Next-Generation Sequencing: Technology, Advancements, and Applications

2

Gourja Bansal, Kiran Narta, and Manoj Ramesh Teltumbade

2.1 Introduction

The most comprehensive way of obtaining information about the genome of any living organism is to determine the precise order of nucleotides, known as sequencing, in its complete DNA sequence. Earlier, traditional methods which include Sanger's chain termination method and Maxam-Gilbert's chemical degradation method have been used for DNA sequencing. However, it is quite expensive and time consuming to sequence whole genome of an organism using the traditional method. Demand for low-cost and highly efficient sequencing gave rise to massively parallel sequencing technology which is known as "Next-Generation Sequencing (NGS)" (Koboldt et al. 2013). NGS refers to the high-throughput sequencing technologies which can simultaneously sequence millions or billions of DNA molecules.

Completion of Human Genome Project (HGP) in 2003 has marked a history in the field of genetics. Later the significant advancements have been made in sequencing technologies which decreased the cost of sequencing per base and increased the number of bases sequenced effectively per unit time (Metzker 2010).

Using the traditional sequencing method, it took nearly 15 years to sequence J.C. Venter genome as part of Human Genome Project, whereas, with the advent of NGS techniques, same can be completed in a couple of hours (Venter et al. 2015). In the current era, sequencing technologies have advanced to such a level that one can study the genome at a cellular level too. Single-cell sequencing techniques are now available which enable researchers to study cells individually rather than relying on an average signal from aggregate of cells (Wang and Navin 2015). In today's time, NGS methods have been applied to a variety of genomes ranging from singular to multicellular organisms. With the advent of NGS techniques, a number of

G. Bansal (✉) · K. Narta · M. R. Teltumbade
Institute of Integrative and Genomic Biology, New Delhi, India

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_2

applications and methods that leverage the power of genome-wide sequencing have increased at an exponential pace.

Rapid progress in sequencing techniques, as well as synchronized development in bioinformatics tools, has provided the solution to many different problems in the field of genetics and biology (Lelieveld et al. 2016). It allowed and helped researchers from diverse groups across the globe to generate draft genome sequence of any organism of their interest (Grada and Weinbrecht 2013).

Whole genome sequence of any organism cannot be read by traditional sequencers in a single run. Instead in a typical NGS run, thousands or millions of short overlapping sequences are produced concurrently in a single run. Each of the short sequence is called as a read. Reads are usually short (<= 250 bp) and can contain sequencing errors. To rule out sequencing errors, generally a genomic region is sequenced more than once. The number of reads spanning a genomic region determines the depth at which a genome is sequenced.

Along with whole genome sequencing, NGS techniques can also be applied to transcriptome sequencing (RNA-Seq), whole exome sequencing (WES), candidate gene sequencing (CGS), genotyping by sequencing (GBS), and chromatin immunoprecipitation sequencing also called as Chip-Seq (Furey 2012). Whole exome sequencing captures variations in all coding regions, or exons, of known genes and covers more than 95% of the total exons. As exome represents around 2% of the genome, it is a cost-effective alternative to whole genome sequencing if one is interested in finding variations only in coding part of the genome (Rabbani et al. 2014). RNA sequencing provides transcriptional activity of all coding as well as noncoding segments of the genome. As it can quantify alternative splice isoforms, RNA-Seq provides more accurate and precise measurements of gene expression levels than microarray platforms (Van Verk et al. 2013). Methylome sequencing complements genome sequencing as it determines the active methylation sites and provides a list of epigenetic markers that regulate gene expression, differentiation, and disease state (Schubeler 2015). Along with increasing our understanding toward genome sequence, sequencing methods also provide information about genetic variations, differential gene expression, and different aspects of transcriptional regulation.

There are several companies which make machines on which NGS can be done, such as Illumina (http://www.illumina.com), Roche (http://www.454.com), ABI/Life Technologies (http://www.lifetechnologies.com), Helicos BioSciences (http://www.helicosbio.com), Pacific Biosciences (www.pacificbiosciences.com), and Oxford Nanopore Technologies (http://www.nanoporetech.com). The different platforms vary in their sequencing technologies in terms of the sequencing chemistry, read length, number of reads per run, speed of sequencing, and cost per base pair sequenced (Goodwin et al. 2016). Table 2.1 provides a list of various NGS platforms along with their features.

In this chapter, we will be mainly focusing on the technology, advancements, and the applications of next-generation sequencing in the field of "-omics" development.

Table 2.1 Various NGS platforms and their features

Company	Sequencing principle	Detection	System platform	Read Length	No. of reads	Time/run	Throughput/ run	Accuracy (%)
Illumina	Reversible Terminator sequencing by synthesis	Fluorescence/Optical	HiSeq 2500/ 1500	36/50/ 100	3 billion (SE)	2~11 days	600 GB	>99
			Genome					
			Analyzer llx	35/50/75/ 100	320 billion (SE)	2~14 days	95 GB	>99
				25/36/ 100/25				
			MiSeq	0	17 million (SE)	4~27 h	8.5 GB	>99
Roche	Pyrosequencing	Optical	454 GS FLX+	700	1 million	23 h	0.7 GB	99.99
			455 GS Junior	400	1 million	10 h	0.035 GB	>99
Helicos Biosciences	Single molecule sequencing	Fluorescence/Optical	Heliscope	25~55	600~800 million	8 days	37 GB	99.99
ABI Life Technologies	Ligation	Fluorescence/Optical	5500 SOLiD	75 + 35	1.4 billion	7 days	90 GB	99.99
			5500×1 SOLiD	75 + 35	2.8 billion	7 days	180 GB	99.99
		Change in pH detected by Ion Sensitive Field effect	Ion Personal					
	Proton detection	Transistors	Genome Machine	35/200/ 400	12 million	2 h	2 GB	>99

(continued)

Table 2.1 (continued)

Company	Sequencing principle	Detection	System platform	Read Length	No. of reads	Time/run	Throughput/ run	Accuracy (%)
Pacific Bioscience	Real Time single molecule DNA sequencing	Fluorescence/Optical	PacBio RS	Average: 3000	~50 k	2 h	13 GB	84~85
Oxford Nanopore	Nanopore Exonuclease Sequencing	Electrical Conductivity	gridION	Tens of Kb	4~10 million	According to experiment	Tens of GB	96

2.2 History of NGS

The seeds of the genomics era were sown with the identification of DNA as the genetic material in 1952 by Alfred Hershey and Martha Chase (Hershey and Chase 1952). One year later, another breakthrough came with the discovery of the DNA double-helical structure by James Watson, Francis Crick, and Rosalind Franklin (Watson and Crick 1953). The basic knowledge of the genetic material led to the curiosity to know how a four-letter code (nucleotides) could govern all biological processes. The order of occurrence of the nucleotides (sequence) in the DNA seemed to play a major role in encoding the information in living organisms. This sequence in the DNA is conserved across generations of a species and sometimes even across different species. To unravel this information, many scientists focused on developing methods to decode the sequence of the genes and genomes that constitute organisms.

In 1977 Frederick Sanger and Alan Coulson sequenced the first genome, Phage Phi X-174 (PhiX), using a "plus and minus" method of sequencing (3). This technique used polyacrylamide gel for identification of the varied lengths of the amplified products. Since then the sequencing technologies have come a long way and have revolutionized the field of genomics. Aggressive research in the area of NGS has led to the development of novel chemistries and technologies, thereby increasing the speed and reducing the cost and time of sequencing (Grada and Weinbrecht 2013). This has led to an affordable sequencing of the human genome. Veritas Genetics has sequenced the human genome at a price of 1000 USD. This is a huge step in predictive and personalized medicine at an affordable price (Mardis 2006; Service 2006).

The efficiency of a sequencing technique is measured by its accuracy, speed, cost, and automation. Although there is no official classification, depending on the above parameters, the sequencing techniques can be broadly classified into three generations. The first generation includes the earliest sequencers (developed by Maxam-Gilbert and Sanger) where only small stretches of amplified DNA regions were sequenced. After that came the high-throughput sequencing technologies, which could sequence multiple DNA regions from multiple samples in one go and generate huge data output. These include the second-generation and the third-generation sequencers. Examples of second-generation technologies include Roche's 454, Illumina's Hiseq, and Life Technologies' SOLiD. The first- and second-generation sequencing techniques are dependent on amplification steps. This is mainly to ensure sufficient signal detection by the sequencer. The amplification comes with inherent biases and errors, which get incorporated into the resulting sequence. The third-generation sequencers do not require the amplification process. These sequencers can detect signals generated from a single molecule of DNA. They have longer reads, faster turnaround time, and higher output. The examples of third-generation sequencers include Helicos BioSciences' tSMS, Pacific Biosciences' SMRT sequencing, Illumina's Tru-seq Synthetic Long-Read technology, and Oxford Nanopore Technologies' sequencing platform. On the basis of advancements in sequencing technologies, sequencing era has been divided into multiple generations (Fig. 2.1).

