Sains Medika: Jurnal Kedokteran dan Kesehatan

journal homepage: http://jurnal.unissula.ac.id/index.php/sainsmedika/

RESEARCH ARTICLE

Bioinformatics study of the a222V substitution in sars-cov-2 spike protein

Kiky Martha Ariesaka[1*], Moh Mirza Nuryady[2,3], Ika Agus Rini[4,5]

[1]Department of Medicine, Faculty of Medicine, State University of Malang, Malang, Indonesia
[2]Biology Education Department, Faculty of Teacher and Training Education, University of Muhammadiyah Malang, Malang, Indonesia
[3]Institute of Parasitology, Department of Biological Science and Pathobiology, Veterinary Medicine University of Vienna, Vienna, Austria
[4]Biology Study Program, Institut Teknologi Sumatera, Indonesia
[5]Department of Integrative Biotechnology, Sungkyunkwan University Gyeonggi-do, Suwon, Jangan-gu, Cheoncheon-dong, Seobu-ro 2066, South Korea
*Correspondence: Kiky Martha Ariesaka; Address: Jl. Semarang No. 5 Malang - Jawa Timur 65145; Email address: kiky.martha.fk@um.ac.id

ARTICLE INFO

SARS-CoV-2
A222V substitution
Spike protein
Phylogenetic analysis
COVID-19
Mutations

ABSTRACT

COVID-19, caused by the SARS-CoV-2 virus, has resulted in over 200 million infections and 4.3 million global deaths, including more than 3.7 million cases and 110,000 fatalities in Indonesia. SARS-CoV-2, a single-stranded RNA virus, encodes several proteins, including Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N). The S protein, which consists of S1 and S2 domains, plays a critical role in binding to the ACE2 receptor and facilitating membrane fusion. This study investigates the phylogenetic relationship of the A222V substitution in the S protein relative to various global isolates. We analysed 25 complete SARS-CoV-2 sequences sourced from the GISAID database, conducted multiple sequence alignments using Clustal W in MEGA-X, and constructed phylogenetic trees through the neighbour-joining method. The urgency of this research stems from the need to understand how the A222V mutation affects the virus's infectivity and transmission, which is vital for developing targeted treatments and vaccines. Initially identified in Spain in the summer of 2020, the A222V variant has rapidly spread across Europe, raising concerns about its potential to increase transmissibility or undermine vaccine efficacy. This variant appears to affect protein conformation or stability rather than receptor binding or membrane fusion, underscoring the need to understand these alterations, as they could influence the virus's behaviour and the effectiveness of public health measures. Our findings indicate high mutation rates in the S gene, characterised by diverse point mutations. Phylogenetic analysis demonstrated that A222V isolates are closely clustered with D614G variants. In conclusion, although not all mutations result in amino acid changes, the A222V variant has a limited impact on ACE2 binding but may affect protein stability.

1. Introduction

In mid-December 2019, a novel coronavirus (CoV) emerged in Wuhan, China, marking the beginning of a global health crisis (Wong, 2020; Huang *et al.,* 2020; Kaushal *et al.,* 2020). The International Committee on

https://doi.org/10.30659/sainsmed.v16i2.39242
Received 26 August 2024; Received in revised form 31 August 2024; Accepted 13 December 2025

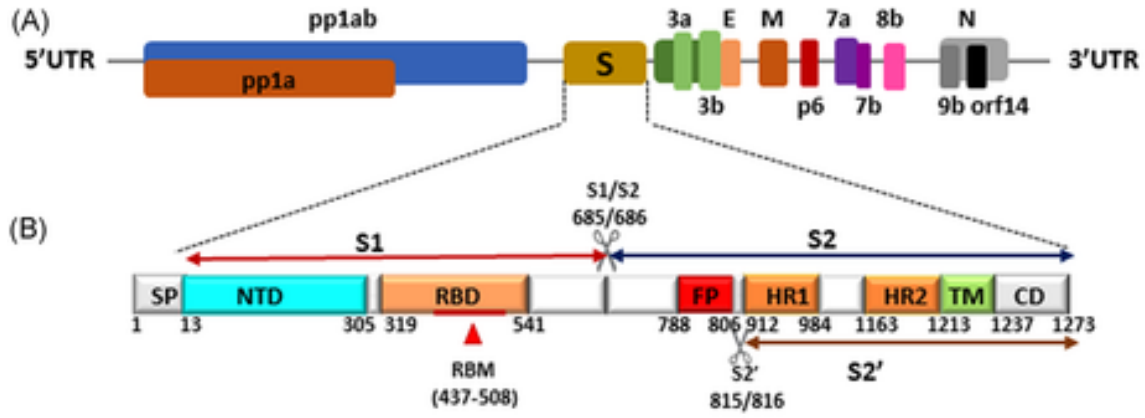


Figure 1 . Schematic structure of the SARS-CoV-2 Spike (S) protein (A) Schematic of the SARS-CoV-2 genome; (B) Schematic of the SARS-CoV-2 S protein (Zhu ***et al.,*** 2021)

Taxonomy of Viruses (ICTV) subsequently designated this virus as "severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)" on February 11, 2020 (WHO, 2020). SARS-CoV-2 is recognised as the causative agent of Coronavirus Disease 2019 (COVID-19) (Huang *et al.,* 2020; Wang *et al.,* 2020; Zhang *et al.,* 2020). As of August 10, 2021, COVID-19 has resulted in over 200 million confirmed cases and 4.3 million deaths worldwide (WHO, 2021). In Indonesia, the number of cases has exceeded 3.7 million, resulting in more than 110,000 fatalities (COVID-19 Task Force, 2021).

