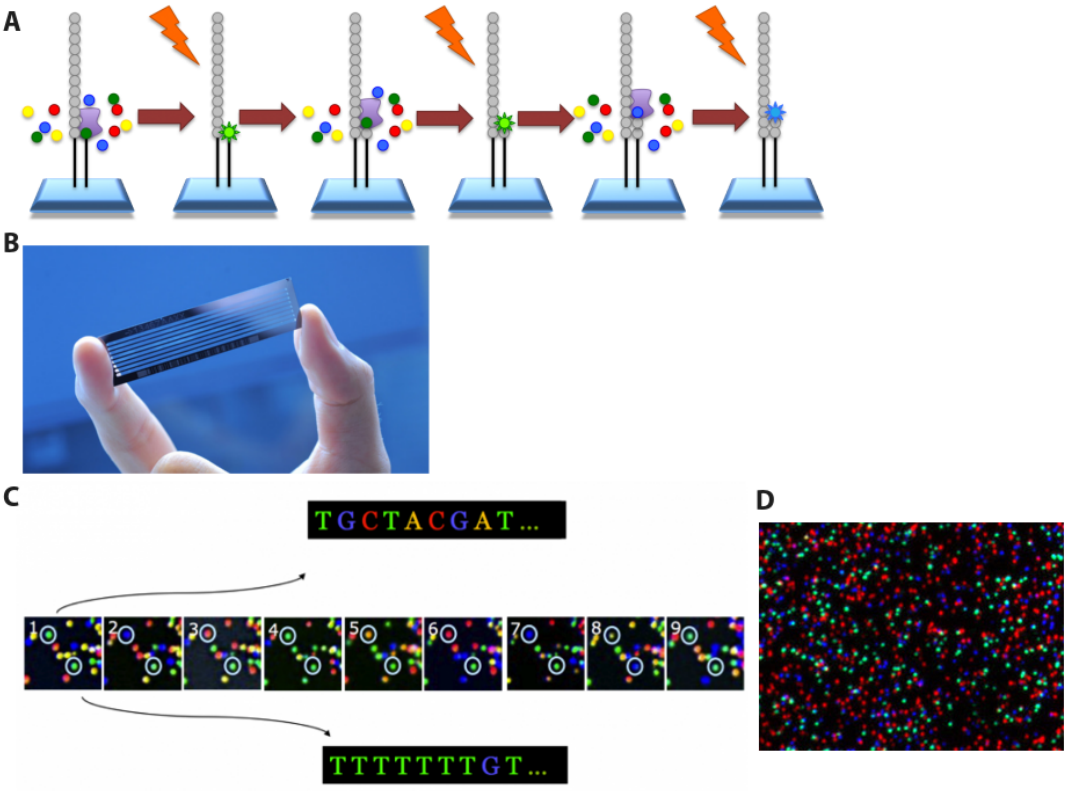

Figure 1-6 Sequencing-by-synthesis. (A) Three cycles of a sequencing-by-synthesis reaction. The polymerase (purple) extends the primer by a single nucleotide. Unincorporated nucleotides are washed away and the fluorescence signal is detected (arrow). The fluorophore and the chain-terminating group are cleaved off and washed away. Then the cycle is repeated. **(B)** Flow cell used in the sequencing-by-synthesis instrument. The DNA to be sequenced is 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 a DNA sequence of 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.

Limitations of the second-generation sequencing technologies

The need to synchronize reactions across molecules limits the speed and read-length of second-generation sequencing technologies. The newest implementation of the SBS technology requires approximately 1 hour to add one additional base to each of the molecules in the flow cell. This is very slow in comparison to the 10-bases-per-second speed (=36'000bp/h), with which a DNA polymerase can incorporate nucleotides under optimal conditions.

This slowness of the SBS 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 SBS 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 i) a substantial waiting time at each reaction step to ensure all molecules have completed the reaction and ii) the time-consuming flushing in and washing out of the reagents that is 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 SBS 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: sequencing individual molecules

Third generation sequencing technologies seek to sidestep the synchrony-based problems of second-generation sequencing techniques by performing the sequencing reaction on individual molecules. We will discuss two recent developments in this area: PacBio and Nanopore.