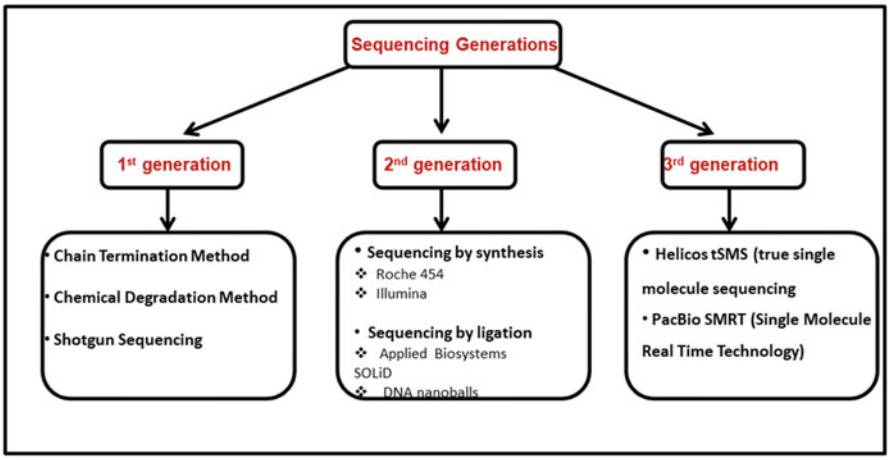

Fig. 2.1 Different sequencing generations

2.3 First-Generation Sequencers

First-generation sequencing techniques mainly used methods that could generate DNA fragments of different sizes. These methods used chemicals/enzymes to cleave DNA at specific sites or modified bases that could terminate the DNA amplification. The fragments hence generated were separated by electrophoresis using polyacrylamide (PAA) gel slabs. The two methods primarily used in the first-generation sequencers were chain termination method developed by Frederick Sanger and the chemical degradation method developed by Maxam-Gilbert (Morey et al. 2013).

2.3.1 Chemical Degradation Method

This approach was developed by Allan Maxam and Walter Gilbert (1977). In this method, each strand is chemically modified randomly, such that the backbone is exposed for degradation by alkali treatment at specific points. This process generates fragments of variable size. The fragmented DNA is terminally (both 3′ and 5′) radiolabeled with 32P and denatured. This is carried out as four reactions depending on the chemical treatment (e.g., G, A + G, C, C + T). The cleaved 32P-ssDNA is run on PAA gel. Then the autoradiograph is obtained from which the sequence of the fragment can be identified. This method involves the use of hazardous chemicals like dimethyl sulfate (for G and with acidic conditions release A), hydrazine, piperidine (for T and C), and NaCl (only C) (Maxam and Gilbert 1977). Also a huge amount of DNA is required. This technique is now obsolete.

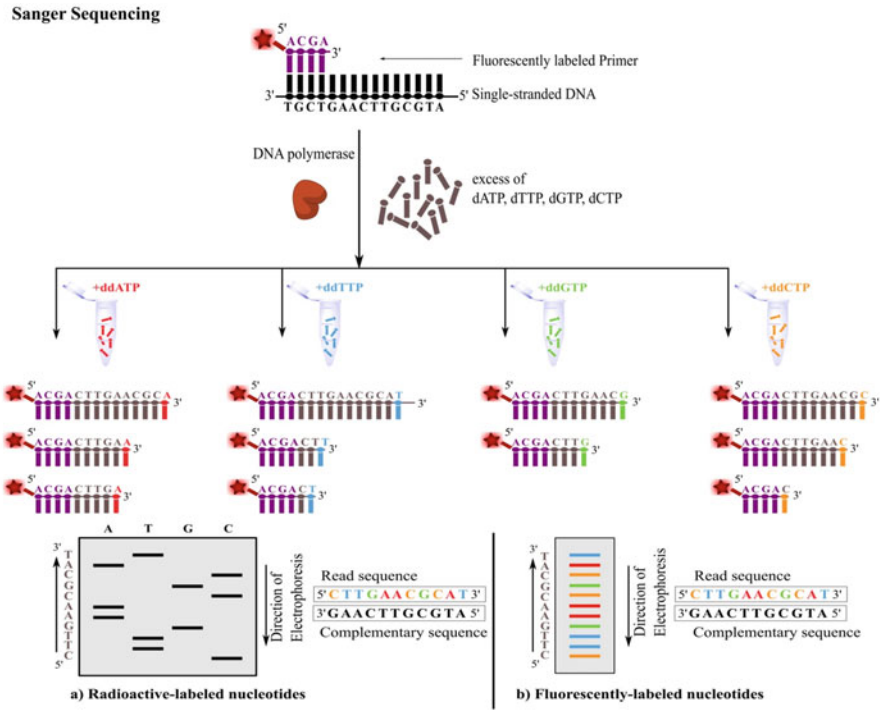

Fig. 2.2 Sanger di-deoxy chain termination method for sequencing: in Sanger sequencing, the template DNA is first primed with a fluorescently labeled (optional) primer. Then a sequencing reaction is carried out. The essential ingredients include a DNA template, primer, DNA polymerase, excess of dNTPs, and chain-terminating ddNTPs. The sequencing reaction can be performed either in four separate tubes, one for each ddNTP (**a**), or if fluorescently labeled ddNTPs are used, then all reactions can be performed in a single tube (**b**). After multiple rounds of template extension, the DNA fragments are denatured. The denatured fragments are run in gel slabs (now capillaries containing polymer) that separate the amplified products depending on their size. The sequence is then deciphered by the relative position of the bands from bottom to top

2.3.2 Chain Termination Method

In 1977, Frederick Sanger and group published the chain termination method of sequencing, also called as sequencing by chain termination (Sanger et al. 1977). This is now known as the Sanger method. Even today variations of this method are widely used by different sequencing techniques.

In the Sanger method (Fig. 2.2), the sample is first denatured by heat, and then four reactions are performed with ssDNA. Each tube contains a primer, DNA polymerase 1(Klenow enzyme), and all four dNTPs and one of four ddNTPs. The ddNTPs have hydrogen (instead of hydroxy-) group at the 3′ terminal. The amplification is carried out by extension of the primer, using single-stranded DNA (ssDNA) as a template. The presence of dNTPs and specific ddNTPs can randomly terminate the extending DNA chain. The DNA is then denatured and run on PAA gel. The

bands hence obtained are combined to form a single sequence (Sanger et al. 1977). Initially, radioisotopes were used to label dNTPs (Fig. 2.2a); later it was completely replaced by fluorescent dyes (Fig. 2.2b) (Smith et al. 1986).

Sanger sequencer ABI 370 was the first automated sequencer and it was launched in 1986 by Applied Biosystems (now Life Technologies). This included some significant improvements in the method like running DNA in capillaries instead of gel slabs, introduction of dye-labeled ddNTPs making way for one tube reactions, multicapillary electrophoresis, and automatic loading of samples (Ansorge et al. 1986; 1987).

Sanger sequencing was widely used in the 1990s to sequence genes and genomes. Even today it plays a very important part in screening genes for disease mutations and validation of data from the next-generation sequencers. The average read length of sequence from Sanger data is still more than most of the second-generation sequencers (Treangen and Salzberg 2011). Sanger sequencing formed the basis of the first draft of the human genome. In 2001 two landmark papers were published, which reported the sequencing of the human genome (Lander et al. 2001; Venter et al. 2001). Celera Genomics used shotgun sequencing in which a large piece of DNA is fragmented mechanically. Each fragment is sequenced independently, using the Sanger method. The sequences obtained are then assembled using the overlapping regions to get a complete sequence (Anderson 1981). Shotgun sequencing can be considered as the bridge between the first-generation and the second-generation sequencers.

After the completion of the Human Genome Project, scientists everywhere realized the enormous potential in identifying the DNA/RNA sequence information of an organism. The primary limitation of the first-generation sequencers was their low output and inability to scale up. The cost per base sequenced is also very high as compared to the high-throughput methods (Mardis 2011). To overcome these issues, automated, faster, and cheaper sequencers were developed. They were primed to sequence longer and large number of DNA molecules parallelly.

2.4 Second-Generation Sequencers

The second-generation sequencers can generate a huge amount of data in one run at a much lower cost and higher speed as compared to the first-generation sequencers. These sequencers use amplified DNA fragments and the sequencing is performed in parallel for millions of DNA fragments which is why it is also called as the massively parallel sequencing. The second-generation sequencers include three major processes: library preparation, amplification, and imaging/sequencing (Mardis 2008). These steps may vary in different sequencers of this generation.

The basic steps in next-generation sequencing are represented in Fig. 2.3.

Library Preparation It primarily involves fragmentation of the DNA and adapter ligation.

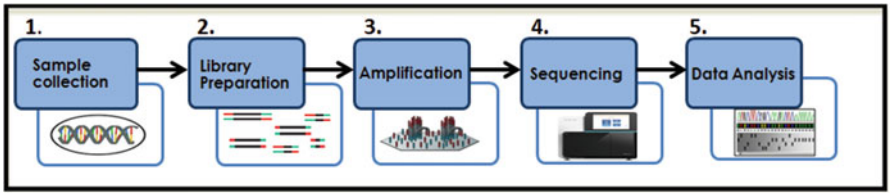

Fig. 2.3 Basic steps in next-generation sequencing

The second-generation sequencing techniques can only sequence small stretches of DNA molecules. Therefore, it is important to fragment the DNA molecules such that they can be sequenced and after that reassembled. Fragmentation can be either mechanical shearing or enzymatic cleavage (Morey et al. 2013).

The fragmented DNA is attached to universal adapters to facilitate amplification and attachment to a surface for sequencing. Adapters are double stranded and have sites for primer binding. They also have index sequences which are used to differentiate the reads coming from various samples, if multiple samples are pooled together during sequencing. Primer binding sites are used to prime the sequencing reaction (Morey et al. 2013). The steps involved in library preparation are shown in Fig. 2.4.

Amplification of Template Most of the second-generation sequencers are not able to detect fluorescence from a single DNA molecule. To overcome this, the DNA molecules are attached/immobilized on a surface, which are then amplified (Morey et al. 2013). This enables the sequencers to capture a clear signal while imaging. Two major amplification techniques are:

(i) Emulsion PCR: Used by 454 (Roche), SOLiD, and Ion Torrent (Thermo Fisher)
(ii) Solid-phase amplification: Illumina

This step produces clonal templates for sequencing (Goodwin et al. 2016).

