School of Biosciences

A Computational study of clades and drug adaptation to remdesivir in SARS-CoV-2

Naoki Ueda

2021

Declaration

No part of this thesis has been submitted in support of an application for any degree or qualification of the University of Kent, or any other University or institute of learning.

Table of Contents

Acknowledgements

I would like to thank all the help and support from my two supervisors Professor Mark Wass and Professor Martin Michaelis. I would like to thank Professor Mark Wass for introducing me to the area of bioinformatics and supported me with all the help I needed. I would like to thank Professor Martin Michaelis for guiding and supporting me in every aspect of scientific research. I would also like to thank the members of the research group for their help and support during the research process.

I would like to acknowledge the collaborators at Goethe University Frankfurt for their constructive research. Without their research and effort, this research could not have been successful.

I would also like to thank all of my friends for their support and for spending a great time with me in this very intense academic year.

At last, I would like to express my deepest appreciation to my family for their understanding and support for this master's course. This accomplishment would not have been possible without them.

Abstract

The first patient of Coronavirus disease 2019 (COVID-19) was detected in December 2019. This infectious disease is caused by a virus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and infecting people globally. Many researchers have been investigating to elucidate the characteristics of this virus from many aspects since the beginning of the outbreak. However, there is still limited information about this virus. In this thesis, we focused on the differences between clades and drug adaptations in SARS-CoV-2.

The second chapter focused on the differences between clades in SARS-CoV-2 by applying the method of differentially conserved positions (DCPs). DCPs can identify the differences between two groups of proteins by only amino acid sequences. We have used nomenclatures introduced by two organizations, GISAID and Nextstrain, to classify into clades. We identified DCPs between clades in SARS-CoV-2, which may reflect the differences of the phenotypes between clades.

The third chapter focused on drug adaptation to remdesivir in SARS-CoV-2 *in vitro*. The collaborator in Frankfurt cultured two strains of the virus until it adapted to remdesivir, and genomic sequencing was performed by Public Health England. By comparing the genomic data between the resistant virus and the control virus, we identified a number of mutations that may be considered as the adaptation factor to remdesivir. We also assessed the effect that may occur on the structure of the proteins by mutations.

The findings in this thesis provided information of DCPs between clades and information about the mutation that may cause an effect to the adaptation to remdesivir for future study of polymerase targeting drugs for SARS-CoV-2.

List of Abbreviations

3CLP	3C-Like Protease
ACE2	Angiotensin-Converting Enzyme 2
BLOSUM	BLOcks SUbstitution Matrix
COVID-19	Coronavirus Disease 2019
DCP	Differentially Conserved Position
ExoN	N-terminal Exonuclease
GISAID	Global Initiative on Sharing Avian Influenza Data
JSD	Jensen-Shannon Divergence
MERS	Middle East Respiratory Syndrome
Mpro	Main protease
MUSCLE	MUltiple Sequence Comparison by Log- Expectation
NGS	Next-Generation Sequencing
NiRAN	Nidovirus RdRp-Associated Nucleotidyltransferase
NSP	Non-Structural Protein
ORF	Open Reading Frame
PANGOLIN	Phylogenetic Assignment of Named Global Outbreak Lineages
PDB	Protein Data Bank
PhyCLIP	Phylogenetic Clustering by Linear Integer Programming
PLP	Papain-Like Protease
RBD	Receptor-Binding Domain
RdRp	RNA dependent RNA polymerase
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
TMPRSS2	Transmembrane Protease Serine 2
WHO	World Health Organisation

List of Figures

Figure in chapter 1

Figure in chapter 2

Figure in chapter 3

Supplementary Figure

List of Tables

Tables in chapter 1

Tables in chapter 2

Tables in chapter 3

Supplementary Table

Chapter 1:

Introduction

This thesis comprises two main research, first identifying differentially conserved positions between clades of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and second the analysis of SARS-CoV-2 adapted to remdesivir.

1.1 SARS-CoV-2

In this thesis, we focused on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 was first detected in December 2019, which is the cause of Coronavirus Disease 2019 (COVID-19) (Gorbalenya et al., 2020; F. Wu et al., 2020). Coronaviruses belong to the family *Coronaviridae* and belong to the order *Nidovirales*. They are divided into four genera called *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus*. Only *Alphacoronaviruses* and *Betacoronaviruses* infect mammalian species, *Gammacoronaviruses* only infect avian species, and *Deltacoronaviruses* infect both mammalian and avian species (F. Li, 2016; Woo et al., 2012). SARS-CoV-2 belongs to the *Betacoronavirus* genus and exhibits 79% genome sequence identity with SARS-CoV (severe acute respiratory syndrome coronavirus) and 50% with MERS (middle east respiratory syndrome) (R. Lu et al., 2020). However, SARS-CoV-2 showed 96% genome sequence identity to bat coronavirus RaTG13 suggesting that bat coronavirus RaTG13 is the closest relative to SARS-CoV-2 (Zhou et al., 2020).

1.1.1 SARS-CoV-2 outbreak

In December 2019, a novel coronavirus named severe acute respiratory syndrome

coronavirus 2 (SARS-CoV-2) was detected and caused an outbreak of unknown viral pneumonia (F. Wu et al., 2020; J. T. Wu et al., 2020). Three patients were admitted to a hospital in Wuhan on 27th December 2019 with severe pneumonia (Zhu et al., 2020). By 2nd January 2020, 41 patients had been admitted to a hospital, and 27 patients had been directly exposed to the Huanan seafood market, which was initially suggested as the source of the outbreak (C. Huang et al., 2020; Q. Li et al., 2020). The virus has spread all over the world and caused a pandemic. As of 15thJuly 2021, the total confirmed cases are 187,827,660 and 4,055,497 deaths globally (www.who.int).

1.1.1.1 Clinical features of COVID-19

The most common symptoms for COVID-19 are fever, fatigue, dry cough (Guan et al., 2020; Hu et al., 2020; C. Huang et al., 2020; Z. Wu & McGoogan, 2020). Olfactory and gustatory disorders are also reported as a symptom of COVID-19 (Giacomelli et al., 2020; Luers et al., 2020). Notably, most young people tend to have mild disease or are asymptomatic, whereas older people have a higher possibility of having a severe disease (Guan et al., 2020; X. Lu et al., 2020; D. Wang et al., 2020). In addition, the incubation period for COVID-19 is typically 5-6 days on average but can be up to 14 days (www.who.int).

1.1.1.2 Diagnostics of COVID-19

Introducing methods to diagnose COVID-19 was urgent in the initial phase of the outbreak. Six technologies are used to diagnose COVID-19; Reverse transcription-quantitative polymerase chain reaction (RT-qPCR), Next-generation sequencing (NGS), isothermal nucleic acid amplification assays, CRISPR-mediated detection, antigen tests, and serological tests (Z. Huang et al., 2020; Mercer & Salit, 2021; Vandenberg et al., 2021). The first case in Wuhan, China, was detected by NGS,

whereas RT-qPCR is mainly used to detect most SARS-CoV-2 infections (Vandenberg et al., 2021). Although antigen tests have lower specificity and sensitivity, these tests are commonly used for mass testing (Mak et al., 2020; Mercer & Salit, 2021; Vandenberg et al., 2021). RT-qPCR is used to detect gene fragments of SARS-CoV-2, whereas antigen tests detect proteins of the virus.

1.1.1.3 Vaccines for SARS-CoV-2

Developing vaccines is an effective strategy to end the COVID-19 pandemic. Many pharmaceutical companies such as Pfizer/BioNTech and Moderna have developed vaccines for COVID-19. As of 20th July 2021, 3.69 billion doses have been administered globally (Mathieu et al., 2021). On 2nd of December 2020, the vaccine developed by Pfizer/BioNTech was approved in the UK, which was the first vaccine approved globally for COVID-19 that has been in large clinical trials (www.gov.uk; Ledford, Cyranoski, & Van Noorden, 2020). Some pharmaceutical companies such as Pfizer/BioNTech and Moderna have used mRNA as a vaccine, which was the first time in history that this has been done (Baden et al., 2020; Pardi et al., 2018; Polack et al., 2020). On the other hand, other companies use different technologies such as DNA, protein, and inactivated viruses (Rawat et al., 2021). Since the receptor-binding domain (RBD) of the spike protein binds to Angiotensin-converting enzyme 2 (ACE2), the spike protein is the target for most vaccines (Dai & Gao, 2020).

1.1.1.4 Antiviral drugs

As of 20th July 2021, there are three recommendations for therapeutics published by WHO, which are IL-6 receptor blockers, remdesivir, and corticosteroids (Rochwerg et al., 2021). However, WHO does not recommend using remdesivir in addition to usual care. Additionally, other antiviral drugs such as favipiravir and camostat are thought

to be potential drugs for treatment (Asselah et al., 2021). The potential targets of SARS-CoV-2 for antiviral drugs are the main 3C-like protease (3CLP), RNA-dependant RNA polymerase (RdRp), helicase, and the papain-like protease (PLP) (Mohamed et al., 2021).

1.1.2 Genes and proteins of SARS-CoV-2

SARS-CoV-2 is a positive single-stranded RNA virus with a genome length of ~ 29.9 kb (Senior et al., 2020). The genes of SARS-CoV-2 are ORF1a, ORF1b, spike, ORF3a, ORF3b, envelope, membrane, ORF6, ORF7a, ORF7b, ORF8, nucleocapsid, ORF9b, ORF9c, and ORF10 (Davidson et al., 2020; F. Wu et al., 2020; Zhou et al., 2020). ORF1a is translated into a polyprotein (pp) 1a, which is comprised of NSP1~NSP10, and ORF1a and ORF1b together produces pp1ab containing NSP1~NSP16 (Rohaim et al., 2021). (-1) ribosomal frameshift enables the production of pp1ab by overreading the stop codon of ORF1a (Brierley et al., 1989). Cleavage of pp1a and pp1ab is facilitated by NSP3 and NSP5 (V'kovski et al., 2020). Additionally, other genes encode spike protein, NS3a, NS3b, envelope protein, membrane protein, NS6, NS7a, NS7b, NS8, nucleocapsid protein, NS9b, and NS9c. However, there is no evidence that ORF10 is expressed or plays any role.

1.1.2.1 Structure of SARS-CoV-2

SARS-CoV-2 comprises four structural proteins: envelope protein, membrane protein, nucleocapsid protein, and spike protein (Figure 1.1). The virus has enveloped virions of 50-200 nm diameters with the spike protein protruding from the surface, giving the virus a crown-look appearance (Chen et al., 2020).

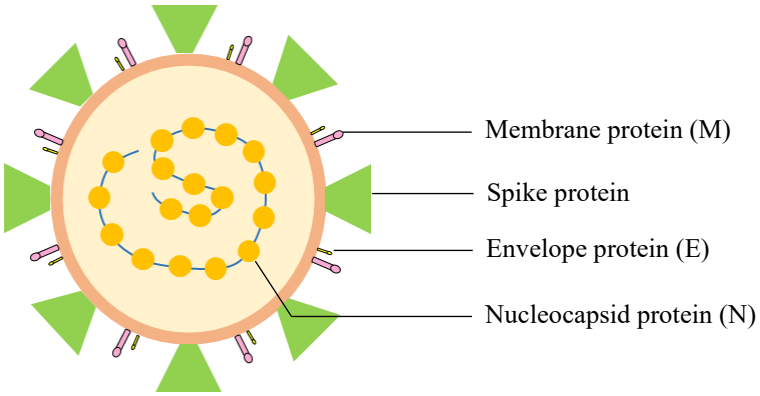


Figure 1.1. Structure of SAR-CoV-2. The figure is reproduced from W. Zhou & Wang, 2021.

1.1.2.2 Structural proteins

As mentioned above, SARS-CoV-2 is constructed of four proteins called structural proteins.

The envelope protein is the smallest structural protein consisting of just 75 amino acids, and its genetic sequence is 98.8% conserved in SARS-CoV-2 strains (Rahman et al., 2021). The envelope protein has three domains, which are the hydrophilic amino terminus, the hydrophobic transmembrane domain, and the hydrophilic carboxyl-terminus domain (Schoeman & Fielding, 2019). The functions of the envelope protein are assembly, budding, envelope formation, and pathogenesis (Rahman et al., 2021).

The membrane protein is the essential protein for assembling, and it interacts with every other structural protein (Schoeman & Fielding, 2019). Interaction with the spike protein is necessary for retention of the spike protein in the ER-Golgi intermediate compartment (ERGIC)/Golgi complex and to incorporate into new virions. By binding with the nucleocapsid protein, it stabilises the nucleocapsid protein and promotes viral assembly. Interaction with the envelope protein is necessary for producing and releasing the virus (Schoeman & Fielding, 2019).

The nucleocapsid protein binds to the SAR-CoV-2 RNA genome. This protein is essential for viral replication/transcription, assembly, and forming helical ribonucleoproteins during the packaging of the viral RNA genome (Kang et al., 2020; Savastano et al., 2020). The Nucleocapsid protein has five domains which are: the intrinsically disordered N-terminal domain (NTD), the RNA-binding domain (RBD), the intrinsically disordered central linker (LINK), a dimerisation domain, and a disordered C-terminal domain (CTD) (Cubuk et al., 2021).

The Spike protein is the most researched protein, as it has a pivotal role in entering the host cell. The amino acid length is 1273 and consists of a signal peptide, the S1 subunit, and the S2 subunit. The S1 subunit consists of an N-terminal domain (NTD) and a receptor-binding domain (RBD), whereas the S2 subunit consists of the fusion peptide (FP), heptapeptide repeat sequence 1 (HR1), HR2, TM domain, and a cytoplasmic domain. S1 binds via the RBD to ACE2, while S2 plays a role in fusing with the host cell membrane (Y. Huang et al., 2020). After the RBD in S1 binds to ACE2, the host proteases cleave the spike protein into the two subunits, S1 and S2. Host cell proteases, such as Transmembrane protease serine 2 (TMPRSS2) and trypsin, are essential to catalysing the cleavage of the spike protein (Y. Huang et al., 2020). Recent studies show that the specific furin cleavage site is located only on the cleavage site of SARS-CoV-2 but not in SARS-CoV (Coutard et al., 2020; Xia et al., 2020). In addition, many mutations have been observed in the spike protein, such as D614G, N501Y, and E484K (*Investigation of Novel SARS-CoV-2 Variant Variant of Concern 202012/01 Technical Briefing 5*, n.d.; Isabel et al., 2020; Santos & Passos, n.d.). Since the spike protein has been the target for most vaccines, mutations in this protein need to be tracked carefully

(Dai & Gao, 2020).

1.1.2.3 Non-structural proteins (NSP)

Proteins encoded on ORF1a and ORF1b are called Non-structural proteins (NSP). These proteins mainly have a role in replicating the virus. ORF1a and ORF1b comprise approximately two-third of the RNA genome of the SARS-CoV-2. Translation of ORF1a and ORF1b in the genomic RNA generates polyproteins pp1a and pp1ab, which contains NSP1-NSP11 and NSP1-NSP16, respectively (Rohaim et al., 2021; V'kovski et al., 2020). The study of ribosome profiling shows that frameshift efficiency between ORF1a and ORF1b is 45% to 70%, which means pp1a is expressed 1.4-2.2 times more than pp1ab (Irigoyen et al., 2016; V'kovski et al., 2020).

NSP1 has a role in blocking innate immune responses of the host by blocking the expression of type I and III interferons and of other host proteins. The function of NSP2 is not known yet. However, it has been reported that eliminating the cleavage site between NSP1 and NSP2 produce infectious virus in mouse hepatitis virus (MHV) (Denison et al., 2004). NSP3 is the largest multi-domain protein in SARS-CoV-2. This protein plays various roles, such as cleaving the viral polyproteins, blocking the host innate immune response and acting as a scaffold protein to bind to other NSPs or host proteins (Imbert et al., 2008). NSP4 is a protein to form a complex with NSP3 and NSP6 that rearrange endoplasmic reticulum into double-membrane vesicles (Angelini et al., 2013; Mariano et al., 2020). NSP5 is regarded as the main protease (Mpro), responsible for cleavage within polyprotein 1a/1ab at 11 sequence-specific sites (Hilgenfeld, 2014). NSP6 is predicted to have seven transmembrane regions. However, only six of them function as membrane-spanning helices (Krogh et al., 2001). NSP7

and NSP8 work as a cofactor of NSP12, which is the main protein of RdRp (Gao et al., 2020)(Hillen et al., 2020). NSP9 is the second replicase cleavage protein after NSP5 (Egloff et al., 2004). NSP10 interacts with NSP14 and NSP16 and stimulates the activity of N-terminal exonuclease (ExoN) and 2′-O-methyltransferase (2′-O-MTase) (Bouvet et al., 2012; Rohaim et al., 2021). As mentioned before, NSP12 is the main protein of RdRp and consists of two domains, the nidovirus RdRp-associated nucleotidyltransferase (NiRAN) and the canonical RdRp domain C-terminal (Lehmann et al., 2015). NSP13 functions as an RNA 5′- triphosphatase activity (V'kovski et al., 2020). NSP14 has two main functions as an N7- methyltransferase (N7 MTase) and an N-terminal exonuclease (ExoN). The N7 MTase function has a role in adding 5' cap structures to RNA, and ExoN has a role in proofreading the viral genome (Rohaim et al., 2021). In addition, NSP15 and NSP16 work as an endoribonuclease and ribose 2′-O-methyltransferase, respectively (V'kovski et al., 2020).

1.1.2.4 Accessory proteins

Proteins other than pp1a, pp1ab, and structural proteins are called accessory proteins. At least five of the accessory proteins, which are ORF3a, ORF6, ORF7a, ORF7b and ORF8, have been reported for SARS-CoV-2 (Davidson et al., 2020; Kim et al., 2020). In addition, ORF3b, ORF9b, ORF9c, and ORF10 have been reported in some studies (Davidson et al., 2020; F. Wu et al., 2020; Zhou et al., 2020). However, not all of these have not been experimentally identified, and the exact number of accessory proteins are still debated (Davidson et al., 2020; Kim et al., 2020; V'kovski et al., 2020).

1.1.3 The cycle of SARS-CoV-2 infection

The initial step of SARS-CoV-2 infection is to enter the human cells. Angiotensin-

converting enzyme 2 (ACE2) is identified as the host cell's receptor for SARS-CoV-2 entry (Hoffmann et al., 2020; Zhou et al., 2020). By fusing the spike protein to the host cell membrane, the virus releases the RNA genome into the cell cytoplasm. After entering the host cell, the two large open reading frames of the genomic RNA, ORF1a and ORF1b, are translated into polyproteins, pp1a and pp1ab. These proteins are processed into the individual NSPs, which form the viral replication and transcription complex. The proteins replicate their genomic RNA to produce copies of full-length RNA, which leads to reproducing the new virus. RNA synthesis is performed by the RNA dependant RNA polymerase (RdRp), which comprises NSP12 as a primary protein and NSP7 and NSP8 as cofactors. Moreover, NSP14 assists RNA synthesis by providing the proofreading function.

The assembly and budding of structural proteins have not yet been assessed in SARS-CoV-2. In general, the structural proteins of coronavirus are suggested to be assembled at the endoplasmic reticulum to the Golgi compartment and shed from the infected cell by exocytosis (V'kovski et al., 2020). However, a recent study shows that betacoronaviruses are instead released from infected cells via the lysosomal trafficking pathway (Ghosh et al., 2020). The study shows that viral interference with lysosomal acidification, enzyme activity, and antigen presentation has occurred during this process.

1.1.4 The nomenclature systems for SARS-CoV-2

There are several nomenclatures to classify variants and clades within SARS-CoV-2 (Table 1.1). Global Initiative on Sharing Avian Influenza Data (GISAID), Nextstrain, World Health Organisation (WHO), and Phylogenetic Assignment of Named Global

Outbreak Lineages (PANGOLIN) software team has introduced the nomenclatures. As of 11th August 2021, GISAID and Nextstrain classify the virus into ten clades and twenty clades, respectively (www.gisaid.org; www.nextstrain.org). PANGOLIN software team has introduced a nomenclature of having a more detailed classification such as B.1.1.7 and B.1351. The original strain is classed as the A strain in this nomenclature (Rambaut et al., 2020). As of 22nd July 2021, WHO has also labelled the virus as Variants of Concern and Variants of Interest using Greek alphabet letters (www.who.int).

Table 1.1. Corresponding nomenclature of SARS-CoV-2. (www.nextstrain.org; www.gisaid.org, www.who.int; https://covariants.org/)

Nextstrain	GISAID	PANGOLIN	WHO
19A	L, V	B	
19B	S	A	
20A	G	B.1	
20B	GR	B.1.1	
20C	GH	B.1	
20D	GR	C	
20E	GV	B.1.177	
20F	GR	D.2	
20G	GH	B.1.2	
20H	GH	B.1.351	Beta
20I	GRY	B.1.1.7	Alpha
20J	GR	P.1	Gamma
21A	GK	B.1.617.2	Delta
21B	G	B.1.617.1	Kappa
21C	GH	B.1.427, B.1.429	Epsilon
21D	G	B.1.525	Eta
21E	GR	P.3	Theta
21F	GH	B.1.526	Iota
21G	GR	C.37	Lamda
21H	GH	B.1.621	Mu

1.1.4.1 GISAID clade

As of 11th August 2021, the ten divisions for GISAID clades are L, S, V, G, GH, GR, GV, GRY, GK and sequences not included in these clades are O. Clade L is considered as an early clade marker in WIV04-reference sequence. These letters stand for the significant marker mutations. For example, clade S has a substitution of L84S in NS8, which leads the clade naming S. The nomenclature for GISAID clade is defined by using Phylogenetic Clustering by Linear Integer Programming (PhyCLIP) and merged small lineages to larger clades shared with marker mutations (Han et al., 2019; Maurer-Stroh et al., 2020). Marker mutations are described in Table 1.2, and the simplified phylogenetic tree is described in Figure 1.2.

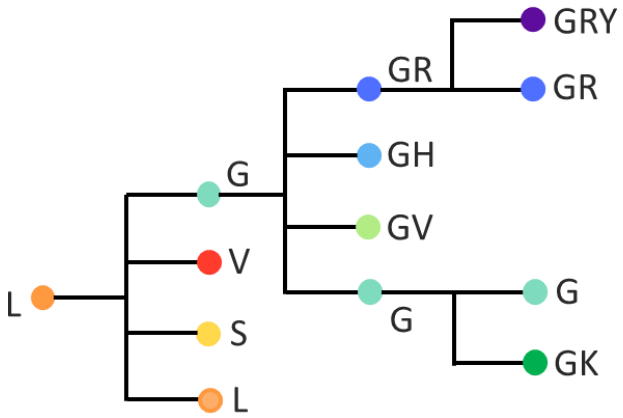


Figure 1.2. Simplified phylogenetic tree by using GISAID clade nomenclature. This figure describes which clade is derived from which clade. This figure does not describe the genetic similarity and timelines.

Table 1.2. Marker mutations of GISAID clade.

Clade	Marker mutation	Protein	Amino acid
L	WIV04-reference (ID: EPI_ISL_402124).		
S	C8782T	NSP4(ORF1a)	-
S	T28144C	NS8(ORF8)	L84S

V	G11083T	NSP6(ORF1a)	L37F
V	G26144T	NS3(ORF3a)	G251V
G	C241T	-	-
G	C3037T	NSP3(ORF1a)	-
G	A23403G	Spike	D614G
GH	C241T	-	-
GH	C3037T	NSP3(ORF1a)	-
GH	A23403G	Spike	D614G
GH	G25563T	NS3(ORF3a)	Q57H
GR	C241T	-	-
GR	C3037T	NSP3(ORF1a)	-
GR	A23403G	Spike	D614G
GR	G28882A	N	G204R
GV	C241T	-	-
GV	C3037T	NSP3(ORF1a)	-
GV	A23403G	Spike	D614G
GV	C22227T	Spike	A222V
GRY	C241T	-	-
GRY	C3037T	NSP3(ORF1a)	-
GRY	21765-21770del	Spike	H69-V70del
GRY	21991-21993del	Spike	Y144del
GRY	A23063T	Spike	N501Y
GRY	A23403G	Spike	D614G
GRY	G28882A	N	G204R
GK	C241T	-	-
GK	C3037T	NSP3(ORF1a)	-
GK	A23403G	Spike	D614G
GK	C22995A	Spike	T478K

1.1.4.2 Nextstrain clade

As of 11th August 2021, clades defined by Nextstrain are 19A, 19B, 20A, 20B, 20C, 20D, 20E, 20F, 20G, 20H, 20I, 20J, 21A, 21B, 21C, 21D, 21E, 21F, 21G, and 21H (www.nextstrain.org). The strategy of the nomenclature is based on the year defined. For example, 19A and 20A are the clade first defined in 2019 and 2020, respectively.

Nextstrain defines a clade as a major clade by three criteria. First, when a clade reaches 20% or more globally for more than two months. Second, when a clade reaches 30% or more in a region for more than two months. Third, when a variant of concern is recognised (Bedford et al., 2021). Marker mutations introduced by Nextstrain are described in Table 1.3. However, other sources show other definitions for 20H, 20I, and 20J, previously introduced as 20H/501.V2, 20I/501Y.V1, and 20J/501Y.V3, respectively (Chand et al., 2020; Faria et al., 2021; Tegally et al., 2020; www.cdc.gov). 19A is considered as the root clade in the Nextstrain nomenclature. The simplified phylogenetic tree of Nextstrain clade nomenclature is shown in Figure 1.3.

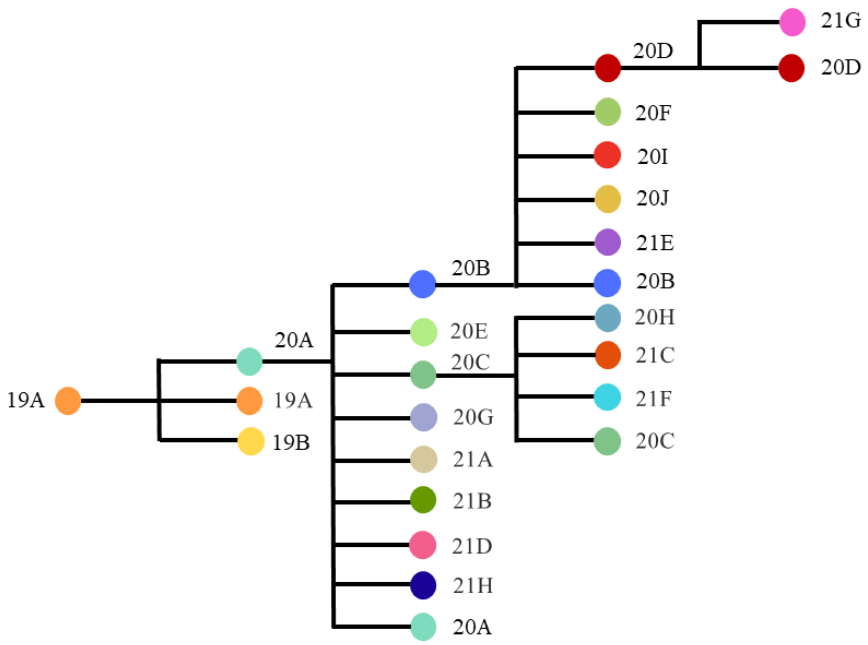


Figure 1.3. Simplified phylogenetic tree by using Nextstrain clade nomenclature. This figure describes which clade is derived from which clade. This figure does not describe the genetic similarity and timelines.

Table 1.3. Marker mutations of Nextstrain clade.

Clade	Marker mutation	Protein	Amino acid
19A	Root clade		
19B	C8782T	NSP4 (ORF1a)	-

19B	T28144C	NS8 (ORF8)	L84S
20A	C14408T	NSP12 (ORF1b)	P323L (314)
20A	A23403G	Spike	D614G
20B	C14408T	NSP12 (ORF1b)	P323L (314)
20B	A23403G	Spike	D614G
20B	G28881A	N and NS9c(ORF14)	R203K and G50N
20B	G28882A	N and NS9c(ORF14)	R203K and G50N
20B	G28883C	N and NS9c(ORF14)	G204R and G50N
20C	C1059T	NSP2 (ORF1a)	T85I (265)
20C	C14408T	NSP12 (ORF1b)	P323L (314)
20C	A23403G	Spike	D614G
20C	G25563T	NS3 (ORF3a)	Q57H
20D	C4002T	NSP3 (ORF1a)	T428I (1246)
20D	G10097A	NSP5 (ORF1a)	G15S (3278)
20D	C13536T	NSP12 (ORF1b)	-
20D	C23731T	Spike	-
20E	C14408T	NSP12 (ORF1b)	P323L (314)
20E	A23403G	Spike	D614G
20E	C22227T	Spike	A222V
20E	C28932T	N and NS9c(ORF14)	A220V and L67F
20E	G29645T	ORF10	V30L
20F	A1163T	NSP2 (ORF1a)	I120F (300)
20F	T7540C	NSP3 (ORF1a)	-
20F	G16647T	NSP13 (ORF1b)	-
20F	C18555T	NSP14 (ORF1b)	-
20F	G22992A	Spike	S477N
20F	G23401A	Spike	-
20G	C10319T	NSP5 (ORF1a)	L89F (3352)
20G	A18424G	NSP14 (ORF1b)	N129D (1653)
20G	C21304T	NSP16 (ORF1b)	R216C (2613)
20G	G25907T	NS3 (ORF3a)	G172V
20G	C27964T	NS8 (ORF8)	S24L
20G	C28472T	N	P67S
20G	C28869T	N and NS9c(ORF14)	P199L and stop
20H (Beta)	C1059T	NSP2 (ORF1a)	T85I (265)
20H (Beta)	C14408T	NSP12 (ORF1b)	P323L (314)

20H (Beta)	A23403G	Spike	D614G
20H (Beta)	G25563T	NS3 (ORF3a)	Q57H
20H (Beta)	A23063T	Spike	N501Y
20H (Beta)	G23012A	Spike	E484K
20I (Alpha)	C14408T	NSP12 (ORF1b)	P323L (314)
20I (Alpha)	A23403G	Spike	D614G
20I (Alpha)	G28881A	N	R203K
20I (Alpha)	G28882A	N	R203K
20I (Alpha)	A23063T	Spike	N501Y
20I (Alpha)	C14676T	NSP12 (ORF1b)	-
20I (Alpha)	C15279T	NSP12 (ORF1b)	-
20J (Gamma)	T733C	NSP2 (ORF1a)	-
20J (Gamma)	C2749T	NSP3 (ORF1a)	-
20J (Gamma)	C3828T	NSP3 (ORF1a)	S370L (1188)
20J (Gamma)	A5648C	NSP3 (ORF1a)	K977Q (1795)
20J (Gamma)	C12778T	NSP9 (ORF1a)	-
20J (Gamma)	C13860T	NSP12 (ORF1b)	-
21A (Delta)	T22917G	Spike	L452R
21A (Delta)	C22995A	Spike	T478K
21A (Delta)	T27638C	NS7a (ORF7a)	V82A
21A (Delta)	G28881T	N and NS9c(ORF14)	R203M and G50W
21A (Delta)	G29402T	N	D377Y
21B (Kappa)	G17523T	NSP13 (ORF1b)	M429I (1352)
21B (Kappa)	T22917G	Spike	L452R
21B (Kappa)	G23012C	Spike	E484Q
21B (Kappa)	T27638C	NS7a (ORF7a)	V82A
21B (Kappa)	G28881T	N and NS9c(ORF14)	R203M and G50W
21B (Kappa)	G29402T	N	D377Y
21C (Epsilon)	G17014T	NSP13 (ORF1b)	D260Y (1183)
21C (Epsilon)	G21600T	Spike	S13I
21C (Epsilon)	G22018T	Spike	W152C
21C (Epsilon)	T22917G	Spike	L452R
21D (Eta)	C14407T	NSP12 (ORF1b)	P323F (314)
21D (Eta)	A21717G	Spike	Q52R
21D (Eta)	T24224C	Spike	F888L
21D (Eta)	C24748T	Spike	-

21E (Theta)	C12049T	NSP7 (ORF1a)	-
21E (Theta)	T23341C	Spike	-
21E (Theta)	C23604A	Spike	P681H
21E (Theta)	T24187A	Spike	-
21E (Theta)	G24836A	Spike	E1092K
21F (Iota)	A16500C	NSP13 (ORF1b)	Q88H (1011)
21F (Iota)	A20262G	NSP15 (ORF1b)	-
21F (Iota)	C21575T	Spike	L5F
21F (Iota)	A22320G	Spike	D253G
21G (Lambda)	G21786T	Spike	G75V
21G (Lambda)	C21789T	Spike	T76I
21G (Lambda)	T22917A	Spike	L452Q
21G (Lambda)	T23031C	Spike	F490S
21H	A11451G	NSP6 (ORF1a)	Q160R (3729)
21H	A13057T	NSP10 (ORF1a)	-
21H	C17491T	NSP13 (ORF1b)	P419S (1342)
21H	T21995A	Spike	Y145N
21H	A21993C	Spike	Y144S
21H	G22599A	Spike	R346K

1.1.5 Mutations and variants of SARS-CoV-2

Mutations may affect the virus by changing the pathogenicity, infectivity, transmissibility, and antigenicity (Harvey et al., 2021). Since SARS-CoV-2 is an RNA virus, the mutation rate is higher than DNA viruses (Sanjuán et al., 2010). However, SARS-CoV-2 has a proofreading mechanism that results in a lower mutation rate than other RNA viruses, such as influenza A virus and Human immunodeficiency virus type 1 (HIV-1) (Manzanares-Meza & Medina-Contreras, 2020; Sanjuán et al., 2010). Although the mutation rate is low, some significant mutations have been observed throughout the pandemic, which leads to a new variant or clade. As of 11th August 2021, the Delta variant, or clade 21A, is predominating and reported to have higher transmissibility (Campbell et al., 2021). Although studies have reported that vaccines

are still effective, a reduction of neutralization has been observed (C. Liu et al., 2021; J. Liu et al., 2021). Additionally, more mutations are observed in the spike protein than in other proteins (Vilar & Isom, 2020). Many mutations in the spike proteins, such as D614G, N501Y, and E484K, have been reported to affect the binding ability to ACE2 (*Investigation of Novel SARS-CoV-2 Variant Variant of Concern 202012/01 Technical Briefing 5*, n.d.; Isabel et al., 2020; Santos & Passos, n.d.). Since this protein is the target for vaccines, the mutations must be carefully monitored.

