

UNIVERSITI TEKNOLOGI MARA

**GENOMIC CHARACTERISATION
OF SARS-COV-2 BY
WHOLE GENOME SEQUENCING
AND ITS ASSOCIATION WITH
DISEASE SEVERITY**

NORAZIMAH BINTI TAJUDIN

Thesis submitted in fulfilment
of the requirements for the degree of
**Master of Science
(Medicine)**

Faculty of Medicine

November 2024

ABSTRACT

In this post-genomic era, whole genome sequencing is widely used to understand viral evolution and enhance epidemiological research. This study aims to describe the genome characterization of the SARS-CoV-2 virus and to associate the patient's clinical data (sociodemographic and co-morbidities) with the disease severity. A cross-sectional study was conducted using archived samples of laboratory-confirmed SARS-CoV-2 infection with Ct value ≤ 30 stored in viral transport media (VTM) from the Hospital Al-Sultan Abdullah (HASA) for one year (June 2021 until June 2022). The clinical data were retrieved from the UNIMEDS database and documented information from the medical records. Before being subjected to whole genome sequencing, the Ct values of the archived samples were validated by RT-PCR. The whole genome sequencing was performed using the Oxford Nanopore Technology (ONT) and assembled using bioinformatic tools EPI2ME software and Minimapv2.24 with quality scores Q10. Lineage assignment was done using Pangolin and NextClade with $>80\%$ coverage. Multiple logistic regression was performed to determine the association between the variables and the disease severity. Results showed that out of 129 samples, 86 were successfully sequenced. Among these, 1/86, (1.16%) was the Beta variant (B.1.354), Delta variant was 51/86, (59.3%) comprised of sublineages AY.59, AY.76 and AY.79. The Omicron variant has 34/86, (39.53%) which include subvariants BA.1.1, BA.1.17, BA.2, BA.2.3, and BA.2.10. Omicron was the most predominant variant detected, with BA.2 (21/51, 24.4%) being the most frequent subvariant, followed by Delta sub lineage AY.59 (21/34, 24.4%). The D614G mutation, was the predominant mutation that was found in all three variants, while mutations like K417N, N501Y were present in Beta and Omicron, and G142D, T478K, and D950N were observed in Delta and Omicron. Phylogenetic analysis reveals three distinct clades (GH, GK and GRA) with seven clusters indicating possible outbreaks occur within the area. Five variables were significantly associated with the disease severity with a p-value <0.05 . They were the VOC Delta (aOR = 22.354; 95% CI = 1.694, 294.921), age 49 years old (aOR = 2.143; 95% CI = 1.010, 4.546), hypertension (aOR = 33.600; 95% CI = 6.595, 171.175) chronic kidney disease (aOR = 25.840; 95% CI = 1.301, 513.20), and chronic illness (aOR = 28.677; 95% CI = 1.924, 427.334). At the point of the study, the Omicron variant's continued dominance could be explained due to the acquisition of mutations from other variants owing to the dynamic nature of SARS CoV-2 evolution. For example, mutations like D614G and N501Y were linked to increased transmissibility and were also found in Beta and Alpha variants. Since its emergence, Omicron has highly diversified and transformed into other Omicron subvariants. Based on the phylogenetic tree, the gradual displacement process of each variant was like other parts of Malaysia which corresponded with the reference strain used in this study. The clinical factors including the Delta variant as well as patients' comorbidity such as hypertension were associated with severe diseases and were consistent with other studies. Thus, the study of the SARS-CoV-2 genome and genomic surveillance will have a significant impact on viral trekking and monitoring to help in making informed decisions for public health preventive and containment strategies.

ACKNOWLEDGEMENT

Foremost, I offer my deepest gratitude to Allah for the strength and wisdom bestowed upon me throughout my Master's degree journey and for allowing me to complete this long and challenging journey after working as a government servant for more than a decade.

I am profoundly indebted to my supervisor, Dr. Siti Farah Alwani Md Naw, for her unwavering support and invaluable guidance. My sincere thanks extend to my co-supervisor, Associate Professor Dr. Wang Seok Mui, for her kind advice and guidance in doing the laboratory work, and Associate Professor Dr. Mariam Mohamad, who teach me most in Biostatistical analysis.

