Analysis of the full genome sequences of SARS-CoV-2 isolates to determine antigenic proteins and epitopes to be used for the development of a vaccine or a diagnostic approach for COVID-19

Hüseyin Can
Ege University
Ahmet Efe Köseoğlu
Ege University
Sedef Erkunt Alak
Ege University
Mervenur Güvendi
Ege University
Mert Döşkaya
Ege University
Muhammet Karakavuk
Ege University
Adnan Yüksel Gürüz
Ege University
Cemal Ün (cemaluen@gmail.com)
Ege University

Research Article

Keywords: SARS-CoV-2, reverse vaccinology in silico approach, vaccine development, diagnostic development

Posted Date: May 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-28142/v1

Version of Record: A version of this preprint was published on December 28th, 2020. See the published version at https://doi.org/10.1038/s41598-020-79645-9.

Abstract

In genome of SARS-CoV-2, the 5'-terminus encodes a polyprotein (pp1ab), which is further cleaved into 15 non-structural proteins (nsp-1 to nsp-10 and nsp-12 to nsp-16) whereas the 3' terminus encodes four structural proteins (spike, envelope, membrane, and nucleocapsid) and eight accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14). Among these 27 proteins, the present study aimed to discover likely antigenic proteins and epitopes to be used for during the development of a vaccine or serodiagnostic assay using a reverse vaccinology *in silico* approach. For this purpose, after the full genome analyses of SARS-CoV-2 isolates, viral surface proteins including spike, envelope and membrane proteins as well as proteins with predicted signal peptide were determined as probable vaccine candidates whereas the remaining were considered as possible antigens to be used during development of serodiagnostic assays. According to results, the phylogenetic analysis of SARS-CoV-2 isolates from 31 different countries showed two significant clusters in which one was clustered with China-Wuhan and the other one with China-Yunnan isolates. During the analyses, 105 SNPs were identified that resulted in change in 70 amino acid positions. Among the 27 proteins, 26 of them were predicted as probable antigen, except nsp-16. In 26 proteins, spike protein was selected as the best vaccine candidate because of having a signal peptide, negative grand average of hydropathicity value, one transmembrane helix, moderate aliphatic index, a big molecular weight, a long-estimated half-life, beta wrap motifs as well as having a stable, soluble and non-allergic features. In addition, orf7a, orf8 and nsp-10 proteins were considered as potential vaccine candidates because of having signal peptides. Nucleocapsid protein and a highly antigenic GGDGKMKD epitope of nucleocapsid protein were identified as ideal antigens to be used in development of serodiagnostic assays. Moreover, considering MHC-I alleles, highly antigenic KLNDLCFTNV and ITLCFTLKRK epitopes belonging to spike and orf7a proteins can be used to develop an epitope-based peptide vaccine or used as antigen for development of serodiagnostic assay.

Introduction

Coronaviruses belonging to the family Coronaviridae and the order Nidovirales are a large family of enveloped positive-strand RNA viruses. Coronaviruses are zoonotic pathogens that infect both animals and humans, and may cause diseases in intestinal, liver, respiratory, and nervous systems. It has been stated that, among known coronaviruses, CoV-229E (alpha coronavirus), CoV-NL63 (alpha coronavirus), CoV-OC43 (beta coronavirus), CoV-HKU1 (beta coronavirus), severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), and current SARS-CoV-2 can infect humans[1].

In genome of SARS-CoV-2, the 5'-terminus encodes a polyprotein (pp1ab), which is further cleaved into 15 non-structural proteins (nsp-1 to nsp-10 and nsp-12 to nsp-16) whereas the 3' terminus encodes four structural proteins including spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins and eight accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14)[1]. Comparative genomic analyses have revealed that SARS-CoV-2 shared more nucleotide homology with SARS-CoV than MERS-CoV[2,3]. SARS-CoV-2 has been also suggested as more adaptive to humans with a higher mutation rate than SARS-

CoV[4]. The novel coronavirus, SARS- CoV-2 causing COVID-19 has been first reported in Wuhan City of Hubei Province in China on December, 2019 and then, COVID-19 has spread from China to 211 different countries with more than 2.8 million cases including more than 195 thousand deaths in record time[5].

Since it has been known that no specific therapeutic agents that target SARS-CoV-2 are currently available, the development of an urgent vaccine against SARS-CoV-2 is inevitable. In vaccine development, traditional or recombinant vaccine methods are being used. Traditional approaches, based on inactivated or live attenuated viruses, can be applied for vaccine development but it has been reported that these approaches have some limitations such as being time-consuming, having problems in the production of non-abundant proteins and pathogens[6]. This condition prevents the development of new vaccines against pathogens causing the outbreaks leading to pandemic. On the other hand, to overcome these problems, new recombinant vaccine development strategies allowing several genes obtained from different pathogenic agents to be cloned, expressed and purified to be used as vaccine candidates are applied[7]. During new recombinant vaccine design, reverse vaccinology (RV) *in silico* approach provides detailed preliminary prediction about vaccine candidates by using genome sequences that can be ultimately translated into proteins. Utilisation of RV *in silico* approach is rather crucial because of offering a prediction for the antigenicity, the epitope regions of B and T cells as well as other parameters such as signal peptide, subcellular localisation, and solubility about targeted proteins[6,8]. Currently, docking analysis demonstrating the binding among predicted epitopes and selected alleles of MHC-I and MHC-II became an important part of RV *in silico* approach[6]. Result obtained from *in silico* prediction has utmost importance for preventing the failures that can be encountered at the end of wet-lab studies or even late stages of clinical trials.

In this context, the present study aimed to analyse the full genome of SARS-CoV-2 isolates from different countries compared with a seed genome (reference isolate Wuhan-Hu-1; Accession number: NC_045512.2) in order to determine the nucleotide variation(s). Secondly, we aimed to discover likely antigenic proteins and epitopes to be used for the development of a vaccine candidate or serodiagnostic assay using a RV *in silico* approach as previously described[6,9–11]. For this purpose, surface proteins including S, E and M proteins as well as proteins that were predicted to have a signal peptide were identified as probable vaccine candidates whereas the remaining were considered as probable antigens to be used in development of a serodiagnostic assay. During *in silico* analyses, physico-chemical parameters, secondary structure, subcellular localisation, transmembrane helices, antigenicity and signal peptide were predicted for 27 proteins of the seed genome. Also, signal peptide and antigenicity predictions were conducted for variant proteins whose reference proteins are structural protein and/or have a signal peptide. Variant proteins were determined according to amino acid alterations detected at least in two countries. Thereafter, allergenicity, BetaWrap motifs, similarity with host proteins, post translational modifications and B/T cell epitopes were predicted for proteins which are structural, variants of structural proteins and/or have a signal peptide. Finally, selected epitopes were docked with receptors of MHC-I/II alleles.

Results

Variations. The full genomes of 41 isolates from six continents were compared with a reference seed genome and a lot of codon alterations were detected in all open reading frames (orfs), except orf10. Orf1ab that contain 15 different coding regions was the largest fragment in genome and codon alterations were detected in 66 positions, resulting in 39 amino acid changes. For orf3a, codon alterations were detected in eight positions and seven of them were found to cause the amino acid changes. Fragments of orf6, orf7a and orf8 contained codon alterations in one, two and two positions, respectively. All codon alterations were detected to cause amino acid changes, except one in orf6.