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 obvious. 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 1-7). The Pacific Biosciences (PacBio) method uses special fluorescently labeled nucleotides. The polymerase cleaves the attached fluorophore during incorporation into the growing DNA strand, allowing it to diffuse away from the sensor area before the next labelled dNTP is incorporated. 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.

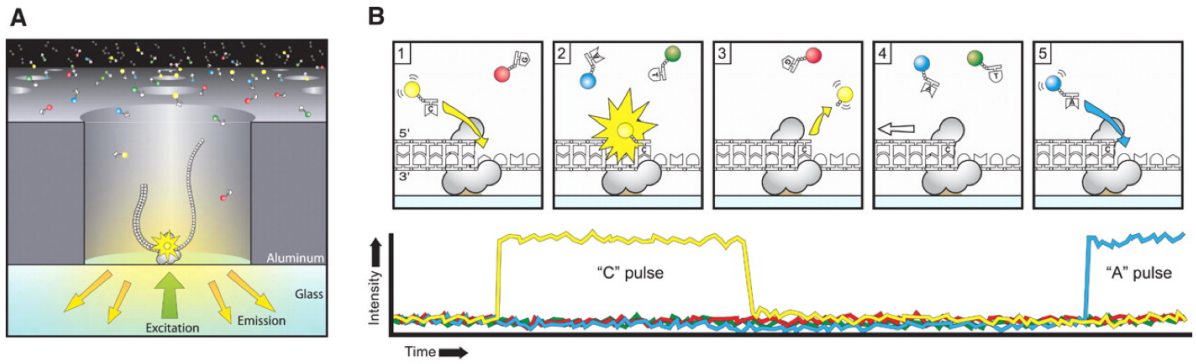

Figure 1-7 Principle of the PacBio single-molecule real-time sequencing technology. (A) Individual DNA polymerase molecules are immobilized at the bottom of 100 nm diameter nanowells etched into the bottom of the reaction chamber. **(B)** Individual fluorescently labeled nucleotides diffuse in and out 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 optics. The color of the fluorescence (trace bottom right) reveals which base is incorporated. (from Eid *et al., Science,* 2016)

The challenge for such an approach is that the signal that can be obtained from a single fluorophore molecule is very limited. Also the fluorescence noise 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 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 developed by the PacBio scientists is to keep the volume around the polymerase, in which the fluorescence is monitored to the extremely small size of a few zeptoliters (10^{-21} liters). Because this 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 synthesis reaction. Nucleotides will briefly diffuse into the monitored volume, but will diffuse out again very quickly on the time scale of a few microseconds. Only when the nucleotide recognizes and binds to the template strand's complementary base in the polymerase active site, the nucleotide and its attached chromophore remain in the monitored reaction volume long enough to generate a significant fluorescent signal. However, this period lasts only the few milliseconds it takes for the chemical step that forms the phosphodiester-backbone bond. At this point, the gamma-phosphate attached fluorophore is released and diffuses out of the monitored volume.

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.

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. The porin protein is embedded into a polymer-based membrane that separates two chambers containing electrolyte solutions (see figure 1-8). By applying a small electrical voltage across the membrane that separates 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. 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 CTGAC, will give a different signal than a G in the context of the sequence TCGCC. 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.

