ELUCIDATING CONSERVED SUB-UNITS FROM BETA-CORONAVIRUS SPECIES FOR THE CREATION OF NOVEL, CROSS-CLADE, AND STABLE CHIMERA SARS-COV-2 SPIKE PROTEINS AS FUTURE PROOF VACCINE CANDIDATES

WANG SUSU

UNIVERSITI SAINS MALAYSIA

2023

ELUCIDATING CONSERVED SUB-UNITS FROM BETA-CORONAVIRUS SPECIES FOR THE CREATION OF NOVEL, CROSS-CLADE, AND STABLE CHIMERA SARS-COV-2 SPIKE PROTEINS AS FUTURE PROOF VACCINE CANDIDATES

by

WANG SUSU

Dissertation submitted in partial fulfilment of
the requirements of the degree of
Master of Science (Biomedicine) Mixed Mode

AUGUST 2023

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who helped me during the writing of this dissertation. My deepest gratitude goes first and foremost to my supervisors Dr. Ezzeddin Kamil Bin Mohamed Hashim, Dr. Nik Yusnoraini Binti Yusof, for their constant encouragement and guidance. They helped me through all the stages of the writing of this dissertation and always be patient with me. Without their consistent and illuminating instruction, I cannot finish this dissertation. Also, I would like to thank all my professors who taught me lessons during this remarkable year. I learned a lot from you.

I also want to say thanks to Dr. Wong Weng Kin who has guided me in essay writing and patiently taught us a lot of knowledge and skills related to essay writing requirements and formats.

My heartfelt appreciation also goes to my fellow course-mates for their helpful advice and motivations throughout the challenging period of completing the research project and writing this thesis.

Last but not least, I also would like to dedicate this thesis to my parents for their unconditional love, support and inspiration. I hope I have made them proud.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS ABBREVIATIONS

WHO	World Health Organization
%	Percentage
>	More than
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
COVID-19	Coronavirus Disease 2019
ACE2	Angiotensin-converting Enzyme 2
NCBI	National Center for Biotechnology Information
MD	Molecular Dynamics
FP	Fusion Peptide
HR1	Heptad Repeat 1
HR2	Heptad Repeat 2
pLDDT	predicted Local Distance Difference Test
PAE	Predicted Aligned Error
pTM	predicted TM Score
ipTM	interface predicted TM Score
RBD	Receptor Binding Domain
RBM	Receptor Binding Motif
CTD	C-terminal Domain
NTD	N-terminal Domain
aa	Amino acid
RMSD	Root Mean Square Deviation
bp	Base pair

MENJELASKAN SUB-UNIT TERPELIHARA DARIPADA SPESIES BETA-CORONAVIRUS UNTUK MENCIPTA PROTEIN SPIKE NOVEL, RENTAS-CLADE DAN STABIL CHIMERA SARS-COV-2 SPIKE SEBAGAI CALON VAKSIN KALIS MASA DEPAN

ABSTRAK

Pandemik COVID-19 yang disebabkan oleh wabak SARS-CoV-2 telah mencetuskan penyelidikan saintifik yang meluas. Dalam tesis ini, protein pepaku (S) daripada SARS-CoV-2 telah dikaji secara mendalam untuk mendapatkan pandangan berguna untuk mencipta vaksin yang terbukti pada masa depan. Maklumat struktur dan kestabilan struktur protein beta-coronavirus chimeric telah disiasat menggunakan teknik bioinformatik, peramalan struktur dan simulasi dinamik molekul. Kajian ini pada mulanya menyasarkan keseluruhan protin S bagi beta-coronavirus cimera, tetapi disebabkan kekangan tertentu, kajian ini beralih arah kepada kawasan terpelihara subunit S2 sahaja untuk menghasilkan jujukan cimera yang mewakili coronavirus beta yang berbeza, demi mendapatkan calon vaksin dengan liputan imun yang meluas. Untuk lebih praktikal, sepuluh jujukan S2 cimera telah dipilih secara teliti untuk mengkaji struktur protin virus cimera. Kajian ini menggunakan AlphaFold2 untuk meramal struktur 3D jujukan protin S2 cimera tersebut. Simulasi dinamik molekul menjelaskan lagi kestabilan strukturnya. Hasil kajian ini meletakkan asas untuk reka bentuk vaksin baru yang terbukti pada masa depan yang baru.

ELUCIDATING CONSERVED SUB-UNITS FROM BETA-CORONAVIRUS SPECIES FOR THE CREATION OF NOVEL, CROSS-CLADE, AND STABLE CHIMERA SARS-COV-2 SPIKE PROTEINS AS FUTURE PROOF VACCINE CANDIDATES

ABSTRACT

The COVID-19 pandemic caused by the SARS-CoV-2 outbreak triggered extensive scientific research. In this thesis, the spike (S) protein of the SARS-CoV-2 was studied in depth to gain useful insights to create a future proof vaccine. The structural information and structural stability of chimeric beta-coronavirus S proteins were investigated using bioinformatics techniques, structural predictions and molecular dynamics simulations. The study initially targeted the entire S protein of chimeric beta-coronaviruses, but due to certain constraints, the study shifted the direction onto conserved regions of the S2 subunit only to generate chimeric sequences from different beta coronaviruses, in order to obtain vaccine candidates with broad immune coverage. For practicality, ten chimeric S2 sequences were carefully selected to study chimeric viral protein structures. This study used AlphaFold2 to predict 3D structures of the chimeric S2 protein sequences. Molecular dynamics simulations further elucidated their structural stabilities. The results lay the foundation for novel future proof vaccine design.

CHAPTER 1

INTRODUCTION

The emergence of the novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and its ensuing global pandemic (Coronavirus Virus Disease in 2019 (COVID-19)) has highlighted the urgent need for effective prevention and treatment strategies. This research has focused on the spike (S) protein of SARS-CoV-2, which is a crucial factor in the virus' capacity to enter and infect human cells and is the subject of efforts to control the spread and effects of COVID-19.

The S protein facilitates viral entry into host cells by attaching to ACE2 receptors on the surface of target cells. This connection sets off a chain of events that leads to viral fusion and the release of the viral genome into the cytoplasm of the host cell. The S protein has emerged as a key target for the development of vaccines and treatments due to its crucial function in the viral life cycle.

Researchers have concentrated on taking use of the potential of S proteins in response to the urgent need for an efficient countermeasure against SARS-CoV-2. The creation of neutralising monoclonal antibodies (nAbs) that precisely target the S protein is one strategy that has advanced significantly. These nAbs have demonstrated a potential capacity to prevent viral entrance and lessen viral replication in preclinical and clinical investigations. This demonstrates the potential of S proteins as therapeutic targets and as the basis for new vaccination candidates.