Sequencing The amplified products from the previous step are sequenced in this step. A sequencing primer is added to the templates to start the addition of bases and their simultaneous imaging. These steps are carried out in a cyclic fashion. Sequencing can be on the basis of one of the following two principles (Goodwin et al. 2016):

(a) *Sequencing by synthesis*: This technique makes use of the DNA polymerase to add bases sequentially. The major platforms which use this approach are 454, Illumina, Qiagen, and Ion Torrent.
(b) *Sequencing by ligation*: In this technique, a fluorophore-bound probe is hybridized to the template and ligated to the former previous base. Once ligation is complete, the probes are imaged to identify the bases. SOLiD and Complete Genomics use this method of sequencing.

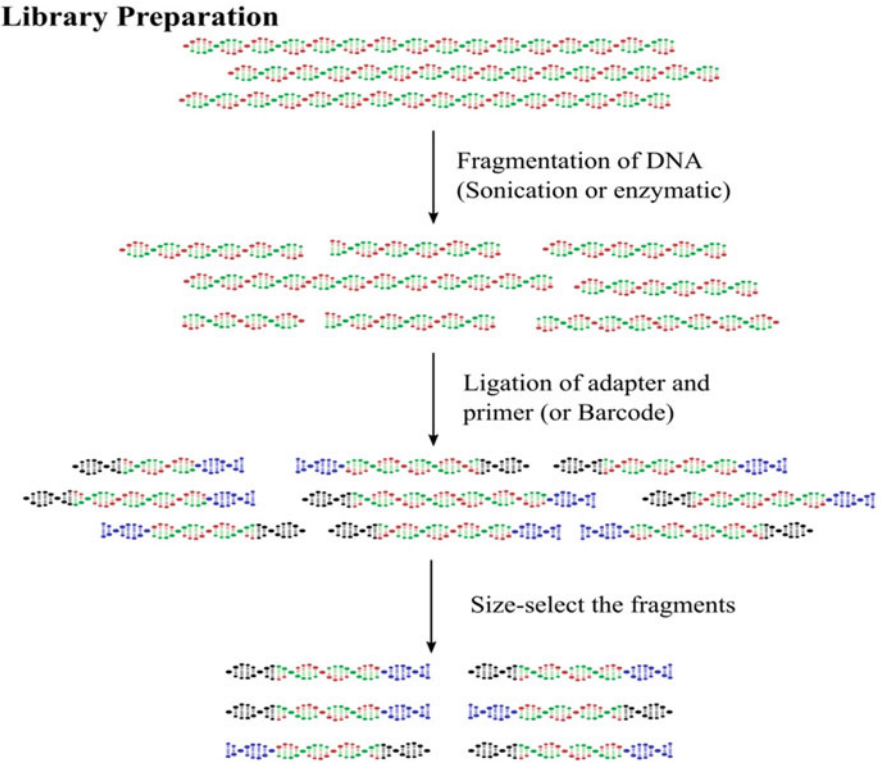

Fig. 2.4 Diagrammatic view of library preparation: The first step involves fragmentation of the DNA molecule. It can be carried out by (a) enzymatic methods by nonspecific endonuclease treatments or tagmentation using transposase and (b) physical methods using sonication or acoustic shearing using Covaris. In the case of RNA sequencing, the fragmentation is usually done by heating in the presence of divalent cations and then the fragmented RNAs are converted to cDNA After fragmentation the ends are repaired, i.e., the ends are blunted, the 5′ ends are phosphorylated, and the 3′ ends are adenylated. The adapters are ligated to the fragments (represented here by blue and black). The adapters are barcoded, using index sequences, so that multiple samples can be sequenced together. The part of the template between the adapters (they are of constant length) is called the insert, and it determines the library size. The insert size is in turn determined by the application and the technology used for sequencing
Size selection involves the steps to obtain the library size within a desired range and to remove adapter dimers. This is usually done with the help of magnetic beads (e.g., Agencourt AMPure XP beads) or the library is run on agarose gel. Once the library is prepared, its quality and quantity is checked before moving on to sequencing

The basic principle involves the cyclic process of addition of bases to the primer until the required number of bases is read (imaged) from the template. All the reactions are performed in parallel.

Imaging of Bases and Data Analysis After the sequencing is performed, the information from images is converted to bases. This is generally carried out using platform-specific software which generates raw files of sequence data for further processing and analysis.

This is the step from where data comes out of the experimental lab to a high configuration computer. Sequence analysis is done by a bioinformatician who analyzes the data and draws meaningful insights out of it. Different open-source algorithms are available for each step of analysis workflow.

Some of the widely used second-generation sequencers are discussed here:

2.4.1 Roche 454 Genome Sequencer

Genome Sequencer GS20 was the first commercialized second-generation sequencer launched in 2005 (www.454.com). It was developed in 1996 at the Stockholm Royal Institute of Technology *and* has been used to sequence Neanderthal, barley, and *Helicobacter pylori* genomes (Rothberg and Leamon 2008).

Library Preparation
Nebulizer randomly fragments the DNA. These fragments are then flanked by two types of adapters on different sides and denatured. One of the adapters contains the sequencing primer binding site and the other adapter has a biotin label. Only the fragments containing different adapters are selected and mixed with capture beads. The capture beads have probes complementary to adapter containing the biotin label so DNA fragments can bind to them. Excess of capture beads are added to bind only one DNA molecule to each bead (Fig. 2.5b).

Amplification
This technique uses emulsion PCR to amplify the DNA fragments bound to the beads clonally. By the end of this process, there are millions of clonal molecules on the beads, which are denatured to obtain only ssDNA molecule. Sequencing primer is attached to the ssDNA adapters.

Sequencing
The beads are then loaded on a picotiter plate containing wells. The dimensions of wells are such that only one bead enters the individual well. Later smaller packing beads containing immobilized sulfurylase and luciferase are added to the wells. 454 sequencing is based on pyrosequencing (Fig. 2.5b). In this technique, as the nucleotides are incorporated, pyrophosphate (PPi) is released. PPi is then converted to ATP using ATP sulfurylase and adenosine phosphosulfate. This ATP combines with luciferase to convert luciferin to oxyluciferin. This generates a light signal, which is detected by a charge-coupled device (CCD) at the bottom of the plate. Each nucleotide is given one at a time in a predesigned order. Incorporation of more than one nucleotide on the same molecule is read as stronger intensity. Therefore, the

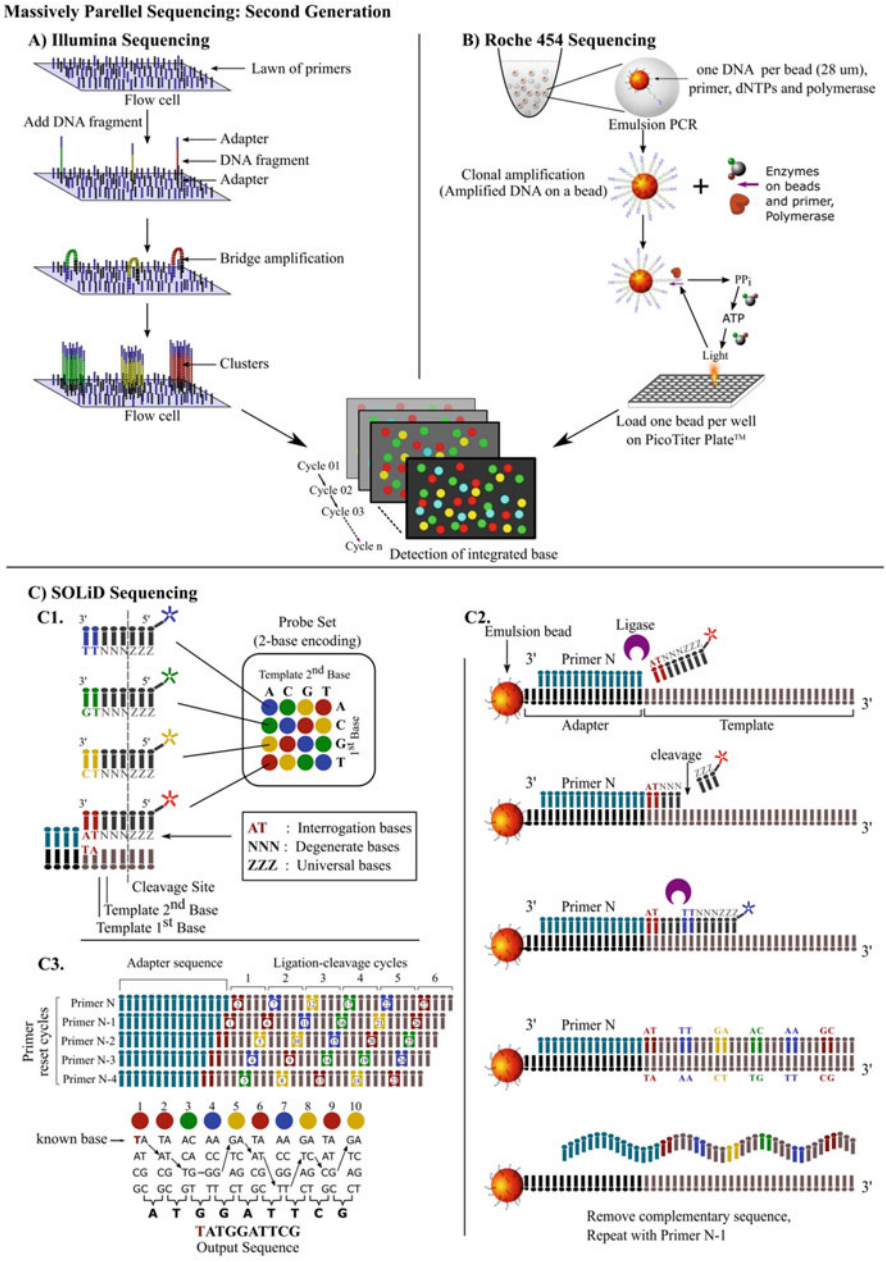

Fig. 2.5 Different platforms for second-generation sequencing
(**a**) Illumina sequencing: The prepared library is distributed over a flow cell. The flow cell contains a lawn of primers that are complementary to the ends of the adapters, as a result of which the DNA fragments bind to them. Solid-phase bridge amplification involves amplification of the attached templates in the presence of unlabeled nucleotides, polymerase, and buffer. The dsDNA is then

amount of light can be regressed to calculate the number of similar nucleotides incorporated. After the completion of required number of cycles, the sequence is acquired by combining the information generated (Rothberg and Leamon 2008).

