Vietnam Journal of Biotechnology **22**(3): 482-506, 2024. DOI: 10.15625/vjbt-21493

TAILORING POTENTIAL ANTIGENIC REGIONS ON PANDEMIC SARS SPIKE PROTEIN

Le Thanh Hoa[1,3], Le Nhat Thong[2,3], Le Minh Thong[1,3]✉

[1]*School of Biotechnology, International University, Ho Chi Minh City, Vietnam.*

[2]*Research Center for Infectious Diseases, International University, Ho Chi Minh City, Vietnam.*

[3]*Vietnam National University, Ho Chi Minh City, Vietnam.*

✉To whom correspondence should be addressed. E-mail: lmthong@hcmiu.edu.vn

Received: 10.07.2024
Accepted: 20.09.2024

ABSTRACT

Coronavirus-associated severe acute respiratory syndrome (SARS) pandemics have devastated lives, economies, and societies worldwide. Given the higher severity of the latter pandemic, the constant mutation, and vaccine escape, new and more dangerous pandemics could emerge. Therefore, it is imperative to identify conserved vaccine candidates for stable effectiveness in future pandemics. This study aimed to tailor potential, conserved peptide-based vaccine candidates for the upcoming Coronavirus pandemic based on the sequences of the spike protein of SARS-CoV-1 and SARS-CoV-2 viruses, using bioinformatic approaches, HLA epitope prediction software and literature support. Epitopes were selected based on sequence analysis of 166 sequences of pandemic strains, binding affinity to HLA molecules of different alleles and reactive antibodies collected from protein database. Seven candidate epitopes were chosen with conservation scores over 76/100 and up to 15 predicted HLA-DRB1 alleles. Over half of the residues of the epitopes interact with antibodies, suggesting good B-cell uptake. The epitopes possess at least 1 HLA-DRB1 allele that potentially provides protection against disease severity and play at least 1 role in the function of the spike protein. A combination of four candidate epitopes was estimated to cover nearly 90% of the world's population. The epitopes could be either modified to adapt to future pandemic strains, improve antigenicity, or used as booster immunization against the currently circulating SARS-CoV-2 variant. This study demonstrates that there is still room for improvement and promising discoveries in vaccine design to deter upcoming SARS pandemics.

Keywords: ACE2, antibody, HLA, reverse vaccinology, SARS-CoV-2, spike protein

INTRODUCTION

Two Coronavirus pandemics that caused severe acute respiratory syndrome (SARS) have occurred in 2002-2004 (Chen *et al.*, 2021) and from 2019 till the state of global

health emergency has been lifted (World Health Organization, 2023), respectively. The SARS-CoV-1 pandemic infected more than 8000 patients, with a rate of mortality of approximately 10% (Zhong *et al.*, 2003). The latter SARS-CoV-2 pandemic resulted in over 775 million cases and roughly 7 million deaths (World Health Organization, forthcoming). The pandemic also resulted in worldwide economic withdrawal, traveling restrictions, social distancing, and a burden on the healthcare system. In the case of the most recent pandemic, comorbidities (e.g., age (Mueller *et al.*, 2020), hypertension, obesity, and diabetes (Ng *et al.*, 2021), complications (e.g., septic shock (Li *et al.*, 2020), coagulation disorders (Vinayagam, Sattu, 2020), cytokine storm (Song *et al.*, 2020)) and persistent post-COVID syndrome (Oronsky *et al.*, 2023) have been widely reported. This fact demonstrated an increased pattern in the disease's pathological complexity, severity, and deadliness if newer strains emerge. Furthermore, the close relationship between these two strains (Chen *et al.*, 2021) implies the possible emergence of a new pandemic strain based on the gradual accumulation of mutations on the genome of the previous pandemic strain. Given the tremendous consequences of the previous pandemics and the possibility of pandemic re-emergence, it is, therefore, imperative to prepare vaccines to contain upcoming pandemics and mitigate mortality.

However, vaccine design is complicated by mutations in the pandemic SARS viruses. The mutations are driven by continuous co-evolution with the host organisms to improve the abilities of the viruses to replicate, transmit, infect, and escape vaccination (Chen *et al.*, 2020; Harvey *et al.*, 2021; Mlcochova *et al.*, 2021; Zhou *et al.*, 2021). For vaccine escape, its relationship with mutations was confirmed by reduced neutralization against new variants of SARS-CoV-2 in recipients of COVID-19 vaccines (Lazarevic *et al.*, 2021; Mlcochova *et al.*, 2021). Consequently, vaccine development efforts could waste money, time, labor, and resources if pathogen mutation and subsequent vaccine escape were not considered during vaccine development. This fact calls for identifying viral protein epitopes that are conserved across multiple variants of SARS pandemic strains to maintain the efficacy of the vaccine despite the genetic variation of the upcoming pandemic strain.

Among the proteins of Coronavirus, the spike protein (S protein) is highly immunogenic given a large number of anti-spike antibodies. Its location on the viral surface and its crucial role in cellular infection lend it a special attention in vaccine development. The S protein is composed of S1 and S2 subunits. To facilitate infection, the SARS-CoV S1 subunit recognizes the host cell receptor ACE2, and its S2 subunit fuses the viral envelope with the host cellular membrane (Huang *et al.*, 2020). There are two advantages to choosing spike protein for vaccine design. Firstly, as cellular infection, driven by S protein, occurs in the early stage of pathogenesis, vaccines that target S protein could prevent cellular infection and disease onset. Secondly, vaccines based on spike protein could elicit strong immune responses to block the S protein and then arrest viral infection. However, mutations across the spike protein can lead to immune evasion and vaccine escape.

The development of vaccines requires a careful curation of immunogens that could elicit the most effective and robust response. One of the usual design strategies is to

predict epitopes with high binding affinity to HLA molecules and incorporate these epitopes into the vaccines. Involved in the early stage of immune response and central in the development of adaptive immunity, the HLA molecules (human leukocyte antigen) (also known as Major Histocompatibility Complex - MHC) present immunogens to the T-cells. HLA class II molecules activate $CD4^+$ T-cells, which prime B-cells to produce antibodies (Janeway *et al.*, 2001). Moreover, long-lasting immunity can be achieved via the differentiation of some effector T-cells into long-lasting memory T-cells (Abbas *et al.*, 2014). Hence, vaccine design usually takes advantage of the HLA. Various programs have been dedicated to the prediction of HLA epitopes (Bassani-Sternberg, Gfeller, 2016; O'Donnell *et al.*, 2018; Peters, Sette, 2005; Rammensee *et al.*, 1999; Reynisson *et al.*, 2020) and formed the basis of the method in vaccine design studies for COVID-19, particularly those at the start of the pandemic (Dar *et al.*, 2020; Rahman *et al.*, 2020; Yazdani *et al.*, 2020).

Since T-cell epitopes are linear, HLA-based vaccines usually comprise immunogenic peptides. Peptide-based vaccines are also generally considered to have fewer adverse effects (Åsjö *et al.*, 2002; Elliott *et al.*, 2008; Gahery *et al.*, 2006; Kran *et al.*, 2004) than attenuated viral vaccines (Saeed *et al.*, 2021) and full-length protein-encoding vaccines such as mRNA vaccines or DNA vector vaccines (Andrzejczak-Grządko *et al.*, 2021; Meo *et al.*, 2021).

This study aimed to identify potential HLA-presented $CD4^+$ T-cell epitopes of the spike protein (referred as candidate epitopes) for vaccine development against future pandemic SARS Coronaviruses by employing a hybrid, structure-informed combination of bioinformatics and a reverse vaccinology approach applied for infectious diseases. HLA class II molecules were the focus of this study to generate antibodies against spike protein, thereby blocking the cellular infection and the onset of the disease. Among the class II HLA loci, DR alleles were chosen in this study to obtain potential epitope candidates for multiple reasons. At first, they are widely distributed across most populations (Arrieta-Bolaños *et al.*, 2023; Southwood *et al.*, 1998). Secondly, they were found to be associated with protection against COVID-19 (Astbury *et al.*, 2022; Langton *et al.*, 2021; Lehmann *et al.*, 2023; Littera *et al.*, 2020). Thirdly, as HLA class II molecules, DR molecules could stimulate antibody production to block infection-driven viral molecules and prevent the onset of the disease.