Studies showing cross-population immunity in people who had previously received the SARS-CoV-2 vaccination after having an early SARS-CoV-1 infection reveal the potential for inducing a strong and broad immune response against related coronaviruses. Interestingly, this reaction was effective against a variety of pan-coronavirus strains, indicating that there may be conserved S protein areas that might be the focus of future vaccine research. This cross-clade enrichment supports the idea of developing stable and functional chimeric SARS-CoV-2 spiking proteins that could be used as future vaccine candidates against a wider range of coronaviruses in addition to shedding light on the potential role of S proteins in eliciting protective immune responses.

Additionally, bioinformatics has evolved over the past several years into a crucial tool for advancing our comprehension of viral biology and directing the development of innovative therapies. The creation of efficient preventative strategies against infectious illnesses has been substantially hastened by the use of computational tools to examine viral genomes, predict protein structures, and simulate molecular interactions. The work of Simpson and Kasson shows the potential of bioinformatics in this area. (Simpson & Kasson, 2023). Their work is an excellent example of how computational tools and experimental methods may work together to solve difficult viral protein problems. This study sought to further the work of Simpson and Carson (2023) and to give a thorough examination of the found conserved subunits and their influence on the stability, functioning, and immunogenicity of the chimeric SARS-CoV-2 spike protein. This work focuses on an integrated approach using AlphaFold2 modelling followed by molecular dynamics simulations to assess the relative stability of chimeric spike proteins. The objective of this work is to get a deeper knowledge of the molecular underpinnings of the S2 region's cross-species protection and its potential to be developed into an effective vaccine candidate.

These chimeric structures are anticipated to trigger strong immune reactions that can neutralise a variety of coronaviruses, helping to build a flexible and robust defence mechanism against possible viral epidemics in the future. This research aims to build a strong basis for the creation of novel vaccine candidates for the ever-evolving coronavirus infection landscape by thoroughly examining and modifying the structural and functional features of the S proteins.

The molecular structure of S proteins, their conserved regions in -coronavirus species, the rationale behind the creation of chimeric spike proteins, an overview of the experimental techniques used, and the anticipated contribution and impact of this study on the creation of vaccine candidates that can stop future coronavirus outbreaks and recurrences are covered in the following sections.

1.1 Overview of the coronaviruses

Infectious bronchitis virus in chickens was originally discovered by researchers in the 1960s, followed by the identification of human coronaviruses such HCoV-229E and HCoV-OC43. A comparable virus was discovered by Hamre et al.（1966） in human embryonic kidney cells, and the representative strain was given the designation 229E virus(Hamre & Procknow, 1966a). Using human embryo tracheal cultures, Mclntosh et al. (1967)identified a collection of viruses from cold patients in 1967; the typical strain was OC43(McIntosh, Becker & Chanock, 1967). The foundation for understanding the virology, transmission, and pathophysiology of these agents was formed by these early research. The advent of the Middle East Respiratory Syndrome

Coronavirus (MERS-CoV) in 2012 and the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) in 2002, however, brought the right attention to this viral family. The potential for coronaviruses to spread from animal reservoirs to human hosts and cause serious and occasionally deadly infections was highlighted by these outbreaks.

1.1.1 Nomenclature of coronaviruses

The term "coronaviruses" comes from the elevated, spiky protein trimers on their envelopes, which, under the microscope, resemble a crown. Coronaviruses are a type of positive-sense single-stranded RNA viruses. Coronaviruses, which are common in animals and birds, primarily cause respiratory or intestinal illnesses, however they can also occasionally cause neurological or hepatic conditions. When understanding this family of viruses, it is crucial to comprehend the coronavirus nomenclature. The nomenclature, which is based on the genetic properties of the virus, facilitates effective communication between scientists and medical experts regarding the many strains and species. The family Coronaviridae, which consists of four genera and four species of coronavirus, is where coronaviruses are classified. Based on a combination of the viral host, geographic origin, and sequence similarity, the species are named (Fenner, 1976). For example, the species name of SARS-CoV is "Severe Acute Respiratory Syndrome-associated Coronavirus", reflecting its association with outbreaks of Severe Acute Respiratory Syndrome. Different strains or variants may emerge as a result of genetic mutation and evolution within a species. These can be identified using a variety of methods, including gene sequencing. For example, different variants of SARS-CoV-2, such as the Alpha, Beta, Gamma and Delta, represent different variants of the same virus with specific genetic changes. While scientific nomenclature provides accurate information about viruses, common names are often used to communicate with the public. For example, the virus that caused the COVID-19 pandemic is commonly referred to as "SARS-CoV-2", while the disease it caused is referred to as "COVID-19".

1.1.2 Classification of coronaviruses

The International Committee on Classification of Viruses classified coronaviruses into the order Nidoviridae, the family Coronaviridae and the subfamily Coronaviridae. Subsequently, on the basis of serological and genomic evidence, the Coronaviridae family was subdivided into four genera, i.e., α-coronavirus, β-coronavirus, γ-coronavirus and δ-coronavirus.(Bačenková et al., 2021). Of these, coronaviruses of genera α and δ are capable of infecting humans and mammals, in contrast to γ-coronaviruses, which are mainly isolated from avian hosts. To date, seven coronaviruses are known to infect humans, of which four are more common in the population, HCoV-NL63 (van der Hoek *et al.*, 2004), HCoV-229E (Hamre & Procknow, 1966b), HCoV-OC43 (Vijgen *et al.*, 2005), and HCoV-HKU1 (Woo *et al.*, 2005), which usually cause upper respiratory infections, with mild symptoms at the onset time of the disease. Meanwhile the other three coronaviruses are Middle East Respiratory Syndrome Coronavirus (MERS-CoV) which was first identified in the Middle East in 2013 with a mortality rate of up to 35% (Zumla, Hui & Perlman, 2015), Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) (Drosten *et al.*, 2003) with a mortality rate of up to 10% as well as the SARS-CoV-2 (Zhu *et al.*, 2020) which was first identified in Wuhan (within Hubei Province of China in 2019) with a more severe symptomatology at the time of disease onset, both of which can infect the lower respiratory tract and cause severe pneumonia.

