We are IntechOpen, the world’s leading publisher of Open Access books. Built by scientists, for scientists.

6,600 Open access books available
177,000 International authors and editors
195M Downloads

154 Countries delivered to
Top 1% of most cited scientists
12.2% Contributors from top 500 universities

Our authors are among the top 1% of most cited scientists. Contributors from top 500 universities.

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us? Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected. For more information visit www.intechopen.com.
High-Throughput Sequencing and Metagenomic Data Analysis

Ahmed Shuikan, Sulaiman Ali Alharbi, Dalal Hussien M. Alkhalifah and Wael N. Hozzein

Abstract

Metagenomic approaches are a growing branch of science and have many applications in different fields. Metagenomics seems to be the ideal culture-independent technique for unraveling the biodiversity of soils and to study how this biodiversity is affected with continuously changing conditions. In addition, its application in clinical and diagnostic approaches was reported. The emergence of several next-generation sequencing (NGS) strategies enriched the metagenomics. The combination between NGS and metagenomic approaches helped the investigators resolve several issues regarding the microbial diversity and the functions and relationships among different microbial flora. A number of NGS approaches were developed including Roche/454 pyrosequencing, Illumina/Solexa sequencing, and Applied Biosystems/SOLiD sequencing. In this chapter, different NGS platforms are discussed in terms of principle, advantages, and limitations. In addition, third-generation sequencing technologies are also addressed.

Keywords: high throughput, metagenomics workflow, sequencing approaches, metagenomic data analysis

1. Introduction

The development of next-generation sequencing (NGS) techniques provides high-throughput sequence analysis with the ability to simultaneously and independently sequence billions of DNA molecules. The combination between such technologies and metagenomic approaches helped the investigators study the microbial diversity and understand the functions and relationships among different microbial flora [1]. The use of metagenomic NGS by microbiologists overcomes several limitations and secured the unbiased methods to study the microbial flora in any given environment [2]. Thus, the dynamic of complex communities particularly those with non-cultivable microorganisms can be resolved [3, 4]. In addition, metagenomic NGS found its way in the field of clinical and diagnostic approaches [5, 6]. In the clinical field, NGS was used to inform the real-time incidence and prevention response to human parainfluenza 3 virus infections [7] and for cerebrospinal fluid diagnostics [8].

Several NGS platforms were developed since 2006 with numerous applications in genetic and biological research fields. Of these platforms, the most commonly used include Roche/454 pyrosequencing, Illumina/Solexa sequencing, and Applied Biosystems/SOLiD sequencing. The principle of all these NGS depends on the detection of luminescent signals released by the base incorporation during the
sequencing process [4]. They also share the same workflow which in the order, DNA extraction, library construction, DNA template preparation, and automated sequence analysis [9]. In this chapter, different NGS platforms are discussed in terms of principle, advantages, and limitations. In addition, third-generation sequencing technologies are also addressed.

2. Workflow of metagenomics

2.1 The sampling process and library construction for metagenomic analysis

Metagenomic analysis is a sophisticated process and involves several steps. Of these steps, the sampling process is very crucial for the downstream applications. Sample collection, preparation, and storage should be handled carefully to prevent lysis and decomposition of the sample compositions. Multiple freezing–thawing cycles may cause changes in the microbial community profile under investigation [10]. As well, a suitable DNA extraction protocol should be adopted to cope with the different chemical and physical characteristics of each sample. For instance, soils contain many substances that are co-extracted with the genomic DNA and may have inhibitory effects on the downstream experiments. Examples include humic and fulvic acids [11]. Therefore, optimization and comparison between different extraction methods are usually required for each type of samples [12, 13, 14, 15].

The extracted DNA is used to construct the DNA library. This is usually achieved by connecting specific adaptors to one or both ends of the DNA fragments [16]. The reason for utilizing DNA adaptor is to deal with the pool of samples and then connect them to its original sample. Handling DNA at this stage should be careful to avoid chemical, physical, or enzymatic damage of DNA molecules [17]. The construction of a DNA library is usually achieved through two approaches. The first one is called meta-pair where the library is characterized by long fragment insert. The second approach is called paired-end libraries with short fragment insert. In both approaches, the DNA is fragmented into different fragment sizes that would allow for their cloning. The DNA fragments obtained from such processes are cloned into the proper cloning vector. The size of the resulting fragments determines the suitable vector for the cloning process. The small DNA fragments are usually cloned into plasmid vectors, whereas fragments up to 40 kbp are cloned into cosmid or fosmid vectors. Bacterial artificial chromosome (BAC) vectors are usually used to clone inserts with sizes that exceed 40 Kbp [18]. Finally, the free adaptor, dimers of the adaptor, and any other artifacts must be removed to avoid noisy sequencing data [17].