This technique can generate reads up to 1 Kb in length and faces major problems in identifying homopolymer stretches in the genome. Moreover, it has lower output and sequence per base cost is higher than its competitors (Mardis 2008; Margulies et al. 2005).

2.4.2 Illumina

It is one of the most successful sequencing platforms. The technology, developed by Solexa, was first commercialized as Genome Analyzer (GA) in 2006. One year later,

Fig. 2.5 (continued) denatured and the original template is washed away, leaving behind the primer and the elongated strand. This process is repeated for a number of cycles to generate clusters from each attached template. After cluster generation, the strands are sequenced. The sequencing is initiated once the sequencing primers are attached and the reagents containing labeled nucleotides and polymerase are added. As each base is added, it is imaged after removing the unincorporated bases. The fluorescent signal is then removed from the incorporated nucleotides. Multiple such cycles are carried out to obtain sequential images of bases of each cluster
(**b**) Roche 454 sequencing: In this, the adapter-ligated fragments are attached to the beads. Then amplification process takes place by emulsion PCR. One bead per well is distributed onto a 454 picotiter plate. The amplification is then carried out by pyrosequencing in which, unlike Illumina, only one type of deoxynucleotide triphosphate base is provided to be incorporated by the polymerase into a cycle. This addition is accompanied by the release of pyrophosphate (PPi). With the help of enzymes (ATP sulfurylase, luciferase, and luciferin attached to the bead), the PPi is converted to light, which is detected. The base addition in each well is recorded for the desired number of cycles, and simultaneously the sequences on the template are deciphered
(**c**) SOLiD sequencing: The emulsion PCR (emPCR) amplification process is similar to Roche. Then the beads with amplified template are deposited on a slide. SOLiD uses the sequencing by ligation method. (c1) SOLiD uses two-base encoding and uses four different colored fluorescence probes. In the probe color matrix, each color represents 4 out of 16 possible combinations. The probes are made up of eight bases. The first two are the template bases, which match read positions on the sequence to be read. The next three are degenerate bases that match the three unread bases upstream to the template bases. The identity of these three bases is not needed for sequencing. Finally, the three universal bases can bind to any of the nucleotides. They have a fluorescent dye attached at 5′ end and has cleavage site at the 3′ end. (c2) The basis of SOLiD is that the labeled probes get ligated to the primer only if they are perfectly matched. Once ligated, the three universal bases at the 5′ end are removed. The remaining ligated probe acts as a primer for the next probe. This process is carried out for the desired number of ligation cleave cycles. Then the extension product is removed and the template is reset with shorter $n-1$ primer revealing a thymidine (T) at the adapter for next round. (c3) The numbers enclosed in white circles represent the sequence of base position in the template. The cycle numbers are denoted on the top. It requires five iterations (n, $n-1$, $n-2$, n-3, $n-4$) to decipher the complete sequence and to fill the gaps (of the two template bases and three degenerate bases) formed during the first iteration. Also represented here is the walk-through to obtain sequence information from multiple iterations

Solexa was purchased by Illumina (Treangen and Salzberg 2011). It uses reversible terminator reactions to carry out sequencing by synthesis.

Library Preparation DNA is first randomly broken into fragments, by mechanical shearing (using Covaris) or by enzymes/transposomes. Generally 200–300 bp fragments are selected using gel or SPRI/Ampure beads and adapters are ligated to them. These are then denatured and injected onto a solid surface called flow cell. The adapters are ligated to the end of the fragments such that one end serves to attach to the flow cell and other is used for sequencing (Fig. 2.5a).

Amplification The surface of a flow cell has a lawn of probes, which has sequences complementary to one end of the adapters bound to the DNA fragments. When the adapter-ligated fragments are distributed over a flow cell, they bind to complementary probes. To obtain clonal copies of the DNA, a process called bridge amplification is used. By the end of this process, thousands of copies are generated for each attached fragment, which corresponds to a cluster. Reverse strands are cleaved and the single strands are primed for sequencing (Adessi et al. 2000).

Sequencing by Synthesis The amplified products are sequenced by a process similar to the Sanger's chain termination reaction. Here the nucleotides are labeled with four different fluorescent dyes and are capable of reversible termination. During sequencing, the flow cell is flushed with all four nucleotides. A complementary fluorescently labeled base is added to the primed template by the polymerase bound to the template. After addition of a single base, the reaction is terminated. To obtain signal from the bound nucleotide, the remaining unbound nucleotides are washed away. The bound-labeled nucleotides are then illuminated with the help of lasers. Imaging is performed to identify location and type of base incorporated in each cluster. Then the fluorescent label is cleaved thereby exposing the 3′-OH group, to which a base can be integrated. This process is carried out with the help of Tris (2-carboxyethyl) phosphine (TCEP, reducing agent).

This cycle of addition of dNTPs, imaging, and cleavage is carried out until all the bases are read. Then all the images can be simultaneously combined to create the sequence for each cluster.

Illumina provides with the ability to sequence DNA from one end (single-end sequencing) or both ends of the fragment (paired-end sequencing) and mate pair sequencing (used to sequence the ends of long fragments, ignoring the bases in between) (Fuller et al. 2009; Pettersson et al. 2009; Rothberg and Leamon 2008).

Over the past decade, this technology is continuously improving its sequencers in terms of efficiency and accuracy. It generates one of the highest outputs with lowest reagent cost among all the sequencers present to date (Liu et al. 2012).

Applied Biosystems' SOLiD

George Church in 2005 developed small oligonucleotide ligation and detection (SOLiD) system for high-throughput DNA sequencing (Shendure et al. 2005). It

was commercialized by Applied Biosystems (now Life Technologies) in 2007. Its principle involves sequencing by ligation (SBL).

Sample Preparation and Amplification
SOLiD uses emulsion PCR, similar to Roche's 454. Fragmentation is achieved by nebulization/sonication or digestion. Universal adapters are attached to the ends of the fragmented template which are then deposited onto microbeads. The templates undergo clonal amplification reaction in water/oil emulsion microdroplets. The microbeads are then distributed on a glass slide to which they bind covalently. Based on application, slides may have one, four, or eight compartments**.**

Sequencing and Imaging:
The sequencing reaction starts by annealing of primer to the amplified template. In SOLiD sequencing, each cycle constitutes the following four steps:

1. The chemical reaction involves the binding of eight-nucleotide-long probes. Only the first two bases (di-base) at 3′end have a known sequence. The rest of the probe is degenerate. The probe is fluorescently labeled at the last base on 5′ end such that it corresponds to a specific di-base. The probe binds to complementary sequence next to the primer. Due to restriction in available fluorescent dyes, the complementary and reverse di-bases are encoded by the same color (FAM for AA, CC, GG, TT; Cy3 for AC, CA, TG, GT; TXR for AG, GA, TC, CT; Cy5 for AT, TA, CG, GC) (Fig. 2.5c1). After the primer is ligated to the adapter, octamer probes with same fluorescent label are added (Fig. 2.5c2).
2. When a complementary probe binds to the template, DNA ligase hybridizes the probe to the primer. During this process, a fluorescent signal is emitted, which is captured by the detector.
3. After ligation, the three bases (including the dye) at the 5′end of the probe are cleaved.
4. These steps are repeated with the three remaining fluorescent dye pool of probes. After each successful annealing, each probe is ligated to the previous probe in the second step. So at the end of one cycle, two bases are read per three skipped bases per probe. This process can be carried out for the desired number of times (Fig. 2.5c2).

After this, the primer along with all the probes is removed. Then a new primer is added such that it anneals to the penultimate base from the adapter-template junction ($n-1$). The abovementioned steps are repeated. This cycle is carried out four times (for $n-2$, $n-3$, and $n-4$ also) and every time primer shifted one base toward 5′ end (Fig. 2.5c3) (Mardis 2008; Valouev et al. 2008).

Data Analysis: Exact Call Chemistry It uses eight-base interrogation system, with four different colored primers to map possible combinations in sequences.

SOLiD sequencers are known to have problems while handling palindromic sequences. However, they are less error prone, as each base is read twice as compared to other second-generation sequencers. They are flexible, allowing for sequencing in different applications. It detects single nucleotide variants (SNVs) and insertion/deletion (indels) with ease.

Despite a high output and multi-sample processing by the second-generation sequencers, there is still scope for improvement. Second-generation sequencers face issues related to errors due to amplification and need for repeated "wash and scan" cycles (Metzker 2010) which are not only time consuming but also lead to asynchronous (dephasing) sequencing (Whiteford et al. 2009). These issues result in erroneous base calls and also limited read lengths (Metzker 2010).