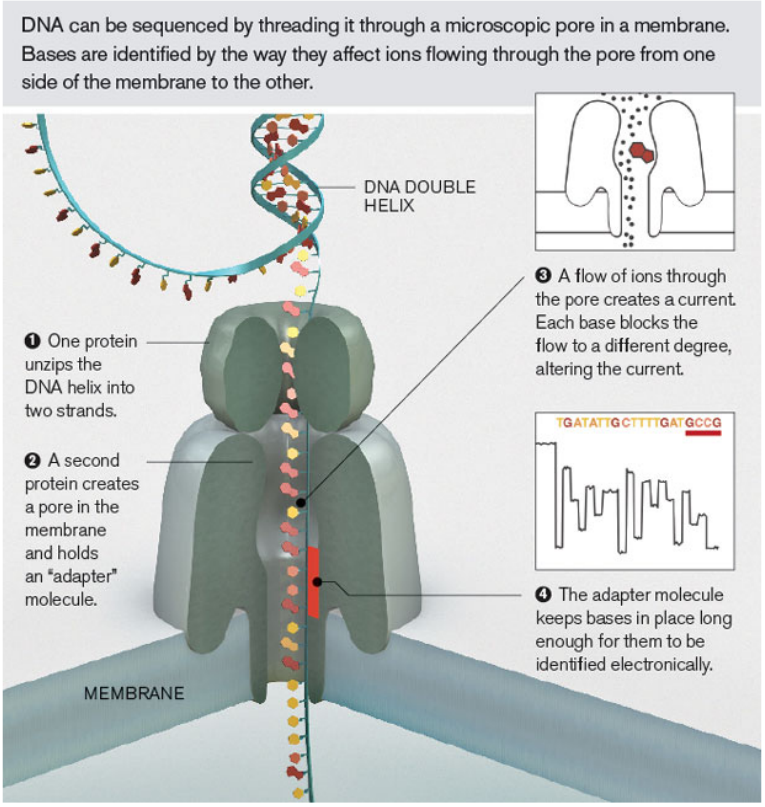

Figure 1-8 Operating principle of the Nanopore sequencing technology. A protein nanopore is embedded in a membrane that separates two chambers containing electrolyte solutions. A voltage applied across this membrane pulls ions and the DNA strand through the pore. The presence of the DNA in the pore restricts 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. (A. Schaffer, *MIT Technology Review*, 2012)

Nanopore technology really stands out concerning the speed with which individual DNA molecules are

read. As of fall 2016, this speed is at $\sim$ 500 bases per second - already very close to the fundamental read-speed limit of DNA-synthesis based sequencing. 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, fiber optics, optical grating etc. of the fluorescence based instruments are needed. As a result, the technology for a Nanopore sequencer can be packed into a device the size of a candy-bar - much smaller than the kitchen-sized enclosure needed to house the newest PacBio instrument.

How to sequence very long stretches of DNA

The methods discussed so far can generate continuous sequences of DNA a few hundred to a few thousand bases long. But the stretches of DNA to be sequenced are often many thousands of times longer than this (e.g., a whole human chromosome or an entire bacterial genome). This means that the initial full-length stretch of DNA has to be broken up into parts that are of the appropriate size for a given sequencing technique. This step is referred to as library generation. After the sequence of each of these smaller pieces has been obtained, these stretches of sequence have to be put back together in the correct order to obtain the sequence of the full-length stretch of DNA. This step is called sequence assembly. As we will see, library generation and sequence assembly are intimately linked parts of what is called a sequencing strategy.

There are two strategies for sequencing long sequences (e.g., whole genomes), approaching the task in very different ways: map-based sequencing and shotgun sequencing. The two strategies each have advantages and disadvantages. In practice most large-scale sequencing projects currently use a hybrid of the two strategies.

Map-based sequencing

The first strategy that was developed for sequencing very long pieces of DNA is called map-based sequencing. In this strategy, the initial piece of DNA is systematically broken up into consecutively smaller fragments. This results in a library of DNA fragments where each fragment is stored in its own test tube and the location of each of the fragments in the initial piece of DNA is known. These fragments are then sequenced individually and the sequences of the fragments can be assembled in their known order to obtain the sequence of the full-length DNA.