The design strategy was based on (1) sequence analysis of the spike protein of different variants for identification of conserved regions to reduce immune escape; (2) the prediction of interaction between all possible segments of the spike protein and HLA, herein HLA-DRB1 to attain broad-spectrum vaccine coverage for the world population; (3) comparison with the literature to support the level of confidence of the selected conserved epitopes for protective ability against disease severity, B-cell uptake propensity, immunogenicity and the sequences conservation of the candidate epitopes (sequence conservation for short) in the future pandemic strains. The resulting peptide-based candidate epitopes are expected to be safer compared to those using attenuated viruses and full-length protein-encoding molecules and more reliable than those without support from structural data or literature.

MATERIALS AND METHODS

Data collection

The sequences of SARS-CoV-1 spike protein were gathered from the NCBI Protein database (Wheeler *et al.*, 2000). Initially, we surveyed the list of proteins linked with all the entries of SARS—related strains of the NCBI taxonomy database. The protein records containing "complete genome" or "spike" were retained. The records with partial sequence or unpublished were removed. For all sequences published in journals, their corresponding articles were scanned for information specific to host organisms. Spike protein sequences reported from the articles that specified the host organism to be human were collected. These sequences were then clustered at 100% sequence similarity using CD-Hit (Li, Godzik, 2006) to remove redundant sequences. The sequences of SARS-CoV-2, including the reference sequences and those of variants of concern (VOCs), were collected from the reference alignment of spike proteins provided by the GISAID database (2022-08-15) (Khare *et al.*, 2021).

Sequence analysis

Initially, a multiple sequence alignment for SARS-CoV-1 spike sequences was generated. For SARS-CoV-2, a multiple sequence alignment containing the reference sequence and all VOCs was adapted from the GISAID data. A profile-to-profile alignment was made between the two above-mentioned multiple alignments of the spike protein of SARS-CoV-1 and SARS-CoV-2. All alignments were generated using ClustalX 2.1 (Larkin *et al.*, 2007) with substitution matrix BLOSUM62, gap penalty -1, gap extension penalty -0.5 and iteration per alignment step. Due to the sheer number of sequences present in the alignment, a summarized version was produced with matplotlib (Hunter, 2007) containing only the reference sequences of SARS-CoV-1, SARS-CoV-2, and the parental sequences of VOCs in the GISAID repository.

At each position (column) of the alignment, a conservation score was estimated for the degree of conservation, i.e., the unlikeliness for mutations between physicochemically dissimilar amino acids to occur at this position. The scores were calculated with the Scorecons server at the European Bioinformatics Institute website using the Valdar01 scoring method (Valdar, 2002), BLOSUM62 matrix, Karlin-like matrix transformation, and gapphilia 0. The scores were multiplied by 100 to yield a range of scores from 0.0 to 100.

For intuitive visualization of the potential peptide region, the distribution of the conservation scores was illustrated combined with other contexts across the sequence of the spike protein. Firstly, a distribution of the scores was graphed across the length of S protein sequence. In this distribution, the x-axis represents the residue position on the S protein sequence and the y-axis denotes the magnitude of the conservation scores. Additionally, a spatial distribution of the conservation scores was also visualized on the structure of the SARS-CoV-2 wild-type spike homotrimer, using PyMOL (Schrödinger, LLC, 2015) and PDB structure 6VXX (see below for detail). The magnitude of the conservation score of each amino acid in the structure was presented as color intensity, e.g., the higher the score, the darker the color intensity. Finally, a probability-like distribution of the number of positions concerning the different ranges of

conservation scores summarized the general degree of conservation of the spike protein. This summarized alignment was used as input to an evolutionary analysis program to select one SARS Coronavirus strain representing all the pandemic strains. This representative strain would be the input for HLA-DRB1 epitope prediction software for predicting the putatively high pathogenic epitope. The determination of the representative strain was carried out with MEGA11 (Tamura *et al.*, 2021). Pairwise estimation of evolutionary divergence of all pairs of strains of SARS-CoV-1, SARS-CoV-2, and VOCs was conducted using the p-distance model. The sequence of the spike protein with the least average divergence estimates was selected as the representative strain.

Prediction of potential epitopes binding to HLA-DRB

Among the sequence of S protein, peptide-based epitopes with the potential to become vaccine candidates were predicted by netMHCIIpan 4.0 (Reynisson *et al.*, 2020). This program predicts whether HLA-DRB1 molecules could present short peptides derived from antigenic proteins. It also informs potential alleles that encode the HLA-DRB1 molecules. Selected DRB1 alleles for prediction were the most common, as suggested by the Allele Frequency Net Database (Gonzalez-Galarza *et al.*, 2020) (www.allelefrequencies.net/top10dist.asp) (40 alleles excluding HLA-DRB1*04:140 and HLA-DRB1*04:155, as these alleles were not included in netMHCIIpan 4.0). We also added the allele HLA-DRB1*08:01 to this list. The spike protein was slid into windows of 15-residue peptides. The binding affinity of each peptide to each allele was graphed into one figure, and each curve represents a binding affinity profile of an HLA-DRB1 allele. A peptide will be deemed an epitope, and its associated HLA-DRB1 allele will be considered positive if the predicted binding affinity between them is under 50 nM. The total number of strong binding alleles of each peptide epitope was also visualized in the figure as the binding affinity.

Structural analysis for identifying antibody-reacted epitopes and selection of potential epitopes

In addition to epitopes identified by HLA-DRB interaction analysis, the potential list can be incorporated by antibody-reacted epitopes. Those can be extracted from 221 structures of complexes between antibodies and spike protein of SARS-CoV-2 deposited in the RCSB Protein Data Bank (Rose *et al.*, 2017). The complexes were analyzed by the Arpeggio program (Jubb *et al.*, 2017) to identify non-bonded interactions between pairs of residues, one from spike protein and one from the antibody of the same complex as the spike protein. The residues of the spike protein involved in the interactions (excluding those classified as clashes) were considered part of the antibody-reacted epitopes. From the list of the epitope residues, each residue was counted for the number of complexes where it was part of the antibody epitope and mapped into the profile. The higher this number is, the more likely the residue corresponds to binding to B-cell receptors, being taken up by the B-cells, and eliciting immunogenicity.

The epitopes with the most significant numbers of favorable alleles and many residues overlapping with antibody-reacted epitopes were selected as candidate epitopes for vaccine design. These epitopes are

reported in a table, along with their binding affinity to favorable HLA-DRB1 alleles and their conservation scores. The degree of conservation and the immunogenicity of each epitope were represented by the median conservation score, and median number of reactive antibodies, respectively. These medians were calculated from 15 residues that constitute each epitope. The selected epitopes were subjected to further cross-checking, as described below, to support their likelihood of protection against disease severity and sequence conservation.

Literature validation of potential epitopes

To validate the possible protection against SARS diseases of the candidate epitopes, protective favorable HLA-DRB1 alleles of each candidate epitope were identified by searching in the literature, using Google Scholar and keywords "immunogenetic," "population," "cohort," and "protective". If that allele was reported by any HLA-typing studies to be associated with protection against disease severity in a population or a cohort, the epitope related to that allele was deemed likely to hold protective properties.

For the functional constraint on the sequence conservation, the candidate epitopes were also searched on Google Scholar for research articles that reported structural and functional features related to the epitopes. Since the function of the spike protein is cellular infection, the search was focused on how the epitopes, at the level of molecular interactions, help the spike protein infect cells.

Estimation of coverage of the candidate epitopes across the human world population

With the list of potential candidate epitopes and their respective HLA-DRB1 alleles, the population coverage of these epitopes worldwide was estimated using the IEDB Population Coverage program (Bui *et al.*, 2006) (http://tools.iedb.org/population/). The program also generated a distribution graph for the proportion of the population covered by the combination of all epitope candidates (input into the program) in terms of the number of HLA alleles. It also reported results in the form of summarized statistics.

RESULTS AND DISCUSSION

Sequence alignment and identification of representative sequence

After data collection and clustering, 155 SARS-CoV-2 and 11 SARS-CoV-1 spike sequences were obtained and aligned. From the complete alignment containing 166 sequences, conservation scores were calculated to express the degree of conservation at each alignment position. The higher the score for one position is, the less likely that position is to mutate into physicochemically dissimilar amino acids. The score distribution was visualized in various contexts mentioned in Section 2.2. Note that sequence conservation allows a position on the alignment to have different amino acids as long as substitutions among them hold positive scores in the substitution matrices. That is because favorable substitutions between amino acids of similar physicochemical properties tend to correlate with a positive score on the substitution matrices (Ng and Henikoff, 2006; Pearson, 2018). This correlation can be seen in the PAM matrix (Jones *et al.*, 1992) or the BLOSUM matrix (Zimmermann, Gibrat,

2010) used in the alignment step of this study.