2.5 Third-Generation Sequencing

Given a very rapid evolution of the successive sequencing technologies over the past few years, and therefore small time lapse, there has been a continuum of improvements among the successive next-generation sequencers. The next-generation sequencers are considered as the third generation primarily on the basis of the following features: First, they do not require amplification of template DNA. Second, the sequencing is performed in real time. They also do not require repeated "wash and scan" cycles. These sequencers generate longer reads and have higher speed and accuracy with lower cost and effort (Gut 2013; Heather and Chain 2016; Morey et al. 2013; Niedringhaus et al. 2011; Pareek et al. 2011; Schadt et al. 2010).

Three of the third-generation techniques are discussed below.

2.5.1 Helicos: tSMS (True Single-Molecule Sequencing)

Helicos BioSciences' tSMS was the first commercially available third-generation single-molecule sequencer (Heather and Chain 2016).

Library Preparation The DNA is broken down into 100–200 bp fragments, and a poly A sequence of approximately 50 bp is attached to 3′ end of each fragment. The fragments are labeled with fluorescent adenosine. These labeled fragments serve as templates for sequencing and are hybridized on the surface of a poly-T-containing flow cell. The flow cell containing 25 channels has oligo-dT (50 bases) primer attached to the surface. The 3′OH of the tailed molecules are blocked by terminal transferase and dideoxynucleotides to prevent extension.

Sequencing Before sequencing begins, the location of each fluorescently labeled template is captured, by illuminating with a laser. After imaging the fluorescent label is washed away. To start sequencing reaction reversible terminator fluorescently

labeled nucleotides (Bowers et al. 2009) and DNA polymerase is added to the flow cell. The sequencing chemistry is similar to Illumina, where signals are captured after a laser illuminates the flow cell. In the case of Helicos, a single type of nucleotide is added at a time (e.g., A). The camera records the addition of each nucleotide on a single DNA fragment. After imaging, the labels are cleaved and washed. This process takes place for the remaining three bases also (C, G, T). This cycle of sequencing is repeated until the required read length is achieved (Fig. 2.6a).

This method requires shorter sample preparation time and can be used to sequence degraded molecules also. A higher accuracy is achieved, as there is no PCR amplification step, but the sequencing time is long due to repeated cleaving and washing steps and also per base cost is high.

2.5.2 Single-Molecule Real-Time Technology (SMRT)

Single-molecule real-time technology was developed by Pacific Biosciences**.**

Library Preparation The DNA fragmentation is performed depending on required insert size, with a range from 500 bp to 10 kb. End repair is carried out to create blunt ends and addition of dA tail. Then SMRTbell hairpin loop adapters are ligated to both ends of the double-stranded fragments. A SMRT library is prepared after purification steps which ensure that only the fragments having adapter ligated to both ends are selected. A Φ29 DNA polymerase is attached to the DNA molecules of the library. This enzyme also has a strand displacement property, so the double-stranded DNA can be opened up into circular template (Eid et al. 2009).

Sequencing The sequencing reactions are carried out in a chip containing small wells (10^–21 L). Each of these reaction cells, also called zero mode waveguide (ZMW), has a molecule of Φ29 DNA polymerase attached to the bottom. ZMW are small pores surrounded by metal film and silicon dioxide (Foquet et al. 2008)**.** Once the template is added, it binds to the DNA polymerase. Then the fluorescently labeled dNTPs are added to the wells. All four nucleotides are phospho-linked and have different colored fluorophores (Korlach et al. 2008a). In this method, the fluorescence is attached to the terminal phosphate of the nucleotide (instead of the base as in previous cases).

During sequencing, the complementary dNTP enters the polymerase and emits a fluorescence signal in the ZMW. This signal is detected as a light pulse in the detection volume of 20 zeptoliters (Korlach et al. 2008b). The fluorescence label is released after cleaving the phosphate chain. Then a new base is incorporated (Fig. 2.6b).

This is a high-speed process as ~10 nucleotides can be added in a second.

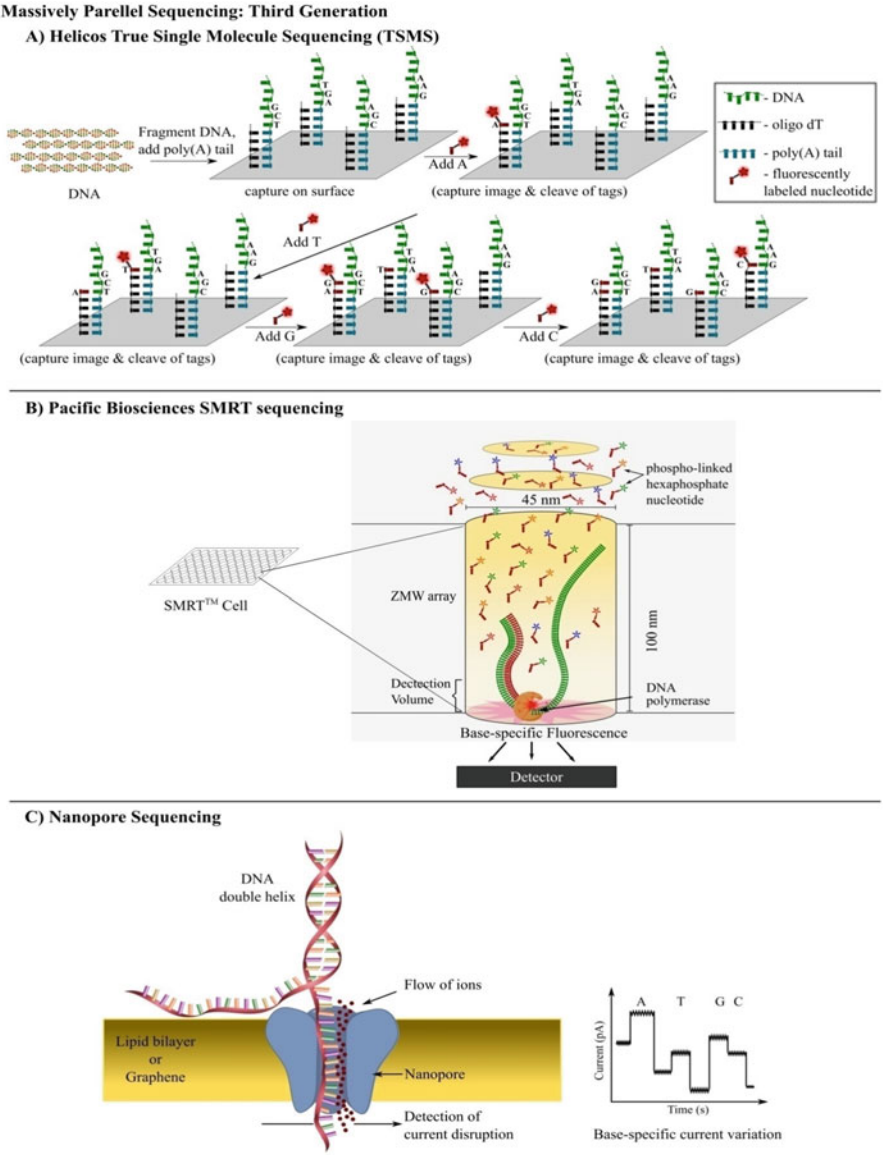

Fig. 2.6 Detailed view of third-generation sequencing platforms
(**a**) TSMS: The poly A tailed fragmented strands are hybridized to the poly-T-bound Helicos flow cell plate. Then the fluorescent labeled nucleotides are added one at a time. The addition is done in a cycle of "quads," where a quad consists of adding each base (A, T, G, and C) once. The labeled bases are then illuminated by a laser, and the images are taken, which helps in detecting the strands that have bound nucleotides. Before adding the new labeled bases, the labels from the hybridized *bases are cleaved*
(**b**) SMRT: A SMRT cell with ZMW nanostructures has a DNA polymerase immobilized to the bottom of the well. The fluorescently labeled phospho-linked nucleotides are added to the primed

2.5.3 Nanopore Sequencing

Nanopore-based chemical and biological molecule detection is among the most advanced sequencing technologies available today (Morey et al. 2013). The nanopores can be synthetic solid-state or biological in nature. The biological nanopores are modeled on the transporters and ion channels present inside the living cell (Haque et al. 2013). The nanopores decode a sequence, as the string of DNA is transported through it. They capture the modulation in the ion flow through these channels or optical signals in real time (Astier et al. 2006).

The α-hemolysin ion channels were the first ones to be used for this purpose (Kasianowicz et al. 1996) and were commercialized by Oxford Nanopore Technologies (Schadt et al. 2010). The modified α-hemolysin protein is embedded in a lipid bilayer and has an exonuclease on the outer surface, and a cyclodextrin sensor is attached on the inside. The exonuclease cleaves each base as it enters the pore. As the base crosses the channel, the variation in current is detected, which correlates with the specific parameters of the nucleotide (Fig. 2.6c).

Improvements in this technique and various other approaches have led to more accurate and a variety of nanopore sequencing platforms. These include:

(a) Using *Mycobacterium smegmatis* porin A (MspA) protein to sequence an intact ssDNA (Derrington et al. 2010).
(b) Optical detection in nanopore sequencing via multi-colored readouts using synthetic DNA (McNally et al. 2010).
(c) Synthetic material has also been incorporated for improvements in this technology; among them solid-state graphene nanopores and carbon nanotubes are of particular interest (Bayley 2010; Liu et al. 2010; Schneider et al. 2010; Zhao et al. 2012).

The nanopore sequencing is inexpensive as there is no addition of modified/fluorescent bases. Nanopore sequencing is marketed by Oxford Nanopore Technologies through their sequencing platform GridION along with a portable device MinION and the scalable PromethION (https://nanoporetech.com).

Fig. 2.6 (continued) DNA template (green). A signal is recorded when a base is bound to the template in the active site. The fluorophores are activated by the lasers only when they are in the detection volume. As the detection volume is minimal, bottom 20–30 nm, therefore only the correctly bound nucleotide is detected. After the phosphodiester-bond formation, the template is translocated so that next base can be attached. The location of the detector for the optical image is under the nanostructure. (**c**) A voltage-biased membrane (lipid bilayer/graphene) separates two aqueous electrolytes containing chambers. A flow of ionic current occurs through pores (blue) present across the membrane. The passage of the DNA is controlled by the enzymes present in the nanopore, as a result of which there is a disruption in the passage of the ions, which is measured by the very sensitive ammeter. A record showing the measurement of the passage of ions corresponding to the type of nucleotide crossing the pore is also represented alongside

The third-generation sequencing techniques promise to deliver longer reads than any of the previous technology. These long reads (5–15 kb) are proving to be important in many areas (Lee et al. 2016)**.** They are instrumental to fill the gaps in the human genome (Chaisson et al. 2015; Pendleton et al. 2015) *and also* used to get highly accurate reassembly and reconstructs of many bacterial, plant, and animal genomes (Berlin et al. 2015; Chen et al. 2014; Gordon et al. 2016; Koren et al. 2013; Loman et al. 2015). The third-generation sequencers are of particular importance in deciphering the diversity of the metagenome and identifying novel transcript isoforms and gene fusion events (Oulas et al. 2015; Sharon et al. 2013).

2.6 Bioinformatics Analysis Pipeline

Through next-generation sequencing technologies, it is now possible to describe methylated regions in the genome sequence, sequencing whole genomes, transcriptome, catalog noncoding RNA, and protein-DNA interaction sites. Each of these applications generates gigabases of sequence information which imposed an increasing demand on statistical methods and bioinformatics tools for analysis and management of enormous data produced by different sequencing platforms (Grada and Weinbrecht 2013).

The first step in sequence data analysis is to produce short nucleotide sequence also referred as reads and their associated quality scores from raw light intensity signals. This is called as base calling and the related software are usually provided by the manufacturer of the sequencing platforms (McGinn and Gut 2013). For example, CASAVA is a base calling software provided by Illumina for converting intensity files to human readable file format, e.g., FASTQ. Short reads generated are stored in the short read archive (SRA) in FASTQ format.

SRA compact design allows storage and retrieval of sequence data including metadata from experiments and reads with associated quality scores in a very effective manner. We can convert SRA to FASTQ file using SRA Toolkit.

FASTQ is a text-based format for storing biological sequences. It is basically a FASTA file associated with the quality score for each base (Mills 2014). A FASTQ file normally uses four lines per sequence.

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
Line 2 is the raw sequence letters.
Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
Line 4 encodes the quality values for the sequence in Line 2 and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

@SEQ_ID
TTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACT
CACAGTTT
+
*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

Phred Score:
A Phred score is a measure of quality of nucleotide bases identified using DNA sequencing (Table 2.2). It is commonly defined as

$$Q = -10\log 10P$$

where P is the probability that base is incorrectly called.

Quality Filtering
Before doing any downstream analyses, it is always advisable to check for read quality. Sequencing artifacts like base calling errors, poor quality reads, adapter contamination, PCR duplication, and GC biasedness are common factors that need to be checked. Filtering is the most crucial step to remove low-quality reads as during further analysis one cannot have control on quality of reads (Watson 2014). Alignment or mapping is the next step in NGS analysis. Two different ways are possible for mapping millions of reads. One is a comparative mapping of reads with the reference genome (DNA sequence of species under consideration), and another is de novo assembly. Reference genome is the DNA sequence database of an organism representing species set of genes. Mapping of reads onto reference genome provides a tentative map indicating regions from where the reads belong. A reference genome for individual organisms can be accessed from various web resources like NCBI, Ensembl, or UCSC genome browser. In the absence of reference genome, de novo genome assembly is done with the help of overlapping reads to stitch consecutive regions in the genome (Fonseca et al. 2012). A variety of tools are available for both comparative mapping and for de novo assembly. Alignment results are stored in BAM (Binary Alignment Map)/SAM (Sequence Alignment Map) file (Li et al. 2009). Best mapping hits can be filtered out using multiple parameters like mapping quality score. For every study, filtering and alignment are common steps in sequence

Table 2.2 Phred score and corresponding incorrect base call probability

Phred Quality Score	Probability of incorrect base call	Base call accuracy (in %)
10	1 in 10	90
20	1 in 100	99
30	1 in 1000	99.90
40	1 in 10,000	99.99
50	1 in 100,000	100.00

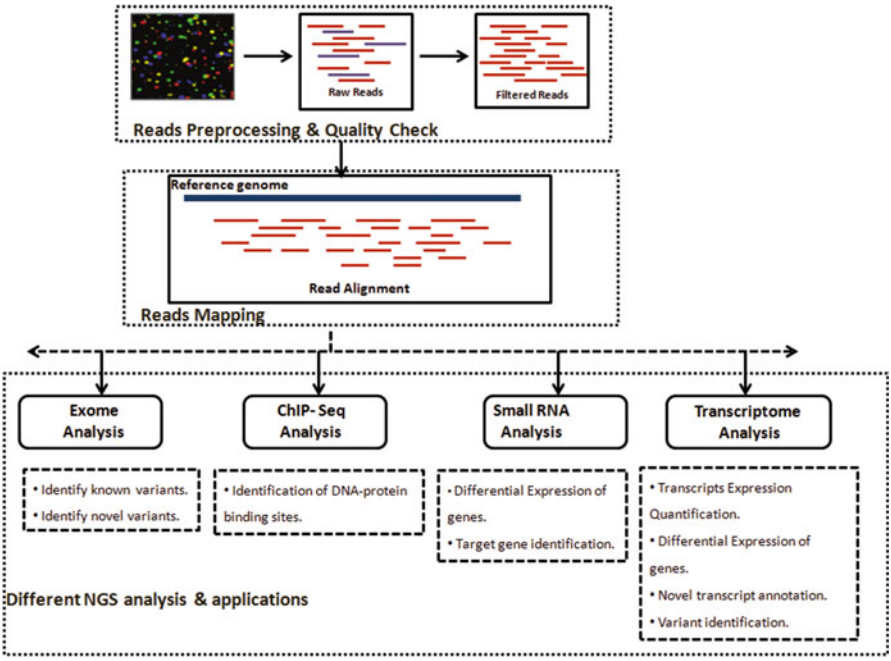

Fig. 2.7 Basic workflow in NGS data analysis and their applications

data analysis. Real fate of sequence analysis is decided only after alignment. Using different kinds of high-throughput data, various analyses can be done (Fig. 2.7).

Exome or whole genome sequencing can be used to detect structural variations that can give rise to different phenotypes in different individuals. Somatic as well as germline mutations can be identified with high precision using exome sequencing. Genes that are expressed differentially in various conditions/groups can be identified through transcriptome analysis. Different software are available for each step of data analysis. Table 2.3 describes freely available commonly used software for various purposes.

For data visualization, a graphical interface usually called as genome browser is required to display analysis results. Comparative analysis with other genomic resources (dbSNP, 1000 genome), expression changes, and peak folding is also possible on these browsers. The real strength of analysis is reflected from the power to display the results in an easy-to-interpret manner. Common IGV, UCSC, Tablet, and MapView are examples of genomic browsers.

Table 2.3 List of software commonly used in NGS data analysis

Commonly used softwares in next generation data analysis	
Reads quality control softwares	*Web URL*
fastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
filteR	http://scbb.ihbt.res.in/SCBB_dept/filter.php
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
FastX toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
DNA/RNA alignment	
BWA	http://bio-bwa.sourceforge.net/
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
Stampy	http://www.well.ox.ac.uk/stampy
TopHat	https://ccb.jhu.edu/software/tophat/index.shtml
STAR	https://github.com/alexdobin/STAR
HiSAT2	http://ccb.jhu.edu/software/hisat2/index.shtml
Denovo assembly DNA/RNA	
Vcake	https://sourceforge.net/projects/vcake/
Velvet	https://www.ebi.ac.uk/~zerbino/velvet/
Trinity	https://sourceforge.net/projects/trinityrnaseq/
Trans-Abyss	https://github.com/bcgsc/transabyss
Variant detection and annotation	
GATK	https://software.broadinstitute.org/gatk/
VarScan	http://varscan.soureeforge.net/
SnpEff	http://snpeff.sourceforge.net/
SeattleSeq	http://snp.gs.washington.edu/SeattleSeqAnnotation138/
Differential expression	
Limma	http://bioconductor.org/packages/release/bioc/html/limma.html
Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/
DeSeq2	http://bioconductor.org/packages/release/bioc/html/DESeq2.html
EBSeq	http://bioconductor.org/packages/release/bioc/html/EBSeq.html
Metagenomics	
QIIME	http://qiime.org/
RDP-Pyro	https://rdp.cme.msu.edu/
Visualization	
IGV	http://software.broadinstitute.org/software/igv/
Circos	http://www.circos.ca/
Tablet	https://omictools.com/tablet-tool
Brig	http://brig.sourceforge.net/brig-in-action/
Cytoscape	http://www.cytoscape.org/
Web resources for NGS data	
Ensembl	http://asia.ensembl.org/index.html
ExAC Browser	http://exac.broadinstitute.org/

(continued)

Table 2.3 (continued)

Commonly used softwares in next generation data analysis	
1000 Genome	http://www.internationalgenome.org/
TCGA	http://cancergenome.nih.gov/
Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/
Genome 10K	https://genome10k.soe.ucsc.edu

2.7 Applications of Next-Generation Sequencing

Earlier, hundreds of publications have been published in which next-generation sequencing is applied for a variety of applications in the field of genomics and transcriptomics. In accelerating biological and biomedical research, NGS technologies have been useful in a number of ways including whole genome sequencing, targeted sequencing, gene expression profiling, novel gene discovery, chromatin immunoprecipitation, etc. (Buermans and den Dunnen 2014).

