

Chapter 2

Emergence of Next-Generation Sequencing

Although in the past few years the genome of several species, as well as humans, were sequenced using the automated Sanger method, the above-mentioned limitations of this method indicated a need to develop new and improved sequencing technologies to sequence the large number of human genomes and to find answers to biological problems of interest that could not be addressed before [1, 2]. For example, advances in sequencing technology would help in the development in fields such as:

1. Comparative genomics, which involves comparing the genome of distinct organisms to learn about their molecular programs.
2. Biomedical research, through which so many problems concerning the genetic basis of susceptibility to diseases, multi-factorial diseases, and cancer therapy can be investigated. The detection of different genomic and epigenomics alterations, such as single nucleotide mutations, small insertions and deletions, chromosomal rearrangements, copy number variations, and DNA methylation can be facilitated using advanced sequencing technologies [3].
3. Personal genome projects (individual genome sequencing) that impact human health, by becoming one of the major components of personalized health care by providing accurate diagnosis, prognosis and guidance for treatments [3, 4].

Subsequently, several research centers initiated the designing of new sequencing technologies not needing gels, which would allow sequencing large numbers of samples in parallel [5]. These technologies are known as Next-generation DNA sequencing (NGS) methods in which the bacterial cloning steps have been removed (in comparison with the Sanger method). Three major NGS methods that are routinely used in many laboratories today include:

1. The Roche/454 FLX (<http://www.454.com>)
2. The Illumina/Solexa Genome Analyzer (<http://www.illumina.com>)
3. The Applied Biosystems SOLiDTM System (<http://marketing.appliedbiosystems.com>)

Another three massively parallel technologies that have been introduced more recently include the Polonator (Dover/Harvard), the HeliScope Single Molecule Sequencer technology (Helicos; Cambridge, MA, USA) [6, 7], and the Ion Semiconductor (Torrent Ion Sequencing). The single molecule real time (SMRT) [Pacific Biosciences] and Nanopore Sequencing [8] are another two newly introduced technologies that are based on the sequencing of single molecules. The biochemical reaction principles of each of the above-mentioned methods will be described later.

De novo sequencing versus resequencing

There are two major types of sequencing projects in terms of application. In de novo sequencing, the genome of an organism is sequenced for the first time. In resequencing projects, the whole genome of an organism—or parts of it—is sequenced while the reference sequenced genome for the species of that organism is already available.

Sequencing depth (coverage) versus sequencing breadth

Maximum sequencing efficiency is achieved as a consequence of both depth (coverage) and uniform read distribution (breadth). Sequencing depth or coverage concerns the average number of times each base in the genome is sequenced. For example, to sequence a 3 Gb human genome with 10x coverage, 30 Gb of sequenced data is needed. Sequencing breadth refers to the percentage of the genome that is covered by sequenced reads.

2.1 454 Pyrosequencing

454 Life Science (Branford, CT, USA) was the first next-generation sequencing technology that was commercially available (in 2005) (later acquired by Roche) [7] and was also the first NGS technology to sequence a complete human genome, that of Dr. James D. Watson [9]. This technique was first successfully validated by sequencing of the entire 580,069 bp of the *Mycoplasma genitalia* genome at 96 % coverage and 99.96 % accuracy in a single run. This method is based on using an alternative sequencing technique called “pyrosequencing,” which was first introduced by PålNyrén and Mostafa Ronaghi at Stockholm’s Royal Institute of Technology in 1996. The maximum length of sequenced reads by the 454 System was 450 bp at first and has now been increased to about 700 bp. (Note that most of the road’s length presented here for discussed platforms is subjected to rapid changes and are appropriate just for comparison between different platforms.) In pyrosequencing, which uses the “sequencing by synthesis” concept (it differs from the Sanger sequencing method in that it depends on the detection of pyrophosphate release on nucleotide incorporation rather than chain termination with dideoxynucleotides) [10], the incorporation of each of four nucleotides which are added in a certain order in each step results in the release of pyrophosphate, which in the presence of adenosine 5′ phosphosulfate can be converted to ATP by enzyme ATP

sulfurylase. This ATP in turn acts as a substrate for a chemiluminescent enzyme, luciferase, which converts luciferine to oxyluciferine and visible light that can be detected by a charged coupled device (CCD) camera for further analyses. The amount of generated light is proportional to the number of incorporating nucleotides. Since dATP can act directly as a substrate for Luciferase and subsequently generate light without necessarily needing for incorporation of dATP, in order to prevent any bias, dATP α S, which can precisely incorporate DNA strand while not a substrate for luciferase enzyme, is used instead of dATP. In each step, unincorporated nucleotides are degraded, using an enzyme called apyrase to exclude the possibility of incorporation of nucleotides remaining from the previous steps. The process can be continued with another step of nucleotides additions [11].

In the 454 approaches, genomic DNA is first fragmented into smaller pieces and short adaptors are then ligated to blunt ends (3' and 5' ends) of the single-stranded fragments (Fig. 2.1a). These adaptors flanked fragments are then mixed with small 28- μ m streptavidin-coated beads whose surfaces have been covered with short sequences complementary to one of the ligated adaptors. Fragments are then hybridized to their corresponding beads in such a way that each bead carries a unique fragment. Fragment amplification must be performed for intensifying the light signal that is required for precise detection of added bases by CCD camera [5]. Amplification of DNA fragments are performed through emulsion PCR in which all required reagents are provided in droplets of water in oil mixture which acts as a micro-reactor for every single bead; consequently, it guarantees that every bead is covered by single-type strands at the end of the amplification step. After breaking the emulsions, the beads whose surfaces have been now covered with about one million copies of each amplified fragment are first pre-incubated with the DNA polymerase enzyme and then loaded onto a PicoTiterPlate device so that each single bead is deposited in a well with the dimensions that can contain only one bead [6]. The wells are then filled with smaller beads (1- μ m) that carry immobilized enzymes required for pyrosequencing (sulfurylase and leuciferase) and also help beads to deposit in the wells gravitationally [12]. PicoTiterPlates are centrifuged to ensure that the enzymes are in close contact with the DNA beads [13]. The loaded plates are placed in a 454 FLX instrument in which sequencing reagents (dNTPs and buffers) are delivered to the wells of the plate. The pyrosequencing technique is initiated with the sequential addition of four nucleotides, as previously explained (Fig. 2.1b).

When a nucleotide is added to the growing chain, a signal is recorded by a CCD camera as a result of the generation of light. As mentioned, the signal strength is proportional to the number of nucleotides, and in the case of homopolymers (the stretches of only one nucleotide type), the signal intensity is higher than that of a single nucleotide. However, signal strength can be precisely detected only for less than 10 consecutive nucleotides; after that, the signal declines rapidly [7]. As a result of this drawback, the insertion and deletion error rates are the most frequently observed errors in this method. The substitution error rate of this method (10^{-3} – 10^{-4}) is higher than the rates for the traditional Sanger method. However,

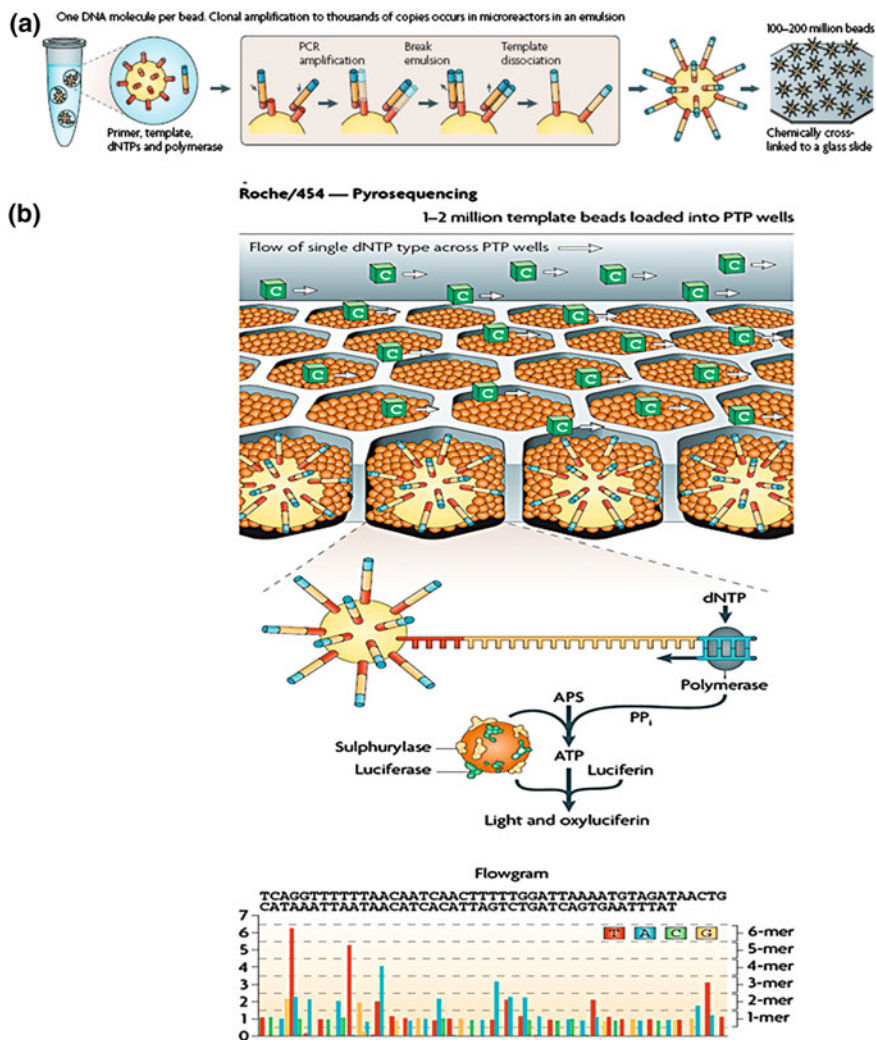


Fig. 2.1 A schematic representation of 454 pyrosequencing technology. **a** Construction of adaptor flanked DNA fragment, DNA amplification on the beads' surface and formation of "clone beads." **b** Deposition of beads into "picotiter plate" wells. In the pyrosequencing method, incorporation of each nucleotide is determined by the detection of light generated by the activity of luciferase enzyme in the presence of ATP and luciferine [5]

the background error rate of the Sanger method that arises from *in vitro* amplification steps that introduce the error in the sample before performing sequencing, is higher [12].