At the end of 2020, the World Health Organization (WHO) classified mutants of the COVID virus (SARS-CoV-2), which poses a greater risk to global public health, into variants of interest (VOI) and variants of concern (VOC) in order to prioritise global detection and research and ultimately to inform the current response to the New Crown Pneumonia Pandemic (NCPP)(Amato *et al.*, 2022). According to ICTV, SARS-CoV-2 is classified as follows: Coronavirus genus Nucleovirus, Coronaviridae, Adenoviridae, Adenovirales, Adenovirales, Coronavirales, Coronaviridae, Coronaviridae, Coronaviridae, Coronaviridae, Beta Coronavirus, Sabellovirus subgenera, SARS-CoV-2 (Bačenková et al., 2021). Figure 1.1 shows the classification of SARS-CoV-2.

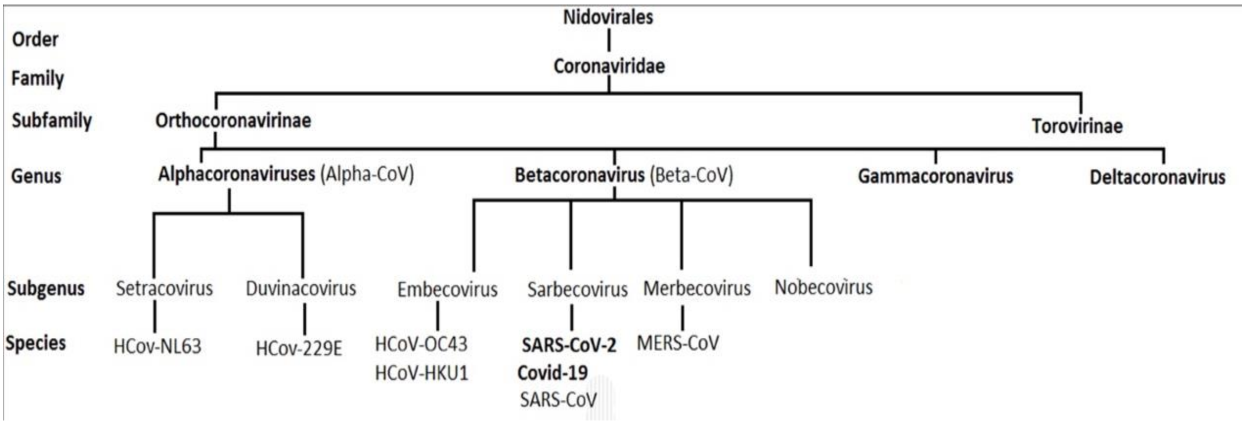

Figure 1.1 Taxonomy of HCoVs (Bačenková et al., 2021).

1.2 SARS-CoV-2

At the end of December 2019, a novel coronavirus pneumonia outbreak and rapid global spread occurred in Wuhan, Hubei Province, China, causing widespread concern in the community with its high contagiousness and transmission rate. As of July 2023, a cumulative total of 768,237,788 people has been infected by the virus globally, with a cumulative total of 6,950,000 deaths (https://covid19.who.int/). In February 2020, the World Health Organisation (WHO) coined it as Coronavirus Disease 19 (COVID-

19), a novel coronavirus pneumonia. In February 2020, WHO and the International Committee for the Classification of Viruses (ICTV) Coronavirus Study Group officially named it as "SARS-CoV-2". The new coronavirus is highly contagious and spreads mainly through respiratory droplets. Common symptoms of COVID-19 patients include fever, dry cough, loss of sense of smell, or accompanied by dyspnoea, shortness of breath, and severe symptoms leading to acute respiratory distress syndrome (ARDS) (Huang *et al.*, 2020)(Jiang *et al.*, 2020). In the field of infectious diseases, few diseases have captured the world's attention and profoundly altered peoples' daily life as much as the COVID-19 pandemic caused by the SARS-CoV-2. This virus, although is novel in its impact, is a member of the large family of coronaviruses.

1.2.1 Genome structure of coronaviruses

SARS-CoV2 is a single-stranded, encapsulated, complete, positive-sense RNA virus containing approximately 30,000 nucleotides and encoding 9860 amino acids. (Chan *et al.*, 2020). Figure 1.2 shows the genome structure of the human coronavirus. Coronavirus genomic RNA contains the eukaryotic 5' cap structure with a 3' polyadenylate tail that can function as a messenger RNA (mRNA). The 5' end of the viral genome is 65-98 nucleotide (nt) tract leading sequence, followed closely by a 200-400 nt untranslated region (UTR). The 3' end also contains a 200-500 base pair (bp) of untranslated region. The untranslated region is involved in viral replication and transcriptional functions (Makino, Stohlman & Lai, 1986; Yang & Leibowitz, 2015). The coronavirus genome usually contains 7-14 open reading frames (ORFs) (Brian & Baric, 2005; Perlman & Netland, 2009; Liu, 2014). The first 2/3 or so of the genomic sequence are encoded by two overlapping open reading frames, named ORF1a and ORF1b, which translated into viral transcription-replication complex polyprotein precursor and protease. Sheared by its own protease, it forms a mature RNA-dependent RNA polymerase (RdRp) with a variety of auxiliary function non-structural proteins (nsp) (Sola *et al.*, 2011; Snijder, Decroly & Ziebuhr, 2016). The other 1/3 of the genomic sequence mainly encodes four viral structural proteins i.e.: spike protein (S), envelope protein (E), membrane protein (M) and nucleocapsid protein (N). In addition, different coronaviruses encode different amounts of auxiliary proteins in this region (Forni *et al.*, 2017), and some of the viruses also encode haemaglutinin esterase (HE) proteins (Yokomori, Banner & Lai, 1991; Langereis *et al.*, 2010).

