In silico discovery of antigenic proteins and epitopes of SARS-CoV-2 for the development of a vaccine or a diagnostic approach for COVID-19

Hüseyin Can1, Ahmet Efe Köseoğlu1, Sedef Erkunt Alak1, Mervenur Güvendi2, Mert Döşkaya2, Muhammet Karakavuk3, Adnan Yüksel Gürüz2 & Cemal Ün1

In the genome of SARS-CoV-2, the 5′-terminus encodes a polyprotein, which is further cleaved into 15 non-structural proteins whereas the 3′ terminus encodes four structural proteins and eight accessory proteins. Among these 27 proteins, the present study aimed to discover likely antigenic proteins and epitopes to be used for the development of a vaccine or serodiagnostic assay using an in silico approach. For this purpose, after the full genome analysis of SARS-CoV-2 Wuhan isolate and variant proteins that are detected frequently, surface proteins including spike, envelope, and membrane proteins as well as proteins with signal peptide were determined as probable vaccine candidates whereas the remaining were considered as possible antigens to be used during the development of serodiagnostic assays. According to results obtained, among 27 proteins, 26 of them were predicted as probable antigen. In 26 proteins, spike protein was selected as the best vaccine candidate because of having a signal peptide, negative GRAVY value, one transmembrane helix, moderate aliphatic index, a big molecular weight, a long-estimated half-life, beta wrap motifs as well as having stable, soluble and non-allergic features. In addition, orf7a, orf8, and nsp-10 proteins with signal peptide were considered as potential vaccine candidates. Nucleocapsid protein and a highly antigenic GGDGKMKD epitope were identified as ideal antigens to be used in the development of serodiagnostic assays. Moreover, considering MHC-I alleles, highly antigenic KLNDLCFTNV and ITLCFTLKRK epitopes can be used to develop an epitope-based peptide vaccine.

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 humans1.

In the 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-CoV2,3. Also, it has been reported that RBD of SARS-CoV-2 binds to the human ACE2 receptor in a higher affinity than that of SARS-CoV as well as the presence of an insertion in the polybasic cleavage region in SARS-CoV-2 may increase the infectivity of the

1Department of Biology Molecular Biology Section, Faculty of Science, Ege University, Bornova, İzmir, Turkey. 2Department of Parasitology, Faculty of Medicine, Ege University, Bornova, İzmir, Turkey. 3Ödemiş Vocational School, Ege University, İzmir, Turkey. *email: cemaluen@gmail.com
Antigenicity. All structural proteins were predicted as probable antigen. Antigenicity value varied from 0.4661 to 0.6025. E protein had the highest antigenicity value whereas S protein had the lowest antigenicity value. Antigenicity values did not dramatically change among the original Wuhan and variant proteins. 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. According to Wuhan orf8 antigenicity value, variant V62L had a higher antigenicity value whereas variant L84S had a lower antigenicity value (Table 3).

Solubility. According to solubility prediction, S, E, and N proteins were soluble. Among non-structural proteins, orf3a as well as nsp-2, 4, 5, 7, 10, 12, 13, 14, 15, and 16 proteins encoded by orf1ab were predicted as insoluble whereas the remaining orf9 to orf10 and nsp-1, 6, 8, 9 were predicted as soluble. The solubility prediction of another protein, nsp-3 encoded by orf1ab, could not be retrieved due to large fragment size (Table 3).

Results
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 1). Among non-structural proteins, except orf1ab, the number 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 1). Each non-structural protein that is encoded by orf1ab was also analysed, and the number of amino acids was the non-structural proteins that were encoded by orf1ab (Table 1). Among non-structural proteins, except orf1ab, the number 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 1).

Secondary structure. According to results obtained from structural proteins, the alpha helix was between ~22 and 47%, that of the extended strand was between ~10 and 22%, and that of the random coil was between ~40 and 60%. For non-structural proteins, the alpha helix varied between 0 and 69%, that of the extended strand varied between ~3 and 47%, and that of the random coil varied between ~28 and 58% (Table 2).

Antigenicity. All structural proteins were predicted as probable antigen. Antigenicity value varied from 0.4661 to 0.6025. E protein had the highest antigenicity value whereas S protein had the lowest antigenicity value. Antigenicity values did not dramatically change among the original Wuhan and variant proteins. 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. According to Wuhan orf8 antigenicity value, variant V62L had a higher antigenicity value whereas variant L84S had a lower antigenicity value (Table 3).

Solubility. According to solubility prediction, S, E, and N proteins were soluble. Among non-structural proteins, orf3a as well as nsp-2, 4, 5, 7, 10, 12, 13, 14, 15, and 16 proteins encoded by orf1ab were predicted as insoluble whereas the remaining orf9 to orf10 and nsp-1, 6, 8, 9 were predicted as soluble. The solubility prediction of another protein, nsp-3 encoded by orf1ab, could not be retrieved due to large fragment size (Table 3).
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 3). When subcellular localisation predictions were examined, S, M, and N proteins were predicted to be in the host endoplasmic reticulum. E as well as M proteins were also predicted to be in the host cell membrane. Non-structural proteins were predicted to locate in cell membrane, endoplasmic reticulum, cytoplasm, nucleus (Table 3).

Signal peptide. According to the 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 4). Any variations did not change the results of signal peptide.

Allergenicity. None of the proteins including variant proteins showed allergenic properties for MEME/MAST motif and IgE epitopes (Table 5).

BetaWrap motifs. Among all proteins analysed, only S protein and its variant (D614G) were predicted to contain BetaWrap motifs (Table 5).

Similarity with host proteome. No significant similarity was predicted between analysed viral proteins including variant proteins and host proteins (Table 5).

B cell epitopes. A lot of linear B cell epitopes were predicted for S, variant S (D614G), N and variants of N (P13L, S194L, S197L, and R203K/G204R), E, and variant E (L37H), orf8 and variants of orf8 (S24L, V62L and L84S) 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 6. Obtained predictions showed that nearly all epitopes had more antigenicity value than those of their own proteins. Interestingly, an epitope (VDEAGSKS) corresponding to variant orf8 (S24L) was predicted non-antigen because of adding valine amino acid to lead of epitope as differ-

---

Table 1. Physico-chemical parameter results predicted by ExPASyProtParam. *GRAVY means the Grand average of hydropathicity.
A specific epitope (AEGSRGGASRGGGSQASSRSSSRSRNS) with a high antigenicity value that was not predicted for N or other variant N 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 6.

**MHC-I and MHC-II epitopes.** A lot of MHC-I epitopes were predicted as probable antigen (Table 7). 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, KWPWY1WLGF, FLAFVFVLLV, FARTRSMWSF, and RNRFLYIIKL, AQFPAPSAAS and LGIITTV AAF epitopes belonging to S (including variant D614G), E (including variant L37H), M (including variant T175M), N (including variants P13L, S194L, S197L, and R203K/G204R) and orf8 (including S24L, V62L, and 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. Also, T175M and S194L variations in M and N proteins caused the prediction of additional epitopes that are specific to own themselves.

Similarly, a lot of MHC-II epitopes were predicted as probable antigen (Table 8). Also, nearly all epitopes had higher antigenicity values than those of their own proteins. Among structural proteins, PTPNSIVTTEILPV 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 (IVFITCFLKIKTE) 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 two epitopes (SKWYIYGKSMFL and KWYIRVAGKSAI) that had an IC50 value lower than 10, indicating a strong binding among epitope and MHC-II alleles, was detected in orf8 and its variants. In addition, variant M protein (T175M) and orf8 variants (S24L and V62L) had specific epitopes with high antigenicity values as different from original Wuhan M and orf8 proteins.

**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 (including T175M), E (including L37H), orf7a, and nsp10 proteins were predicted to have N-glycosylated and phosphorylated sites

![Table 2. Secondary structures results predicted by GOR IV.](https://www.nature.com/scientificreports/)
while orf7a was predicted to have an acetylation site. Orf8 and its variants were predicted to have N-glycosylated and phosphorylated sites whereas two additional phosphorylation sites, one of which locate in the exposed surface and the other one is buried, were predicted in only variant L84S. In addition, N protein and its four variants were predicted to have N-/O-glycosylated, phosphorylated and acetylated sites. When N protein and its variant were compared, the number of O-glycosylation, acetylation, or phosphorylation sites showed minor alterations. In addition, post-translational modifications within significant epitopes were shown in Table 9.

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, KWPWYIWLF, KLNDLCFTNV, FLAFVFLLVF, LIFLWLLWPV, MEVTPSGTWL, FLIVAAIVFI, and LEYHDVRVVL epitopes belonging to S (including variant D614G), E (including variant L37H), M (including variant T175M, N (including variants P13L, S194L, S197L, and R203K/G204R), orf7a and orf8 (including variants S24L, V62L, and L84S), respectively, were docked with receptors of selected MHC-I alleles (Figs. 1, 2, and 3).

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

Solvent-exposure positions in epitopes. The amino acids exposed to the solvent were detected in all significant epitopes. Among these epitopes, the whole of only two epitopes (FLAFVFLLVF and KWPWYIWLF) were in the solvent accessible region on protein structure (Table 9).

### Table 3. Solubility, transmembrane helices, localisation and antigenicity results predicted by SolPro, TMHMM, Virus-mPLoc and Vaxijen, respectively. *Could not be retrieved.

| Proteins | SolPro | TMHMM | Virus-mPLoc | Vaxijen v2.0 value |
|----------|--------|--------|-------------|-------------------|
| Structural proteins | | | | |
| S protein | Soluble | 1 Host endoplasmic reticulum | 0.4661 (probable antigen) |
| E protein | Soluble | 1 Host endoplasmic reticulum | 0.6025 (probable antigen) |
| M protein | Insoluble | 3 Host cell membrane, endoplasmic reticulum | 0.5102 (probable antigen) |
| N protein | Soluble | 0 Host cell membrane | 0.5059 (probable antigen) |
| Non-structural proteins | | | | |
| orf3a | Insoluble | 3 Host cell membrane | 0.4945 (probable antigen) |
| orf6 | Soluble | 0 Host endoplasmic reticulum | 0.6131 (probable antigen) |
| orf7a | Soluble | 1 Host endoplasmic reticulum | 0.6441 (probable antigen) |
| orf7b | Soluble | 1 * | 0.8462 (probable antigen) |
| orf8 | Soluble | 0 Host cell membrane, endoplasmic reticulum, cytoplasm | 0.6502 (probable antigen) |
| orf10 | Soluble | 0 * | 0.7185 (probable antigen) |
| Pp1ab | | | | |
| nsp-1 | Soluble | 0 Host cytoplasm | 0.4064 (probable antigen) |
| nsp-2 | Insoluble | 0 Host cytoplasm | 0.4034 (Probable antigen) |
| nsp-3 | * | 4 Host cytoplasm | 0.5142 (probable antigen) |
| nsp-4 | Insoluble | 4 Host cell membrane, endoplasmic reticulum | 0.4691 (probable antigen) |
| nsp-5 | Insoluble | 0 Host cell membrane | 0.4159 (probable antigen) |
| nsp-6 | Soluble | 8 Host cell membrane | 0.5813 (probable antigen) |
| nsp-7 | Insoluble | 0 Host cytoplasm | 0.4167 (probable antigen) |
| nsp-8 | Soluble | 0 Host cytoplasm | 0.4008 (probable antigen) |
| nsp-9 | Soluble | 0 Host cell membrane, cytoplasm | 0.6476 (Probable antigen) |
| nsp-10 | Insoluble | 0 Host cell membrane, endoplasmic reticulum | 0.4039 (probable antigen) |
| nsp-12 | Insoluble | 0 Host cytoplasm | 0.4064 (probable antigen) |
| nsp-13 | Insoluble | 0 Host nucleus | 0.4480 (probable antigen) |
| nsp-14 | Insoluble | 0 Host cytoplasm, nucleus | 0.4138 (probable antigen) |
| nsp-15 | Insoluble | 0 Host cytoplasm | 0.5554 (probable antigen) |
| nsp-16 | Insoluble | 0 Host cytoplasm | 0.3809 (probable non-antigen) |
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. 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 the development of a vaccine or serodiagnostic assay as described by Dangi et al. and Goodswen et al. 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. The 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. 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, fewer than two transmembrane helices facilitating cloning, expression, and purification, 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, orf8 and nsp-10 proteins were predicted to have a signal peptide.

### Table 4

The signal peptide results predicted by Signal-BLAST. Bold indicates proteins that have a signal peptide.

| Protein name | Sensitivity | Specificity | Balanced prediction | Cleavage site |
|--------------|-------------|-------------|---------------------|--------------|
| S protein    | Yes         | Yes         | Yes                 | Yes          |
| Variant D614G| Yes         | Yes         | Yes                 | Yes          |
| E protein    | No          | No          | No                  | Yes          |
| Variant L37H | No          | No          | No                  | No           |
| M protein    | No          | No          | No                  | No           |
| Variant M T175M | No      | No          | No                  | No           |
| N protein    | No          | No          | No                  | Yes          |
| Variant N P13L | No     | No          | No                  | Yes          |
| Variant N S194L | No    | No          | No                  | Yes          |
| Variant NS197L | No      | No          | No                  | Yes          |
| Variant N R203K/G204R | No | No          | No                  | Yes          |
| orf3a        | No          | No          | No                  | No           |
| orf6         | No          | No          | No                  | Yes          |
| orf7a        | Yes         | Yes         | Yes                 | Yes          |
| ORF7b        | No          | No          | No                  | Yes          |
| orf8         | Yes         | Yes         | Yes                 | Yes          |
| Variant orf8 S24L | Yes | Yes         | Yes                 | Yes          |
| Variant orf8 V62L | Yes | Yes         | Yes                 | Yes          |
| Variant orf8 L84S | Yes | Yes         | Yes                 | Yes          |
| orf10        | No          | No          | No                  | Yes          |
| nsp-1        | No          | No          | No                  | Yes          |
| nsp-2        | No          | No          | No                  | Yes          |
| nsp-3        | No          | No          | No                  | No           |
| nsp-4        | No          | No          | No                  | Yes          |
| nsp-5        | Yes         | No          | Yes                 | Yes          |
| nsp-6        | Yes         | No          | Yes                 | Yes          |
| nsp-7        | No          | No          | No                  | Yes          |
| nsp-8        | No          | No          | No                  | No           |
| nsp-9        | Yes         | No          | Yes                 | Yes          |
| nsp-10       | Yes         | Yes         | Yes                 | Yes          |
| nsp-12       | No          | No          | No                  | Yes          |
| nsp-13       | No          | No          | No                  | No           |
| nsp-14       | No          | No          | No                  | No           |
| nsp-15       | No          | No          | No                  | No           |
| nsp-16       | No          | No          | No                  | No           |

**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. 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 the development of a vaccine or serodiagnostic assay as described by Dangi et al. and Goodswen et al.

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. The 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. 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, fewer than two transmembrane helices facilitating cloning, expression, and purification, 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, orf8 and nsp-10 proteins were predicted to have a signal peptide. This feature is an important
parameter which indicates that the protein can be destined towards the secretory pathway. Moreover, the signal peptide promotes protein secretion, and thus, the signal peptide is used to improve the protein secretion level in recombinant techniques. For example, a study showed that vaccination with an unconventionally secreted viral nonstructural protein (NS1) protected mouse from murine norovirus. Also, it has been reported that proteins with the signal peptide should be taken into consideration as vaccine candidates both they have been targeted to the secretory pathway and have high antigenicity and specificity. Accordingly, these probable secreted and antigenic three proteins (orf7a, orf8, and nsp-10) can also be considered as potential vaccine candidates.

### Table 5. Allergenicity, BetaWrap motifs and host proteome similarity results predicted by AlgPred, BetaWrap and BlastP, respectively.

| Proteins     | AlgPred IgE epitopes | AlgPred MEME/MAST motif | AlgPred BetaWrap motifs | AlgPred BlastP |
|--------------|----------------------|-------------------------|-------------------------|----------------|
| Structural proteins |                      |                         |                         |                |
| S protein    |                      |                         |                         |                |
| Variant S (D614G) |                | Non allergen             | P value: 0.014            | No             |
| E protein    |                      |                         |                         |                |
| Variant E (L37H) |                | Non allergen             | No                      | No             |
| M protein    |                      |                         |                         |                |
| Variant M (T175M) |               | Non allergen             | No                      | No             |
| N protein    |                      |                         |                         |                |
| Variant N (P13L) |               | Non allergen             | No                      | No             |
| Variant N (S194L) |              | Non allergen             | No                      | No             |
| Variant N (S197L) |              | Non allergen             | No                      | No             |
| N protein    |                      |                         |                         |                |
| Variant N (R203K/204R) |          | Non allergen             | No                      | No             |
| Non-structural proteins |              |                          |                         |                |
| orf7a        |                      |                         |                         |                |
| orf8         |                      |                         |                         |                |
| Variant orf8 (S24L) |           | Non allergen             | No                      | No             |
| Variant orf8 (V62L) |           | Non allergen             | No                      | No             |
| Variant orf8 (L84S) |          | Non allergen             | No                      | No             |
| Pp1ab        |                      |                         |                         |                |
| nsp-10       |                      |                         |                         |                |

### Table 6. B cell epitopes predicted by both Bcepred and IEDB and antigenicity value predicted by Vaxijen v2.0.

| Proteins     | Antigenicity value | B cell epitopes | Antigenicity value for epitopes |
|--------------|--------------------|----------------|--------------------------------|
| Vaccine candidate epitopes |                      |                |                                |
| S protein    | 0.4661 (probable antigen) | VYYHKNNKSW YAWNRKRISN GDEVQR2 NLDKYV | 0.4497 (probable antigen) 0.5855 (probable antigen) 0.6701 (probable antigen) 0.7443 (probable antigen) |
| Variant (D614G) | 0.4638 (probable antigen) |                |                                |
| E protein    | 0.6025 (probable antigen) | SrvKlnnssrvp  | 0.5572 (probable antigen) |
| Variant E (L37H) | 0.6298 (probable antigen) |                |                                |
| M protein    | 0.5102 (probable antigen) | ND              | -                              |
| Variant M (T175M) | 0.4990 (probable antigen) |                |                                |
| orf8         | 0.6502 (probable antigen) |                |                                |
| Variant orf8 (V62L) | 0.6734 (probable antigen) |                |                                |
| Variant orf8 (L84S) | 0.6063 (probable antigen) |                |                                |
| Variant orf8 (S24L) | 0.6581 (probable antigen) |                |                                |
| nsp-10       | 0.4039 (probable antigen) | Thtgtgq      | 1.0789 (probable antigen) |
| Diagnostic epitopes |                      |                |                                |
| N protein    | 0.5059 (probable antigen) | NgpQnQrnap NtNssppdqi GdGKmrd | 0.5058 (probable antigen) 0.4913 (probable antigen) |
| Variant N (P13L) | 0.5119 (probable antigen) |                |                                |
| Variant N (S194L) | 0.4989 (probable antigen) |                |                                |
| Variant N (S197L) | 0.5139 (probable antigen) |                |                                |
| Variant N (R203K/G204R) | 0.5068 (probable antigen) |                |                                |
| Variant N (P13L) | 0.5119 (probable antigen) | Aeg-SrgQssqRsSrsNsS | 0.9061 (probable antigen) |

ND not detected. aVaccine candidate epitopes can also be used to develop serodiagnostic assays.
recombinant technology, eukaryotic expression systems such as yeast, insect or mammalian should be preferred instead of bacterial systems23.

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 continents24,25. 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 countries7,26. According to the results obtained from these studies, S and M proteins were reported to induce a strong immune response. For the vaccine development against SARS-CoV-2, it has been stated that S protein is a promising candidate because it plays a role in viral attachment, fusion, and entry27–29. In addition, a report showing that antibodies against S protein of SARS-CoV inhibit the SARS-CoV-2

| Promin | Allele | Start | End | Peptide | IC50 | Percentile rank | Antigenicity |
|--------|--------|-------|-----|---------|------|----------------|-------------|
| S protein | Variant S (D614G) | HLA-A*02:01 | 386 | 395 | KLNDLCFTNV | 15.27 | 0.14 | 2.6927 (probable antigen) |
| | | HLA-B*07:02 | 680 | 689 | SPRRARSV AS | 18.4 | 0.23 | 0.8406 (probable antigen) |
| | | HLA-B*15:01 | 49 | 58 | HSTQDLFLPF | 27.79 | 0.17 | 0.5162 (probable antigen) |
| | | HLA-A*02:01 | 1211 | 1220 | KWPWYIWLGF | 18.4 | 0.27 | 0.5390 (probable antigen) |
| | | HLA-B*15:01 | 488 | 497 | CYFPLQSYGF | 27.59 | 0.04 | 0.7775 (probable antigen) |
| | | HLA-A*02:01 | 268 | 277 | GLAQFRYLLL | 36.12 | 0.4 | 0.7535 (probable antigen) |
| | | HLA-B*15:01 | 406 | 415 | QTSCPFVYPK | 27.8 | 0.04 | 0.5417 (probable antigen) |
| | Variant E (L37H) | HLA-A*02:01 | 20 | 29 | FLAFVVFLLV | 9.95 | 0.1 | 0.5651 (probable antigen) |
| | | HLA-B*27:05 | 71 | 80 | YKDFVTRVTK | 39.44 | 0.18 | 0.7087 (probable antigen) |
| | | HLA-A*03:01 | 724 | 733 | TEILPVSMTK | 36.12 | 0.4 | 0.7535 (probable antigen) |
| | | HLA-B*15:01 | 7 | 16 | LGIITTV AAF | 6.51 | 0.02 | 0.7595 (probable antigen) |
| | | HLA-A*02:01 | 101 | 110 | FLIV AAIVFI | 13.43 | 0.13 | 0.6283 (probable antigen) |
| | | HLA-B*40:01 | 109 | 118 | LEYHDVRVVL  | 41.00 | 0.12 | 0.9860 (probable antigen) |
| | | HLA-A*03:01 | 408 | 417 | RQIAPGQTGK | 38.0 | 0.15 | 1.3140 (probable antigen) |
| | | HLA-B*15:01 | 56 | 65 | LTCFSTQFAF | 29.06 | 0.18 | 1.1543 (probable antigen) |
| | | HLA-A*03:01 | 23 | 32 | CVCLTTYLKL | 37.99 | 0.22 | 0.6175 (probable antigen) |
| | | HLA-B*15:01 | 57 | 66 | LLWPVTLACF | 11.12 | 0.05 | 0.9864 (probable antigen) |

Table 7. Epitopes specific to selected MHC-I alleles. *Vaccine candidate epitopes can also be used to develop serodiagnostic assays.
Scientific Reports

Table 8. Epitopes specific to selected MHC-II alleles. aVaccine candidate epitopes can also be used to develop serodiagnostic assays.

| Proteins | Allele | Start | End | Peptide | IC50 | Percentile rank | Antigenicity |
|----------|--------|-------|-----|---------|------|-----------------|--------------|
| S protein | HLA-DRB3*02:02 | 115 129 | QSLLIVNNATNVVVIK | 13.08 | 0.02 | 0.4343 (probable antigen) |
| D614G    |        |       |     |         |      |                 |              |
| Variant N (S197L) | HLA-DRB5*01:01 | 894 908 | LQIFEMQMYRGNFGP | 20.96 | 0.59 | 0.7205 (probable antigen) |
| Variant N (P13L) | HLA-DRB1*07:01 | 715 729 | PTNPNTSTTVTIEKLP | 25.32 | 0.63 | 1.1349 (probable antigen) |
| Variant N (R203K/G204R) | HLA-DRB3*02:02 | 1091 1105 | REGFVSNTHWVFTN | 25.75 | 0.2 | 0.4441 (probable antigen) |
| M protein |        |       |     |         |      |                 |              |
| Variant M (T175M) | HLA-DRB5*01:01 | 175 189 | TLSYLYKGLASQVRAG | 16.67 | 0.30 | 0.4376 (probable antigen) |
| orf7a |        |       |     |         |      |                 |              |
| Variant M (T175M) | HLA-DRB5*01:01 | 100 114 | FRFLTRKMSWSFNP | 28.34 | 1.20 | 0.8873 (probable antigen) |
| orf8 |        |       |     |         |      |                 |              |
| Variant orf8 (S24L) | HLA-DRB5*01:01 | 34 48 | LQFAYANNRFLFNY | 26.30 | 1.10 | 0.7387 (Probable antigen) |
| Diagnostic epitopes |        |       |     |         |      |                 |              |
| N protein |        |       |     |         |      |                 |              |
| Variant N (P13L) | HLA-DRB1*07:01 | 303 317 | QAQQAFPASAFFGGM | 32.18 | 1.10 | 0.4032 (probable antigen) |
| Variant N (S197L) | HLA-DRB1*07:01 | 264 278 | ATKAYNVTQAFGRRG | 29.21 | 1.40 | 0.7146 (probable antigen) |
| N protein |        |       |     |         |      |                 |              |
| Variant N (R203K/G204R) | HLA-DRB1*01:01 | 305 319 | AQFAPASAFFGMSR | 41.34 | 2.07 | 0.5266 (probable antigen) |

S protein-mediated entry into cells encourages the use of this molecular target for vaccination. Currently, a lot of companies or research groups target the S protein to develop a vaccine against SARS-CoV-2 using various recombinant vaccine technologies. For example, Inovio using S protein with a DNA vaccine technology is in Phase I. Another company, Moderna, is in Phase I/II with an RNA based vaccine targeting S protein. Consequently, these findings of recent studies and our in silico study support that only S proteins can be a strong vaccine candidate protein in the development of a 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 studies for the previous coronavirus (SARS-CoV) supporting our predictions. For example, a previous study reported a strong antibody response against recombinant N protein in 10 of 12 SARS patients. In a different study, a B cell epitope region between 156 and 175 positions of N protein reacted strongly with sera from SARS patients.

Among structural proteins, subcellular localisations of S and E proteins were predicted as endoplasmic reticulum in silico methods in the current study, and this result was found to be compatible with the results of SARS-CoV performed with an in vitro immunofluorescent analysis showing the localisation of S protein in several compartments of host secretory pathway from the endoplasmic reticulum to cell membrane as well as E protein in endoplasmic reticulum. However, subcellular localisation of M protein was predicted as the host cell membrane and endoplasmic reticulum in the current in silico analysis while it was shown in the Golgi apparatus in the same in vitro analysis. In fact, it was also thought to be compatible with sRNA results because endoplasmic reticulum, Golgi, and cell membrane are parts of the same host secretory pathway and all surface proteins may be detected in each part of the pathway.
Table 9. Additional analysis performed for significant epitopes. Only underlined amino acid shows the moderate solvent exposure intensity. Both underlined and bold amino acids show the high solvent exposure intensity. Analysed common lineages are A, A.1, A.2, A.3, A.5, B, B.1, B.1.1, B.2, B.3, B.4.

| Epitopes         | N glycosylation region | Acetylation region | Phosphorylation region | Solvent-exposure amino acids | Lineages          |
|------------------|------------------------|--------------------|------------------------|-----------------------------|-------------------|
| THTGTGQ         | -                      | -                  | -                      | THTGTGQ                     | All lineages      |
| GGDGKMKD        | -                      | -                  | -                      | GGDGKMKD                    | All lineages      |
| KLNDLCFTNV      | KLNDLCFTNV             | -                  | -                      | KLNDLCFTNV                  | All lineages      |
| ITLCTFLKKKR     | ITLCTFLKKKR            | -                  | -                      | ITLCTFLKKKR                 | All lineages      |
| PTNFTISVTTEILPV | PTNFTISVTTEILPV        | -                  | -                      | PTNFTISVTTEILPV             | All lineages      |
| IVFITHCLFTLKRKTE| IVFITHCLFTLKRKTE       | -                  | -                      | IVFITHCLFTLKRKTE            | All lineages      |
| KWPWYIWLG       | -                      | KWPWYIWLG          | -                      | KWPWYIWLG                   | All lineages      |
| FLAFVVELV       | -                      | -                  | -                      | FLAFVVELV                   | All lineages      |
| LGHTTVAAP       | -                      | -                  | -                      | LGHTTVAAP                   | All lineages      |
| FARTSMWFSF      | -                      | -                  | -                      | FARTSMWFSF                  | All lineages      |
| RNRFYIIKL       | RNRFYIIKL              | -                  | -                      | RNRFYIIKL                   | All lineages      |
| SKWYIRVGARSAPI  | SKWYIRVGARSAPI         | -                  | -                      | SKWYIRVGARSAPI              | All lineages      |
| KKWYIRVGARSA    | KKWYIRVGARSA           | -                  | -                      | KKWYIRVGARSA                | All lineages      |

Figure 1. (A) Predicted KWPWYIWLG and KLNDLCFTNV epitopes docking to MHC-I alleles. (B) Docking results of epitopes with a 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).
In this study, the immunological effects of prevalent variant proteins belonging to E, M, N, S, and orf8 proteins were also analysed. Accordingly, the comparison of reference S protein and its variant (D614G) showed no difference in antigenicity values, epitope regions, and antigenicity values of epitopes. However, detecting D614G variation as prevalent has been associated with selection advantage and random founder effect\(^3\)\(^4\),\(^5\),\(^6\). In addition to these, a study reported that the D614G variant was more stable and enhanced its infectious nature\(^7\) whereas another study reported that there was not enough evidence to express that the variant is more infectious\(^8\). For N protein, among five variations (P13L, S194L, S197L, R203K/G204R), P13L and S197L variations were predicted to increase the antigenicity value of N protein and thus, utilise of P13L and S197L variants was thought to be a better antigen for studies conducted in countries harboring SARS-CoV-2 isolates with P13L or S197L variant. A similar result was also detected in E protein and a higher antigenicity value was predicted in variant L37H. Variant orf8 (L84S) had a lower antigenicity value whereas a higher antigenicity value was predicted in variant orf8 (V62L) compared to orf8 of Wuhan isolate. Also, variant M protein (T175M) had a lower antigenicity value. As depending on these results, since a higher antigenicity value is associated with a stronger immune response in the host, selection of the proteins with high antigenicity values in vaccinological or serodiagnostic studies would be useful.

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. 1) 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 a neutralising antibody response or can be used in the 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 neutralising epitopes has been reported to bind with S protein of SARS-CoV\(^{38–40}\). For example, in a study conducted in mice, a major neutralisation determinant was reported in receptor-binding domain (RBD) of S protein in SARS-CoV\(^{38}\). Another study reported that the epitope NYNWKR in S protein had a neutralising effect against SARS-CoV\(^{39}\). 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 a cellular immune response. Among these epitopes, only one of them (VGGNYNYLYRLFRKS; between 445 and 459 positions) was inside RBD, five of them (YNYKLPDDFTGCVIA; DDFTGCVIAWSNNL; VVLSFELLHAPATVC; LLHAPATCGPKKST; KNKCVNFNGLTG) were located nearby RBD whereas the remaining three (SFQPSAPHVVFLH; PHGVVFLHVYYVPQ; FTFA-PAICHDGKAHF) were inside S2 segment of S protein\(^{41}\). Interestingly, in our study, an epitope (CYFPLQSYGF; between 488 and 497 positions) with a relatively lower antigenicity value was predicted in RBD of S protein and four epitopes (NLDSKV, KLNDLCFTNV, RQIAPGQTGK, GDEVRO) were also predicted in a very close region. Docking results supported that the epitope KLNDLCFTNV was targeted by HLA-A*02:01 allele (Fig. 1). These findings indicate that the above-mentioned epitopes may have a promising neutralising effect against SARS-CoV-2. In a previous in silico study, five different epitopes (SYGFQPTNGVGYQPY; SQSHAYTMSLGAEN; IPTNPTISVTTEILP; AAYYYVGYLPQRTFL; APHVYVFLHVYYVPQ) related to both MHC-I and II were predicted in S protein\(^{42}\) and only one of them overlapped with a highly antigenic epitope (PNTFTISVTTEILPV) predicted in our study. In another in silico study, 14 epitopes were predicted in S protein for T cells\(^{43}\) 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.

Docking analysis confirming the interaction among predicted epitopes and MHC-I/II alleles allows the prediction of more reliable epitopes that can be used for wet lab studies. In this study, docking analysis could not be performed for each of the all MHC-I or MHC-II epitopes due to the lack of 3D structures of some MHC-I/II alleles in protein database or server. This situation limits the docking analysis part of this study as preventing the analysis of all epitopes. Therefore, it was thought that increasing the number of 3D models of MHC-I or II alleles in PDB or servers would be useful for the analysis of a more robust epitope.

Development of computer-based methodologies enhances the credibility of in silico approaches in biological studies. Based on that, the methods of predicting vaccine candidate proteins are always favored even though they are not expressed in vitro. In silico methods also have the advantage of being able to make a fast and cost-efficient analysis. The other advantage of these methods is that they make predictions depending on the structure of vaccine candidate proteins and constitute a major way for vaccine design. Thus, in silico methods can be used at a

---

**Figure 3.** (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).
very early stage in the vaccine development process and this makes in silico methods essential as a pre-analysis approach before starting wet lab studies. Although the multiple numbers of proteins can be analysed by in silico methods for vaccine design studies, there are some limitations that should be taken into consideration. The lack of information in databases, inaccuracies of software algorithms, and usage of inappropriate tools for data are known limitations in terms of in silico studies. Therefore, it is important to select the right tools for analysis and utilize different parameters to find the correct results for in silico-based studies.

Conclusion

As spike protein has important roles such as viral attachment, fusion and, entry, it is a very significant strategic target for vaccine studies and a lot of companies and research groups use the protein for vaccine development. Our reverse vaccinology in silico approach also supports that S protein is the best vaccine candidate protein. In addition, probable secreted orf7a, orf8, and nsps-10 proteins with signal peptide can be promising vaccine candidates. Epitopes predicted in S protein and other proteins having a signal peptide may have a potential neutralising effect and can be used to develop an epitope-based peptide vaccine or a serodiagnostic assay. In the future, in addition to the currently studied S protein, antigenicity of orf7a, orf8, and nsps-10 proteins as well as significant epitopes selected in this study should be checked by wet lab studies and antigenic proteins/epitopes should be studied as vaccine or serodiagnostic candidates.

Methods

SARS-CoV-2 Wuhan isolate and variant proteins. NCBI (National Center for Biotechnology Information) (https://www.ncbi.nlm.nih.gov) was used to obtain the full genome of SARS-CoV-2 isolate (reference isolate Wuhan-Hu-1; Accession number: NC_045512.2 and sequenced in December 2019) and alignment and editing of the full genome was performed by MEGA 7 and BioEdit (Version 7.2)44,45. In addition to this reference genome, according to Nextstrain results showing the prevalent variations, variant D614G corresponding for S protein, variant T175M corresponding for M protein, variants P19L, S194L, S197L, R203K, and G204R corresponding for N protein, and variants S24L, V62L, and L84S corresponding for orf8 were selected as additional proteins to investigate the immunological effects associated with antigenicity, signal peptide, BetaWrap motifs, PTMs and, epitope regions. Variants were selected for only structural proteins and the proteins that were predicted to have a signal peptide.

Prediction of physico-chemical parameters and secondary structures. The reference genome proteins were investigated using Expasy ProtParam online server (https://web.expasy.org/protparam/) for the prediction of physico-chemical properties46. The prediction of solubility was performed by SolPro (http://scrat.proteomics.ics.uci.edu/)47. 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)48.

Prediction of antigenicity. The reference 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 0.44.

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/)50. For the prediction of the number of transmembrane helices, TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) was used.

Prediction of signal peptide. The reference genome proteins and variant proteins were analysed by Signal-BLAST (http://bsg.temple.edu/services.came.sbg.ac.at/_Signalblast.html)52.

Prediction of allergenicity. The allergenicity of the reference genome structural proteins, variant proteins, and the proteins that 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 epitopes53.

Prediction of BetaWrap motifs. The prediction of BetaWrap motifs of the reference genome structural proteins, variant proteins, and the proteins that have a signal peptide was carried out by BetaWrap online server (http://cb.csail.mit.edu/cb/betawrap/betawrap.html)54.

Prediction of similarity with host proteome. The reference genome structural proteins, variant proteins, and the proteins that 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 the reference genome structural proteins, variant proteins, and the proteins that have a signal peptide were carried out using NetNGlyc 1.0 server (http://www.cbs.dtu.dk/services/NetNGlyc/)55, NetOGlyc 4.0 server (http://www.cbs.dtu.dk/services/NetOGlyc/)56, NetPhos 3.1 server (http://www.cbs.dtu.dk/services/NetPhos/)57 and, GPS-MSP and GPS-PAIL running under CSS-Palm Online Service (http://csspalm.biocuckoo.org/online
e.php)39. 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 proteins39.

**Prediction of B cell epitopes.**  Linear B cell epitopes of the reference genome structural proteins, variant proteins, and the proteins that have a signal peptide were predicted by Bcepred (http://crdd.osdd.net/raghava/bcepred/)39 and Bepipred Linear Epitope Prediction 2.0 running under IEDB (the immune epitope database, https://www.iedb.org/)39 online servers.

**Prediction of MHC-I and MHC-II epitopes.**  The prediction of MHC-I and MHC-II epitopes of the reference genome structural proteins, variant proteins, and the proteins that have a signal peptide were analysed by IEDB (https://www.iedb.org/)39. 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 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)39. In addition, epitopes that have the highest antigenicity value were 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)39 and visualised on UCSF Chimera 1.14 tool40. For docking analyses conducted with MHC-II alleles, each epitope that was selected based on low IC50 value and being probable antigen was docked to its specific MHC-II allele by selecting specific alleles from the EpiDock Server (http://www.ddg-pharmfac.net/epidock/EpiDockPage.html)39.

**Prediction of solvent-exposure positions in epitopes.**  Each protein was aligned with its variants by Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)40. Then, a 3D structure pdb file for each protein was downloaded from the Protein Data Bank (PDB; http://www.rcsb.org/pdb/) or constructed by modelling using Swiss-Model (https://swissmodel.expasy.org). As an input for each protein, an alignment file and a 3D protein model file were uploaded and run on ESPript 3.0 (http://escript.ibcp.fr/EScript/EScript/)39 to predict amino acid solvent-exposure (accessibility) properties for epitopes.

**Comparison of significant epitopes among major lineages.**  For representation of each of 11 major lineage (including A, A1, A2, A3, A5, B, B1, B1.1, B2, B3, B4) given in a study39, 11 SARS-CoV-2 genome sequences were retrieved from GISAID database (http://www.gisaid.org) and checked for lineage analysis using Pangolin (https://pangolin.cog-uk.io). Then, protein coding regions for S, M, N, E, orf7a, orf8, and nsp-10 were downloaded and compared to find epitope differences among 11 SARS-CoV-2 sequences representing major lineages. Accession numbers for major lineages are MT049951, EPI_ISL_420879, LC528233, EPI_ISL_416538, EPI_ISL_530117, MT020781, EPI_ISL_420910, EPI_ISL_418263, MT033890, EPI_ISL_529598, NC_045512.

Received: 27 April 2020; Accepted: 10 December 2020
Published online: 28 December 2020

**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. Sci. China Life Sci. 63(3), 457–460. https://doi.org/10.1007/s11427-020-1837-5 (2020).

3. Zheng, J. SARS-CoV-2: An emerging coronavirus that causes a global threat. Int. J. Biol. Sci. 16(10), 1678–1685. https://doi.org/10.7150/ijbs.45053 (2020).

4. Boni, M. F. et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. https://doi.org/10.1038/s41564-020-0771-4 (2020).

5. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. Nat. Med. 26(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9 (2020).

6. Ye, Z. W. et al. Zoonotic origins of human coronaviruses. Int. J. Biol. Sci. 16(10), 1686–1697. https://doi.org/10.7150/ijbs.45472 (2020).

7. WHO (World Health Organization). at https://www.who.int, Accessed 13 Sept 2020.

8. Dangi, M., Kumari, R., Singh, B. & Chhillar, A. K. Advanced in silico tools for designing of antigenic epitope as potential vaccine candidates against coronavirus. In Bioinformatics: Sequences, Structures, Phylogeny (ed. Shanker, A.) 329–357 (Springer, Singapore, 2018).

9. Nascimento, I. P. & Leite, L. C. Recombinant vaccines and the development of new vaccine strategies. Braz. J. Med. Biol. Res. 45(12), 1102–1111. https://doi.org/10.1590/s1516-8913201200500142 (2012).

10. Can, H., Alak, S. E., Köseoğlu, A. E., Döşkaya, M. & Ün, C. Do Toxoplasma gondii apicoplast proteins have antigenic potential? An in silico study. Comput. Biol. Chem. 84, 107158. https://doi.org/10.1016/j.compbiolchem.2019.107158 (2020).

11. Meunier, M. et al. Identification of novel vaccine candidates against campylobacter through reverse vaccinology. J. Immunol. Res. 2016, 5715790. https://doi.org/10.1155/2016/5715790 (2016).

12. Nazir, Z., Afridi, S. G., Shah, M., Shams, S. & Khan, A. Reverse vaccinology and subtractive genomics-based putative vaccine targets identification for Burkholderia pseudomallei Bp1651. Microb. Pathog. 125, 219–229. https://doi.org/10.1016/j.micpath.2018.09.033 (2018).
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., M.K., A.Y.G. and C.U. conducted review and editing.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to C.U.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020