The codon alterations were also detected in gene fragments encoding structural proteins including S, E, M and N proteins. These codon alterations detected in only structural proteins were depicted comprehensively in Table 1, 2 and 3. In the gene fragment of S, codon alterations were detected in eleven positions and nine of them were detected to cause amino acid changes. Especially codon alteration caused by the substitution of adenine for guanine at position 1841 was prevalent and detected in 17 different isolates from 14 countries (Table 1). The altered codon caused a change from aspartic acid to glycine. In the gene fragment of N, codon alterations were detected in 12 different positions. Among these codon alterations, three were prevalent and two of them were detected in the same six isolates from five countries (Georgia, Greece, Argentina, Colombia, Nigeria). One of them was caused by consecutive guanine to adenine transition at two positions 608 and 609, and resulted in arginine to lysine substitution whereas the other one was caused by a changing from guanine to cytosine at position 610 and resulted in glycine to arginine substitution. The third prevalent variation caused by substitution of cytosine to thymine was detected in 36 isolates from 29 countries and did not result in amino acid alteration (Table 2).

The fragments belonging to E and M were the smallest when compared to S and N proteins. In the gene fragment of E, two variations caused change in amino acid sequence and one of them was detected in two isolates whereas a codon alteration detected in the gene fragment of M protein did not result in amino acid change (Table 3).

Phylogenetic analysis. SARS-CoV-2 sequences from Turkey (EPI_ISL_424366), Canada- Ontario (EPI_ISL_418384), and Australia-Western Australia (EPI_ISL_420539) were grouped in a cluster while a sequence of Wuhan SARS-CoV-2 (NC_045512) and two sequences from Kuwait (EPI_ISL_416458, EPI_ISL_416541) were grouped in another cluster that two clusters diverge from the same node on the phylogenetic tree. China-Yunnan (MT049951) sequence was clustered with sequences from East and South Asian Countries including New Zealand (EPI_ISL_416538), Australia-Queensland (EPI_ISL_420879), India (MT050493), Japan (LC528233), and Singapore (EPI_ISL_414379) except a sequence from Chile (EPI_ISL_415661). Sequences from Chile (EPI_ISL_415661) and Japan (LC528233) were clustered together. Two sequences from Greece (EPI_ISL_418263, EPI_ISL_418264) were clustered together while two sequences from Georgia (EPI_ISL_420140, EPI_ISL_420144) were clustered in separate from each other. (Fig. 1).

Physico-chemical parameters. The number of amino acids varied from 75 to 1273 among structural proteins. The largest one was S protein with ~142 kDa whereas E protein with ~8.4 kDa was the smallest one (Table 4). Among non-structural proteins, except orf1ab, the amount of amino acids varied from 43 to 275. Orf3a with ~32 kDa was the largest protein whereas orf7b with 5.2 kDa was the smallest protein (Table 4). Each non-structural protein that is encoded by orf1ab was also analysed, and the amount of amino acids was detected to vary from 83 to 1945. Nsp-3 with ~218 kDa molecular weight was one of the largest proteins whereas the smallest one was nsp-7 with ~9.3 kDa size (Table 4). When all proteins encoded by full genome were analysed, theoretical PI value was between 4.6 and 10.07. Among structural proteins, only S protein was negatively charged whereas E, M and N protein were positively charged. In addition, orf7a, orf10, nsp-6, nsp-9, nsp-13, nsp-14 and nsp-16 proteins were positively charged whereas the remaining proteins were negatively charged except nsp-4 and nsp-8 that were neutral. The estimated half-life was 30 h for all proteins, except proteins that were encoded by orf1ab. Only nsp-1 in orf1ab had 30 h estimated half-life. According to instability index, N protein was found as instable while S, E, M structural proteins and most of the non-structural proteins were found as stable. Aliphatic index showed a significant variation ranging between 52.53 to 144 among all proteins. Grand average of hydropathicity value was found negative in S and N proteins as well as in most of the non-structural proteins that were encoded by orf1ab (Table 4).

Secondary structure. According to results obtained from structural proteins, alpha helix was between ~22 and 47%, that of extended strand was between ~10 and 22%, and that of random coil was between ~40 and 60%. For non-structural proteins, alpha helix varied between 0 and 69%, that of extended strand varied between ~3 and 47%, and that of random coil varied between ~28 and 58% (Table 5).

Antigenicity. All structural proteins were predicted as probable antigen. Antigenicity value varied from 0.4638 to 0.6298. E protein and its variant L37H had the highest antigenicity value whereas S protein had the lowest antigenicity value. Interestingly, all non-structural proteins were also predicted as probable antigen, except nsp-16 encoded by orf1ab. In addition, orf7b had the highest antigenicity value with 0.8462 among all proteins (Table 6).

Solubility. According to solubility prediction, S, E and N proteins were soluble. Among non- structural proteins, orf3a as well as nsp-2, 4, 7, 10, 12, 13, 14, 15 and 16 proteins encoded by orf1ab were predicted as insoluble. The solubility prediction of another protein, nsp-3 encoded by orf1ab, could not be retrieved due to large fragment size (Table 6).

Subcellular localisation and transmembrane helices. The number of transmembrane helices varied from 0 to 3 among structural proteins. The number of transmembrane helices was the lowest in N protein whereas it was the highest in M protein. Among non-structural proteins, although the number of transmembrane helices varied from 0 to 8, most of them had no transmembrane helices (Table 6). When subcellular localisation predictions were examined, S, M and N proteins were predicted to be in host endoplasmic reticulum. E as well as M proteins were also predicted to be in host cell membrane. Non-

structural proteins were predicted to locate in cell membrane, endoplasmic reticulum, cytoplasm, nucleus (Table 6).

Signal peptide. According to prediction of signal peptide based on four different parameters, only S protein and its variant were predicted to have a signal peptide during the analyses of structural proteins. Among non-structural proteins, orf7a, orf8, variant of orf8 and nsp-10 were predicted to have a signal peptide (Table 7).

Allergenicity. None of the proteins analysed showed allergenic properties for MEME/MAST motif and IgE epitopes (Table 8).

BetaWrap motifs. Among all proteins analysed, only S protein and its variant (D614G) were predicted to contain BetaWrap motifs (Table 8).

Similarity with Host Proteome. No significant similarity was predicted between analysed viral proteins and host proteins (Table 8).

B cell epitopes. A lot of linear B cell epitopes were predicted for S, variant S (D614G), N and variants of N (S197L and R203K/G204R), E and variant E (L37H), orf8 and nsp-10 proteins using Bcepred and IEDB. Epitopes that were predicted in both Bcepred and IEDB, and detected as probable antigen were presented in Table 9. Obtained predictions showed that nearly all epitopes had more antigenicity value than those of their own proteins. Among these analysed proteins, the highest antigenicity value (1.4530) was predicted for an epitope (GGDGKMKD) belonging to N protein and its variants. Another epitope (THTGTGQ) that had a high antigenicity value of 1.0789 was predicted in Nsp-10 encoded by orf1ab. Also, any antigenic epitope was not predicted for M and orf7a proteins. All predicted probable antigenic epitopes were depicted in Table 9.

MHC-I and MHC-II epitopes. A lot of MHC-I epitopes were predicted as probable antigen (Table 10). Antigenicity values belonging to epitopes were generally predicted higher than those of their own proteins. Among structural proteins, an epitope (KLNDLCFTNV) that had the highest antigenicity value (2.6927) was predicted in S protein and its variant (D614G). For non- structural proteins, an epitope (ITLCFTLKRK) in orf7a had the highest antigenicity value (2.5150). Any antigenic epitope was not predicted for nsp-10. On the other hand, KWPWYIWLGF, FLAFVVFLLV, FARTRSMWSF and RNRFLYIIKL, AQFAPSASAF and LGIITTVAAF epitopes belonging to S (including variant D614G), E (including variant L37H), M, N (including S197L and R203K/G204R) and orf8 (including L84S), respectively, had an IC50 value lower than 10 and a percentile rank varying from 0.02 from 0.1, indicating a strong binding among the epitope and MHC-I alleles.