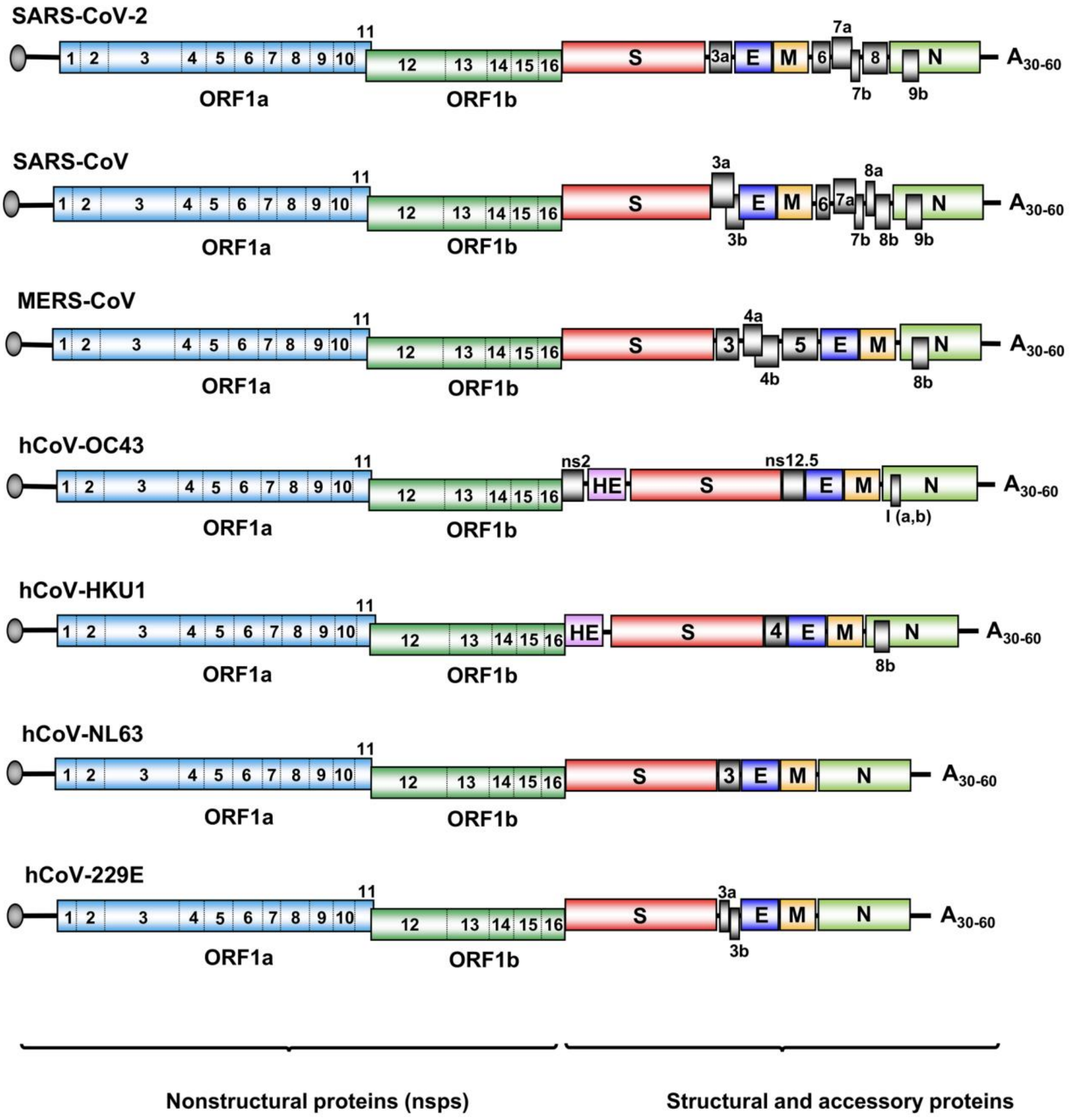

Figure 1.2 Genome structure of human coronavirus(Brant et al., 2021).

1.2.2 Main structural proteins and functions of coronaviruses

The spike protein (S) of SARS-COV-2 is a trimeric protein consisting of 1,273 amino acid residues, which is functionally responsible for specifically recognizing host cells and mediating cell membrane fusion into host cells and is structurally divided into two subunits: the S1 subunit and the S2 subunit (Wrapp et al., 2020). Figure 1.3 shows the structure of SARS-CoV-2. The S1 subunit consists of a receptor binding domain (RBD), an N-terminal domain, and a signal peptide, which is mainly involved in receptor recognition and enables the S protein to interact with host cells. The S2 subunit contains two heptapeptide repeats, HR-N and HR-C, which mainly promotes the fusion of cell membranes and facilitates the entry of coronaviruses into the host cells. Both S1 and S2 subunits have synergistic roles in their functions (Walls *et al.*, 2020). The SARS-CoV-2 S proteins promote viral infection of host cells upon binding to host cellular cathepsin-converting enzyme 2 (ACE2), thus, blocking the S proteins are crucial in preventing viral infection of human hosts.

Membrane glycoproteins (M) are the most abundant structural proteins in SARS-COV-2, with the main components being an amino (N)-terminal structural domain, multiple transmembrane structural domains, and a carboxyl (C)-terminal structural domain. M proteins are key players in the assembly and release of viral particles, and they also play an important role in immune regulation (de Haan, Vennema & Rottier, 2000). Envelope protein (E) is a section membrane protein, which acts as an ion channel and plays a major role in pathogenesis, viral assembly and release (Nieto-Torres *et al.*, 2014).

Nuclear coat protein (N) is a multifunctional structural protein that is translated in the cytoplasm and wrapped in the daughter genomic RNA to form a nuclear coat with three highly conserved structural domains: the N-terminal structural domain, the RNA-binding structural domain, and the C-terminal structural domain, which can interact with the membrane proteins during the assembly of viral particles, and plays an important role in enhancing viral transcription and assembly (McBride, van Zyl & Fielding, 2014).

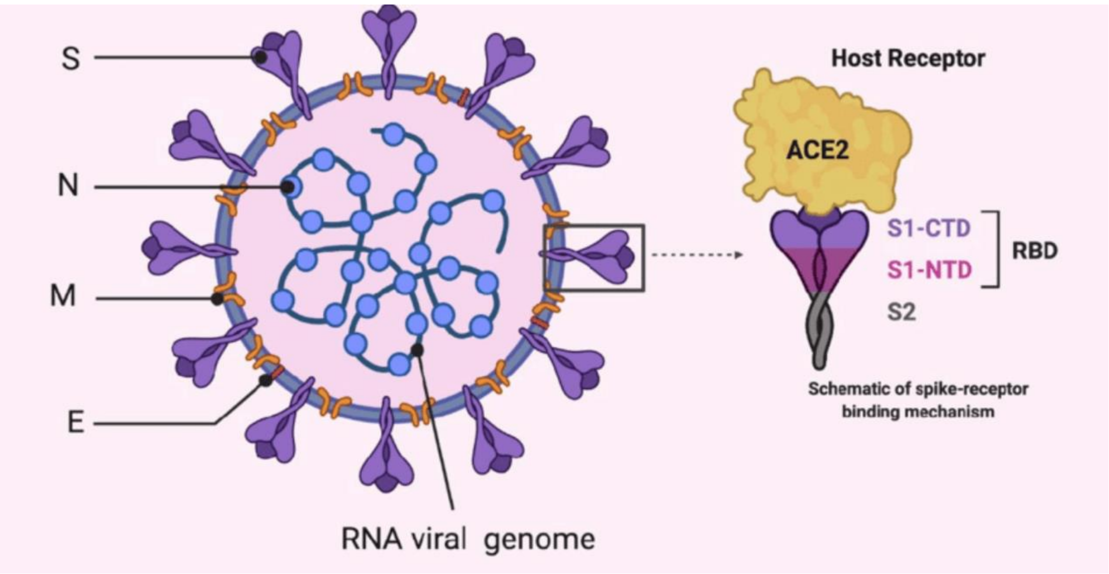

Figure 1.3 Structure of SARS-COV-2 virus and interaction of the receptor binding domain (RBD) with human angiotensin converting enzyme 2 (hACE-2) (Mahmood et al., 2020).