1.2 Differentially conserved positions (DCPs)

Amino acid sequences are important for the structure and functions of the protein. Thus, conserved amino acid positions within protein families are the most obvious predictor to find significant residues for functions from sequence information alone (Chagoyen et al., 2016). Moreover, within large protein families, positions that are differently conserved between sub-families are called differentially conserved positions (DCPs or Specificity Determining Positions) (Figure 1.4) and they have traditionally been thought to have a role determining functional specificity (e.g. enzyme-substrate specificity), and they have been shown to be enriched in ligand binding and protein-protein interface sites (Rausell et al., 2010).

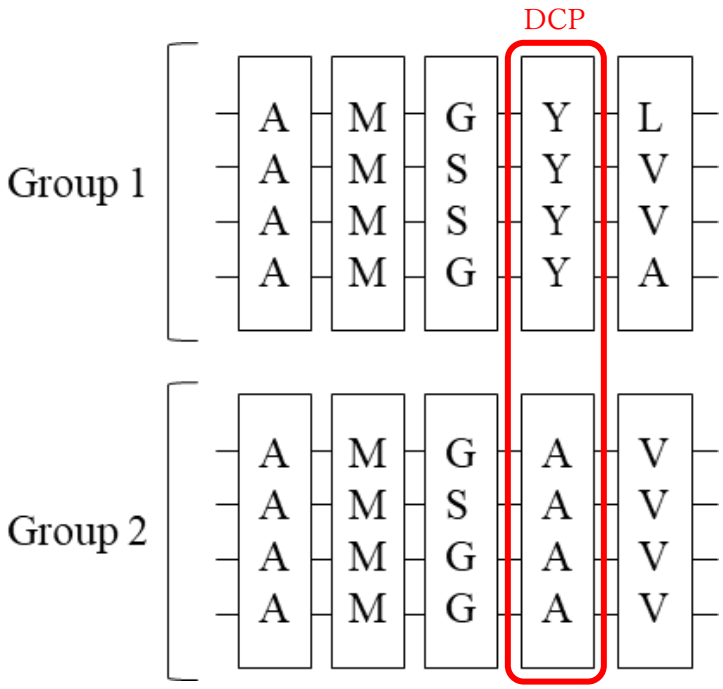


Figure 1.4. Representation of Differentially Conserved Positions (DCPs). In this example, the fourth position is the DCP. Group 1 has tyrosine in the position, which is conserved, and the group has alanine which is also conserved within the group.

More recently, research in the Wass-Michaelis group has applied the identification of DCPs for the analysis of related virus species that exhibit different phenotypes. They have reported that a few mutations could change the pathogenicity of Ebola virus by applying this method (Pappalardo et al., 2016). Moreover, they have identified DCPs between SARS-CoV and SARS-CoV-2 and suggested that the differences in clinical behaviour of these two viruses could be explained by DCPs (Bojkova et al., 2021).

1.3 Jensen-Shannon divergence (JSD)

The Jensen-Shannon divergence is a score that quantifies the similarity between two probability distributions (Capra & Singh, 2007). Unlike the Kullback-Leibler divergence, the JSD is a measure defined to be symmetric. The score of JSD is nonnegative and equal to zero when the probability distribution is the same. If the value of JSD increases, the two probability distributions move away from each other

(Lin, 1991).

1.4 BLOSUM score

BLOSUM matrices, representing Blocks Substitution Matrix, score the likelihood of the substitutions of amino acid. The results are expressed as log-odds, and BLOSUM score B is defined as

$$B_{ij} = \frac{1}{\lambda} \log \frac{P_{ij}}{f_i f_j}$$

where P_{ij} is a probability of finding amino acid i and j aligned in the homologous alignment. f_i and f_j represent the probability of the occurrence of amino acids i and j in the protein. Thus, a positive score is given to more likely substitutions, and a negative score is given to less likely substitutions. BLOSUM matrices are identified based on the minimum percentage identity of the aligned amino acid sequences (Lesk, Arthur M., 2019; Zomaya, 2006).

1.5 Remdesivir

Remdesivir (GS5734) is a prodrug, and the active metabolite (GS-441524) works as a nucleotide analogue in replicating RNA. Remdesivir was originally developed by Gilead Sciences and emerged from a collaboration project of Gilead Sciences, the U.S. Centers for Disease Control and Prevention (CDC), and the US Army Medical Research Institute of Infectious Diseases (USAMRIID) (Eastman et al., 2020; www.gilead.com). They were aiming to develop drugs to target RNA-based viruses, which had the potential of causing a global pandemic. GS-441524 and its S-acyl-2-thioethyl monophosphate prodrug were reported in 2012 to show broad activity against various RNA viruses such as yellow fever virus (YFV), Dengue virus type 2 (DENV-2), influenza A, parainfluenza 3, and SARS-CoV (Cho et al., 2012). When the Ebola

virus outbreak occurred in 2014, a study showed that remdesivir suppressed the replication of the Ebola virus in rhesus monkeys (Warren et al., 2016a). In addition, antiviral activity against coronaviruses was validated both *in vitro* and *in vivo* (Eastman et al., 2020; Jordan et al., 2017).

Remdesivir is a polymerase inhibitor targeted to inhibit the viral RNA dependent RNA polymerase (RdRp – NSP12 in SARS-CoV-2) and thus inhibiting replication of the viral genome. Remdesivir is converted to remdesivir triphosphate (RTP) in the cell. The structure of RTP is similar to adenine and inhibits the RNA synthesis by the specific mechanism of delayed chain termination. The study shows that RTP forms a phosphodiester bond with the nucleotide next to it but terminates the synthesis of three nucleotides downstream. RTP inhibits the synthesis at position four nucleotide downstream because a steric clash is formed between the 1′-CN substituent of the incorporated remdesivir and residue Ser-861 of the RdRp (Gordon et al., 2020; Saha et al., 2020). Hence, it decreases the production of the viral RNA. In addition, no significant inhibition has been confirmed to human RNA Pol II and human mitochondrial RNA polymerase for the use of remdesivir *in vitro* (Warren et al., 2016b). In the clinical trial against Ebola Virus Disease, the result did not show benefits despite the success in animal models (Mulangu et al., 2019; Warren et al., 2016a).

1.5.1 Use of remdesivir in treatment for COVID-19

Remdesivir has shown effectiveness against SARS-CoV-2 *in vitro* (M. Wang et al., 2020). In addition, the results of a clinical trial conducted by the National Institute of Allergy and Infectious Diseases (NIAID) has demonstrated that remdesivir can speed up the recovery from COVID-19 (Beigel et al., 2020). However, the Solidary clinical

trial conducted by WHO does not show differences in mortality and the recovery speed using remdesivir ("Repurposed Antiviral Drugs for Covid-19 — Interim WHO Solidarity Trial Results," 2021). Various clinical trials have been conducted, and most of them did not show significant differences (CD et al., 2020; JD et al., 2020; Y. Wang et al., 2020). Therefore, the effect of remdesivir for use in treatment to COVID-19 is still uncertain. Remdesivir has been authorised for temporary use in approximately 50 countries globally (www.gilead.com). In the European Union, remdesivir was approved for treatment for patients with severe disease due to COVID-19 (www.ema.europa.eu). The US Food and Drug Administration has approved remdesvir for use in adults and pediatric patients who need hospitalisation (www.fda.gov).

1.6 The organisation of this thesis

This thesis comprises of two main research: Differentially conserved positions (DCPs) between clades of SARS-CoV-2, Drug adaptation to remdesivir of SARS-CoV-2.

This thesis is divided into the following four chapters:

Chapter 1 is an introduction part and describes information about SARS-CoV-2, remdesivir, and the methods to analyse the Differentially conserved positions.

Chapter 2 describes the analysis of DCPs between clades in SARS-CoV-2.

Chapter 3 introduces the analysis of drug adaptation to remdesivir in SARS-CoV-2

Chapter 4 is a discussion chapter.

Chapter 2:

Differentially conserved positions between clades in SARS-CoV-2

2.1 Introduction

Studies have reported that some variants (or clades) in SARS-CoV-2 have different transmissibility and reduce neutralization (Souza et al., 2021; P. Wang et al., 2021). For example, 20I/501Y.V1 (Nextstrain nomenclature) or the Alpha variant (WHO label) have higher transmissibility than earlier clades (Davies et al., 2021). Differences in characteristics of organisms can be explained by analysing proteins. Amino acid sequences are important for the structure and the function of a protein. Thus, conserved positions within protein families are considered functionally significant residues. In addition, positions that are differentially conserved, called Differentially conserved positions (DCPs), within sub-families tend to determine functional specificity (Rausell et al., 2010). Here, we applied the method of identifying DCPs to investigate which protein and amino acid could be the potential factor of the differences of characteristics between clades in SARS-CoV-2.

2.2 Methods

2.2.1 Data

Amino acid sequences for 429,535 SARS-CoV-2 were downloaded from Global Initiative on Sharing Avian Influenza Data (GISAID). The metadata was also downloaded from GISAID (www.gisaid.org). These data were downloaded on 31st January 2021. Thus, we used the nomenclature which was proposed at that time. As of 31st January 2021, GISAID and Nextstrain classified the virus into eight clades and

twelve clades, respectively (Bedford et al., 2021; Maurer-Stroh et al., 2020). GISAID introduced clades defined as L, S, V, G, GH, GR, GV, and O (Figure 2.1) (Maurer-Stroh, 2020). Clades defined by Nextstrain were 19A, 19B, 20A, 20B, 20C, 20D, 20E, 20F, 20G, 20H/501Y.V2, 20I/501Y.V1, and 20J/501Y.V3 (Figure 2.2) (Bedford et al., 2021).

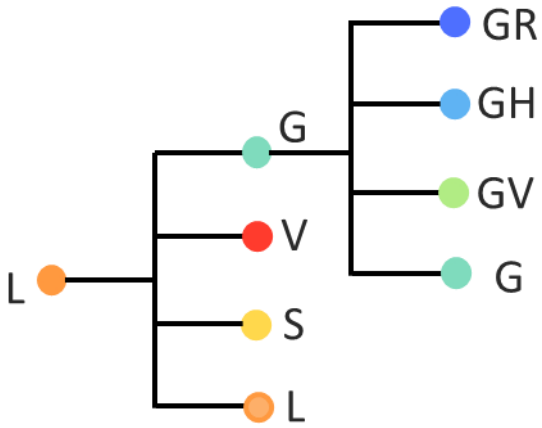


Figure 2.1. Simplified phylogenetic tree by using GISAID clade nomenclature. This is the tree as of 31st January 2021.

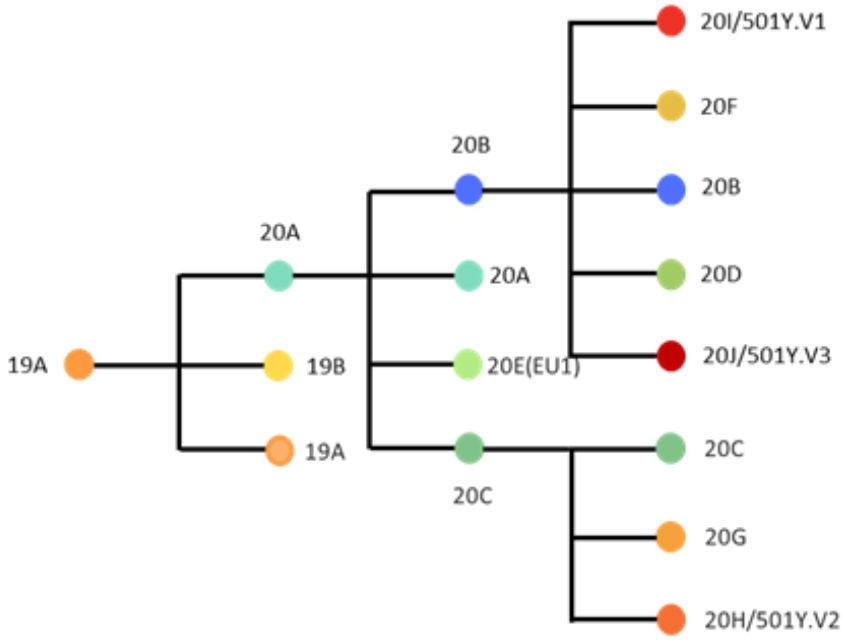


Figure 2.2. Simplified phylogenetic tree by using Nextstrain clade nomenclature. This is the tree as of 31st January 2021.

2.2.2 Creating FASTA format file

Genomes with lengths less than 29000 bases were excluded as they were incomplete. The remaining sequences were divided into clade groups. We used the clade nomenclatures of both GISAID and Nextstrain. GISAID clades are L, S, V, G, GH, GR, GV, and O. Nextstrain clades are 19A, 19B, 20A, 20B, 20C, 20D, 20E, 20F, 20G, 20H/501Y.V2, 20I/501Y.V1, and 20J/501Y.V3.

The genomes were translated into their respective proteins, resulting in sequences for the 27 proteins encoded in the genome (data divided into 27 FASTA format files for each clade).

2.2.3 Multiple sequencing alignment and identification of differentially conserved positions

We compared the amino acid sequence between each clade. Twenty-eight combinations of comparison were conducted in the GISAID nomenclature, and 66 combinations were undertaken in the Nextstrain nomenclature. Multiple sequencing alignments were generated using MUltiple Sequence Comparison by Log- Expectation (MUSCLE) (Edgar, 2004) with default settings. Sequences that contain 'X' were excluded from the alignment. Differentially conserved positions (DCPs) were identified by calculating the Jensen Shannon divergence score for each position in the multiple sequence alignment. The positions with conservation scores higher than 0.8 for each group (and containing different amino acids) were considered to be differentially conserved positions (DCPs). This was equivalent to the approach used previously (Martell et al., 2019).

2.2.4 Comparison of amino acid frequency in each position

From the multiple sequence alignment, the frequency of each amino acid was counted

for each position in the sequence. The count of each amino acid was converted to the ratio of amino acid in each position. Subsequently, the positions that the most frequent amino acids are different between clades were extracted.

2.3 Results

2.3.1 Data

The 429,535 sequences comprise 5117 L clade, 8892 S clade, 6145 V clade, 6085 O clade, 66065 G clade, 91240 GH clade, 150121 GR clade, and 95526 GV clade in GISAID nomenclature. Moreover, the 429,535 sequences are also classified into 16122 19A clade, 8974 19B clade, 95059 20A clade, 97259 20B clade, 46628 20C clade, 4942 20D clade, 97280 20E clade, 12614 20F clade, 13402 20G clade, 715 20H/501Y.V2 clade, 36256 20I/501Y.V1 clade, and 35 20J/501Y.V3 clade in Nextstrain nomenclature.

2.3.2 Identifying differentially conserved positions (DCPs) in GISAID clade

By comparing the amino acid sequences, we have identified the DCPs between GISAID clades (Table 2.1). Three positions were observed, which were L37F in NSP6 found between L and V, G251V in NS3 between L and V, S and V, GR and V, and GV and V, and D614G in the Spike protein between L and G, L and GH, L and GR, and L and GV.

The substitution L37F was only observed between the clades of L and V. This suggests that the substitution of L37F occurred when clade V was derived from clade L. However, L37F in NSP6 is a marker substitution in the GISAID nomenclature. Furthermore, G251V in NS3 was observed in this analysis between L and V, S and V, GR and V, and GV and V. The BLOSUM score for this is -3, indicating that this is a

substitution that rarely occurs during evolution. However, this position is also a marker mutation. In addition, D614G was found between L and G, L and GH, L and GR, L and GV. This position has been widely researched and is widely known as the marker substitution of clade G.

In this analysis, we found 3 DCPs when using the GISAID clade nomenclature. However, all of the positions were marker substitutions used to define the clades and did not define further conserved differences between the clades, which could have a functional significance.

Table 2.1. Differentially conserved positions (DCPs) in GISAID clades. Each row shows the results of a comparison of the protein between clade 1 and clade 2. The conservation score describes how the amino acid is conserved in the clade. The BLOSUM score shows the ratio of having the same amino acid substitution by chance.

Clade1	Clade2	Protein	DCP	Position		Conservation score		BLOSUM
				Clade1	Clade2	Clade1	Clade2	
L	V	NS3	G251V	251	251	0.81	0.82	-3
L	V	NSP6	L37F	37	37	0.82	0.87	0
L	G	Spike	D620G	614	614	0.86	0.82	-1
L	GH	Spike	D617G	614	614	0.86	0.82	-1
L	GR	Spike	D616G	614	614	0.86	0.82	-1
L	GV	Spike	D614G	614	614	0.86	0.82	-1
S	V	NS3	G252V	251	251	0.81	0.82	-3
V	GR	NS3	V255G	251	251	0.82	0.81	-3
V	GV	NS3	V255G	251	251	0.82	0.8	-3

2.3.3 Identifying differentially conserved positions (DCPs) in Nextstrain clade

On the other hand, overall 33 positions in SARS-CoV-2 were DCPs when comparing Nextstrain clades described in Table 2.3. Seventeen DCPs were in the Spike protein, and five DCPs were in the N protein. In addition, NSP2, NSP3, and NS3 had 2 DCPs,

and NSP5, NSP13, NS8, NS9b, and NS9c had one DCP (Table 2.2). By comparing the pairs of clades described in Table 2.3, 157 DCPs were identified in total (Table 2.3). In this result, 121 of these DCPs were not marker mutations, whereas 36 of them were marker mutations that define the different clades. Moreover, 103 DCPs were observed when comparing clade 20J/501Y.V3 and any other clades. Interestingly, non-marker DCPs were only identified when the following clades were compared, 20H/501Y.V2, 20I/501Y.V1 or 20J/501Y.V3 (Table 2.3). The full results are described in Supplementary Table 1.

In the N protein, the DCPs were P80R, R203K, G204R, I205T, and F235, as well as R203K and G204R, which are marker mutations for some clades. In NS3, S253P was the only DCP identified, while E92K was the only DCP in NS8. Q77E and V49L were the only DCPs in NS9b and NS9c, respectively. Between clade 20H/501Y.V2 and clade 20J/501Y.V3, N837K was identified as a DCP in NSP3, the only DCP in this protein.

In the spike protein, L18F, T20N, P26S, D80A, D138Y, R190S, K417N, K417T, S477N, E484K, N501Y, H655Y, A701V, T716I, S982A, T1027I, D1118H, and V1176F were the DCPs identified throughout this analysis. However, S477N in clade 20F, E484K and N501Y in clade 20H/501Y.V2, and N501Y in clade 20I/501Y.V1 are considered as marker substitutions.

Table 2.2. The overall amount of DCPs in each protein in Nextstrain clade (including the comparison of any combination of clades).

Protein	DCPs
Spike	17
N	5
NSP2	2
NSP3	2
NSP5	1
NSP13	1
NS3	2
NS8	1
NS9b	1
NS9c	1
Total	33

Table 2.3. Overall DCPs in Nextstrain clade. Markers are described in the method section.

clade1	clade2	DCPs	DCPs (not marker)	DCPs (marker)
19A	20F	2	0	2
19A	20H	5	3	2
19A	20J	9	8	1
19B	20F	2	0	2
19B	20H	3	3	0
19B	20J	8	7	1
20A	20H	1	1	0
20A	20I	1	0	1
20A	20J	4	3	1
20B	20H	1	1	0
20C	20E	2	0	2
20C	20F	1	0	1
20C	20J	6	5	1
20D	20F	1	0	1
20D	20H	3	2	1
20D	20I	5	3	2
20D	20J	9	8	1
20E	20H	1	1	0

20E	20J	6	6	0
20F	20G	1	0	1
20F	20H	9	3	6
20F	20I	4	2	2
20F	20J	16	13	3
20G	20H	3	2	1
20G	20I	4	4	0
20G	20J	12	11	1
20H	20I	5	5	0
20H	20J	21	19	2
20I	20J	12	11	1
	Total	157	121	36

2.3.4 Differences of frequencies of amino acid in each position in GISAID clade

Although we have found DCPs throughout the analysis, some substitutions that should have been detected as markers, such as Q57H in NS3 between clade GH and other clades, were not detected when we might have expected to observe them. Thus, we decided to look into the frequencies of amino acids in each position because DCPs were only classified at positions where the amino acids present were not shared between the two groups. For example, if at a position one clade predominantly had an alanine, while the other clade had a glycine, this would not be DCP if there was a single sequence in the second clade that contained alanine.

In this analysis, we have found six positions where the frequency of given amino acids differs between the clades but not sufficiently to be a DCP. Table 2.4 shows the protein and position that changed in any of the clades. We excluded the marker mutations which defines the clades. The positions where the frequency of the amino acid differs were found in N, NS9c, NSP2, and NSP12 (Table 2.4). In N protein, R203K and A220V were identified as changes of the most frequent amino acid. For NSP9c, two

changes, G54N and L71F, were observed in this analysis. Moreover, NSP2 and NSP12 had the substitution of T85I and P323L, respectively. Interestingly, there were no differences in the most frequent amino acid between GISAID clade in the spike protein, which has the highest tendency to mutate.

Table 2.4. The positions having differences of the most frequent amino acid in GISAID clade.

Protein	Changes of the most frequent amino acid
N	R203K, A220V
NS9c	G54N, L71F
NSP2	T85I
NSP12	P323L

2.3.5 Differences of frequencies of amino acid in each position in Nextstrain clade

We also looked into the differences of amino acid frequencies in Nextstrain clade. Table 2.5 shows the protein and the position that changed the most frequent amino acid in any of the clades. The differences of the most frequent amino acids were identified at a total of 42 positions in 11 proteins: E, N, NS3, NS8, NS9b, NS9c, NSP3, NSP5, NSP6, NSP13, and the spike protein (Table 2.5). In this analysis, we have excluded the marker mutations described in the method section. 24 out of 42 positions were the positions that were not identified as a differentially conserved position (DCP). Compared with the analysis using GISAID clades, there were more positions and proteins that had the differences between clades.

Interestingly, there were 10 positions that had different amino acid frequencies other than the DCPs in the spike protein by analysing Nextstrain clades, although there were

no differences in GISAID clades.

Table 2.5. The positions having differences of the most frequent amino acid in Nextstrain clade. The positions having a bracket were identified as a DCP in the previous analysis.

Protein	Changes of the most frequent amino acid (Identified as a DCP)
E	P71L
N	D3L, (P80R), (T205I), (S235F)
NS3	S175L, (S258P)
NS8	(E94K)
NS9b	(Q77E)
NS9c	(V53L), L56F
NSP3	T189I, (K843N), A896D, I1422T
NSP5	K90R
NSP6	F38L, S105del, L106del, S107del, G108S, F109L
NSP13	(E341D)
Spike	(L19F), (T21N), (P27S), I70del, H71del, V72I, (D82A), (D140Y), Y146del, (R192S), A227V, L246del, L247del, A248del, (K424N) and (K424T), A570D, (H664Y), P690H, (A710V), (T725I), (S991A), (T1036I), (D1127H), (V1185F)

2.4 Discussion

Here, we conducted research of identifying differentially conserved positions (DCPs) between clades of SARS-CoV-2.

We identified three DCPs by comparing each GISAID clade, which was L37F in NSP6 found between clade L and V, G251V in NS3 found between clades L and V, S and V,

GR and V, and GV and V, and D614G in the Spike protein found between clade L and G, L and GH, L and GR, and L and GV. This result of having a small number of substitutions agrees with previous research that suggests coronaviruses have a low mutation rate (~1.0×10^{-6} per site per cycle) compared with other RNA viruses such as Influenza virus (~3×10^{-5} per site per cycle) because of the proofreading mechanism (Bar-On et al., 2020; Manzanares-Meza & Medina-Contreras, 2020; Sanjuán et al., 2010).

In addition, 33 positions were DCPs in the Nextstrain clade, and the three clades that emerged later was primarily involved in the comparison, which are 20H/501Y.V2, 20I/501Y.V1, and 20J/501Y.V3. Interestingly, non-marker DCPs were only identified when the three clades were compared, 20H/501Y.V2, 20I/501Y.V1 or 20J/501Y.V3. However, according to some other sources, most of these DCPs are also the definition of clades (Chand et al., 2020; Faria et al., 2021; Tegally et al., 2020; www.cdc.gov). Considering these sources, just a few DCPs were not a definition of clades, which were V1176F in the spike protein, S253P in NS3, Q77E in NS9b, and V49L in NS9c. V1176F in the spike protein was identified in the comparison of clade 20J/501Y.V3 and five other clades (Table 2.4). S253P in NS3 was identified between 20H/501Y.V2 and 20J/501Y.V3. Q77E in NS9b was considered a DCP in comparison between 20J/501Y.V3 and six other clades. However, this amino acid is encoded in the same position as P80R of the N protein, which is a mutation that defines clades. Moreover, V49L in NS9c was identified by comparing 20E and 20J/501Y.V3, 20G and 20J/501Y.V3.

The differences in the number of DCPs between the analysis using GISAID clades and

Nextstrain clades can be attributed to the fact that Nextstrain clades have a larger number of clades than GISAID clades. Moreover, another possible reason is that the Nextstrain clade nomenclature includes clades reported as Variants of concern (VOC), 20H/501Y.V2, 20I/501Y.V1, and 20J/501Y.V3. Studies have suggested that clade 20I/501Y.V1, or the Alpha variant using WHO labels, initially appeared in South East England, have higher transmissibility than earlier variants (Davies et al., 2021). Additionally, 20H/501Y.V2, or the Beta variant using WHO labels, is more resistant to neutralisation by convalescent sera and post-vaccinated sera than earlier variants, and 20J/501Y.V3, or the Gamma variant using WHO labels, also reduces neutralisation by antibodies after natural infection and vaccination (Souza et al., 2021; P. Wang et al., 2021). These differences in phenotypes are consistent with the result of having more DCPs compared to earlier clades.

This analysis was conducted within species adapting in an individual outbreak. Thus, the number of DCPs was not as much as previous research, such as the comparison of Ebola virus species (Pappalardo et al., 2016) and comparisons of SARS-CoV and SARS-CoV-2 (Bojkova et al., 2021). For Ebolaviruses (approximately 19000 bases), 189 DCPs were identified between Reston virus (an ebolavirus that does not cause disease in humans) and human pathogenic Ebolaviruses, and the comparison of SARS-CoV and SARS-CoV-2 (approximately 30000 bases) had 891 DCPs (Bojkova et al., 2021; Pappalardo et al., 2016). By comparing different species, many more DCPs are be observed instead of comparing the sequence within species, as has been done here in our comparison of the SARS-CoV-2 clades.

Another possible reason for not having many DCPs in this analysis may be the duration

of the virus tracking. The data was downloaded in January 2021, and the outbreak occurred in December 2019, which means that the duration of tracking the virus genome is only one year. However, the virus is continuously mutating, and which may result in significant differences over time. As of 29 July 2021, the new Delta variant, or clade 21A, is predominating, which needs further research (www.who.int).

The frequency analysis also shows that most of the positions having different amino acid frequencies are clade 20H/501Y.V2, 20I/501Y.V1, and 20J/501Y.V3, which may suggest that there may be multiple factors that determine the characteristics of this virus.

Although this study has identified DCPs, this analysis could not reveal which amino acid affects the characteristics of SARS-CoV-2. Moreover, when comparing the virus within species, analysis of sequences over a longer time period is needed to identify DCPs. Further studies will be needed to identify which amino acid is significant for the virus pathogenicity.

In conclusion, we have revealed the DCPs between clades within SARS-CoV-2 in this study. These positions indicate the differences of phenotype between clades.

Chapter 3:

Drug adaptation to remdesivir in SARS-CoV-2

3.1 Introduction

SARS-CoV-2 has been spreading all around the world and infecting people globally. To end this pandemic, developing antiviral drugs is essential. Remdesivir is a broad-spectrum antiviral drug and has shown effectiveness against this virus *in vitro* (M. Wang et al., 2020). In addition, a clinical trial reported that remdesivir might increase the speed to recover from COVID-19 (Beigel et al., 2020). However, the mechanisms of resistance that may occur to remdesivir in SARS-CoV-2 is still unknown. Here, we analysed the changes that occurred to SARS-CoV-2 by adapting to remdesivir *in vitro*.

3.2 Methods

3.2.1 Virus culture and remdesivir adaptation – performed by collaborators at Goethe University Frankfurt

Two SARS-CoV-2 strains were used for this assay, FFM3 and FFM7. The virus was serially passaged with increasing concentration (starting concentration - 500nM) of remdesivir in Caco-2 cells. Viral replication was monitored by observation for any cytopathogenic effect present in the culture. Infected cultures were frozen and stored at −80°C and thawed once for subsequent passaging. The virus was serially passaged by using one aliquot of viral stock from the preceding passage to infect fresh Caco-2 cells (MOI of 0.1) in the presence of increasing concentrations of the compound. The drug concentrations used in the selection protocol varied, depending on the level of viral replication present in the preceding passage. The selection was carried out for a total of 30 passages with remdesivir. Drug concentrations ranged from 500 nM to 2

μM by passage 30. A further passage was carried out up to 60 passages with remdesivir. Drug concentrations ranged from 2 μM to 4 μM in this further passage. Drug sensitivity was determined by an antiviral assay in Caco-2 and Calu-3 cells as described above. For clarity, we described the virus as $FFM3_{LOW}$: without remdesivir (low passage), $FFM3_{HIGH}$: without remdesivir (high passage), $FFM3_{remLOW}$: with remdesivir (low concentration), $FFM3_{remHIGH}$: with remdesivir (high concentration).

3.2.2 Genome analysis

The sequence data were obtained using Next Generation Sequencing (NGS). This was performed by Public Health England. Five data sets were obtained from virus culture. Bases were called if one nucleotide was present in > 90% of the reads at that position. All other positions were classed as having a mixed population.