Similarly, a lot of MHC-II epitopes were predicted as probable antigen (Table 11). Also, nearly all epitopes had higher antigenicity values than those of their own proteins. Among structural proteins, PTNFTISVTTEILPV and VTLAILTAHRLCAYC epitopes predicted in S protein (including variant D614G) and variant L37H had the highest antigenicity value. For non- structural proteins, orf7a had an epitope

(IVFITLCFTLKRKTE) that was predicted as a probable antigen with a high antigenicity value (1.8597). Any antigenic epitope was not predicted for nsp-10. Among MHC-II epitopes, although there were a lot of epitopes with low percentile rank, only one epitope that had an IC50 value lower than 10, indicating a strong binding among epitope and MHC-II alleles, was detected in orf8.

Post-translational modifications. S protein and its variant (D614G) were predicted to have highly N-glycosylated and phosphorylated sites as well as a few O-glycosylated and acetylated sites. M, E (including L37H), orf7a, and nsp10 proteins were predicted to have N-glycosylated and phosphorylated sites while orf7a was predicted to have an acetylation site. Orf8 and its variant (L84S) were predicted to have N-glycosylated and phosphorylated site whereas two additional phosphorylation sites, one of which locate in exposed surface and the other one is buried, were predicted in only variant L84S. In addition, N protein and its two variants (S197L and R203K/G204R) were predicted to have N-/O-glycosylated, phosphorylated and acetylated sites. When N protein and its variant were compared, O-glycosylation or phosphorylation sites were not detected in variant S197L whereas an extra acetylation site was detected in variant R203K/G204R.

Docking Analysis. All probable antigenic epitopes that have a low IC50 value and percentile rank could not be docked with their MHC-I or MHC-II alleles because of limitations associated with available MHC-I and MHC-II alleles variations in data bank or server. Accordingly, KWPWYIWLGF, KLNDLCFTNV, FLAFVVFLLV, LIFLWLLWPV, MEVTPSGTWL, FLIVAAIVFI and LEYHDVRVVL epitopes belonging to S (including variant D614G), E (including variant L37H), M, N (including variants S197L and R203K/G204R), orf7a and orf8 (including variant L84S), respectively, were docked with receptors of selected MHC-I alleles (Figs. 2, 3, and 4).

During docking analysis conducted by MHC-II alleles, in S protein, core regions of PTNFTISVTTEILPV, SIIAYTMSLGAENSV, and GYFKIYSKHTPINLV epitopes were docked with receptor of HLA-DRB1*07:01. Also, core region of another epitope (QDLFLPFFSNVTWFH) in S protein was docked with receptor of HLA-DRB1*15:01. In M protein, core regions of ASFRLFARTRSMWSF, RTLSYYKLGASQRVA and PKEITVATSRTLSYY epitopes were docked with receptor of HLA-DRB1*07:01. Also, core region of an epitope (QIAQFAPSASAFFGM) in N protein was docked with receptor of HLA- DRB1*07:01. Similarly, core region of an epitope (VTLAILTAHRLCAYC) in variant L37H was docking to receptor of HLA-DRB1*1501.

Discussion

Reverse vaccinology plays an important role in the development of recombinant vaccines by allowing in silico analyses of the genome of pathogens. *In silico* analyses enables identifying the highly antigenic and secreted proteins which are crucial in vaccine development before the beginning of the wet lab studies[35,6]. Using this approach, the present study aimed to discover likely antigenic proteins as well as epitope regions that are targeted by both B and T cell arms of the adaptive immune response for

development of a vaccine or serodiagnostic assay as described by Dangi at al. (2018)[6] and Goodswen et al. (2013)[35].

For this purpose, firstly full genome sequences belonging to 41 isolates from 31 different countries were compared in order to find the variations causing codon alterations and/or amino acid substitutions, and the association with each other. Viral distribution caused by human mobility or migration was predicted according to phylogenetic tree drawn. Accordingly, isolates from Turkey (EPI_ISL_424366), Canada-Ontario (EPI_ISL_418384), and Australia-Western Australia (EPI_ISL_420539) shared a common ancestor with isolates from Wuhan SARS-CoV-2 (NC_045512), and Kuwait (EPI_ ISL_416458, EPI_ISL_416541). Also, it was thought that the isolate from Chile (EPI_ISL_415661) might be sourced from Japan (LC_528233) because of their closest clustering.

Variants were detected in S, N, E and orf8 proteins. Accordingly, in S proteins, D614G variation was detected to be prevalent (Table 1). The comparison of reference S protein and its variant (D614G) showed no difference in antigenicity values, epitope regions and antigenicity values of epitopes (Table 9, 10 and 11). In addition, three variations (S197L, R203K/G204R) causing amino acid alterations were detected in N protein (Table 2). S197L variation was predicted to increase antigenicity value of N protein (Table 9) and thus, utilise of S197L variant was thought to be a better antigen for serodiagnostic studies conducted in countries harboring SARS-CoV-2 isolates with S197L variant. Similar result was also detected in E protein and higher antigenicity value was predicted in variant L37H (Table 3 and 9). Contrary to these results, for variant orf8 (L84S) detected in Japan, China (Yunnan), India, Chile, Australia- Queensland and New Zealand, a lower antigenicity value was predicted (Table 9). These results show the importance of assessment of sequence data obtained from regional isolates before initiated vaccine studies.

All proteins of SARS-CoV-2, except nsp-16 encoded by orf1ab, were predicted as probable antigen. Although, there was no major difference between predicted antigenicity values for probable vaccine candidate proteins, S protein was selected as a better vaccine candidate protein compared to others depending on *in silico* analyses results. Physico-chemical analysis showed that S protein had a negative GRAVY value indicating that S protein is hydrophilic and has a better interaction with surrounding water molecules[36]. Also, it had stable and soluble characteristics which are important parameters for biophysical studies on epitope-based vaccine design. Moreover, S protein had a moderate aliphatic index which indicates stability in a wide spectrum of temperature[37], fewer than two transmembrane helices facilitating cloning, expression and purification[9] and a big molecular weight and long estimated half-life (more than 10 h). These properties show that S protein can be used as a vaccine candidate antigen. In addition to these physico-chemical properties, other predicted parameters such as the presence of a signal peptide that increase the immune response and the presence of betawrap motifs that are a virulence factor, as well as a non-allergic property also showed that S protein was a better vaccine candidate. In addition, orf7a, 8 and nsp-10 proteins were predicted to have a signal peptide and it is known that the presence of a signal peptide on any protein is an important parameter which indicates

that the protein can be destined towards the secretory pathway[38,39]. Accordingly, these probable secreted three proteins (orf7a, 8 and nsp-10) were also considered as potential vaccine candidate proteins. As S, orf7a, orf8 and nsp-10 proteins were examined with regard to secondary structure, random coil were detected higher than 49%. The presence of this highly predicted random coil shows that these proteins can be preferably recognised by an antibody[40]. Another critical point for these proteins was the prediction of post-translational modifications. The presence of these modifications indicates that if these proteins are produced by recombinant technology, eukaryotic expression systems such as yeast, insect or mammalian should be preferred instead of bacterial systems[41].

In previous vaccine studies, S and M proteins have been used for the development of DNA or recombinant protein vaccines against SARS-CoV that affected 30 countries in five continents[42,43]. Also, S protein has been used to develop a vaccine against MERS CoV which is another zoonotic pathogen that has infected approximately 2500 people in over 25 countries[5,44]. According to the results obtained from these studies, S and M proteins were reported to induce a strong immune response. Consequently, these findings of recent studies and our *in silico* study support that only S proteins can be strong vaccine candidate protein in development of recombinant vaccine against SARS-CoV-2 causing COVID-19.