Regarding the subunits and domains of the spike protein, Figure 1A showed a fluctuating level of conservation in the S1 subunit and its receptor binding domain (RBD) and high conservation across the S2 subunit. This skew can be explained by the fact that the S1 subunit is more exposed at the viral surface than the S2 subunit (Walls *et al.*, 2020). Among the S1 subunit, the highly mutated RBD is located on the top of the spike trimer, constituting a protruding region well-exposed to the host immune surveillance (Li, 2016). As a result, it is the main target site of host antibodies, leading to a high mutation rate for immune evasion.

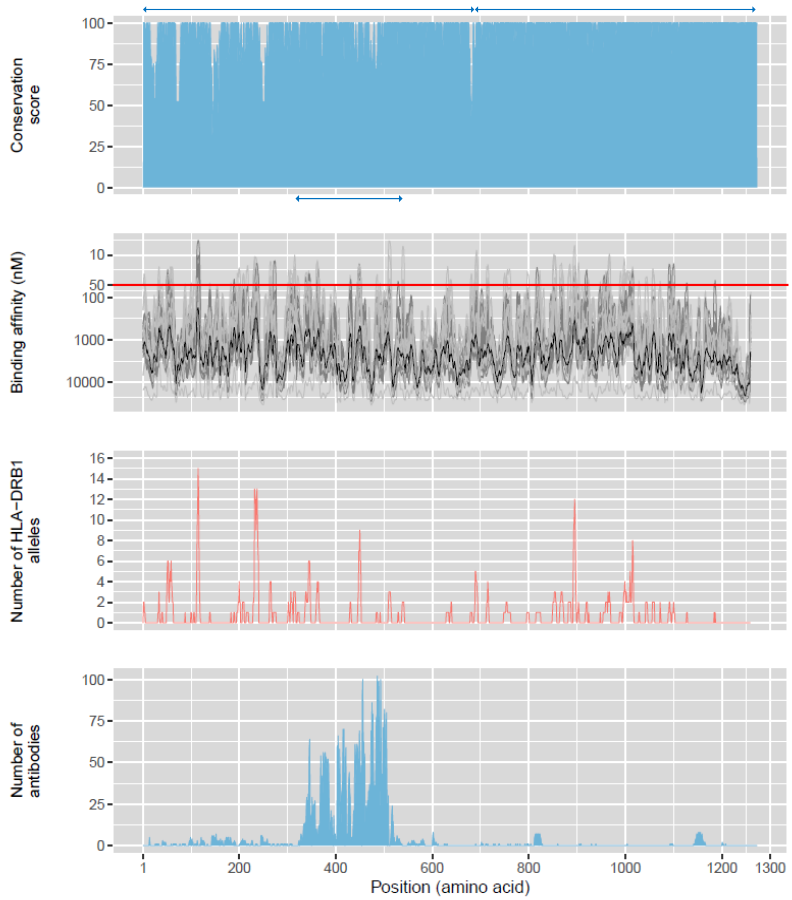

Figure 1. Overview of the distributions across the sequence of SARS-CoV-2 spike protein. A) conservation scores derived from the sequence alignment, B) binding affinity of S-derived peptides with HLA molecules encoded by 41 HLA-DRB1 alleles (with the average binding affinity of all the alleles as the black curve), C) number of favorable HLA-DRB1 alleles concerning the position of the peptide on the spike protein and D) number of antibody epitopes per position analyzed from PDB structures of spike protein-antibody complexes.

Figure S1 illustrates the distribution of the conservation score across the 3D structure of the spike protein. Conserved areas span the majority of the surface of the spike homotrimer (Figure S1A). Variable residues are mostly situated at the top of the spike

protein (Figure S1B), the most protruding area frequently targeted by antibodies and consequently under strong pressure to mutate for immune escape.

For an overview, a proportional distribution of the conservation score is provided in Figure S2. The spike protein has a high level of sequence conservation, with the distribution skewed towards the score of 100. Nearly 95% of the positions in the alignment have a conservation score equal to or higher than 70, while a conservation score of 100 (i.e., constituted of only one type of amino acid) makes up roughly 69% of the alignment, which is the highest proportion among the scores. The median conservation score is 100 with an interquartile range of 5.1. This result shows high conservation throughout most of the spike sequence across different pandemic strains and variants. This fact may look counterintuitive initially, given the 75% sequence identity between SARS-CoV-1 and SARS-CoV-2 (Huang *et al.*, 2020). However, it should be noted that the conservation score reflects sequence identity and sequence similarity. The latter notion is dictated by substitution matrices. Substitutions with positive scores tend to be between amino acids with similar physicochemical properties. Therefore, such substitutions can improve the conservation score without attaining sequence identity.

To determine the representative for HLA-DRB1 epitope prediction, multiple sequence alignment was used to estimate evolutionary divergence between pairs of strains and variants. The result is reported in Table S1. SARS-CoV-2 was estimated to have the least degree of divergence, so it was chosen as the input sequence for predicting candidate epitopes for downstream analysis.

Identification of high antigenic epitopes by prediction of HLA-DRB1 interactions

After gathering and analysis, data exported from netMHCIIpan 4.0 were plotted into Figures 1B and C and summarized in Table 1. In these figures, positive epitopes with high binding affinity and a high number of favorable HLA-DRB1 alleles can be seen across the sequence of the spike protein. Most of the functional regions of the spike protein (NTD, RBD, heptapeptide repeat 1 (HR1) and heptapeptide repeat 2 (HR2), except fusion peptide (FP)) were predicted to be able to be presented by HLA-DRB1 molecules. Hence, they can be immunogenic. Table 1 provided statistics for the binding affinity and the number of favorable alleles of these functional regions, calculated from the respective data of each region's epitopes. With the Kruskal-Wallis statistical test, no regions outperformed others in terms of binding affinity to HLA-DRB1 alleles. However, some regions are more promiscuous than others. The Mann-Whitney test revealed NTD to be more promiscuous than HR2 (p-value = 0.002). The RBD has a higher level of promiscuity than HR1 (p = 0.039) and HR2 (p-value = 0.013). These promiscuity statistics agree with the maximum number of favorable alleles in Figure 1. The FP has no positive HLA-DRB1 molecules; hence, it is deemed non-immunogenic. These analyses demonstrated the high immunogenicity of NTD and RBD, two domains of the S1 subunit, making them suitable for subunit vaccine design.

Table 1. Summary of the positive HLA-DRB1 alleles of each functional region of the spike protein in terms of the binding affinity and the highest number of the positive alleles.

Position	Functional regions	IC_{50} (nM)			Highest number of positive alleles
		Median	IQR	Min	
13-305	N-terminal Domain	31.8	18.4	4.6	15
319-541	Receptor Binding Domain	34.5	17.5	4.6	9
788-806	Fusion Peptide	-	-	-	-
912-984	Heptapeptide Repeat 1	30.5	17.1	9.6	3
1163-1213	Heptapeptide Repeat 2	44.0	4.2	39.8	1

IQR: interquartile range

Min: minimum

A list of epitopes with high potential to be vaccine candidates (Table 2) was tailored with Figure 1 and based on criteria. The criteria were the number of favorable HLA-DRB1 alleles, conservation score, and epitope residues overlapping with the antibody-reacted epitope. The last criterion is not present in Table 2 and will be reported in Table 3. All the selected epitopes belong to the NTD and RBD, both of which were found above to be the most promiscuous regions.

Table 2. List of the epitopes selected to be vaccine candidates.

Epitopes	Sequence	Positive HLA-DRB1 alleles				Conservation score	
		Number	IC_{50} (nM)			Med	IQR
			Med	IQR	Min		
59-73	FSNVTWFHAIHVSGT	4	34.6	16.5	19.3	89.1	47.9
115-129	QSLLIVNNATNVVIK	15	32.9	26.0	4.6	100	7.7
200-214	YFKIYSKHTPINLVR	4	21.5	26.0	20.2	82.1	21.6
237-251	RFQTLLALHRSYLTP	13	25.1	19.8	15.3	76.7	19.8
344-358	ATRFASVYAWNRKRI	6	35.4	3.4	23.6	100	3.7
364-378	DYSVLYNSASFSTFK	4	21.7	3.5	21.4	100	17.5
450-464	NYLYRLFRKSNLKPF	9	36.0	18.5	18.7	100	20.5

Med: median

IQR: interquartile range

Min: minimum

Although epitope 59-73 has a low number of favorable HLA-DRB1 alleles, it was selected due to its unique HLA-DRB1 alleles to raise vaccine coverage of the world population to nearly 90% (see Section 3.6). Epitopes 200-214 and 364-378 have the same limitation but were selected for reasons stated in Section 3.3.