These data sets were compared with each other (Figure 3.1). First, the original virus sequence was compared with the virus sequence passaged without a drug (Figure 3.1 ①). The positions where base changes (including changes to mixed populations) occurred were extracted. Second, the high passaged virus was compared with the original sequence (Figure 3.1 ②). Third, the virus adapted to remdesivir of low concentration was compared with the original sequence (Figure 3.1 ③). Fourth, the virus adapted to the high concentration of remdesivir was also compared with the original virus (Figure 3.1 ④). Finally, the viruses adapted to the two different concentrations of remdesivir were compared (Figure 3.1 ⑤). These analyses were conducted in two strains, FFM3 and FFM7. Bases obtained from this analysis were converted to amino acids using the UCSC Genome Browser (Fernandes et al., 2020). By using the RNA codon table and the codon obtained from UCSC Genome Browser,

substations of amino acids were identified.

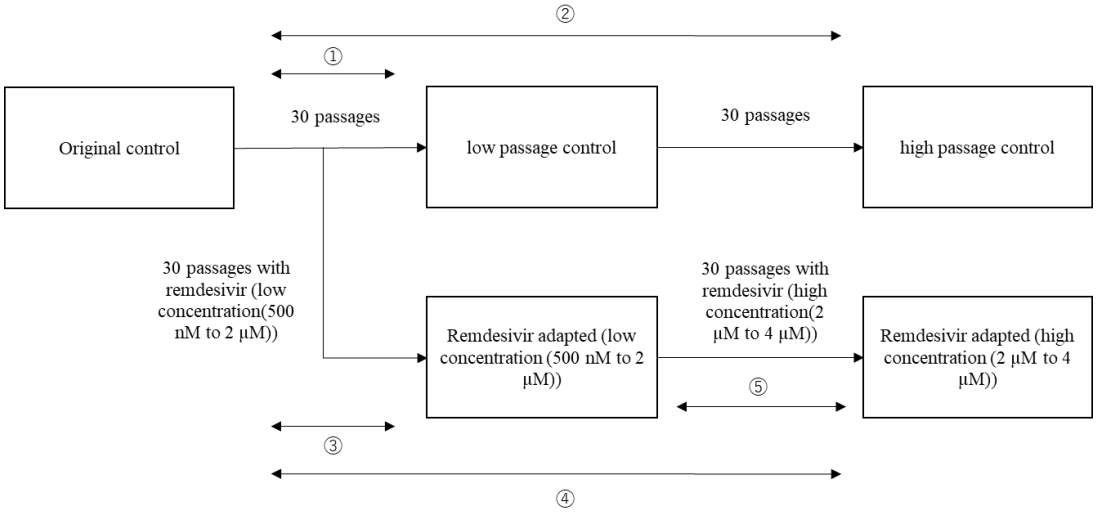


Figure 3.1 The procedure of the virus culture passage. ① is a process of the virus culture without remdesivir. ② is a process of virus culture without virus with additional 30 passages. In the process ③, the virus was cultured with remdesivir until adapted (30 passages). ④ is a process of culture with higher concentration of remdesivir.

3.2.3 Structural analysis

Structures of proteins for SARS-CoV-2 were obtained from Protein Data Bank (PDB). For the RNA dependant RNA polymerase (RdRp), the structures 7bv2 (PDB code) and 6yyt were used for the structure analysis. Changes of amino acids were mapped onto the protein structure using PyMOL. The PyMOL mutagenesis tool was used to analyse the different amino acids introduced. The impact of changes of amino acids was analysed manually based on hydrogen bonding and clashes.

3.3 Results

The aim of this study was to investigate the mechanisms of resistance that may occur to remdesivir in SARS-CoV-2. We identified a number of mutations that occurred in the remdesivir adapted virus in two strains, FFM3 and FFM7. Additionally, we assessed the effect on protein structure caused by the mutations.

3.3.1 Comparison between original virus and virus cultured without remdesivir

Each strain was cultured without remdesivir as a control to identify changes that may occur due to the adaptation to the cell culture. We refer to $FFM3_{LOW}$ and $FFM3_{HIGH}$ to represent the two different numbers of passaged strains without remdesivir. For the FFM3 strain, two base changes were present in $FFM3_{LOW,}$ which were C21789T and A28649T. In addition, 20 positions had mixed populations in $FFM3_{LOW}$ (Table 3.1; A mixed population is any position where one base is not present in >90% of reads).

Table 3.1 Comparison between original control and $FFM3_{LOW}$. Bases were called if one nucleotide was present in >90% of the reads at that position. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM3_{LOW}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{LOW}$ (aa)
510	G	(G, del)	NSP1	82	G	(G, V)
511	T	(T, del)	NSP1	82	G	(G, V)
512	C	(C, del)	NSP1	83	H	(H, del)
513	A	(A, del)	NSP1	83	H	(H, del)
514	T	(T, del)	NSP1	83	H	(H, del)
515	G	(G, del)	NSP1	84	V	(V, del)
516	T	(T, del)	NSP1	84	V	(V, del)
517	T	(T, del)	NSP1	84	V	(V, del)
518	A	(A, del)	NSP1	85	M	(M, del)
519	T	(T, del)	NSP1	85	M	(M, del)
520	G	(G, del)	NSP1	85	M	(M, del)
521	G	(G, del)	NSP1	86	V	(V, V)
522	T	(T, del)	NSP1	86	V	(V, V)
11083	G	(G, T, del)	NSP6	37	L	(L, F, F)
11750	C	(C, T)	NSP6	260	L	(L, F)
11916	C	(C, T)	NSP7	25	S	(S, L)
20480	C	(C, T)	NSP15	287	S	(S, L)
20573	T	(T, C)	NSP15	318	V	(V, A)
21789	C	T	Spike	76	T	I

22264	C	(C, T)	Spike	234	N	N
27131	C	(C, T)	M	203	N	N
28649	A	T	N	126	N	Y

By contrast, there were no base changes in the FFM7 strain (i.e. none where a different base was present in 90% of reads), although 15 positions with mixed populations of bases were observed in $FFM7_{LOW}$ (Table 3.2). Of the changes observed, there was only overlap between $FFM3_{LOW}$ and $FFM7_{LOW}$ for base change C21789, which had completely changed in $FFM3_{LOW}$ and was a mixed population in $FFM7_{LOW}$ (51% T21789 and 49% C21789).

Table 3.2 Comparison between original control and $FFM7_{LOW}$. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM7_{LOW}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM7_{LOW}$ (aa)
44	C	(T, C)	-	-	-	-
835	C	(T, C)	NSP2	10	F	F
2509	C	(C, T)	NSP2	568	P	P
9298	C	(C, T)	NSP4	248	Y	Y
11399	A	(G, A)	NSP6	143	M	(V, M)
16616	C	(C, T)	NSP13	127	T	(T, I)
21789	C	(T, C)	Spike	76	T	(I, T)
22032	T	(T, C)	Spike	157	F	(F, S)
23179	(C, T)	(T, C)	Spike	539	V	V
23271	C	(T, C)	Spike	570	A	(V, A)
23280	C	(C, T)	Spike	573	T	(T, I)
24130	C	(C, T)	Spike	856	N	N
27585	T	(T, G)	ORF7a	64	A	A
28311	C	(T, C)	N	13	P	(L, P)
28899	G	(G, T)	N	209	R	(R, I)

For $FFM3_{LOW}$, both C21789T and A28649T resulted in nonsynonymous mutations (Table 3.1). Base 21789 results in the amino acid change of T76I in the spike protein. A28649T is present in N and results in the amino acid change N126Y. Since C21789T is not a base change in $FFM7_{LOW}$, there were no nonsynonymous mutations in $FFM7_{LOW}$ (Table 3.2).

However, in both $FFM3_{LOW}$ and $FFM7_{LOW}$, many positions where different bases were present in a mixed population compared with the original virus were nonsynonymous mutations (Table 3.1, 3.2). Position 510 – 522 in $FFM3_{LOW}$ had a significant change in amino acid sequence, causing a frameshift. For $FFM7_{LOW}$, four of the positions in which the predominated bases (>50% of the read) in mixed populations differed from the original virus were nonsynonymous mutations.

3.3.2 Comparison between original virus and virus cultured without remdesivir in high passage

For the FFM3 strain, the two changes present in $FFM3_{LOW}$ (C21789T and A28649T) were also present in $FFM3_{HIGH}$ (Table 3.3). Additionally, 72 bases mixed populations were observed in $FFM3_{HIGH}$ (Table 3.3). Position 508 – 522 had a mixed population of having a deletion predominantly (>50% of the read). In addition, 14408 – 14414 also had a continuous deletion as a mixed population. Moreover, Position 29729 – 29759, which were in a non-coding region, were mixed population.

Table 3.3 Comparison between original control and $FFM3_{HIGH}$. Position 5962 lacked data and was less than 100% read. For simplicity, the mutation of positions 29729 – 29759, which are the non-coding region, were excluded from this table. The full table is shown in the supplementary

table. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM3_{LOW}$	$FFM3_{HIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{LOW}$ (aa)	$FFM3_{HIGH}$ (aa)
508	T	T	(del, T)	NSP1	81	H	H	H
509	G	G	(del, G)	NSP1	82	G	G	(del, G)
510	G	(G, del)	(del, G)	NSP1	82	G	(G, V)	(del, G)
511	T	(T, del)	(del, G)	NSP1	82	G	(G, V)	(del, G)
512	C	(C, del)	(del, C)	NSP1	83	H	(H, del)	(del, H)
513	A	(A, del)	(del, A)	NSP1	83	H	(H, del)	(del, H)
514	T	(T, del)	(del, T)	NSP1	83	H	(H, del)	(del, H)
515	G	(G, del)	(del, G)	NSP1	84	V	(V, del)	(del, V)
516	T	(T, del)	(del, T)	NSP1	84	V	(V, del)	(del, V)
517	T	(T, del)	(del, T)	NSP1	84	V	(V, del)	(del, V)
518	A	(A, del)	(del, A)	NSP1	85	M	(M, del)	(del, M)
519	T	(T, del)	(del, T)	NSP1	85	M	(M, del)	(del, M)
520	G	(G, del)	(del, G)	NSP1	85	M	(M, del)	(del, M)
521	G	(G, del)	(del, G)	NSP1	86	V	(V, V)	(H, V)
522	T	(T, del)	(del, T)	NSP1	86	V	(V, V)	(H, V)
1288	C	C	(C, T)	NSP2	161	C	C	C
2062	C	C	(C, T)	NSP2	419	A	A	A
5962	T	T	(T)	NSP3	1081	Y	Y	(Y)
6045	A	A	(A, del)	NSP3	1109	N	N	(N, I)
6046	T	T	(T, del)	NSP3	1109	N	N	(N, I)
6047	T	T	(T, del)	NSP3	1110	F	F	(F, I)
7521	C	C	(C, T)	NSP3	1601	T	T	(T, I)
8290	C	C	(C, T)	NSP3	1857	L	L	L
11760	A	A	(A, G)	NSP6	263	K	K	(K, R)
12016	T	T	(T, G)	NSP7	58	V	V	V
12334	(A, G, T)	(A, G, T)	A	NSP8	81	A	A	A
14408	T	T	(del, T)	NSP12	323	P	P	(L, P)
14409	T	T	(del, T)	NSP12	323	P	P	(L, P)
14410	A	A	(del, A)	NSP12	324	T	T	(del, T)
14411	C	C	(del, C)	NSP12	324	T	T	(del, T)
14412	A	A	(del, A)	NSP12	324	T	T	(del, T)

14413	A	A	(del, A)	NSP12	325	S	S	(L, S)
14414	G	G	(del, G)	NSP12	325	S	S	(L, S)
17146	A	A	(A, G)	NSP13	304	I	I	(I, V)
20178	C	C	(C, T)	NSP15	186	V	V	V
20480	C	(C, T)	(C, T)	NSP15	287	S	(S, L)	(S, L)
20573	T	(T, C)	(T, C)	NSP15	318	V	(V, A)	(V, A)
21789	C	T	T	Spike	76	T	I	I
22264	C	(C, T)	(C, T)	Spike	234	N	N	N
23271	C	C	(C, T)	Spike	570	A	A	(A, V)
25688	C	C	(C, T)	ORF3a	99	A	A	(A, V)
27131	C	(C, T)	(C, T)	M	203	N	N	N
28649	A	T	T	N	126	N	Y	Y

For $FFM7_{HIGH}$, three positions had changed bases compared with the original virus (C44T, C835T, and C23271T; Table 3.4). Moreover, mixed populations of bases were observed at 16 positions in $FFM7_{HIGH.}$ In comparison between the strains, there were no further overlaps between the two strains. Only one of the three base changes was nonsynonymous (C23271T). This mutation resulted in the amino acid change A570V in the spike protein.

Table 3.4 Comparison between original control and $FFM7_{HIGH}$. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM7_{LOW}$	$FFM7_{HIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM7_{LOW}$ (aa)	$FFM7_{HIGH}$ (aa)
44	C	(T, C)	T	-	-	-	-	-
835	C	(T, C)	T	NSP2	10	F	F	F
2509	C	(C, T)	(C, T)	NSP2	568	P	P	P
5299	T	T	(T, C)	NSP3	860	T	T	T
6255	C	C	(C, T)	NSP3	1179	A	A	(A, V)
11399	A	(G, A)	(G, A)	NSP6	143	M	(V, M)	(V, M)
12334	(A, G, T, C)	(A, G, T)	A	NSP8	81	A	A	A
16616	C	(C, T)	(C, T)	NSP13	127	T	(T, I)	(T, I)

17678	C	C	(T, C)	NSP13	481	T	T	(T, M)
19955	C	C	(T, C)	NSP15	112	T	T	(I, T)
21789	C	(T, C)	(T, C)	Spike	76	T	(I, T)	(I, T)
22032	T	(T, C)	(C, T)	Spike	157	F	(F, S)	(S, F)
23179	(C, T)	(T, C)	T	Spike	539	V	V	V
23271	C	(T, C)	T	Spike	570	A	(V, A)	V
23542	T	T	(T, C)	Spike	660	Y	Y	Y
27972	C	C	(T, C)	ORF8	27	Q	Q	(Stop, Q)
28311	C	(T, C)	(C, T)	N	13	P	(L, P)	(P, L)
28887	C	C	(C, T)	N	205	T	T	(T, I)
28899	G	(G, T)	(T, G)	N	209	R	(R, I)	(I, R)

3.3.3 Analysis of virus adapted to low concentration of remdesivir

Three base changes were present in both $FFM3_{remLOW}$ (C7321G, G15451A, and C18687T; Table 3.5) and $FFM7_{remLOW}$ (C3768T, C12459T, and C14786T; Table 3.6). Additionally, six and 12 positions had mixed populations of bases in $FFM3_{remLOW}$ and $FFM7_{remLOW}$, respectively. Furthermore, there were no overlaps between the full base changes between the two strains. However, C14786T is a base change in $FFM7_{remLOW}$ and had a mixed population in $FFM3_{remLOW}$ (51% of T14786, 48% of C14786, and 1% of deletion; Table 3.5, 3.6).

Table 3.5 Comparison between original control and $FFM3_{remLOW}$. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM3_{remLOW}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{remLOW}$ (aa)
1104	T	(C, T)	NSP2	100	I	(I, T)
7321	C	G	NSP3	1534	S	R
12334	(A, G, T)	A	NSP8	81	A	A
13961	T	(C, T)	NSP12	174	V	(V, A)
14786	C	(T, C, -)	NSP12	449	A	(A, V, V)
15451	G	A	NSP12	671	G	S

18687	C	T	NSP14	216	C	C
25603	C	(C, T)	ORF3a	71	L	L
29473	G	(T, G)	N	400	L	(F, L)

Table 3.6 Comparison between original control and FFM7$_{remLOW}$. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	FFM7$_{remLOW}$	Protein	Amino acid (aa) position	Original control (aa)	FFM7$_{remLOW}$ (aa)
335	C	(C, T)	NSP1	24	R	(R, C)
3768	C	T	NSP3	350	T	I
5386	T	(G, T)	NSP3	889	A	A
6323	G	(G, A)	NSP3	1202	E	(E, K)
12459	C	T	NSP8	123	T	I
13626	T	(T, C)	NSP12	62	D	D
14408	T	(T, C)	NSP12	323	L	(L, P)
14786	C	T	NSP12	449	A	V
21765	T	(del, T)	Spike	68	I	I
21766	A	(del, A)	Spike	68	I	I
21767	C	(del, C)	Spike	69	H	(del, H)
21768	A	(del, A)	Spike	69	H	(del, H)
21769	T	(del, T)	Spike	69	H	(del, H)
21770	G	(del, G)	Spike	70	V	(I, V)
24712	G	(G, T)	Spike	1050	M	(M, I)
29901	A	n/a	-	-	-	-
29902	A	n/a	-	-	-	-
29903	A	n/a	-	-	-	-

For FFM3$_{remLOW}$, C18687T was a synonymous mutation, while C7321G and G15451A were nonsynonymous. These mutations result in the amino acid changes S1534R in NSP3 and G671S in NSP12, respectively. In addition, four of the positions changed to mixed population had changes of predominant bases (>50% of the reads), and these were all nonsynonymous changes (C1104T, T13961C, C14786T, and G29473T; Table

3.5).

For $FFM7_{remLOW}$, all of the three base changes resulted in nonsynonymous changes, which were C3768T, C12459T, and C14786T (Table 3.6). C3768T is coded in NSP3, and the change of amino acid is T350I. For C12459T, this mutation refers to T123I of NSP8. Moreover, as mentioned before, C14786T is a change of amino acid of A449V in NSP12. Interestingly, although the positions are a mixed population, position 21765-21770 had a deletion which results in deletion of amino acid position 69 in the spike protein.

3.3.4 Analysis of virus adapted to the high concentration of remdesivir

Ten base changes occurred in both $FFM3_{remHIGH}$ (Table 3.7) and $FFM7_{remHIGH}$ (Table 3.8). Only C14786T overlapped between the two strains (Table 3.7, 3.8). Additionally, there were 8 and 21 positions with mixed populations of bases in $FFM3_{remHIGH}$ and $FFM7_{remHIGH}$, respectively.

Table 3.7 Comparison between original control and $FFM3_{remHIGH}$. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM3_{remLOW}$	$FFM3_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{remLOW}$ (aa)	$FFM3_{remHIGH}$ (aa)
703	C	C	(A, C)	NSP1	146	G	G	G
1001	G	G	A	NSP2	66	E	E	K
1093	C	C	(C, T)	NSP2	96	P	P	P
1104	T	(C, T)	C	NSP2	100	I	(I, T)	T
2107	T	T	(T, C)	NSP2	434	T	T	T
6205	G	G	(G, A)	NSP3	1162	K	K	K
6649	T	T	(T, A)	NSP3	1310	A	A	A
7321	C	G	G	NSP3	1534	S	R	R

13961	T	(C, T)	C	NSP12	174	V	(V, A)	A
14786	C	(T, C, del)	T	NSP12	449	A	(A, V, V)	V
15451	G	A	A	NSP12	671	G	S	S
16044	A	A	(A, G)	NSP12	868	P	P	P
18687	C	T	T	NSP14	216	C	C	C
20178	C	C	C	NSP15	186	V	V	V
24763	T	T	(T, C)	Spike	1067	Y	Y	Y
25003	A	A	(A, G)	Spike	1147	S	S	S
26541	A	A	G	M	7	T	T	A
29473	G	(T, G)	T	N	400	L	(F, L)	F

Table 3.8 Comparison between original control and FFM7$_{remHIGH}$. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	FFM7$_{remLOW}$	FFM7$_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	FFM7$_{remLOW}$ (aa)	FFM7$_{remHIGH}$ (aa)
1009	C	C	T	NSP2	68	S	S	S
3768	C	T	T	NSP3	350	T	I	I
3798	T	T	(G, T)	NSP3	360	F	F	(F, C)
4180	G	G	(G, T)	NSP3	487	K	K	(K, N)
4320	C	C	(C, T)	NSP3	534	A	A	(A, V)
5386	T	(G, T)	G	NSP3	889	A	A	A
8097	C	C	(C, T)	NSP3	1793	T	T	(T, I)
8764	T	T	G	NSP4	70	D	D	E
12334	(A, G, T, C)	(A, G, C)	A	NSP8	81	A	A	A
12459	C	T	T	NSP8	123	T	I	I
12694	G	G	(A, G)	NSP9	3	E	E	E
12796	A	A	(A, G)	NSP9	37	G	G	G
13626	T	(T, C)	(C, T)	NSP12	62	D	D	D
14120	C	C	(C, T)	NSP12	227	P	P	(P, L)
14786	C	T	T	NSP12	449	A	V	V
15699	C	C	(C, T)	NSP12	753	F	F	F
18555	C	C	(C, T)	NSP14	172	D	D	D
20915	G	G	(G, A)	NSP16	86	R	R	R
21765	T	(del, T)	(T, del)	Spike	68	I	I	I

21766	A	(del, A)	(A, del)	Spike	68	I	I	I
21767	C	(del, C)	(C, del)	Spike	69	H	(del, H)	(H, del)
21768	A	(del, A)	(A, del)	Spike	69	H	(del, H)	(H, del)
21769	T	(del, T)	(T, del)	Spike	69	H	(del, H)	(H, del)
21770	G	(del, G)	(G, del)	Spike	70	V	(I, V)	(V, del)
23179	(C, T)	C	C	Spike	539	V	V	V
24872	G	G	(G, A)	Spike	1104	V	V	(V, I)
27502	T	T	G	ORF7a	37	S	S	A
28742	A	A	G	N	157	I	I	V
29891	A	A	(A, G)	-	-	-	-	-

For $FFM3_{remHIGH}$, there were 8 nonsynonymous changes (positions - G1001A, T1104C, C7321G, T13961C, C14786T, G15451A, A26541G, and G29473T; Table 3.7). G1001A and T1104C are in NSP2 and result in the amino acid changes E66K and I100T. As mentioned before, C7321G results in amino acid change S1534R in NSP3. T13961C, C14786T, and G15451A are present in NSP12 and lead to the amino acid changes V174A, A449V, and G671S, respectively. A26541G is a change of T7A in the M protein. Moreover, G29473T is a change in the N protein at L400F.

For $FFM7_{remHIGH}$, there were six nonsynonymous mutations. The mutations are C3768T, T8764G, C12459T, C14786T, T27502G, and A28742G (Table 3.8). C3768T was encoded in NSP3 and resulted in the T350I amino acid change. The mutation of T8764G was present in NSP4, which led to D70E. C12459T was converted to T123I in NSP8, which is a cofactor of RdRp. As mentioned earlier, C14786T, which overlaps with $FFM3_{remHIGH}$, was converted to A449V present in NSP12, the main protein of RdRp. T27502G was present in ORF7a, which led to an amino acid change of S37A. The mutation of A28742G was converted to I157V in N protein.

3.3.5 Comparison of analysis between low concentration and high concentration of remdesivir

When comparing the original virus, $FFM3_{remLOW}$ and $FFM3_{remHIGH}$, we observed two positions where the base had not changed in $FFM3_{remLOW}$, but it had upon further adaption in $FFM3_{remHIGH}$. Additionally, for three positions with a mixed population of bases in $FFM3_{remLOW}$, the base has fully changed in $FFM3_{remHIGH}$ (Table 3.9). For a further eight positions, no change occurred in $FFM3_{remLOW}$ but there was a mixed population in $FFM3_{remHIGH}$. All of these changes may reflect the ongoing adaptation to remdesivir during passaging. There was also one position (25603) where a mixed population was present in $FFM3_{remLOW}$ but was back to C25603 in $FFM3_{remHIGH}$.

Table 3.9 Comparison between $FFM3_{remLOW}$ and $FFM3_{remHIGH}$. The positions where the bases are different between the concentration of remdesivir. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM3_{remLOW}$	$FFM3_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{remLOW}$ (aa)	$FFM3_{remHIGH}$ (aa)
703	C	C	(A, C)	NSP1	146	G	G	G
1001	G	G	A	NSP2	66	E	E	K
1093	C	C	(C, T)	NSP2	96	P	P	P
1104	T	(C, T)	C	NSP2	100	I	(I, T)	T
2107	T	T	(T, C)	NSP2	434	T	T	T
6205	G	G	(G, A)	NSP3	1162	K	K	K
6649	T	T	(T, A)	NSP3	1310	A	A	A
13961	T	(C, T)	C	NSP12	174	V	(V, A)	A
14786	C	(T, C, del)	T	NSP12	449	A	(A, V, V)	V
16044	A	A	(A, G)	NSP12	868	P	P	P
24763	T	T	(T, C)	Spike	1067	Y	Y	Y
25003	A	A	(A, G)	Spike	1147	S	S	S

25603	C	(C, T)	C	ORF3a	71	L	L	L
26541	A	A	G	M	7	T	T	A
29473	G	(T, G)	T	N	400	L	(F, L)	F

Similarly, by comparing the original virus, $FFM7_{remLOW}$ and $FFM7_{remHIGH}$, we observed four positions where the base had not changed in $FFM7_{remLOW}$ but had changed in $FFM7_{remHIGH}$. Moreover, one position having a mixed population in $FFM7_{remLOW}$ had a base change in $FFM7_{remHIGH}$. For a further four positions, a mixed population of bases in $FFM7_{remLOW}$ was back to the original base in $FFM7_{remHIGH}$. Furthermore, 12 positions had no changes in $FFM7_{remLOW,}$ but there was a mixed population in $FFM7_{remHIGH}$. Additionally, seven positions had mixed populations in both $FFM7_{remLOW}$ and $FFM7_{remHIGH}$, and six of these are consecutive positions. The last three bases (29901, 29902, and 29903) were not sequenced in $FFM7_{remLOW}$.

Table 3.10 Comparison between $FFM7_{remLOW}$ and $FFM7_{remHIGH}$. This table shows the positions where the nucleotide is different between $FFM7_{remLOW}$ and $FFM7_{remHIGH}$. (aa) represents amino acid. The bracket means that the position had mixed population.

Position	Original control	$FFM7_{remLOW}$	$FFM7_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM7_{remLOW}$ (aa)	$FFM7_{remHIGH}$ (aa)
335	C	(C, T)	C	NSP1	24	R	(R, C)	R
1009	C	C	T	NSP2	68	S	S	S
3798	T	T	(G, T)	NSP3	360	F	F	(F, C)
4180	G	G	(G, T)	NSP3	487	K	K	(K, N)
4320	C	C	(C, T)	NSP3	534	A	A	(A, V)
5386	T	(G, T)	G	NSP3	889	A	A	A
6323	G	(G, A)	G	NSP3	1202	E	(E, K)	E
8097	C	C	(C, T)	NSP3	1793	T	T	(T, I)
8764	T	T	G	NSP4	70	D	D	E
12694	G	G	(A, G)	NSP9	3	E	E	E

12796	A	A	(A, G)	NSP9	37	G	G	G
13626	T	(T, C)	(C, T)	NSP12	62	D	D	D
14120	C	C	(C, T)	NSP12	227	P	P	(P, L)
14408	T	(T, C)	T	NSP12	323	L	(L, P)	L
15699	C	C	(C, T)	NSP12	753	F	F	F
18555	C	C	(C, T)	NSP14	172	D	D	D
20915	G	G	(G, A)	NSP16	86	R	R	R
21765	T	(del, T)	(T, del)	Spike	68	I	I	I
21766	A	(del, A)	(A, del)	Spike	68	I	I	I
21767	C	(del, C)	(C, del)	Spike	69	H	(del, H)	(H, del)
21768	A	(del, A)	(A, del)	Spike	69	H	(del, H)	(H, del)
21769	T	(del, T)	(T, del)	Spike	69	H	(del, H)	(H, del)
21770	G	(del, G)	(G, del)	Spike	70	V	(del, V)	(V, del)
24712	G	(G, T)	G	Spike	1050	M	(M, I)	M
24872	G	G	(G, A)	Spike	1104	V	V	(V, I)
27502	T	T	G	ORF7a	37	S	S	A
28742	A	A	G	N	157	I	I	V
29891	A	A	(A, G)	-	-	-	-	-
29901	A	n/a	A	-	-	-	-	-
29902	A	n/a	A	-	-	-	-	-
29903	A	n/a	A	-	-	-	-	-

The bases which had not changed in $FFM3_{remLOW}$ but had changed upon further adaptation in $FFM3_{remHIGH}$ were G1001A and A26541G (Table 3.9). Moreover, the three bases changed from a mixed population in $FFM3_{remLOW}$ to fully base change by increasing the concentration is T1104C, T13961C, and G29473T (Table 3.9). The positions which had no change in $FFM3_{remLOW}$ but were mixed population in $FFM3_{remHIGH}$ are 703, 1093, 2107, 6205, 6649, 16044, 24763, and 25003.

For FFM7 strain, the four base changes observed in $FFM7_{remHIGH}$ but not in $FFM7_{remLOW}$ were C1009T, T8764G, T27502G, and A28742G (Table 3.10).

Furthermore, position 5386 changed from a mixed population in $FFM7_{remLOW}$ to a full base change by increasing the concentration of remdesivir in the FFM7 strain (Table 3.10). Position 335, 6323, 14408, and 24712 had a mixed population in $FFM7_{remLOW}$ but reverted in $FFM7_{remHIGH}$. Additionally, position 3798, 4180, 4320, 8097, 12694, 12796, 14120, 15699, 18555, 20915, 24872, and 29891 had no changes in $FFM7_{remLOW}$ but changed to mixed population in $FFM7_{remHIGH}$. Interestingly, 21765 – 21770 had a mixed population in $FFM7_{remLOW}$ and $FFM7_{remHIGH}$. None of the changes overlapped between the FFM3 strain and the FFM7 strain.

3.3.6 Summary of nonsynonymous mutations when bases fully changed

In summary, we observed two positions of nonsynonymous mutations in the FFM3 strain passaging without remdesivir. The virus adapted to remdesivir had eight positions of fully base changes which were nonsynonymous mutations in the FFM3 strain. Additionally, for the FFM7 strain, there was only one nonsynonymous mutation that occurred during passage without remdesivir. When adapted to remdesivir, six nonsynonymous mutations occurred in the FFM7 strain.

The two nonsynonymous mutations in the FFM3 strain passaged without remdesivir were C21789T and A28649T (Table 3.11). These were converted to T76I in the spike protein and N126Y in the N protein, respectively. The eight nonsynonymous mutations that occurred when adapted to remdesivir in the FFM3 strain were G1001A, T1104C, C7321G, T13961C, C14786T, G15451A, A26541G, and G29473T (Table 3.12). G1001A and T1104C are present in NSP2 and converted to E66K and I100T, respectively. The nonsynonymous mutation C7321G is converted to amino acid change of S1534R in NSP3. T13961C, C14786T, and G15451A are present in NSP12

and converted to V174A, A449V, and G671S, respectively. These mutations are likely to be the factor of resistance to remdesivir because the RdRp is the target of remdesivir and binds to this protein. Additionally, A26541G is encoded in M and converted to T7A. G29473T is an amino acid change of L400F present in N.

Table 3.11 Nonsynonymous mutations in $FFM3_{LOW}$ and $FFM3_{HIGH}$. (aa) represents amino acid. The upper half of the column "Low" in each row is the nucleotide or amino acid of the virus cultivated without remdesivir ($FFM3_{LOW}$). The lower half is the remdesivir adapted virus ($FFM3_{remLOW}$). For example, the upper cell of "Low" in position 21789 is T which means that the nucleotide in this position changed from C to T during passage without remdesivir. The same applies to column "High".

Position	Original control	Low	High	Protein	Position (aa)	Original control (aa)	Low (aa)	High (aa)
21789	C	T	T	Spike	76	T	I	I
		C	C				T	T
28649	A	T	T	N	126	N	Y	Y
		A	A				N	N

Table 3.12 Nonsynonymous mutations in $FFM3_{remLOW}$ and $FFM3_{remHIGH}$. (aa) represents amino acid. The upper half of the column "Low" in each row is the nucleotide or amino acid of the virus cultivated without remdesivir ($FFM3_{LOW}$). The lower half is the remdesivir adapted virus ($FFM3_{remLOW}$). The same applies to column "High".

