CHAPTER I

INTRODUCTION

1.1. Background

China's authority reported patients associated with pneumonia derived from unknown etiology in Wuhan back in December 2019. Soon, it was identified as a new type of coronavirus and was successfully isolated and fully sequenced in 10 January 2020 named as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). It enters the body through receptors called Angiotensin-Converting Enzyme-2 (ACE-2) widely expressed in human organs including lower respiratory tract organs such as lungs (Astuti & Ysrafil, 2020). Following entry, the human body will trigger protective responses and eventually cause acute respiratory failure with more serious complications. This disease is termed as Coronavirus Disease 2019 (COVID-19). As of February 2021, the World Health Organization (WHO) reported 111 million confirmed COVID-19 cases contributing to 2.4 million fatalities worldwide (WHO, 2021).

Researchers rally the study as a newly emerging COVID-19 led to a high number of fatalities everyday since its early identification. Examination of SARS-CoV-2 morphological structure showed the spherical shape of its particle with diameters ranging from 60 - 140 nm and spikes length of 8 - 12 nm (Zheng, 2020). Through morphological observation, researchers saw its consistency with the viruses under *Coronaviridae* family and hence, it was classified as the member of the *Coronaviridae* family. Genetic materials study through structural observation and viral genomic sequencing indicated a large single-stranded RNA genome posing the length between 26 - 32 kilobases composed of 13 to 15 Open Reading Frames (ORF) consisting of around 30,000 nucleotides each, while the genome contains 11 protein-coding genes and 38% GC content (Naqvi et al, 2020; Astuti & Ysrafil, 2020). Understanding of morphological and viral genome characteristics of SARS-CoV-2 are providing valuable insights to turn the table off the worsening pandemic situation. However,

transmission and anti-viral treatments might induce mutations and consequently, generate more virulent strains with higher fatalities or resistance to available treatment and vaccines (Koyama, Platt & Parida, 2020). One study has conducted data sciences analysis towards SARS-CoV-2 genome submissions between February and May 2020 and revealed that several variants exist with D614G, where adenine substitution to guanine happens at position 23,403. It is the most common variant discovered since December 2019 (Koyama, Platt & Parida, 2020). SARS-CoV-2 variant identification is pivotal in providing insight on viral infectivity, severity, and also to study the evolutionary analysis of SARS-CoV-2

COVID-19 pandemic has brought computational biology with Next Generation Sequencing (NGS) to the frontline as it revolutionized the biological sciences in the past decades with its high throughput and tremendous ability to study biological systems through a wide variety of applications. NGS enables researchers to conduct Whole Genome Sequencing (WGS), the construction of a complete DNA sequence belonging to an organism's genome at a single time. Application of WGS is capable of understanding the transmission pattern, gain insight on outbreak control decisions, and discover new variants of viruses (Oude Munnink et al, 2020). This was proven when WGS was capable in helping public health decision making strategy during the 2014-2016 West African Ebola outbreak. Therefore, WGS studies during the ongoing COVID-19 pandemic is an active area of research. The first complete genome of SARS-CoV-2 was fully recovered back in 10 January 2020 through de-novo assembly using metagenomic RNA sequencing (Wu et al, 2020). Afterwards, 431.757 whole genome sequences of SARS-CoV-2 are submitted to Global Initiative on Sharing Avian Influenza Data (GISAID) data sharing with Indonesia reported 460 complete genomes of SARS-CoV-2 as of January 2021.

NGS technologies are heavily influenced by Illumina® as the prominent player of second generation NGS. All Illumina's NGS platforms were built based on bridge

amplification with ease support and applicable to genomic sequencing, exome sequencing, targeted sequencing, metagenomics and RNA sequencing (Slatko, Gardner & Ausubel, 2018). Responding to COVID-19 pandemic, Illumina published a guideline as the improvement for target enrichment workflow in detecting respiratory viruses using the NGS platform. The workflows are highly sensitive and able to characterize common respiratory viruses including coronavirus strains without the need to map raw NGS data to the human genome (Illumina, 2020). Target enrichment has been widely used long before the COVID-19 pandemic, it utilizes hybrid-capture methods to capture genomic regions of interest using biotinylated oligonucleotide probes designed to hybridize regions of interest (Mamanove et al, 2010). Furthermore, its sensitive detection excludes the need of high read depth required for shotgun metagenomics sequencing (Gaudin & Desnues, 2018).