SARS-CoV-2 is a single-stranded RNA virus with a genome size ranging from 29.8 to 29.9 kb, encoding several proteins crucial for its replication and pathogenesis, including Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N) proteins, along with various open reading frames (ORFs) such as ORF1a, ORF1b, ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF14, and ORF10 (Alouane *et al.,* 2020). The S protein consists of 1,273 amino acids and is divided into S1 and S2 domains (Huang *et al.,* 2020). The S1 domain mediates attachment to the ACE2 receptor and is further divided into two subdomains: the N-terminal domain (NTD) and the receptor-binding domain (RBD), as illustrated in Figure 1. The S2 domain comprises a fusion peptide, a heptad repeat, a transmembrane domain, and a cytoplasmic domain (Zhu *et al.,* 2021). This structural protein, located on the viral surface, plays a critical role in binding to the angiotensin-converting enzyme 2 (ACE2) receptor on host cells, thereby initiating membrane fusion, a crucial step in the infection process (Zhang *et al.,* 2020).

Several mutations in the Spike protein have been identified that may increase viral infection efficiency (Harvey *et al.,* 2021; Kaushal *et al.,* 2020). Understanding these mutations and their effects on the virus is essential for developing effective therapeutic strategies and public health measures in combating COVID-19.

Currently, there are three primary nomenclature systems for categorizing SARS-CoV-2: (1) the Nextstrain system, which comprises 11 main clades—19A, 19B, and 20A-20I; (2) the PANGOLIN lineage established by Rambaut *et al.,* consisting of five principal lineages—A, B, A.1, B.1.1, and B.1.177; and (3) the Global Initiative on Sharing All Influenza Data (GISAID), which classifies SARS-CoV-2 sequences into seven main clades based on distinct mutation signatures. According to the GISAID classification, clade L is recognised as the wild-type (WT) variant, containing early isolates from China, including the reference genome, as well as isolates from other countries involved in the initial stages of the pandemic. Clade S is distinguished by the L84S mutation in NS8, while clade V is characterised by the G251V mutation in NS3. The D614G mutation in the Spike protein is a notable feature of clade G, which has become the dominant isolate since spring 2020. From clade G, three additional clades—GH, GR, and GV—have emerged, each sharing identical mutations except for D614G. The GH clade is marked by the Q57H mutation in NS3, the GR clade by the G204R mutation in the N-protein, and the GV clade by the A222V mutation in the S-protein. Furthermore, the GISAID nomenclature includes an eighth clade, designated clade 'O', though it remains ill-defined and is often grouped with clade V in various GISAID reports (Alouane *et al.,* 2020; Gaffar *et al.,* 2021; Shahhosseini *et al.,* 2021). The SARS-CoV-2 variant 20A.EU1 (S: A222V) first appeared in early summer 2020 in Spain and then spread to several locations in Europe. During the summer, the frequency increases in parallel in many countries. S: A222V is in the A domain of the Spike protein, which is also referred to as the NTD (Hodcroft *et al.,* 2021). S: A222V likely does not play a direct role in receptor binding or membrane fusion, but mutations can sometimes mediate long-term effects

on protein conformation or stability (McCallum *et al.,* 2020; Tortorici *et al.,* 2020). This study aims to analyse the phylogenetic relationship of S: A222V SARS-CoV-2 with various isolates from several countries, analysing mutations and their effect on amino acid changes.

2. Materials and Methods

2.1 Data Collection

A total of 25 complete SARS-CoV-2 sequences were obtained from 17 randomly selected countries, spanning different time periods throughout the pandemic and encompassing all clades identified in the GISAID database (https://www.gisaid.org/, accessed on August 10, 2021). Each sequence was named according to its accession ID, followed by the country of origin. The pandemic periods were categorised as follows: period 1—December 2019; period 2—January to June 2020; period 3—July to December 2020; period 4—January to March 2021; and period 5—April to June 2021. Sequences that were similar in clade classification, isolation time, and geographic origin were excluded from the final analysis. Additionally, the reference sequence of the first SARS-CoV-2 isolate from Wuhan (NC_045512.2) was retrieved from GenBank (https://www.ncbi.nlm.nih.gov/, accessed on August 10, 2021) to serve as the wild-type (WT) variant.

2.2 Multiple Sequence Alignment

All selected sequences were saved in FASTA format. Multiple sequence alignment (MSA) was performed using the Clustal W method in MEGA-X (Molecular Evolutionary Genetics Analysis) software. The opening gap and extension penalties were set at 15.00 and 6.66, respectively. Each sample's sequence was trimmed based on the A222V SNP sequence (EPI_ISL_2450736, Italy) to remove any truncated sequences from other samples. The alignment results were subsequently saved in FASTA format for further analysis of amino acid changes. The protein structure of the A222V sample (EPI_ISL_2450736, Italy) was visualised using GISAID (https://www.gisaid.org/, accessed on August 11, 2021) by entering the nucleotide sequence after MSA.