1.2.3 Mechanisms of SARS-COV-2 infection

All coronaviruses encode a surface glycoprotein, Spike, which binds to host cell receptors and mediates viral entry. Fusion of their envelopes with host cell membranes usually in acidic endosomal compartments or less frequently at the plasma membrane results in delivery of their nucleocapsids into the host cytoplasm. Spike glyco-protein (S) drives viral entry and is the major viral determinant of cytophagy. It is a class I fusion protein that is essential for relevant receptor binding on the host cell surface and for mediating fusion between the host and viral membranes in a process driven by significant conformational changes in the Spike protein. For β-coronaviruses, the receptor-binding domain (RBD) region of the Spike protein species mediates interaction with host cell receptors. Upon binding to the receptor, proximal host proteases cleave the spines and release spine fusion peptides to facilitate viral entry (Sanyal, 2020). Since the viral particles of coronaviruses are surrounded by S protein tract membranes, it can use the highly glycosylated S proteins to recognise host cells and enter them to cause viral infection (Lu, Wang & Gao, 2015). While S protein, as a trimeric class I fusion protein, exists in a sub-stable pre-fusion conformation and undergoes a large number of structural rearrangements to fuse the viral membrane with the host cell membrane (Bosch et al., 2003; Li, 2016). The surface-exposed portion of the S protein consists of two structural domains, S1 binds to the host cell receptor angiotensin-converting enzyme 2 (ACE2), whereas S2 catalyzes fusion of the virus and host cell. The S1 subunit contains the receptor binding domain (RBD) and the N-terminal structural domain (NTD), in which the RBD can directly bind to ACE2, so it also becomes the receptor binding domain; the S2 subunit mainly contains three structural domains of the fusion peptide (FP), HR-1/2 (heptad repeat region 1 and heptad repeat-2), The presence of the enzymatic cleavage site S1/S2 between the S1 and S2 subunits and the enzymatic cleavage site S2' within S2 is important for membrane fusion during viral invasion of cells (Wrapp *et al.*, 2020). Figure 1.4 shows conformational structure of SARS-CoV-2 before fusion. The RBD will attach to the body of the S protein through a soft region, and exists in a downward conformation after attachment, in which the RBD in the accessible upward conformation can bind with the ACE2, while making it cleavable by host proteases to trigger the S2 conformational change required for viral entry into host cells. Hoffmann et al.(2020) found that the S1 subunit on the S protein of SARS-CoV-2 binds to the human ACE2 receptor protein to cause infection, and further confirmed that the RBD region in the S1 subunit plays a key role in the binding process (Hoffmann *et al.*, 2020).

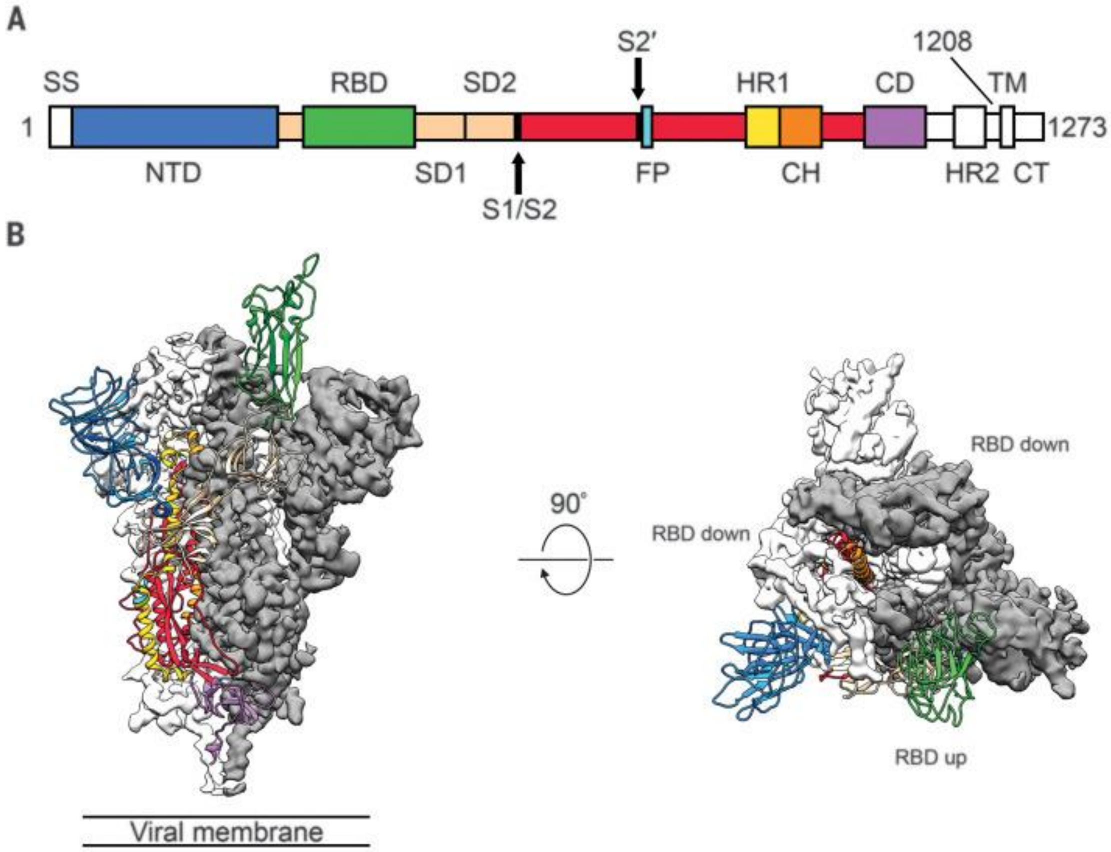

Figure 1.4 Conformational structure of SARS-CoV-2 before fusion (Wrapp et al., 2020)