The full-length DNA is broken into large fragments of about 100'000 to 200'000 bp length using restriction enzymes. These large fragments are inserted into bacterial artificial chromosomes (BAC), circular DNA constructs that allow the insertion and propagation of long stretches of DNA (like bacterial plasmids do). The position of each of these large fragments in the initial full-length DNA sequence is determined by physical mapping (figure 1-9). Then, a subset of these BACs are selected that spans the entire sequence. The selected BACs are each digested further into fragments of about 10'000 bases. Again, physical mapping is used to determine the position of each of these smaller fragments within their respective BAC. These fragments in the plasmids are now sequenced using a primer that is complementary to a known portion of vector sequence directly adjacent to the 5' end of the inserted fragment. The sequence generated in this reaction will then reveal a substantial stretch of the sequence of this insert. This sequence can then be used to generate the primer for the next sequencing reaction and so forth (this process is called primer-walking). The process is repeated until the appearance of sequence stemming from the plasmid indicates that the entire sequence of the insert has been obtained.

It is important to realize that this process represents a very substantial logistical challenge. A diploid human genome, for example, contains 6 billion base pairs, which corresponds to a minimum of 600'000 fragments that each needs to be stored in a separate container and for which separate primers need to be designed and synthesized. The upside of the map-based approach is that there should be no ambiguity about how to assemble the sequences of the individual fragments back to obtain the full-length sequence.