My heartfelt appreciation goes to the Institute of Molecular Medicine (IMMB), UiTM Sungai Buloh, my fellow postgraduate classmates, dedicated IMMB Staff, and seniors in the Faculty of Medicine. I extend my special thanks to from Universiti Malaysia Pahang (UMP), and laboratory staff from IPromise, UiTM Puncak Alam, and my colleagues in the Molecular Laboratory, Virology Section at MKAK Sungai Buloh. Their support and collaboration were appreciated in helping me in the most challenging phases of my lab work and bioinformatics analysis. I owe a lot to the experts who always encouraged and helped me in this tough journey.

Finally, this thesis is dedicated to my supportive husband, to my dear daughter and son for sacrificing their time to understand my vision and determination to complete this journey, and to both my parents, who have always been my backbone, encouraging me and never giving up on their daughter's pursuit of education.

To all of you, I dedicate this piece of victory. Alhamdulillah.

TABLE OF CONTENTS

	Page
CONFIRMATION BY PANEL OF EXAMINERS	ii
AUTHOR'S DECLARATION	iii
ABSTRACT	iv
ACKNOWLEDGEMENT	v
TABLE OF CONTENTS	vi
LIST OF TABLES	x
LIST OF FIGURES	xii
LIST OF SYMBOLS	xiii
LIST OF ABBREVIATIONS	xiv
LIST OF NOMENCLATURE	xv
CHAPTER 1 INTRODUCTION	1
1.1 Research Background	1
1.2 Motivation for This Work	2
1.3 Problem Statement	2
1.4 Research Objectives	3
1.4.1 General Objective	3
1.4.2 Specific Objectives	3
1.5 Research Question	4
1.6 Significance of Study	4
1.7 Limitations	5
1.8 Hypothesis	5
1.9 Thesis Scope	5
CHAPTER 2 LITERATURE REVIEW	7
2.1 Introduction	7
2.2 Genome Arrangement of SARS-CoV-2	8
2.3 Mutation	10
2.4 Mutation on Spike Gene	11

CHAPTER 1

INTRODUCTION

1.1 Research Background

In December 2019, a novel severe acute respiratory syndrome-related coronavirus, which originally started in Wuhan, China, caused a global outbreak. The majority of patients who were infected in the outbreak at that time had exposure to exotic wildlife animals such as snakes, birds, and small mammals including marmots, bats, frogs and hedgehogs sold at the Huanan Seafood Wholesale Market in Wuhan (Fani et al., 2020; Tiwari et al., 2020a). They exhibited symptoms of typical viral pneumonia, such as fever and shortness of breath, and, in severe cases, developed signs such as bilateral lung infiltration (Gralinski & Menachery, 2020b). The clinical course and prognosis of patients infected with SARS-CoV-2 also cause higher severity and death in those older patients with diabetes (Clark et al., 2020; Holman et al., 2020; Williamson et al., 2020). The patients included in class III obesity ($\text{BMI} \geq 40 \text{ kg/m}^2$) and above were also associated with increased odds for mortality (Tadayon Najafabadi et al., 2023), and is one of the risk factors for the progression of major complications (Demeulemeester et al., 2021). On the 11th of February 2020, The International Committee on Taxonomy of Viruses (ICTV), in communication with the World Health Organization (WHO), officially classified this novel virus as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and the disease it caused was named Coronavirus Disease 2019 (COVID-19). Whole genome sequencing (WGS) was one of the surveillance tools that played a role in identifying genomic variation, evolution, and mutational profile and characterising the epidemiology of SARS-CoV-2 concerning the regional and worldwide virus lineages. The SARS-CoV-2 variants of concern (VOC) share many types of mutations and have become the biggest concern in public health measures as they are capable of causing rapid transmission, increasing their virulence and evading immune responses. In Malaysia, a consortium was established between MOH-MOSTI-KPT to coordinate WGS activities for the country so that more genomic data submissions would be available to keep track of the variant of interest (VOI) and VOC transmission (Muhammad Azami et al.,