2.3 Phylogenetic Tree Construction

The results from the MSA were used to construct a phylogenetic tree to assess the evolutionary relationships between the A222V sample and other international isolates, using MEGA-X. The neighbour-joining method was selected for phylogenetic tree construction due to its efficiency, as the dataset lacked truncated sequences. Bootstrap analysis, crucial for evaluating the robustness of the phylogenetic model, was performed with 1,000 replications; sequences with low bootstrap values were excluded from further analysis to enhance the reliability of the phylogenetic tree. The substitution model employed was determined based on the model selection criteria provided by MEGA-X, with the model exhibiting the lowest Bayesian Information Criterion (BIC) value selected.

3. Results and Discussion

3.1 Multiple Sequence Alignment Analysis

The nucleotide and amino acid sequences of the Spike protein were analysed using Multiple Sequence Alignment (MSA) conducted with the Clustal W method. This approach allows systematic alignment of nucleotide and amino acid sequences, facilitating

	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	
	5	7	7	7	7	7	7	8	9	1	2	3	3	6	7	7	7	7	7	8	9	1	1	2	3	4	6	7	8	9	0	0	0	5	5	7	7	8	1	2	2	2	3	
	8	1	2	3	5	5	6	9	6	1	0	3	9	5	2	2	4	9	9	8	8	3	3	5	3	3	0	3	3	3	7	8	9	1	7	3	9	2	8	3	5	9	3	
	9	9	7	4	1	2	7	6	1	1	2	3	5	1	1	6	6	7	8	9	2	5	7	9	8	9	2	1	1	1	1	3	2	0	8	2	6	2	9	9	5	7	5	
EPI ISL 2450736 Italy	C	G	A	C	G	G	C	T	T	C	G	G	G	T	C	G	G	A	A	T	C	G	G	C	C	C	G	C	T	G	C	T	C	S	C	T	G	T	G	A	C	G	A	Σ
EPI ISL 3274295 Italy	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	4
EPI ISL 3239841 Germany	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	T	G	.	G	.	.	.	.	.	4
EPI ISL 1722078 Germany	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	4
EPI ISL 3276946 India	.	.	.	.	.	.	.	.	C	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	5
EPI ISL 3332727 Japan	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	4
EPI ISL 3321272 Chile	.	.	.	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	G	.	G	C	G	.	.	T	.	.	8
EPI ISL 1595846 Japan	.	.	.	.	T	.	.	.	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	7
EPI ISL 3322859 USA	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	T	.	5
EPI ISL 3320811 Norway	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	T	.	5
EPI ISL 767938 Australia	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	T	.	6
EPI ISL 2990761 USA	.	.	.	.	.	T	.	.	C	.	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	C	.	G	.	G	.	G	.	.	.	T	.	9
EPI ISL 3281606 Spain	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	4
EPI ISL 3322841 USA	.	.	.	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	G	.	G	.	G	.	.	.	.	.	6
EPI ISL 3253279 Ukraine	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	G	.	G	.	G	.	.	.	.	.	7
EPI ISL 1573461 Germany	T	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	G	.	G	.	G	.	G	.	.	.	7
EPI ISL 529213 China	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	5
EPI ISL 3190068 Spain	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	T	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	6
EPI ISL 3003279 UK	.	T	.	T	.	.	A	.	C	.	.	.	T	.	.	.	.	.	.	.	.	T	A	.	G	.	.	.	C	.	.	.	.	G	.	G	.	G	.	.	.	.	.	12
EPI ISL 3274165 Egypt	.	.	.	.	.	.	.	.	C	.	T	.	.	.	.	.	.	.	C	.	.	.	A	T	.	.	.	T	.	.	.	.	.	G	.	G	.	G	T	.	.	.	T	11
EPI ISL 529216 China	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	5
EPI ISL 3105864 Phillippines	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	5
EPI ISL 1312639 Latvia	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	5
EPI ISL 2591211 Finland	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	5
EPI ISL 2984742 Turkey	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	7
NC 045512.2 China	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	G	.	G	.	.	.	.	.	5
Σ	1	1	1	1	1	1	1	1	22	1	1	1	1	1	1	2	1	1	2	1	1	1	10	1	1	1	1	1	1	1	1	1	2	25	1	25	1	25	1	1	1	4	1	151

Figure 2. Mutation point. Calculations based on MSA of Spike protein nucleotides (21,563-25,384nt) from 26 samples.