Since N protein does not locate at the surface of SARS-CoV-2, it was thought that N protein may not be a proper vaccine candidate but could be a good antigen for serodiagnosis of COVID-19 because of having a negative GRAVY value and soluble characteristics and not transmembrane helices. There were several previous coronavirus (SARS-CoV) related studies supporting our predictions. For example, a previous study reported strong antibody response against recombinant N protein in 10 of 12 SARS patients[45]. In a different study, a B cell epitope region between 156 and 175 positions of N protein reacted strongly with sera from SARS patients[46].

In the second part of our study, epitope regions specific to B and T cells were predicted in all structural proteins, variants of structural proteins and non-structural proteins that have a signal peptide and, antigenicity control was performed for all predicted epitopes. Results associated with B cell epitopes showed that there were a lot of highly antigenic epitopes. Antigenicity value was very high for GGDGKMKD, THTGTGQ, and NLDSKV epitopes corresponding to N, nsp-10 encoded by orf1ab and S proteins. Similarly, epitopes that have high antigenicity values were also predicted for MHC-I and II alleles. Among these predicted epitopes, for MHC-I alleles, KLNDLCFTNV (Fig. 2) and ITLCFTLKRK epitopes belonging to S and orf7a proteins had very high antigenicity values whereas for MHC-II alleles, PTNFTISVTTEILPV and IVFITLCFTLKRKTE epitopes belonging to S and orf7a proteins also had significant antigenicity values.

These findings indicate that a cocktail/mixture composed of these epitopes may induce neutralizing antibody response or can be used in development of an epitope-based peptide vaccine because of their association with both B and T cells. Also, it was thought that they can be used as antigens that capture IgM and IgG antibodies against SARS-CoV-2 during viral infection in ELISA or Western blotting tests. In previous wet lab studies, the presence of neutralizing epitopes has been reported to bind with S protein of

SARS-CoV[47–49]. For example, in a study conducted in mice, major neutralisation determinant was reported in receptor binding domain (RBD) of S protein in SARS-CoV[47]. Another study reported that NYNWKR epitope in S protein had a neutralizing effect against SARS-CoV[48]. There are also some new studies using wet lab techniques and *in silico* approaches associated with SARS-CoV-2. In a study, splenocytes were stimulated with plenty of T cell epitopes belonging to S protein and nine of them were reported to induce cellular immune response. Among these epitopes, only one of them (VGGNYNYLYRLFRKS; between 445 and 459 positions) was inside RBD, five of them (YNYKLPDDFTGCVIA; DDFTGCVIAWNSNNL;mVVLSFELLHAPATVC; LLHAPATVCGPKKST; KNKCVNFNFNGLTGT) were located nearby RBD whereas the remaining three (SFPQSAPHGVVFLHV; PHGVVFLHVTYVPAQ; FTTAPAICHDGKAHF) were inside S2 segment of S protein[50]. Interestingly, in our study, an epitope (CYFPLQSYGF; between 488 and 497 positions) with a relatively lower antigenicity value were predicted in RBD of S protein and four epitopes (NLDSKV, KLNDLCFTNV, RQIAPGQTGK, GDEVRQ) were also predicted in a very close region. Docking results supported that KLNDLCFTNV epitope was targeted by HLA-A*02:01 allele (Fig. 2). These findings indicate that the above-mentioned epitopes may have promising neutralizing effect against SARS-CoV-2. In a previous *in silico* study, five different epitopes (SYGFQPTNGVGYQPY; SQSIIAYTMSLGAEN; IPTNFTISVTTEILP; AAAYYVGYLQPRTFL; APHGVVFLHVTYVPA) related to both MHC-I and II were predicted in S protein[51] and only one of them overlapped with a highly antigenic epitope (PTNFTISVTTEILPV) predicted in our study. In another *in silico* study, 14 epitopes were predicted in S protein for T cells[52] and six of them were detected to overlap with epitopes predicted in our study. However, none of these overlapped epitopes were among the significant epitopes identified in this study.

Conclusion

In conclusion, our study shows that vaccine candidate proteins can be selected from among a lot of proteins of SARS-CoV-2 using a reverse vaccinology *in silico* analyses approach. Depending on our *in silico* results, S protein was the best vaccine candidate protein. In addition, orf7a, orf8 and nsp-10 proteins were promising vaccine candidate proteins because of having a signal peptide. Moreover, epitopes predicted in S protein and other proteins having a signal peptide may have a potential neutralizing effect and can be used to develop an epitope-based peptide vaccine or a serodiagnostic assay.

Methods

Strains of SARS-CoV-2. NCBI (National Center for Biotechnology Information) (https://www.ncbi.nlm.nih.gov) and GISAID (Global Initiative on Sharing All Influenza Data) (https://www.gisaid.org) were used to obtain the full genome of different SARS-CoV-2 isolates, with at least 5 samples from each continent to represent the whole world. Accession number of each isolate was given in Supplementary S1.

Genome and Proteome Variations. Whole genome sequences of SARS-CoV-2 isolates from different countries were aligned by MEGA 7 software with a seed genome (Accession number: NC_045512.2) to detect variations corresponding to amino acid sequences[12]. For phylogenetic analysis, aligned sequences were trimmed and edited by MEGA 7 and BioEdit (Version 7.2)[13] softwares, respectively. A phylogenetic tree was reconstructed by MEGA 7 software based on Neighbour Joining method using Tamura-Nei Gamma distribution (TN93+G) model[14] with 500 Bootstrap replications.

Prediction of physico-chemical parameters and secondary structures. Seed genome proteins were investigated using Expasy ProtParam online server (https://web.expasy.org/protparam/) for the prediction of physico-chemical properties[15]. The prediction of solubility was performed by SolPro (http://scratch.proteomics.ics.uci.edu/)[16]. Also, prediction of secondary structures was performed by GOR IV online server (https://npsa-prabi.ibcp.fr/cgi- bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html)[17].

Prediction of Antigenicity. Seed genome proteins as well as variant proteins and predicted epitopes were analysed by Vaxijen v2.0 online server (http://www.ddg- pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) for the prediction of antigenicity using a threshold value of 0.4[18].

Prediction of Subcellular Localisation and Number of Transmembrane Helices. The subcellular localisation of virus proteins in infected host cells were predicted by Virus-mPLoc (http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/)[19]. For the prediction of the number of transmembrane helices, TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) was used[20].

Prediction of Signal Peptide. Seed genome proteins and variant proteins whose reference protein is a structural protein and/or has a signal peptide were analysed by Signal-BLAST (http://sigpep.services.came.sbg.ac.at/signalblast.html)[21].

Prediction of Allergenicity. The allergenicity of proteins which are structural, variants of structural and/or have a signal peptide was predicted by Algpred online server (http://crdd.osdd.net/raghava/algpred/) using a prediction approach of MEME/MAST motif and IgE epitopes[22].

Prediction of BetaWrap Motifs. The prediction of BetaWrap Motifs of proteins which are structural, variants of structural and/or have a signal peptide was carried out by BetaWrap online server (http://cb.csail.mit.edu/cb/betawrap/betawrap.html)[23].

Prediction of Similarity with Host Proteome. The proteins which are structural, variants of structural and/or have a signal peptide were examined by BlastP (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) to predict the similarity with the host proteome. In analysis, *Homo sapiens* was selected as a host organism.

Prediction of post-translational modifications. The prediction of post-translational modifications of proteins which are structural, variants of structural and/or have a signal peptide were carried out using

NetNGlyc 1.0 server (http://www.cbs.dtu.dk/services/NetNGlyc/)[24], NetOGlyc 4.0 server (http://www.cbs.dtu.dk/services/NetOGlyc/)[25], NetPhos 3.1 server (http://www.cbs.dtu.dk/services/NetPhos/)[26] and, GPS-MSP and GPS-PAIL running under CSS- Palm Online Service (http://csspalm.biocuckoo.org/online.php)[27]. In addition, NetSurfP 2.0 (http://www.cbs.dtu.dk/services/NetSurfP/) was used for the prediction of surface accessibility of post-translational modification sites in proteins[28].