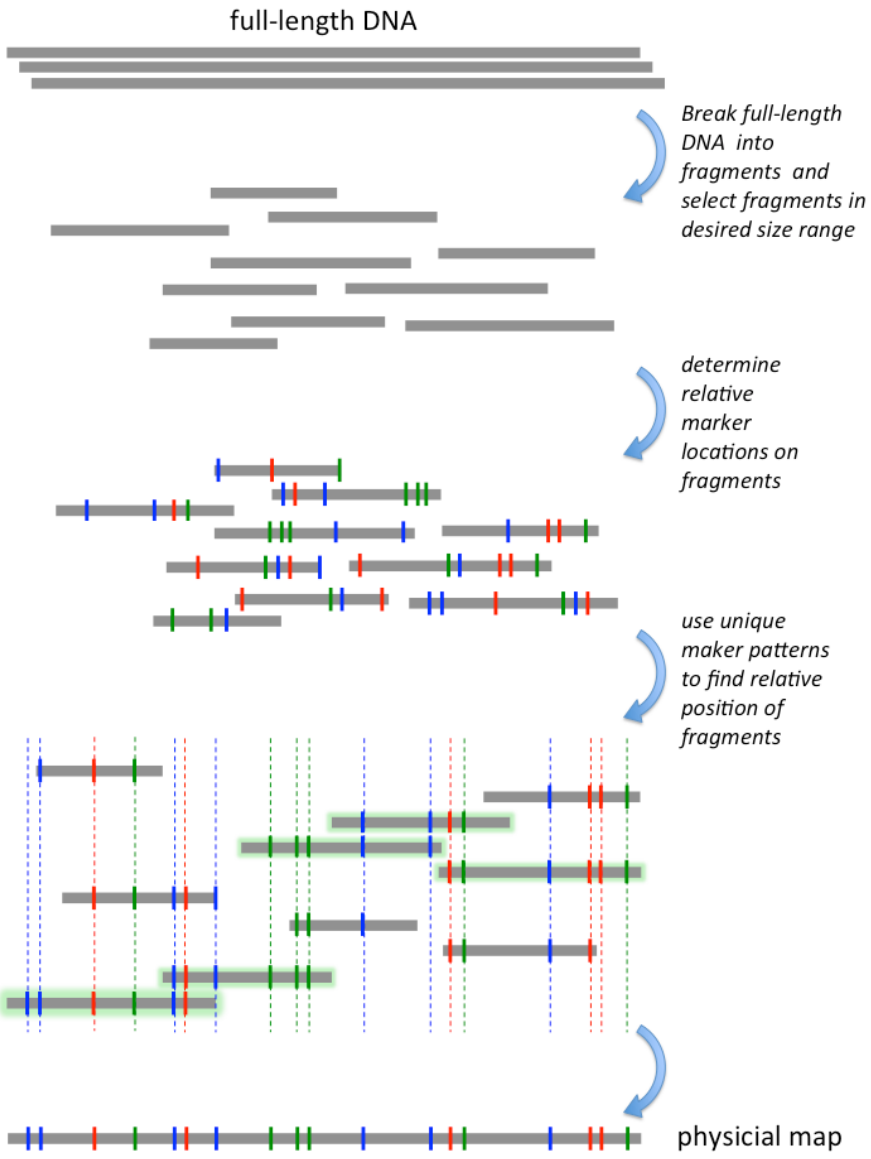

Figure 1-9 A physical map indicates the position of sequence-specific markers (blue, green and red) on a continuous piece of DNA (long grey bar). Traditionally, physical maps were often based on naturally occurring restriction-enzyme sites. The positions of these restriction sites were determined by analyzing the sizes of DNA fragments generated by individual and combined digestions with multiple restriction enzymes. Increasingly, fluorescently labeled oligo probes combined with optical-distance measurements are replacing restriction enzymes as the mapping technology of choice. A physical map can be used to determine the relative position of different fragments (short grey bars) within the full-length DNA. Out of all the fragments that were generated, a unique set of fragments (indicated by green glow) spanning the entire DNA sequence is selected. These fragments are cloned and can then be sequenced or mapped in further detail.

Shotgun sequencing

The alternative to map-based sequencing is the shotgun strategy. Here, the full-length piece of DNA is sheared into fragments of a size that can be sequenced in an individual read. These fragments are then sequenced individually. In this approach, all of the subcloning, mapping and primer-walking steps are eliminated. Obviously, this simplifies the library preparation and sequencing process dramatically. But, the resulting library is completely unordered.

How can the sequences of the individual fragments in this library be assembled to yield the full-length sequence? The answer lies in the fragment sequences themselves. Because multiple copies of the initial piece of DNA were sheared randomly, most of the fragments will overlap partially with several of the neighboring fragments. The overlapping portions of the neighboring fragment will therefore have identical sequences. One can use these overlapping sequence sections to find the neighbors of a particular fragment and then the neighbors of those fragments, etc. Using substantial computing power, this ultimately allows the assembly of the complete sequence for the initial full-length DNA (figure 1-10).

Assuming random sequences, an overlap of just 16 bases (4^{16} = 4.3 billion combinations) would be sufficient to uniquely identify a neighboring fragment in a library derived from an entire haploid human genome (3.2 billion bases). Unfortunately, many of the DNA sequences of interest contain duplicated genes or long stretches of repeat sequences. The presence of such sequence elements will result in ambiguities about the way in which the fragments need to be assembled. As a result of these ambiguities, it is typically not possible to accurately assemble entire mammalian genome sequences using a pure shotgun approach and the shotgun sequencing data needs to be complemented by traditional mapping data.

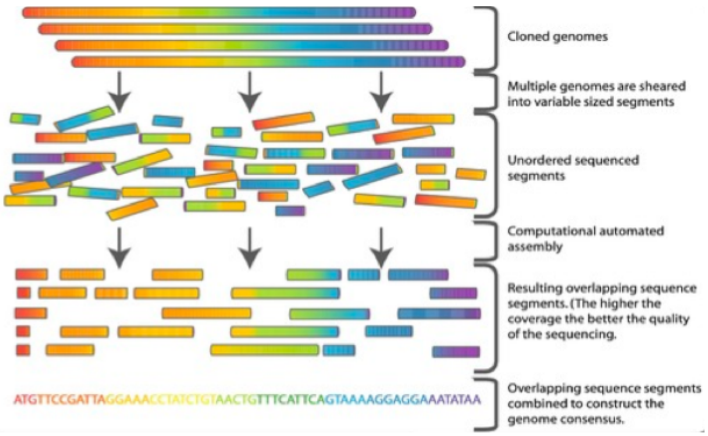

Figure 1-10 Conceptual workflow for the shotgun-sequencing strategy. Multiple copies of the initial full-length DNA (e.g., a full genome) are broken up into small fragments and sequenced. Overlaps between the fragment sequences can then be used to assemble the full-length sequence. (adapted from Commis *et al.*, 2009)