Position	Original control	Low	High	Protein	Position (aa)	Original control (aa)	Low (aa)	High (aa)
1001	G	G	G	NSP2	66	E	E	E
		G	A				E	K
1104	T	T	T	NSP2	100	I	I	I
		(C, T)	C				(I, T)	T
7321	C	C	C	NSP3	1534	S	S	S
		G	G				R	R

13961	T	T (C, T)	T C	NSP12	174	V	V (V, A)	V A
14786	C	C (T, C, del)	C T	NSP12	449	A	A (V, A)	A V
15451	G	G A	G A	NSP12	671	G	G S	G S
26541	A	A A	A G	M	7	T	T T	T A
29473	G	G (T, G)	G T	N	400	L	L (F, L)	L F

In addition, the only nonsynonymous mutation observed in FFM7 strain passaging without remdesivir is C23271T which was converted to A570V in the spike protein (Table 3.13). The six nonsynonymous mutations that occurred during the adaptation to remdesivir in the FFM7 strain were C3768T, T8764G, C12459T, C14786T, T27502G, and A28742G (Table 3.14). C3768T was present in NSP3 and converted to T350I. The mutation T8764G was converted to amino acid change of D70E in NSP4. C12459T was encoded in NSP8 as an amino acid change of T123I. The mutation C14786T, which overlapped with the FFM3 strain, was an amino acid change of A449V in NSP12. T27502G was present in ORF7a and converted to S37A. A28742G was encoded in the N protein as an amino acid change of I157V.

Table 3.13 Nonsynonymous mutations in $FFM7_{LOW}$ and $FFM7_{HIGH}$. (aa) represents amino acid. The upper half of the column "Low" in each row is the nucleotide or amino acid of the virus cultivated without remdesivir ($FFM7_{LOW}$). The lower half is the remdesivir adapted virus

($FFM7_{remLOW}$). The same applies to column "High".

Position	Original control	Low	High	Protein	Position (aa)	Original_control (aa)	Low (aa)	High (aa)
23271	C	(T, C)	T	Spike	570	A	(V, A)	V
		C	C				A	A

Table 3.14 Nonsynonymous mutations in $FFM7_{remLOW}$ and $FFM7_{remHIGH}$. (aa) represents amino acid. The upper half of the column "Low" in each row is the nucleotide or amino acid of the virus cultivated without remdesivir ($FFM7_{LOW}$). The lower half is the remdesivir adapted virus ($FFM7_{remLOW}$). The same applies to column "High".

Position	Original control	Low	High	Protein	Position (aa)	Original_control (aa)	Low (aa)	High (aa)
3768	C	C	C	NSP3	350	T	T	T
		T	T				I	I
8764	T	T	T	NSP4	70	D	D	D
		T	G				D	E
12459	C	C	C	NSP8	123	T	T	T
		T	T				I	I
14786	C	C	C	NSP12	449	A	A	A
		T	T				V	V
27502	T	T	T	ORF7a	37	S	S	S
		T	G				S	A
28742	A	A	A	N	157	I	I	I
		A	G				I	V

3.3.7 Structural analysis

Throughout the genome analysis, we identified that the FFM3 strain showed eight nonsynonymous mutations in five different proteins. There were three nonsynonymous mutations in NSP12, two in NSP2, and one in each of NSP3, M, and N. Additionally, we observed six nonsynonymous mutations in the FFM7 strain in six different proteins. There was one change in each protein: NSP3, NSP4, NSP8, NSP12, ORF7a, and N. We used structural modelling to investigate the effect of the resulting amino acid change may have on protein structure and function and how they could have a role in remdesivir resistance. Since the target of remdesivir is RNA dependent RNA polymerase (RdRp), we focused on NSP12 and NSP8, which are the components of RdRp.

3.3.7.1 Structural analysis of NSP12

NSP12 is the main protein of RNA dependent RNA polymerase (RdRp) which has a role in replicating the RNA of the virus. Additionally, this protein is the target of remdesivir and binds to this protein. By binding to the protein, remdesivir terminates the replication process of the RNA. Thus, amino acid changes in this protein may have a role in remdesivir resistance.

We observed an amino acid change of V174A in the FFM3 strain. V174A is on the nidovirus RdRp-associated nucleotidyltransferase (NiRAN) domain of NSP12. This change of valine to alanine is a conservative change. They both have small hydrophobic amino side chains. Despite the change of size of the amino acid, clashes did not occur with other amino acids. In addition, there were no other possible rotamers by using PyMOL (Figure 3.2).

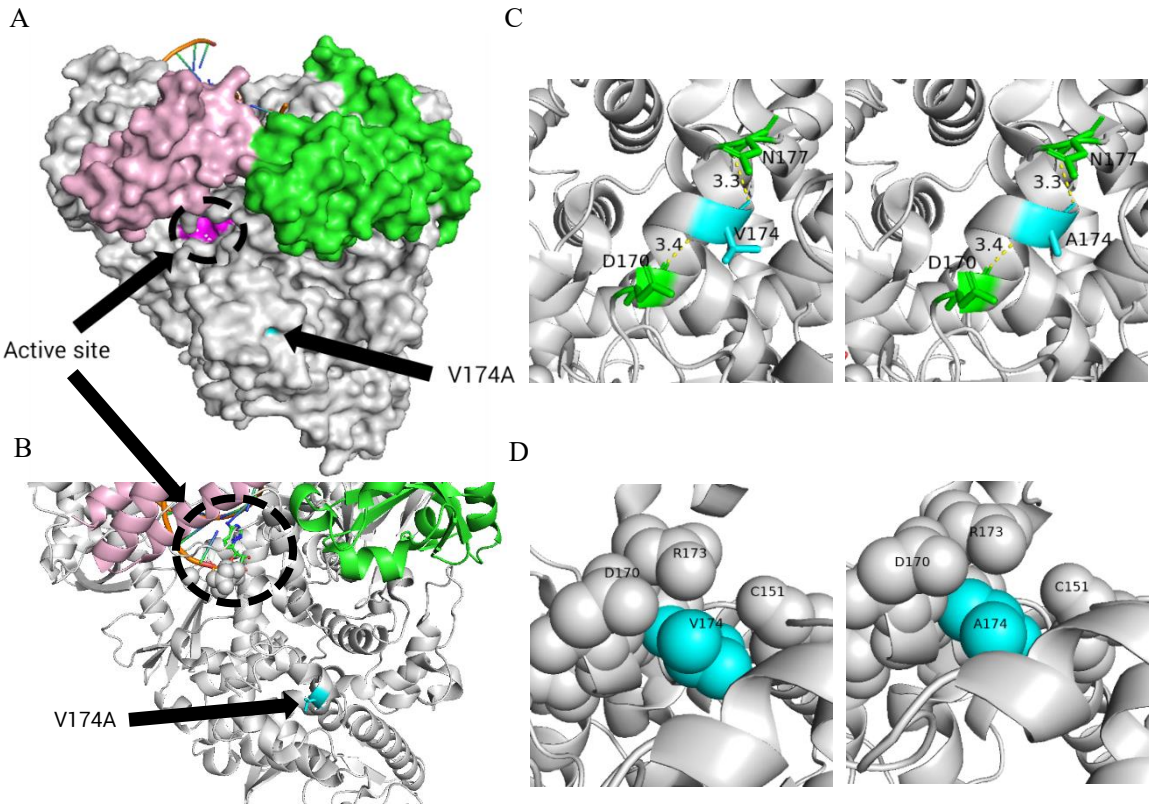


Figure 3.2 Structural analysis of V179A in NSP12. Change of V174A. NSP7, NSP8, and NSP12 are shown in pink, green, and grey, respectively. A174V is shown in cyan. PDB code: 7bv2. (A) The location of V174A in the RdRp. The active site of RdRp is shown in a black circle. (B) Closer view of V174A. (C) The left side is the original amino acid. The right side is the amino acid which changed to alanine. Polar contacts are shown in yellow lines. (D) The left side is the original amino acid, and the right side is the amino acid that changed to alanine.

Interestingly, the substitution of A449V was identified in both strains, FFM3 and FFM7. The change of A449V is also conservative, and alanine and valine are both hydrophobic. Three rotamers were observed in the analysis using PyMOL (Figure 3.3). Besides, due to the increase of the amino acid size from alanine to valine, A449V clashes between L544 in two of the possible rotamers, which could result in some conformational change. L544 is close to the active site of RdRp and may have an effect on the affinity of remdesivir. However, the structure model is static, so it is not possible to investigate this.

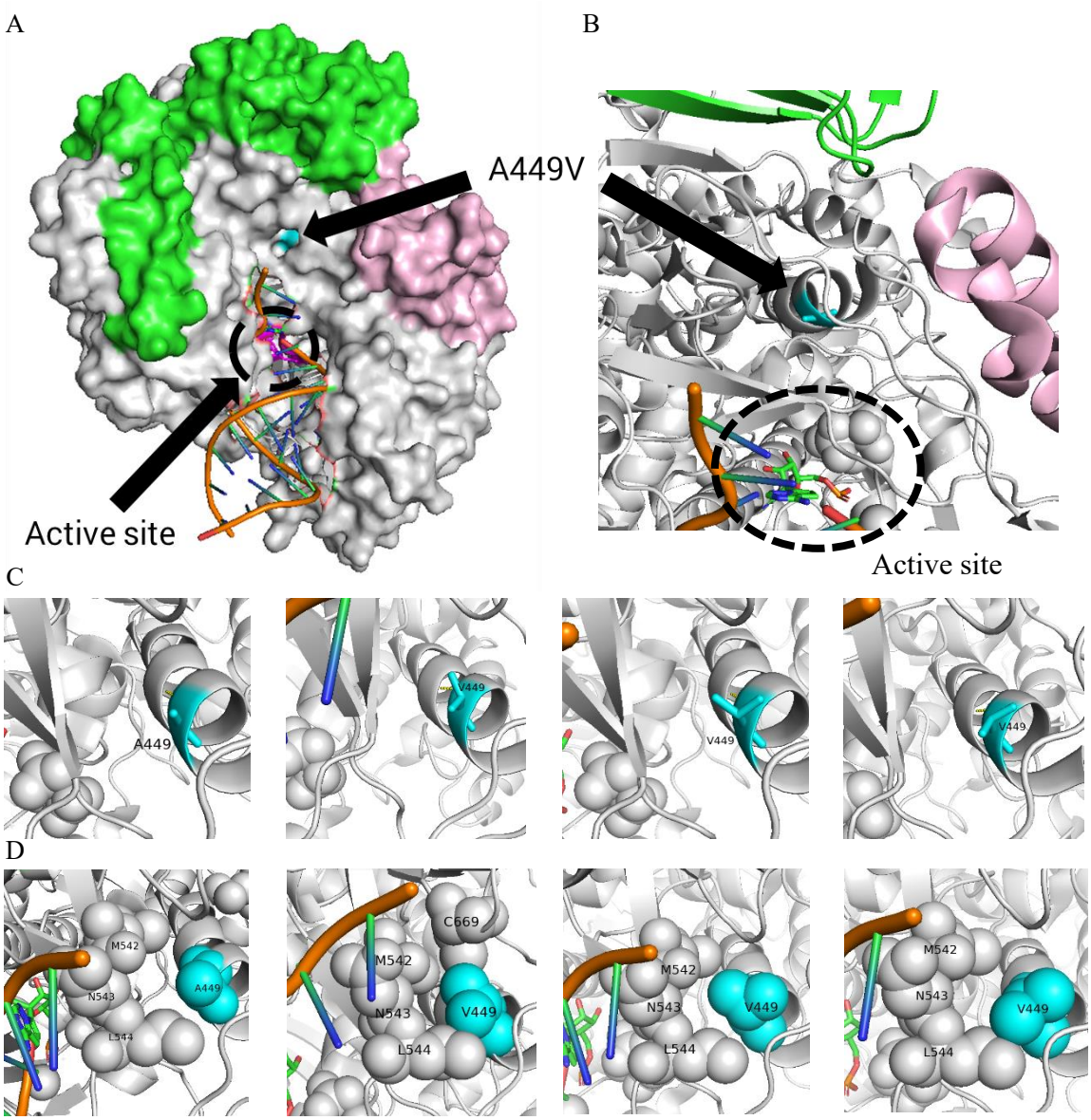
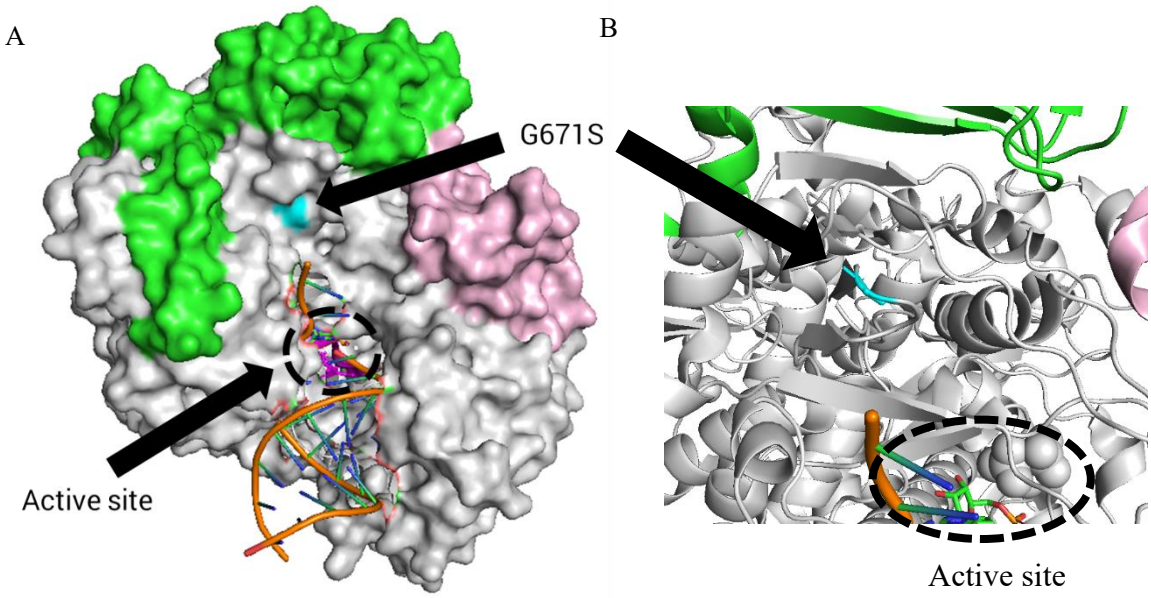
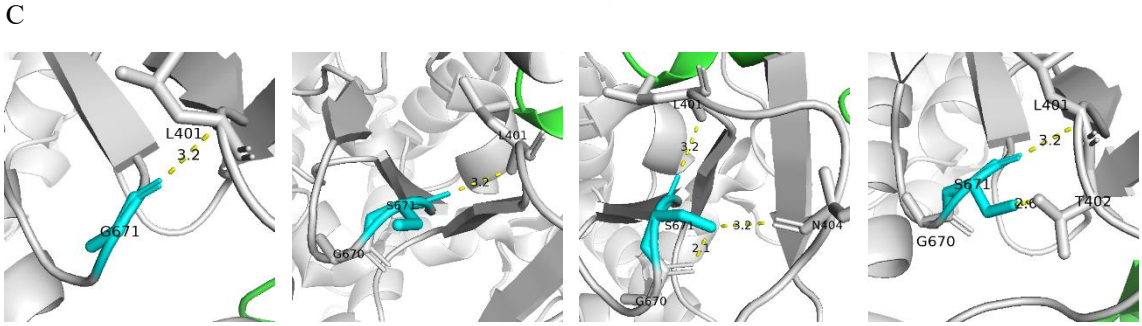


Figure 3.3 Structural analysis of A449V in NSP12. Changes of A449V. NSP7, NSP8, and NSP12 are shown in pink, green, and grey, respectively. A449V is described in cyan. PDB code: 7bv2. (A) The location of V174A in the RdRp. The active site of RdRp is shown in a black circle. (B) The closer view of A449V. (C) The original amino acid, A449, is shown in the left end. The other three figures are the possible rotamers of V449. (D) The figures correspond with C. The second figure from the left and the figure on the right end clash with L544.

The substitution of G671S was observed only in the FFM3 strain, not in the FFM7 strain. G671S is located on the surface of RdRp and close to A449V (Figure 3.4). This position changed from glycine to serine which is a hydrophilic uncharged side chain.

This change of properties may affect the structure of the protein. The backbone of the glycine has a hydrogen bond with the backbone of L401. Three possible rotamers were observed in the substitution from glycine to serine. The side chain of serine forms additional hydrogen bonds with other amino acids in two of the possible rotamers. One of them formed hydrogen bonds with the backbone of N404 and G670, and the other possible rotamer binds with the side chain of T402. T402 is on a different element of secondary structure as G671S, which may have an effect on protein flexibility or stability.





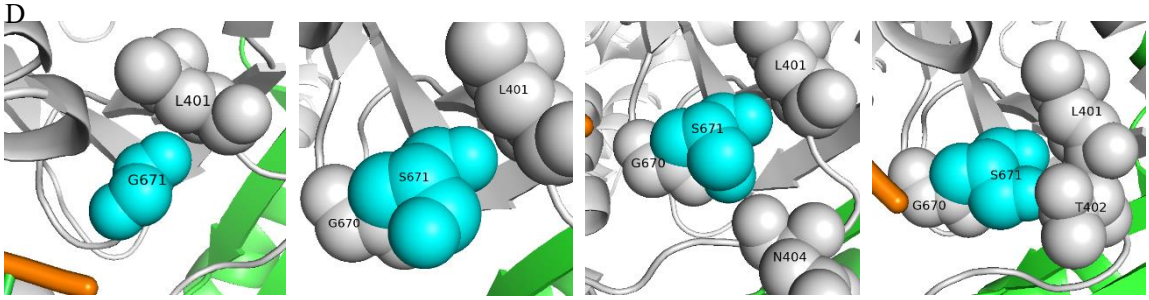
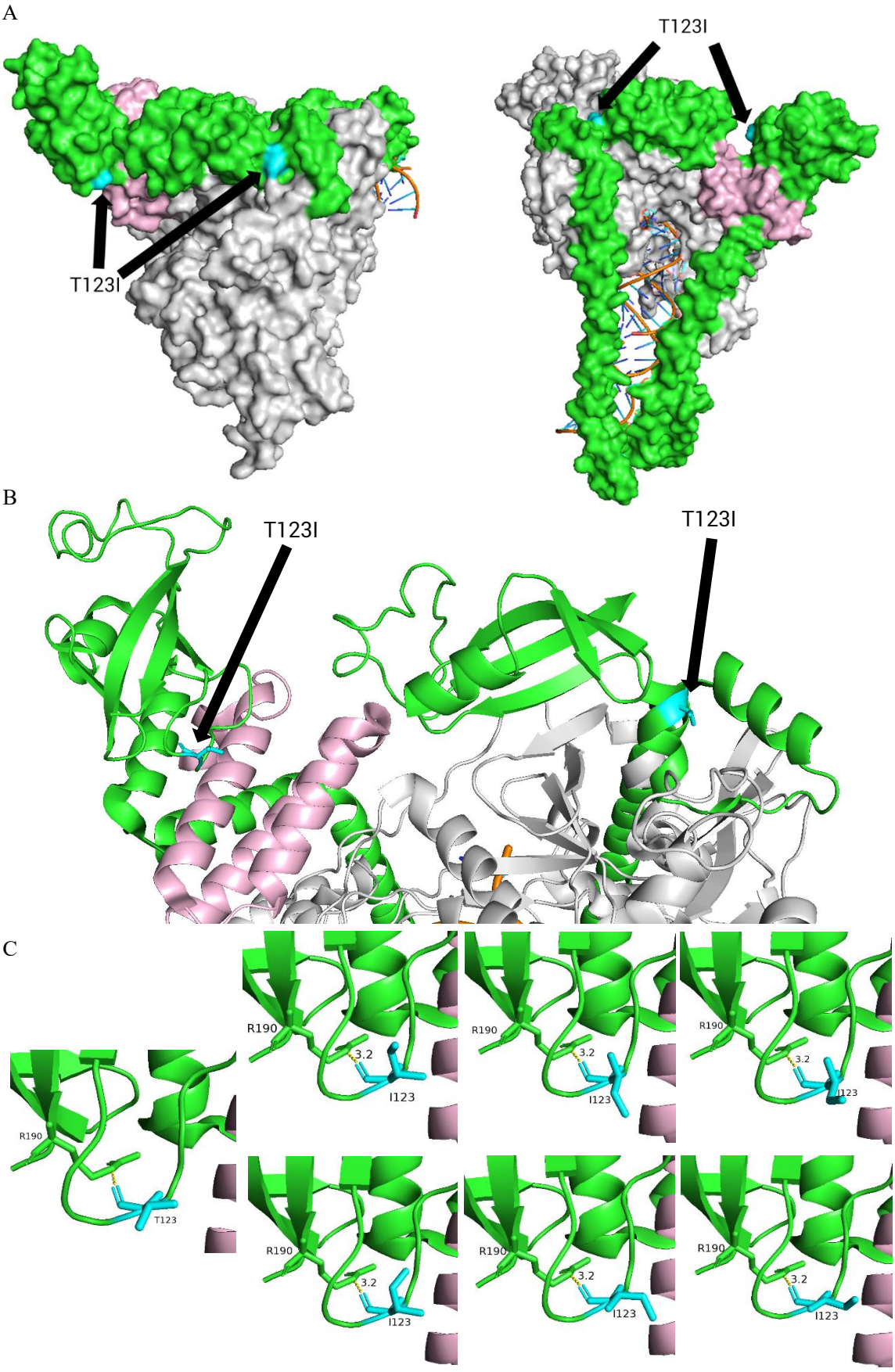


Figure 3.4 Structural analysis of G671S in NSP12. Change of G671S. NSP7, NSP8, and NSP12 are shown in pink, green, and grey, respectively. G671S is described in cyan. PDB code: 7bv2. (A) The location of V174A in the RdRp. The active site of RdRp is shown in a black circle. (B) The closer view of G671S. (C) The original amino acid, G671, is shown on the left end. The other three figures are the possible rotamers of S671. Polar contacts are shown in a yellow line. (D) The figures correspond with C. Every possible rotamer clash with G670. One of the rotamers also clashes with T402.

3.3.7.2 Structural analysis of NSP8

The RNA dependent RNA polymerase is a complex of NSP12, NSP7, and two NSP8. By adapting to remdesivir, the substitution of T123I was identified in NSP8. This change was only observed in the FFM7 strain and not in the FFM3 strain. Since the RdRp consists of two NSP8, T123I were observed in two parts of RdRp (Figure 3.5 A). The change (T123I) closer to NSP7 had six possible rotamers (Figure 3.5 C). T123 forms a hydrogen bond with the backbone of R190, but no other hydrogen bonds were observed. Every possible rotamer clashed with other amino acids due to the amino acid substitution. Three of the possible rotamers clashed with A125 and A126. Two of them clashed with T124, and one clashed with P121 (Figure 3.5 D).

Another T123, which was further from NSP7, binds three hydrogen bonds with the backbone of I119 and I120 (Figure 3.5 E). One of them was binding the side chain of T123 and the backbone of I119. There were three possible rotamers in this amino acid change. I123 lost one of the hydrogen bonding with I119, which was the binding between the side chain of T123 and the backbone of I119. Additionally, many clashes were observed in this amino acid change. One of the possible rotamers clashed with I106, and one of them also clashed with I119 and T124. However, one of the possible rotamers did not clash with other amino acids, which may indicate that this amino acid change can be accommodated at this position (Figure 3.5 F).


A
T123I
T123I
B
T123I
T123I
C
R190
3.2
I123
R190
T123

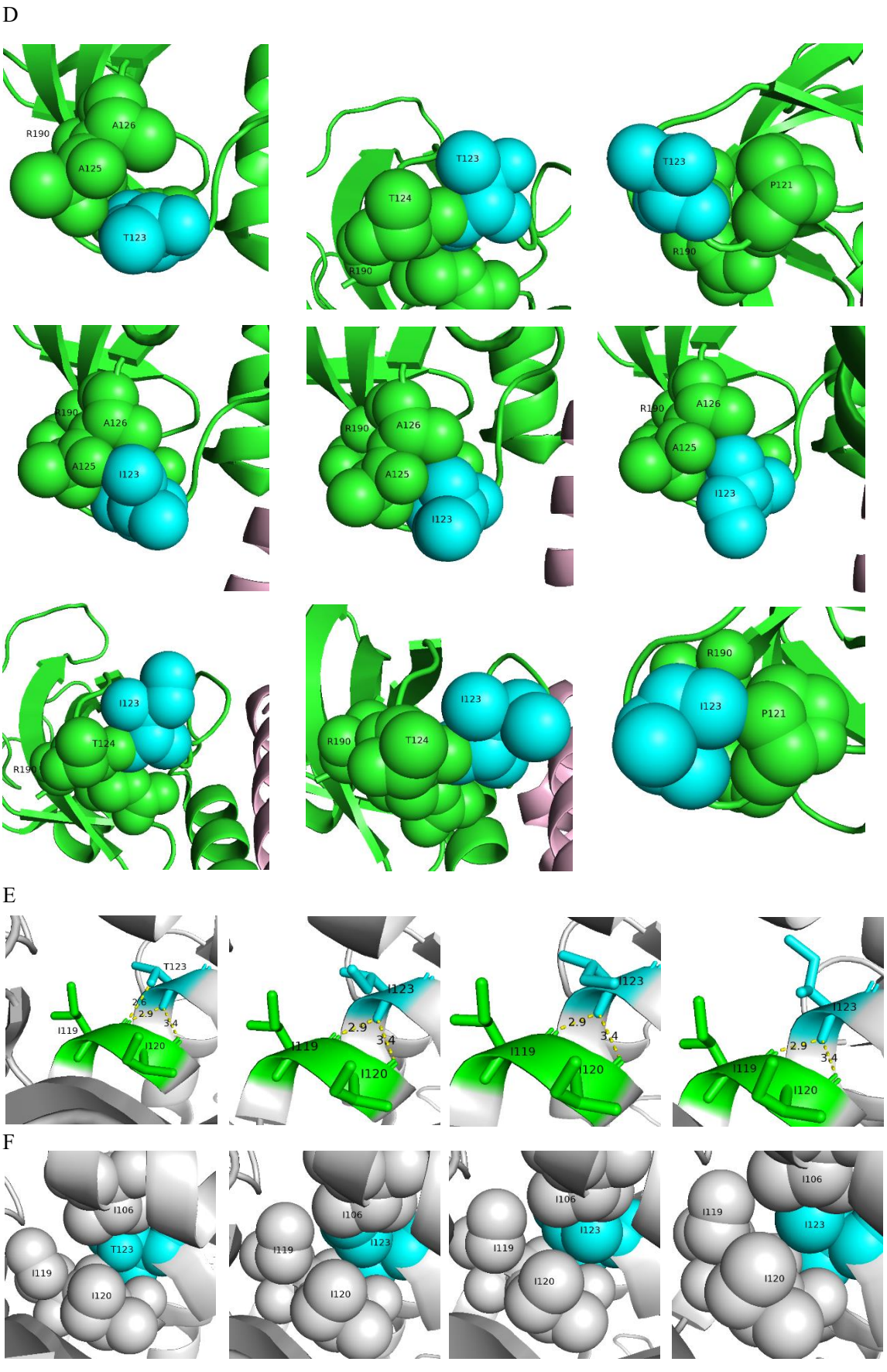


Figure 3.5 Structural analysis of I123T in NSP8. Change of T123I. NSP7, NSP8, and NSP12

are shown in pink, green, and grey, respectively. T123I is described in cyan. PDB code: 6yyt. (A) Location of T123I in RdRp. (B) T123I in a closer view. Two T123I are shown in the figure. (C) Hydrogen bonding in one of the T123I, which is on the left side of B. The figure on the left end is the original amino acid which is T123. Other figures are the possible rotamers of I123. (D) Clashes are shown in the same amino acid as C. The first row is the original amino acid T123 shown from different angles. The second row and the third row correspond with the possible rotamer shown in C. (E) Hydrogen bonding in one of the T123I, which is on the right side of B. The figure on the left end is the original amino acid, T123, and others are the possible rotamers of amino acid I123. (F) Clashes occur with I123. The figure on the left end is the original amino acid, T123, and others are the possible rotamers of amino acid I123.

3.4 Discussion

Here, we identified the mutations of SARS-CoV-2, which adapted to remdesivir, a broad-spectrum antiviral drug. Moreover, we have analysed the effect of the mutations on the structure of the RNA dependent RNA polymerase (RdRp), the target protein of remdesivir.

In the RNA sequence analysis, we identified three base changes for both strains FFM3 and FFM7 in comparison of initial sequence and the remdesivir adapted virus in low concentration. Additionally, ten base changes were observed by comparing initial virus and remdesivir adapted virus in high concentration for FFM3 and FFM7. Since the target of remdesivir is RdRp, we have looked into the protein and identified four amino acid substitutions in the FFM3 strain and FFM7 strain together.

We predicted to have mutations in the interacting region with RNA to inhibit remdesivir terminating the RNA synthesis (Hillen et al., 2020; Padhi et al., 2021; Yin et al., 2020). However, there were no mutations in the positions which directly interact with remdesivir and RNA.

Moreover, one study had reported that the substitution of E802D in NSP12 may be significant to confer reduction of sensitivity to remdesivir in SARS-CoV-2 (Szemiel et al., 2021a). However, we only had overlaps in the nucleotide positions 513 and 11750 with their report. This mutation was observed in the virus passaged without remdesivir in our analysis. These results may suggest that there are multiple factors that affect the resistance to remdesivir in SARS-CoV-2.

In this analysis, amino acid position 449 changed in both strains FFM3 and FFM7, which suggests that it may be a factor of resistance to remdesivir. However, the substitution was a conservative change, which was a change of alanine to valine. A clash was found between L544 in two of the possible rotamers. Remdesivir monophosphate form (RMP) forms interactions with side chains from K545, next to L544 (Yin et al., 2020). This may suggest that a clash between V449 and L544 may affect the interaction between RMP and K545, which may lead to a decrease of affinity to remdesivir binding to a new RNA during RNA synthesis.

Moreover, the amino acid change of G671S was observed in NSP12 in the FFM3 strain. For this change, one of the rotamers gained an additional hydrogen bond with T402 and a clash with this amino acid at the same time. T402 is on a different element of secondary structure as G671S, which may affect protein flexibility or stability.

Additionally, no base changes were observed in NSP7, which may suggest that NSP7 is not the factor of resistance to remdesivir in SARS-CoV-2. Only one change was identified in NSP8, which was T123I. This mutation causes clashes between other amino acids in the structure, which may affect the conformation. Interestingly, there

were no nonsynonymous mutations that occurred due to remdesivir adaptation in the spike protein. This may suggest that the spike protein does not have an effect on the remdesivir resistance.

These results may indicate that both strains we used, FFM3 and FFM7, does not mutate the interacting region and influence the remdesivir termination. Instead, the substitutions may have an effect by mutating the regions close to the binding region and cause clashes or additional hydrogen bonds. The mutations that cause resistance need the RdRp to still be active but to affect the binding of remdesivir. This may be the reason that we observed only four substitutions in the RdRp, as many possible mutations could have a significant effect on RdRp function.

We acknowledged certain limitations in this study. The protein structure we used to analyse the effect of amino acid substitution is static. However, the actual protein is dynamic, which the assessment of the effect would be limited. Moreover, further research is needed to analyse the effect of this mutation in experimental methods.

In conclusion, we identified base changes when SARS-CoV-2 is adapted to remdesivir *in vitro*. These base changes are likely to be the factor of remdesivir adaptation in SARS-CoV-2. We have also assessed the effect on the structure of RdRp by these mutations.

Chapter 4:

Discussion

In this thesis, we presented two main topics of SARS-CoV-2: Differentially conserved positions (DCPs) between clades of SARS-CoV-2 and drug adaptation of SARS-CoV-2. In the first study, we identified differentially conserved positions (DCPs) between clades using nomenclatures introduced by GISAID and Nextstrain within SARS-CoV-2. In the second study, we identified mutations that occur when adapted to remdesivir *in vitro* and assessed how the mutations may affect SARS-CoV-2.