2.2 Illumina (Solexa) Genome Analyzer

The Genome Analyzer first introduced by Solexa in 2006 (San Diego, CA, USA) and then further developed by Illumina. At first the resulting reads were very short (36 bp or less) compared to Sanger methods. Since then, many technical improvements have been introduced to this method, resulting in increased read length to up to 100 bp [12]. Since millions of reads can be generated in parallel (simultaneously), reassembling of these reads using efficient algorithms results in useful outputs.

In this method, a DNA library can be constructed by any method that generates adaptor-flanked fragments up to several hundred base pairs in length (Fig. 2.2a) [6]. Template amplification is performed with a method called “bridge PCR,” in which both forward and reverse primers that are complementary to the adaptors’ sequences are attached to a solid surface called “flow cell,” containing eight independent channels that allow performing PCR amplification on its surface. The name “bridge PCR” refers to the fact that during the annealing step, the extension product from one bound primer forms a bridge to the other bound primer and its complementary strand is synthesized. Since each initial single fragment has been tethered to a distinct channel on the surface, the amplified fragments form clusters, each of which contains approximately one million copies of the initial fragment that would be sufficient for sequencing reaction (Fig. 2.2b). Moreover, each cluster contains both forward as well as reverse strands of the original sequences, but to have homogenous populations of strands in each cluster, which can be sequenced precisely without the interference of the complementary strand, one of the strands must be removed before initiating the sequencing process [12]. The sequencer adopts the technology of sequencing by synthesis. In this method, four reversible blocked nucleotides (3'-OH is chemically blocked), which then transiently block the extension of nucleotides, are added to the reaction. Following the incorporation of each nucleotide, the imaging step is carried out. After the acquisition of images in each cycle, the 3'-OH blocking group is chemically removed and another cycle can subsequently be initiated. The steps of adding nucleotides and other PCR reagents and capturing image can be continued for a specific number of times and ultimately generate 100 bp reads.

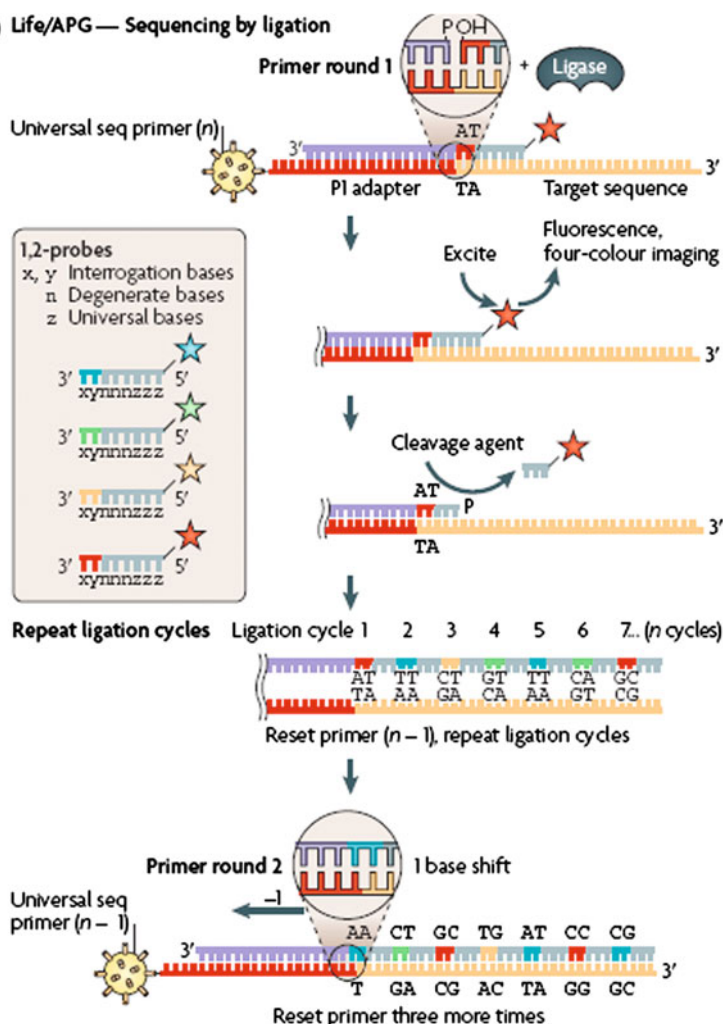
The first Illumina sequencer, GA, as already described, was introduced in 2006. The performance of GA was improved from 1G/run to 85G/run in the GAIIX series, which was released in 2009. In early 2010, Illumina launched HiSeq 2000, which adopts the same sequencing strategy with GA. Its initial output was 200G per run, and improved to 600G per run currently, which could be finished in 8 days. In 2011, MiSeq, a bench-top Illumina sequencer, was introduced; it shared most technologies with HiSeq. MiSeq is especially convenient for amplicon and bacterial sample sequencing [14].

2.3 Applied Biosystems SOLiD Sequencing

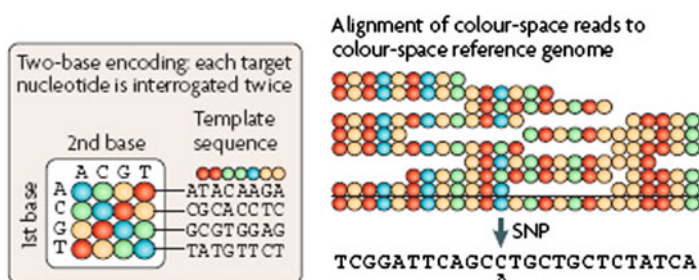
The SOLiD system, which became commercially available in late 2007, was developed by J.S. and colleagues in collaboration with McKernan and colleagues at Agencourt Personal Genomics, later acquired by Applied Biosystems and now part of Life Technologies (Carlsbad, CA, USA). In this method, the library of adaptor-flanked DNA fragments is constructed similarly to the previously mentioned NGS methods. Clonal amplification is performed, as mentioned earlier, for the 454 pyrosequencing platform through emulsion PCR using small magnetic beads (1- μ m in diameter) to which DNA fragments can be bound, so that each bead carries an individual fragment. After amplification of DNA fragments, “bead clones” that have now been covered by nucleic acid species, are selectively isolated and are tethered to a solid surface—a “flow cell,” via 3' modification of DNA strands. This flow cell is basically a microscope slide that can be serially exposed to liquids. A universal primer, which is complementary to the common adaptor in each fragment, is annealed to the DNA fragment. In the SOLiD method, unlike other NGS technologies, which base detection of DNA fragments, is performed through polymerase reaction; this is achieved by sequencing by ligation. A library of fluorescently labeled octamers (four distinct fluorescent dyes) is constructed with the following composition: each octamer has a specific dinucleotide (out of 16 possible combinations), which in different versions of ABI-SOLiD sequencing chemistry is different (for example, in ABI SOLiD v.2.0 positions 1 and 2 and in the original version, bases 4 and 5 are utilized). The rest of the bases have completely random (degenerate) composition. Each octamer also holds one of four fluorescent dyes—most often at position 6, which correlates with the composition of dinucleotide in that octamer. The sequencing process can be conducted as follows:

- (1) A universal primer with “n” nucleotides anneals to each DNA fragment, and then a mixture of octamers with specific dinucleotide composition is added and one whose specific dinucleotide is matched with a template then can be hybridized, which will be followed by the process of ligation in the presence of enzyme ligase.
- (2) In this step, an image is taken (the specific fluor is detected, which corresponds to the specific dinucleotide composition; in fact, each fluorescent dye is in correlation with four dinucleotide combinations—see Fig. 2.4). The unextended fragments are capped in the presence of the same mixture of nonfluorescent dinucleotides.
- (3) In the next step, the last three bases of octamer (i.e., the bases 6, 7 and 8) along with the fluorescent dye are chemically removed, so five bases of octamer are left behind and the above steps are repeated for base 6 and 7. This step is typically repeated 10 times [12] and at the end, the fluorescent colors, corresponding to 10 dinucleotides (1/2, 6/7, 11/12, 16/17, 21/22, 26/27, 31/32, 36/37, 41/42 and 46/47) is determined.