2.2 Sequencing approaches

During the 1970s, the first-generation sequencing techniques, chain termination [19], and chemical sequencing approaches [20] were developed. In contrast to the chemical sequencing approach, the Sanger sequencing method ultimately prevailed and found immense applications due to its simplicity and is more amenable to being scaled up [21]. Simply, the basis of Sanger sequencing depends on the incubation of a specific primer and the template DNA in the presence of DNA polymerase. The reaction is accomplished by the addition of a mixture of deoxyribonucleotide triphosphates and dNTPs’ dideoxyribonucleotide triphosphates for chain termination, one of which was labeled with phosphorus-32. The resulting pool of DNA amplicons will be with the same 5‘ residue and different dNTP residues at the 3‘ end (Figure 1).

This pool of DNA fragments is then fractionated by denaturing polyacrylamide gel electrophoresis giving a band pattern. In this way, DNA decoding can be achieved
High-Throughput Sequencing and Metagenomic Data Analysis
DOI: http://dx.doi.org/10.5772/intechopen.89944

Currently, the use of fluorescent dNTPs associated with the capillary electrophoresis provides full automation of the Sanger approach. This modification allows retrieving up to 96 sequences per run with an average 800–1000 bp size of DNA fragments [21, 23, 24]. Although the Sanger sequencing was the mainstay of the original human genome project, this approach still has some limitations. These limitations include high cost and low throughput, and it is inadequate for studying unculturable organisms in complex environments [25].

2.2.1 Next-generation sequencing (NGS)

Due to the limitations of Sanger sequencing technique, next-generation sequencing emerged in 2005 [26]. Indeed, next-generation sequencing has made it possible to study and identify organisms directly from their habitats without prior preparations [27]. Compared to the first-generation sequencing, NGS can generate several hundred thousand to millions of sequencing reads in parallel. As well, sequencing can be generated without some conventional steps such as vector-based cloning procedure and hence reduces the chance of DNA contamination from other organisms [28]. Therefore, several next-generation sequencing platforms have been introduced including Roche 454, Illumina®, Applied Biosystems SOLiD sequencer, and Ion Torrent. All next-generation sequencing or real-time sequencing (Roche 454, Illumina®, and AB SOLiD) utilized optical sensors that detect luminescent signal, which are produced during incorporation of bases in the sequence. The principles and characteristics of NSG, SGS, and TGS are summarized in Table 1 [21]. In the subsequent sections, the features and limitations of each of the NGS techniques are discussed.

2.2.1.1 Roche 454 genome sequence

Roche/454 pyrosequencing is the first NGS technology that launched and became commercially available in 2005. It uses real-time sequencing-by-synthesis
**Metagenomics - Basics, Methods and Applications**

(SBS) pyrosequencing technology, and it depends on the detection of pyrophosphate (PPi) molecule that is initiated from the incorporation of a nucleotide in the DNA polymerase (Figure 2) [29]. Briefly, the 454 pyrosequencing technology is proceeding as follows: (i) the library fragments are connected to beads that carry oligonucleotides complementary to adapter sequence ligated at the ends, (ii) amplifying the library fragments by emulsion PCR resulting in DNA beads that carry millions of copies of DNA fragments on their surface, and (iii) the amplified beads are inserted into picotiter plate (PTP) that consists of millions of wells. Each well can hold only one amplified bead and contains diluted pyrosequence enzyme beads, DNA amplified beads, PPiase beads, and pyrosequence beads. Finally, the light emission from PTP is recorded by a CCD camera and is translated to nucleotide sequences [29]. In comparison with other NGS platform, 454 pyrosequencing has

<table>
<thead>
<tr>
<th></th>
<th>First generation</th>
<th>Second generation</th>
<th>Third generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fundamental technology</strong></td>
<td>Size separation of specifically end-labeled DNA fragments, produced by SBS or degradation</td>
<td>Wash-and-scan SBS</td>
<td>SBS, by degradation, or direct physical inspection of the DNA molecule</td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>Averaged across many copies of the DNA molecule being sequenced</td>
<td>Averaged across many copies of the DNA molecule being sequenced</td>
<td>Single-molecule resolution</td>
</tr>
<tr>
<td><strong>Current raw accuracy</strong></td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
</tr>
<tr>
<td><strong>Current read length</strong></td>
<td>Moderate (800–1000 bp)</td>
<td>Short, generally much shorter than Sanger sequencing</td>
<td>Long, 1000 bp, and longer in commercial systems</td>
</tr>
<tr>
<td><strong>Current throughput</strong></td>
<td>Low</td>
<td>High</td>
<td>Moderate</td>
</tr>
<tr>
<td><strong>Current cost</strong></td>
<td>High cost per base</td>
<td>Low cost per base</td>
<td>Low-to-moderate cost per base</td>
</tr>
<tr>
<td><strong>RNA sequencing method</strong></td>
<td>cDNA sequencing</td>
<td>cDNA sequencing</td>
<td>Direct RNA sequencing and cDNA sequencing</td>
</tr>
<tr>
<td><strong>Time from start of sequencing reaction to result</strong></td>
<td>Hours</td>
<td>Days</td>
<td>Hours</td>
</tr>
<tr>
<td><strong>Sample preparation</strong></td>
<td>Moderately complex, PCR amplification not required</td>
<td>Complex, PCR amplification required</td>
<td>Ranges from complex to very simple depending on technology</td>
</tr>
<tr>
<td><strong>Data analysis</strong></td>
<td>Routine</td>
<td>Complex because of large data volumes and because short reads complicate assembly and alignment algorithms</td>
<td>Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges</td>
</tr>
<tr>
<td><strong>Primary results</strong></td>
<td>Base calls with quality values</td>
<td>Base calls with quality values</td>
<td>Base calls with quality values, potentially other base information such as kinetics</td>
</tr>
</tbody>
</table>

**Table 1.**
The features and principles of first-generation sequencing, SGS, and TGS.
High-Throughput Sequencing and Metagenomic Data Analysis
DOI: http://dx.doi.org/10.5772/intechopen.89944

the longest reading (up to 1000–1200 bp). On the other hand, 454 pyrosequencing has the highest cost per base and the lowest output [30].

2.2.1.2 Illumina sequencing (Solexa genome analyzer)

Illumina, formerly known as Solexa, has been introduced commercially in 2007. Illumina technology utilizes bridge PCR amplification coupled with SBS in the flow cell (Figure 3). Simply, the principle of Illumina sequencing is that the DNA fragments with barcoding primer (adaptor) are attached to the flow cell. The sequencing reaction is performed in the flow cell by adding labeled nucleotides. When the nucleotide is incorporated, a luminescent signal is generated and then recorded by optical sensors. After that, the fluorescent molecules are removed and the next labeled nucleotide incorporated. However, the DNA fragment can be sequenced on one side that is called single-end (SE) or from both sides known as paired-end (PE). Nowadays, the most common sequencing used is PE due to the ability to generate two reads for one DNA fragment which is useful in order to determine the distance between two ends of the DNA fragment [31]. In fact, due to its low cost per base and high yield, Illumina becomes the most widely used and popular NGS platform. The output of Illumina sequencing is the highest among all NGS, making it suitable for multiplexing hundreds of samples at the same time [32].

2.2.1.3 Applied biosystems (AB) SOLiD sequencer

AB SOLiD refers to sequencing by oligonucleotide ligation and detection. It has been developed by Applied Biosystems (Life Technology) and became commercially available in 2007. The AB SOLiD sequencing approach differs from the other
two next-generation sequencing technologies, Illumina, and 454 pyrosequencing. AB SOLiD platform relies on sequencing-by-oligo-ligation (SBL) (Figure 4), whereas others rely on sequencing-by-synthesis (SBS) [33]. In SOLiD sequencer,
the DNA library is prepared from the sample, and specific adaptor is then amplified by emPCR [34]. Instead of utilizing DNA polymerase, short nucleotides marked by DNA ligase known as interrogation probes are used. The interrogation probe contains six universal bases and two-base encoded probe. The universal bases are attached to the fluorescent label. When an integrated probe is ligated with primers using DNA ligase, fluorescent light is generated and detected. After the 5’ end that is linked to the fluorescent label by cleavable linkage is cleaved and removed, thereby the next interrogation probe is connected. This process is repeated several times until the targeted DNA is completely sequenced. In fact, the read length of SOLiD is short about 85 bp leading to inaccurate read assembly as it requires more time for sequencing but it has the highest accuracy among other NGS [35]. Application of SOLiD includes whole genome sequencing, targeted sequencing, transcriptome, and epigenome [35].

2.2.1.4 Ion torrent sequencing

Ion Torrent has been launched in 2010 by Life Technology. Some authors have classified the Ion Torrent platform as a technique between the next-generation and the third-generation sequencing. This could be attributed to the dependence of this approach on optical sensors. However, it relies on chemical sensors that detect the hydrogen-ion concentration change that occurred during the incorporation of a nucleotide in the sequence [21]. Ion Torrent sequencing quality is high and stable due to the utilizing of a chemical sensor instead of fluorescence and camera. In addition, the Ion Torrent approach is characterized by its high speed and low cost compared with pyrosequencing and Illumina [35].

2.2.2 Third-generation sequencing

The major limitations of NGS are that the short-read length and the PCR bias are introduced by clonal amplification and the fluorescent-based signaling detection [21]. Therefore, the third-generation sequencing or single-molecule-sequencing technologies (SMS) overcome these limitations by dispensing PCR before sequencing, and the signal is captured in real time by monitoring the enzymatic reaction [36, 21]. The following sections discuss some TGS platforms.

2.2.2.1 Helicos biosciences (HeliScope)

The first single-molecule-sequencing (SMS) that has been introduced in 2008 is HeliScope. It is a fluorescent-based, single-molecule-sequencing platform. In HeliScope platform, the preparation step depends on preparing a single-strand DNA, and there is no need for PCR amplification in the preparation step. During sequencing, repetitive cycles of DNA polymerase and one labeled nucleotide are flowed, resulting in DNA template extension which depends on the flow of nucleotides. The labeled nucleotides are modified by attaching a poly-A tail in order to stop polymerase extension until the fluorescence that generates from the incorporated nucleotide is recorded by a CCD camera. Then unincorporated nucleotides are washed out and the fluorescent labels on the strand chemically removed, allowing for next base incorporation [37, 38]. HeliScope Genetic Analysis System platform allows the sequencing of RNA, and there is no need for converting them to cDNA. Furthermore, HeliScope Genetic Analysis System platform is in its infancy due to small read length (24–70 bases) and low data output (20 GB) [39].
2.2.2.2 PacBio technology/SMRT sequencer

Pacific Bioscience has launched a single-molecule real-time (SMRT) technology in 2010. It is a real-time, fluorescent-based, and single-molecule-sequencing platform. In SMRT, there is no need for PCR amplification during DNA preparation [36]. In this platform, a nanostructure known as zero-mode waveguide (ZMW) is utilized for real-time observation of DNA synthesis. During the sequencing process, a single-stranded template is used to synthesize the complementary. Unlike other NGS platforms, four different colored fluorescent labels are attached to the terminal phosphate group instead of attaching to a nucleotide, resulting in the release of a fluorescent signal during nucleotide incorporation [40]. Then the camera captures the fluorescent signal in real time (like a movie) [41]. In SMRT, the washing step between nucleotide flows is not required, resulting in increasing the nucleotide incorporation and improving the quality of sequencing [42]. SMRT has several advantages including fast sample preparation (hours instead of days like NGS), no need for PCR amplification during the preparation step, and longer-read length than any other next-generation sequencing platform [42].

2.2.2.3 Oxford Nanopore technology

Nanopore sequencing, developed by Oxford Nanopore Technology, relies on passing the DNA sequence through a 1 nm diameter hole (nanopore) where electric current is applied. The electrical current of the pore is altered for each nucleotide, and signal is detected in real time [39]. Like other third-generation sequencing approaches, this technology does not require PCR amplification or chemical labeling of the sample [43]. In May 2015, Oxford Nanopore Technologies has introduced commercially the MinION. The MinION is a pocket-size portable, real-time detection of bases (fluorescent tag-free), has long-read length, and is a low-cost technology [44, 41, 45]. Interestingly, by utilizing this technology, samples can be sequenced in the field directly, instead of collecting samples and sequencing them in the lab, which means nanopore sequencing will make all other sequencing machines redundant [46, 44].

2.3 Metagenomic data analysis

Several bioinformatic tools were developed to analyze the metagenomic data at the molecular level (e.g., 16S rRNA), species level, and strain level. 16S rRNA sequence strategy is among the most common approaches to understand microbial taxonomy and phylogeny. This could be attributed to the stable functions of 16S rRNA gene over time, the existence of 16S rRNA in nearly all microorganisms, and its size which is enough for bioinformatics analysis [47, 48]. A number of bioinformatics tools are available for the analysis of 16S rRNA: QIIME, MOTHUR, DADA2, UPARSE, and minimum entropy decomposition (MED) [49]. The QIIME software is designed to analyze data generated on the Illumina or other NGS platforms via graphics and statistics. This involves the demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations [50, 51]. QIIME depends on the use of the PyCogent toolkit to identify misinterpretations and database deposition using raw sequencing results [51]. Operational taxonomic units (OTUs) can be generated from NGS data by UPARSE [52]. The UPARSE software acts by filtering and trimming reads into equals lengths, removing singleton reads and clustering the remaining reads [52].
Community sequence data can be analyzed by a flexible and comprehensive software package called MOTHUR. The MOTHUR package includes the following algorithms: DOTUR, SONS, TreeClimber, LIBSHUFF, D-LIBSHUFF, and UniFrac [50]. DADA2 is a suitable approach for correcting amplicon errors with no option to generate OTUs [53]. DADA2 uses a new quality-aware model of Illumina amplicon errors to improve the DADA algorithm [53]. MED is applied to solve the limitations of fine-scale resolution descriptions of microbial communities [54]. MED acts through partitioning the data set of amplicon sequences into homogenous OTUs for alpha- and beta-diversity analyses [54].

For species-level metagenomic data analysis, there are at least six metagenomic analysis software including MetaPhlAn2 [55], Kraken [56], CLARK [57], FOCUS [58], SUPERFOCUS [59], and MG-RAST [60]. All of these software programs can be used to profile organisms in metagenomic samples and to score their abundance. MetaPhlAn2 applies Bowtie2 and UCLUST [52, 61] as its main algorithms, whereas k-mers (DNA words of length k) is the core algorithm for Kraken and CLARK. On the other hand, FOCUS uses the NNLS (nonnegative least squares) to identify the microbial profile [49].

3. Conclusion

At the beginning, the metagenomic workflow was complicated that it requires many steps, sophisticated equipment, and qualified technicians to perform. Likewise, it was very expensive that not all scientists or laboratories were able to afford its cost. However, nowadays, due to the presence of many different competing companies and laboratories that led to the development of more efficient sequencing approaches, the metagenomic workflow became easier. It is easy now to study and identify organisms directly from their habitats without prior preparations. In terms of cost, NGS is also much cheaper, and with the appearance of third-generation sequencing approaches, it is not required to conduct sample sequencing. Surprisingly, sequencing can be carried out in the field by utilizing a pocket-size portable sequencer. The advancements in the field of metagenomics are amazing, and it became easier, cheaper, and faster.

Acknowledgements

This work was funded by the Deanship of Scientific Research at Princess Nourah Bint Abdulrahman University, through the Research Groups Program Grant no. RGP-1438-0004.

Conflict of interest

The authors declare that there are no conflicts of interest.
Author details

Ahmed Shuikan\textsuperscript{1}, Sulaiman Ali Alharbi\textsuperscript{4}, Dalal Hussien M. Alkhalifah\textsuperscript{2} and Wael N. Hozzein\textsuperscript{3,4}\textsuperscript{*}

1 Botany and Microbiology Department, College of Science, King Saud University, Riyadh, Saudi Arabia

2 Biology Department, Faculty of Science, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

3 Bioproducts Research Chair, Zoology Department, College of Science, King Saud University, Riyadh, Saudi Arabia

4 Botany and Microbiology Department, Faculty of Science, Beni-Suef University, Beni-Suef, Egypt

*Address all correspondence to: hozzein29@yahoo.com; whozzein@ksu.edu.sa

IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
References


[13] Lim NY, Roco CA, Frostegård Å. Transparent DNA/RNA co-extraction workflow protocol suitable for inhibitor-rich environmental samples that focuses on complete DNA removal for transcriptomic analyses. Frontiers in Microbiology. 2016;7:1588


bacterial DNA extraction methods from non-polluted and polluted soils. Folia Microbiologica. 2018;63(1):85-92


The sequence of the human genome. Science. 2001;291:1304-1351


and libraries for bead-based sequencing. US20080003571; 2008


community supported software for describing and comparing microbial communities. Applied and Environmental Microbiology. 2009;75:7537-7541