The following section lists the immunogenic profile of the candidate epitopes obtained from the structural analysis of spike protein-antibody complexes.

Incorporating antibody-reacted epitopes

The chance of selecting candidate epitopes that can contact B-cell receptors and elicit immunogenicity could be improved by incorporating the antibody-reacted epitopes. The antibody-reacted epitopes were obtained by analyzing the 3D structure of spike protein-antibody complexes. There are two reasons behind this approach. Firstly, the contact between the candidate epitopes and B-cell receptors triggers cellular uptake of the epitopes, paving the way to antigen processing and immunogenicity. Secondly, the pieces of experimental data could elevate the confidence in the predicted results of this study. For these reasons, the result of HLA-DRB1 epitope prediction (Figure 1B, C) was combined with the distribution of the number of antibodies per position of the spike protein in Figure 1D. Table 3 provided further immunogenic data about the antibody epitopes of each candidate epitope.

Figure 1D shows the abundant antibody-reacted epitopes in RBD (among residues 319-541). Notably, the number of antibodies per position substantially climbs up around and within the receptor binding motif (among residues 438-506). This motif is where direct contact with ACE2 is made, making it a perfect and favorite target for neutralizing antibodies. Apart from RBD, only NTD is also targeted by the antibodies, albeit to a lesser extent.

From Table 3, the residues being part of antibody epitopes make up more than half the length of 3 candidate epitopes. All residues, or almost all of them, are involved in antibody-reacted epitopes for the other four epitopes. The higher the number of candidate-epitope residues participating in antibody epitopes, the better the chance for cellular uptake and the more binding interactions of the antibodies induced by the candidate epitopes towards the designated epitopes. All the candidate epitopes overlap much of the antibody epitopes, so they are likely to bind well with B-cell receptors and trigger immunogenicity.

Table 3. Immunogenicity revealed from structural analysis of the antibody-spike protein complexes. The immunogenicity of each candidate epitope is measured as the number of its residues, which are also part of the antibody epitopes, and the median number of antibodies per residue of the epitopes.

Epitopes	Number of candidate-epitope residues being part of antibody-reacted epitope	Number of antibodies per candidate-epitope residue		
		Median	IQR	Maximum

59-73	8	1	1	1
115-129	8	1	2.5	5
200-214	14	2	2	4
237-251	8	1	5	6
344-358	14	22	8	64
364-378	15	37	27.5	56
450-464	14	48	37.5	100

IQR: interquartile range

On the other hand, the number of antibodies per candidate epitope reflects the strength of immunogenicity. The closer a candidate epitope is to RBD (among residues 319-541), the higher the number of antibodies per candidate-epitope residue is. The RBM-housed epitope 450-464 has the highest number, which agrees with the correlation between RBM and the number of antibodies.

Despite having a low number of favorable HLA-DRB1 alleles (table 2), epitopes 59-73, 200-214, and 364-378 were chosen for their high number of residues being a part of antibody epitopes. This fact raises the chance for the vaccine that incorporates them to be taken in by B-cells.

Protective properties revealed by literature

Many studies designed vaccines against SARS-CoV-2 using HLA epitope prediction tools but did not verify their proposed epitopes for protective ability against the disease. Consequently, the proposed vaccines may not be effective. They may even result in disease severity, as some HLA-DRB1 alleles were associated with adverse outcomes in SARS-CoV-2 patients (Lehmann *et al.*, 2023; Littera *et al.*, 2020). These issues could be resolved by incorporating immunogenetic HLA-typing population studies to inform the design strategy. Such studies attempted to determine the HLA alleles that contribute to the protection against and the severity of COVID-19 (Astbury *et al.*, 2022; Langton *et al.*, 2021; Lehmann *et al.*, 2023; Littera *et al.*, 2020). The information from these studies can be utilized to select candidate epitopes capable of triggering protective immune responses rather than harmful side effects such as antibody-dependent enhancement. Therefore, incorporating immunogenetic studies into the candidate selection process is a valuable contribution to vaccine design.

The protective alleles were searched in the literature and compared with the list of predicted alleles of each candidate epitope exported from the result of netMHCIIpan 4.0. Many HLA-DRB1 alleles of the candidate epitopes were confirmed by previous studies to be protective against SARS-CoV-2, so they are likely to be presented by the HLA-DRB1 complex to T-cell receptors for subsequent development of immunity against the spike protein. Protective alleles reported in the literature and their corresponding linear epitopes predicted by our study are listed in Table 4.

Table 4. Protective HLA-DRB1 alleles of the epitope candidates predicted to bind to them and the reference associated with the alleles.

Linear epitope	Protective HLA-DRB1 alleles	Reference
59-73	DRB1*01:01 DRB1*15:01	(Lehmann *et al.*, 2023) (Astbury *et al.*, 2022)
115-129	DRB1*01:01 DRB1*04:01 DRB1*13:01	(Lehmann *et al.*, 2023) (Langton *et al.*, 2021) (Lehmann *et al.*, 2023)
200-214	DRB1*01:01	(Lehmann *et al.*, 2023)
237-251	DRB1*13:01 DRB1*15:01 DRB1*16:01	(Lehmann *et al.*, 2023) (Astbury *et al.*, 2022) (Littera *et al.*, 2020)
344-358	DRB1*01:01 DRB1*13:01 DRB1*16:01	(Lehmann *et al.*, 2023) (Lehmann *et al.*, 2023) (Littera *et al.*, 2020)
364-378	DRB1*01:01	(Lehmann *et al.*, 2023)
450-464	DRB1*01:01 DRB1*13:01 DRB1*16:01	(Lehmann *et al.*, 2023) (Lehmann *et al.*, 2023) (Littera *et al.*, 2020)

All candidate epitopes selected in this study were associated with at least one protective allele. Therefore, those epitopes can potentially trigger protective immune responses against SARS.

Likelihood for sustained conservation of the chosen epitopes

Many residues of the candidate epitopes have been conserved between SARS-CoV-1 and the SARS-CoV-2 pandemics. Such long-term sequence conservation implies an essential role of the epitopes in the function of the spike protein. These epitopes should be less likely to mutate in future strains to maintain the spike protein's function. This hypothesis is based on the observations that deleterious mutations cause reduction or loss of protein function. The organisms' fitness and viability are maintained with proper protein function. Although mutations can be advantageous, most yield neutral or deleterious effects (Eyre-Walker, Keightley, 2007; Soskine, Tawfik, 2010). As a result, regions of protein sequences that are important to the protein's function are more likely to be conserved.

All functional constraints of the candidate epitopes are summarized in Table 5. The constraints in the table also describe the candidate epitopes' contributions to the spike protein's function. Figures 2 and 3 illustrate the location of the epitopes near functional sites of the spike structure with PDB ID 6VXX.

Table 5. List of functional constraints placed on the candidate epitopes. The "residues of the epitope" column informs the residues on which the functional constraints are directly applied.

Epitopes	Residues of the epitope	Functional constraints	Reference
59-73	61-65	Beta sheet	PDB: 6VXX
	All	Interaction with sialic acid	(Unione et al., 2022)
115-129	117-122	Beta sheet	PDB: 6VXX
	125-129	Beta sheet	PDB: 6VXX
	122-125	Stabilizing a binding pocket of sialic acid	(Behloul *et al.*, 2020)
200-214	200-209	Beta sheet	PDB: 6VXX
237-251	237-244	Beta sheet	PDB: 6VXX
	241-251	Interaction with sialic acid	(Unione *et al.*, 2022)
344-358	346-349	Part of heparin binding site	(Clausen *et al.*, 2020)
	353-357	Part of heparin binding site	
364-378	All	Situated close to sialic acid binding pockets	(Li *et al.*, 2021)
	375, 378	Stabilizing sialic acid binding pockets	
450-464	453-458, 462	Binding to ACE2	(Yang *et al.*, 2021; Zhang *et al.*, 2021; 2022)

For epitopes 59-73, 115-129, 200-214, and 237-251 situated in the NTD, many residues of theirs form beta sheets, a crucial element to the tertiary structure formation and stability of the NTD (Figure 2). As side chains are involved in the stability of beta sheets (Hutchinson *et al.*, 1998; Merkel *et al.*, 1999), mutation in this epitope is likely discouraged as it could disrupt the beta sheets and affect the overall structure of NTD. Therefore, conservation (in the forms of sequence identity or favorable substitutions in terms of physicochemical properties) of the residues inside this epitope is likely desirable to maintain the structure and function of NTD.