(a) Life/APG — Sequencing by ligation



(b)



◀ **Fig. 2.3** Schematic representation of the steps involved in AB-SOLiD sequencing. **a** After initial primer (n-nucleotides) annealing, seven cycles of oligonucleotide hybridization and ligation, imaging and cutting the 5' three nucleotides along with fluorescent dye is performed. Then the initial primer and extended strand are removed, which is followed by annealing a new primer with n-1 nucleotides (primer reset). Following that, all the above-mentioned steps are repeated. Repeating primer resets four times results in 35-base generation. **b** Dinucleotide encoding by four fluorescent dyes and decoding of each base using two dinucleotides. Alignment of the color-space reads of the particular sequence with color-space reads of the reference genome will result in SNP identification [1, 7]

- (4) In the next step, the initial primer and all ligated portions are melted and discarded, and another round of primer annealing, “primer reset,” is started with a primer with “n-1” nucleotides via repeating steps 1–3, and this time the position of nucleotides 0/1, 5/6, 10/11, 15/16, 20/21, 25/26, 30/31, 35/36, 40/41 and 45/46 is detected. After five primer reset cycles, approximately 50 contiguous nucleotides have been generated. In each of which nucleotides have been detected by two separate dinucleotides (Fig. 2.3a).

Since there are 16 possible combinations of dinucleotides (4^2) and only four fluorescent dyes (Fig. 2.3b), nucleotide identity detection would not be possible only from data color obtained from the incorporation of dinucleotides. The dinucleotide composition of the first dinucleotide, i.e., nucleotides 0/1, is readily deduced from its fluorescent signal because the first base of newly synthesized strand corresponds to the last base of the universal primer, whose composition is obvious and will help decode the composition of other dinucleotides from their corresponding fluorescent signals [12]. An alternative procedure to primer reset after the first round discussed above is to use octamers with different positions of dinucleotides. For example, in the second round, different positions of octamer mixture (e.g., bases 4 and 5) can be correlated with the identity of fluorescent dye [8].

2.4 Ion Semiconductor (Ion Torrent Sequencing)

Ion Torrent sequencing technology uses simple sequencing chemistry in which no enzymatic reaction, no fluorescence, no optic, and no light are used to determine the sequence of a DNA fragment (Life technologies). This technology, which is one of the least expensive methods, has been introduced to research and clinical laboratories as a personal genomic machine. The biochemical basis of this system is very simple and involves the release of hydrogen ion following the incorporation of a nucleotide into a strand of DNA by DNA polymerase (Fig. 2.4a). To carry out this biochemical process in a highly parallel way, Ion Torrent uses a high-density array of micro-machined wells that is provided on the Ion Torrent proprietary microchips. There is an Ion-sensitive field effect transistor (ISFET) below the wells that detects the change in pH as a result of hydrogen release. This change is

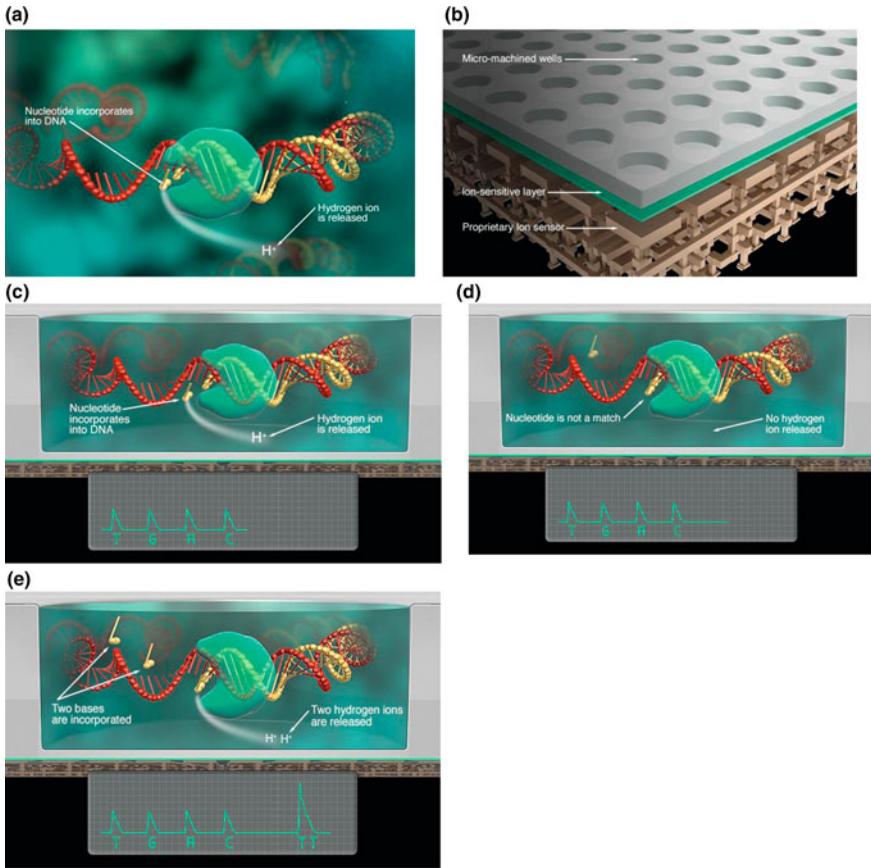


Fig. 2.4 Schematic work flow of Ion Torrent sequencing technology. **a** Release of a hydrogen ion as a byproduct following the incorporation of a nucleotide. **b** High-density array of micro-machined wells. A different DNA template is held in each well. Beneath the wells is an ion-sensitive layer and beneath that a proprietary Ion sensor. **c** Following the incorporation of the added nucleotide and release of hydrogen ion, change of the pH (chemical signal) is translated into digital information. **d** In the case of no incorporation, no voltage change is recorded. **e** If there are two identical bases on the DNA strand, the voltage will be double (taken from Life Technologies™)

recorded as a potential change (ΔV) by an ion sensor layer that indicates the nucleotide incorporation event [15] (Fig. 2.4b).

Calling the bases by this method goes like this: Following the addition of a certain base to a DNA template ("A," for example), if it is incorporated in to the DNA strand, a hydrogen ion will be released and consequently the charge of that ion will change the pH of the solution that can be detected by the ion sensor (Fig. 2.4c). The chip will be flooded by introducing one nucleotide after another. If no incorporation occurs following the addition of a nucleotide, no voltage change is

recorded, and as a result no base will be called. In the case of the presence of two consecutive identical bases on the DNA strand, the voltage will be doubled, and as a result, two identical bases will be recorded. The size of DNA fragments that can be sequenced by this method was first 100 bp and now has been improved to 400 bp.

2.5 Polonator Technology

Polony sequencing technology was first introduced by Dr. George Church's group at the Harvard Medical School. Unlike other sequencing technologies, it has open-source software and free downloadable protocols. The polony sequencing method is begun by paired-end-tag library construction (Fig. 2.5). The Target DNA sequence is randomly sheared, and then the fragments—about 1 kb in size—are selected. After making the ends of these fragments blunt and A-tailing, in which an A is added to the 3' ends of fragments, the fragments are circularized, using 30 bp-synthesized oligonucleotides (T30) with two outward facing recognition sites for a type II restriction enzyme (MmeI); then amplification will occur, using rolling circle replication. In the next step, the amplified circularized DNA is subjected to MmeI, which cuts at a distance of 17–18 bp after detecting the recognition site, and this results in the generation of a fragment of about 70 bp from which 30 bp belongs to T30 and the rest belongs to two 17–18 bp flanking regions or tags. The resulting fragments then end repaired and two emulsion-PCR primers will be attached to their 3' and 5' ends, resulting in the production of a 135 bp fragment that is then subjected to amplification. These 135 bp fragments construct a paired-insert library. Emulsion PCR is then performed on these fragments, using streptavidin-coated beads that have a complementary oligonucleotide of forward primer in each 135 bp fragment. Following that, the resulting emulsion droplets are broken. As a result of inequality in the distribution of the templates on the surface of the beads, the emulsion PCR yields empty, clonal and non-clonal beads [8], and an enrichment procedure must consequently be conducted to produce a dense population of clonal beads. The beads are then immobilized on the surface of the flow-cell for sequencing.

The following steps are carried out during polonator sequencing: after inserting the flow cell in the polonator instrument, each flow cell is subjected to a stream of a mixture of anchored primers where they can be hybridized to the 3 or 5 ends of the two 18 bp flanking oligonucleotides (a total of four possible positions are available).