	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	
	2	2	2	2	2	3	4	4	5	5	5	6	6	6	7	7	7	9	0	0	1	2	2	4		
	4	4	4	5	9	2	0	4	5	7	7	0	3	6	1	1	8	4	2	2	3	9	6	7	1	
	0	3	5	1	9	1	1	5	1	4	6	0	0	1	2	3	0	4	4	8	1	3	6	4	9	
EPI ISL 2450736 Italy	**W**	**I**	**T**	**G**	**I**	**F**	**R**	**D**	**C**	**P**	**A**	**M**	**L**	**S**	**R**	**V**	**L**	**L**	**T**	**V**	**A**	**S**	**V**	**H**	**R**	**∑**
EPI ISL 3274295 Italy	.	F	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	2
EPI ISL 3239841 Germany	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	F	.	Q	.	2
EPI ISL 1722078 Germany	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	1
EPI ISL 3276946 India	.	.	.	.	.	L	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	3
EPI ISL 3332727 Japan	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	2
EPI ISL 3321272 Chile	.	.	.	.	.	L	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	L	Q	W	5
EPI ISL 1595846 Japan	.	.	.	*	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	2
EPI ISL 3322859 USA	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	2
EPI ISL 3320811 Norway	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	2
EPI ISL 767938 Australia	.	.	.	.	.	L	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	Q	.	3
EPI ISL 2990761 USA	.	.	.	V	.	L	.	.	G	.	.	.	P	.	.	.	.	.	.	A	.	.	.	Q	.	6
EPI ISL 3281606 Spain	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Q	.	2
EPI ISL 3322841 USA	.	.	.	.	T	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	E	.	.	Q	.	4
EPI ISL 3253279 Ukraine	.	.	.	.	.	L	.	.	.	.	T	.	.	F	.	.	.	.	I	.	.	.	.	Q	.	5
EPI ISL 1573461 Germany	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	E	.	.	Q	.	3
EPI ISL 529213 China	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	3
EPI ISL 3190068 Spain	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	3
EPI ISL 3003279 UK	L	.	I	.	.	L	.	.	.	.	.	.	.	.	M	M	V	P	.	.	.	.	.	Q	.	8
EPI ISL 3274165 Egypt	.	.	.	.	.	L	L	.	.	.	.	L	.	.	.	M	.	.	.	.	.	.	.	Q	.	5
EPI ISL 529216 China	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	3
EPI ISL 3105864 Phillippines	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	3
EPI ISL 1312639 Latvia	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	3
EPI ISL 2591211 Finland	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	3
EPI ISL 2984742 Turkey	.	.	.	.	.	L	.	.	.	.	T	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	4
NC 045512.2 China	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Q	.	3
∑	**1**	**1**	**1**	**1**	**1**	**22**	**1**	**1**	**1**	**1**	**2**	**2**	**1**	**1**	**1**	**10**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**25**	**1**	**82**

Figure 3. Protein changes. Calculations based on MSA of Spike protein (7,187-8,461) from 26 samples.

comparison of mutations in the sample sequences with the A222V reference sequence. Clustal W utilises a progressive alignment strategy, aligning the most similar sequences first, then progressively more dissimilar ones, to achieve a comprehensive global alignment. The algorithm initiates the process by calculating a distance matrix for each pair of sequences based on pairwise sequence alignment scores, subsequently employing the maximum likelihood method to generate a guide tree that underpins the global alignment (Gaffar *et al.,* 2021).

The alignment results revealed that all analysed samples exhibited distinct point mutations relative to the A222V variant, with each sequence containing 4–12 point mutations. This finding indicates a notable mutation rate across various positions in the Spike protein gene. In total, 151 nucleotide mutations were identified within the Spike protein across all samples (Figure 2). Out of these 151 mutations, 82 resulted in significant amino acid changes (Figure 3). Notably, these results demonstrate that not all nucleotide mutations alter the corresponding amino acids; furthermore, some mutations that do result in amino acid changes may preserve the protein's functional properties.

The observed high mutation rate highlights the dynamic evolutionary nature of SARS-CoV-2, especially within the Spike protein, which is vital for viral entry and infectivity. Mutations in the Spike protein can influence virus transmissibility, pathogenicity, and the effectiveness of vaccine-induced immunity. Understanding how these mutations are distributed and their impact is essential for ongoing surveillance and public health strategies to control the spread of COVID-19. Future research should examine the functional effects of these mutations, particularly those that alter amino acid properties, to better understand their role in viral evolution and adaptation.

3.2 Phylogenetic Analysis

Before constructing the phylogenetic tree and performing bootstrap analysis, the optimal substitution model was selected using MEGA-X. A variety of models are available for this purpose, including General Time Reversible (GTR), Hasegawa-Kishino-Yano (HKY), Tamura-Nei (TN), Tamura 3-parameter (T93), Kimura 2-parameter (K2P), and Jukes-Cantor (JC). The best substitution model is determined based on the model with the lowest Bayesian Information Criterion (BIC) score. To account for potential non-uniformity in evolutionary rates across sites, a discrete gamma distribution (+G) with five rate categories can be employed, along with the assumption that a small proportion of sites remains evolutionarily invariable (+I) (Kumar *et al.,* 2018; Nei & Kumar, 2000).

The model selection results for the 26 samples indicated that the GTR model with a discrete gamma distribution (GTR+G) had the lowest BIC, suggesting