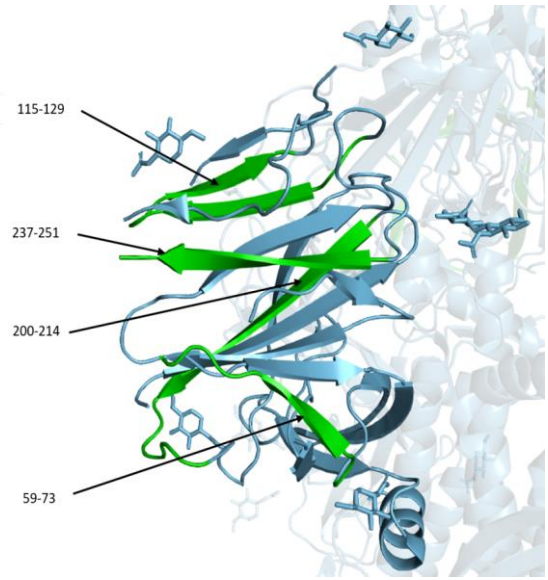

Figure 2. Location of 4 potential epitopes (green) in the NTD (opaque blue) of the spike protein (transparent blue). Note the beta-sheet segments (thin, long green sheets) inside these epitopes.

The sialic acid associated with host cells helps bring the pandemic SARS Coronaviruses close to the surface of the host cells (Seyran *et al.*, 2021). Epitope 115-129 contains the loop N122-N125 (which is also a NxxN conserved motif). The loop was predicted to be inside a binding pocket of sialic acid and stabilize it (Behloul *et al.*, 2020). Epitopes 59-73 and 237-251 overlapped with two regions on NTD that were predicted to interact with sialic acid during a 1.75-microsecond molecular dynamic simulation (Bò *et al.*, 2021). Finally, epitope 364-378 is situated close to two putative sialic acid binding pockets, which were determined by 200 ns of molecular dynamics (Li *et al.*, 2021). This study also reported a stabilizing role of residues 375 and 378 in the binding pocket. Because of this, these residues and others of this epitope that are adjacent to or spatially close to them should be physicochemically conserved to avoid unwanted destabilizing effects to the binding affinity to sialic acid of these pockets.

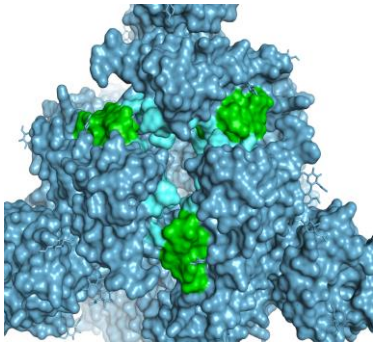

Figure 3. The location of three copies of epitope 364-378 (green) (each on one monomer of the spike protein (teal blue)) near putative sialic acid binding pockets (cyan).

Another host-cell molecule, heparin, contributes to cellular infection. Inside one putative heparin binding site on RBD are residues from 346 to 349 and from 353 to 357 of epitope 344-358 (Clausen *et al.*, 2020). Using transmission electron microscopy, the predisposition of RBD to bind to ACE2 when it is bound to heparin was detected. Also, in the study, when treated with SARS-CoV-2 virus and a mixture of heparin lyases, Vero E6 cells were less infected by nearly five times. This observation indicated the role of heparin in cellular infection.

The receptor binding domain (RBD) of the S1 subunit is the domain that interacts with ACE2 (Li, 2016). Epitopes 450-464 in the receptor binding motif of the receptor binding domain are conserved across SARS-CoV-2 VOCs and contribute to the interaction between RBD and ACE2. Conservation across this epitope suggests that its residues are essential to the function of RBD, which is to bind to ACE2. This fact was confirmed by molecular dynamics studies where about half of the residues of the epitope (Y453, R454, L455, F456, R457, H/K458, R/K462) (Yang *et al.*, 2021; Zhang *et al.*, 2021; Zhang *et al.*, 2022) exhibited negative decomposed binding free energy when they are in binding conformation with ACE2. This functional constraint should keep them conserved in future pandemic strains, as mutations in these positions might negatively affect the spike protein's binding affinity to ACE2.

Since multiple candidate epitopes are located near the binding site of infection-facilitating molecules (i.e., ACE2, sialic acid, and heparin), antibodies or designed therapeutics that recognize such epitopes may be able to reduce infection and prevent disease onset.

Predicted protective coverage of the candidate epitopes across the world population

The IEDB Population Coverage tool was used to estimate the protective coverage of the selected peptides for the world population. The tool also generated the probability distribution of the number of HLA alleles recognized by the population (Figure 4) and summarized it with the maximum and average values (Table 6). A coverage of 89.65% of the world's population can be obtained using all seven epitopes. The maximum number of alleles was 13 (Figure 4), and the average value was 3.87 (Table 6). Interestingly, the population coverage is still the same with the combination of only four epitopes 59-73, 115-129, 237-251, and 450-464. However, the four epitopes have a lower profile of recognized HLA alleles than seven epitopes (maximum number of 8 and 2.86 for average number). If epitopes 59-73 were removed from the four epitopes mentioned earlier, the remaining three were predicted to bind to 13-15 HLA-DRB1 alleles per epitope and could cover 87.65% of the world population. This result indicates these epitopes as must-have candidates for vaccine design.

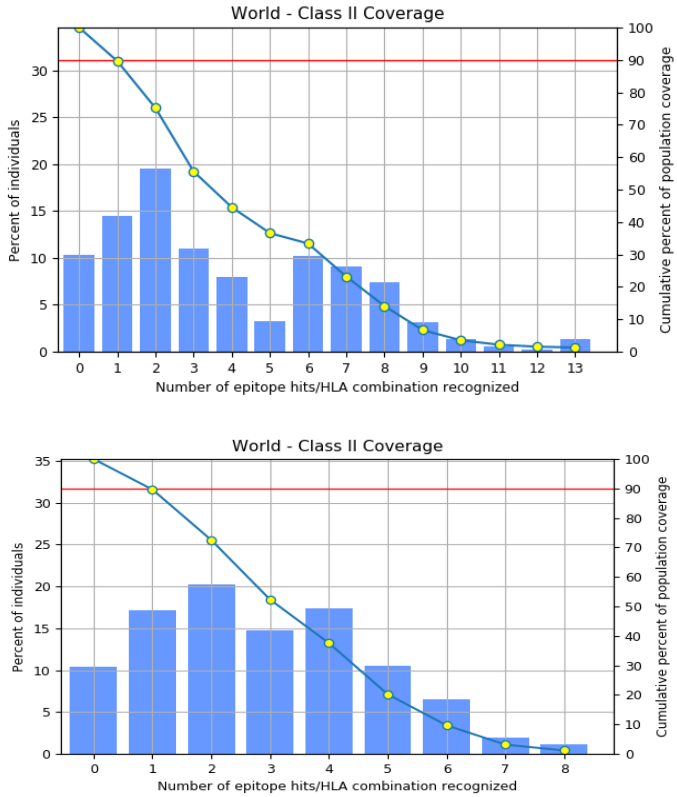

Figure 4. Distributions of the proportion of world population coverage (vertical bars) with all candidate epitopes listed in Table 2 (top graph) versus the combination of 4 candidate epitopes 59-73, 115-129, 237-251, and 450-464 (bottom graph). Cumulative distribution of the coverage percentage of the world population is provided as the blue curve with yellow dots. The graphs were provided by the IEDB Population Coverage tool.

Table 6. Comparison of different combinations of HLA-DRB1 epitopes in terms of the percentage of coverage and the average number of HLA-DRB1 alleles recognized by the world population.

Epitopes	**Coverage [a]**	**Average hit [b]**
All	89.65%	3.87
59-73, 115-129, 237-251 and 450-464	89.65%	2.86

[a] projected population coverage
[b] average number of epitope hits / HLA combinations recognized by the population

Future direction and recommendation

We have identified conserved, immunogenic peptide-based epitopes for vaccine design against future pandemic Coronavirus. The epitopes could make potential and indispensable candidates for vaccine design for several reasons. Firstly, they are peptide-based, and are generally less allergenic and toxic than attenuated vaccines or those based

on the expression of full-length antigenic proteins. Next, a large proportion of the world population could be covered with just a few of the candidate epitopes. Finally, compared to the epitopes of other reverse vaccinology studies during the height of the COVID-19 pandemic, our candidate epitopes are supported by structural, immunogenic and functional data. We propose them to be utilized as immunogenic scaffolds, by modifying their sequences either to adapt to the sequence of future pandemic Coronavirus or to improve their immunogenicity. Alternatively, the epitopes can be administered for booster vaccination against the currently circulating SARS-CoV-2 variant.

Due to the theoretical nature of our study, the following next steps should be taken to translate the candidate epitopes into practical solutions for global health. The first step is molecular dynamics simulation to estimate the binding affinity between the epitopes and antibodies. Good binding affinity to antibodies helps the epitopes to bind well with B-cell receptors for B-cell uptake. Because after the uptake, the epitopes are processed and presented to HLA molecules, we will measure the binding affinity between the epitopes and HLA molecules. Finally, we would like to collaborate with clinical researchers for clinical trials to assess the immunogenicity, protection against disease severity and safety of the vaccine candidates.

The strength of this reverse vaccinology study was lent from vast amount of and the variety of biological data. Hence, it can be argued that reverse vaccinology will benefit from the accumulation of a variety of biological data. There are two ways the data can augment the effectiveness of this vaccine design approach. One, continuous expansion of the public database of experimental HLA-epitope affinity characterization can improve the predictive power of artificial intelligence models behind the HLA epitope prediction methods. Two, each type of data guides the identification and selection of vaccine candidates in different ways. Sequence data helps identify epitopes conserved across a large number of variants. Antibody-antigen structural data inform B-cell uptake ability and immunogenicity of the epitopes. Immunogenetic data gathered from population HLA typing studies can point to protective HLA alleles against disease severity. And functional data support sequence conservation of the vaccine candidates. Therefore, vaccine design projects should utilize different types of biological data to determine the most advantageous vaccine candidates. As a result, continuous development of biological databases is highly recommended and can accelerate vaccine development in response to the emergence of any pandemic.

CONCLUSION

The increase in severity, deadliness, and socioeconomic impact of the latter pandemic Coronavirus strains compared to the former one and the threat of pandemic re-emergence prompt the search for vaccine candidates for future pandemics. The desirable vaccines require its epitopes to have a generally good degree of conservation and a high binding affinity to the antigen-presenting HLA molecules. Using sequence analysis and reverse vaccinology, we have detected short conserved epitopes on the spike protein of pandemic strains and variants, with high binding affinity to numerous HLA-DRB1 alleles. These epitopes were cross-checked with structural, immunogenetic and

functional data from many studies published throughout the COVID-19 pandemic. Compared to other vaccine design studies (to our knowledge), the advantage and novelties of this study are less adverse effect compared to traditional or full-length protein-encoding vaccines and the cross-referencing with previous studies. Our study contributes a list of potential epitopes for vaccine design against the SARS pandemic viruses. The study also demonstrates that tailoring potential antigenic regions for vaccine candidates to prepare for a possibly upcoming pandemic should be an ongoing effort that can lead to attractive and promising results.

ACKNOWLEDGEMENTS

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

We would like to give special thanks to Dr. Ly Le at International University, Viet Nam National University for her advice and suggestion for the method and the discussion sections of this article.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Abbas AK, Lichtman AH, Pillai S (2014) *Cellular and Molecular Immunology, 8th Edition*. Elsevier Health Sciences.

Andrzejczak-Grządko S, Czudy Z, Donderska M (2021) Side effects after COVID-19 vaccinations among residents of Poland. *European review for medical and pharmacological sciences* 25: 4418–4421. https://doi.org/10.26355/eurrev_202106_26153

Arrieta-Bolaños E, Hernández-Zaragoza DI, Barquera R (2023) An HLA map of the world: A comparison of HLA frequencies in 200 worldwide populations reveals diverse patterns for class I and class II. *Front Genet* 14: 866407. https://doi.org/10.3389/fgene.2023.866407

Åsjö B, Stavang H, Sørensen B, Baksaas I, Nyhus J, Langeland N (2002) Phase I Trial of a Therapeutic HIV Type 1 Vaccine, Vacc-4x, in HIV Type 1-Infected Individuals with or without Antiretroviral Therapy. *AIDS Research and Human Retroviruses* 18(18): 1357–1365. https://doi.org/10.1089/088922202320935438

Astbury S, Reynolds CJ, Butler DK, Muñoz-Sandoval DC, Lin K-M, Pieper FP, Otter A, Kouraki A, Cusin L, Nightingale J, Vijay A, Craxford S, Aithal GP, Tighe PJ, Gibbons JM, Pade C, Joy G, Maini M, Chain B, Semper A, Brooks T, Ollivere BJ, McKnight Á, Noursadeghi M, Treibel TA, Manisty C, Moon JC, Investigators* Covid, Valdes AM, Boyton RJ, Altmann DM (2022) HLA-DR polymorphism in SARS-CoV-2 infection and susceptibility to symptomatic COVID-19. *Immunology* 166(1): 68–77. https://doi.org/10.1111/imm.13450

Bassani-Sternberg M, Gfeller D (2016) Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions. *The Journal of Immunology* 197(6): 2492–2499. https://doi.org/10.4049/jimmunol.1600808

Behloul N, Baha S, Shi R, Meng J (2020) Role of the GTNGTKR motif in the N-terminal receptor-binding domain of the SARS-CoV-2 spike protein. *Virus Research* 286: 198058. https://doi.org/10.1016/j.virusres.2020.198058

Bò L, Miotto M, Di Rienzo L, Milanetti E, Ruocco G (2021) Exploring the Association Between Sialic Acid and SARS-CoV-2 Spike

Protein Through a Molecular Dynamics-Based Approach. *Front Med Technol* 2: 614652. https://doi.org/10.3389/fmedt.2020.614652

Bui H-H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A (2006) Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics* 7(1): 153. https://doi.org/10.1186/1471-2105-7-153

Chen J, Wang R, Wang M, Wei G-W (2020) Mutations Strengthened SARS-CoV-2 Infectivity. *Journal of Molecular Biology* 432(19): 5212–5226. https://doi.org/10.1016/j.jmb.2020.07.009

Chen Z, Boon SS, Wang MH, Chan RWY, Chan PKS (2021) Genomic and evolutionary comparison between SARS-CoV-2 and other human coronaviruses. *Journal of Virological Methods* 289: 114032. https://doi.org/10.1016/j.jviromet.2020.114032

Clausen TM, Sandoval DR, Spliid CB, Pihl J, Perrett HR, Painter CD, Narayanan A, Majowicz SA, Kwong EM, McVicar RN, Thacker BE, Glass CA, Yang Z, Torres JL, Golden GJ, Bartels PL, Porell RN, Garretson AF, Laubach L, Feldman J, Yin X, Pu Y, Hauser BM, Caradonna TM, Kellman BP, Martino C, Gordts PLSM, Chanda SK, Schmidt AG, Godula K, Leibel SL, Jose J, Corbett KD, Ward AB, Carlin AF, Esko JD (2020) SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2. *Cell* 183(4): 1043-1057.e15. https://doi.org/10.1016/j.cell.2020.09.033

Dar HA, Waheed Y, Najmi MH, Ismail S, Hetta HF, Ali A, Muhammad K (2020) Multiepitope Subunit Vaccine Design against COVID-19 Based on the Spike Protein of SARS-CoV-2: An In Silico Analysis. *Journal of Immunology Research* 2020(1): 8893483. https://doi.org/10.1155/2020/8893483

Elliott SL, Suhrbier A, Miles JJ, Lawrence G, Pye SJ, Le TT, Rosenstengel A, Nguyen T, Allworth A, Burrows SR, Cox J, Pye D, Moss DJ, Bharadwaj M (2008) Phase I Trial of a CD8+ T-Cell Peptide Epitope-Based Vaccine for Infectious Mononucleosis. *Journal of Virology* 82(3): 1448–1457. https://doi.org/10.1128/jvi.01409-07

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* 8(8): 610–618. https://doi.org/10.1038/nrg2146

Gahery H, Daniel N, Charmeteau B, Ourth L, Jackson A, Andrieu M, Choppin J, Salmon D, Pialoux G, Guillet J-G (2006) New CD4+ and CD8+ T Cell Responses Induced in Chronically HIV Type-1-Infected Patients After Immunizations with an HIV Type 1 Lipopeptide Vaccine. *AIDS Research and Human Retroviruses* 22(7): 684–694. https://doi.org/10.1089/aid.2006.22.684