Following that, a completely degenerate mixture of fluorescently labeled nonamers (in each of which the identity of one position corresponds to the identity of the fluorescent dye that is bound to it: 5' Cy5-NNNNNNNT, 5' Cy3-NNNNNNNA, 5' TexasRed-NNNNNNNC, 5' 6FAM-NNNNNNNG) and T4 DNA ligase flow into the cell. A fluorescent signal is generated when ligation selectively occurs between the anchored primer and corresponding nonamer. In the following step, the

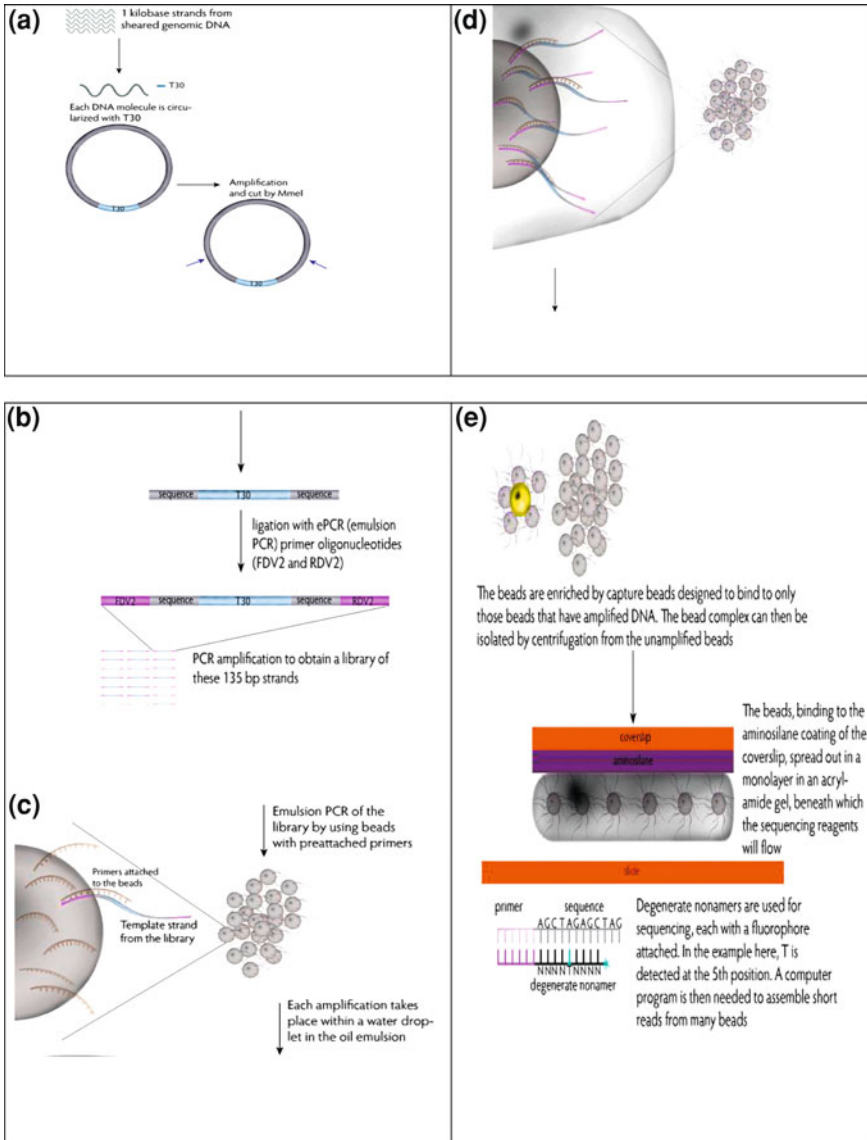


Fig. 2.5 Steps for Polony sequencing method. **a** Paired genomic tag construction using universal linker (T30) for circularizing DNA fragment and MmI for making a cut 17–18 bp outside T30. **b** Ligation of two emulsion PCR primers at the ends of the resulted fragment. **c** In the next step, emulsion PCR is performed, resulting in empty, clonal and non-clonal beads. **d** The enrichment step, which involves bringing together the clonal beads. **e** Conducting sequencing using continuous cycles of occurring ligation process between anchored primers and fully degenerate fluorescently labeled nonamers in each of which the identity of only one base is known and a fluorescent label has been attached to each nonamer according to that base

imaging step is conducted, and after that the primer-nonamer complex and ligase are removed and then another cycle of sequencing can be begun with a new mixture of nonamers in which the query position is one base shifted further into the 17–18 bp tags (5'Cy5-NNNNNNNTN, 5'Cy3-NNNNNNNAN, 5'TexasRed-NNNNNNNCN, 5' 6FAM- NNNNNNNGN) and new primers. Using DNA ligase, the query position can be precisely determined when the distance between it and the ligation site is 7 bases in 5' to 3' direction and 6 bases in 3' to 5' direction. As a result, totally 26 bp from two 17–18 bp tags (13 bp from each) can be sequenced with a 4–5 bp gap within each tag.

2.6 Heliscope (Single Molecule Sequencing)

Heliscope, which was one of the first techniques for sequencing a single DNA molecule, was introduced by Braslavsky et al. in 2003 and licensed by Helicos biosciences in 2007. Using this method, large research projects can be conducted in less time, and fewer errors and lower expenses since there is no need for DNA sample amplification [16]. This method, which is carried out on single DNA molecules, is conducted on an instrument called the Heliscope sequencer and uses the sequencing by synthesis technique. The DNA sample is first sheared into small pieces (100–200 bp) and then becomes adaptor-flanked. This adaptor is usually a poly-A tail, and adaptor-flanked fragments can consequently be tethered to a surface of flow cells on which poly-T oligonucleotides have been bound (Fig. 2.6). After the attachment of fragments, a mixture of a single-labeled nucleotide and polymerase are streamed into the surface and the polymerase will add nucleotide wherever it is complementary to the first positions of the attached fragments. Then the image is taken from all bound fragments that contain incorporated labeled nucleotides. The next cycle of nucleotide incorporation can be performed when the label attached to incorporated nucleotides is cleaved off (this is somewhat similar to the Illumina technology, in which reversible blocked nucleotides were utilized). Nucleotide incorporation can be continued up to 25–45 bases.

2.7 Latest Developments in Next-Generation Sequencing Methods

Since introducing the Helicos system, many attempts have been made to generate efficient sequencing technologies relying on single molecules, excluding the need for an amplification step. The two most promising such technologies are nanopore DNA sequencing and Pacific single molecule real time (SMRT) DNA sequencing.

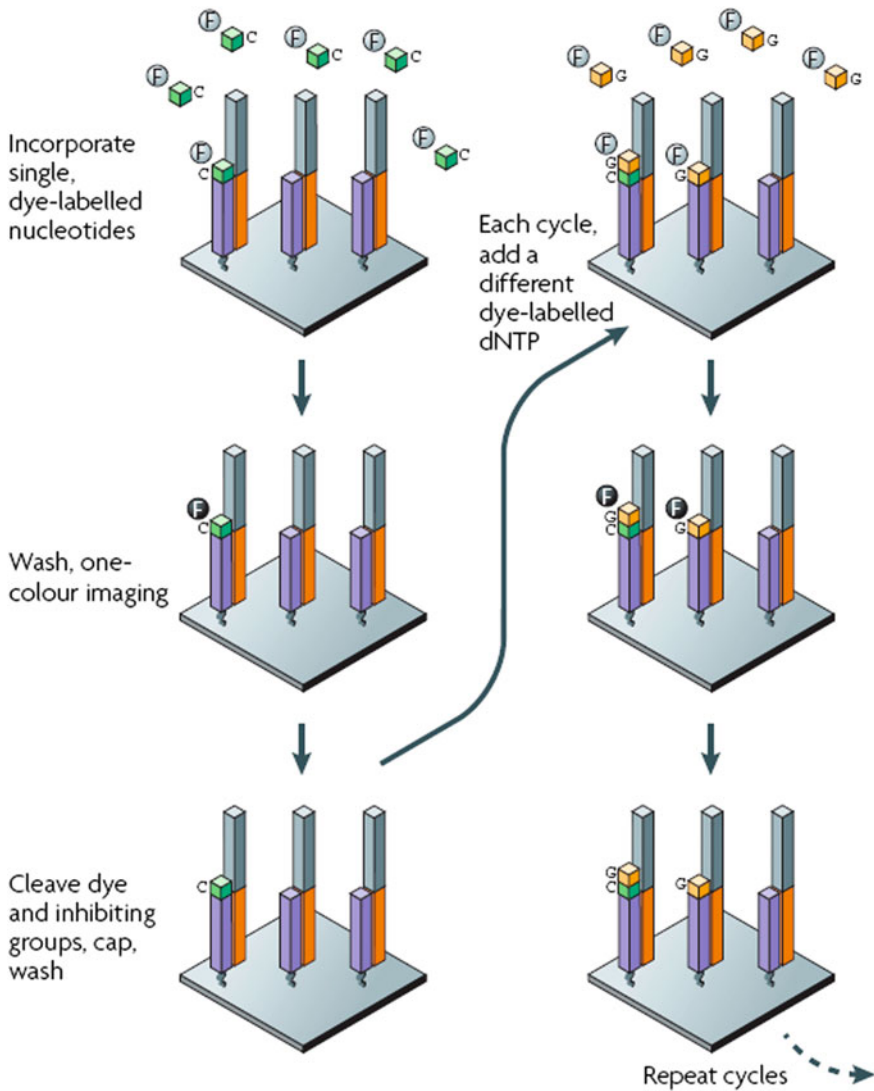


Fig. 2.6 Schematic of Helicos sequencing method. After shearing a single DNA molecule into small pieces and poly-A tailing (shown in *orange*), these fragments are bound to the surface of the flow cell using poly-T already attached to the surface (shown in *dark blue*). Once polymerase and labeled single base flow into the surface, nucleotide incorporation occurs wherever complementarities exist between the added base and the first position of attached DNA fragments. Unincorporated bases are washed from the flow cell. Then a camera scans the entire surface to realize on which fragments the new bases have been added. In the next step, the labeling dye and the inhibitor attached to incorporated bases are cleaved off and consequently another cycle of introducing new labeled nucleotide and polymerase can be begun [1]