2.7.1 Genomic Applications

NGS technology enabled a new era of genomic research through massively high-throughput sequencing data which has solved various research problems. It has provided the most comprehensive view of genomic information and associated biological implications. Using comparative genomics, one can obtain a correlation between variations and its associated clinical features. Through international efforts like UK10K and 1000 genome projects, it is now possible to find variations present in normal healthy individuals belonging to different ethnicities. Various applications of NGS in genomics can be described as:

- *Whole genome sequencing*: With the advent of NGS technologies, it is now possible to sequence genome of simple as well as complex organisms at a faster rate with much low cost. Personalized treatment plans can be offered by healthcare providers using WGS. Based on the variations present in genomic sequence with respect to controls or reference genome, one can predict probable predispositions toward disease in future. Healthcare professionals can suggest any lifestyle-related changes to avoid future complications (Shapiro et al. 2013).
- *Detection of rare variations*: Many international efforts are going on with a sole aim to catalog various kinds of variations like SNPs, mutations, indels, and copy number variations present in the genome. One such global effort is 1000 genome consortium which has cataloged more than 79 million variations in around 2500 individuals (Consortium 2015). Through exome sequencing, one can find mutations in genes which are responsible for rare Mendelian disorders such as sickle cell anemia, Miller's syndrome, as well as common diseases like obesity, diabetes, etc. Structural variations including copy number variations and indels

have been successfully identified in diseased vs. non-diseased individuals. Genome-wide variations identified using NGS techniques help in understanding why some people respond to some therapy easily while many cannot (Boycott et al. 2013).

- *Prenatal diagnosis of genetic diseases*: Sequencing technologies have been applied to detect biomarkers for genetic disorders including Down's syndrome (Palomaki et al. 2011), Edwards' syndrome, and many others. It is now possible to test for genetic abnormalities before the birth itself (Cram and Zhou 2016). Maternal cell-free plasma sequencing is done to detect various chromosomal anomalies in the fetus (Canick et al. 2013). This technique successfully detected 22q11.2 deletion syndrome, Down's syndrome, myotonic dystrophy, and various single gene disorders.
- *Transplantation*: The human leukocyte antigen (HLA) system is a gene complex encoding the major histocompatibility complex (MHC) proteins in humans. These proteins are responsible for the regulations of the immune system in humans. Differences in HLAs are the major cause of organ transplant rejection. Mapping the variations is important to identify the possible course of patient body in accepting or rejecting the transplant. Nowadays, doctors go for HLA-typing to find the suitable match for transplantation (Lan and Zhang 2015).
- *Forensics*: Genome sequencing can be used to find the suspected criminal from the proof like blood and hair obtained from the crime site. As every individual has unique DNA sequence, patterns obtained from sample can be used as proof to identify criminal. Similarly, DNA sequencing has been applied to find paternity of the child (Yang et al. 2014).
- *Population adaptation*: People are adapted to diverse environmental conditions. It is possible with NGS to catalog variations which help them to survive under extreme environments (Long et al. 2015). Classic example of one such kind of gene is EGLN1 whose variations are reported in literature which makes a person able to adapt in low oxygen conditions (Aggarwal et al. 2015).
- *Disease gene identification*: Different gene panels for disease like cancer are available which can detect presence of specific tumor in patients and help in planning a proper treatment for the same.

2.7.2 Transcriptomics Applications

All transcripts expressed by the genome in different tissues at different time points can be captured using RNA sequencing. It is now possible to map all transcribed regions in the genome with a great precision. Currently, two important publicly available databases, the Encyclopedia of DNA Elements (ENCODE) and Genotype-Tissue Expression (GTEx; The GTEx Consortium 2013), are used to map functional elements that can regulate gene expression in different human tissues. Various applications of transcriptome data sequencing are:

- *Gene expression quantification*: NGS technologies have been successfully applied to measure gene expression of thousands of genes at any given point of time. Through RNA sequencing, it is possible to measure expression levels of different transcripts as well. It is now possible to find quantitative expression in different biological conditions, in different cells as well as in different tissues. Genes which are expressed differently in diseased conditions as compared to a normal control can be identified using high-throughput expression quantification techniques (Chen et al. 2012).
- *Noncoding RNA quantification*: Noncoding RNAs (ncRNA) which include transfer RNA (tRNA), ribosomal RNA (rRNA), small nucleolar RNA, micro-RNA (miRNA), and small interfering RNA (siRNA) are not translated into proteins. However, ncRNA plays an important role in various posttranscriptional modifications. It is now possible to measure ncRNA with great precision. Long noncoding RNAs can be easily identified and are found to be associated with various neurological diseases like Alzheimer's and different cancer types (Brunner et al. 2012).
- *Transcript annotation*: RNA sequencing is capable of detecting novel transcript isoforms, promoter elements, and untranscribed regions which can be of functional importance in the genome (Trapnell et al. 2010).
- *Variant detection*: Allele-specific expression detection is very useful to find causal variations in various case control studies. It is now possible to detect tissue-specific transcript variants in different samples accurately.
- *Fusion detection*: A fusion transcript is a chimeric RNA containing exons from two or more different genes and has the potential to code for novel proteins. Through different RNA sequencing experiments, fusion transcripts have been found to be associated with different cancer types including breast and prostate cancer (Bao et al. 2014).

2.7.3 Epigenetics

The study of heritable gene regulation that does not involve DNA sequence itself is called epigenetics. Two major kinds of epigenetic modifications are DNA methylation and histone tail modifications. Epigenetic modifications are of prime importance in oncogenesis and development. These changes decide whether the genes will be turned on or off and ensure proper production of proteins in specific cells only (Holliday 2006).

- *DNA methylation*: Methylomics is the study of genome-wide DNA methylation patterns and their effect on gene regulation. In methylation, when methyl groups are added to a particular gene, that gene is turned off, and no protein is produced from it. Bisulfite sequencing is done to determine methylation patterns of DNA. Bisulfite treatment of DNA converts cytosine to uracil but leaves

5-methylcytosine residues unaffected. Therefore, only methylated residues will be retained. The Human Epigenome Project is an initiative to identify, catalog, and interpret genome-wide methylation patterns of all human genes in all major tissues (Eckhardt et al. 2004). Different methylation clusters are found to be present in cancer patients as compared to control (Soto et al. 2016).

- *Histone tail modification*: Histones are the proteins which package and order the DNA into structural units called nucleosomes. Chromatin immunoprecipitation sequencing (Chip-Seq) is used to analyze histone modifications which determine the accessibility of DNA to transcriptional regulators. It is widely used in gene regulatory networks to find transcription factors and any other protein interactions with DNA on a genome-wide scale. Transcription factors controlling the progression of disease in an individual have been identified through Chip-Seq; for example, GABP is a transcription factor and is a promoter for TERT gene and is found to be associated with multiple cancer types (Messier et al. 2016).

2.7.4 Metagenomics

Metagenomics is the branch of genomics which involves genetic analysis of microbial genomes contained within an environmental sample. NGS-based metagenome analysis has revolutionized our understanding of ecology around us. To reveal the importance of microorganisms that surrounds us, various international efforts such as the Human Microbiome Project (HMP) (Turnbaugh et al. 2007) and Human Gut Microbiome Project have been initiated worldwide. The main goal of all these efforts is to find out the association of changes in the human microbiome with human health and diseases. Various studies have shown the applications of metagenomics to microbial ecology and industrial biotechnology.

- *Human health*: Diet and nutrition intake are the most important identifiers of human health and both govern human microbiome too. Gut microbiome plays a major role in metabolic, nutritional, physiological, and immunological processes in the human body. Studies have shown that perturbations in intestinal microbiome have been associated with various diseases including obesity (Flint et al. 2014), inflammatory bowel syndrome (Kostic et al. 2014), and celiac disease (David Al Dulaimi 2015).
- *Bioremediation*: Biosurfactants are low molecular weight surface-active compounds mainly produced by bacteria, yeast, and fungi. They are used in agriculture for plant pathogen elimination and for increasing the bioavailability of nutrients for beneficial plant microbes. As metagenomics is culture-independent technique, it is used these days to find novel compounds associated with natural ecosystems (Edwards and Kjellerup 2013).
- *Ecology*: Microorganisms play an integral part in history and function of life on earth. Studies in metagenomics have provided valuable insights into the

functional ecology of the microbial community. For example, bacterial communities found in defecations of sea lions in Australia suggest that nutrient-rich feces of these lions are an important nutrient source for coastal ecosystems (Lavery et al. 2012).