Gonzalez-Galarza FF, McCabe A, Santos EJM dos, Jones J, Takeshita L, Ortega-Rivera ND, Cid-Pavon GMD, Ramsbottom K, Ghattaoraya G, Alfirevic A, Middleton D, Jones AR (2020) Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research* 48(D1): D783–D788. https://doi.org/10.1093/nar/gkz1029

Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ, Robertson DL (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 19(7): 409–424. https://doi.org/10.1038/s41579-021-00573-0

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

Hunter JD (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9(03): 90–95. https://doi.org/10.1109/MCSE.2007.55

Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN (1998) Determinants of strand

register in antiparallel β-sheets of proteins. *Protein Science* 7(11): 2287–2300. https://doi.org/10.1002/pro.5560071106

Janeway CA Jr, Travers P, Walport M, Shlomchik MJ (2001) B-cell activation by armed helper T cells. In *Immunobiology: The Immune System in Health and Disease. 5th Edition*. Garland Science

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8(3): 275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL (2017) Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology* 429(3): 365–371. https://doi.org/10.1016/j.jmb.2016.12.004

Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, Ho J, Lee RT, Yeo W, Curation Team GC, Maurer-Stroh S (2021) GISAID's Role in Pandemic Response. *China CDC Wkly* 3(49): 1049–1051. https://doi.org/10.46234/ccdcw2021.255

Kran A-MB, Sørensen B, Nyhus J, Sommerfelt MA, Baksaas I, Bruun JN, Kvale D (2004) HLA- and dose-dependent immunogenicity of a peptide-based HIV-1 immunotherapy candidate (Vacc-4x). *AIDS* 18(14): 1875–1883. https://doi.org/10.1097/00002030-200409240-00003

Langton DJ, Bourke SC, Lie BA, Reiff G, Natu S, Darlay R, Burn J, Echevarria C (2021) The influence of HLA genotype on the severity of COVID-19 infection. *HLA* 98(1): 14–22. https://doi.org/10.1111/tan.14284

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21): 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Lazarevic I, Pravica V, Miljanovic D, Cupic M (2021) Immune Evasion of SARS-CoV-2 Emerging Variants: What Have We Learnt So Far? *Viruses* 13(7): 1192. https://doi.org/10.3390/v13071192

Lehmann C, Loeffler-Wirth H, Balz V, Enczmann J, Landgraf R, Lakowa N, Gruenewald T, Fischer JC, Doxiadis I (2023) Immunogenetic Predisposition to SARS-CoV-2 Infection. *Biology* 12(1): 37. https://doi.org/10.3390/biology12010037

Li B, Wang L, Ge H, Zhang X, Ren P, Guo Y, Chen W, Li J, Zhu W, Chen W, Zhu L, Bai F (2021) Identification of Potential Binding Sites of Sialic Acids on the RBD Domain of SARS-CoV-2 Spike Protein. *Front Chem* 9: 659764. https://doi.org/10.3389/fchem.2021.659764

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

Li H, Liu L, Zhang D, Xu J, Dai H, Tang N, Su X, Cao B (2020) SARS-CoV-2 and viral sepsis: observations and hypotheses. *The Lancet* 395(10235): 1517–1520. https://doi.org/10.1016/S0140-6736(20)30920-X

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13): 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Littera R, Campagna M, Deidda S, Angioni G, Cipri S, Melis M, Firinu D, Santus S, Lai A, Porcella R, Lai S, Rassu S, Scioscia R, Meloni F, Schirru D, Cordeddu W, Kowalik MA, Serra M, Ragatzu P, Carta MG, Del Giacco S, Restivo A, Deidda S, Orrù S, Palimodde A, Perra R, Orrù G, Conti M, Balestrieri C, Serra G, Onali S, Marongiu F, Perra A, Chessa L (2020) Human Leukocyte Antigen Complex and Other

Immunogenetic and Clinical Factors Influence Susceptibility or Protection to SARS-CoV-2 Infection and Severity of the Disease Course. The Sardinian Experience. *Front. Immunol.* 11 https://doi.org/10.3389/fimmu.2020.605688

Meo SA, Bukhari IA, Akram J, Meo AS, Klonoff DC (2021) COVID-19 vaccines: comparison of biological, pharmacological characteristics and adverse effects of Pfizer/BioNTech and Moderna Vaccines. *Eur Rev Med Pharmacol Sci* 25(3): 1663–1669. https://doi.org/10.26355/eurrev_202102_24877

Merkel JS, Sturtevant JM, Regan L (1999) Sidechain interactions in parallel β sheets: the energetics of cross-strand pairings. *Structure* 7(11): 1333–1343. https://doi.org/10.1016/S0969-2126(00)80023-4

Mlcochova P, Kemp SA, Dhar MS, Papa G, Meng B, Ferreira IATM, Datir R, Collier DA, Albecka A, Singh S, Pandey R, Brown J, Zhou J, Goonawardane N, Mishra S, Whittaker C, Mellan T, Marwal R, Datta M, Sengupta S, Ponnusamy K, Radhakrishnan VS, Abdullahi A, Charles O, Chattopadhyay P, Devi P, Caputo D, Peacock T, Wattal C, Goel N, Satwik A, Vaishya R, Agarwal M, Mavousian A, Lee JH, Bassi J, Silacci-Fegni C, Saliba C, Pinto D, Irie T, Yoshida I, Hamilton WL, Sato K, Bhatt S, Flaxman S, James LC, Corti D, Piccoli L, Barclay WS, Rakshit P, Agrawal A, Gupta RK (2021) SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* 599(7883): 114–119. https://doi.org/10.1038/s41586-021-03944-y

Mueller AL, McNamara MS, Sinclair DA (2020) Why does COVID-19 disproportionately affect older people? *Aging* 12(10): 9959–9981. https://doi.org/10.18632/aging.103344

Ng PC, Henikoff S (2006) Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annual Review of Genomics and Human Genetics* 7(Volume 7, 2006): 61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630

Ng WH, Tipih T, Makoah NA, Vermeulen J-G, Goedhals D, Sempa JB, Burt FJ, Taylor A, Mahalingam S (2021) Comorbidities in SARS-CoV-2 Patients: A Systematic Review and Meta-Analysis. *mBio* 12(1): 10.1128/mbio.03647-20. https://doi.org/10.1128/mbio.03647-20

O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J (2018) MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *cels* 7(1): 129-132.e4. https://doi.org/10.1016/j.cels.2018.05.014

Oronsky B, Larson C, Hammond TC, Oronsky A, Kesari S, Lybeck M, Reid TR (2023) A Review of Persistent Post-COVID Syndrome (PPCS). *Clinic Rev Allerg Immunol* 64(1): 66–74. https://doi.org/10.1007/s12016-021-08848-3

Pearson WR (2018) Selecting the Right Similarity-Scoring Matrix. *Current Protocols in Bioinformatics* https://doi.org/10.1002/0471250953.bi0305s43

Peters B, Sette A (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* 6(1): 132. https://doi.org/10.1186/1471-2105-6-132

Rahman MS, Hoque MN, Islam MR, Akter S, Alam ASMRU, Siddique MA, Saha O, Rahaman MM, Sultana M, Crandall KA, Hossain MA (2020) Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2, the etiologic agent of COVID-19 pandemic: an in silico approach. *PeerJ* 8: e9572. https://doi.org/10.7717/peerj.9572

Rammensee H-G, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50(3): 213–219. https://doi.org/10.1007/s002510050595

Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC

eluted ligand data. *Nucleic Acids Research* 48(W1): W449–W454. https://doi.org/10.1093/nar/gkaa379

Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao Y-P, Valasatava Y, Voigt M, Westbrook JD, Woo J, Yang H, Young JY, Zardecki C, Berman HM, Burley SK (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research* 45(D1): D271–D281. https://doi.org/10.1093/nar/gkw1000

Saeed BQ, Al-Shahrabi R, Alhaj SS, Alkokhardi ZM, Adrees AO (2021) Side effects and perceptions following Sinopharm COVID-19 vaccination. *International Journal of Infectious Diseases* 111: 219–226. https://doi.org/10.1016/j.ijid.2021.08.013

Schrödinger, LLC (2015) *The PyMOL Molecular Graphics System, Version 1.8*.