ACE2 receptor protein is a type I membrane protein that promotes the conversion of angiotensin I to angiotensin II, thus playing a role in regulating vasoconstriction and controlling blood pressure, and ACE2 receptor protein is widely found in the lungs, heart, kidneys, and intestines; in addition, ACE2 is also a membrane transport chaperone to help amino acid transporter protein B0AT1 membrane transport, and experiments have shown that the interaction between B0AT1 and ACE2 enables it to exist stably on the cell membrane (Bosch *et al.*, 2003). ACE2 can be stabilised in the cell membrane through interaction (Bosch *et al.*, 2003). In order to study the binding of S protein to ACE2, Zhou Qiang's team at Westlake University expressed the B0AT1-ACE2 complex, and then added the structural domain of S protein, RBD, to obtain the ternary complex of RBD-ACE2-B0AT1, which was resolved by cryo-electron microscopy to confirm that the RBD structural domain can recognise and bind human ACE2 protein (Yan *et al.*, 2020). The specific binding mode is that the PD structural domain of ACE2 binds to the RBD structural domain of S protein, and when the two RBDs are in the downward conformation, only one of them which is in the upward conformation can bind to the PD, and the two PD structures on the complex can each bind to an S protein, and the two PD structural domains on the B0AT1-ACE2 complex can each bind to an S protein.

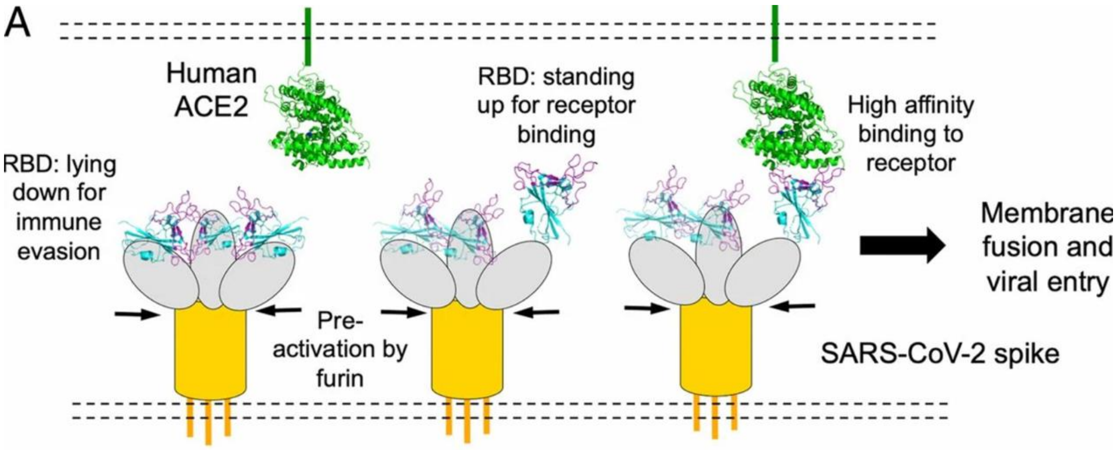
A
Human ACE2
RBD: standing up for receptor binding
High affinity binding to receptor
RBD: lying down for immune evasion
Membrane fusion and viral entry
Pre-activation by furin
SARS-CoV-2 spike

Figure 1.5 Schematic representation of the binding mode of RBD and ACE2(Shang *et al.*, 2020)

1.3 SARS-CoV-2 vaccine

Based on the pathogenesis of the SARS-CoV-2 Spike and ACE2 proteins, the S1 and RBD structural domains are important vaccine targets. Currently, researchers around the world have invested a great deal of effort in the research and development of SARS-CoV-2 vaccines, which mainly include five types: inactivated virus vaccine, recombinant protein vaccine, nucleic-acid vaccine, and live attenuated vaccine. vaccine, nucleic-acid vaccine, live attenuated vaccine, and viral vector-based vaccine. Inactivated viral vaccines, which are safer and more effective, are made by culturing and amplifying the new coronavirus, killing it, and then injecting the inactivated virus particles into the human body to induce an immune response.

The three current inactivated virus vaccines in China are produced by Sinopharm Zhongsheng Beijing, Sinopharm Zhongsheng Wuhan, and Beijing Kexing Zhongwei Biotechnology (Xia *et al.*, 2020). Recombinant protein vaccines are prepared by expressing purified pathogen antigenic proteins in engineered cells by means of genetic engineering and then prepared as vaccines. Novavax (Gao *et al.*, 2020) in the United States developed a SARS-CoV-2 vaccine using recombinant S protein nanoparticle vaccine; Another viral particle-like protein is a modified poxvirus vector used by GeoVax to activate the immune response by co-expression with the S protein. The recombinant neo-collagenic vaccine (CHO cells), jointly developed by the Institute of Microbiology of the Chinese Academy of Sciences and Anhui Zhifeilongkema, has been approved for use. The adenovirus vector vaccine developed by Concinol in collaboration with academician Chen Wei's team has been put into use (Callaway, 2020). The adenovirus vector vaccine is like a car carrying a cargo of "neo-coronavirus antigenic genes". When delivered to human cells, the "neo-coronavirus antigen gene" neither replicates itself nor has disease-causing ability but can express protein antigens of neo-coronaviruses and stimulate the human immune system to produce antibodies. However, these vaccines have problems such as multiple vaccinations and antiviral vector reactions (Zarkesh *et al.*, 2023). In addition to the problems with the vaccines, the mutation of the coronavirus itself is also of great concern, for example, in a study by Korber *et al.* (2020), it was found that a large number of amino acids i.e. 614 aa of the S protein, aspartic acid, were mutated to glycine, which may lead to the change of the conformation of the S protein and thus make the S protein more contagious and lethal (Dubey *et al.*, 2022). Upon further studies, more and more mutation sites have been found (Focosi *et al.*, 2023). Mutations in viral amino acids can affect the effectiveness of the vaccine and can even lead to inactivation of the vaccine. Since the safety issues caused by the continuous mutation of SARS-CoV-2 remain to be solved, and the existing vaccines have reduced immune effects or even inactivated vaccines against the mutants of SARS-CoV-2, the research of universal vaccines can provide a future proof solution.