2.7.1 Nanopore Sequencing

A few nanopore sequencing methods have been introduced; in all of them, the sequence of a DNA fragment can be identified based on the translocation of a single-stranded DNA molecule through a thin membrane. One of the main features of nanopore techniques is the ability to thread extremely long DNA molecules. Moreover, solid state sub-5 nm pores can be parallelized in the form of a very condensed array [17]. Two nanopore sequencing techniques are described here:

The technique that was introduced by Oxford Nanopore Technologies seems to be able to achieve the US National Institutes of Health's goal of decreasing the cost for sequencing of the entire human genome to \$1,000 in the near future. In this sequencing platform, there is no need for any DNA fragment shearing, amplification step, fluorescently labeled nucleotides and optical instrumentation for detecting fluorescent labels. This technology has its origin in threading a single-stranded DNA molecule through a nanoscale protein pore (staphylococcal α -haemolysin) created in a membrane under an applied potential (Fig. 2.7). Passing a DNA strand through the pore results in fluctuation of the ionic current. The translocation of each base through the pore causes a decrease in current intensification which is specific for each kind of base. In fact, while each of the four bases passes through the pore, a different amount of current can translocate, and this is the key feature in base identification. However, there is difficulty in the base registration step as a result of the high speed with which DNA strand passage takes place, which in turn results in the current resolution that is essential for the precise detection of bases. This problem may be overcome by recent works that reduce the speed of DNA translocation through the pore [18].

In another nanopore sequencing technique proposed by McNally, the target DNA is first subjected to a biochemical preparation step in which each base of the sequence is converted into a form that can easily be read using a solid state nanopore. In fact, each of the four bases (A, C, G, and T) in the target DNA is converted to a predefined sequence of oligonucleotides, which is hybridized with a molecular beacon that carries a specific fluorophore (McNally). Molecular beacons are oligonucleotide probes that can report the presence of specific nucleic acids in homogenous solutions. Molecular beacons are hairpin-shaped molecules with an internally quenched fluorophore whose fluorescence is restored when they bind to a target nucleic acid [19]. For a two-color readout (i.e., two types of fluorophores), the four sequences are combinations of two predefined unique sequences, bit "0" and bit "1," such that an A would be "1 1," a G would be "1 0," a T would be "0 1" and finally a C would be "0 0" (Fig. 2.8a). Two types of molecular beacons, carrying two types of fluorophores, hybridize specifically to the "0" and "1" sequences. The converted DNA and hybridized molecular beacons are electrophoretically passed through a solid-state pore where the beacons are sequentially removed. Each time a beacon is removed, a new fluorophore is lighted, which results into a burst of photons recorded at the location of the pore (Fig. 2.8b). This method allows wide field imaging and spatially fixed pores that make possible the simultaneous detection of

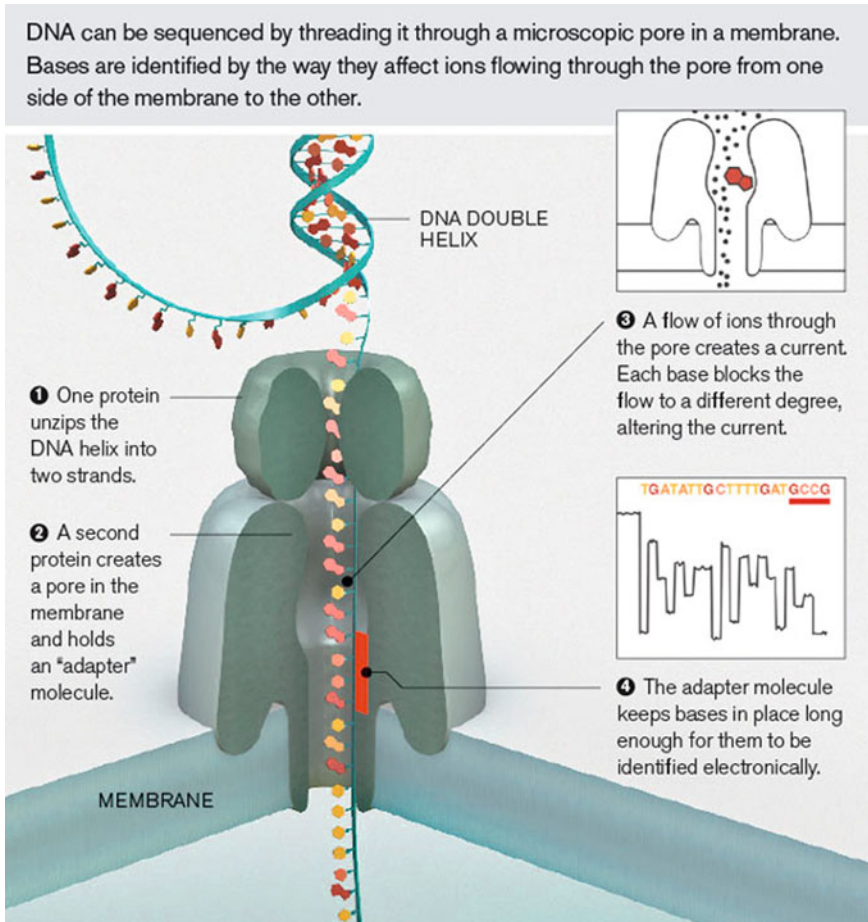


Fig. 2.7 Schematic representation of nanopore sequencing system. (1) The upper protein is used to make the DNA molecule single stranded. (2) The second protein forms a nanopore in a membrane. It also contains an adaptor molecule. (3) Each base obstructs the flow to a different degree. (4) The adaptor is basically used to reduce the speed of passing DNA through the pore, which is necessary for the exact identification of the DNA strand base composition (picture from MIT's Technology Review)

several pores using a special camera, Electron multiplying charged coupled device (EM-CCD) [17].

2.7.2 Single Molecule Real Time DNA Sequencing

This technique was introduced by Pacific Biosciences in 2009. It is based on the observation of the performance of polymerase during DNA synthesis. On this

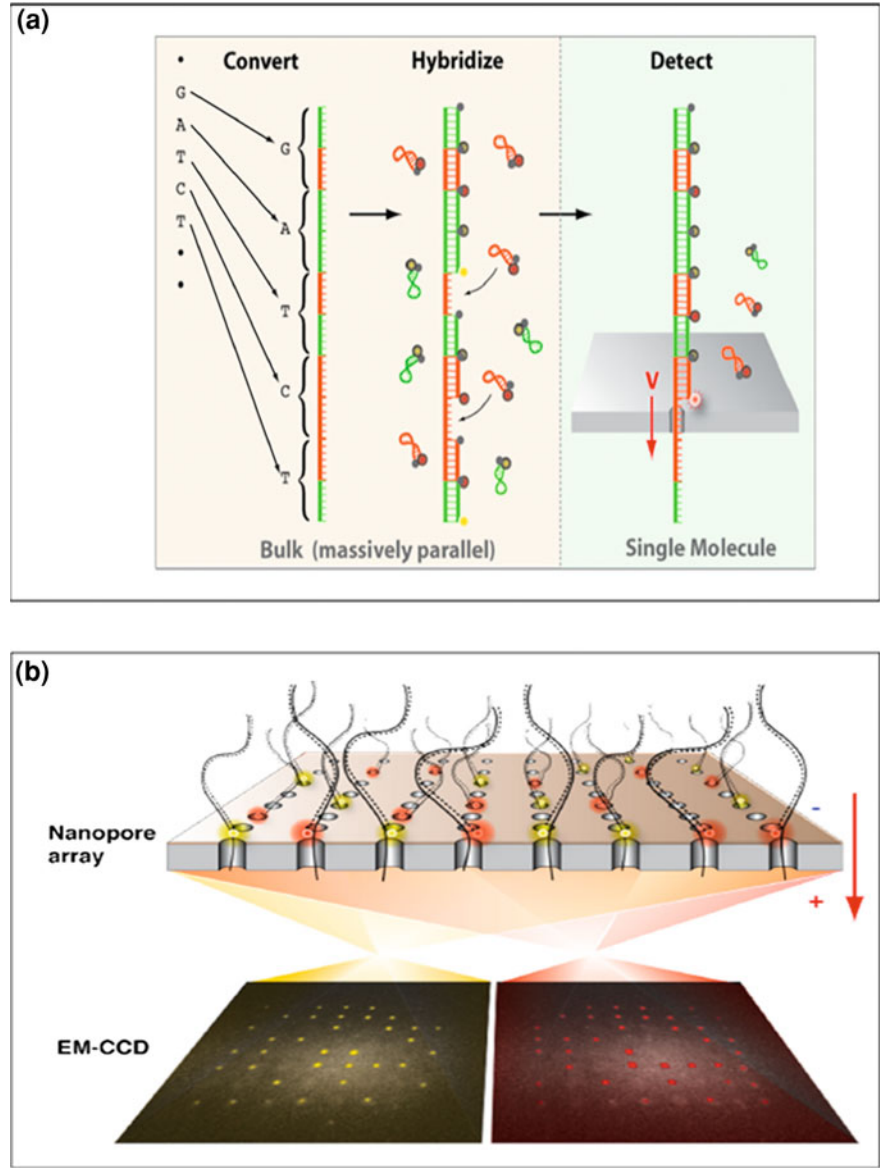


Fig. 2.8 Another version of nanopore sequencing methods introduced by McNally. **a** The biochemical preparation step involving conversion of each base of the sequence into an oligonucleotide that can be hybridized with a molecular beacon. **b** The threading of the beacon hybridized oligonucleotides through a nanopore makes it possible to detect optical signals. The signals are projected onto a wide field imaging screen that is very useful in the simultaneous detection of several pores using an EM-CCD camera [17]

