OPTIMIZING VARIANT CALLING PERFORMANCE FOR HOTSPOT CASES BASED ON ION TORRENT SEQUENCING TECHNOLOGY

Basma Nasser Abd Elsalam Abd Elfattah

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Biomedical Engineering & Systems

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT

2018

OPTIMIZING VARIANT CALLING PERFORMANCE FOR HOTSPOT CASES BASED ON ION TORRENT SEQUENCING TECHNOLOGY

by

Basma Nasser Abd-Elsalam Abd-Elfattah

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Biomedical Engineering & Systems

Under the Supervision of

Prof.Dr. Ayman Eldeib
.....
Professor of Biomedical Engineering
Systems & Biomedical Engineering
Faculty of Engineering, Cairo University

Dr. Mohamed I. Abouelhoda
.....
Associate Professor
Systems & Biomedical Engineering
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

Acknowledgments

I would like to thank Prof. Ayman Eldeib for providing me with the vision and giving me time to hear me. Great thanks to Dr. Mohamed I. Abouelhoda, for his guidance and patience. Without his persistent help, this thesis would not have been possible.

Thanks for all my colleagues and friends who help and support me. Special thanks for my family for their support over the last couple of years.

Dedication

Table of Contents

List of Tables

List of Figures

Abbreviations

BAM	Binary Sequence Alignment/Map.
BRCA	BReast CAncer genes.
DNA	Deoxyribonucleic Acid.
ENCODE	Encyclopedia of DNA Elements.
GATK	Genome Analysis Toolkit.
GIAB	Genome in a bottle
Hapmap	Haplotype Map.
HC	HaplotypeCaller
InDel	Insertions and DELetion.
mTVC	Modified torrent variant caller script.
NGS	Next Generation Sequencing.
PCR	Polymerase chain reaction
PE	Paired-end.
PGM	Personal genome machine.
SAM	Sequence Alignment/Map.
SNP	Single nucleotide polymorphism.
TMAP	Torrent Mapping Alignment Program.
TVC	Torrent Variant Caller.
VCF	Variant call Format.
VQSR	Variant Quality Score Recalibration

Abstract

In last few years, Next generation sequencing (NGS) revolutionized in DNA sequencing research. The latest major technologies released are Illumina and Ion Torrent. In this study, we analyze the performance of calling variants using both platforms. In order to compare between both platforms, we used two sequenced data sets from ion community, which contain flow spaces required by Ion Torrent. We are concerned with the execution time of both platforms. Ion torrent detects genome variants faster than Illumina but with low accuracy. Moreover, we found that Ion Torrent called slightly more variants but with higher false positive rate. The hotspot option provided by ion torrent variant caller provides more accurate variant calling positions as the filtration restricted to specific positions in the genome, but this leads to time consumption. Here, we enhanced the execution time to attain the accurate positive variants using torrent variant caller by using the NOCALL variants as hotspot regions to torrent variant caller (TVC). We developed two dependent packages the first one to create the new hotspot file (mHotspot) required by TVC and the second one to rerun the TVC with mHotspot file generated to get the exact positions of variants.

Chapter 1

Introduction

Sequencing technologies are evolving rapidly; in 2011, a lot of sequencing technologies were developed. The advances of sequencing technologies make it possible to detect enormous number of potential disease-causing variants; also, next generation sequencing (NGS) data has been used for diagnostic purposes. This is due to the improvements made to the bioinformatics tools used to analyze the massive data produced by NGS instruments.

NGS technologies use same workflow for searching for mutations in a patient. Align the raw data to human reference genome, and identify single nucleotide variants and INDELs that cause the phenotype of interest. We have to decide the best tools to use for each stage of the pipeline. There are various tools for sample sequencing, but the most important parts are to select the proper aligner and variant caller to achieve the optimal results for SNPs and INDEL variants.

Different pipelines with different combinations between aligners and variant callers identify Gene variants (SNPs and INDELs). The most widely used aligners are BWA. GATK and torrent variant caller (TVC) are popular tools for variant calling [1].

In this thesis, we describe the pipelines and typical tools for read alignment and variant callers. In addition, we will elaborate in details the results of using ion torrent workflow for detecting variants from two different samples (NA12878 and BRCA2). In addition, we will describe new workflow for TVC in order to enhance time of execution using our plugin mTVC.

1.1 Problem Definition

Several studies have conducted comparison between different variant callers. From these studies, quail et al. BMC Genomics 2012 [2] mentioned that ion torrent has poor performance with only 65% of genome is covered with high quality reads compared to other platforms. It was mentioned that the overall rate of SNP Calling was slightly higher in ion torrent than Illumina, $\tilde{8}2\%$ of SNPs being called correctly. But far less false positives ion torrent from Illumina for SNPs and InDels. They noted that the inbuilt automatic variant calling in ion torrent calling only 1.4% of variants and to detect small InDels using proton platform we need to develop more accurate variant calling technique [4]. From torrent variant caller [5], using hotspot and target file, TVC evaluates the variants in specific locations and regions in order to achieve high accuracy and increase the true

positive variants. However, using hotspot file takes days to process and generate variants report. Although, TVC takes minutes to generate the variants without using hotspot file.

1.2 Thesis Objective

The main purpose of the thesis is to optimize the processing time of Torrent Variant Caller (TVC) to call variants when using hotspot file. We achieved this by using the NOCALL variants and reuse it as hotspot positions to TVC. Additionally, we present solution to automate the generation of new hotspot file and enhance the run time of torrent variant caller to detect variants.

1.3 Thesis Organization

In chapter 2, we will introduce briefly the biological terms related to the problem and DNA sequencing generations and technologies. As well as, we defined an overview of NGS. A detailed review of variant calling pipelines presented in chapter 3. Chapter 4 review parallel computing implementation types, advantages, and the contribution also elaborate the steps and tools used to achieve. The results are discussed in chapter 5. Then finally, chapter 6 concludes and discusses our future work. Appendix A and B summarize the common steps to run variant calling pipelines. Appendix C shows the steps to run the developed package.

Chapter 2

Biological Background

In this chapter, we introduce the basic concepts of biology and bioinformatics that are needed in the thesis. We will show the main topics of bioinformatics such as sequencing technologies, biological sequence databases and sequences format.

2.1 Genetic Background

In 1857, Gregor Mendel performed an experiment with plants that increased the interest in the study of genetics. After that, Friedrich Miescher in 1869 discovered a substance he called "nuclein" or “nucleic acid” that exists only in the chromosome later he isolated sample of material known now as “DNA”. James Watson and Francis Crick began to examine the DNA’s structure guided by X-ray diffraction photos of DNA fibers taken by Maurice Wilkins. They created double helix model with little bars called nucleotides connecting the two strands.
Our human body is build of billions of cells. Each cell varies according to the organ function. DNA (deoxyribonucleic acid) considered as the blueprint for our life, which exists in the nucleus of the cell. It is present in all living organisms from the smallest bacterium to the largest whale. DNA determines the genetic characteristics and the behavior of organism including the diseases that may develop.
The DNA structure includes four main nucleotide bases, Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), attached to sugar and phosphate string. Each DNA bases bound to each other using hydrogen bonds to form units called base pairs, A with T and C with G. Each base attached to a sugar and a phosphate molecule as shown in figure 2.1. Human DNA contains around 3 billion bases. These bases sequenced according to the organ and information that needs to be transmitted.
The nucleus contains different number of chromosomes for each organism. The set of chromosomes make up a genome. The human genome arranged into 46 chromosomes. The sequence of pieces of DNA called genes that act as instructions to make molecules called proteins. According to Human Genome Project estimation, humans have between 20,000 and 25,000 genes. Every human has pairs of each gene, inherited from each parent.

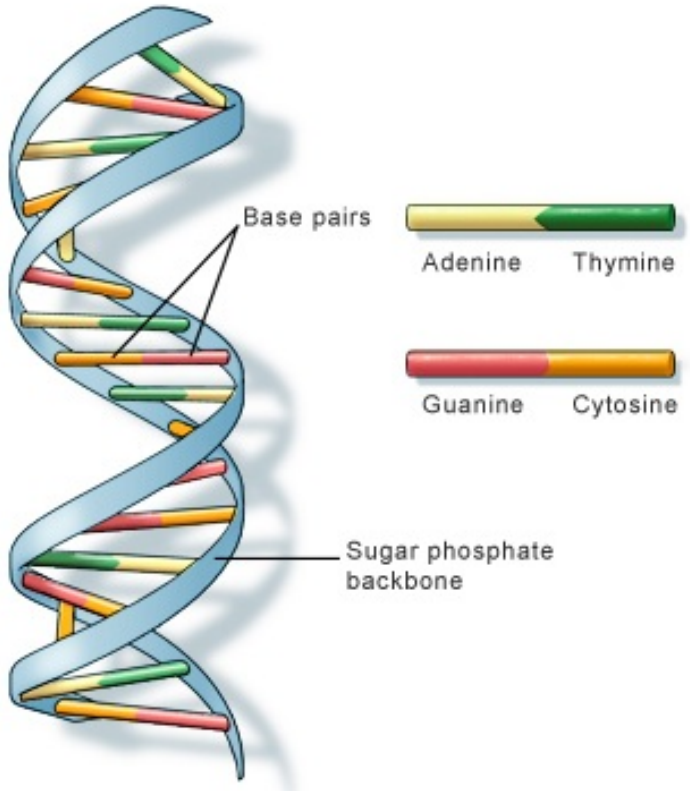

Figure 2.1: DNA Helix strand.

2.2 DNA sequencing

DNA sequencing is a laboratory method to determine the order of the nucleotides 4 bases (Adenine (A),thymine (T), cytosine (C), guanine (G)) in a piece of DNA strand. we cannot sequence a genome from start to end. We have to break the DNA strand into many smaller fragments, then sequencing the pieces and label each base individually with different colors. The detailed flow described in the next chapter. Previously, to sequence one or two genes we may spend several years. DNA sequencing technologies have been developed rapidly since the completion of Human Genome Project. We can sequence the entire genome in few hours and with less money.
DNA sequencing can tell the scientists the genetic information in particular DNA segment.

In addition, sequencing can be used in molecular biology to allow researchers to identify the changes in gene that may cause disease and identify potential drug targets. DNA sequencing can be used to study how different organisms related to each other and how they evolved. Moreover, we can determine the risk of genetic diseases.