4.1 Differentially conserved positions between clades in SARS-CoV-2

In the analysis of the comparison between SARS-CoV-2 clades, we identified three differentially conserved positions (DCPs) when using the GISAID nomenclature. In addition, we identified 33 DCPs in Nextstrain nomenclature. Since this analysis is a comparison within a single species, we found that there are few differences between clades compared with the comparisons using DCPs performed in previous research, such as comparing human pathogenic Ebolavirus species with Reston viruses, conducted by Wass-Michaelis group (Pappalardo et al., 2016) and the comparison between SARS-CoV and SARS-CoV-2 (Bojkova et al., 2021). As we did not find many DCPs, we looked into the frequencies of amino acids at each position. When analysing the clades from the GISAID nomenclature, there were six positions that had significant differences of amino acid frequencies between clades. By contrast, we identified 42 positions where amino acid frequencies, differed significantly when using the Nextstrain nomenclature. Most of the positions had differences in either clade

20H/501Y.V2, 20I/501Y.V1, or 20J/501Y.V3. These results may be relevant to the reports that propose these variants have different characteristics with higher transmissibility and reduced neutralisation ability (Davies et al., 2021; Souza et al., 2021; P. Wang et al., 2021). These differences between clades/variants may be caused by the substitutions observed in this analysis.

4.2 Drug adaptation to remdesivir in SARS-CoV-2

In chapter 3, we identified the mutations that occurred during the adaptation process to remdesivir in two strains of SARS-CoV-2, FFM3 and FFM7. We found that there are three full base changes in both strains and 10 full base changes for both strains with further adaptation to the higher concentration of remdesivir. We also modelled the changes of the amino acid in the structure to assess the effect caused by the mutations.

Four mutations occurred in the RNA dependant RNA polymerase (RdRp) in either FFM3 or FFM7. Interestingly, A449V was observed in the adapted virus of both strains, which may indicate that the mutation may be the factor of the resistance to remdesivir. This amino acid change causes clashes with L544, a proximate amino acid of K545 that interacts with the RNA. However, there were no other overlaps between these two strains. Additionally, Szemiel et al. reported that the substitution of E802D in NSP12 was sufficient to confer decreased sensitivity to remdesivir but this position was not observed in our analysis. Moreover, the changes in nucleotide position 513 and position 11750 overlapped between the report of Szemiel et al. and this analysis, but these positions only changed when the virus passaged without remdesivir in this analysis. These results of having different mutations may suggest that there may be multiple different mutations that can affect the resistance to remdesivir in SARS-CoV-

2. The other possibility is that there may be different positions for different strains that cause resistance to remdesivir. Further studies are needed to elucidate which mutations are causing the resistance and the possibility of having multiple factors.

4.3 Limitation of this study

There are some limitations in this study. Since the pandemic is ongoing, we could not cover all of the clades and variants which are currently predominating. During this project the 21A clade, or delta variant, has become predominate and is suggested to have a difference between earlier clades but this clade was not analysed as it was only appearing during the work and we decided to freeze the set of sequences at a certain point to avoid repeatedly performing the same analysis as new sequences are continuously obtained. Moreover, some clades, especially 20J/501Y.V3, did not have many sequences compared with earlier ones. In addition, we identified DCPs but we could not identify which amino acid is contributing to the difference in characteristics of clades.

Additionally, in chapter three, the structural analysis was conducted using static models that cannot assess the actual interaction of proteins. This analysis also could not reveal which mutation is causing a significant effect on the resistance of remdesivir.

In this thesis, research was conducted using computational methods, and ideally these results need to be validated using wet laboratory methods.

4.4 Future work

First, many more sequences and data of variants are now available to investigate SARS-CoV-2. This may give more details about the difference between clades and

variants. Second, experimental research can be performed to validate the results we found throughout the analysis of adaptation to remdesivir. This can be performed by changing the amino acids observed in this study, such as V174A, A449V, and G671S in NSP12.

References

Angelini, M. M., Akhlaghpour, M., Neuman, B. W., & Buchmeier, M. J. (2013). Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *MBio*, *4*(4). https://doi.org/10.1128/MBIO.00524-13

Arthur M Lesk. (2019). Introduction to bioinformatics. Fifth edition. Oxford : Oxford University Press.

Asselah, T., Durantel, D., Pasmant, E., Lau, G., & Schinazi, R. F. (2021). COVID-19: Discovery, diagnostics and drug development. *Journal of Hepatology*, *74*(1), 168–184. https://doi.org/10.1016/J.JHEP.2020.09.031

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., … Zaks, T. (2020). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa2035389

Bar-On, Y. M., Flamholz, A., Phillips, R., & Milo, R. (2020). Sars-cov-2 (Covid-19) by the numbers. *ELife*, *9*. https://doi.org/10.7554/ELIFE.57309

Bedford, T., Hodcroft, E. B., & Neher, R. A. (2021). *Updated Nextstain SARS-CoV-2 clade naming strategy*. https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming

Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., Hohmann, E., Chu, H. Y., Luetkemeyer, A., Kline, S., Lopez de Castilla, D., Finberg, R. W., Dierberg, K., Tapson, V., Hsieh, L., Patterson, T. F., Paredes,

R., Sweeney, D. A., Short, W. R., … Lane, H. C. (2020). Remdesivir for the Treatment of Covid-19 — Final Report. *New England Journal of Medicine*, *383*(19), 1813–1826. https://doi.org/10.1056/nejmoa2007764

Bojkova, D., McGreig, J. E., McLaughlin, K.-M., Masterson, S. G., Antczak, M., Widera, M., Krähling, V., Ciesek, S., Wass, M. N., Michaelis, M., & Cinatl, J. (2021). Differentially conserved amino acid positions may reflect differences in SARS-CoV-2 and SARS-CoV behaviour. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab094

Bouvet, M., Imbert, I., Subissi, L., Gluais, L., Canard, B., & Decroly, E. (2012). RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proceedings of the National Academy of Sciences*, *109*(24), 9372–9377. https://doi.org/10.1073/PNAS.1201130109

Brierley, I., Digard, P., & Inglis, S. C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell*, *57*(4), 537–547. https://doi.org/10.1016/0092-8674(89)90124-4

Campbell, F., Archer, B., Laurenson-Schafer, H., Jinnai, Y., Konings, F., Batra, N., Pavlin, B., Vandemaele, K., Kerkhove, M. D. Van, Jombart, T., Morgan, O., & Waroux, O. le P. de. (2021). Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance*, *26*(24), 2100509. https://doi.org/10.2807/1560-7917.ES.2021.26.24.2100509

Capra, J. A., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, *23*(15), 1875–1882. https://doi.org/10.1093/bioinformatics/btm270

CD, S., RL, G., GJ, C., JR, A. L., AM, C., A, S. V., O, O., P, M., KM, M., A, C., LYA, C., M, R., OTY, T., E, B., P, L. T., SC, C., D, S., RH, H., AO, O., … FM, M. (2020). Effect of Remdesivir vs Standard Care on Clinical Status at 11 Days in Patients With Moderate COVID-19: A Randomized Clinical Trial. *JAMA*, *324*(11), 1048–1057. https://doi.org/10.1001/JAMA.2020.16349

Centers for Disease Control and Prevention (CDC). www.cdc.gov

Chagoyen, M., García-Martín, J. A., & Pazos, F. (2016). Practical analysis of specificity-determining residues in protein families. *Briefings in Bioinformatics*, *17*(2), 255–261. https://doi.org/10.1093/bib/bbv045

Chand, M., Hopkins, S., Dabrera, G., Achison, C., Barclay, W., Ferguson, N., Volz, E., Loman, N., Rambaut, A., & Barrett, J. (2020). Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01 Technical Briefing 3. *Gov.Uk*, *December*, 1–11. https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet (London, England)*, *395*(10223), 507. https://doi.org/10.1016/S0140-6736(20)30211-7

Cho, A., Saunders, O. L., Butler, T., Zhang, L., Xu, J., Vela, J. E., Feng, J. Y., Ray, A. S., & Kim, C. U. (2012). Synthesis and antiviral activity of a series of 1′-substituted 4-aza-7,9-dideazaadenosine C-nucleosides. *Bioorganic & Medicinal Chemistry Letters*, *22*(8), 2705–2707. https://doi.org/10.1016/J.BMCL.2012.02.105

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N. G., & Decroly, E.

(2020). The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Research*, *176*, 104742. https://doi.org/10.1016/J.ANTIVIRAL.2020.104742

Cubuk, J., Alston, J. J., Incicco, J. J., Singh, S., Stuchell-Brereton, M. D., Ward, M. D., Zimmerman, M. I., Vithani, N., Griffith, D., Wagoner, J. A., Bowman, G. R., Hall, K. B., Soranno, A., & Holehouse, A. S. (2021). The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nature Communications 2021 12:1*, *12*(1), 1–17. https://doi.org/10.1038/s41467-021-21953-3

Dai, L., & Gao, G. F. (2020). Viral targets for vaccines against COVID-19. *Nature Reviews Immunology*. https://doi.org/10.1038/s41577-020-00480-0

Davidson, A. D., Williamson, M. K., Lewis, S., Shoemark, D., Carroll, M. W., Heesom, K. J., Zambon, M., Ellis, J., Lewis, P. A., Hiscox, J. A., & Matthews, D. A. (2020). Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Medicine 2020 12:1*, *12*(1), 1–15. https://doi.org/10.1186/S13073-020-00763-0

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., Zandvoort, K. van, Silverman, J. D., Group1‡, C. C.-19 W., Consortium‡, C.-19 G. U. (COG-U., Diaz-Ordaz, K., … Edmunds, W. J. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, *372*(6538), eabg3055. https://doi.org/10.1126/SCIENCE.ABG3055

Denison, M. R., Yount, B., Brockway, S. M., Graham, R. L., Sims, A. C., Lu, X., &

Baric, R. S. (2004). Cleavage between Replicase Proteins p28 and p65 of Mouse Hepatitis Virus Is Not Required for Virus Replication. *Journal of Virology*, *78*(11), 5957–5965. https://doi.org/10.1128/JVI.78.11.5957-5965.2004

Eastman, R. T., Roth, J. S., Brimacombe, K. R., Simeonov, A., Shen, M., Patnaik, S., & Hall, M. D. (2020). Remdesivir: A Review of Its Discovery and DevelopmentLeading to Emergency Use Authorization for Treatment of COVID-19. *ACS Central Science*, *6*(5), 672. https://doi.org/10.1021/ACSCENTSCI.0C00489

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 1–19. https://doi.org/10.1186/1471-2105-5-113

Egloff, M.-P., Ferron, F., Campanacci, V., Longhi, S., Rancurel, C., Dutartre, H., Snijder, E. J., Gorbalenya, A. E., Cambillau, C., & Canard, B. (2004). The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proceedings of the National Academy of Sciences*, *101*(11), 3792–3796. https://doi.org/10.1073/PNAS.0307877101

European Medicines Agency. www.ema.europa.eu

Faria, N. R., Claro, I. M., Candido, D., Franco, L. A. M., Andrade, P. S., Thais, M., Silva, C. A. M., Sales, F. C., Erika, R., Aguiar, R. S., Gaburo, N., Cecília, C., Fraiji, N. A., Crispim, M. A. E., Carvalho, P. S. S., & Rambaut, A. (2021). *Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings*. Virological.Org. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-

findings/586

Fernandes, J. D., Hinrichs, A. S., Clawson, H., Gonzalez, J. N., Lee, B. T., Nassar, L. R., Raney, B. J., Rosenbloom, K. R., Nerli, S., Rao, A. A., Schmelter, D., Fyfe, A., Maulding, N., Zweig, A. S., Lowe, T. M., Ares, M., Corbet-Detig, R., Kent, W. J., Haussler, D., & Haeussler, M. (2020). The UCSC SARS-CoV-2 Genome Browser. In *Nature Genetics* (Vol. 52, Issue 10, pp. 991–998). Nature Research. https://doi.org/10.1038/s41588-020-0700-8

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., Wang, T., Sun, Q., Ming, Z., Zhang, L., Ge, J., Zheng, L., Zhang, Y., Wang, H., Zhu, Y., Zhu, C., Hu, T., Hua, T., Zhang, B., … Rao, Z. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*, *368*(6492), 779–782. https://doi.org/10.1126/science.abb7498

Ghosh, S., Dellibovi-Ragheb, T., Pak, E., Qiu, Q., Fisher, M., Takvorian, P., Bleck, C., Hsu, V., Fehr, A., Perlman, S., Achar, S., Straus, M., Whittaker, G., Haan, C. de, Altan-Bonnet, G., & Altan-Bonnet, N. (2020). β-Coronaviruses use lysosomal organelles for cellular egress. *BioRxiv*, 2020.07.25.192310. https://doi.org/10.1101/2020.07.25.192310

Giacomelli, A., Pezzati, L., Conti, F., Bernacchia, D., Siano, M., Oreni, L., Rusconi, S., Gervasoni, C., Ridolfo, A. L., Rizzardini, G., Antinori, S., & Galli, M. (2020). Self-reported olfactory and taste disorders in patients with severe acute respiratory coronavirus 2 infection: A cross-sectional study. *Clinical Infectious Diseases*, *71*(15), 889–890. https://doi.org/10.1093/cid/ciaa330

Gilead Sciences, Inc. www.gilead.com

Global Initiative on Sharing All Influenza Data (GISAID). www.gisaid.org

Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva,

A. A., Haagmans, B. L., Lauber, C., Leontovich, A. M., Neuman, B. W., Penzar, D., Perlman, S., Poon, L. L. M., Samborskiy, D. V., Sidorov, I. A., Sola, I., & Ziebuhr, J. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. In *Nature Microbiology* (Vol. 5, Issue 4, pp. 536–544). Nature Research. https://doi.org/10.1038/s41564-020-0695-z

Gordon, C. J., Tchesnokov, E. P., Woolner, E., Perry, J. K., Feng, J. Y., Porter, D. P., & Götte, M. (2020). Remdesivir is a direct-acting antiviral that inhibits RNA-dependent RNA polymerase from severe acute respiratory syndrome coronavirus 2 with high potency. *Journal of Biological Chemistry*, *295*(20), 6785–6797. https://doi.org/10.1074/JBC.RA120.013679

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., … Zhong, N. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *Https://Doi.Org/10.1056/NEJMoa2002032*, *382*(18), 1708–1720. https://doi.org/10.1056/NEJMOA2002032

Han, A. X., Parker, E., Scholer, F., Maurer-Stroh, S., & Russell, C. A. (2019). Phylogenetic clustering by linear integer programming (PhyCLiP). *Molecular Biology and Evolution*, *36*(7), 1580–1595. https://doi.org/10.1093/molbev/msz053

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 2021 19:7*, *19*(7), 409–424. https://doi.org/10.1038/s41579-021-00573-0

Hilgenfeld, R. (2014). From SARS to MERS: crystallographic studies on coronaviral proteases enable antiviral drug design. *The FEBS Journal*, *281*(18), 4085–4096. https://doi.org/10.1111/FEBS.12936

Hillen, H. S., Kokic, G., Farnung, L., Dienemann, C., Tegunov, D., & Cramer, P. (2020). Structure of replicating SARS-CoV-2 polymerase. *Nature*, *584*(7819), 154–156. https://doi.org/10.1038/s41586-020-2368-8

Hodcroft, E., Aksamentov, I., Neher, R., Bedford, T., Hadfield, J., Zuber, M., Scottbrown, J., Sanderson, T., babarlelephant, Bloom, J., Roemer, C., acx01b, rasenschachmatt, Goater, R. https://covariants.org/, CoVariants

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Mü, M. A., Drosten, C., & Pö, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, *181*, 271-280.e8. https://doi.org/10.1016/j.cell.2020.02.052

Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2020). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology 2020 19:3*, *19*(3), 141–154. https://doi.org/10.1038/s41579-020-00459-7

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 2020 41:9*, *41*(9), 1141–1149. https://doi.org/10.1038/s41401-020-0485-4

Huang, Z., Tian, D., Liu, Y., Lin, Z., Lyon, C. J., Lai, W., Fusco, D., Drouin, A., Yin, X., Hu, T., & Ning, B. (2020). Ultra-sensitive and high-throughput CRISPR-p owered COVID-19 diagnosis. *Biosensors and Bioelectronics*, *164*, 112316. https://doi.org/10.1016/J.BIOS.2020.112316

Imbert, I., Snijder, E. J., Dimitrova, M., Guillemot, J. C., Lécine, P., & Canard, B. (2008). The SARS-Coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein. *Virus Research*, *133*(2), 136–148. https://doi.org/10.1016/J.VIRUSRES.2007.11.017

Investigation of novel SARS-CoV-2 variant Variant of Concern 202012/01 Technical briefing 5. (n.d.). Retrieved July 27, 2021, from www.gov.uk/government/publications/nervtag-paper-on-covid-19-variant-of-concern-b117

Irigoyen, N., Firth, A. E., Jones, J. D., Chung, B. Y.-W., Siddell, S. G., & Brierley, I. (2016). High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. *PLOS Pathogens*, *12*(2), e1005473. https://doi.org/10.1371/JOURNAL.PPAT.1005473

Isabel, S., Graña-Miraglia, L., Gutierrez, J. M., Bundalovic-Torma, C., Groves, H. E., Isabel, M. R., Eshaghi, A. R., Patel, S. N., Gubbay, J. B., Poutanen, T., Guttman, D. S., & Poutanen, S. M. (2020). Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *Scientific Reports*, *10*(1), 14031. https://doi.org/10.1038/s41598-020-70827-z

JD, G., DCB, L., DS, H., KM, M., R, B., R, M., CD, S., M, G., MY, A., RG, N., YS, C., D, S., RH, H., AO, O., H, C., C, B., X, W., A, G., DM, B., … A, S. (2020). Remdesivir for 5 or 10 Days in Patients with Severe Covid-19. *The New England Journal of Medicine*, *383*(19), 1827–1837.

https://doi.org/10.1056/NEJMOA2015301

Jordan, R., Hogg, A., Warren, T., De Wit, E., Sheahan, T., Lo, M., Soloveva, V., Weidner, J., Gomba, L., Feldmann, F., Cronin, J., Sims, A., Cockrell, A., Feng, J., Trantcheva, I., Babusis, D., Porter-Poulin, D., Bannister, R., Mackman, R., … Bavari, S. (2017). Broad-spectrum Investigational Agent GS-5734 for the Treatment of Ebola, MERS Coronavirus and Other Pathogenic Viral Infections with High Outbreak Potential. *Open Forum Infectious Diseases*, *4*(suppl_1), S737–S737. https://doi.org/10.1093/OFID/OFX180.008

Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, X., He, S., Zhou, Z., Zhou, Z., Chen, Q., Yan, Y., Zhang, C., Shan, H., & Chen, S. (2020). Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharmaceutica Sinica B, 10*(7), 1228–1238. https://doi.org/10.1016/J.APSB.2020.04.009

Kim, D., Lee, J. Y., Yang, J. S., Kim, J. W., Kim, V. N., & Chang, H. (2020). The Architecture of SARS-CoV-2 Transcriptome. *Cell*, *181*(4), 914-921.e10. https://doi.org/10.1016/J.CELL.2020.04.011

Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/JMBI.2000.4315

Ledford, H., Cyranoski, D., & Van Noorden, R. (2020). The UK has approved a COVID vaccine - here's what scientists now want to know. *Nature*, *588*(7837), 205–206. https://doi.org/10.1038/D41586-020-03441-8

Lehmann, K. C., Gulyaeva, A., Zevenhoven-Dobbe, J. C., Janssen, G. M. C., Ruben, M., Overkleeft, H. S., van Veelen, P. A., Samborskiy, D. V., Kravchenko, A.

A., Leontovich, A. M., Sidorov, I. A., Snijder, E. J., Posthuma, C. C., & Gorbalenya, A. E. (2015). Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. *Nucleic Acids Research*, *43*(17), 8416–8434. https://doi.org/10.1093/NAR/GKV838

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annual Review of Virology*, *3*(1), 237–261. https://doi.org/10.1146/annurev-virology-110615-042301

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., … Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine*, *382*(13), 1199–1207. https://doi.org/10.1056/nejmoa2001316

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE TRANSACTIONS ON INFORMATION THEORY*, *37*.

Liu, C., Ginn, H. M., Dejnirattisai, W., Supasa, P., Wang, B., Tuekprakhon, A., Nutalai, R., Zhou, D., Mentzer, A. J., Zhao, Y., Duyvesteyn, H. M. E., López-Camacho, C., Slon-Campos, J., Walter, T. S., Skelly, D., Johnson, S. A., Ritter, T. G., Mason, C., Costa Clemens, S. A., … Screaton, G. R. (2021). Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell*, *184*(16), 4220-4236.e13. https://doi.org/10.1016/j.cell.2021.06.020

Liu, J., Liu, Y., Xia, H., Zou, J., Weaver, S. C., Swanson, K. A., Cai, H., Cutler, M., Cooper, D., Muik, A., Jansen, K. U., Sahin, U., Xie, X., Dormitzer, P. R., & Shi, P. Y. (2021). BNT162b2-elicited neutralization of B.1.617 and other

SARS-CoV-2 variants. *Nature*. https://doi.org/10.1038/s41586-021-03693-y

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., … Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, *395*(10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8

Lu, X., Zhang, L., Du, H., Zhang, J., Li, Y. Y., Qu, J., Zhang, W., Wang, Y., Bao, S., Li, Y., Wu, C., Liu, H., Liu, D., Shao, J., Peng, X., Yang, Y., Liu, Z., Xiang, Y., Zhang, F., … Wong, G. W. K. (2020). SARS-CoV-2 Infection in Children. *Https://Doi.Org/10.1056/NEJMc2005073*, *382*(17), 1663–1665. https://doi.org/10.1056/NEJMC2005073

Luers, J. C., Rokohl, A. C., Loreck, N., Wawer Matos, P. A., Augustin, M., Dewald, F., Klein, F., Lehmann, C., & Heindl, L. M. (2020). Olfactory and gustatory dysfunction in coronavirus disease 2019 (COVID-19). *Clinical Infectious Diseases*, *71*(16), 2262–2264. https://doi.org/10.1093/cid/ciaa525

Mak, G. C., Cheng, P. K., Lau, S. S., Wong, K. K., Lau, C. S., Lam, E. T., Chan, R. C., & Tsang, D. N. (2020). Evaluation of rapid antigen test for detection of SARS-CoV-2 virus. *Journal of Clinical Virology*, *129*, 104500. https://doi.org/10.1016/J.JCV.2020.104500

Manzanares-Meza, L. D., & Medina-Contreras, O. (2020). SARS-CoV-2 and influenza: A comparative overview and treatment implications. *Boletin Medico Del Hospital Infantil de Mexico*, *77*(5), 262–273. https://doi.org/10.24875/BMHIM.20000183

Mariano, G., Farthing, R. J., Lale-Farjat, S. L. M., & Bergeron, J. R. C. (2020). Structural Characterization of SARS-CoV-2: Where We Are, and Where We

Need to Be. *Frontiers in Molecular Biosciences*, *0*, 344. https://doi.org/10.3389/FMOLB.2020.605236

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour 2021*, 1–7. https://doi.org/10.1038/s41562-021-01122-8

Maurer-Stroh, S. (2020). *GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses*. https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/

Mercer, T. R., & Salit, M. (2021). Testing at scale during the COVID-19 pandemic. *Nature Reviews Genetics 2021 22:7*, *22*(7), 415–426. https://doi.org/10.1038/s41576-021-00360-w

Mohamed, K., Yazdanpanah, N., Saghazadeh, A., & Rezaei, N. (2021). Computational drug discovery and repurposing for the treatment of COVID-19: A systematic review. *Bioorganic Chemistry*, *106*, 104490. https://doi.org/10.1016/J.BIOORG.2020.104490

Mulangu, S., Dodd, L. E., Richard T. Davey, J., Mbaya, O. T., Proschan, M., Mukadi, D., Manzo, M. L., Nzolo, D., Oloma, A. T., Ibanda, A., Ali, R., Coulibaly, S., Levine, A. C., Grais, R., Diaz, J., Lane, H. C., Muyembe-Tamfum, J.-J., & Group, the P. W. (2019). A Randomized, Controlled Trial of Ebola Virus Disease Therapeutics. *Https://Doi.Org/10.1056/NEJMoa1910993*, *381*(24), 2293–2303. https://doi.org/10.1056/NEJMOA1910993

Nextstrain. www.nextstrain.org

Padhi, A. K., Shukla, R., Saudagar, P., & Tripathi, T. (2021). High-throughput

rational design of the remdesivir binding site in the RdRp of SARS-CoV-2: implications for potential resistance. *IScience*, *24*(1), 101992. https://doi.org/10.1016/j.isci.2020.101992

Pappalardo, M., Julia, M., Howard, M. J., Rossman, J. S., Michaelis, M., & Wass, M. N. (2016). Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses. *Scientific Reports*, *6*(March). https://doi.org/10.1038/srep23743

Pardi, N., Hogan, M. J., Porter, F. W., & Weissman, D. (2018). mRNA vaccines-a new era in vaccinology. In *Nature Reviews Drug Discovery* (Vol. 17, Issue 4, pp. 261–279). Nature Publishing Group. https://doi.org/10.1038/nrd.2017.243

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., … Gruber, W. C. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615. https://doi.org/10.1056/nejmoa2034577

Rahman, M. S., Hoque, M. N., Islam, M. R., Islam, I., Mishu, I. D., Rahaman, M. M., Sultana, M., & Hossain, M. A. (2021). Mutational insights into the envelope protein of SARS-CoV-2. *Gene Reports*, *22*, 100997. https://doi.org/10.1016/J.GENREP.2020.100997

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Rausell, A., Juan, D., Pazos, F., & Valencia, A. (2010). Protein interactions and

ligand binding: From protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(5), 1995–2000. https://doi.org/10.1073/pnas.0908044107

Rawat, K., Kumari, P., & Saha, L. (2021). COVID-19 vaccine: A recent update in pipeline vaccines, their design and development strategies. *European Journal of Pharmacology*, *892*, 173751. https://doi.org/10.1016/J.EJPHAR.2020.173751

Repurposed Antiviral Drugs for Covid-19 — Interim WHO Solidarity Trial Results. (2021). *New England Journal of Medicine*, *384*(6), 497–511. https://doi.org/10.1056/nejmoa2023184

Rochwerg, B., Siemieniuk, R., & Jacobs, D. (2021). *Guideline Therapeutics and COVID-19: living guideline*.

Rohaim, M. A., El Naggar, R. F., Clayton, E., & Munir, M. (2021). Structural and functional insights into non-structural proteins of coronaviruses. In *Microbial Pathogenesis* (Vol. 150, p. 104641). Academic Press. https://doi.org/10.1016/j.micpath.2020.104641

Saha, A., Sharma, A. R., Bhattacharya, M., Sharma, G., Lee, S. S., & Chakraborty, C. (2020). Probable Molecular Mechanism of Remdesivir for the Treatment of COVID-19: Need to Know More. *Archives of Medical Research*, *51*(6), 585–586. https://doi.org/10.1016/J.ARCMED.2020.05.001

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733–9748. https://doi.org/10.1128/JVI.00694-10

Santos, J. C., & Passos, G. A. (n.d.). *The high infectivity of SARS-CoV-2 B.1.1.7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation*. https://doi.org/10.1101/2020.12.29.424708

Savastano, A., Opakua, A. I. de, Rankovic, M., & Zweckstetter, M. (2020). Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nature Communications 2020 11:1*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-19843-1

Schoeman, D., & Fielding, B. C. (2019). Coronavirus envelope protein: current knowledge. *Virology Journal 2019 16:1*, *16*(1), 1–22. https://doi.org/10.1186/S12985-019-1182-0

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Souza, W. M., Amorim, M. R., Sesti-Costa, R., Coimbra, L. D., Brunetti, N. S., Toledo-Teixeira, D. A., Souza, G. F. de, Muraro, S. P., Parise, P. L., Barbosa, P. P., Bispo-dos-Santos, K., Mofatto, L. S., Simeoni, C. L., Claro, I. M., Duarte, A. S. S., Coletti, T. M., Zangirolami, A. B., Costa-Lima, C., Gomes, A. B. S. P., … Proença-Módena, J. L. (2021). Neutralisation of SARS-CoV-2 lineage P.1 by antibodies elicited through natural SARS-CoV-2 infection or vaccination with an inactivated SARS-CoV-2 vaccine: an immunological study. *The Lancet Microbe*, *0*(0). https://doi.org/10.1016/S2666-5247(21)00129-4

Szemiel, A. M., Merits, A., Orton, R. J., MacLean, O., Pinto, R. M., Wickenhagen, A., Lieber, G., Turnbull, M. L., Wang, S., Mair, D., Filipe, A. da S., Willett, B. J., Wilson, S. J., Patel, A. H., Thomson, E. C., Palmarini, M., Kohl, A., & Stewart, M. E. (2021a). In vitro evolution of Remdesivir resistance reveals

genome plasticity of SARS-CoV-2. *BioRxiv*, 2021.02.01.429199. https://doi.org/10.1101/2021.02.01.429199

Szemiel, A. M., Merits, A., Orton, R. J., MacLean, O., Pinto, R. M., Wickenhagen, A., Lieber, G., Turnbull, M. L., Wang, S., Mair, D., Filipe, A. da S., Willett, B. J., Wilson, S. J., Patel, A. H., Thomson, E. C., Palmarini, M., Kohl, A., & Stewart, M. E. (2021b). In vitro evolution of Remdesivir resistance reveals genome plasticity of SARS-CoV-2. *BioRxiv*, 2021.02.01.429199. https://doi.org/10.1101/2021.02.01.429199

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., Gottberg, A. von, Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Zyl, G. Van, … Oliveira, T. de. (2020). Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *MedRxiv*, *10*, 2020.12.21.20248640. https://doi.org/10.1101/2020.12.21.20248640

UK government. www.gov.uk

U.S. Food and Drug Administration. www.fda.gov

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., & Thiel, V. (2020). Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology 2020 19:3*, *19*(3), 155–170. https://doi.org/10.1038/s41579-020-00468-6

Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A., & Kozlakidis, Z. (2021). Considerations for diagnostic COVID-19 tests. *Nature Reviews Microbiology*, *19*(3), 171–183. https://doi.org/10.1038/s41579-020-00461-z

Vilar, S., & Isom, D. G. (2020). One Year of SARS-CoV-2: How Much Has the

Virus Changed? *BioRxiv*, 2020.12.16.423071. https://doi.org/10.1101/2020.12.16.423071

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., Zhao, Y., Li, Y., Wang, X., & Peng, Z. (2020). Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA*, *323*(11), 1061–1069. https://doi.org/10.1001/JAMA.2020.1585

Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., Shi, Z., Hu, Z., Zhong, W., & Xiao, G. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. In *Cell Research* (Vol. 30, Issue 3, pp. 269–271). Springer Nature. https://doi.org/10.1038/s41422-020-0282-0

Wang, P., Nair, M. S., Liu, L., Iketani, S., Luo, Y., Guo, Y., Wang, M., Yu, J., Zhang, B., Kwong, P. D., Graham, B. S., Mascola, J. R., Chang, J. Y., Yin, M. T., Sobieszczyk, M., Kyratsous, C. A., Shapiro, L., Sheng, Z., Huang, Y., & Ho, D. D. (2021). Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature 2021 593:7857*, *593*(7857), 130–135. https://doi.org/10.1038/s41586-021-03398-2

Wang, Y., Zhang, D., Du, G., Du, R., Zhao, J., Jin, Y., Fu, S., Gao, L., Cheng, Z., Lu, Q., Hu, Y., Luo, G., Wang, K., Lu, Y., Li, H., Wang, S., Ruan, S., Yang, C., Mei, C., … Wang, C. (2020). Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet*, *395*(10236), 1569–1578. https://doi.org/10.1016/S0140-6736(20)31022-9

Warren, T. K., Jordan, R., Lo, M. K., Ray, A. S., Mackman, R. L., Soloveva, V., Siegel, D., Perron, M., Bannister, R., Hui, H. C., Larson, N., Strickley, R.,

Wells, J., Stuthman, K. S., Tongeren, S. A. Van, Garza, N. L., Donnelly, G., Shurtleff, A. C., Retterer, C. J., … Bavari, S. (2016a). Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature 2016 531:7594*, *531*(7594), 381–385. https://doi.org/10.1038/nature17180

Warren, T. K., Jordan, R., Lo, M. K., Ray, A. S., Mackman, R. L., Soloveva, V., Siegel, D., Perron, M., Bannister, R., Hui, H. C., Larson, N., Strickley, R., Wells, J., Stuthman, K. S., Van Tongeren, S. A., Garza, N. L., Donnelly, G., Shurtleff, A. C., Retterer, C. J., … Bavari, S. (2016b). Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature*, *531*(7594), 381–385. https://doi.org/10.1038/nature17180

Woo, P. C. Y., Lau, S. K. P., Lam, C. S. F., Lau, C. C. Y., Tsang, A. K. L., Lau, J. H. N., Bai, R., Teng, J. L. L., Tsang, C. C. C., Wang, M., Zheng, B.-J., Chan, K.-H., & Yuen, K.-Y. (2012). Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavi. *Journal of Virology*, *86*(7), 3995–4008. https://doi.org/10.1128/jvi.06540-11

World Health Organization (WHO). www.who.int

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C., & Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak

originating in Wuhan, China: a modelling study. *The Lancet*, *395*(10225), 689–697. https://doi.org/10.1016/S0140-6736(20)30260-9

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, *323*(13), 1239–1242. https://doi.org/10.1001/JAMA.2020.2648

Xia, S., Lan, Q., Su, S., Wang, X., Xu, W., Liu, Z., Zhu, Y., Wang, Q., Lu, L., & Jiang, S. (2020). The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. *Signal Transduction and Targeted Therapy 2020 5:1*, *5*(1), 1–3. https://doi.org/10.1038/s41392-020-0184-0

Yin, W., Mao, C., Luan, X., Shen, D. D., Shen, Q., Su, H., Wang, X., Zhou, F., Zhao, W., Gao, M., Chang, S., Xie, Y. C., Tian, G., Jiang, H. W., Tao, S. C., Shen, J., Jiang, Y., Jiang, H., Xu, Y., … Xu, H. E. (2020). Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science*, *368*(6498), 1499–1504. https://doi.org/10.1126/science.abc1560

Zhou, P., Yang, X. Lou, Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L., Chen, H. D., Chen, J., Luo, Y., Guo, H., Jiang, R. Di, Liu, M. Q., Chen, Y., Shen, X. R., Wang, X., … Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A Novel Coronavirus from Patients with Pneumonia in China,

2019. *Https://Doi.Org/10.1056/NEJMoa2001017*, *382*(8), 727–733. https://doi.org/10.1056/NEJMOA2001017

Zomaya, A. Y. (2006). Handbook of Nature-Inspired and Innovative Computing Integrating Classical Models with Emerging Technologies. *Implementing Neural Models in Silicon*, *May*. http://www.springerlink.com/index/10.1007/0-387-27705-6

Appendix 1: Chapter 2 supplementary material

Supplementary Table 1. The full results of Differentially conserved positions (DCPs) in Nextstrain clade.