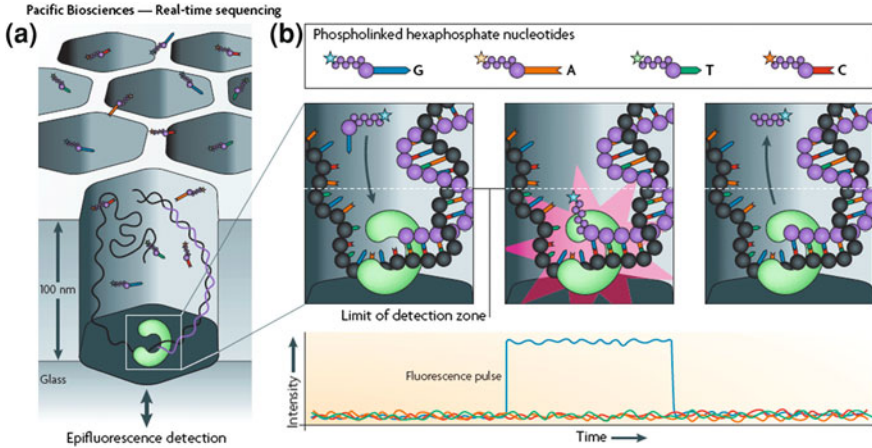


Fig. 2.9 **a** The zero-mode waveguide (ZMW) design with an attached polymerase at its bottom. **b** Each incorporated phospholinked nucleotide will reside on the enzyme's active site for a few milliseconds, which is enough time for a fluorescent signal to be recorded. The released labeled pentaphosphates will diffuse quickly [1]

platform, SMRT cells are used, with each cell having thousands of zero-mode waveguides (ZMWs), which are holes in a surface that acts as a nanoscale chamber. In each ZMW (which is tens of nanometers in diameter), a single molecule of DNA polymerase is attached to the bottom surface (Fig. 2.9a).

The accuracy and the speed of performance of the polymerase depend on high concentrations of nucleotides, and since the nucleotides are fluorescently labeled, this will lead to the background noise that creates difficulties in nucleotide incorporation detection. To overcome this problem, the detection volume in SMRT has been reduced to 20 zeptoliters (10^{-21} liters). This considerable reduction of detection volume can reduce the effect of background noise. One of the main differences between SMRT and previously described methods is the site of attachment of the fluorescent label. In other systems, the fluorescent label is attached to the base in nucleotides and consequently the labels remain attached after nucleotide incorporation, which leads to an increase in background noise. Moreover, incorporation of multiple bases will also lead to the creation of a steric hindrance as a result of the physical bulk of several dye molecules, which in turn leads to the limitation of enzyme activity. In SMRT technology, the fluorescent label is attached to the phosphate chain, and as a result of nucleotide incorporation the pentaphosphate-label couple will be removed from the nucleotides and will diffuse out of the reaction volume (Fig. 2.9b) [20], (Pacific Biosciences, 2009. Single Molecule Real Time (SMRTTM) DNA Sequencing) [1]).

2.8 Comparison of Available Next-Generation Sequencing Techniques

Already available next-generation sequencing methods are distinguished from each other based on various characteristics including maximum read length, number of reads in each run (degree of parallelization), run time, and many others. Table 2.1 compares the performances of various next-generation sequencing instruments [9]. Note that most of the information provided here is subject to rapid changes and it is suitable just for general comparison between techniques.

2.9 DNA Sequencing Costs

For many years, the National Human Genome Research Institute (NHGRI) has tracked the costs associated with DNA sequencing performed at the sequencing centers funded by the Institute. This information serves as an important benchmark for assessing improvements in DNA sequencing technologies and for establishing the DNA sequencing capacity of the NHGRI Genome Sequencing Program. The cost-accounting data presented by NHGRI is on the basis of (1) “Cost per megabase” (Mb: 1 million base) of DNA sequence (Fig. 2.10a); and (2) “Cost per genome” (human-size genome) (Fig. 2.10b). In each represented graph, a comparison has been made between the cost data and Moore’s Law, which describes a long-term trend in the computer hardware industry that involves the doubling of “compute power” every 2 years. Technology improvements that “keep up” with Moore’s Law are widely regarded to be doing exceedingly well, making it useful for comparison (National Human Genome Research Institute.genome.gov). The reduction of the cost of DNA sequencing per megabase and per genome from 2001 until 2012 is also represented in Table 2.2.

2.10 Sequencing Status

The number of completed sequenced genomes and the number of genome sequencing projects that are in progress are reported daily by Genome online database (GOLD). Table 2.3 shows the sequencing status until March 29, 2013.

Table 2.1 Comparison of available next-generation sequencing methods’ characteristics

Instrument	Read length (nucleotides)	No. of reads ^a	Output (Gb) ^a	No.of samples ^{a,b}	Runtime	Advantages	Disadvantages
Roche 454 GS FLX+	700 ^c	1×10^6	0.7	192 ^d	23 h	Long reads, short run time	Homopolymer errors, expensive
Illumina HiSeq2000	100 ^e	3×10^9	600	384	11 days ^f	High yield	No. of index tags limiting
Life technologies SOLiD 5500xl	75 ^g	15×10^9	180	1,152	14 days ^f	Inherent error correction	Short reads ⁹
Roche 454 GS junior	400 ^c	1×10^5	0.035	132	9 h	Long reads	Homopolymer errors, expensive
Illumina MiSeq	150	5×10^6	15	96	27 h	Short run time, ease of use	Expensive per base
Ion torrent PGM ion 316 chip	> 100 ^h	1×10^6	0.1	16	2 h	Short run time, low reagent cost	Not well evaluated
Helicos biosciences HeliScope	35 ^h	1×10^9	35	4,800	8 days	SMS, sequences RNA	Short reads, high error rate
Pacific biosciences PacBio RS	>1,000 ^h	1×10^5	0.1	1	90 min	SMS, long reads, short run time	High error rate, low yield

^a Numbers calculated for two flow cells on HiSeq2000 and SOLiD 5500xl
^b Calculated as no. of index tags (provided by the sequencing company) x no. of divisions on solid support
^c Average for single-end sequencing, paired-end reads are shorter
^d No. of reads decreases when the PicoTiterPlate is divided
^e 36 nucleotides for mate-pair reads
^f Run time depends on the read length, and on whether one or two flow cells are used
^g Second read in paired-end sequencing is limited to 35 nucleotides, and mate-pair reads to 60 nucleotides
^h Average
SMS = single molecule sequencing