Table 1. Results of substitution model analysis

Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G	58	85261,597	84591,874	-42237,932	n/a	0,05	1,41	0,299	0,322	0,183	0,196	0,046	0,022	0,070	0,043	0,154	0,082	0,036	0,270	0,018	0,107	0,135	0,017
TN93+G	55	85263,740	84628,657	-42259,325	n/a	0,05	1,80	0,299	0,322	0,183	0,196	0,055	0,031	0,069	0,051	0,175	0,034	0,051	0,307	0,034	0,106	0,055	0,031
GTR+G+I	59	85270,658	84589,388	-42235,689	0,49	0,05	1,41	0,299	0,322	0,183	0,196	0,046	0,022	0,070	0,043	0,155	0,083	0,035	0,272	0,017	0,106	0,136	0,016
TN93+G+I	56	85273,599	84626,970	-42257,481	0,50	0,08	1,80	0,299	0,322	0,183	0,196	0,055	0,031	0,069	0,051	0,175	0,034	0,051	0,308	0,034	0,105	0,055	0,031
GTR	57	85274,939	84616,762	-42251,377	n/a	n/a	1,40	0,299	0,322	0,183	0,196	0,046	0,022	0,070	0,043	0,153	0,082	0,037	0,269	0,018	0,108	0,134	0,017
GTR+I	58	85275,577	84605,854	-42244,922	0,50	n/a	1,40	0,299	0,322	0,183	0,196	0,046	0,022	0,070	0,043	0,153	0,082	0,036	0,269	0,018	0,107	0,135	0,017
TN93	54	85276,466	84652,930	-42272,461	n/a	n/a	1,80	0,299	0,322	0,183	0,196	0,055	0,031	0,070	0,051	0,174	0,034	0,051	0,306	0,034	0,107	0,055	0,031
TN93+I	55	85277,260	84642,177	-42266,085	0,50	n/a	1,80	0,299	0,322	0,183	0,196	0,055	0,031	0,070	0,051	0,175	0,034	0,051	0,307	0,034	0,107	0,055	0,031
HKY+G	54	85284,727	84661,191	-42276,592	n/a	0,05	1,80	0,299	0,322	0,183	0,196	0,055	0,031	0,129	0,051	0,120	0,034	0,051	0,211	0,034	0,196	0,055	0,031
HKY+G+I	55	85294,457	84659,374	-42274,683	0,47	0,05	1,79	0,299	0,322	0,183	0,196	0,055	0,032	0,128	0,051	0,120	0,034	0,051	0,211	0,034	0,196	0,055	0,032
HKY	53	85297,450	84685,461	-42289,727	n/a	n/a	1,80	0,299	0,322	0,183	0,196	0,055	0,031	0,129	0,051	0,120	0,034	0,051	0,211	0,034	0,196	0,055	0,031
HKY+I	54	85298,142	84674,605	-42283,299	0,50	n/a	1,80	0,299	0,322	0,183	0,196	0,055	0,031	0,129	0,051	0,120	0,034	0,051	0,211	0,034	0,196	0,055	0,031
T92+G	52	85303,341	84702,898	-42299,446	n/a	0,05	1,80	0,310	0,310	0,190	0,190	0,053	0,033	0,124	0,053	0,124	0,033	0,053	0,204	0,033	0,204	0,053	0,033
T92+G+I	53	85312,982	84700,993	-42297,493	0,48	0,05	1,79	0,310	0,310	0,190	0,190	0,053	0,033	0,124	0,053	0,124	0,033	0,053	0,203	0,033	0,203	0,053	0,033
T92	51	85316,128	84727,232	-42312,613	n/a	n/a	1,80	0,310	0,310	0,190	0,190	0,053	0,033	0,124	0,053	0,124	0,033	0,053	0,204	0,033	0,204	0,053	0,033
T92+I	52	85316,920	84716,477	-42306,235	0,50	n/a	1,80	0,310	0,310	0,190	0,190	0,053	0,033	0,124	0,053	0,124	0,033	0,053	0,204	0,033	0,204	0,053	0,033
K2+G	51	87090,962	86502,066	-43200,030	n/a	0,05	1,80	0,250	0,250	0,250	0,250	0,045	0,045	0,161	0,045	0,161	0,045	0,045	0,161	0,045	0,161	0,045	0,045
K2+G+I	52	87098,478	86498,036	-43197,014	0,56	0,05	1,81	0,250	0,250	0,250	0,250	0,045	0,045	0,161	0,045	0,161	0,045	0,045	0,161	0,045	0,161	0,045	0,045
K2	50	87105,344	86527,995	-43213,994	n/a	n/a	1,80	0,250	0,250	0,250	0,250	0,045	0,045	0,161	0,045	0,161	0,045	0,045	0,161	0,045	0,161	0,045	0,045
K2+I	51	87105,596	86516,700	-43207,347	0,50	n/a	1,80	0,250	0,250	0,250	0,250	0,045	0,045	0,161	0,045	0,161	0,045	0,045	0,161	0,045	0,161	0,045	0,045
JC+G	50	87177,894	86600,545	-43250,269	n/a	0,05	0,50	0,250	0,250	0,250	0,250	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083
JC+G+I	51	87185,556	86596,660	-43247,327	0,56	0,05	0,50	0,250	0,250	0,250	0,250	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083
JC	49	87192,224	86626,422	-43264,208	n/a	n/a	0,50	0,250	0,250	0,250	0,250	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083
JC+I	50	87192,585	86615,236	-43257,615	0,50	n/a	0,50	0,250	0,250	0,250	0,250	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+*G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+*I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 26 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. There were a total of 29408 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

it is the most appropriate model for subsequent phylogenetic analysis. However, given the substantial number of samples and nucleotides in this dataset, a decision was made to utilise the next lowest BIC model, the Tamura-Nei model with a gamma distribution (TN93+G) (Table 1).