Clade1	Clade2	Protein	DCP	Position		Conservation score		BLOSUM
				Clade1	Clade2	Clade1	Clade2	
19A	20F	Spike	S477N	477	477	0.84	0.86	1
19A	20F	NSP2	I120F	120	120	0.84	0.86	0
19A	20H	Spike	D80A	80	80	0.85	0.83	-2
19A	20H	Spike	K417N	417	414	0.84	0.86	0
19A	20H	Spike	E484K	484	481	0.84	0.84	1
19A	20H	Spike	N501Y	501	498	0.85	0.86	-2
19A	20H	Spike	A701V	701	698	0.81	0.82	0
19A	20J	Spike	T20N	20	20	0.85	0.87	0
19A	20J	Spike	R190S	190	190	0.84	0.84	-1
19A	20J	Spike	K417T	417	417	0.84	0.84	-1
19A	20J	Spike	E484K	484	484	0.84	0.84	1
19A	20J	Spike	N501Y	501	501	0.85	0.86	-2
19A	20J	Spike	T1027I	1027	1027	0.84	0.83	-1
19A	20J	NSP3	K977Q	977	977	0.84	0.86	1
19A	20J	N	P80R	80	80	0.85	0.86	-2
19A	20J	NS9b	Q77E	77	77	0.86	0.84	2
19B	20F	Spike	S477N	477	477	0.84	0.86	1
19B	20F	NSP2	I120F	120	120	0.84	0.86	0
19B	20H	Spike	D80A	80	80	0.85	0.83	-2
19B	20H	Spike	K417N	417	414	0.84	0.86	0
19B	20H	Spike	A701V	701	698	0.81	0.82	0
19B	20J	Spike	T20N	20	20	0.85	0.87	0
19B	20J	Spike	P26S	26	26	0.86	0.84	-1
19B	20J	Spike	D138Y	138	138	0.86	0.88	-3
19B	20J	Spike	K417T	417	417	0.84	0.84	-1
19B	20J	Spike	T1027I	1027	1027	0.84	0.83	-1
19B	20J	NSP3	K977Q	977	977	0.84	0.86	1
19B	20J	N	P80R	80	80	0.85	0.86	-2
19B	20J	NS9b	Q77E	77	77	0.81	0.84	2

20A	20H	Spike	K422N	417	414	0.84	0.86	0
20A	20I	N	R203K	203	203	0.82	0.83	2
20A	20J	Spike	T20N	20	20	0.84	0.87	0
20A	20J	Spike	K422T	417	417	0.84	0.84	-1
20A	20J	NSP3	K978Q	977	977	0.84	0.86	1
20A	20J	N	P80R	80	80	0.86	0.86	-2
20B	20H	Spike	D80A	80	80	0.85	0.83	-2
20C	20E	NS3	H58Q	57	57	0.86	0.85	0
20C	20E	NSP2	I85T	85	85	0.84	0.84	-1
20C	20F	NSP2	I120F	120	120	0.84	0.86	0
20C	20J	Spike	T21N	20	20	0.84	0.87	0
20C	20J	Spike	R191S	190	190	0.84	0.84	-1
20C	20J	Spike	K418T	417	417	0.84	0.84	-1
20C	20J	N	P80R	80	80	0.86	0.86	-2
20C	20J	NS9b	Q77E	77	77	0.86	0.84	2
20C	20J	NSP3	K977Q	977	977	0.84	0.86	1
20D	20F	NSP2	I120F	120	120	0.83	0.86	0
20D	20H	Spike	D80A	80	80	0.85	0.83	-2
20D	20H	Spike	K417N	417	414	0.84	0.86	0
20D	20H	Spike	N501Y	501	498	0.85	0.86	-2
20D	20I	Spike	N501Y	501	498	0.85	0.86	-2
20D	20I	Spike	T716I	716	713	0.85	0.84	-1
20D	20I	Spike	S982A	982	979	0.83	0.81	1
20D	20I	Spike	D1118H	1118	1115	0.85	0.88	-1
20D	20I	NSP5	S15G	15	15	0.84	0.82	0
20D	20J	Spike	T20N	20	20	0.84	0.87	0
20D	20J	Spike	P26S	26	26	0.86	0.84	-1
20D	20J	Spike	D138Y	138	138	0.86	0.88	-3
20D	20J	Spike	R190S	190	190	0.84	0.84	-1
20D	20J	Spike	K417T	417	417	0.84	0.84	-1
20D	20J	Spike	N501Y	501	501	0.85	0.86	-2
20D	20J	Spike	T1027I	1027	1027	0.84	0.83	-1
20D	20J	Spike	V1176F	1176	1176	0.82	0.85	-1
20D	20J	NSP3	K977Q	977	977	0.84	0.86	1
20E	20H	Spike	D80A	80	80	0.84	0.83	-2
20E	20J	Spike	T20N	20	20	0.84	0.87	0
20E	20J	Spike	R190S	190	190	0.84	0.84	-1

20E	20J	Spike	K417T	417	417	0.84	0.84	-1
20E	20J	N	P80R	80	80	0.86	0.86	-2
20E	20J	N	R203K	203	203	0.83	0.84	2
20E	20J	NS9c	V49L	49	49	0.8	0.8	1
20F	20G	NSP2	I120F	120	120	0.86	0.84	0
20F	20H	Spike	D80A	80	80	0.85	0.83	-2
20F	20H	Spike	K417N	417	414	0.84	0.86	0
20F	20H	Spike	N477S	477	474	0.86	0.84	1
20F	20H	Spike	E484K	484	481	0.84	0.84	1
20F	20H	Spike	N501Y	501	498	0.85	0.86	-2
20F	20H	Spike	A701V	701	698	0.81	0.82	0
20F	20H	NS3	Q57H	57	57	0.86	0.87	0
20F	20H	NSP2	T85I	85	85	0.85	0.84	-1
20F	20H	NSP2	F120I	120	120	0.86	0.84	0
20F	20I	Spike	N477S	477	474	0.86	0.84	1
20F	20I	Spike	N501Y	501	498	0.85	0.86	-2
20F	20I	Spike	S982A	982	979	0.83	0.81	1
20F	20I	Spike	D1118H	1118	1115	0.85	0.88	-1
20F	20J	Spike	L18F	18	18	0.82	0.87	0
20F	20J	Spike	T20N	20	20	0.85	0.87	0
20F	20J	Spike	P26S	26	26	0.86	0.84	-1
20F	20J	Spike	D138Y	138	138	0.86	0.88	-3
20F	20J	Spike	R190S	190	190	0.84	0.84	-1
20F	20J	Spike	K417T	417	417	0.84	0.84	-1
20F	20J	Spike	N477S	477	477	0.86	0.84	1
20F	20J	Spike	E484K	484	484	0.84	0.84	1
20F	20J	Spike	N501Y	501	501	0.85	0.86	-2
20F	20J	Spike	H655Y	655	655	0.87	0.86	2
20F	20J	Spike	T1027I	1027	1027	0.84	0.83	-1
20F	20J	Spike	V1176F	1176	1176	0.82	0.85	-1
20F	20J	NS8	E92K	92	92	0.85	0.85	1
20F	20J	NSP2	F120I	120	120	0.86	0.84	0
20F	20J	NSP3	K977Q	977	977	0.84	0.86	1
20F	20J	NSP13	E341D	341	341	0.84	0.85	2
20G	20H	Spike	D80A	80	80	0.85	0.83	-2
20G	20H	Spike	K417N	417	414	0.84	0.86	0
20G	20H	Spike	E484K	484	481	0.83	0.84	1

20G	20I	Spike	S984A	982	979	0.83	0.81	1
20G	20I	Spike	D1120H	1118	1115	0.85	0.88	-1
20G	20I	N	G204R	204	204	0.82	0.84	-2
20G	20I	N	S235F	235	235	0.83	0.85	-2
20G	20J	Spike	T20N	20	20	0.84	0.87	0
20G	20J	Spike	R190S	190	190	0.84	0.84	-1
20G	20J	Spike	K417T	417	417	0.84	0.84	-1
20G	20J	Spike	E484K	484	484	0.83	0.84	1
20G	20J	Spike	T1029I	1027	1027	0.84	0.83	-1
20G	20J	Spike	V1178F	1176	1176	0.82	0.85	-1
20G	20J	NSP3	K982Q	977	977	0.83	0.86	1
20G	20J	N	P80R	80	80	0.86	0.86	-2
20G	20J	N	G204R	204	204	0.82	0.85	-2
20G	20J	NS8	E92K	92	92	0.85	0.85	1
20G	20J	NS9b	Q77E	77	77	0.86	0.84	2
20G	20J	NS9c	V49L	49	49	0.81	0.8	1
20H	20I	Spike	A80D	80	78	0.83	0.85	-2
20H	20I	Spike	N417K	414	414	0.86	0.84	0
20H	20I	Spike	T716I	713	713	0.85	0.84	-1
20H	20I	Spike	S982A	979	979	0.83	0.81	1
20H	20I	Spike	D1118H	1115	1115	0.85	0.88	-1
20H	20J	Spike	T20N	20	20	0.84	0.87	0
20H	20J	Spike	P26S	26	26	0.86	0.84	-1
20H	20J	Spike	A80D	80	80	0.83	0.85	-2
20H	20J	Spike	D138Y	138	138	0.86	0.88	-3
20H	20J	Spike	R190S	190	190	0.84	0.84	-1
20H	20J	Spike	N417T	414	417	0.86	0.84	0
20H	20J	Spike	H655Y	652	655	0.87	0.86	2
20H	20J	Spike	V701A	698	701	0.82	0.81	0
20H	20J	Spike	T1027I	1024	1027	0.84	0.83	-1
20H	20J	Spike	V1176F	1173	1176	0.82	0.85	-1
20H	20J	NSP3	N837K	837	837	0.88	0.86	0
20H	20J	NSP3	K977Q	977	977	0.84	0.86	1
20H	20J	NSP13	E341D	341	341	0.84	0.85	2
20H	20J	N	P80R	80	80	0.86	0.86	-2
20H	20J	N	R203K	203	203	0.84	0.84	2
20H	20J	N	G204R	204	204	0.82	0.85	-2

20H	20J	N	I205T	205	205	0.83	0.84	-1
20H	20J	NS3	H57Q	57	57	0.87	0.86	0
20H	20J	NS3	S253P	253	253	0.83	0.94	-1
20H	20J	NS8	E92K	92	92	0.85	0.85	1
20H	20J	NS9b	Q77E	77	77	0.86	0.84	2
20I	20J	Spike	T20N	20	20	0.84	0.87	0
20I	20J	Spike	K417T	414	417	0.84	0.84	-1
20I	20J	Spike	I716T	713	716	0.84	0.85	-1
20I	20J	Spike	A982S	979	982	0.81	0.83	1
20I	20J	Spike	T1027I	1024	1027	0.84	0.83	-1
20I	20J	Spike	H1118D	1115	1118	0.88	0.85	-1
20I	20J	Spike	V1176F	1173	1176	0.82	0.85	-1
20I	20J	NSP3	K977Q	977	977	0.84	0.86	1
20I	20J	N	P80R	80	80	0.86	0.86	-2
20I	20J	N	F235S	235	235	0.85	0.84	-2
20I	20J	NS8	E92K	92	92	0.85	0.85	1
20I	20J	NS9b	Q77E	77	77	0.86	0.84	2

Differences of frequencies of amino acid in GISAID clade

In the frequency analysis of GISAID clade, we identified six positions in four proteins that the most frequent amino acid is different in any of the clade; R203K and A220V in N protein, G54L and L71F in NS9c, T85I in NSP2, P323L in NSP12.

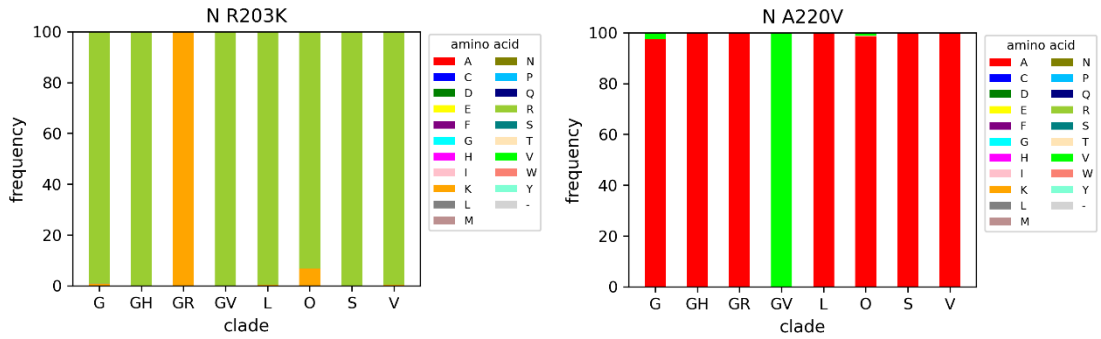
N protein

In the N protein, amino acid positions 203 and 220 demonstrated different frequencies of specific amino acids (Supplementary Figure 1). Position 203 was arginine in most clades, whereas it was predominantly lysine in the GR clade (Supplementary Table 2). In addition, position 220 in N protein was alanine in most clades. However, valine was the predominant amino acid in the GV clade. These positions were not DCPs because the amino acids present were shared between the two groups, even if the predominant

amino acids are different. For example, for position 203, lysine is predominant in clade GR, but there were 11 sequences out of 134691 sequences that were arginine.

Supplementary Table 2. The most frequent amino acid in each clade of N protein. Amino acids are described in one-letter code.

Clade / Position	L	S	V	O	G	GH	GR	GV
203	R	R	R	R	R	R	K	R
220	A	A	A	A	A	A	A	V



Supplementary Figure 1. Frequencies of amino acid of position 203 and 220 in N protein of each clade. The frequencies are described in percentage. Most of the clades in position 203 have arginine as the most frequent amino acid, but lysine is predominant in clade GR. Alanine is the most frequent amino acid in position 220, but valine is predominant in clade GV.

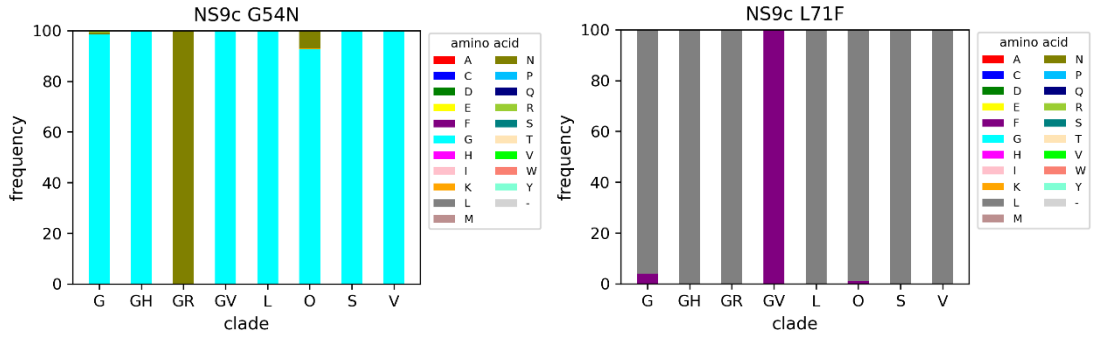
NS9c

In NS9c, residue 54 was glycine in most clades, but valine was the most frequent amino acid in the GR clade. Moreover, leucine was the predominant amino acid in most clades at residue 71. However, phenylalanine was the most frequent amino acid at this position of NS9c in clade GV (Supplementary Table 3). Clade G also had approximately 5% of phenylalanine in position 71 (Supplementary Figure 2).

Supplementary Table 3. The most frequent amino acid in each clade of NS9c. Amino acids are

described in one-letter code.

Clade / Position	L	S	V	O	G	GH	GR	GV
54	G	G	G	G	G	G	N	G
71	L	L	L	L	L	L	L	F



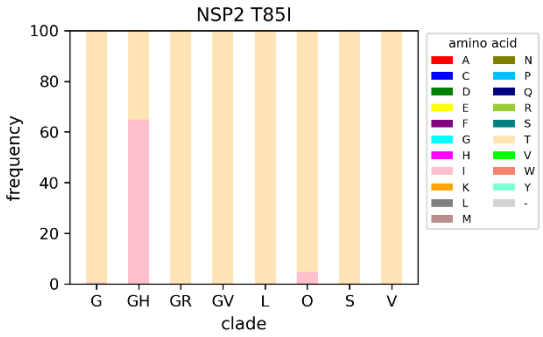
Supplementary Figure 2. Frequencies of amino acid of position 54 and 71 in NS9c protein of each clade. The frequencies are described in percentage. Most of the clades in position 54 have proline as the most frequent amino acid, but asparagine is predominant in clade GR. Lysine is the most frequent amino acid in position 71, but phenylalanine is predominant in clade GV.

NSP2

In NSP2, only one position was different when comparing clades, which was residue 85. Threonine was the predominant amino acid in most clades, whereas tyrosine was the most frequent in the GH clade (Supplementary Table 4). However, tyrosine was approximately 65% of the clade GH, and 35% was threonine, which is the predominant amino acid at this position (Supplementary Figure 3).

Supplementary Table 4. The most frequent amino acid in each clade of NSP2. Amino acids are described in one-letter code.

Clade / Position	L	S	V	O	G	GH	GR	GV
85	T	T	T	T	T	I	T	T



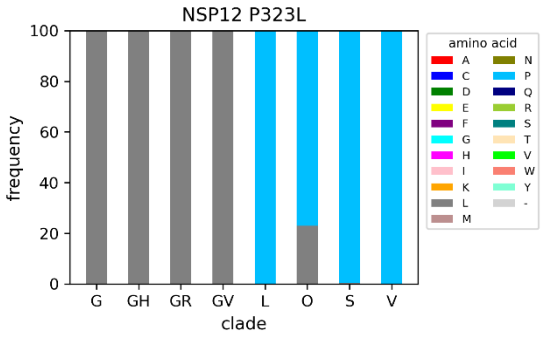
Supplementary Figure 3. Frequencies of amino acid of position 85 in NSP2 protein of each clade. The frequencies are described in percentage. In position 85 of NSP2, threonine is the predominant amino acid in most of the clades. Clade GH has Isoleucine for approximately 60% of frequencies, and the rest is threonine.

NSP12

In NSP12, the amino acid position 323 was proline in clades L, S, V, and O, whereas the most frequent amino acid was leucine in clade G, GH, GR, and GV (Supplementary Table 5). The frequencies were nearly 100% in each of the clades except clade O (Supplementary Figure 4). The clade O represents 'others', as this clade can be considered mixed clades. Therefore, position 323 may be substituted between clade L and clade G because clade G is derived from clade L.

Supplementary Table 5. The most frequent amino acid in each clade of NSP12. Amino acids are described in one-letter code.

Clade / Position	L	S	V	O	G	GH	GR	GV
323	P	P	P	P	L	L	L	L



Supplementary Figure 4. Frequencies of amino acid of position 323 in NSP12 protein of each clade. Clade G, GH, GR, and GV has leucine as the most frequent amino acid. Proline is the most frequent amino acid for clade L, O, S, and V.

Differences of frequencies of amino acid in Nextstrain clade

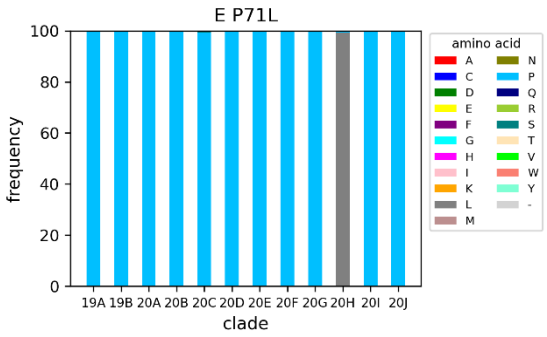
We performed the analysis in Nextstrain clade in the same way as the GISAID clade. The differences of amino acids were identified in 11 proteins: E, N, NS3, NS8, NS9b, NS9c, NSP3, NSP5, NSP6, NSP13, and the spike protein.

E protein

In E, residue 71 was identified as a difference between clades (Supplementary Table 6). Proline was the most abundant amino acid in all clades except the 20H clade, where leucine is the most frequent amino acid. In other clades, proline had approximately 100% of the frequencies (Supplementary Figure 5).

Supplementary Table 6. The most frequent amino acid in each clade of E protein. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
71	P	P	P	P	P	P	P	P	P	L	P	P



Supplementary Figure 5. Frequencies of amino acid of position 71 in E protein of each clade. Leucine is the most frequent amino acid in clade 20H. In other clades, proline is the predominant amino acid in position 71.

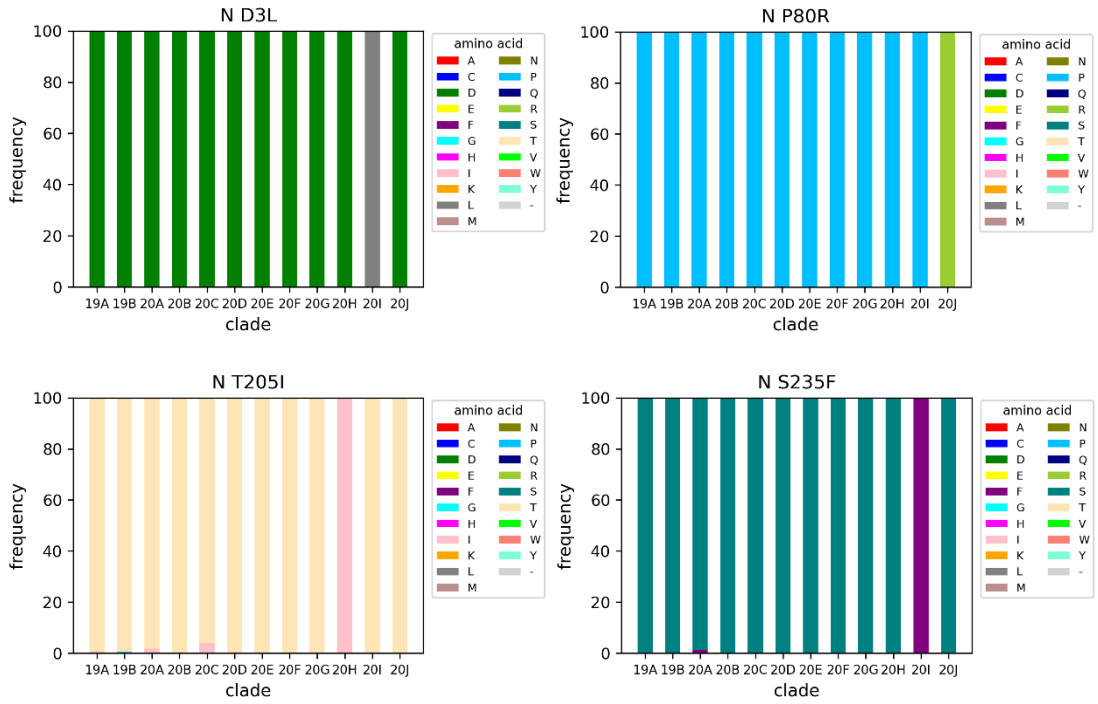
N protein

Four positions were identified in N, where clades had different frequencies of amino acids (Supplementary Table 7). At residue 3, clade 20I/501Y.V1 had mostly leucine, and other clades had aspartic acid. Although these clades have almost different amino acids, which could be a DCP, a few sequences shared the same amino acids between the two groups, resulting in not considering as a DCP. In addition, proline was the predominant amino acid except clade 20J/501Y.V3 at position 80. On the other hand, arginine was the most frequent amino acid in clade 20J/501Y.V3. Moreover, in position 205 of N, clade 20H/501Y.V2 had isoleucine which is different from the other clades with threonine. Clade 19A, 20A, and 20C also had isoleucine for a small percentage of sequences (Supplementary Figure 6). Furthermore, position 235 was identified as well. This position in N was serine in most clades, but it was substituted to phenylalanine in clade 20I/501Y.V1.

Supplementary Table 7. The most frequent amino acid in each clade of N protein. Amino acids are described in one letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
3	D	D	D	D	D	D	D	D	D	D	L	D

80	P	P	P	P	P	P	P	P	P	P	P	R
205	T	T	T	T	T	T	T	T	T	I	T	T
235	S	S	S	S	S	S	S	S	S	S	F	S



Supplementary Figure 6. Frequencies of amino acid of position 3, 80, 205, and 235 in N protein of each clade. The frequency is described in percentage. For amino acid position 3, leucine is the most frequent amino acid in clade 20I, and aspartic acid is the most frequent amino acid for other clades. In position 80, arginine has approximately 100% of the frequency in clade 20J. For position 205, although threonine is predominant in most clades, isoleucine is approximately 100% frequent in clade 20H and clade 20A and 20C has also isoleucine for a few percent. For position 235, phenylalanine is the most frequent amino acid in clade 20I, but other clades have serine as the most frequent amino acid.

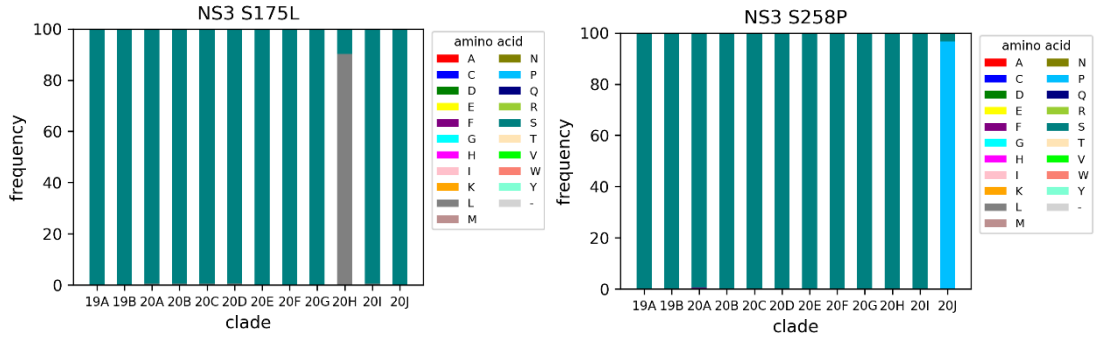
NS3

In NS3, two positions (175 and 258) were identified as having amino acid differences between clades (Supplementary Table 8). In most clades, serine was the predominant amino acid at position 175, but leucine was the most frequent amino acid in clade 20H/501Y.V2. However, approximately 10% of clade 20H/501Y.V2 had serine in this

position as well. In addition, serine was also the most frequent amino acid at position 258 in most clades. Clade 20J/501Y.V3 had proline in more than 90% of the sequences in the position (Supplementary Figure 7).

Supplementary Table 8. The most frequent amino acid in each clade of NS3. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
175	S	S	S	S	S	S	S	S	S	L	S	S
258	S	S	S	S	S	S	S	S	S	S	S	P



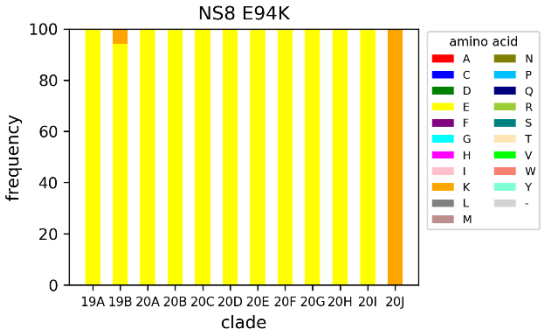
Supplementary Figure 7. Frequencies of amino acid of position 175 and 258 in NS3 protein of each clade.

NS8

In NS8, an amino acid change was observed at amino acid position 94 (Supplementary Table 9). The most frequent amino acid, which was glutamate, was replaced by lysine in clade 20J/501Y.V3. In addition, lysine accounted for approximately 5% of clade 19B (Supplementary Figure 8). Since clade 19B and clade 20J/501Y.V3 is not a relatively close clade, this might suggest that this substitution occurs frequently.

Supplementary Table 9. The most frequent amino acid in each clade of NS8. Amino acids are described in one letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
94	E	E	E	E	E	E	E	E	E	E	E	K



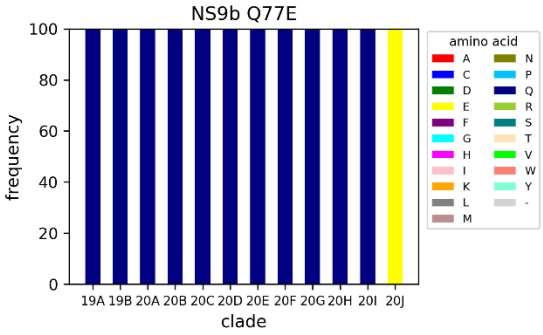
Supplementary Figure 8. Frequencies of amino acid of position 94 in NS8 protein of each clade.

NS9b

Only one position was identified in NS9b, which was the amino acid position 77 (Supplementary Table 10). Glutamine was the predominant amino acid in most of the clades. However, in clade 20J/501Y.V2, glutamic acid accounted for 100% of the clade (Supplementary Figure 9). Thus, this position was a DCP in the comparison of clade 20J/501Y.V2 and clade 19A, 19B, 20C, 20G, 20H/501Y.V2, and 20I/501Y.V1. However, in comparison between clade 20J/501Y.V2 and clades 20A, 20B, 20D, 20E, and 20F were not considered DCP because a few sequences in these clades presented glutamic acid in position 77.

Supplementary Table 10. The most frequent amino acid in each clade of NS9b. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
77	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	E



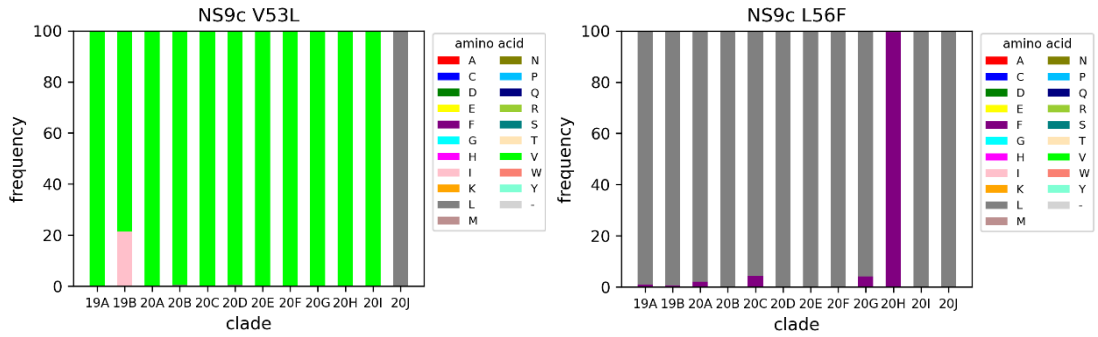
Supplementary Figure 9. Frequencies of amino acid of position 77 in NS9b protein of each clade.

NS9c

By analysing NS9c, two positions were found with different amino acid frequencies (Supplementary Table 11). In position 53, valine was substituted to leucine in clade 20J/501Y.V3. In the analysis of DCPs, this position was considered as V49L when comparing 20J/501Y.V3 and 20E, 20J/501Y.V3 and 20G, respectively. However, when comparing between 20J/501Y.V3 and clades other than 20E and 20G, it was not a DCP because other clades had leucine, the same amino acid as 20J/501Y.V3, in a few sequences. Interestingly, clade 19B had isoleucine in this position for more than 20% of the sequences. In addition, L56F is another difference in NS9c. Nearly 100% of the 20H/501Y.V2 clade had phenylalanine in this position. Clade 19A, 20A, 20C, and 20G also had phenylalanine in more than 1% of the sequences (Supplementary Figure 10).

Supplementary Table 11. The most frequent amino acid in each clade of NS9c. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
53	V	V	V	V	V	V	V	V	V	V	V	L
56	L	L	L	L	L	L	L	L	L	F	L	L



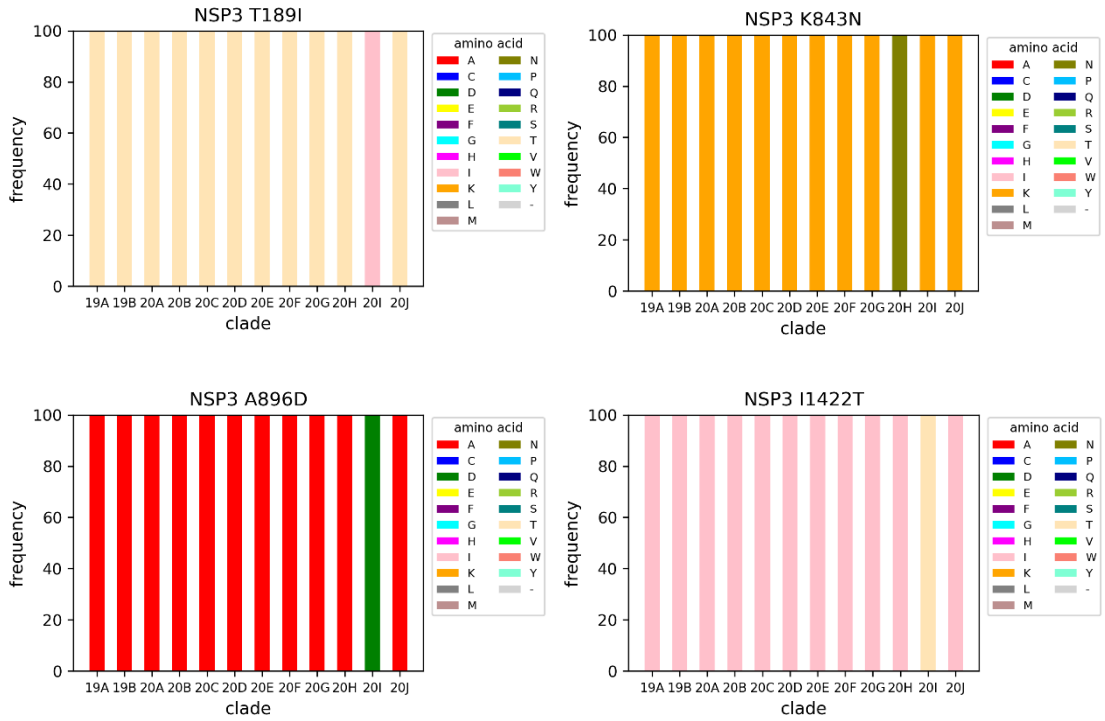
Supplementary Figure 10. Frequencies of amino acid of position 53 and 56 in NS9c protein of each clade.

NSP3

NSP3 had four positions that had the differences of most frequent amino acids between clades (Supplementary Table 12). Three of the positions were different in clade 20I/501Y.V1, which were T189I, A896D, and I1422T. Also, for K843N, 20H/501Y.V2 was different from the other clades and had lysine of approximately 100% frequencies (Supplementary Figure 11). These positions were not DCPs, according to the DCP analysis.

Supplementary Table 12. The most frequent amino acid in each clade of NSP3. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
189	T	T	T	T	T	T	T	T	T	T	I	T
843	K	K	K	K	K	K	K	K	K	N	K	K
896	A	A	A	A	A	A	A	A	A	A	D	A
1422	I	I	I	I	I	I	I	I	I	I	T	I



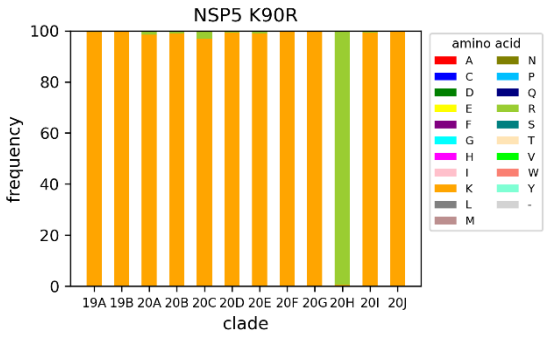
Supplementary Figure 11. Frequencies of amino acid of position 189, 843, 896, and 1422 in NSP3 protein of each clade.

NSP5

In NSP5, only one position had a difference between clades, which was position 90 (Supplementary Table 13). Lysine was substituted to arginine in most sequences in clade 20H/501Y.V2. However, other clades also included some sequences with arginine in amino acid position 90 (Supplementary Figure 12). Thus position 90 was not considered as a DCP.

Supplementary Table 13. The most frequent amino acid in each clade of NSP5. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
90	K	K	K	K	K	K	K	K	K	R	K	K



Supplementary Figure 12. Frequencies of amino acid of position 90 in NSP5 protein of each clade.

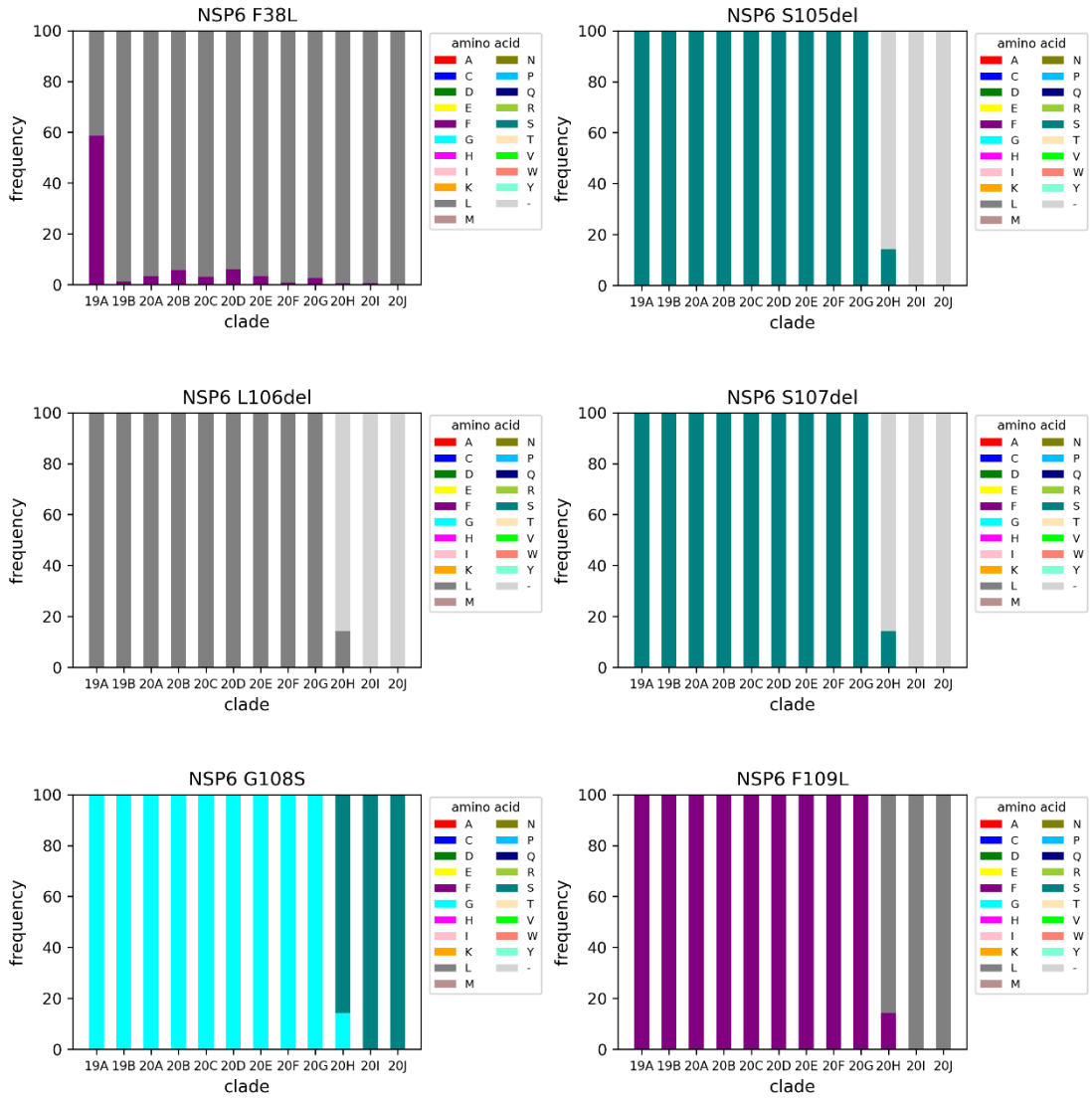
NSP6

In NSP6, six amino acid positions were identified by the analysis. Position 38, 105, 106, 107, 108, and 109 were the identified positions (Supplementary Table 14). In position 38 of clade 19A, approximately 60% of the sequences were phenylalanine, whereas the rest of the sequences were mostly leucine which is the most frequent amino acid in other clades (Supplementary Figure 13). Other clades also included phenylalanine in this position, suggesting that these amino acids in this position are not significant to the pathogenicity of the virus. In addition, clade 20H/501Y.V2, 20I/501Y.V1, and 20J/501Y.V3 had a deletion in positions 105, 106, and 107. Moreover, clade 20H/501Y.V2, 20I/501Y.V1, and 20J/501Y.V3 had a different predominant amino acid compared with the other clades in the positions 108 and 109.

Supplementary Table 14. The most frequent amino acid in each clade of NSP6. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
38	F	L	L	L	L	L	L	L	L	L	L	L
105	S	S	S	S	S	S	S	S	S	-	-	-
106	L	L	L	L	L	L	L	L	L	-	-	-
107	S	S	S	S	S	S	S	S	S	-	-	-

108	G	G	G	G	G	G	G	G	G	S	S	S
109	F	F	F	F	F	F	F	F	F	L	L	L



Supplementary Figure 13. Frequencies of amino acid of position 38, 105, 106, 107, 108, and 109 in NSP6 protein of each clade.

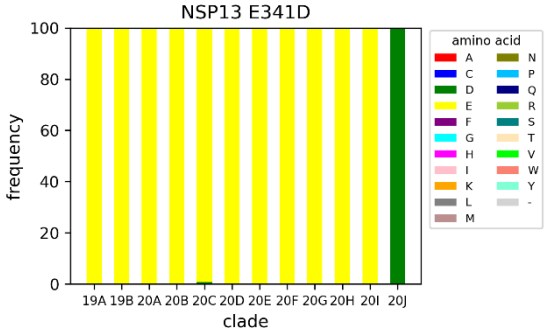
NSP13

In NSP13, we found one position with different predominant amino acids compared with other clades, which was position 341 (Supplementary Table 15). In addition, clade 20C has glutamic acid for more than 1% of frequency. Moreover, throughout the DCP analysis, comparison of 20F and 20J/501Y.V3, 20H and 20J/501Y.V3 were considered

as DCP, respectively. However, other clades also had aspartic acid instead of glutamic acid in this position in some sequences of the clades. Thus, those comparisons other than 20F and 20J/501Y.V3, 20H and 20J/501Y.V3 were not considered as DCP.

Supplementary Table 15. The most frequent amino acid in each clade of NSP13. Amino acids are described in one-letter code.

Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
341	E	E	E	E	E	E	E	E	E	E	E	D



Supplementary Figure 14. Frequencies of amino acid of position 341 in NSP13 protein of each clade.

Spike protein

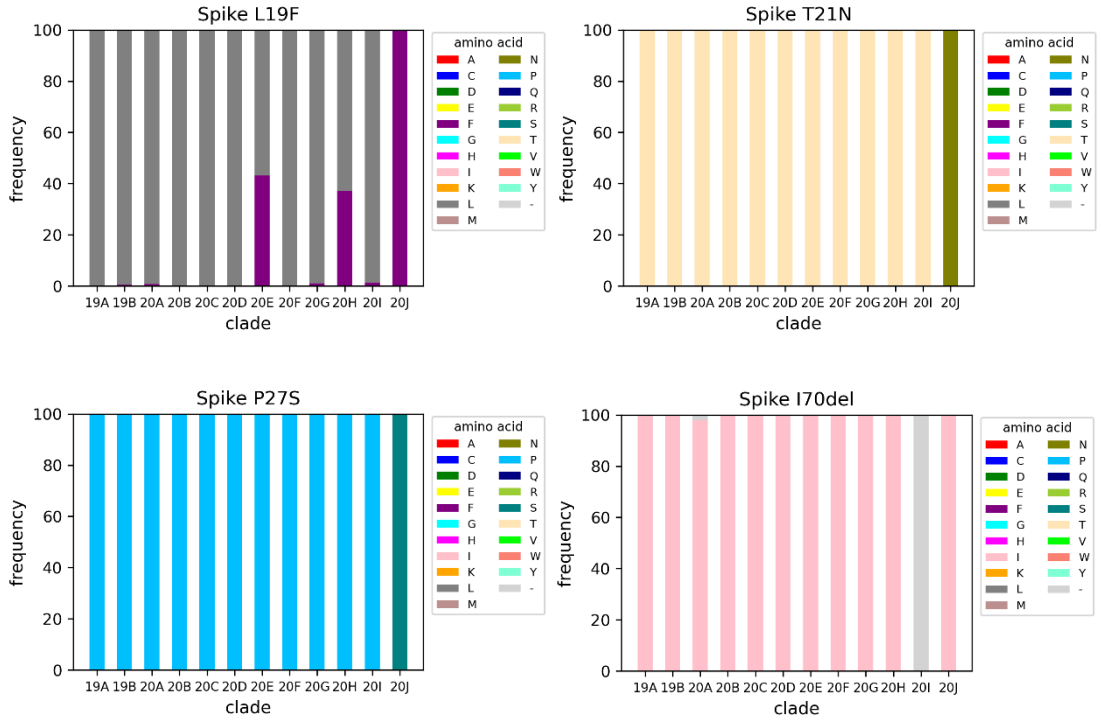
In the spike protein, 24 positions were identified as the differences between clades, which were positions 19, 21, 27, 70, 71, 72, 82, 140, 146, 192, 220, 246, 247, 248, 424, 579, 664, 690, 710, 725, 991, 1036, 1127, and 1185 (Supplementary Table 16). In position 19, clade 20J/501Y.V3 had 100% phenylalanine, and clade 20E and 20H/501Y.V2 also had phenylalanine for approximately 40% in the clades (Supplementary Figure 15). Interestingly, clade 20E, 20H/501Y.V2, and 20J/501Y.V3 are not genetically close clades according to the phylogenetic tree. Thus, the substitution of leucine to phenylalanine in amino acid position 19 may occur frequently.

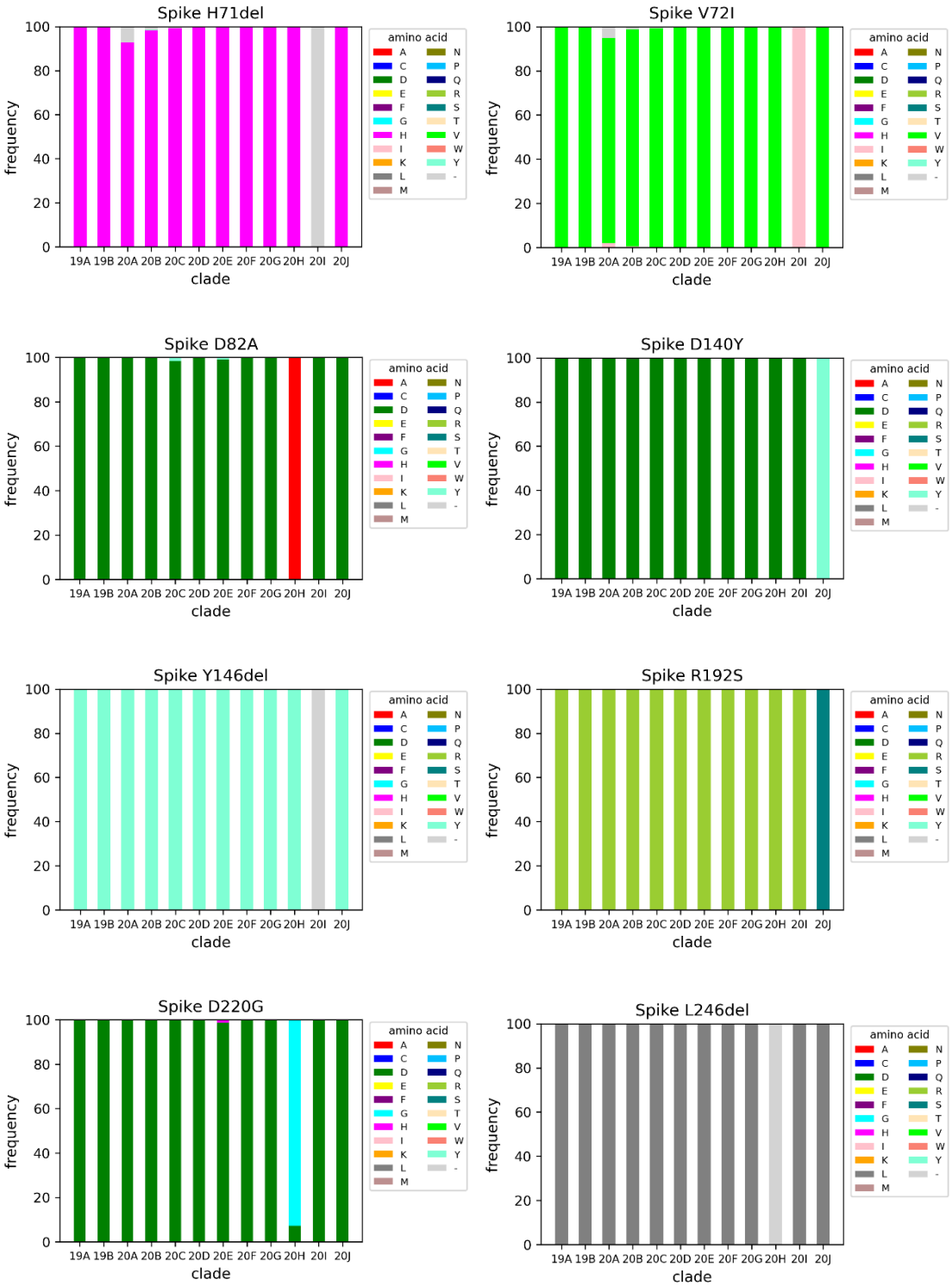
In addition, positions 21, 27, 140, 192, 664, 1036, and 1185 had different amino acids in clade 20J/501Y.V3. Moreover, in clade 20I/501Y.V1, the most frequent amino acid in positions 70, 71, 72, 146, 579, 690, 725, and 1127 was different from other clades. Furthermore, in positions 246, 247, and 248, the clade 20H/501Y.V2 had a deletion instead of leucine or alanine as the predominant amino acid. Interestingly, in position 424, most of the clades had lysine as the most frequent amino acid, whereas clade 20H/501Y.V2 and 20J/501Y.V3 had asparagine and threonine, respectively.

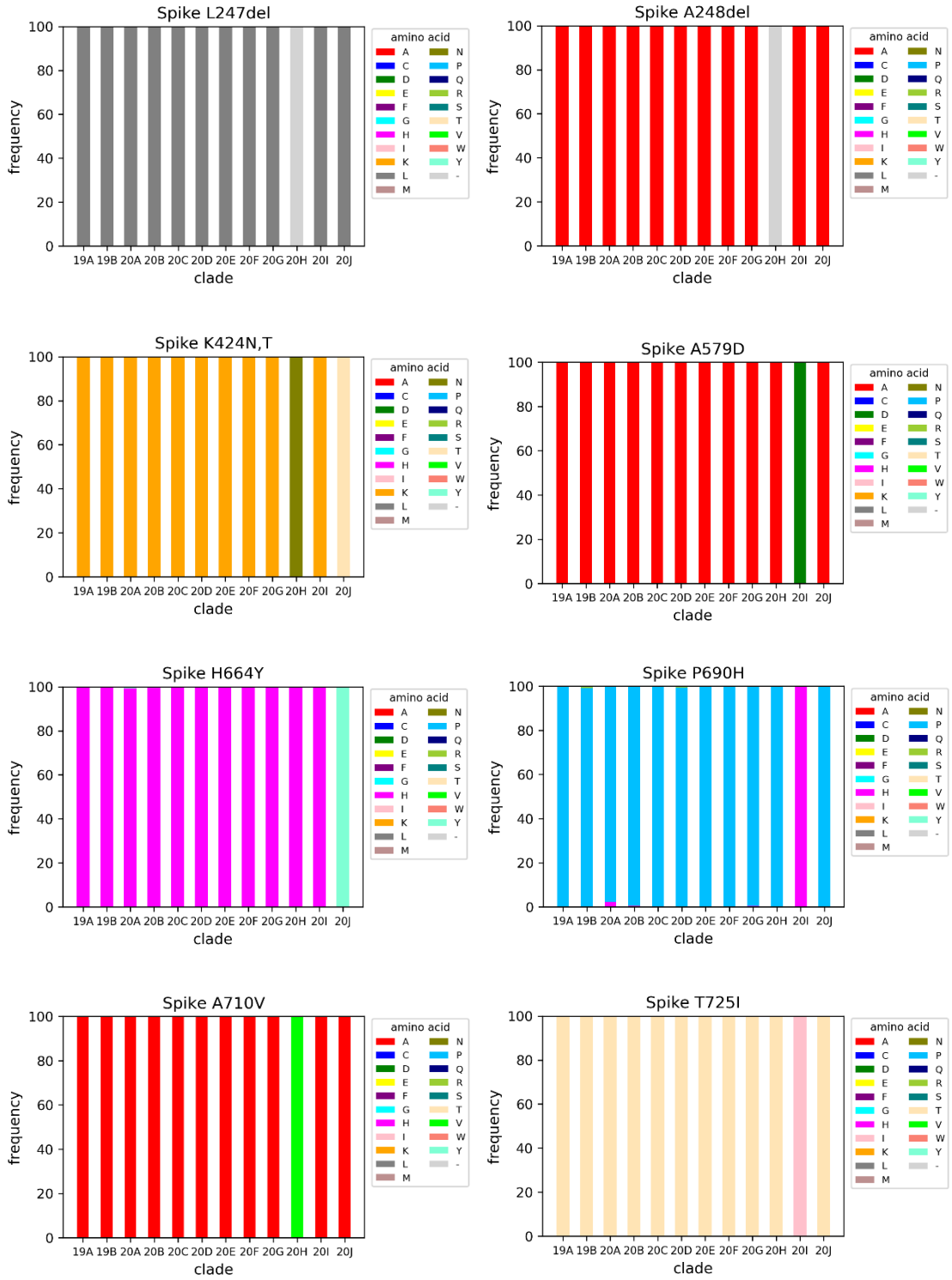
Supplementary Table 16. The most frequent amino acid in each clade of Spike protein. Amino acids are described in one-letter code.

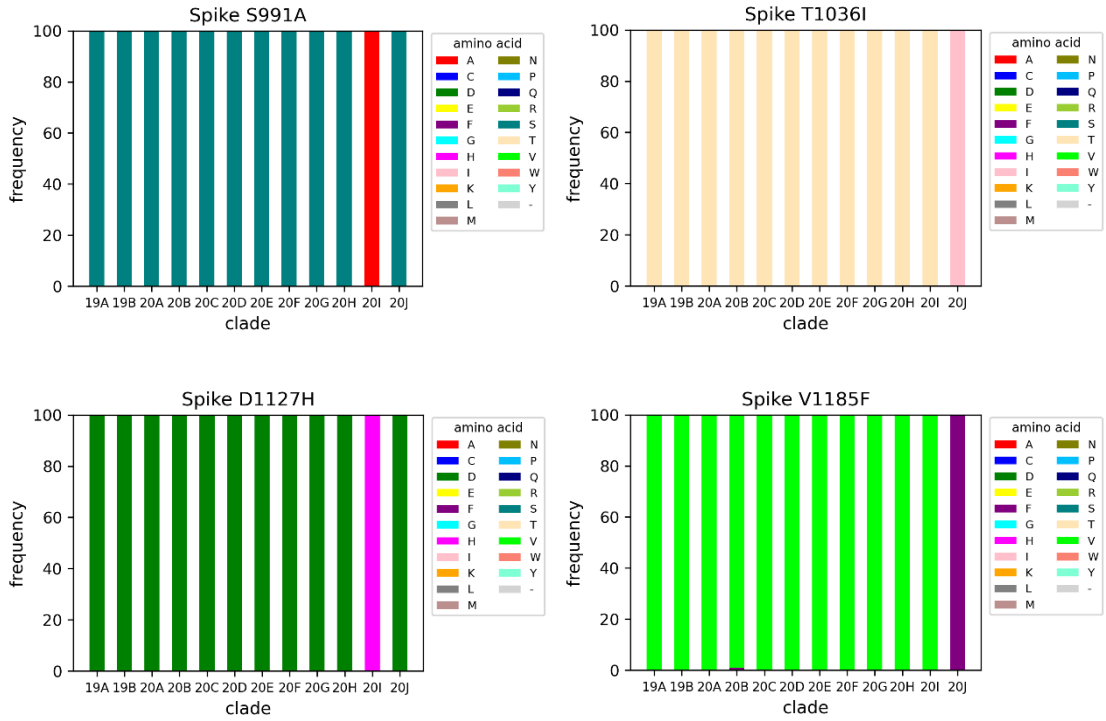
Clade / Position	19A	19B	20A	20B	20C	20D	20E	20F	20G	20H	20I	20J
19	L	L	L	L	L	L	L	L	L	L	L	F
21	T	T	T	T	T	T	T	T	T	T	T	N
27	P	P	P	P	P	P	P	P	P	P	P	S
70	I	I	I	I	I	I	I	I	I	I	-	I
71	H	H	H	H	H	H	H	H	H	H	-	H
72	V	V	V	V	V	V	V	V	V	V	I	V
82	D	D	D	D	D	D	D	D	D	A	D	D
140	D	D	D	D	D	D	D	D	D	D	D	Y
146	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
192	R	R	R	R	R	R	R	R	R	R	R	S
220	D	D	D	D	D	D	D	D	D	G	D	D
246	L	L	L	L	L	L	L	L	L	-	L	L
247	L	L	L	L	L	L	L	L	L	-	L	L
248	A	A	A	A	A	A	A	A	A	-	A	A

424	K	K	K	K	K	K	K	K	K	N	K	T
579	A	A	A	A	A	A	A	A	A	A	D	A
664	H	H	H	H	H	H	H	H	H	H	H	Y
690	P	P	P	P	P	P	P	P	P	P	H	P
710	A	A	A	A	A	A	A	A	A	V	A	A
725	T	T	T	T	T	T	T	T	T	T	I	T
991	S	S	S	S	S	S	S	S	S	S	A	S
1036	T	T	T	T	T	T	T	T	T	T	T	I
1127	D	D	D	D	D	D	D	D	D	D	H	D
1185	V	V	V	V	V	V	V	V	V	V	V	F




Spike H71del
Spike V72I
Spike D82A
Spike D140Y
Spike Y146del
Spike R192S
Spike D220G
Spike L246del
frequency
clade
amino acid
A C D E F G H I K L M N P Q R S T V W Y -
19A 19B 20A 20B 20C 20D 20E 20F 20G 20H 20I 20J


Spike L247del
Spike A248del
Spike K424N,T
Spike A579D
Spike H664Y
Spike P690H
Spike A710V
Spike T725I
frequency
clade
19A 19B 20A 20B 20C 20D 20E 20F 20G 20H 20I 20J
amino acid
A C D E F G H I K L M N P Q R S T V W Y -



Supplementary Figure 15. Frequencies of amino acid of position 19, 21, 27, 70, 71, 72, 82, 140, 146, 192, 220, 246, 247, 248, 424, 579, 664, 690, 710, 725, 991, 1036, 1127, and 1185 in Spike protein of each clade.

Appendix 2: Chapter 3 supplementary material

The full results of the sequence analysis are shown in this section.

Supplementary Table 17. Comparison between original control and $FFM3_{LOW}$. The percentages are the ratio of the nucleotides in the reads. Bases were called if one nucleotide was present in >90% of the reads at that position. The percentages of amino acid substitution correspond to the percentages of the base. (aa) represents amino acid.

Position	Original control	$FFM3_{LOW}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{LOW}$ (aa)
510	G	(G: 86.1621, del: 13.5361)	NSP1	82	G	(G, V(frameshift))
511	T	(T: 85.7487, del: 14.0584)	NSP1	82	G	(G, V(frameshift))
512	C	(C: 85.9727, del: 13.9674)	NSP1	83	H	(H, del)
513	A	(A: 86.2357, del: 13.6253)	NSP1	83	H	(H, del)
514	T	(T: 80.8193, del: 19.1015)	NSP1	83	H	(H, del)
515	G	(G: 80.9692, del: 18.9535)	NSP1	84	V	(V, del)
516	T	(T: 78.9138, del: 20.6612)	NSP1	84	V	(V, del)
517	T	(T: 79.7661, del: 20.0504)	NSP1	84	V	(V, del)
518	A	(A: 78.6081, del: 20.8314)	NSP1	85	M	(M, del)
519	T	(T: 87.6773, del: 12.1702)	NSP1	85	M	(M, del)
520	G	(G: 87.9632, del: 11.9448)	NSP1	85	M	(M, del)
521	G	(G: 89.3565, del: 10.5973)	NSP1	86	V	(V, V(frameshift))
522	T	(T: 88.535, del: 11.1489)	NSP1	86	V	(V, V(frameshift))
11083	G	(G: 88.2997, T: 9.59596, del: 2.10438)	NSP6	37	L	(L, F, F(frameshift))
11750	C	(C: 80.6886, T: 19.2116)	NSP6	260	L	(L, F)
11916	C	(C: 82.8148, T: 17.1313)	NSP7	25	S	(S, L)
20480	C	(C: 88.9352, T: 11.0648)	NSP15	287	S	(S, L)
20573	T	(T: 89.3473, C: 10.6527)	NSP15	318	V	(V, A)

21789	C	T	Spike	76	T	I
22264	C	(C: 86.3636, T: 13.6364)	Spike	234	N	N
27131	C	(C: 69.5341, T: 30.0717)	M	203	N	N
28649	A	T	N	126	N	Y

Supplementary Table 18. Comparison between original control and $FFM7_{LOW}$.

Position	Original control	$FFM7_{LOW}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM7_{LOW}$ (aa)
44	C	(T: 52.8866, C: 47.0169999999999)	-	-	-	-
835	C	(T: 54.1016, C: 45.5078)	NSP2	10	F	F
2509	C	(C: 89.6681, T: 10.3029999999999)	NSP2	568	P	P
9298	C	(C: 89.6388, T: 10.3612)	NSP4	248	Y	Y
11399	A	(G: 51.3368, A: 48.6425)	NSP6	143	M	(V, M)
16616	C	(C: 78.4488, T: 21.5418)	NSP13	127	T	(T, I)
21789	C	(T: 51.076, C: 48.924)	Spike	76	T	(I, T)
22032	T	(T: 60.7193, C: 39.1396)	Spike	157	F	(F, S)
23179	(C: 56.6181, T: 43.2826)	(T: 87.1688, C: 12.7615)	Spike	539	V	V
23271	C	(T: 85.0681, C: 14.9319)	Spike	570	A	(V, A)
23280	C	(C: 87.2469, T: 12.6554)	Spike	573	T	(T, I)
24130	C	(C: 88.4389, T: 11.1039)	Spike	856	N	N
27585	T	(T: 87.8068, G: 11.3347)	ORF7a	64	A	A
28311	C	(T: 67.3919, C: 32.5791)	N	13	P	(L, P)
28899	G	(G: 89.7273, T: 9.89471)	N	209	R	(R, I)

Supplementary Table 19. Comparison between original control and $FFM3_{HIGH}$

Position	Original control	$FFM3_{LOW}$	$FFM3_{HIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{LOW}$ (aa)	$FFM3_{HIGH}$ (aa)
508	T	T	(del: 52.7194, T: 47.1872)	NSP1	81	H	H	H

509	G	G	(del: 53.301, G: 46.6304)	NSP1	82	G	G	(del, G)
510	G	(G: 86.1621, del: 13.5361)	(del: 65.7674, G: 33.7306)	NSP1	82	G	(G, V(frameshift))	(del, G)
511	T	(T: 85.7487, del: 14.0584)	(del: 66.3943, G: 33.4024)	NSP1	82	G	(G, V(frameshift))	(del, G)
512	C	(C: 85.9727, del: 13.9674)	(del: 66.8244, C: 33.1282)	NSP1	83	H	(H, del)	(del, H)
513	A	(A: 86.2357, del: 13.6253)	(del: 66.7544, A: 33.1499)	NSP1	83	H	(H, del)	(del, H)
514	T	(T: 80.8193, del: 19.1015)	(del: 66.6022, T: 33.268)	NSP1	83	H	(H, del)	(del, H)
515	G	(G: 80.9692, del: 18.9535)	(del: 66.7297, G: 33.236)	NSP1	84	V	(V, del)	(del, V)
516	T	(T: 78.9138, del: 20.6612)	(del: 67.4066, T: 32.4898)	NSP1	84	V	(V, del)	(del, V)
517	T	(T: 79.7661, del: 20.0504)	(del: 67.0348, T: 32.9092)	NSP1	84	V	(V, del)	(del, V)
518	A	(A: 78.6081, del: 20.8314)	(del: 67.3911, A: 32.4729)	NSP1	85	M	(M, del)	(del, M)
519	T	(T: 87.6773, del: 12.1702)	(del: 67.1115, T: 32.8006)	NSP1	85	M	(M, del)	(del, M)
520	G	(G: 87.9632, del: 11.9448)	(del: 66.958, G: 32.9531)	NSP1	85	M	(M, del)	(del, M)
521	G	(G: 89.3565, del: 10.5973)	(del: 65.7345, G: 34.2222)	NSP1	86	V	(V, V(frameshift))	(H, V)
522	T	(T: 88.535, del: 11.1489)	(del: 65.3789999999999, T: 33.7467)	NSP1	86	V	(V, V(frameshift))	(H, V)
1288	C	C	(C: 84.1061, T: 15.8384)	NSP2	161	C	C	C
2062	C	C	(C: 84.1517, T: 15.8314)	NSP2	419	A	A	A
5962	T	T	(T: 81.1249)	NSP3	1081	Y	Y	(Y)

6045	A	A	(A: 89.5812, del: 10.276)	NSP3	1109	N	N	(N, I)
6046	T	T	(T: 89.0053, del: 10.7702)	NSP3	1109	N	N	(N, I)
6047	T	T	(T: 88.7258, del: 10.6844)	NSP3	1110	F	F	(F, I)
7521	C	C	(C: 53.9715, T: 46.0285)	NSP3	1601	T	T	(T, I)
8290	C	C	(C: 86.141, T: 13.757)	NSP3	1857	L	L	L
11760	A	A	(A: 68.9241, G: 31.0069999999999)	NSP6	263	K	K	(K, R)
12016	T	T	(T: 86.5919, G: 13.157)	NSP7	58	V	V	V
12334	(A: 89.0675, G:4.91803, T:4.31126)	(A: 89.6299, G: 4.64974, T: 3.84419)	A	NSP8	81	A	A	A
14408	T	T	(del: 68.3007, T: 31.2949)	NSP12	323	P	P	(L(frameshift), P)
14409	T	T	(del: 68.5469, T: 31.4461)	NSP12	323	P	P	(L(frameshift), P)
14410	A	A	(del: 68.9873, A: 30.9636)	NSP12	324	T	T	(del, T)
14411	C	C	(del: 69.4855, C: 30.5003)	NSP12	324	T	T	(del, T)
14412	A	A	(del: 69.5594, A: 30.4052)	NSP12	324	T	T	(del, T)
14413	A	A	(del: 68.6659, A: 31.208)	NSP12	325	S	S	(L(frameshift), S)
14414	G	G	(del: 65.0086, G: 28.7953)	NSP12	325	S	S	(L(frameshift), S)
17146	A	A	(A: 74.8634, G: 25.1115)	NSP13	304	I	I	(I, V)
20178	C	C	(C: 61.9078, T: 38.0279)	NSP15	186	V	V	V

20480	C	(C: 88.9352, T: 11.0648)	(C: 71.5445, T: 28.393)	NSP15	287	S	(S, L)	(S, L)
20573	T	(T: 89.3473, C: 10.6527)	(T: 79.0348, C: 20.8979)	NSP15	318	V	(V, A)	(V, A)
21789	C	T	T	Spike	76	T	I	I
22264	C	(C: 86.3636, T: 13.6364)	(C: 80.3188, T: 19.6812)	Spike	234	N	N	N
23271	C	C	(C: 52.5979, T: 47.38)	Spike	570	A	A	(A, V)
25688	C	C	(C: 54.5479, T: 45.3431)	ORF3a	99	A	A	(A, V)
27131	C	(C: 69.5341, T: 30.0717)	(C: 75.4056, T: 24.5081)	M	203	N	N	N
28649	A	T	T	N	126	N	Y	Y

Supplementary Table 20. Comparison between original control and $FFM7_{HIGH}$.

Position	Original control	$FFM7_{LOW}$	$FFM7_{HIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM7_{LOW}$ (aa)	$FFM7_{HIGH}$ (aa)
44	C	(T: 52.8866, C: 47.0169999999999)	T	-	-	-	-	-
835	C	(T: 54.1016, C: 45.5078)	T	NSP2	10	F	F	F
2509	C	(C: 89.6681, T: 10.3029999999999)	(C: 55.491, T: 44.343)	NSP2	568	P	P	P
5299	T	T	(T: 82.5574, C: 17.3391)	NSP3	860	T	T	T
6255	C	C	(C: 55.4576, T: 44.5004)	NSP3	1179	A	A	(A, V)
11399	A	(G: 51.3368, A: 48.6425)	(G: 59.3008, A: 40.674)	NSP6	143	M	(V, M)	(V, M)
12334	(A: 89.4929, G: 4.73853, T: 4.21928, C: 1.36765)	(A: 89.3829, G: 5.10838, T: 3.70473)	A	NSP8	81	A	A	A

16616	C	(C: 78.4488, T: 21.5418)	(C: 77.1102, T: 22.8522)	NSP13	127	T	(T, I)	(T, I)
17678	C	C	(T: 60.7103, C: 39.2365)	NSP13	481	T	T	(T, M)
19955	C	C	(T: 55.1192, C: 44.8591)	NSP15	112	T	T	(I, T)
21789	C	(T: 51.076, C: 48.924)	(T: 69.2797, C: 30.6939)	Spike	76	T	(I, T)	(I, T)
22032	T	(T: 60.7193, C: 39.1396)	(C: 81.1349, T: 18.715)	Spike	157	F	(F, S)	(S, F)
23179	(C: 56.6181, T: 43.2826)	(T: 87.1688, C: 12.7615)	T	Spike	539	V	V	V
23271	C	(T: 85.0681, C: 14.9319)	T	Spike	570	A	(V, A)	V
23542	T	T	(T: 87.3556, C: 12.5822)	Spike	660	Y	Y	Y
27972	C	C	(T: 68.3281, C: 31.6045)	ORF8	27	Q	Q	(Stop, Q)
28311	C	(T: 67.3919, C: 32.5791)	(C: 71.8957, T: 27.9873)	N	13	P	(L, P)	(P, L)
28887	C	C	(C: 69.8426, T: 30.1291)	N	205	T	T	(T, I)
28899	G	(G: 89.7273, T: 9.89471)	(T: 64.3299, G: 35.395)	N	209	R	(R, I)	(I, R)

Supplementary Table 21. Comparison between original control and $FFM3_{remLOW}$.

Position	Original control	$FFM3_{remLOW}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{remLOW}$ (aa)
1104	T	(C: 63.7076, T: 36.2924)	NSP2	100	I	(I, T)
7321	C	G	NSP3	1534	S	R
12334	(A: 89.0675, G:4.91803, T:4.31126)	A	NSP8	81	A	A
13961	T	(C: 56.9699, T: 42.9823)	NSP12	174	V	(V, A)

14786	C	(T: 51.1807, C: 47.5988, -: 1.22048)	NSP12	449	A	(A, V, V(frameshift))
15451	G	A	NSP12	671	G	S
18687	C	T	NSP14	216	C	C
25603	C	(C: 60.7398, T: 39.1761)	ORF3a	71	L	L
29473	G	(T: 69.5636, G: 30.1135)	N	400	L	(F, L)

Supplementary Table 22. Comparison between original control and $FFM7_{remLOW}$.

Position	Original control	$FFM7_{remLOW}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM7_{remLOW}$ (aa)
335	C	(C: 77.9874, T: 22.0126)	NSP1	24	R	(R, C)
3768	C	T	NSP3	350	T	I
5386	T	(G: 67.5405, T: 32.4595)	NSP3	889	A	A
6323	G	(G: 68.9617, A: 31.0383)	NSP3	1202	E	(E, K)
12459	C	T	NSP8	123	T	I
13626	T	(T: 85.0358, C: 14.9391)	NSP12	62	D	D
14408	T	(T: 89.0866, C: 9.96441)	NSP12	323	L	(L, P)
14786	C	T	NSP12	449	A	V
21765	T	(-: 68.9655, T: 31.0345)	Spike	68	I	I
21766	A	(-: 74.5763, A: 24.9153)	Spike	68	I	I
21767	C	(-: 74.4932, C: 25.5068)	Spike	69	H	(del, H)
21768	A	(-: 75.7732, A: 24.2268)	Spike	69	H	(del, H)
21769	T	(-: 75.1286, T: 24.8714)	Spike	69	H	(del, H)
21770	G	(-: 73.2441, G: 24.5819)	Spike	70	V	(I, V)
24712	G	(G: 73.8287, T: 26.1347)	Spike	1050	M	(M, I)
29901	A	n/a	-	-	-	-
29902	A	n/a	-	-	-	-
29903	A	n/a	-	-	-	-

Supplementary Table 23. Comparison between original control and $FFM3_{remHIGH}$

Position	Original control	$FFM3_{remLOW}$	$FFM3_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{remLOW}$ (aa)	$FFM3_{remHIGH}$ (aa)

703	C	C	(A: 52.7693, C: 46.9631)	NSP1	146	G	G	G
1001	G	G	A	NSP2	66	E	E	K
1093	C	C	(C: 86.0135, T: 13.9135)	NSP2	96	P	P	P
1104	T	(C: 63.7076, T: 36.2924)	C	NSP2	100	I	(I, T)	T
2107	T	T	(T: 56.0339, C: 43.9567)	NSP2	434	T	T	T
6205	G	G	(G: 86.5395, A: 13.3565)	NSP3	1162	K	K	K
6649	T	T	(T: 89.3391, A: 10.3714)	NSP3	1310	A	A	A
7321	C	G	G	NSP3	1534	S	R	R
13961	T	(C: 56.9699, T: 42.9823)	C	NSP12	174	V	(V, A)	A
14786	C	(T: 51.1807, C: 47.5988, -: 1.22048)	T	NSP12	449	A	(A, V, V(frameshift))	V
15451	G	A	A	NSP12	671	G	S	S
16044	A	A	(A: 84.5488, G: 15.3846)	NSP12	868	P	P	P
18687	C	T	T	NSP14	216	C	C	C
20178	C	C	C	NSP15	186	V	V	V
24763	T	T	(T: 83.8842, C: 16.0866)	Spike	1067	Y	Y	Y
25003	A	A	(A: 85.6416, G: 14.3099)	Spike	1147	S	S	S
26541	A	A	G	M	7	T	T	A
29473	G	(T: 69.5636, G: 30.1135)	T	N	400	L	(F, L)	F

Supplementary Table 24. Comparison between the original control of FFM3 strain and $FFM3_{HIGH}$.

Position	Original control	FFM3LOW	FFM3HIGH	Protein	Amino acid (aa) position	Original control (aa)	FFM3LOW (aa)	FFM3HIGH (aa)
508	T	T	(del: 52.7194, T: 47.1872)	NSP1	81	H	H	H
509	G	G	(del: 53.301, G: 46.6304)	NSP1	82	G	G	(del, G)
510	G	(G: 86.1621, del: 13.5361)	(del: 65.7674, G: 33.7306)	NSP1	82	G	(G, V(frameshift))	(del, G)
511	T	(T: 85.7487, del: 14.0584)	(del: 66.3943, G: 33.4024)	NSP1	82	G	(G, V(frameshift))	(del, G)
512	C	(C: 85.9727, del: 13.9674)	(del: 66.8244, C: 33.1282)	NSP1	83	H	(H, del)	(del, H)
513	A	(A: 86.2357, del: 13.6253)	(del: 66.7544, A: 33.1499)	NSP1	83	H	(H, del)	(del, H)
514	T	(T: 80.8193, del: 19.1015)	(del: 66.6022, T: 33.268)	NSP1	83	H	(H, del)	(del, H)
515	G	(G: 80.9692, del: 18.9535)	(del: 66.7297, G: 33.236)	NSP1	84	V	(V, del)	(del, V)
516	T	(T: 78.9138, del: 20.6612)	(del: 67.4066, T: 32.4898)	NSP1	84	V	(V, del)	(del, V)
517	T	(T: 79.7661, del: 20.0504)	(del: 67.0348, T: 32.9092)	NSP1	84	V	(V, del)	(del, V)
518	A	(A: 78.6081, del: 20.8314)	(del: 67.3911, A: 32.4729)	NSP1	85	M	(M, del)	(del, M)
519	T	(T: 87.6773, del: 12.1702)	(del: 67.1115, T: 32.8006)	NSP1	85	M	(M, del)	(del, M)
520	G	(G: 87.9632, del: 11.9448)	(del: 66.958, G: 32.9531)	NSP1	85	M	(M, del)	(del, M)
521	G	(G: 89.3565, del: 10.5973)	(del: 65.7345, G: 34.2222)	NSP1	86	V	(V, V(frameshift))	(H, V)
522	T	(T: 88.535, del: 11.1489)	(del: 65.3789999999999, T: 33.7467)	NSP1	86	V	(V, V(frameshift))	(H, V)

1288	C	C	(C: 84.1061, T: 15.8384)	NSP2	161	C	C	C
2062	C	C	(C: 84.1517, T: 15.8314)	NSP2	419	A	A	A
5962	T	T	(T: 81.1249)	NSP3	1081	Y	Y	(Y)
6045	A	A	(A: 89.5812, del: 10.276)	NSP3	1109	N	N	(N, I)
6046	T	T	(T: 89.0053, del: 10.7702)	NSP3	1109	N	N	(N, I)
6047	T	T	(T: 88.7258, del: 10.6844)	NSP3	1110	F	F	(F, I)
7521	C	C	(C: 53.9715, T: 46.0285)	NSP3	1601	T	T	(T, I)
8290	C	C	(C: 86.141, T: 13.757)	NSP3	1857	L	L	L
11760	A	A	(A: 68.9241, G: 31.0069999999999)	NSP6	263	K	K	(K, R)
12016	T	T	(T: 86.5919, G: 13.157)	NSP7	58	V	V	V
12334	(A: 89.0675, G:4.91803, T:4.31126)	(A: 89.6299, G: 4.64974, T: 3.84419)	A	NSP8	81	A	A	A
14408	T	T	(del: 68.3007, T: 31.2949)	NSP12	323	P	P	(L(frameshift), P)
14409	T	T	(del: 68.5469, T: 31.4461)	NSP12	323	P	P	(L(frameshift), P)
14410	A	A	(del: 68.9873, A: 30.9636)	NSP12	324	T	T	(del, T)
14411	C	C	(del: 69.4855, C: 30.5003)	NSP12	324	T	T	(del, T)
14412	A	A	(del: 69.5594, A: 30.4052)	NSP12	324	T	T	(del, T)
14413	A	A	(del: 68.6659, A: 31.208)	NSP12	325	S	S	(L(frameshift), S)

14414	G	G	(del: 65.0086, G: 28.7953)	NSP12	325	S	S	(L(frameshift), S)
17146	A	A	(A: 74.8634, G: 25.1115)	NSP13	304	I	I	(I, V)
20178	C	C	(C: 61.9078, T: 38.0279)	NSP15	186	V	V	V
20480	C	(C: 88.9352, T: 11.0648)	(C: 71.5445, T: 28.393)	NSP15	287	S	(S, L)	(S, L)
20573	T	(T: 89.3473, C: 10.6527)	(T: 79.0348, C: 20.8979)	NSP15	318	V	(V, A)	(V, A)
21789	C	T	T	Spike	76	T	I	I
22264	C	(C: 86.3636, T: 13.6364)	(C: 80.3188, T: 19.6812)	Spike	234	N	N	N
23271	C	C	(C: 52.5979, T: 47.38)	Spike	570	A	A	(A, V)
25688	C	C	(C: 54.5479, T: 45.3431)	ORF3a	99	A	A	(A, V)
27131	C	(C: 69.5341, T: 30.0717)	(C: 75.4056, T: 24.5081)	M	203	N	N	N
28649	A	T	T	N	126	N	Y	Y
29729	T	T	(T: 89.1544, del: 10.2824)	-	-	-	-	-
29730	C	C	(C: 89.3007, del: 10.5428)	-	-	-	-	-
29731	A	A	(A: 88.709, del: 10.8333)	-	-	-	-	-
29732	C	C	(C: 89.0981, del: 10.8333)	-	-	-	-	-
29733	C	C	(C: 88.4508, del: 10.8131)	-	-	-	-	-
29734	G	G	(G: 88.1356, del: 10.9485)	-	-	-	-	-
29735	A	A	(A: 87.5155, del: 11.3761)	-	-	-	-	-
29736	G	G	(G: 87.9207, del: 11.4633)	-	-	-	-	-

29737	G	G	(G: 87.2632, del: 11.6399)	-	-	-	-	-
29738	C	C	(C: 87.5606, del: 11.7986)	-	-	-	-	-
29739	C	C	(C: 87.2713, del: 12.2713)	-	-	-	-	-
29740	A	A	(A: 86.1179, del: 13.1402)	-	-	-	-	-
29741	C	C	(C: 86.2311, del: 13.4971)	-	-	-	-	-
29742	G	G	(G: 84.9052, del: 14.542)	-	-	-	-	-
29743	C	C	(C: 84.8207, del: 14.924)	-	-	-	-	-
29744	G	G	(G: 83.6476, del: 15.6467)	-	-	-	-	-
29745	G	G	(G: 83.1499, del: 16.2682)	-	-	-	-	-
29746	A	A	(A: 82.7334, del: 16.771)	-	-	-	-	-
29747	G	G	(G: 82.3358, del: 17.5056)	-	-	-	-	-
29748	T	T	(T: 82.4847, del: 17.2158)	-	-	-	-	-
29749	A	A	(A: 79.858, del: 19.432)	-	-	-	-	-
29750	C	C	(C: 82.5932, del: 13.9281, T: 3.38648)	-	-	-	-	-
29751	G	G	(G: 85.5162, del: 14.1876)	-	-	-	-	-
29752	A	A	(A: 84.9252, del: 14.5903)	-	-	-	-	-
29753	T	T	(T: 84.3123, del: 15.0192)	-	-	-	-	-

29754	C	C	(C: 83.782999999999, del: 15.9897)	-	-	-	-	-
29755	G	G	(G: 83.6713, del: 16.2058)	-	-	-	-	-
29756	A	A	(A: 83.2451, del: 16.4902)	-	-	-	-	-
29757	G	G	(G: 83.3696, del: 16.4748)	-	-	-	-	-
29758	T	T	(T: 82.8544, del: 16.7193)	-	-	-	-	-
29759	G	G	(G: 83.4019, del: 16.1392)	-	-	-	-	-

Supplementary Table 25. Comparison between original control and FFM7$_{remHIGH}$.

Position	Original control	FFM7$_{remLOW}$	FFM7$_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	FFM7$_{remLOW}$ (aa)	FFM7$_{remHIGH}$ (aa)
1009	C	C	T	NSP2	68	S	S	S
3768	C	T	T	NSP3	350	T	I	I
3798	T	T	(G: 71.7268, T: 28.1433)	NSP3	360	F	F	(F, C)
4180	G	G	(G: 87.8939, T: 12.0869)	NSP3	487	K	K	(K, N)
4320	C	C	(C: 53.0214, T: 46.9618)	NSP3	534	A	A	(A, V)
5386	T	(G: 67.5405, T: 32.4595)	G	NSP3	889	A	A	A
8097	C	C	(C: 86.8797, T: 13.0799)	NSP3	1793	T	T	(T, I)
8764	T	T	G	NSP4	70	D	D	E
12334	(A: 89.4929, G: 4.73853, T: 4.21928, C: 1.36765)	(A: 88.5248, G: 5.26468, C: 1.76615)	A	NSP8	81	A	A	A

12459	C	T	T	NSP8	123	T	I	I
12694	G	G	(A: 69.977, G: 29.9959)	NSP9	3	E	E	E
12796	A	A	(A: 76.1811, G: 23.4402)	NSP9	37	G	G	G
13626	T	(T: 85.0358, C: 14.9391)	(C: 69.0219, T: 30.9313)	NSP12	62	D	D	D
14120	C	C	(C: 58.6855, T: 41.272)	NSP12	227	P	P	(P, L)
14786	C	T	T	NSP12	449	A	V	V
15699	C	C	(C: 62.4526, T: 37.4457)	NSP12	753	F	F	F
18555	C	C	(C: 66.766, T: 33.1755)	NSP14	172	D	D	D
20915	G	G	(G: 84.4288, A: 15.5388)	NSP16	86	R	R	R
21765	T	(-: 68.9655, T: 31.0345)	(T: 78.6009, -: 21.1187)	Spike	68	I	I	I
21766	A	(-: 74.5763, A: 24.9153)	(A: 76.0182, -: 23.3138)	Spike	68	I	I	I
21767	C	(-: 74.4932, C: 25.5068)	(C: 76.8336, -: 23.1179)	Spike	69	H	(del, H)	(H, del)
21768	A	(-: 75.7732, A: 24.2268)	(A: 76.0258, -: 23.6598)	Spike	69	H	(del, H)	(H, del)
21769	T	(-: 75.1286, T: 24.8714)	(T: 75.8322, -: 23.7017)	Spike	69	H	(del, H)	(H, del)
21770	G	(-: 73.2441, G: 24.5819)	(G: 74.247, -: 23.8286)	Spike	70	V	(I, V)	(V, del)
23179	(C: 56.6181, T: 43.2826)	C	C	Spike	539	V	V	V
24872	G	G	(G: 66.3453, A: 33.6088)	Spike	1104	V	V	(V, I)
27502	T	T	G	ORF7a	37	S	S	A
28742	A	A	G	N	157	I	I	V
29891	A	A	(A: 85.7143, G: 14.2857)	-	-	-	-	-

Supplementary Table 26. Comparison between $FFM3_{remLOW}$ and $FFM3_{remHIGH}$.

Position	Original control	$FFM3_{remLOW}$	$FFM3_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM3_{remLOW}$ (aa)	$FFM3_{remHIGH}$ (aa)
703	C	C	(A: 52.7693, C: 46.9631)	NSP1	146	G	G	G
1001	G	G	A	NSP2	66	E	E	K
1093	C	C	(C: 86.0135, T: 13.9135)	NSP2	96	P	P	P
1104	T	(C: 63.7076, T: 36.2924)	C	NSP2	100	I	(I, T)	T
2107	T	T	(T: 56.0339, C: 43.9567)	NSP2	434	T	T	T
6205	G	G	(G: 86.5395, A: 13.3565)	NSP3	1162	K	K	K
6649	T	T	(T: 89.3391, A: 10.3714)	NSP3	1310	A	A	A
13961	T	(C: 56.9699, T: 42.9823)	C	NSP12	174	V	(V, A)	A
14786	C	(T: 51.1807, C: 47.5988, -: 1.22048)	T	NSP12	449	A	(A, V, V(frameshift))	V
16044	A	A	(A: 84.5488, G: 15.3846)	NSP12	868	P	P	P
24763	T	T	(T: 83.8842, C: 16.0866)	Spike	1067	Y	Y	Y
25003	A	A	(A: 85.6416, G: 14.3099)	Spike	1147	S	S	S
25603	C	(C: 60.7398, T: 39.1761)	C	ORF3a	71	L	L	L
26541	A	A	G	M	7	T	T	A
29473	G	(T: 69.5636, G: 30.1135)	T	N	400	L	(F, L)	F

Supplementary Table 27. Comparison between $FFM7_{remLOW}$ and $FFM7_{remHIGH}$.

Position	Original control	$FFM7_{remLOW}$	$FFM7_{remHIGH}$	Protein	Amino acid (aa) position	Original control (aa)	$FFM7_{remLOW}$ (aa)	$FFM7_{remHIGH}$ (aa)
335	C	(C: 77.9874, T: 22.0126)	C	NSP1	24	R	(R, C)	R
1009	C	C	T	NSP2	68	S	S	S
3798	T	T	(G: 71.7268, T: 28.1433)	NSP3	360	F	F	(F, C)
4180	G	G	(G: 87.8939, T: 12.0869)	NSP3	487	K	K	(K, N)
4320	C	C	(C: 53.0214, T: 46.9618)	NSP3	534	A	A	(A, V)
5386	T	(G: 67.5405, T: 32.4595)	G	NSP3	889	A	A	A
6323	G	(G: 68.9617, A: 31.0383)	G	NSP3	1202	E	(E, K)	E
8097	C	C	(C: 86.8797, T: 13.0799)	NSP3	1793	T	T	(T, I)
8764	T	T	G	NSP4	70	D	D	E
12694	G	G	(A: 69.977, G: 29.9959)	NSP9	3	E	E	E
12796	A	A	(A: 76.1811, G: 23.4402)	NSP9	37	G	G	G
13626	T	(T: 85.0358, C: 14.9391)	(C: 69.0219, T: 30.9313)	NSP12	62	D	D	D
14120	C	C	(C: 58.6855, T: 41.272)	NSP12	227	P	P	(P, L)
14408	T	(T: 89.0866, C: 9.96441)	T	NSP12	323	L	(L, P)	L
15699	C	C	(C: 62.4526, T: 37.4457)	NSP12	753	F	F	F
18555	C	C	(C: 66.766, T: 33.1755)	NSP14	172	D	D	D
20915	G	G	(G: 84.4288, A: 15.5388)	NSP16	86	R	R	R

21765	T	(del: 68.9655, T: 31.0345)	(T: 78.6009, del: 21.1187)	Spike	68	I	I	I
21766	A	(del: 74.5763, A: 24.9153)	(A: 76.0182, del: 23.3138)	Spike	68	I	I	I
21767	C	(del: 74.4932, C: 25.5068)	(C: 76.8336, del: 23.1179)	Spike	69	H	(del, H)	(H, del)
21768	A	(del: 75.7732, A: 24.2268)	(A: 76.0258, del: 23.6598)	Spike	69	H	(del, H)	(H, del)
21769	T	(del: 75.1286, T: 24.8714)	(T: 75.8322, del: 23.7017)	Spike	69	H	(del, H)	(H, del)
21770	G	(del: 73.2441, G: a24.5819)	(G: 74.247, del: 23.8286)	Spike	70	V	(del, V)	(V, del)
24712	G	(G: 73.8287, T: 26.1347)	G	Spike	1050	M	(M, I)	M
24872	G	G	(G: 66.3453, A: 33.6088)	Spike	1104	V	V	(V, I)
27502	T	T	G	ORF7a	37	S	S	A
28742	A	A	G	N	157	I	I	V
29891	A	A	(A: 85.7143, G: 14.2857)	-	-	-	-	-
29901	A	n/a	A	-	-	-	-	-
29902	A	n/a	A	-	-	-	-	-
29903	A	n/a	A	-	-	-	-	-

Supplementary Table 28. Nonsynonymous mutations in FFM3$_{LOW}$ and FFM3$_{HIGH}$

Position	Original control	Low	High	Protein	Position (aa)	Original control (aa)	Low (aa)	High (aa)
21789	C	T	T	Spike	76	T	I	I
		C	C				T	T
28649	A	T	T	N	126	N	Y	Y
		A	A				N	N

Supplementary Table 29. Nonsynonymous mutations in FFM3$_{remLOW}$ and FFM3$_{remHIGH}$.

Position	Original control	Low	High	Protein	Position (aa)	Original control (aa)	Low (aa)	High (aa)
1001	G	G	G	NSP2	66	E	E	E
		G	A				E	K
1104	T	T	T	NSP2	100	I	I	I
		(C: 63.7076, T: 36.2924)	C				(I, T)	T
7321	C	C	C	NSP3	1534	S	S	S
		G	G				R	R
13961	T	T	T	NSP12	174	V	V	V
		(C: 56.9699, T: 42.9823)	C				(V, A)	A
14786	C	C	C	NSP12	449	A	A	A
		(T: 51.1807, C: 47.5988, -: 1.22048)	T				(V, A)	V
15451	G	G	G	NSP12	671	G	G	G
		A	A				S	S
26541	A	A	A	M	7	T	T	T
		A	G				T	A
29473	G	G	G	N	400	L	L	L
		(T: 69.5636, G: 30.1135)	T				(F, L)	F

Supplementary Table 30. Nonsynonymous mutations in FFM7$_{LOW}$ and FFM7$_{HIGH}$

Position	Original control	Low	High	Protein	Position (aa)	Original_control (aa)	Low (aa)	High (aa)
23271	C	(T: 85.0681, C: 14.9319)	T	Spike	570	A	(V, A)	V
		C	C				A	A

Supplementary Table 31. Nonsynonymous mutations in FFM7$_{remLOW}$ and FFM7$_{remHIGH}$

Position	Original control	Low	High	Protein	Position (aa)	Original_control (aa)	Low (aa)	High (aa)
3768	C	C	C	NSP3	350	T	T	T
		T	T				I	I
8764	T	T	T	NSP4	70	D	D	D
		T	G				D	E
12459	C	C	C	NSP8	123	T	T	T
		T	T				I	I
14786	C	C	C	NSP12	449	A	A	A
		T	T				V	V
27502	T	T	T	ORF7a	37	S	S	S
		T	G				S	A
28742	A	A	A	N	157	I	I	I
		A	G				I	V