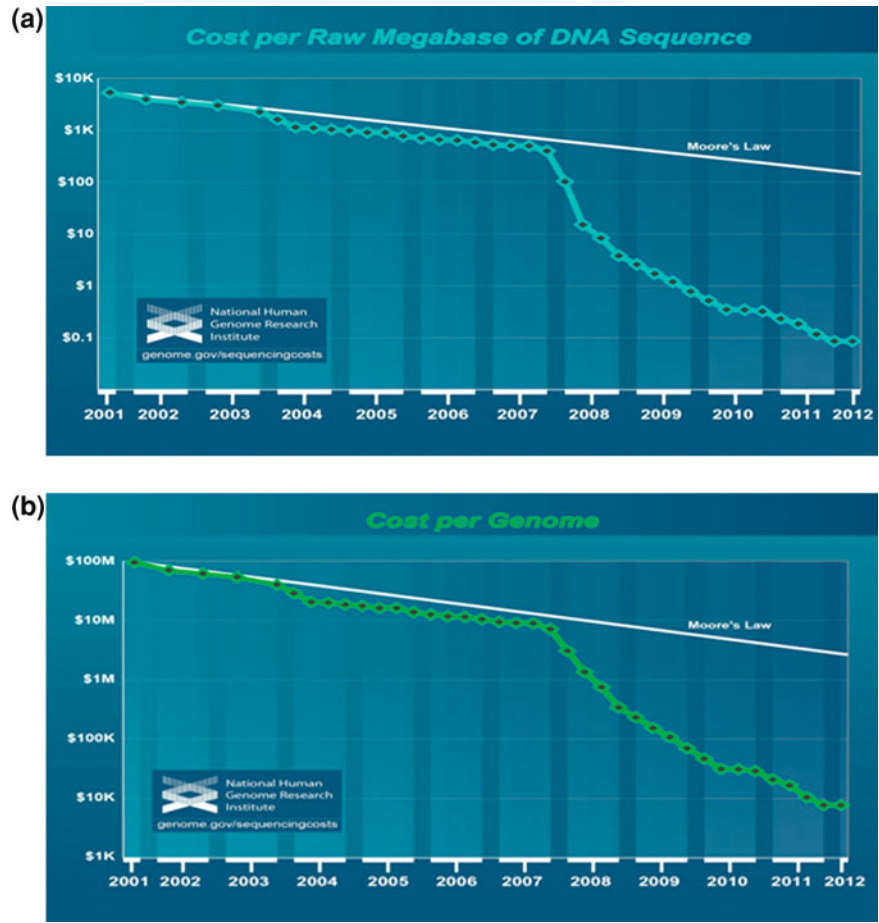


Fig. 2.10 Reduction in the cost of sequencing between 2001 and 2012. **a** The graph of the cost per megabase of DNA and **b** The graph of the cost per human-size genome. Also in each graph is hypothetical data reflecting Moore's Law (National Human Genome Research Institute.genome.gov)

2.11 Shortcoming of NGS Techniques: Short-Reads and Reads Accuracy Issues

The common problem with all current NGS technologies is the short length of reads (sequenced fragments) and higher error rate than those of the traditional Sanger sequencing method. The short-reads that are produced in these methods are especially problematic when large DNA fragments, e.g., a whole genome, is to be sequenced [1]. This drawback is an issue especially in sequencing new genomes

Table 2.2 Reduction in the cost of sequencing per megabase and per genome between 2001 and 2012

Date	Cost per Mb	Cost per genome	Date	Cost per Mb	Cost per genome
Sep-01	\$5,292.39	\$95,263,072	Jul-07	\$495.96	\$8,927,342
Mar-02	\$3,898.64	\$70,175,437	Oct-07	\$397.09	\$7,147,571
Sep-02	\$3,413.80	\$61,448,422	Jan-08	\$102.13	\$3,063,820
Mar-03	\$2,986.20	\$53,751,684	Apr-08	\$15.03	\$1,352,982
Oct-03	\$2,230.98	\$40,157,554	Jul-08	\$8.36	\$752,080
Jan-04	\$1,598.91	\$28,780,376	Oct-08	\$3.81	\$342,502
Apr-04	\$1,135.70	\$20,442,576	Jan-09	\$2.59	\$232,735
Jul-04	\$1,107.46	\$19,934,346	Apr-09	\$1.72	\$154,714
Oct-04	\$1,028.85	\$18,519,312	Jul-09	\$1.20	\$108,065
Jan-05	\$974.16	\$17,534,970	Oct-09	\$0.78	\$70,333
Apr-05	\$897.76	\$16,159,699	Jan-10	\$0.52	\$46,774
Jul-05	\$898.90	\$16,180,224	Apr-10	\$0.35	\$31,512
Oct-05	\$766.73	\$13,801,124	Jul-10	\$0.35	\$31,125
Jan-06	\$699.20	\$12,585,659	Oct-10	\$0.32	\$29,092
Apr-06	651.81	\$11,732,535	Jan-11	\$0.23	\$20,963
Jul-06	\$636.41	\$11,455,315	Apr-11	\$0.19	\$16,712
Oct-06	\$581.92	\$10,474,556	Jul-11	\$0.12	\$10,497
Jan-07	\$522.71	\$9,408,739	Oct-11	\$0.09	\$7,743
Apr-07	\$502.61	\$9,047,003	Jan-12	\$0.09	\$7,666



Table 2.3 Reported number of completed and in-progress genome sequencing projects

	Complete	In progress
<i>Archaea</i>	268	70
<i>Bacteria</i>	6,998	80,771
<i>Eukarya</i>	1,230	1,399

(with no prior knowledge of the relative position of fragments) and in sequencing highly rearranged genome segments such as one might discover in cancer genomes or in regions of structural variation. However, as mentioned before, using a paired-end sequencing approach would be beneficial to the utility of short-reads for sequencing de novo sequence assembly and for sequencing rearranged genomic segments [21]. Besides the paired-end sequencing technique, several new assembly algorithms began to be designed recently in order to overcome this problem [13]. The quality of each base is also lower than that of the Sanger chemistry reads. This problem was solved for the Sanger sequencing method through the validated base quality score originating from Phred programs. The quality scores compress a variety of types of information about the quality of base calls into a readily usable probability-of-error value. To deliver accurate results, many analysis tools and all assemblers require quality score input. To date, the quality scores offered by new

sequencing technology companies are not good substitutes for Phred scores. The accurate quality score is a key issue in various areas, such as sensitive and specific polymorphism detection, enabling accurate statistical modeling of the significance of read alignments and providing a quantitative basis for comparison of sequencing results from different technologies [22].

The template preparation process in all NGS methods involves “shearing the genomic DNA into small fragments” or “creating paired-end libraries.” The common feature that is observed in all NGS methods is that the resulting DNA templates are bound to the solid surface with specific distances from each other that allow millions of sequencing reaction to be performed in parallel [1].

2.12 NGS File Formats

Besides a few archives available for next-gen sequence data like NCBI’s Sequence Read Archive or SRA (<http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>), ENA-Reads at EBI (<http://www.ebi.ac.uk/embl/Documentation/ENA-Reads.html>), and the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/sub/trace_sra-e.html), there is other next-gen sequence data created and hosted by many authors and researchers [23].

Common next-generation sequencing file types among different data analysis and visualization tools are listed in Table 2.4 [23].

FASTQ format is the most common format for raw reads, as is BAM for aligned reads. FASTQ files are only kept before the alignment is done and the BAM file is generated.

2.13 NGS Applications

NGS application areas are defined in Table 2.5 (below) in terms of the particular source of nucleic acids selected for sequencing as well as the analysis strategy chosen to interpret the sequence information. A brief description of this application will be presented later. (Please see references for more details.)

RNA-Seq and small RNA sequencing

RNA-Seq, which was introduced as a new way for gene expression analysis, refers to sequencing the complex mixture of transcripts that were initially converted to cDNA using reverse transcription. RNA-Seq has resulted in the production of new data; for instance, finding novel spliced junctions, antisense regulation mode of action, or intragenic expression are among the many applications of RNA-Seq [24]. The term “small RNA” refers to microRNA and other non-coding RNAs. Sequencing of small RNA allows for accurate detection and quantification of rare small RNA sequences. Differential expression of known microRNAs, as well as finding novel microRNA targets, is also possible using NGS methods.

Table 2.4 NGS data formats

File extension	Description	Reference	Publication
.fasta	Classic DNA sequence file format	http://www.ncbi.nlm.nih.gov/blast/fasta.shtml	n/a
.ace	File format for whole-genome assemblies	Annotated in the documentation for CONSED, currently: http://www.phrap.org/consed/distributions/README.19.0.txt	Gordon et al. (1998)
.wig	A reference-genome indexed data series for “dense” and continuous data (such as %GC)	http://genome.ucsc.edu/goldenPath/help/wiggle.html	Kent & Haussler (2002)
.bed	A reference-genome indexed data series for “sparse” data (such as transcriptome data)	http://genome.ucsc.edu/goldenPath/help/bedgraph.html ,	Kent & Haussler (2002)
.tab	Tab-delimited text	N/A	n/a
.pdf	Portable document format	Either ISO-32000-1 or http://www.adobe.com/devnet/pdf/pdf_reference.html	n/a
.sam	“Sequence Alignment/Map” format	http://samtools.sourceforge.net/SAM1.pdf	Li (2009)
.bam	Binary format of sam	http://samtools.sourceforge.net/SAM1.pdf	Li (2009)
.fastq	Combination of sequences and quality scores in one file; mainly for data from Illumina sequencers in which case quality scores have been transformed.	http://maq.sourceforge.net/fastq.shtml	Li (2008) and Cock (2010)
.csfasta	Life technologies SOLiD colorspace fasta file—containing color calls (0, 1, 2, 3) rather than base calls	See: http://solidsoftw.aretools.com/	n/a
.qual	Per-base quality scores generated during basecalling. All but Illumina scores are scaled to estimate the probability of an incorrect base call, as is in common use for conventional sequencing as Phred quality scores.		n/a
.gff	A flexible format for annotating features (e.g. genes) on a sequence.	http://www.sanger.ac.uk/resources/software/gff/ .	n/a
.srf	“Short Read Format”—a new format proposed for short-read DNA sequence	http://srf.sf.net	n/a
.sff	Standard flowgram Format (specific for Roche/454)	http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=for mat58tm=docBts=format5#header-global	n/a
.gtf	Gene transfer format, an alternate format to GFF for specifying gene features	http://mblab.wustl.edu/GTF22.html	n/a

Table 2.5 A review of NGS applications

Name	Nucleic acid population	Brief analysis strategy
RNA-Seq	RNA (may be poly-A mRNA or total RNA)	Alignment of reads to “genes”; variations for detecting splice junctions and quantifying abundance
Small RNA sequencing	Small RNA (often miRNA)	Alignment of reads to small RNA references (e.g., miRbase), then to the genome; quantify abundance
ChIP-Seq	DNA bound to protein, captured via antibody (ChIP = Chromatin Immunoprecipitation)	Align reads to reference genome, identify peaks & motifs
RIP-Seq	RNA bound to protein, captured via antibody (RIP = RNA Immunoprecipitation)	Align reads to reference genome and/or “genes,” identify peaks and motifs
Methylation analysis	Select methylated genomic DNA regions, or convert methylated nucleotides to alternate forms	Align reads to reference and either identify peaks or regions of methylation
SNP calling/discovery	All or some genomic DNA or RNA	Either align reads to reference and identify statistically significant SNPs, or compare multiple samples to each other to identify SNPs
Structural variation analysis	Genomic DNA, with two reads (mate-pair reads) per DNA template	Align mate-pairs to reference sequence and interpret structural variants
de novo sequencing	Genomic DNA (possibly with external data e.g., cDNA, genomes of closely related species, etc.)	Piece-together reads to assemble contigs, scaffolds, and (ideally) whole-genome sequence
Metagenomics	Entire RNA or DNA from a (usually microbial) community	Phylogenetic analysis of sequences

ChiP-Seq technique and RIP-Seq

In ChIp-Seq (Chromatin immunoprecipitation), which is used to direct mapping of DNA binding site to a reference genome, Microarray has been replaced by NGS techniques [5]. Protein-bound DNA is captured through interaction of an antibody targeting protein. DNA is then treated with DNase to digest parts of DNA that are not bound to protein. In the next step, protein-bound DNA is separated from protein-part, sequenced and matched to the reference genome. RNA binding protein immunoprecipitation sequencing (RIP-Seq) is an analogue to ChiP-Seq, which is used to identify RNA molecules that are bound to a nuclear or cytoplasmic protein.

Methylation analysis

DNA methylation, which is the major form of epigenetic modification, influences cytosines in promoter sequences that are enriched in CGs (hence referred to as CpG islands) and converts it to 5 methyl cytosine. This modification is known to influence gene expression and it has been understood that specific patterns of methylation are implicated in cancer development. Decoding the genome wide methylation profile plays an important role in finding the correlation between DNA methylation and other epigenetic modifications, and NGS methods have helped in deciphering the methylation profile of the genome [25].

SNP calling/discovery

The discovery of SNPs, including single nucleotide mutation, insertion, and deletion in NGS data, is straightforward, using various methods and software tools. SNP discovery has helped population genetic studies due to the generation of relatively inexpensive, high-depth sequencing data through NGS methods.

Structural variation analysis

SNPs have long been thought to be the most common class of genetic variations and have been widely used in linkage and genome-wide association studies. However, structural variation analysis, which is typically concerned with identification of large-scale amplifications, deletions, translocations, or inversions, is now believed to be widespread in the human genome. NGS has been widely used in structural variation analysis [26].

Metagenomics

One of the early applications of NGS was to determine how all organisms live together in a certain ecosystem, such as in a deep mine or in the ocean. This field is called Meta-genomics; in it, all biological communities of an ecosystem are sequenced en masse to obtain information about all the organisms of that ecosystem. This approach enables the study of communities that could not be easily cultured and thereby provides an unbiased view of the composition and state of the community [27].

2.14 Summary

Recent high-throughput sequencing technologies have their origins in the results of the studies of Sanger's and Gilbert's groups, which introduced two distinct procedures for identifying the order of nucleotides in a DNA fragment in the late 1970s. Among these methods, Sanger's, which is also known as "chain termination" or the "dideoxy" method, became commercialized and has been in use for over 30 years. The massive sequencing projects, such as the Human Genome Project, were also based on Sanger's chemistry. The maximum length of the DNA fragment that can be sequenced by Sanger sequencing technology with high fidelity (1–2 % error rate) is about 1000 bp and as a result of using this technique, shearing the larger pieces of DNA is required. Assembling the sequenced reads is faced with difficulties originated from missing fragments, repeats and intrinsic sequencing errors. Moreover, the cost of the Sanger method is now about \$0.5 per 1 kb, which is still too high to sequence a whole human genome for each person (\$10,000–\$100,000) that can revolutionize the future medicine. The NIH goal that was announced in 2004 was to develop new sequencing technologies that will be able to sequence individual human genomes at a lower cost (\$1,000) and, of course, in just a few days.

Since then, many research groups have been working on developing new sequencing systems. First Roche/454 FLX introduced its platform, which was based on a pyrosequencing procedure that was commercially available in 2004. In early 2007, the Illumina (Solexa) Genome Analyzer released its platform, which became highly popular and was based on the bridge PCR. That same year, Applied Biosystems introduced the SOLiD sequencing system, which, in contrast to the two previously mentioned methods, basically works through ligation. All three of these techniques rely on an amplification step before sequencing, which is a costly and time-consuming steps. Ion semiconductor is another NGS method which works on the basis of pH fluctuation. In polonator sequencing technology, sequencing is performed using degenerate nonamers. In 2008, the Helicose sequencing platform, which requires no amplification step and therefore a single DNA molecule that can directly be sequenced became available. Helicose became a prototype for other groups to design new sequencing systems without any amplification step. Nanopore and single molecule real time (SMRT) are the two most promising platforms that are becoming commercialized. With these new sequencing technologies, which have revolutionized the field of genomics, it seems that the \$1,000 genome goal will be feasible in the near future. These high-throughput sequencing methods have introduced new fields in biology, which include personal genomics, analysis of RNA transcript, ChIP-seq (Chromatin Immunoprecipitation sequencing), ChIP-chip (Chromatin Immunoprecipitation coupled to DNA microarray) and many others.

References

1. Metzker, M. L. (2009). Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1), 31–46.
2. Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, 11(1), 378.
3. Dong, H., & Wang, S. (2012). Exploring the cancer genome in the era of next-generation sequencing. *Frontiers of medicine*, 6(1), 48–55.
4. Shendure, J., et al. (2004). Advanced sequencing technologies: Methods and goals. *Nature Reviews Genetics*, 5(5), 335–344.
5. Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4), 195–203.
6. Shendure, J., et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741), 1728–1732.
7. Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, 387–402.
8. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), 1135–1145.
9. Berglund, E. C., Kiialainen, A., & Syvänen, A. C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, 2(1), 1–15.
10. Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of applied genetics*, 52(4), 413–435.
11. Novais, R., & Thorstenson, Y. (2011). The evolution of Pyrosequencing® for microbiology: From genes to genomes. *Journal of Microbiological Methods*, 86(1), 1–7.
12. Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing—concepts and limitations. *BioEssays*, 32(6), 524–536.
13. Nowrousian, M. (2010). Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. *Eukaryotic Cell*, 9(9), 1300–1310.
14. Liu, L., et al. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012.
15. Hui, P. (2012). Next generation sequencing: chemistry, technology and applications. [Without Title], pp. 1–18.
16. Wash, S., & Image, C. (2008). DNA sequencing: generation next-next. *Nature Methods*, 5(3), 267.
17. McNally, B., et al. (2010). Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano Letters*, 10(6), 2237–2244.
18. Clarke, J., et al. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4), 265–270.
19. Tyagi, S., et al. (2000). Molecular beacons: hybridization probes for detection of nucleic acids in homogeneous solutions. *Nonradioactive Analysis of Biomolecules*, 2nd ed. C. Kessler, ed. Springer-Verlag, Berlin, pp. 606–616.
20. Eid, J., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910), 133–138.
21. Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5), 255–264.
22. Brockman, W., et al. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, 18(5), 763–770.
23. Hunicke-Smith, S., Next-generation sequencing data and workflows.
24. Tarazona, S., et al. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12), 2213–2223.

25. Zhang, Y., & Jeltsch, A. (2010). The application of next generation sequencing in DNA methylation analysis. *Genes*, 1(1), 85–101.
26. Xi, R., Kim, T.-M., & Park, P. J. (2010). Detecting structural variations in the human genome using next generation sequencing. *Briefings in Functional Genomics*, 9(5–6), 405–415.
27. Hugenholtz, P., & Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, 455(7212), 481–483.
28. Gordon, D., et al. (1998). Consed: A graphical tool for sequence finishing. *Genome research*, 8(3), 195–202.
29. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, D. (2002). *Genome Research*, 12(6), 996–1006.
30. Li, H., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.
31. Li, R., et al. (2008). SOAP: Short oligonucleotide alignment program. *Bioinformatics*, 24(5), 713–714.
32. Cock, P. J., et al. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6), 1767–1771.

Next Generation Sequencing and Sequence Assembly
Methodologies and Algorithms

Masoudi-Nejad, A.; Narimani, Z.; Hosseinkhan, N.

2013, X, 86 p. 46 illus., 43 illus. in color., Softcover

ISBN: 978-1-4614-7725-9