In the present study, phylogenetic analysis was conducted using the neighbour-joining and bootstrap methods, employing 1000 replications alongside the Tamura-Nei substitution model with a discrete Gamma distribution (TN93+G). The constructed cladogram effectively delineates six distinct clades (I-VI), reflecting the genetic diversity and evolutionary trajectories of the viral isolates. Figure 4 shows the clade analysis and characterisations, including:

- Clade I comprises the isolates EPI_ISL_529213 from China, EPI_ISL_3190068 from Spain, EPI_ISL_3003279 from the UK, and EPI_ISL_3274165 from Egypt. These isolates are unified by the L84S mutation in the NS8 protein, indicating a potentially unique evolutionary adaptation within these geographic regions.
- Clade II encompasses wild-type variants, including isolates EPI_ISL_529216 from China, EPI_ISL_3105864 from the Philippines, EPI_ISL_1312639 from Latvia, EPI_ISL_2591211 from Finland, and NC 045512.2 from China. These isolates represent early variants central to the pandemic's initial spread, underscoring the foundational role of these genotypes in subsequent evolutionary developments.
- Clade III is characterised by the D614G mutation in the S-protein, observed in EPI_ISL_3281606 from Spain, EPI_ISL_3322841 from the USA, and EPI_ISL_1573461 from Germany. This mutation has achieved predominance globally since spring 2020, highlighting its adaptive advantage and widespread dissemination.
- Clade IV includes isolates EPI_ISL_2450736 and EPI_ISL_3274295 from Italy, as well as EPI_ISL_3239841 and EPI_ISL_1722078 from Germany. These isolates harbour both D614G and A222V mutations in the S-protein, suggesting a further evolutionary step within the GV clade, as corroborated by GISAID classifications.
- Clade V features the D614G mutation in the S-protein alongside G204R in the N-protein. It includes EPI_ISL_3276946 from India, EPI_ISL_3332727 from Japan, EPI_ISL_3321272 from Chile, and EPI_ISL_1595846 from Japan, indicating parallel mutations that may infer similar selective pressures across diverse geographical locations.
- Clade VI includes isolates EPI_ISL_3322859 from the USA, EPI_ISL_767938 from Australia, and EPI_ISL_2990761 from the USA, characterised by D614G in the S-protein and Q57H in the NS3 protein. This clade represents a distinct evolutionary pathway which diverges from clades characterised by other significant mutations.

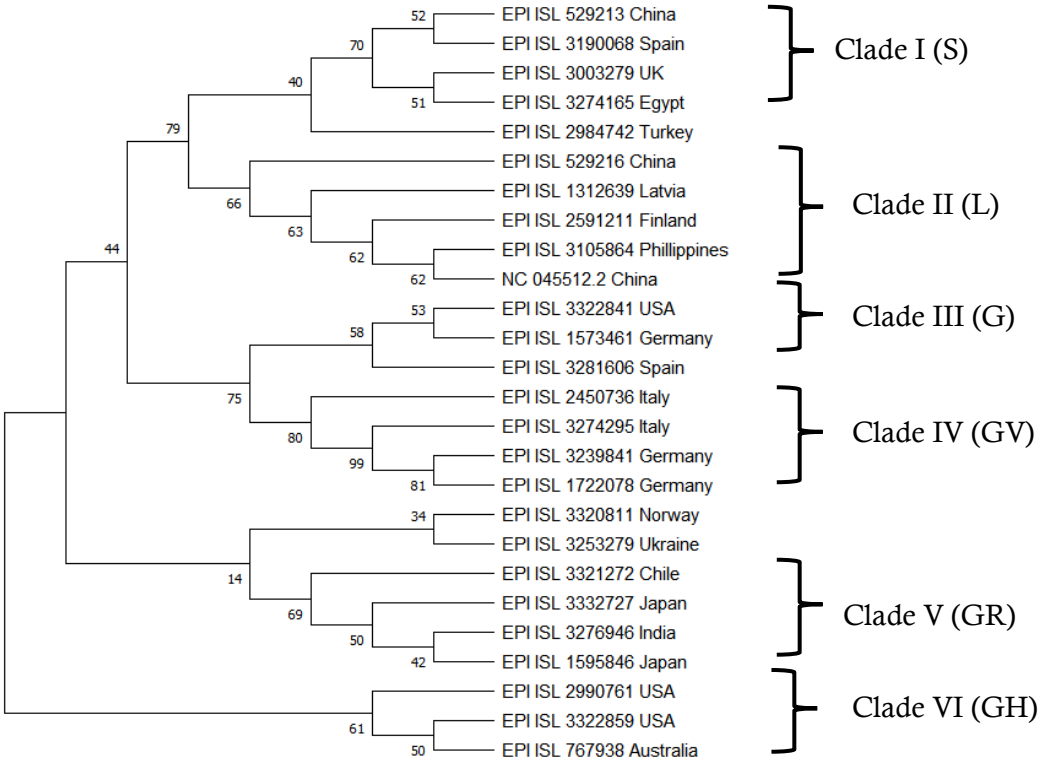
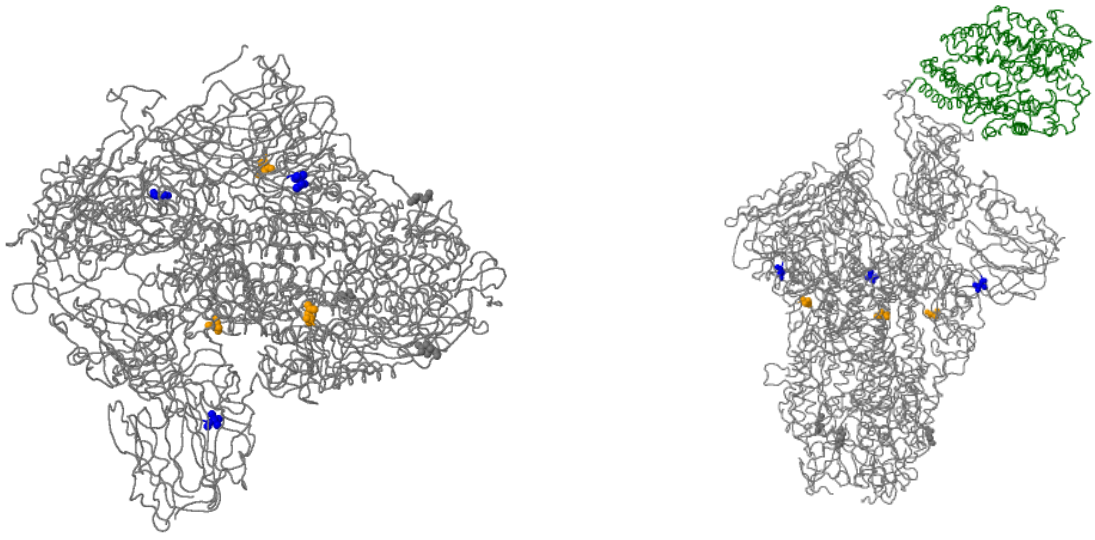