Seyran M, Takayama K, Uversky VN, Lundstrom K, Palù G, Sherchan SP, Attrish D, Rezaei N, Aljabali AAA, Ghosh S, Pizzol D, Chauhan G, Adadi P, Mohamed Abd El-Aziz T, Soares AG, Kandimalla R, Tambuwala M, Hassan SkS, Azad GK, Pal Choudhury P, Baetas-da-Cruz W, Serrano-Aroca Á, Brufsky AM, Uhal BD (2021) The structural basis of accelerated host cell entry by SARS-CoV-2. *The FEBS Journal* 288(17): 5010–5020. https://doi.org/10.1111/febs.15651

Song P, Li W, Xie J, Hou Y, You C (2020) Cytokine storm induced by SARS-CoV-2. *Clinica Chimica Acta* 509: 280–287. https://doi.org/10.1016/j.cca.2020.06.017

Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11(8): 572–582. https://doi.org/10.1038/nrg2808

Southwood S, Sidney J, Kondo A, del Guercio M-F, Appella E, Hoffman S, Kubo RT, Chesnut RW, Grey HM, Sette A (1998) Several Common HLA-DR Types Share Largely Overlapping Peptide Binding Repertoires1 2. *The Journal of Immunology* 160(7): 3363–3373. https://doi.org/10.4049/jimmunol.160.7.3363

Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* 38(7): 3022–3027. https://doi.org/10.1093/molbev/msab120

Unione L, Moure MJ, Lenza MP, Oyenarte I, Ereño-Orbea J, Ardá A, Jiménez-Barbero J (2022) The SARS-CoV-2 Spike Glycoprotein Directly Binds Exogeneous Sialic Acids: A NMR View. *Angewandte Chemie* 134(18): e202201432. https://doi.org/10.1002/ange.202201432

Valdar WSJ (2002) Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics* 48(2): 227–241. https://doi.org/10.1002/prot.10146

Vinayagam S, Sattu K (2020) SARS-CoV-2 and coagulation disorders in different organs. *Life Sciences* 260: 118431. https://doi.org/10.1016/j.lfs.2020.118431

Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181(2): 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 28(1): 10–14. https://doi.org/10.1093/nar/28.1.10

World Health Organization (2023) *WHO Director-General's opening remarks at the media briefing – 5 May 2023*. (Accessed July 2024: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing---5-may-2023)

World Health Organization forthcoming. *COVID-19 cases*. (Accessed June 2024: https://data.who.int/dashboards/covid19/cases)

Yang Y, Zhang Y, Qu Y, Zhang C, Liu X-W, Zhao M, Mu Y, Li W (2021) Key residues of the receptor binding domain in the spike protein of SARS-CoV-2 mediating the interactions with ACE2: a molecular dynamics study. *Nanoscale* 13(20): 9364–9370. https://doi.org/10.1039/D1NR01672E

Yazdani Z, Rafiei A, Yazdani M, Valadan R (2020) Design an Efficient Multi-Epitope Peptide Vaccine Candidate Against SARS-CoV-2: An insilico Analysis. *Infection and Drug Resistance* 13: 3007–3022. https://doi.org/10.2147/IDR.S264573

Zhang Y, He X, Zhai J, Ji B, Man VH, Wang J (2021) In silico binding profile characterization of SARS-CoV-2 spike protein and its mutants bound to human ACE2 receptor. *Briefings in Bioinformatics* 22(6): bbab188. https://doi.org/10.1093/bib/bbab188

Zhang Z-B, Xia Y-L, Shen J-X, Du W-W, Fu Y-X, Liu S-Q (2022) Mechanistic Origin of Different Binding Affinities of SARS-CoV and SARS-CoV-2 Spike RBDs to Human ACE2. *Cells* 11(8): 1274. https://doi.org/10.3390/cells11081274

Zhong N, Zheng B, Li Y, Poon L, Xie Z, Chan K, Li P, Tan S, Chang Q, Xie J, Liu X, Xu J, Li D, Yuen K, Peiris J, Guan Y (2003) Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet* 362(9393): 1353–1358. https://doi.org/10.1016/S0140-6736(03)14630-2

Zhou B, Thao TTN, Hoffmann D, Taddeo A, Ebert N, Labroussaa F, Pohlmann A, King J, Steiner S, Kelly JN, Portmann J, Halwe NJ, Ulrich L, Trüeb BS, Fan X, Hoffmann B, Wang L, Thomann L, Lin X, Stalder H, Pozzi B, de Brot S, Jiang N, Cui D, Hossain J, Wilson MM, Keller MW, Stark TJ, Barnes JR, Dijkman R, Jores J, Benarafa C, Wentworth DE, Thiel V, Beer M (2021) SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* 592(7852): 122–127. https://doi.org/10.1038/s41586-021-03361-1

Zimmermann K, Gibrat J-F (2010) Amino acid 'little Big Bang': Representing amino acid substitution matrices as dot products of Euclidian vectors. *BMC Bioinformatics* 11(1): 4. https://doi.org/10.1186/1471-2105-11-4

APPENDIX

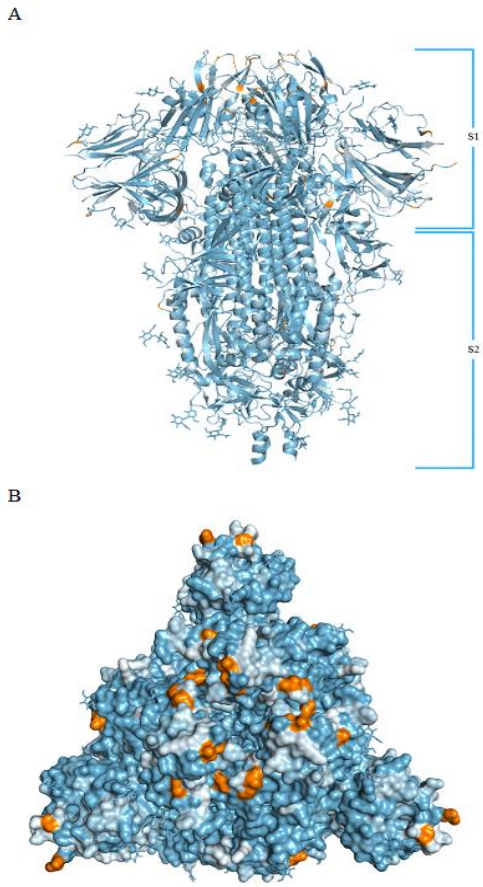

Figure S1. Distribution of the conservation scores throughout the homo-trimeric structure of the spike protein. The residues with conservation score equal to or higher than 70 are colored teal blue. The darker the teal blue, the higher the conservation score. Residues lower than 70 are colored orange and considered as non-conserved. (A) Front view of the spike protein, with brackets to approximately mark S1 and S2 subunits. (B) The surface of S1 subunit in top-down view.

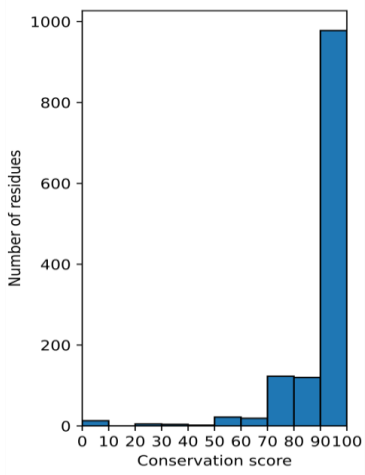

Figure S2. Distribution of conservation score from 0 to 100. Each bar illustrates the number of residues with their conservation scores being within the range of conservation scores of that bar.

Table S1. Estimates of evolutionary divergence between spike protein sequences of Coronavirus. Evolutionary analyses were conducted in MEGA11 (Tamura *et al.*, 2021). All ambiguous positions of the input alignment (Figure S1) were removed for each sequence pair using pairwise deletion option. The number of amino acid differences per position between sequences are shown in the table. The least average value of estimated evolutionary divergence is highlighted in blue color and belongs to SARS-CoV-2.

	SARS-CoV-1	**SARS-CoV-2**	**Alpha**	**Beta**	**Gamma**	**Delta**	**Omicron**
SARS-CoV-1		0.219	0.223	0.222	0.223	0.220	0.229
SARS-CoV-2	0.219		0.006	0.006	0.009	0.007	0.024
Alpha	0.223	0.006		0.008	0.012	0.010	0.024
Beta	0.222	0.006	0.008		0.009	0.011	0.024
Gamma	0.223	0.009	0.012	0.009		0.015	0.027
Delta	0.220	0.007	0.010	0.011	0.015		0.024
Omicron	0.229	0.024	0.024	0.024	0.027	0.024	
Average	0.223	0.045	0.047	0.047	0.049	0.048	0.059