In order to realise a universal vaccine, there are currently two main design ideas: one is a multivalent vaccine, which means that the vaccine is designed to contain antigens that can respond to different viruses, thus eliciting an immune response; the other is to make a vaccine using antigens that can stimulate an immune response against a variety of viruses. Previously, there has been the development of a universal vaccine against influenza virus, and the idea of influenza universal vaccine design is to first conduct sequence comparison to find mutation sites and conserved regions, and then determine potential target proteins in combination with functional changes brought about by the mutations, and then through codon optimization to get the genes that can express the proteins, and then enter into vaccine trials after the antigenic proteins have been expressed and tested to be able to bind to the vaccine, and finally take appropriate optimization solutions to the problems that arise in vaccine trials. Finally, optimized solutions to problems arising in vaccine trials are adopted (Bedi, Bayless & Glanville, 2023). Nowadays, universal vaccines for coronaviruses are also being developed. Ma *et al.* (2014), in a study of universal vaccines for MERS coronaviruses, expressed five RBD fragments of MERS-CoV and confirmed that they possessed the ability to bind to the receptor and elicit an immune response. In a study of a universal vaccine for different coronaviruses, Kovalenko *et al.* (2023) have shown that the RBD-Dimer of MERS-CoV has a more stable structure and is more effective in immunity, and it would be a good idea to apply this strategy to the second design of a universal vaccine, i.e., a universal vaccine against both beta-coronaviruses with only one chimeric antigen. Chimeric antigens can be created by exchanging structural domains such as homologous proteins encoded by different viruses to form a new protein that maintains the overall proximate structure but changes most of the primary sequence of the original antigen. Until now it has not been clearly reported that a universal vaccine has been developed against the new coronavirus mutants, probably due to mutations within the viral RBD binding domain, which further lead to a change in the way the S protein binds to ACE2, and if the RBD structural domain is chosen as the antigen it may not produce good immunity against some mutants of the virus, but although the RBD is the immunodominant structural domain of the S RBD is the immunodominant structural domain of S protein, which triggers a large number of neutralizing antibodies, but the other two main neutralization targets, the N-terminal structural domain (NTD) and the stem structural domain (S2 structural domain), can be used as the main binding domains of chimeric antigens. Therefore, in this experiment, partial fragment of the S2 subunit of the S protein was manipulated to achieve broad-spectrum immunity against beta-coronavirus with SARS-CoV-2.

1.4 Rationale of the study

Researchers have made significant progress in understanding the structure and function of the SARS-CoV-2 spike protein and in identifying mutations that can improve its binding to the human ACE2 receptor. However, there is still a need to develop effective vaccines and therapeutics that provide cross-protective immunity against other β-coronaviruses.

An existing research gap is the lack of understanding of the conserved subunits of β-coronaviruses that could be used to create novel chimeric spike proteins with broad-spectrum immunity. Although some studies have compared the spike proteins of SARS-CoV-2, SARS-CoV, and MERS-CoV, there is still a need for more comprehensive analyses on the utilization of the conserved regions of a wider range of β-coronaviruses. More studies are also needed on the immune responses triggered by chimeric spike proteins and their potential to cross-protect against other β-coronaviruses. Although several studies have demonstrated the efficacy of mRNA vaccines (e.g., BNT162b2) in eliciting neutralizing antibodies against pan-Sarbecovirus (Tan et al., 2021), more studies are needed to assess the long-term durability and breadth of the immune responses induced by these vaccines.

Current vaccines are designed based on the original SARS-CoV-2 spike protein, which may not provide optimal protection against emerging variants with mutations in the spike protein. By identifying functionally conserved subunits of β-coronavirus species, it may be possible to design new chimeric spike proteins that can trigger cross-protective immune responses against a wider range of coronaviruses

In addition, the identification of protective subunits may lead to the development of broadly neutralizing antibodies that can be directed against a wide range of coronaviruses. This approach has already shown promise in the development of pan-Sarbecovirus neutralizing antibodies in individuals using the Pfizer-BioNTech COVID-19 vaccine and previously infected with SARS-CoV-1 (Tan et al., 2021).

Overall, addressing gaps in knowledge regarding preferred mutations and conserved subunits of β-coronavirus species could have a significant impact on the development of more effective vaccines and therapeutics for neocoronaviruses, as well as future preparedness to deal with emerging coronaviruses.

1.5 Objective of the study

In this context, the present study aims to address the urgent need for future vaccine candidates against various beta-coronaviruses. The overall goal of the study was to identify conserved subunits in β-coronaviruses, integrate them into novel chimeric SARS-CoV-2 spike proteins, and computationally assess their stability to develop potential vaccine applications.

To achieve this goal, this study was guided by the following specific objectives:

1. Identify conserved subunits in beta-coronavirus spike proteins:

Identify conserved subunits in beta coronavirus spike proteins which will be used as prime candidates for integration into the SARS-CoV-2 spike proteins.

2. Enumerate conserved subunits and substitute them into the SARS-CoV-2 spike protein backbone:

Enumerate and replace conserved subunits from β-coronaviruses into the SARS-CoV-2 spike protein backbone.

3. Model structures using AlphaFold2:

Predict 3D structures of the novel chimeric spike proteins using AlphaFold software and test the software parameters for optimal prediction results.

4. Assess structure similarity:

Evaluate structure similarity between crystal structure of the SARS-CoV-2 S protein and novel chimeric S protein using Matchmaker tool in ChimeraX software.

5. **Evaluate stability through molecular dynamics:**

Evaluate structure stability of the novel chimeric spike proteins using molecular dynamic simulations.

By successfully achieving these specific goals, this study will endeavour to make substantial contributions to the field of viral immunology and vaccine development. Subsequent sections will provide insight into the methods employed to achieve these goals, computational strategies, and the anticipated impact of the findings on the development of robust, adaptable, and future-proof vaccine candidates against a wide range of coronaviruses.

1.6 Overview of the study

The methodology of the study is as shown below.