Figure 4. Bootstrap phylogenetic tree of SARS-CoV-2 on 26 samples. Bootstrap values (1000 replicates) for each node are shown at the base of the branch.



(a) **(b)**

Figure 5. (a) Spike protein structure of sample A222V (EPI_ISL_2450736 Italy) with the RBD in the lower conformation, accompanied by variations A222V (blue), D614G (yellow), and E1072Z (grey). (b) Spike protein receptor complex of sample A222V (EPI_ISL_2450736 Italy) with ACE2 (green ribbon).

Bootstrap values, ranging from 70 to 100, indicate potential changes within small clades, while values below 70 suggest variations at the larger clade level. This differential in bootstrap values underscores the robustness of the phylogenetic affiliations and provides insights into evolutionary stability and divergence.

The phylogenetic elucidation confirms that isolates with the SNP A222V mutation are collocated within Clade IV. Their evolutionary lineage is proximal to Clade III (S: D614G) with a bootstrap value of 75, while maintaining a distant relationship to Clade VI (Q57H). This concords with established GISAID classifications, affirming that Clade IV (GV) represents an evolutionary derivation from Clade III (G), driven

by the combined D614G and A222V mutations in the spike protein. Through this analysis, the study elucidates the intricate relationships between SARS-CoV-2 isolates across geographical and temporal dimensions, contributing to the broader understanding of viral evolution and adaptation mechanisms.

3.3 Analysis of changes in protein structure

The three-dimensional structure of the protein from the A222V sample (EPI_ISL_2450736 Italy) displays a 99.607% similarity to the hCoV-19/Wuhan/WIV04/2019 sequence. In Figure 5a, variations are evident with A222V marked in blue, D614G in yellow, and E1072Z in grey. Upon interaction with the ACE2 receptor, it appears that the S:A222V mutation may not significantly impact the Spike protein's binding affinity to ACE2, as this mutation does not directly affect the binding interface between the Spike and ACE2 proteins. This observation is consistent with the findings of Hodcroft et al. (2021), who reported that although the S:A222V mutation does not directly affect receptor binding or membrane fusion in SARS-CoV-2, it can still lead to long-term conformational or stability changes in the protein.

The A222V mutation, located in the N-terminal domain (NTD) of the Spike protein, gained attention as it spread across Europe, particularly associated with summer travellers to Spain. This variant, often termed the "Spanish variant" or 20A.EU1, was implicated in the UK's second COVID-19 wave from October to November 2020. To evaluate the potential functional impact of the S:A222V mutation on the Spike protein's ability to facilitate viral entry, Hodcroft et al. conducted experiments using pseudotyped lentiviral particles. These particles, engineered to include or exclude the A222V mutation and alterations in the spike's cytoplasmic tail, revealed that those with the S: A222V mutation exhibited slightly higher titers—approximately 1.3 times greater than those without the mutation. However, this difference was not statistically significant. Unlike the D614G mutation, which is widely recognised for enhancing SARS-CoV-2 virulence, the A222V mutation does not produce a comparable increase in spike pseudotype lentiviral titers. It is crucial to interpret these minor effects with caution, as the implications of mutations on viral transmission in human populations may not align directly with outcomes in laboratory settings (Hodcroft *et al.,* 2021).

4. Conclusions

While not all mutations result in amino acid changes, and some changes may not significantly alter protein properties, the S: A222V mutation has been shown to have minimal impact on the binding affinity of the Spike protein to ACE2, as it does not directly interact with the binding site. Phylogenetic analysis further reveals that clade S: A222V (Clade IV) is most closely related to clade S: D614G (Clade III), underscoring the latter as the ancestral lineage of the S:A222V variant. This relationship highlights the evolutionary trajectory of these variants and suggests that, while S:A222V may exhibit some differences, it does not confer substantial changes in receptor binding. Understanding these relationships is critical for assessing the potential implications of such mutations on viral transmission and pathogenicity.

Conflict of interest

All authors have no conflict of interest in this article.