- *Biofuel*: Due to diminishing fossil fuel reserves and increased CO_2 accumulation in the atmosphere, biofuels have been viewed as an alternative for sustainability and protecting the environment. Biofuels are fuels derived from biomass conversion like in the conversion of cellulose into cellulosic ethanol (Morrison et al. 2009). Lignocellulose represents the largest terrestrial carbon source on earth, but cannot be broken down without a combination of acids, industrial chemicals, and heat. Various fungi and bacteria have been identified that can enzymatically decompose lignocellulose to its monomeric compounds for use as carbon sources. Various metagenomic studies have identified the key genes and enzymes involved in lignocellulose digestion and conversion into biofuels (Chandel and Singh 2011; Hess et al. 2011; Xing et al. 2012).

NGS technologies are now considered a routine part in multi-omics research. Reduction in cost of sequencing per base facilitated sequencing technologies at different genomic centers and private companies. Low-cost and high-throughput methods are providing physicians with the tools to translate genomic knowledge into clinical practice. Due to current NGS technologies, major advances are possible in many areas especially in understanding and diagnosis of complex and rare diseases. As our understanding of genome variability increases, functional annotation of the genome will also rise. However, no advancements will prove fruitful without developing efficient algorithms which can transform sequence reads and data into meaningful information. There is a need for innovative bioinformatics methods for analysis and infrastructure to store available wealth of data. A year-by-year rise in the number of publications related to the field of NGS is a proof of its wide applicability and advancements. In the coming years, novel sequencing solutions are expected from additional sequencing providers.

References

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ et al (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28:E87

Aggarwal S, Gheware A, Agrawal A, Ghosh S, Prasher B, Mukerji M (2015) Combined genetic effects of EGLN1 and VWF modulate thrombotic outcome in hypoxia revealed by Ayurgenomics approach. J Transl Med 13:184

Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res 9:3015–3027

Ansorge W, Sproat BS, Stegemann J, Schwager C (1986) A non-radioactive automated method for DNA sequence determination. J Biochem Biophys Methods 13:315–323

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res 15:4593–4602

Astier Y, Braha O, Bayley H (2006) Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. J Am Chem Soc 128:1705–1710

Bao ZS, Chen HM, Yang MY, Zhang CB, Yu K et al (2014) RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res 24:1765–1773

Bayley H (2010) Nanotechnology: holes with an edge. Nature 467:164–165

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol 33:623–630

Bowers J, Mitchell J, Beer E, Buzby PR, Causey M et al (2009) Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods 6:593–595

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14:681–691

Brunner AL, Beck AH, Edris B, Sweeney RT, Zhu SX et al (2012) Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers. Genome Biol 13:R75

Buermans HP, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. Biochim Biophys Acta 1842:1932–1941

Canick JA, Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE (2013) The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. Prenat Diagn 33:667–674

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M et al (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611

Chandel AK, Singh OV (2011) Weedy lignocellulosic feedstock and microbial metabolic engineering: advancing the generation of 'biofuel'. Appl Microbiol Biotechnol 89:1289–1303

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY et al (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148:1293–1307

Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM et al (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. Cell 158:1187–1198

Consortium GP (2015) A global reference for human genetic variation. Nature 526:68–74

Cram DS, Zhou D (2016) Next generation sequencing: coping with rare genetic diseases in China. Intract Rare Dis Res 5:140–144

David Al Dulaimi M (2015) The role of infectious mediators and gut microbiome in the pathogenesis of celiac disease. Arch Iran Med 18:244

Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M et al (2010) Nanopore DNA sequencing with MspA. Proc Natl Acad Sci U S A 107:16060–16065

Eckhardt F, Beck S, Gut IG, Berlin K (2004) Future potential of the human epigenome project. Expert Rev Mol Diagn 4:609–618

Edwards SJ, Kjellerup BV (2013) Applications of biofilms in bioremediation and biotransformation of persistent organic pollutants, pharmaceuticals/personal care products, and heavy metals. Appl Microbiol Biotechnol 97:9909–9921

Eid J, Fehr A, Gray J, Luong K, Lyle J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Flint HJ, Duncan SH, Louis P (2014) Gut microbiome and obesity. In: Treatment of the obese patient. Springer, New York, pp 73–82

Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28:3169–3177

Foquet M, Samiee KT, Kong X, Chauduri BP, Lundquist PM et al (2008) Improved fabrication of zero-mode waveguides for single-molecule detection. J Appl Phys 103:034301

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T et al (2009) The challenges of sequencing by synthesis. Nat Biotechnol 27:1013–1023

Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 13:840–852
Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351
Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN et al (2016) Long-read sequence assembly of the gorilla genome. Science 352:aae0344
Grada A, Weinbrecht K (2013) Next-generation sequencing: methodology and application. J Invest Dermatol 133:e11
Gut IG (2013) New sequencing technologies. Clini Transl Oncol Off Publ Fed Span Oncol Soc Nat Cancer Inst Mex 15:879–881
Haque F, Li J, Wu HC, Liang XJ, Guo P (2013) Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. Nano Today 8:56–74
Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8
Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol 36:39–56
Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H et al (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331:463–467
Holliday R (2006) Epigenetics: a historical overview. Epigenetics 1:76–80
Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A 93:13770–13773
Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155:27–38
Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM et al (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14:R101
Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT et al (2008a) Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. Nucleosides Nucleotides Nucleic Acids 27:1072–1083
Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL et al (2008b) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A 105:1176–1181
Kostic AD, Xavier RJ, Gevers D (2014) The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology 146:1489–1499
Lan JH, Zhang Q (2015) Clinical applications of next-generation sequencing in histocompatibility and transplantation. Curr Opin Organ Transplant 20:461–467
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
Lavery TJ, Roudnew B, Seymour J, Mitchell JG, Jeffries T (2012) High nutrient transport and cycling potential revealed in the microbial metagenome of Australian sea lion (Neophoca cinerea) faeces. PLoS One 7:e36478
Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S et al (2016) Third-generation sequencing and the future of genomics. bioRxiv:048603
Lelieveld SH, Veltman JA, Gilissen C (2016) Novel bioinformatic developments for exome sequencing. Hum Genet 135:603–614
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
Liu H, He J, Tang J, Liu H, Pang P et al (2010) Translocation of single-stranded DNA through single-walled carbon nanotubes. Science 327:64–67
Liu L, Li Y, Li S, Hu N, He Y et al (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364
Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733–735

Long A, Liti G, Luptak A, Tenaillon O (2015) Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat Rev Genet 16:567–582
Mardis ER (2006) Anticipating the 1,000 dollar genome. Genome Biol 7:112
Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402
Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470:198–203
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380
Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560–564
McGinn S, Gut IG (2013) DNA sequencing – spanning the generations. New Biotechnol 30:366–372
McNally B, Singer A, Yu Z, Sun Y, Weng Z, Meller A (2010) Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. Nano Lett 10:2237–2244
Messier TL, Gordon JA, Boyd JR, Tye CE, Browne G et al (2016) Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. Oncotarget 7:5094–5109
Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11:31–46
Mills L (2014) Common file formats. Current Protocols in Bioinformatics 45:A 1B 1–A 1B18
Morey M, Fernandez-Marmiesse A, Castineiras D, Fraga JM, Couce ML, Cocho JA (2013) A glimpse into past, present, and future DNA sequencing. Mol Genet Metab 110:3–24
Morrison M, Pope PB, Denman SE, McSweeney CS (2009) Plant biomass degradation by gut microbiomes: more of the same or something new? Curr Opin Biotechnol 20:358–363
Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE (2011) Landscape of next-generation sequencing technologies. Anal Chem 83:4327–4341
Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N et al (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinf Biol Insights 9:75–88
Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM et al (2011) DNA sequencing of maternal plasma to detect down syndrome: an international clinical validation study. Genet Med 13:913–920
Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52:413–435
Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O et al (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12:780–786
Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–111
Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. J Hum Genet 59:5–15
Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26:1117–1124
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467
Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240
Schneider GF, Kowalczyk SW, Calado VE, Pandraud G, Zandbergen HW et al (2010) DNA translocation through graphene nanopores. Nano Lett 10:3163–3167
Schubeler D (2015) Function and information content of DNA methylation. Nature 517:321–326
Service RF (2006) Gene sequencing. The race for the $1000 genome. Science 311:1544–1546
Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14:618–630

Sharon D, Tilgner H, Grubert F, Snyder M (2013) A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 31:1009–1014

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679

Soto J, Rodriguez-Antolin C, Vallespin E, de Castro CJ, Ibanez de Caceres I (2016) The impact of next-generation sequencing on the DNA methylation-based translational cancer research. Transl Res J Lab Clin Med 169(1–18):e11

The GTEx Consortium (2013) The genotype-tissue expression (GTEx) project. Nat Genet 45:580–585

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36–46

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI (2007) The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449:804

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S et al (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18:1051–1063

Van Verk MC, Hickman R, Pieterse CM, Van Wees SC (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci 18:175–179

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al (2001) The sequence of the human genome. Science 291:1304–1351

Venter JC, Smith HO, Adams MD (2015) The sequence of the human genome. Clin Chem 61:1207–1208

Wang Y, Navin NE (2015) Advances and applications of single-cell sequencing technologies. Mol Cell 58:598–609

Watson M (2014) Quality assessment and control of high-throughput sequencing data. Front Genet 5:235

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–738

Whiteford N, Skelly T, Curtis C, Ritchie ME, Lohr A et al (2009) Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics 25:2194–2199

Xing M-N, Zhang X-Z, Huang H (2012) Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. Biotechnol Adv 30:920–929

Yang Y, Xie B, Yan J (2014) Application of next-generation sequencing technology in forensic science. Genomics Proteomics Bioinformatics 12:190–197

Zhao Q, Wang Y, Dong J, Zhao L, Rui X, Yu D (2012) Nanopore-based DNA analysis via graphene electrodes. J Nanomater 2012:4