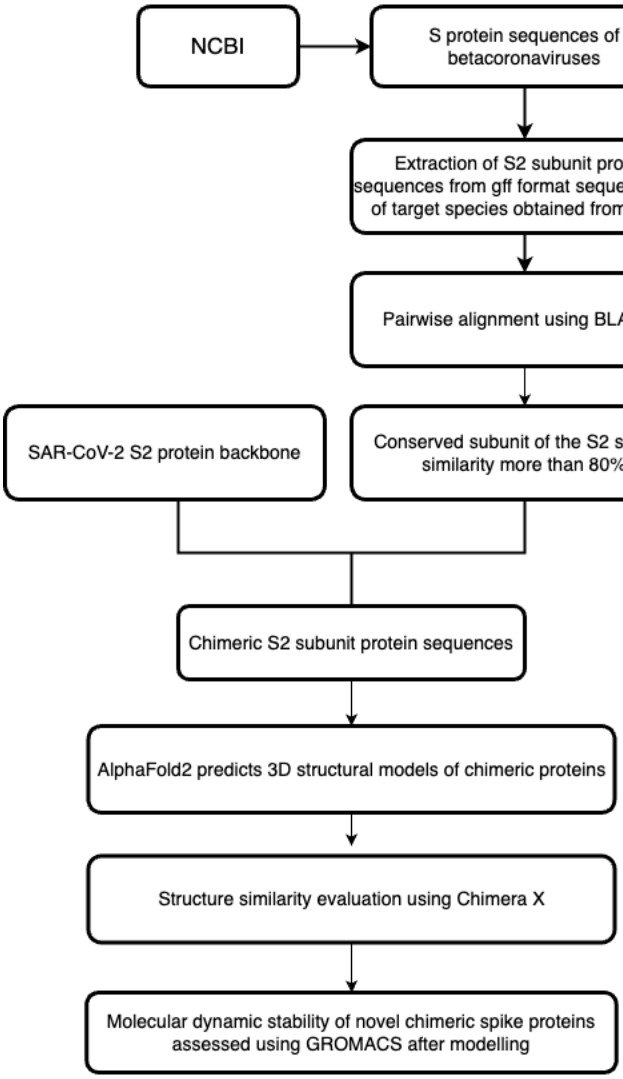
NCBI
S protein sequences of betacoronaviruses
Extraction of S2 subunit protein sequences from gff format sequence files of target species obtained from NCBI
Pairwise alignment using BLASTP
SAR-CoV-2 S2 protein backbone
Conserved subunit of the S2 subunit similarity more than 80%
Chimeric S2 subunit protein sequences
AlphaFold2 predicts 3D structural models of chimeric proteins
Structure similarity evaluation using Chimera X
Molecular dynamic stability of novel chimeric spike proteins assessed using GROMACS after modelling

Figure 1.6 Flowchart of the study.

CHAPTER 2

MATERIALS AND METHODS

2.1 Datasets and Software

2.1.1 Betacoronavirus species

The data for this study was obtained from the Virus Database hosted by the National Centre for Biotechnology Information (NCBI). NCBI provides a range of bioinformatics resources, including databases for biomedical literature (PubMed), genes (Gene), nucleotides (Nucleotide), and proteins (Protein). We curated a collection of S protein sequences from ten distinctive beta coronavirus species, spanning the years 2020 to 2023. Detailed information about these sequences is presented in the Table 2.1:

Table 2.1 List of representative human and non-human coronavirus isolates used for this study.

Coronavirus	Isolates	Year	Genome ID	S Protein ID
HCoV-229E	camel/Riyadh/Ry141/2015	2001	NC_028752.1	YP_009194639.1
SARS-CoV-1	Tor2	2002–2003	NC_004718.3	YP_009825051.1
HCoV- NL63	MP789 genomic sequence	2004	NC_005831.2	YP_003767.1
HCoV-HKU1		2005	NC_006577.2	YP_173238.1
HCoV-OC43	ATCC VR-759	2011	NC_006213.1	YP_009555241.1
MERS-CoV	HCoV-EMC/2012	2011	NC_019843.3	YP_009047204.1
SARS-CoV-2	Wuhan Hu-1	2020	NC_045512.2	YP_009724390.1
Civet-SARS-CoV	Civet007	2004	AY572034.1	AAU04646.1
Bat-SL-CoV	ZXC21	2015	MG772934.1	AVP78042.1
Pangolin-SL-CoV	MP789	2020	OM009282.1	UJZ92542.1

2.1.2 Specific sequence domains of beta coronavirus species used during enumeration domain exchanged –to generate chimeric proteins

Using SARS-CoV-2 as the cytoskeleton, the fusion peptide (FP), Heptad Repeat 1 (HR1) and Heptad Repeat 2 (HR2) sequence fragments of the S2 subunit of the S protein of each of 10 representative beta coronaviruses were exchanged to generate new chimeric protein sequences. The FP, HR1 and HR2 sequence fragments of the S2 subunit of each beta coronavirus are shown in Table 2.2 (refer to NCBI database):

Table 2.2 Site-specific and amino acid sequences of the FP, HR1 and HR2 fragments of the S2 subunit used for enumeration exchange.

	S2 subunit fragments					
Coronavirus	FP		HR1		HR2	
	Site	Sequence	Site	Sequence	Site	Sequence
HCoV-229E	654-672	AFTLANVSSFGDYNLSSVI	785-864	ENQKILAASFNKAMTNIVDAFTGVNDAITQTSQAIQTVATALNKIQDVVNQQGNALNHLTSQLRQNFQAISSSIQAIYDR	1046-1087	YTVPDLGIDQYNQTILNLTSEISTLENKSAELNYTVQRLQTL
SARS-CoV-1	770-788	MYKTPTLKYFGGFNFSQIL	900-965	ENQKQIANQFNKAISQIQESLTTTSTALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSR	1144-1185	PDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEL
HCoV-NL63	839-857	AFSLANVTSFGDYNLSSVL	970-1049	ENQKILAASFNKAINNIVASFSSVNDAITQTAEAIHTVTIALNKIQDVVNQQGSALNHLTSQLRHNFQAISNSIQAIYDR	1231-1272	YVKPNFDLTPFNLTYLNLSSELKQLEAKTASLFQTTVELQGL
HCoV-HKU1	868-892	GVTLSSNLNTNLHFDVDNINFKSLV	1003-1068	KNQKLIATAFNNALLSIQNGFSATNSALAKIQSVVNSNAQALNSLLQQLFNKFGAISSSLQEILSR	1248-1290	IAPNLTLNLHTINATFLDLYYEMNLIQESIKSLNNSYINLKDI
HCoV-OC43	866-890	VTLSTKLKDGVNFNVDDINFSPVL	1002-1067	QNQKLIANAFNNALYAIQEGFDATNSALVKIQAVVNANAEALNNLLQQLSNRFGAISASLQEILSR	1247-1287	VAPDLSLDYINVTFLDLQVEMNRLQEAIKVLNQSYINLKDI
MERS-CoV	856-875	SQSSPIIPGFGGDFNLTLLE	992-1057	ENQKLIANKFNQALGAMQTGFTTTNEAFQKVQDAVNNNAQALSKLASELSNTFGAISASIGDIIQR	1245-1286	SIPNFGSLTQINTTLLDLTYEMLSLQQVVKALNESYIDLKEL
SARS-CoV-2	788-806	IYKTPPIKDFGGFNFSQIL	918-983	ENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN	1162-1203	PDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQE