Chapter 13
Genome Sequencing in the Cloud

Wei Chen
The University of Texas MD Anderson Cancer Center, USA

Bo Peng
The University of Texas MD Anderson Cancer Center, USA

Yun Wan
School of Arts & Science, University of Houston - Victoria, USA

Christopher I Amos
Geisel School of Medicine, Dartmouth College, USA

ABSTRACT

This chapter discussed the latest development of using cloud-computing technology for genome sequencing in bioinformatics field. It introduced the definition of genome sequencing and cloud computing, discussed the current status of NGS in cloud with the example of Nimbix. It also provided a rich source of cloud computing related service providers and technologies for references. Finally, it discussed the challenges of conducting NGS in a cloud environment.

INTRODUCTION

Bioinformatics is an inter-disciplinary field that has grown rapidly over the last two decades along with breakthroughs in both biotechnologies and information technologies. It relates to the study of methods about how to store, retrieve and analyze biological data, such as nucleic acid (DNA/RNA) and protein sequences, structures, functions, pathways and genetic interactions (http://en.wikipedia.org/wiki/Bioinformatics). The high-throughput data generated especially from next generation sequencing (NGS) is a challenge for its management, extraction and analysis. The Human Genome Project began in 1990 and aimed at mapping and identifying the approximately 20,000-25,000 genes in human genome, and determining the sequences of the 3 billion base pairs that comprise human DNA. With the successful completion of the Human Genome Project in 2003, new needs emerged for rapidly and efficiently identifying single nucleotide polymorphisms (SNPs), copy number variations (CNVs), structural variations (SVs) and insertions and deletions (indels). Thus, the focus has shifted to understanding the roles or functions of genetic variations in the context of gene-gene and gene-environment interactions and cell signaling pathways, especially on genes

DOI: 10.4018/978-1-4666-8210-8.ch013

for complex diseases and the development of personalized therapies. Only 1.1-1.4% of the genome is spanned by exons (coding sequence for amino acids) and the remaining sequence comprises introns (non-coding regions), intergenic regions or other sequences such as micro RNA that control expression of genes. Less than 1% of all SNPs lead to variation in proteins possibly causing damage, so the task of determining which SNPs have functional consequences remains an open challenge (Venter et al., 2001).

The increasingly data-intensive nature of life sciences, especially driven by the Human Genome Project and recently next generation sequencing data, demands high-speed data transfer, big data storage, optimized processing and efficient analyses in Petabyte and Exabyte scale. This leaves two options for most research organizations to enhance their computing infrastructure and support such initiatives: either build one's own cluster computing infrastructure, or use a cloud-based approach, that is a private cloud, or publically available cloud computing services like Amazon EC2.

The challenge for both options is there are few off-the-shelf tools to help integrate, analyze and visualize such massive amounts of data, which forces most bioinformatics researchers to search for solutions by themselves. In this chapter, we first introduce genome sequencing and cloud computing technology, then explain the current techniques of genome sequencing in the cloud with a list of available cloud computing service and software for genome sequencing. Then we discuss the challenges of this new trend.

GENOME SEQUENCING

Since the 1970s, the genome sequencing technology has experienced three stage of development. They were typically classified as zero generation sequencing used in the 1970s, first generation sequencing used in the 1980s to 1990s, and second generation sequencing adopted since the 2000s.

The first generation genome sequencing method is called Sanger sequencing and used cycle sequencing to generate a ladder of increased dye-labeled products, which are subjected to high-resolution electrophoretic separation. The four-channel spectrum is used to trace the sequence to fluorescently labeled fragments (Sanger & Coulson, 1978). Second generation genome sequencing is called clonally PCR amplified molecule sequencing (Shendure & Ji, 2008). The sequencing steps include fragmentation, library generation, amplification, sequencing and analysis. The raw sequence reads of whole genome sequencing may be produced in 5 days, followed by 3 weeks of mapping the reads to genome and generating raw genomic features (e.g. SNPs) and finally months of data analysis to uncover the biological meaning. The raw uncompressed sequencing data obtained by one Illumina HiSeq run can be 1.5 Terabyte, or 1400+ CDs of data over 1.7 m tall. Sequencing platforms in this category include Illumina Genome Analyzer/HiSeq (Illumina Inc.), ABI SOLiD (Life Technologies Corp.), and Roche 454 pyrosequencing. The third generation sequencing or next-next generation sequencing is called true single molecule sequencing (Harris et al., 2008). This category includes much more sensitive yet expensive sequencing by synthesis without amplification using true Single Molecule Sequencing (tSMS) or Heliscore technology HeliScopeSingle Molecule Sequencer (Helicos Biosciences Corp.), Ion Torrent technology uses a semiconductor (Life Technologies Corp.), SMRT technology (Pacific Biosciences Inc.), DNA nanoball technology (Complete Genomics Inc.), or Nanopore technology (Oxford Nanopore Technologies Inc.). The third generation sequencing produces longer reads but currently has higher error rates than earlier generations of sequencing approaches. The sequence assembly is easier because of the longer reads.