References

Alouane, T., Laamarti, M., Essabbar, A., Hakmi, M., Bouricha, E. M., Chemao-Elfihri, M. W., Kartti, S., Boumajdi, N., Bendani, H., Laamarti, R., Ghrifi, F., Allam, L., Aanniz, T., Ouadghiri, M., El Hafidi, N., El Jaoudi, R., Benrahma, H., Attar, J. El, Mentag, R., Sbabou, L., Nejjari, C., Amzazi, S., Belyamani, L., Ibrahimi, A. (2020). Genomic diversity and hotspot mutations in 30,983 SARS-CoV-2 genomes: Moving toward a universal vaccine for the "Confined virus"? *Pathogens*, 9(829): 1-19. https://doi.org/10.3390/pathogens9100829

COVID-19 Task Force. (2021). Situasi Virus COVID-19 di Indonesia.

Gaffar, S., Rahmani, S. A. F., & Hardianto, A. (2021). Mutation and phylogenetic analysis of spike glycoprotein of indonesian isolates of severe-acute-respiratory-syndrome-coronavirus-2 (SARS-CoV-2). *Majalah Kedokteran Bandung*, 53(1): 38-47. https://doi.org/10.15395/mkb.v53n1.223

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology,* 19(7): 409–424. https://doi.org/10.1038/s41579-021-00573-0

Hodcroft, E. B., Zuber, M., Nadeau, S., Vaughan, T. G., Crawford, K. H. D., Althaus, C. L., Reichmuth, M. L., Bowen, J. E., Walls, A. C., Corti, D., Bloom, J. D., Veesler, D., Mateo, D., Hernando, A., Comas, I., González Candelas, F., SeqCOVID-SPAIN

consortium, Stadler, T., & Neher, R. A. (2021). Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *MedRxiv: The Preprint Server for Health Sciences* :1-18. https://doi.org/10.1101/2020.10.25.20219063

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223): 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Huang, Y., Yang, C., Xu, X., Xu, W., & Liu, S. (2020). Structural and functional properties of SARS-CoV-2 spike protein: Potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica,* 41(9): 1141–1149. https://doi.org/10.1038/s41401-020-0485-4

Kaushal, N., Gupta, Y., Goyal, M., Khaiboullina, S. F., Baranwal, M., & Verma, S. C. (2020). Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA. *Pathogens,* 9(565): 1-16. https://doi.org/10.3390/pathogens9070565

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution,* 35(6): 1547–1549. https://doi.org/10.1093/molbev/msy096

McCallum, M., Walls, A. C., Bowen, J. E., Corti, D., & Veesler, D. (2020). Structure-guided covalent stabilization of coronavirus spike glycoprotein trimers in the closed conformation. *Nature Structural & Molecular Biology,* 27(10): 942–949. https://doi.org/10.1038/s41594-020-0483-8

Nei M, & Kumar S. (2000). *Molecular Evolution and Phylogenetics.* Oxford University Press.

Shahhosseini, N., Babuadze, G. (Giorgi), Wong, G., & Kobinger, G. P. (2021). Mutation signatures and in silico docking of novel SARS-CoV-2 variants of concern. *Microorganisms,* 9(926): 1-15. https://doi.org/10.3390/microorganisms9050926

Tortorici, M. A., Beltramello, M., Lempp, F. A., Pinto, D., Dang, H. V., Rosen, L. E., McCallum, M., Bowen, J., Minola, A., Jaconi, S., Zatta, F., De Marco, A., Guarino, B., Bianchi, S., Lauron, E. J., Tucker, H., Zhou, J., Peter, A., Havenar-Daughton, C., … Veesler, D. (2020). Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms. *Science*, 370(6519): 950–957. https://doi.org/10.1126/science.abe3354

Wang, M. Y., Zhao, R., Gao, L. J., Gao, X. F., Wang, D. P., & Cao, J. M. (2020). SARS-CoV-2: Structure, biology, and structure-based therapeutics development. *Frontiers in Cellular and Infection Microbiology,* 10(November):1–17. https://doi.org/10.3389/fcimb.2020.587269

Wong, R. S. Y. (2020). The SARS-CoV-2 uutbreak: An epidemiological and clinical perspective. *SN Comprehensive Clinical Medicine,* 2(11):1983–1991. https://doi.org/10.1007/s42399-020-00546-z

World Health Organization. (2020). Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Accessed on 2 May 2021

World Health Organization. (2021). WHO Coronavirus (COVID-19) Dashboard. https://data.who.int/dashboards/covid19/cases. Accessed on 2 June 2021

Zhang, B., Zhou, X., Qiu, Y., Song, Y., Feng, F., Feng, J., Song, Q., Jia, Q., & Wang, J. (2020). Clinical characteristics of 82 cases of death from COVID-19. *PLoS ONE*, 15(7 July): 1–13. https://doi.org/10.1371/journal.pone.0235458

Zhang, H., Penninger, J. M., Li, Y., Zhong, N., & Slutsky, A. S. (2020). Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Medicine,* 46(4): 586–590. https://doi.org/10.1007/s00134-020-05985-9

Zhu, C., He, G., Yin, Q., Zeng, L., Ye, X., Shi, Y., & Xu, W. (2021). Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. *Journal of Medical Virology,* 93:5729-574. https://doi.org/10.1002/jmv.27132