September 2014 Vol. 21 No. 3 84-96
Journal of Northeast Agricultural University (English Edition)

High-throughput Sequencing Technology and Its Application

Zhu Qiang-long, Liu Shi, Gao Peng, and Luan Fei-shi*

College of Horticulture, Northeast Agricultural University, Harbin 150030, China

Abstract: Gene sequencing is a great way to interpret life, and high-throughput sequencing technology is a revolutionary technological innovation in gene sequencing researches. This technology is characterized by low cost and high-throughput data. Currently, high-throughput sequencing technology has been widely applied in multi-level researches on genomics, transcriptomics and epigenomics. And it has fundamentally changed the way we approach problems in basic and translational researches and created many new possibilities. This paper presented a general description of high-throughput sequencing technology and a comprehensive review of its application with plain, concisely and precisely. In order to help researchers finish their work faster and better, promote science amateurs and understand it easier and better.

Key words: high-throughput sequencing, data analysis, genome sequence, transcriptome sequence, bioinformatics

CLC number: S6 **Document code:** A **Article ID:** 1006-8104(2014)-03-0084-13

Introduction

With the indepth study of life sciences and further development of bio-technology, more and more scientists recognize that the whole genome sequencing of a species will be the fundamental basis and important clue to help them reveal the nature of life of the species. The discovery of DNA double helix (Watson and Crick, 1953), cracking genetic code (Nirenberg *et al*., 1966), and the successful completion of the first one complete genome map (Sanger *et al*., 1977) have undoubtedly become a series of important journey milestones in the history of life scientific development, and make more scientists profoundly recognize that sequencing technology plays an important role in life science researches. The rapid sequencing technology would make DNA sequencing become one of the most important methods of molecular analysis (Sanger *et al*., 1992). This technology provides important data for basic biology study, such as disclosure of genetic information and regulation of gene expressions. With the appearence of Roche's 454 technology (2005), Illumina's Solexa technology (2006) and ABI's SOLiD technology (2007), high-throughput sequencing technology has got enormous developments, thus amounts of genetic information is successively revealed, which allow us to explore the essence of life in detail, to uncover the huge diversity of novel genes that are currently inaccessible, to understand nucleic acid therapeutics, to better integrate biological information for a complete picture of health and disease at a personalized level and to move to advance that we can not yet imagine (Kahvejian *et al*., 2008). Therefore, a number of bioinformatics methods and softwares have

Received 29 October 2013
Supported by the National Natural Science Foundations of China (31272186; 31301791)
Zhu Qiang-long (1989-), male, Master, engaged in the research of bioinformatics and watermelon molecular breeding. E-mail: longzhu2011@126.com
* Corresponding author. Luan Fei-shi, professor, supervisor of Ph. D students, engaged in the research of watermelon and melon molecular breeding. E-mail: luanfeishi@sina.com

been created to accelerate high-throughput sequencing technology to be widely applied in aspects of genomics researches on genomics, transcriptomics and epigenetics. High-throughput sequencing technology has fundamentally changed the way we approached problems in basic and translational researches and created many new possibilities. Whereas, it has also brought new challenges for bioinformatics: how to effectively process and analyze these massive data and extract valuable bio-information form it, which have become an important key to decide if high-throughput sequencing technology plays a major role in the scientific exploration. In this article, we intended to present a comprehensive and systematic introduction of high-throughput sequencing technology and its applications to the enthusiast of biological science with plain, concisely and precisely hope to help researchers finish their work faster and better, to promote science amateurs understand it easier and better. Meanwhile, we tried to take data generated from Illumina Hiseq 2000 sequencing platform as an example to present a more complete description of the basic procedure, key methods and existing software of the sequencing data generating process, data processing and analysis.

History of High-throughput Sequencing Technology Development

High-throughput sequencing technology is the second generation sequencing technology launched by Roche/454 Company, Illumina/Solexa Company and ABI/SOLiD Company based on Sanger sequencing and single-molecule sequencing technologies announced by Helicos Heliscope™ and Pacific Biosciences, which is also called as deep sequencing technology (Sultan *et al*., 2008) or the next-generation sequencing technology (NGS) (Schuster, 2008) .

The 1st generation sequencing technology

In 1977, Sanger of Cambridge and Gilbert of Harvard almost simultaneously published their different methods of DNA sequencing in the same magazine (Maxam and Gilbert, 1977; Sanger *et al*., 1977), their inventions first opened a door to study the genetic code of life deeply for researchers, and brought hope to the development of faster and more efficient sequencing technology. Sanger method belongs to dideoxy chain termination method, while Gilbert method is chemical degradation method. The former is more convenient and more suitable for optical automatic detection gradually replaced the latter, and became the most widely applied method of sequencing in the field of life science. Thus, Sanger won the 1980 Nobel Prize in chemistry (Sanger, 1988). Most of the automated DNA sequencers are based on this method. Its principle is as below, when a nucleic acid template is replicating under the presence of DNA polymerase, a pair of primers, four types of single deoxynucleotide triphosphate (dNTP, one of them labeled with a radioactive 32P), join four kinds of dideoxynucleotide triphosphate (ddNTP) into four reactive systems in proportion, because dideoxynucleotide have no 3'-OH, so long as the dideoxynucleotide append to the end of the chain, its extension is stopped, if the single deoxynucleotide triphosphate append to the end of the chain, it can continue to be extended. So that a series of the nucleic acid fragments with the dideoxy nucleotide at the 3' end in different length ranges will be synthesized in each reaction system. After termination of the reaction, different lengths of nucleic acid fragments should be isolated by gel electrophoresis in four lanes, where there is a difference of one nucleotide among near segments. After autoradiography, the order of base in synthetic fragment can be read, according to the dideoxy nucleosides at the 3' end of the fragment (Xie *et al*., 2010). Subsequently, a variety of DNA sequencing technologies based on this technology has been exploited, the most important one of them is fluorescent automated sequencing technology (Fig. 1) (http://en.wikipedia.org/wiki/File:Sanger-sequencing.svg). This generation sequencing technology has played a key role in human genome project, accelerating the completion of human genome project. The sequencer