Prediction of B cell Epitopes. Linear B cell epitopes of proteins which are structural, variants of structural and/or have a signal peptide were predicted by Bcepred (http://crdd.osdd.net/raghava/bcepred/)[29] and Bepipred Linear Epitope Prediction 2.0 running under IEDB (the immune epitope database, https://www.iedb.org/)[30] online servers.

Prediction of MHC-I and MHC-II epitopes. The prediction of MHC-I and MHC-II epitopes of proteins which are structural, variants of structural and/or have a signal peptide were analysed by IEDB (https://www.iedb.org/)[30]. For the prediction of MHC-I epitopes, twelve different MHC-I alleles (A01.01, A02.01, A03.01, A24.02, A26.01, B07.02, B08.01, B27.05, B39.01, B40.01, B58.01 and B15.01) which are HLA super-type representative were utilised in the analysis. For the prediction of MHC-II epitopes, seven different MHC-II alleles (DRB1.03.01, DRB1.07.01, DRB1.15.01, DRB3.01.01, DRB3.02.02, DRB4.01.01 and DRB5.01.01) were used in the analysis.

Docking Analysis with MHC-I and II alleles. For docking analyses conducted with MHC-I alleles, receptor alleles that were specific to each epitope were retrieved from The Protein Data Bank (PDB; http://www.rcsb.org/pdb/). In selection of MHC-I receptor models, the presence of free (undocked) 3D protein structures were considered. Models of epitopes that were selected based on low IC50 value and being probable antigen were predicted by I-TASSER Server (http://zhanglab.ccmb.med.umich.edu/I-TASSER)[31]. In addition, epitopes that have the highest antigenicity value also selected for docking. Each modelled epitope ligand was docked to its specific MHC-I allele receptor by ClusPro Server (https://cluspro.bu.edu/home.php)[32] and visualised on UCSF Chimera 1.14 tool[33]. For docking analyses conducted with MHC-II alleles, epitope models that were selected based on low IC50 values and being probable antigen were predicted by I-TASSER Server (http://zhanglab.ccmb.med.umich.edu/I-TASSER). Each modelled epitope ligand was docked to its specific MHC-II allele by selecting specific alleles from the EpiDock Server (http://www.ddg-pharmfac.net/epidock/EpiDockPage.html)[34].

Declarations

Author contributions: H.C., A.E.K., S.E.A., M.G., M.D., M.K., A.Y.G., and C.U. designed research; H.C., A.E.K., S.E.A., M.D., and C.U. wrote the paper and H.C., A.E.K., S.E.A., M.D., M.K., A.Y.G. and C.U. conducted review and editing.

Competing interests

The authors declare that they have no competing interests.

Funding

This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe.* **27**(3), 325–328, https://doi.org/10.1016/j.chom.2020.02.001 (2020).
2. Xu, X. *et al.* Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *China Life Sci.* **63**(3), 457– 460, https://doi.org/10.1007/s11427-020-1637-5 (2020).
3. Zheng, J. SARS CoV-2: an Emerging Coronavirus that Causes a Global Threat. *J. Biol. Sci.* **16**(10), 1678–1685, https://doi.org/10.7150/ijbs.45053 (2020).
4. Ye, Z. W. *et al.* Zoonotic origins of human coronaviruses. *J. Biol. Sci.* **16**(10), 1686– 1697, https://doi.org/10.7150/ijbs.45472 (2020).
5. *WHO (World Health Organization)*. at, https://who.int (Accessed: 27 April, 2020).
6. Dangi, M., Kumari, R., Singh, B. & Chhillar, A. K. Advanced In Silico Tools for Designing of Antigenic Epitope as Potential Vaccine Candidates Against In *Bioinformatics: Sequences, Structures, Phylogeny* (ed. Shanker, A.) 329–357 (Springer, Singapore, 2018).
7. Nascimento, I. P. & Leite, L. C. Recombinant vaccines and the development of new vaccine strategies. *J. Med. Biol. Res.* **45**(12), 1102–1111, https://doi.org/10.1590/s0100- 879x2012007500142 (2012).
8. Can, H., Alak, S. E., Köseoğlu, A. E., Döşkaya, M. & Ün, C. . Do Toxoplasma gondii apicoplast proteins have antigenic potential? An in silico study. *Biol. Chem.* **84**, 107158, https://doi.org/10.1016/j.compbiolchem.2019.107158 (2020).
9. Meunier, M. *et al.* Identification of Novel Vaccine Candidates against Campylobacter through Reverse Vaccinology. *Immunol. Res.* **2016**, 5715790, https://doi.org/10.1155/2016/5715790 (2016).
10. Nazir, Z., Afridi, S. G., Shah, M., Shams, S. & Khan, A. Reverse vaccinology and subtractive genomics-based putative vaccine targets identification for Burkholderia pseudomallei *Microb. Pathog.* **125**, 219–229, https://doi.org/10.1016/j.micpath.2018.09.033 (2018).
11. Rashid, M. I., Rehman, S., Ali, A. & Andleeb, S. Fishing for vaccines against Vibrio cholerae using in silico pan-proteomic reverse vaccinology approach. *PeerJ* **7**, e6223, https://doi.org/10.7717/peerj.6223 (2019).
12. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Biol. Evol.* **33**(7), 1870–1874, https://doi.org/10.1093/molbev/msw054 (2016).

13. Hall, *Biological sequence alignment editor (BioEdit), version 7.2.5* https://www.mybiosoftware.com/bioedit-7-0-9-biological-sequence-alignment-editor.html (2013).
14. Tamura, & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**(3), 512–526, https://doi.org/10.1093/oxfordjournals.molbev.a040023 (1993).
15. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook* 571–607, https://doi.org/10.1385/1-59259-890-0:571 (Humana Press, 2005).
16. Cheng, J., Randall, A. Z., Sweredoski, M. J. & Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* **33**, 72–76, https://doi.org/10.1093/nar/gki396 (2005).
17. Garnier, J., Gibrat, J. F., & Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540–553, https://doi.org/10.1016/s0076-6879(96)66034-0 (1996).
18. Doytchinova, I. A. & Flower, D. R. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinf.* **8**(1), 4, https://doi.org/10.1186/1471- 2105-8-4 (2007).
19. Shen, H. B. & Chou, K. C. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *Biomol. Struct. Dyn.* **28**(2), 175–186, https://doi.org/10.1080/07391102.2010.10507351 (2010).
20. Krogh, A., Larsson, B., von Heijne, & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**(3), 567–580, https://doi.org/10.1006/jmbi.2000.4315 (2001).
21. Frank, K. & Sippl, M. J. High performance signal peptide prediction based on sequence alignment *Bioinformatics* **24**(19), 2172–2176, https://doi.org/10.1093/bioinformatics/btn422 (2008).
22. Saha, S. & Raghava, G. P. S. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34**, 202–209, https://doi.org/10.1093/nar/gkl343 (2006).
23. Bradley, P., Cowen, L., Menke, M., King, J. & Berger, B. BETAWRAP: successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. *Natl. Acad. Sci. U. S. A.* **98**(26), 14819–14824, https://doi.org/10.1073/pnas.251267298 (2001).
24. Gupta, R., Jung, E. & Brunak, S. *Prediction of N-glycosylation sites in human proteins* http://www.cbs.dtu.dk/services/NetNGlyc/ (2004).
25. Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**(10), 1478–1488, https://doi.org/10.1038/emboj.2013.79 (2013).
26. Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Mol. Biol.* **294**(5), 1351–1362, https://doi.org/10.1006/jmbi.1999.3310 (1999).
27. Ren, J. *et al.* CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng. Des. Sel.* **21**(11), 639–644, https://doi.org/10.1093/protein/gzn039 (2008).

