Chapter 9
Experimental Approaches for Genome Sequencing

Mohd Sayeed Akhtar, Ibrahim A. Alaraidh, and Khalid Rehman Hakeem

Contents

9.1 Introduction

Genome sequencing is developing rapidly as a revolutionizing field due to advances in DNA sequencing technologies and started the new era in the field of molecular biology (Kelley and Salzberg 2010; Ruperao et al. 2014; Khan et al. 2016; Visendi et al. 2016). The scientist working in this arena has gained the popularity by manipulating the DNA molecules for the study of genes and their harness toward the development sparking a new revolution in biological investigations (Fuller et al. 2006; Hsu et al. 2014). These recent advances in genome sequencing served as an important tool in basic and translational research, drug development, and clinical

M. S. Akhtar (✉)
Department of Botany, Gandhi Faiz-e-Aam College, Shahjahanpur, Uttar Pradesh, India

I. A. Alaraidh
Botany and Microbiology Department, King Saud University, Science College, Riyadh, Saudi Arabia

K. R. Hakeem
Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

Princess Dr. Najla Bint Saud Al-Saud Center for Excellence Research in Biotechnology, King Abdulaziz University, Jeddah, Saudi Arabia

K. R. Hakeem et al. (eds.), *Essentials of Bioinformatics, Volume III*,
https://doi.org/10.1007/978-3-030-19318-8_9

trials (Fontana et al. 2012; Readhead and Dudley 2013; Roychowdhury and Chinnaiyan 2016). Nowadays, the sequencing cost and a high-throughput data generation are not limiting factors due to development of modern sophisticated technologies, but a big core facility is still needed to operate the procedures of genome sequencing (Buermans and den Dunnen 2014; Head et al. 2014). Since, the genome analysis is increasingly used to address various problems related to the genotyping, diagnosis, environmental and microbiome profiling, and mutation and evolutionary studies. The number of challenges in genome analysis is associated with sequencing methods. There has been very fast development in genome sequencing. A good example was the completion of human genome project in record time. The human body has about 100 trillion cells. Inside each cell is the nucleus that contains the genome (46 human chromosomes), which governs human development (Kothekar and Nandi 2007). Similarly, the tomato genome identified an esterase responsible for differences in volatile ester content in different tomato species (Goulet et al. 2012). In general, the chromosomes comprise millions of copies of the four-letter genetic code—the DNA bases (A, C, G, and T) which are arranged into genes and noncoding sections (Akhtar et al. 2017). Finding the order or sequence of these four letters is the goal in genomics. The entire human genome is made up of about 3.5 billion bases. To read the DNA sequence, the chromosomes are cut into tiny pieces and read individually. When all the segments have been read, they assembled in the correct order. The properties of a biological system are studied through the expression of many genes simultaneously. Simple interpretation strategies are useful. A typical example is protein p53, an early component in cells which respond to DNA damage (Goeman et al. 2017). Thus, the aim of this chapter is to provide the in-depth knowledge of various experimental approaches used for sequencing of the genome.

9.2 Genome Sequencing Approaches

There are fundamentally two ways to sequence the genome, namely BAC (bacterial artificial chromosome)-to-BAC approach and shotgun approach.

9.2.1 BAC-to-BAC Approach

BAC-to-BAC approach is also referred to as the map-based approach. It was first employed in human genome studies during the late 1980s and continues its expansion till date. The BAC-to-BAC approach first creates a crude physical map of the whole genome (Becker 1998; Bolger et al. 2014). Constructing a map requires cutting the chromosomes into pieces and figuring out the order of the big chunks of DNA before sequencing all the fragments. The BAC sequences were individually assembled and arranged according to the physical map, creating a very high-quality genome sequence (Fig. 9.1a).

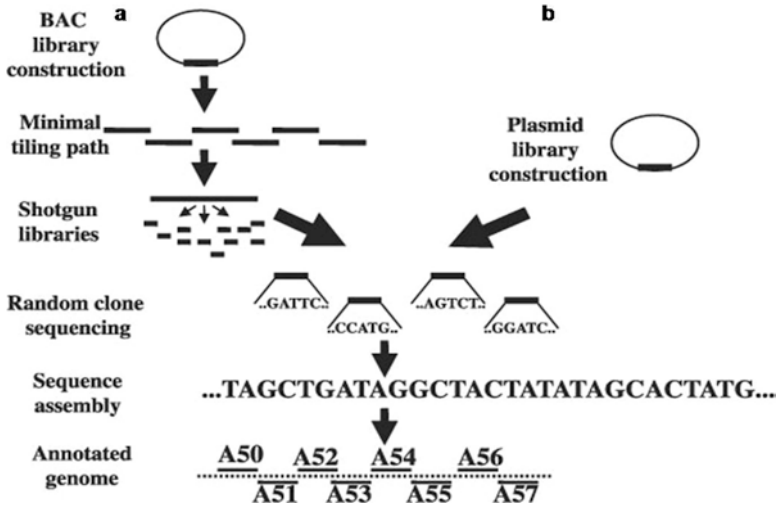

Fig. 9.1 Whole genome sequencing; (**a**) BAC-to-BAC approach; (**b**) Shotgun approach

Several copies of the genome are randomly cut into pieces of about 150,000 base pairs (bp), and each of these fragments is inserted into a BAC. A BAC is an artificial piece of DNA that can replicate inside the bacterial cell. The whole collection of BACs containing the entire human genome is called a BAC library. Each BAC is like a book in the library that can be accessed and copied. Each piece is fingerprinted to give a unique identification tag that determines the order of the fragments. Fingerprinting involves cutting each BAC fragment with a single enzyme and finding common sequence landmarks in overlapping fragments that determine the location of each BAC along the chromosome. Each BAC is then broken randomly into pieces of 1500 bp and is placed in another artificial piece of DNA called M13. The collection is known as M13 library. All the M13 libraries are sequenced. A 500 bp from one end of the fragment is sequenced to generate millions of sequences. Compute algorithms assemble millions of sequenced fragments into a continuous stretch of the chromosome. These sequences are fed into a computer program called PHRAP, which looks for common sequences that join two fragments.

9.2.2 Shotgun Approach

It is a speedy approach to genome sequencing, which may enable the researchers to complete their job in a short time. Venter (1996) developed the shotgun approach at The Institute for Genomic Research (TIGR). The approach of sequencing bypasses the need for a physical map and goes straight into the job of decoding (Fig. 9.1b). This is the main reason for this speedy technique. Multiple copies of the genome are randomly shredded into pieces of 2000 bp by squeezing the DNA through a pressurized syringe. This is done a second time to generate pieces of 10,000 bp long.

Each 2000 and 10,000 bp fragment is inserted into a plasmid, which is a piece of DNA that replicates in bacteria. The two collections of plasmids containing 2000 and 10,000 bp chunks of human DNA are known as plasmid libraries. Both 2000–10,000 bp plasmid libraries are sequenced. A 500 bp from each end of a fragment is decoded to generate millions of sequences. Sequencing both ends is critical for assembling the entire chromosome; computer algorithms assemble millions of sequenced fragments into a continuous stretch of the chromosome.

9.2.3 Other Sequencing Approaches

9.2.3.1 Large-Scale Approach

It includes hybridization and sequencing approaches. Hybridization has evolved from early membrane-based radioactive detection embodiments to parallel quantitative methods using fluorescence detection (Lee 2007).

9.2.3.2 cDNA Microarray Detectors

It is a very sensitive technique. It requires only 2–5 nl of DNA solution coating with poly L-lysine and aminocialines. The cDNA libraries provide a flexible sequence probe. The choice of fluoroprobe is important. The biological samples (or their cDNA derivatives) are hybridized to the range and are referred to as the target. Labeling with fluorescent dyes with different excitation and emission characteristics allows the simultaneous hybridization of two contrasting targets on a single array (Aharoni and Vorst 2001; Campos-De Quiroz 2002). Microarrays can be based on cDNA molecules, and their basic features are represented in tabular form (Table 9.1).

9.2.3.3 PCR Method

PCR is used to amplify the single copy or copies of a DNA segments across several orders of magnitude to generate millions of copies of a desired DNA sequences. It is an easy, cheap, and reliable technique to replicate a focused segment of DNA, and most widely used in molecular biology for biomedical research, criminal forensics, and molecular archaeology. Now, it is commonly used in clinical and research

Table 9.1 Some common features of cDNA microarray

Features	cDNA microarray
Array preparation	Direct or indirect spotting
Target	cDNA
Target labeling	Cy3-dCTP and Cy5-dCTP incorporation through reverse transcription
Type of hybridization	DNA-DNA

laboratories for a broad variety of applications including DNA cloning and manipulation, gene mutagenesis, construction of DNA-based phylogenies, or functional analysis of genes; diagnosis and monitoring of hereditary diseases; amplification of ancient DNA; analysis of genetic fingerprints; and detection of pathogens in nucleic acid tests for the diagnosis of infectious diseases. The majority of PCR methods rely on thermal cycling, which involves exposing the reactants to cycles of repeated heating and cooling, permitting different temperature-dependent reactions specifically DNA melting and enzyme-driven DNA replication to quickly proceed many times in sequence. Primers contain sequences complementary to the target region, along with a DNA polymerase to enable selected and repeated amplifications. In PCR, the choice of template is important. As the PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the original DNA template is exponentially amplified. Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase. If heat-susceptible DNA polymerase is used, it will denature every cycle at the denaturation step. Before the use of Taq polymerase, DNA polymerase had to be manually added every cycle, which was a tedious and costly process. This DNA polymerase enzymatically assembles a new DNA strand from free nucleotides by using single-stranded DNA as a template and DNA oligonucleotides to initiate DNA synthesis. In the first step, the two strands of the DNA double helix are physically separated at a high temperature in a process called DNA melting. In the second step, the temperature is lowered and the two DNA strands become templates for DNA polymerase to selectively amplify the target DNA. Selectivity of PCR results from the use of primers that are complementary to sequence around the DNA region targeted for amplification under specific thermal cycling conditions. PCR has an enormous impact in both basic and diagnostic aspects of molecular science because it produces large amounts of a specific DNA fragments from small amounts of a complex template. PCR represents a form of in vitro cloning that can generate or modify the DNA fragments of definite length and sequence in a simple automated reaction. In addition, PCR plays a critical role in the identification of medically important sequences as well as an important diagnostic one in their detection.

9.2.3.4 Image Analysis

Intensity evaluation (100–400 μ^2 pixels) allows 50–200 sampling at each spot. Data normalization is done. Statistical analysis is important. Assay reliability has to be tested.

9.2.3.5 Functional Proteomics

Phene is the functional protein contained in the total protein data. Phene is to "phenotype" as a gene is to "genotype." In multicellular organisms, the set of proteins would differ from cell type to cell type. Proteins normally undergo large-scale

modifications. Proteome analysis is concerned with biochemical changes like post-translational modifications, phosphorylation, etc. In general, the phosphorylation is a reversible enzymatic reaction and plays an important role in various cellular processes, viz., division, function of target proteins, immunity metabolism, membrane transport, and organelle trafficking (Bolger et al. 2014). It can activate and inhibit enzyme activity through allosteric conformational changes, facilitate the recognition of other proteins, promote protein-protein association or dissociation, and also induce order to disorder transition.

9.3 Conclusions and Future Prospects

Genome sequencing has a major impact on molecular biological research and improvement in the comparatively small period of time. Although a rapid development has been observed in the preparation of library in the past decades with the performances of some small genome studies, still the breakthrough researches are expected. Therefore, in the future, more studies are needed on the deep phenotyping platforms to overcome the issues and the elucidation of mechanisms to complete enormity of the available data.

Acknowledgments The authors (Mohd Sayeed Akhtar and Ibrahim A. Alaraidh) are highly grateful to the Department of Botany, Gandhi Faiz-e-Aam College, Shahajahanpur, U.P., India, and the Botany and Microbiology Department, Science College, King Saud University, Riyadh, Kingdom of Saudi Arabia.

References

Aharoni A, Vorst O (2001) DNA microarrays for functional plant genomics. Plant Mol Biol 48:99–118

Akhtar MS, Swamy MK, Alaraidh IA, Panwar J (2017) Genomic data resources and data mining. In: Hakeem KR, Malik A, Sukan FV, Ozturk M (eds) Plant bioinformatics. Springer International Publishing, Cham, pp 267–278

Becker S (1998) Three-dimensional structure of the Stat3β homodimer bound to DNA. Nature 394:145–151

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KFX (2014) Plant genome sequencing applications for crop improvement. Curr Opin Biotechnol 26:31–37

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

Campos-De Quiroz H (2002) Plant genomics: an overview. Biol Res 35:385–399

Fontana JM, Alexander E, Salvatore M (2012) Translational research in infectious disease: current paradigms and challenges ahead. Transl Res 159:430–453

Fuller DN, Gemmen GJ, Rickgauer JP, Dupont A, Millin R, Recouvreux P, Smith DE (2006) A general method for manipulating DNA sequences from any organism with optical tweezers. Nucleic Acids Res 34:e15. https://doi.org/10.1093/nar/gnj016

Goeman F, Strano S, Blandino G (2017) MicroRNAs as key effectors in the p53 network. Int Rev Cell Mol Biol 333:51–90

Goulet C, Mageroy MH, Lam NB, Floystad A, Tieman DM, Klee HJ (2012) Role of an esterase in flavor volatile variation within the tomato clade. Proc Natl Acad Sci U S A 109:19009–19014

Head SR, Komori HK, LaMere SA, Whisenant T, Nieuwerburgh FV, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: overviews and challenges. Biotechniques 56:61. https://doi.org/10.2144/000114133

Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell 157:1262–1278

Kelley DR, Salzberg SL (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol 11:R28

Khan S, Wajid Ullah M, Siddique R, Nabi G, Manan S, Yousaf M, Hou H (2016) Role of recombinant DNA technology to improve life. Int J Genomics 2016:2405954. https://doi.org/10.1155/2016/2405954

Kothekar V, Nandi T (2007) An introduction to bioinformatics. Duckworth Press, BioScience Publishers, Delhi

Lee MLT (2007) Analysis of microarray gene expression data. Kluwer Academic Publisher, Boston

Readhead B, Dudley J (2013) Translational bioinformatics approaches to drug development. Adv Wound Care 2:470–489. https://doi.org/10.1089/wound.2012.0422

Roychowdhury S, Chinnaiyan AM (2016) Translating cancer genomes and transcriptomes for precision oncology. CA Cancer J Clin 66:75–88

Ruperao P, Chan C-KK, Azam S, Karafiátová M, Hayashi S, Čížková J, Saxena RK, Šimková H, Song C, Vrána J, Chitikineni A, Visendi P, Gaur PM, Millán T, Singh KB, Tara B, Wang J, Batley J, Doležel J, Varshney RK, Edwards D (2014) A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. Plant Biotechnol J 12:778–786

Venter JC, Smith HO, Leed H (1996) A new strategy for genome sequencing. Nature 381:364–366

Visendi P, Berkman PJ, Hayashi S, Golicz AA, Bayer PE, Ruperao P, Hurgobin B, Montenegro J, Chan CKK, Staňková H, Batley J, Šimková H, Doležel J, Edwards D (2016) An efficient approach to BAC based assembly of complex genomes. Plant Methods 12:2. https://doi.org/10.1186/s13007-016-0107-9