Target enrichment workflow through hybrid-capture method able to directly detect respiratory viruses. However, no previous studies evaluate how accurate the target enrichment workflow guideline provided by Illumina in detecting SARS-CoV-2. This study will incorporate different bioinformatics pipelines for target enrichment workflow in detecting SARS-CoV-2 using Illumina NGS system. The aim of this study is to compare different bioinformatics pipelines towards target enrichment workflow by Illumina. The NGS data were obtained from 8 hospitalized patients in Yogyakarta and Central Java, tested positive for SARS-CoV-2 and took Real Time - Polymerase Chain Reaction (RT-PCR) swab test between May - September 2020. Prior to joining the study, patients were given informed consent and the study design was approved by the Medical and Health Research Ethics Committee of the Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada alongside Dr. Sardjito Hospital (KE/FK/0563/EC/2020). The first pipeline dubbed as ‘Fast Pipeline’ will directly map the raw NGS data to the SARS-CoV-2 reference genome. While the second ‘Normal Pipeline’ will map the raw NGS data to the human genome and proceed to map subsequent unmapped reads to the SARS-CoV-2 reference genome. The comparison in-

between pipelines should be observed all the way to identification of nucleotide substitutions and amino acids mutations.

1.2. Objective

The main objective of this study is to gain insight and understanding of target enrichment workflow by Illumina towards different bioinformatics pipelines. This study would utilize different pipelines called 'Fast Pipeline' and 'Normal Pipeline'. Ensuing specific parameters as listed below will be used for comparative analysis.

Detailed objectives of this study are as follows:

- Implementation of 'Fast Pipeline' to a patient's NGS data (quality control, reference assisted assembly using SARS-CoV-2 genome, and followed by implementation in [Branch 1] Nucleotide Substitution Pipeline and [Branch 2] Amino Acid Substitution Pipeline).
- Implementation of 'Normal Pipeline' to a patient's NGS data (quality control, reference assisted assembly using human genome, obtaining unmapped reads, reference assisted assembly using SARS-CoV-2 genome, and followed by implementation in [Branch 1] Nucleotide Substitution Pipeline and [Branch 2] Amino Acid Substitution Pipeline).
- Observation of count reads, coverage depth, nucleotide substitution, amino acids mutations, and runtime execution as the parameters of benchmarking.

1.3. Significance, Scope and Definitions

This study uses Whole Genome Sequencing (WGS) data collected through nasopharyngeal swabs from patients with COVID-19 from Yogyakarta and its surrounding area between May and September 2020. Ethical conduct was fully implemented considering the involvement of human subjects in the study. Written informed consent was obtained from all participants before joining the study. Furthermore, full study design was approved by the Medical and Health Research Ethics Committee of the Faculty of Medicine, Public

Health, and Nursing, Universitas Gadjah Mada alongside Dr. Sardjito Hospital (KE/FK/0563/EC/2020). The Worsening COVID-19 pandemic situation showed by rapid increase of new cases alongside fatalities forcing Indonesian government to take aggressive action through Testing, Tracing, and Treatment (3T) method. Furthermore, Illumina as the well known player in the NGS platform also took action by publishing Illumina enrichment workflow guidelines with the capabilities to directly and sensitively detect common respiratory viruses including SARS-CoV-2. This enrichment workflow eventually eliminates the need to map the reads towards the human genome as part of metagenomic sequencing. However, as mentioned above, no previous studies have been conducted to benchmark the capabilities of Illumina enrichment workflow in detecting SARS-CoV-2 by assessment through different downstream bioinformatics pipelines until Single Nucleotide Polymorphisms (SNPs) identification and variant analysis. Therefore, we constructed pipelines called 'Normal Pipeline' and 'Fast Pipeline' further described in Chapter 3. as a way of comparing the subsequent workflow. Benchmarking should be observed all the way to identification of nucleotide substitution and amino acids mutations. Parameters including count of reads, coverage depth, nucleotide and amino acid substitutions as well as runtime execution of both pipelines used as the main parameters for comparison.