28. Petersen, B., Petersen, T. N., Andersen, P., Nielsen, & Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **9**(1), 51, https://doi.org/10.1186/1472-6807-9-51 (2009).
29. Saha, S. & Raghava, G. P. S. BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In *Artificial Immune Systems, Lecture Notes in Computer Science* 3239, 197–204 (Springer, Berlin Heidelberg, 2004).
30. Vita, R. *et al.* The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* **47**(1), 339–343, https://doi.org/10.1093/nar/gky1006 (2019).
31. Yang, *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**(1), 7–8, https://doi.org/10.1038/nmeth.3213 (2015).
32. Kozakov, D. *et al.* The ClusPro web server for protein-protein docking. *Protoc.* **12**(2), 255–278, https://doi.org/10.1038/nprot.2016.169 (2017).
33. Pettersen, E. F. *et* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**(13), 1605–1612, https://doi.org/10.1002/jcc.20084 (2004).
34. Atanasova, M., Patronov, A., Dimitrov, I., Flower, D. R. & Doytchinova, I. EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. *Protein Eng. Des. Sel.* **26**(10), 631–634, https://doi.org/10.1093/protein/gzt018 (2013).
35. Goodswen, S. J., Kennedy, P. J. & Ellis, T. A guide to in silico vaccine discovery for eukaryotic pathogens. *Brief. Bioinform.* **14**(6), 753–774, https://doi.org/10.1093/bib/bbs066 (2013).
36. Droppa-Almeida, D., Franceschi, E. & Padilha, F. Immune-Informatic Analysis and Design of Peptide Vaccine From Multi-epitopes Against Corynebacterium pseudotuberculosis. *Bioinform. Biol. Insights* **12**, 1177932218755337, https://doi.org/10.1177/1177932218755337 (2018).
37. Shey, R. A. *et al.* In-silico design of a multi-epitope vaccine candidate against onchocerciasis and related filarial diseases. *Rep.* **9**(1), 4409, https://doi.org/10.1038/s41598-019-40833- x (2019).
38. Nakai, K. Protein sorting signals and prediction of subcellular localization. *Adv Protein* **54**, 277–344, https://doi.org/10.1016/s0065-3233(00)54009-1 (2000).
39. Hegde, R. S. & Bernstein, H. D. The surprising complexity of signal sequences. *Trends Biochem. Sci.* **31**(10), 563–571, https://doi.org/10.1016/j.tibs.2006.08.004 (2006).
40. Shaddel, M., Ebrahimi, M. & Tabandeh, M.R. Bioinformatics analysis of single and multi- hybrid epitopes of GRA-1, GRA-4, GRA-6 and GRA-7 proteins to improve DNA vaccine design against Toxoplasma gondii. *Parasit. Dis.* **42**(2), 269–276, https://doi.org/10.1007/s12639-018-0996-9 (2018).
41. Hansson, M., Nygren, P.A. & Stahl, S. Design and production of recombinant subunit *Biotechnol. Appl. Biochem.* **32**(2), 95–107, https://doi.org/10.1042/ba20000034 (2000).
42. Bisht, H. *et al.* Severe acute respiratory syndrome coronavirus spike protein expressed by attenuated vaccinia virus protectively immunizes mice. *Natl. Acad. Sci. U. S. A.* **101**(17), 6641–6646, https://doi.org/10.1073/pnas.0401939101 (2004).

43. Woo, P. C. *et al.* SARS coronavirus spike polypeptide DNA vaccine priming with recombinant spike polypeptide from Escherichia coli as booster induces high titer of neutralizing antibody against SARS coronavirus. *Vaccine* **23**(42), 4959–4968, https://doi.org/10.1016/j.vaccine.2005.05.023 (2005).
44. Al-Amri, S. S. *et al.* Immunogenicity of Candidate MERS CoV DNA Vaccines Based on the Spike Protein. *Rep.* **7**, 44875, https://doi.org/10.1038/srep44875 (2017).
45. Huang, L. R. *et al.* Evaluation of antibody responses against SARS coronaviral nucleocapsid or spike proteins by immunoblotting or ELISA. *Med. Virol.* **(3)**, 338–346, https://doi.org/10.1002/jmv.20096 (2004).
46. Liu, S. J. *et al.* Immunological characterizations of the nucleocapsid protein based SARS vaccine *Vaccine.* **24**(16), 3100–3108, https://doi.org/10.1016/j.vaccine.2006.01.058 (2006).
47. He, Y., Lu, H., Siddiqui, P., Zhou, Y. & Jiang, S. Receptor-binding domain of severe acute respiratory syndrome coronavirus spike protein contains multiple conformation-dependent epitopes that induce highly potent neutralizing antibodies. *Immunol.* **174**(8), 4908–4915, https://doi.org/10.4049/jimmunol.174.8.4908 (2005).
48. Shih, Y. P. *et al.* Identifying Epitopes Responsible for Neutralizing Antibody and DC-SIGN Binding on the Spike Glycoprotein of the Severe Acute Respiratory Syndrome *J. Virol.* **80**(21), 10315–10324, https://doi.org/10.1128/JVI.01138-06 (2006).
49. Berry, J. D. *et al.* Neutralizing epitopes of the SARS CoV S-protein cluster independent of repertoire, antigen structure or mAb technology. **2**(1), 53–66, https://doi.org/10.4161/mabs.2.1.10788 (2010).
50. Trevor, R. F. *et al.* Rapid development of a synthetic DNA vaccine for COVID-19. Preprint at https://researchsquare.com/article/rs-16261/v1 (2020).
51. Fast, E., Altman, R. B. & Chen, B. Potential T-cell and B-cell Epitopes of 2019-nCoV. Preprint at https://biorxiv.org/content/10.1101/2020.02.19.955484v2 (2020).
52. Bojin, F., Gavriliuc, O., Margineanu, M. & Paunescu, V. Design of an Epitope-Based Synthetic Long Peptide Vaccine to Counteract the Novel China Coronavirus (2019-nCoV). Preprint at https://preprints.org/manuscript/202002.0102/v1 (2020).

Tables

Due to technical limitations, Tables 1-11 are only available as a download in the supplemental files section

Additional Information

Supplementary information: Supplementary S1.

Figures

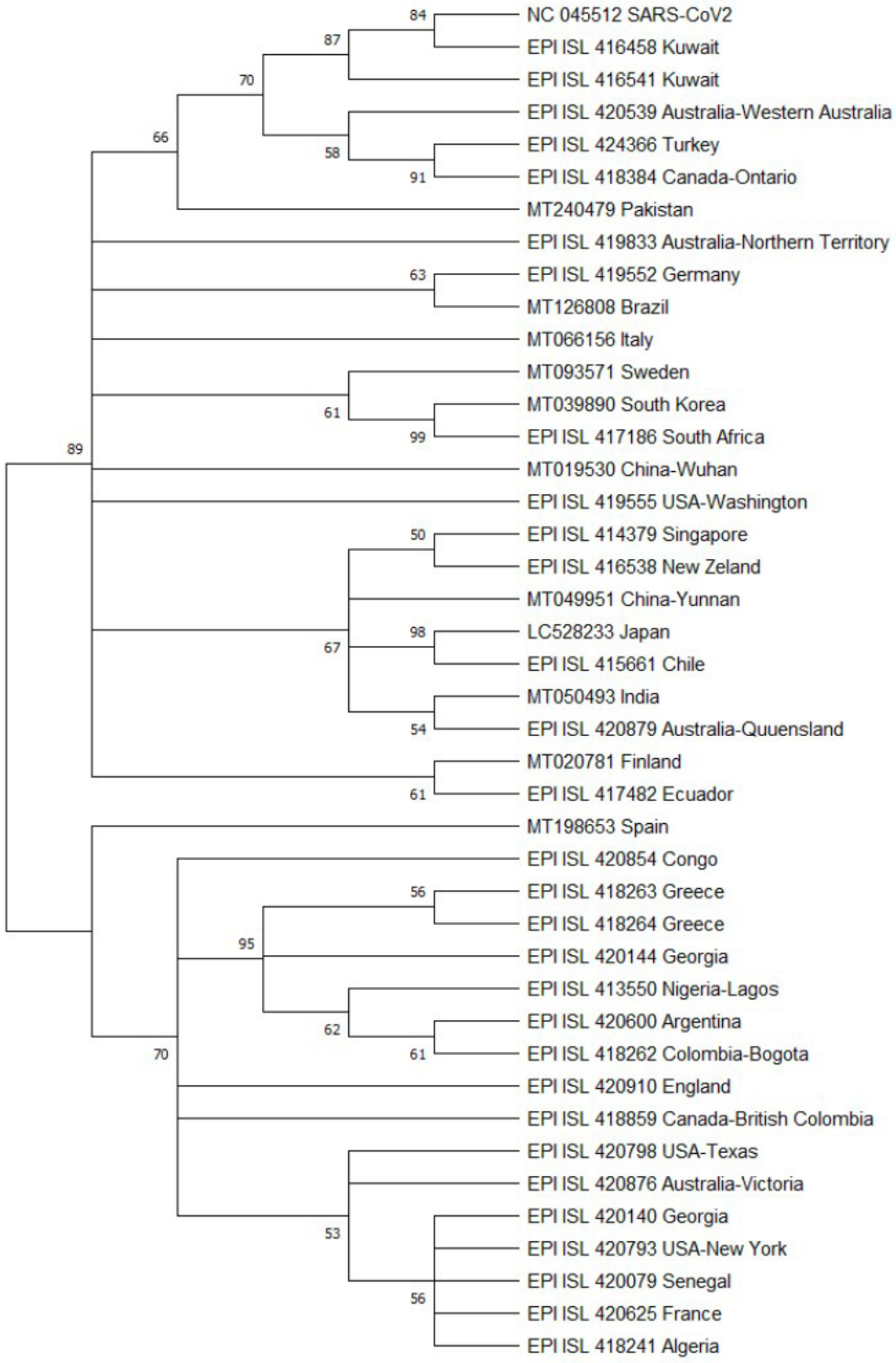

Figure 1

Phylogenetic tree results belonging to 42 SARS-CoV-2 isolates from six continents.

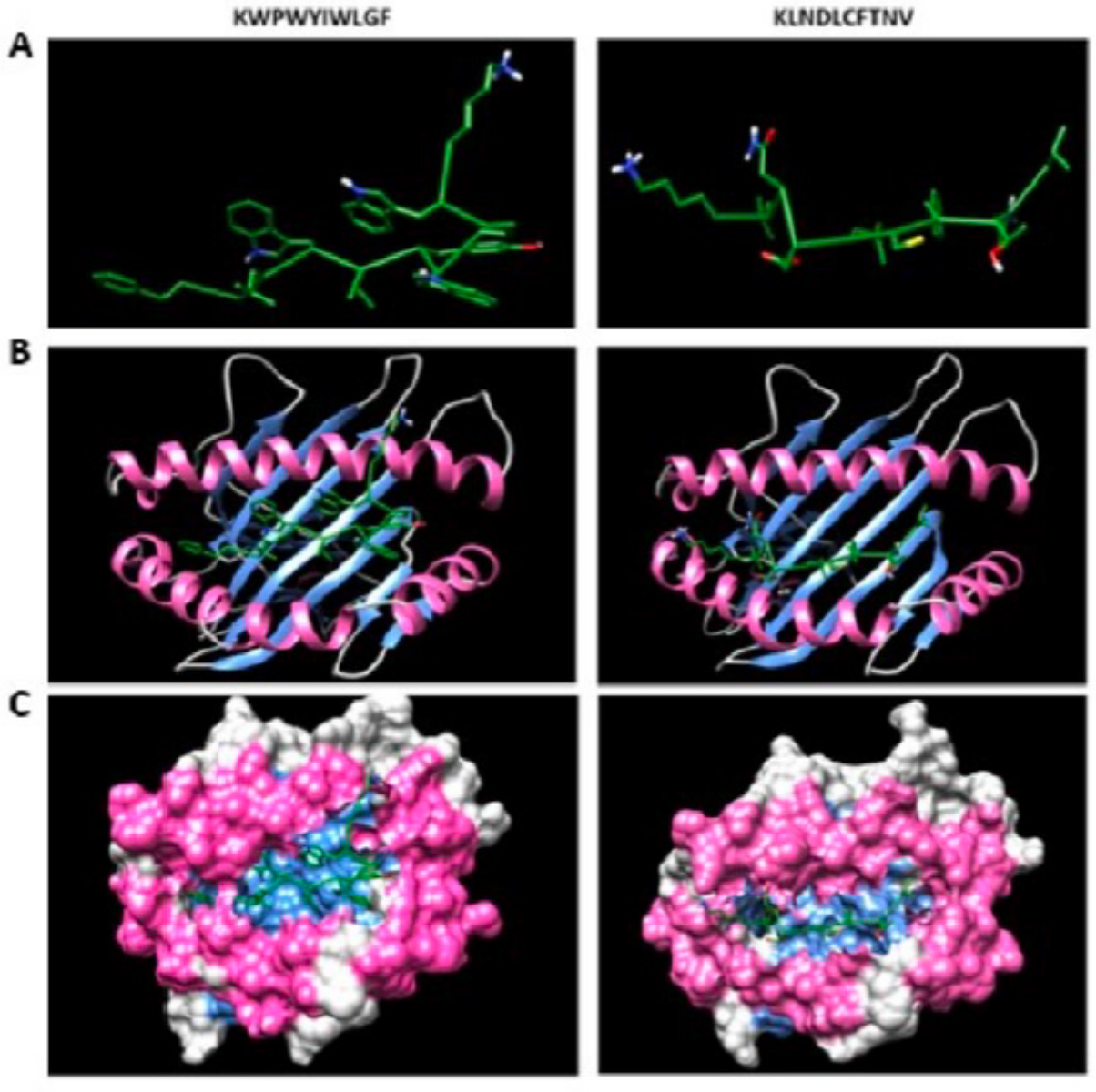

Figure 2

(A) Predicted KWPWYIWLGF and KLNDLCFTNV epitopes docking to MHC-I alleles. (B) Docking results of epitopes with α chain of MHC-I alleles using ClusPro. (C) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualised using Chimera 1.14).

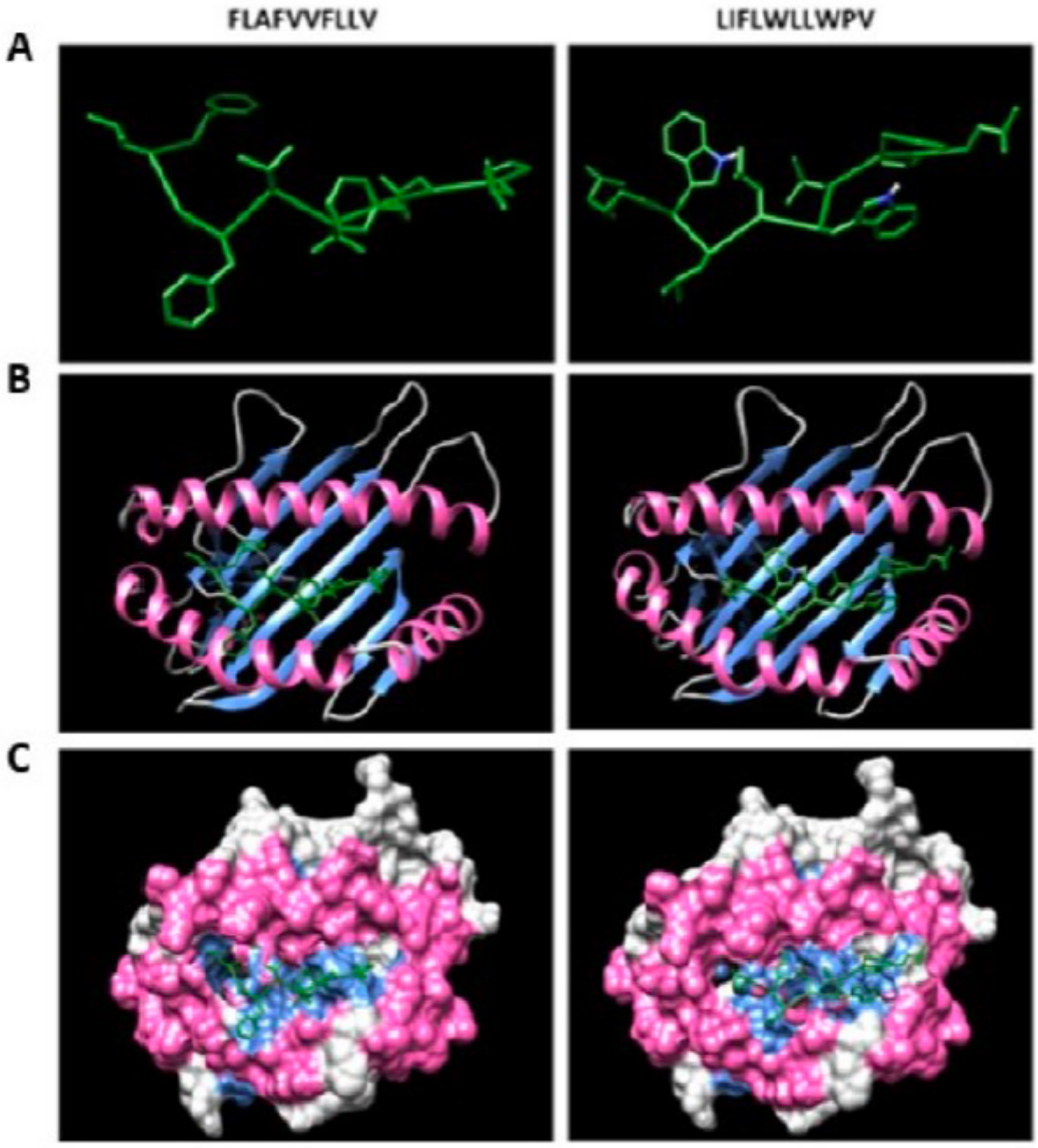

Figure 3

(A) Predicted FLAFVVFLLV and LIFLWLLWPV epitopes docking to MHC-I alleles. (B) Docking results of epitopes with α chain of MHC-I alleles using ClusPro. (C) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualised using Chimera 1.14).

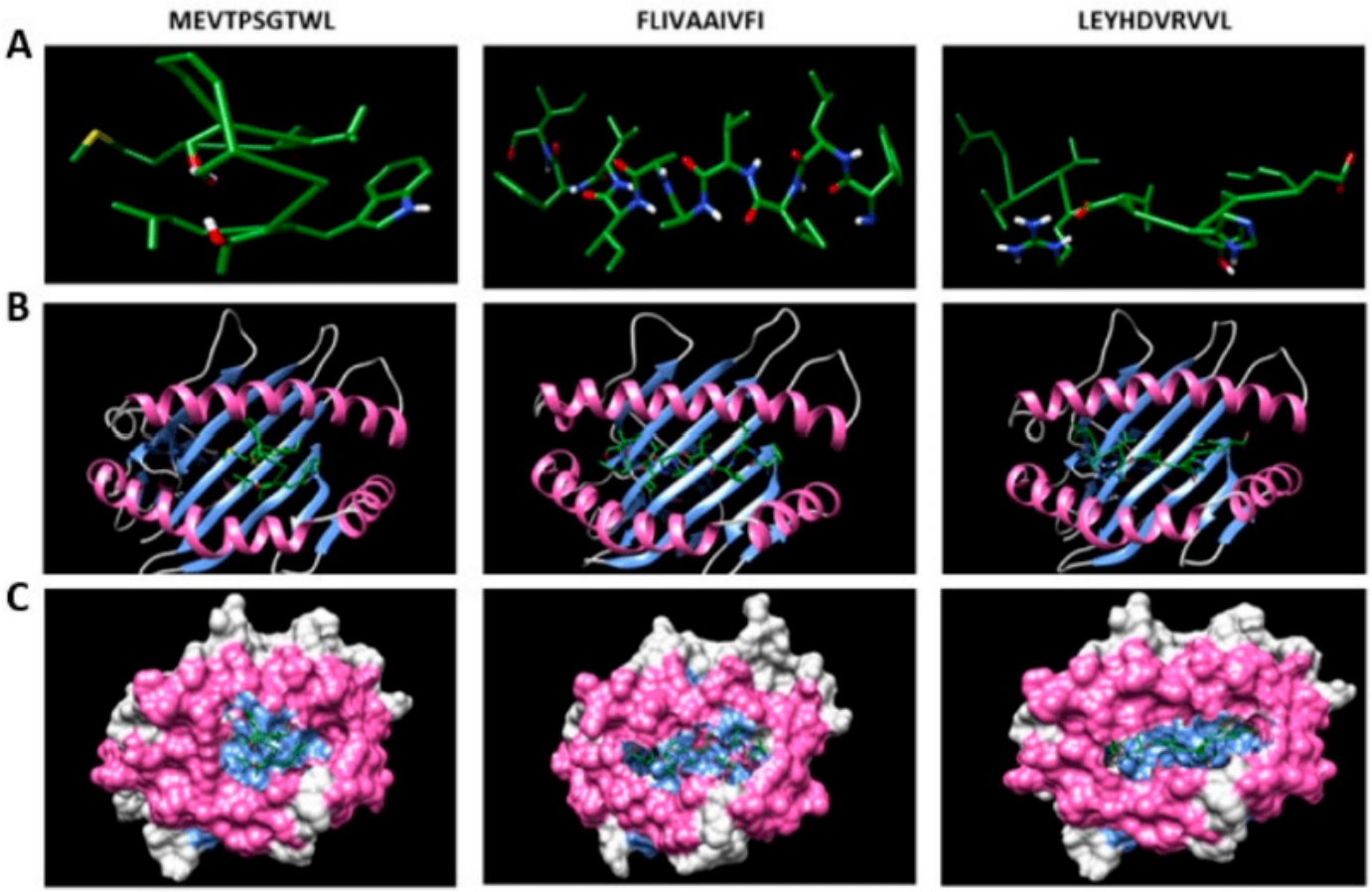

Figure 4

(A) Predicted MEVTPSGTWL, FLIVAAIVFI and LEYHDVRVVL epitopes docking to MHC-I alleles. (B) Docking results of epitopes with α chain of MHC-I alleles using ClusPro. (C) The snapshot representing the epitope docked in the pocket of molecular surface of the receptor (all the structures are visualised using Chimera 1.14).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf
- Tables.pdf