



OPEN

Immunoinformatic identification of B cell and T cell epitopes in the SARS-CoV-2 proteome

Stephen N. Croke, Inna G. Ovsyannikova, Richard B. Kennedy & Gregory A. Poland✉

A novel coronavirus (SARS-CoV-2) emerged from China in late 2019 and rapidly spread across the globe, infecting millions of people and generating societal disruption on a level not seen since the 1918 influenza pandemic. A safe and effective vaccine is desperately needed to prevent the continued spread of SARS-CoV-2; yet, rational vaccine design efforts are currently hampered by the lack of knowledge regarding viral epitopes targeted during an immune response, and the need for more in-depth knowledge on betacoronavirus immunology. To that end, we developed a computational workflow using a series of open-source algorithms and webtools to analyze the proteome of SARS-CoV-2 and identify putative T cell and B cell epitopes. Utilizing a set of stringent selection criteria to filter peptide epitopes, we identified 41 T cell epitopes (5 HLA class I, 36 HLA class II) and 6 B cell epitopes that could serve as promising targets for peptide-based vaccine development against this emerging global pathogen. To our knowledge, this is the first study to comprehensively analyze all 10 (structural, non-structural and accessory) proteins from SARS-CoV-2 using predictive algorithms to identify potential targets for vaccine development.

In December 2019, public health officials in Wuhan, China, reported the first case of severe respiratory disease attributed to infection with the novel coronavirus SARS-CoV-2¹. Since its emergence, SARS-CoV-2 has spread rapidly via human-to-human transmission², threatening to overwhelm healthcare systems around the world and resulting in the declaration of a pandemic by the World Health Organization³. The disease caused by the virus (COVID-19) is characterized by fever, pneumonia, and other respiratory and inflammatory symptoms that can result in severe inflammation of lung tissue and ultimately death—particularly among older adults or individuals with underlying comorbidities^{4–6}. As of this writing, the SARS-CoV-2 pandemic has resulted in 4 million confirmed cases of COVID-19 and over 280,000 deaths worldwide⁷.

SARS-CoV-2 is the third pathogenic coronavirus to cross the species barrier into humans in the past two decades, preceded by severe acute respiratory syndrome coronavirus (SARS-CoV)^{8,9} and Middle-East respiratory syndrome coronavirus (MERS-CoV)¹⁰. All three of these viruses belong to the β -coronavirus genus and have either been confirmed (SARS-CoV) or suggested (MERS-CoV, SARS-CoV-2) to originate in bats, with transmission to humans occurring through intermediary animal hosts^{11–14}. While previous zoonotic spillovers of coronaviruses have been marked by high case fatality rates (~10% for SARS-CoV; ~34% for MERS-CoV), widespread transmission of disease has been relatively limited (8,098 cases of SARS; 2,494 cases of MERS)¹⁵. In contrast, SARS-CoV-2 is estimated to have a lower case fatality rate (~2 to 4%) but is far more infectious and has achieved world-wide spread in a matter of months¹⁶.

As the number of COVID-19 cases continues to grow, there is an urgent need for a safe and effective vaccine to combat the spread of SARS-CoV-2 and reduce the burden on hospitals and healthcare systems. No licensed vaccine or therapeutic is currently available for SARS-CoV-2, although there are over 100 vaccine candidates reportedly in development worldwide. Seven vaccine candidates have rapidly progressed into Phase I/II clinical trials: adenoviral vector-based vaccines (CanSino Biologics, ChiCTR2000030906; University of Oxford, NCT04324606), nucleic-acid based vaccines encoding for the viral spike (S) protein (Moderna, NCT04283461; Inovio Pharmaceuticals, NCT04336410; BioNTech/Pfizer, 2020-001038-36), and inactivated virus formulations (Sinopharm, ChiCTR2000031809; Sinovac (NCT04352608)¹⁷. While the advancement of these vaccine candidates into clinical testing is promising, it is imperative they meet stringent endpoints for safety¹⁸. Preclinical studies of multiple experimental SARS-CoV vaccines have reported a Th2-type immunopathology in the lungs of

Mayo Clinic Vaccine Research Group, Mayo Clinic, Guggenheim Building 611C, 200 First Street SW, Rochester, MN 55905, USA. ✉email: poland.gregory@mayo.edu

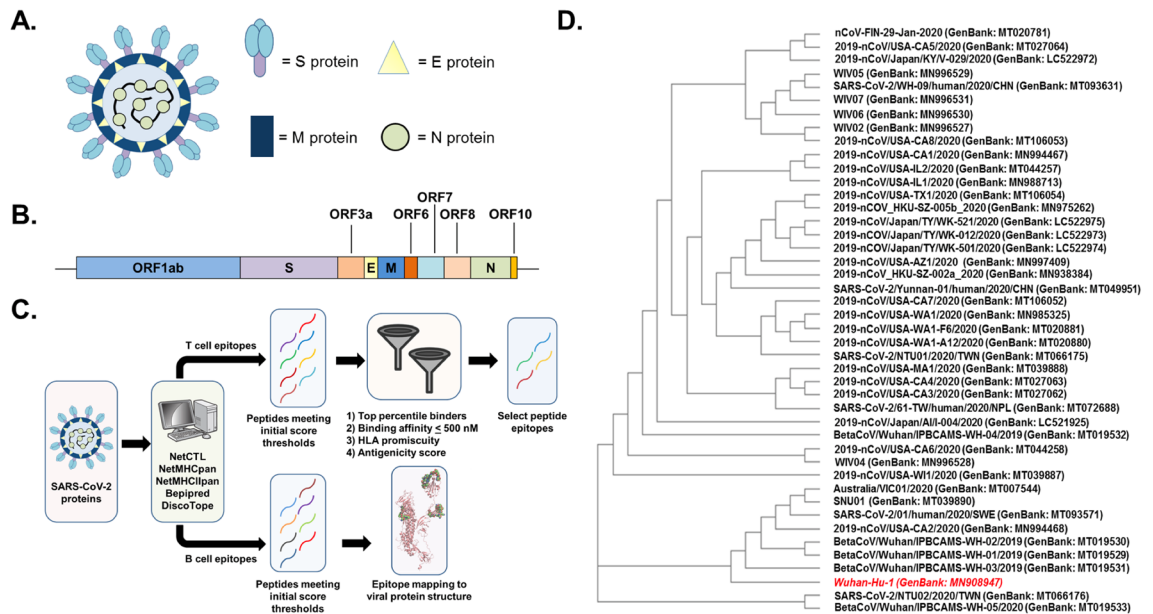


Figure 1. (A) Diagram of SARS-CoV-2 virion structure with the major structural proteins (S, M, N, and E) highlighted. (B) Cartoon representation of the SARS-CoV-2 genome with the 10 major protein-coding regions annotated. The box diagrams are proportional to the protein size. (C) Diagram of peptide identification workflow illustrating the algorithms used^{36,44–47,49–51,58,60} and filtering criterion applied to refine peptide selection. (D) Cladogram illustrating the genetic relationship of SARS-CoV-2 isolates. The original viral isolate and consensus sequence (Wuhan-Hu-1) is highlighted in red.

vaccinated mice following viral challenge, suggesting hypersensitization of the immune response against certain viral proteins^{19–22}. Similarly, a modified vaccinia virus Ankara vector expressing the SARS-CoV S protein induced significant hepatitis in immunizing ferrets²³. These data suggest that candidate coronavirus vaccines that limit the inclusion of whole viral proteins may have more beneficial safety profiles.

The SARS-CoV-2 genome encodes for 10 unique protein products: 4 structural proteins (surface glycoprotein (S), envelope (E), membrane (M), nucleocapsid (N)); 5 non-structural proteins (open reading frame (ORF)3a, ORF6, ORF7a, ORF8, ORF10); and 1 non-structural polyprotein (ORF1ab) (Fig. 1A,B)²⁴. There is currently very little known regarding which epitopes in the SARS-CoV-2 proteome are recognized by the human immune system, although a limited number of studies have recently reported a broad spectrum of cellular immune responses against the structural and non-structural proteins from SARS-CoV-2 among convalescent subjects^{25–27}. Studies of SARS-CoV immune responses suggest that both cellular and humoral responses against structural proteins mediate protection against disease^{19,22,28–30}, and it is likely that cellular immune responses against non-structural viral proteins also play a key role in orchestrating protective antiviral immunity^{31–33}. In lieu of biological data, immunoinformatic algorithms can be employed to predict peptide epitopes based on amino acid properties and known human leukocyte antigen (HLA) binding profiles^{34–36}. These computational approaches represent a validated methodology for rapidly identifying potential T cell and B cell epitopes for exploratory peptide-based vaccine development and have been recently used to identify target epitopes for MERS-CoV³⁷ and SARS-CoV-2, although many of these reports focus solely on structural proteins^{38–41}.

Herein, we employed a comprehensive immunoinformatics approach to identify putative T cell and B cell epitopes across the entire SARS-CoV-2 proteome (Fig. 1C). We independently identified peptides from each viral protein that were restricted to either HLA class I or HLA class II molecules across a subset of the most common HLA alleles in the global population. By filtering this list of peptides on the basis of predicted binding affinity, antigenicity, and promiscuity, we produced 5 HLA class I-restricted and 36 HLA class II-restricted peptides as leading candidates for further study. We also evaluated linear and structural B cell epitopes in the SARS-CoV-2 spike protein, with six antigenic regions identified as potential sites for antibody binding. These selected peptides may serve as initial candidates in the rational and accelerated design of a peptide-based vaccine against SARS-CoV-2.

Methods

Comparison of genome sequences from SARS-CoV-2 isolates. Genomic sequences for reported SARS-CoV-2 isolates were identified and retrieved from the Virus Pathogen Resource (ViPR) database on February 27, 2020 (https://www.viprbrc.org/brc/home.spg?decorator=corona_ncov). Sequences that did not cover the complete viral genome (~29,900 nucleotides) were excluded from further analysis. Remaining sequences were aligned using the Clustal Omega program (version 1.2.4) from the European Bioinformatics Institute⁴² and compared against the first reported genome sequence for SARS-CoV-2 (Wuhan-Hu-1; taxonomy ID: 2697049)¹.

Sequences from Wuhan-Hu-1 viral proteins were determined to be representative of those from all viral isolates and were subsequently used for epitope prediction analyses.

Prediction of SARS-CoV-2 T cell epitopes. Prediction of HLA class I and class II peptide epitopes was carried out with the 10 protein sequences reported for the Wuhan-Hu-1 isolate: E (GenBank accession: QHD43418); M (QHD43419); N (QHD43423); S (QHD43416); ORF3a (QHD43417); ORF6 (QHD43420); ORF7a (QHD43421); ORF8 (QHD43422); ORF10 (QHI42199); ORF1ab (QHD43415). We used standard methods similar to those previously applied to the analysis of SARS-CoV-2 protein sequences^{38,43}.

For CD8⁺ T cell epitope prediction, NetCTL 1.2 (Immune Epitope Database) was initially used to evaluate the binding of nonameric peptides derived from each viral protein to the most common HLA class I supertypes present among the human population^{44,45}. HLA class I molecules preferentially bind 9-mer peptides, and most algorithm training datasets have been based on peptides of this length. The weight placed on C-terminal cleavage and antigen transport efficiency was 0.15 and 0.05, respectively. The antigenic score threshold was 0.75. Peptides with scores above this threshold were subsequently analyzed on the NetMHCpan 4.0 server (Technical University of Denmark) to predict binding affinity and percentile rank across representative alleles of each major HLA class I supertype (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*24:02, HLA-B*07:02, HLA-B*08:01, HLA-B*27:05, HLA-B*40:01, HLA-B*58:01, HLA-B*15:01), which collectively cover the majority of class I alleles present in the human population^{46–48}. Thresholds for defining binding strength were set at 0.5% and 2.0% for strong and weak binders, respectively.

For CD4⁺ T cell epitope prediction, NetMHCIIpan 3.2 server (Technical University of Denmark) was used for predicting the binding affinity and percentile rank of 15-mer peptides derived from each viral protein across a reference panel of 27 HLA class II molecules^{36,49}. Thresholds for defining binding strength were set at 2% and 10% for strong and weak binders, respectively.

HLA class I and class II peptides with high predicted binding affinities (≤ 500 nM), high percentile ranks ($\leq 0.5\%$ for class I; $\leq 2\%$ for class II), and broad HLA coverage (≥ 3 alleles) were independently analyzed on the VaxiJen 2.0 server (Edward Jenner Institute)^{50,51} using a conservative score threshold (0.7) to predict antigenicity. Global population HLA allele coverage for this peptide subset was separately calculated for class I and class II molecules using the Population Coverage tool from IEDB⁵² and the predicted HLA alleles identified in our analyses. The potential toxicity and allergenicity of each peptide were calculated using the ToxinPred⁵³ and AllerCatPro⁵⁴ web tools, respectively. Default parameters were used for all sequence inputs.

Molecular docking of HLA class I peptides. Docking simulations of 5 HLA class I-restricted SARS-CoV-2 peptides with high antigenicity scores and a commonly shared predicted HLA molecule (HLA-DRB1*15:01) were performed using the GalaxyPepDock server (Seoul National University Laboratory of Computational Biology)⁵⁵. The structure of HLA-DRB1*15:01 was accessed from the Protein Data Bank as a co-crystallized structure of the HLA molecule with a nonameric SARS-CoV peptide (PDB ID: 3C9N)⁵⁶. The bound nonamer peptide was removed from the structure using Chimera 1.14 (University of California-San Francisco)⁵⁷ prior to running simulations. Ten models of each peptide-HLA complex were generated on the basis of minimized energy scores, and the top model for each complex was selected for comparative analysis.

Prediction and structural modeling of SARS-CoV-2 B cell epitopes. Linear B cell epitope predictions were performed on the three exposed SARS-CoV-2 structural proteins: S (GenBank accession: QHD43416), M (QHD43419), and E (QHD43418) using the BepiPred 1.0 algorithm⁵⁸. Epitope probability scores were calculated for each amino acid residue using a threshold of 0.35 (corresponding to >0.75 specificity and sensitivity below 0.5), and only epitopes ≥ 5 amino acid residues in length were further analyzed. The structure of the SARS-CoV-2 S protein was accessed from the Protein Data Bank (PDB ID: 6VSB)⁵⁹. Discontinuous (i.e., structural) B cell epitope predictions for the S protein structure were carried out using DiscoTope 1.1⁶⁰ with a score threshold greater than -7.7 (corresponding to >0.75 specificity and sensitivity below 0.5). The main protein structure was modeled in PyMOL (Schrödinger, LLC), with predicted B cell epitopes identified by both BepiPred 1.0 and DiscoTope 1.1 highlighted as spheres.

All data presented and analyzed were retrieved from ViPR, IEDB, and PDB as described. The tables, figures and supplementary files include all data generated and/or analyzed as a part of this study. Files of peptides and protein sequences compiled from ViPR and IEDB are available upon request.

Results

Genetic similarity of SARS-CoV-2 isolates. The primary goal of our study was to identify peptide epitopes that would be broadly applicable in vaccine development efforts against SARS-CoV-2. We identified 72 point mutations and 5 deletions across the genomes of 44 clinical isolates, with the majority of mutations ($n=46$) and deletions ($n=4$) occurring in the ORF1ab polyprotein (Supp. Figure S1, Supp. Table S1). Single-point mutations were also found in the S protein ($n=5$), N protein ($n=5$), ORF8 protein ($n=3$), ORF3a protein ($n=2$), E protein ($n=1$), and M protein ($n=1$). The remaining mutations ($n=10$) and 1 deletion were mapped to the untranslated regions (UTRs) of the SARS-CoV-2 genome. Despite the genetic diversity introduced by these events (Fig. 1D), matrix analysis determined that $>99\%$ sequence identity was maintained across all viral genomes. Based on these findings and for study feasibility, the genome from the original virus isolate (Wuhan-Hu-1; GenBank: MN908947) was selected as the consensus sequence for all further analyses.

Prediction of CD8⁺ T cell epitopes in the SARS-CoV-2 proteome. We next identified potential CD8⁺ T cell epitopes from all proteins in the SARS-CoV-2 proteome. Using the NetCTL 1.2 predictive algo-

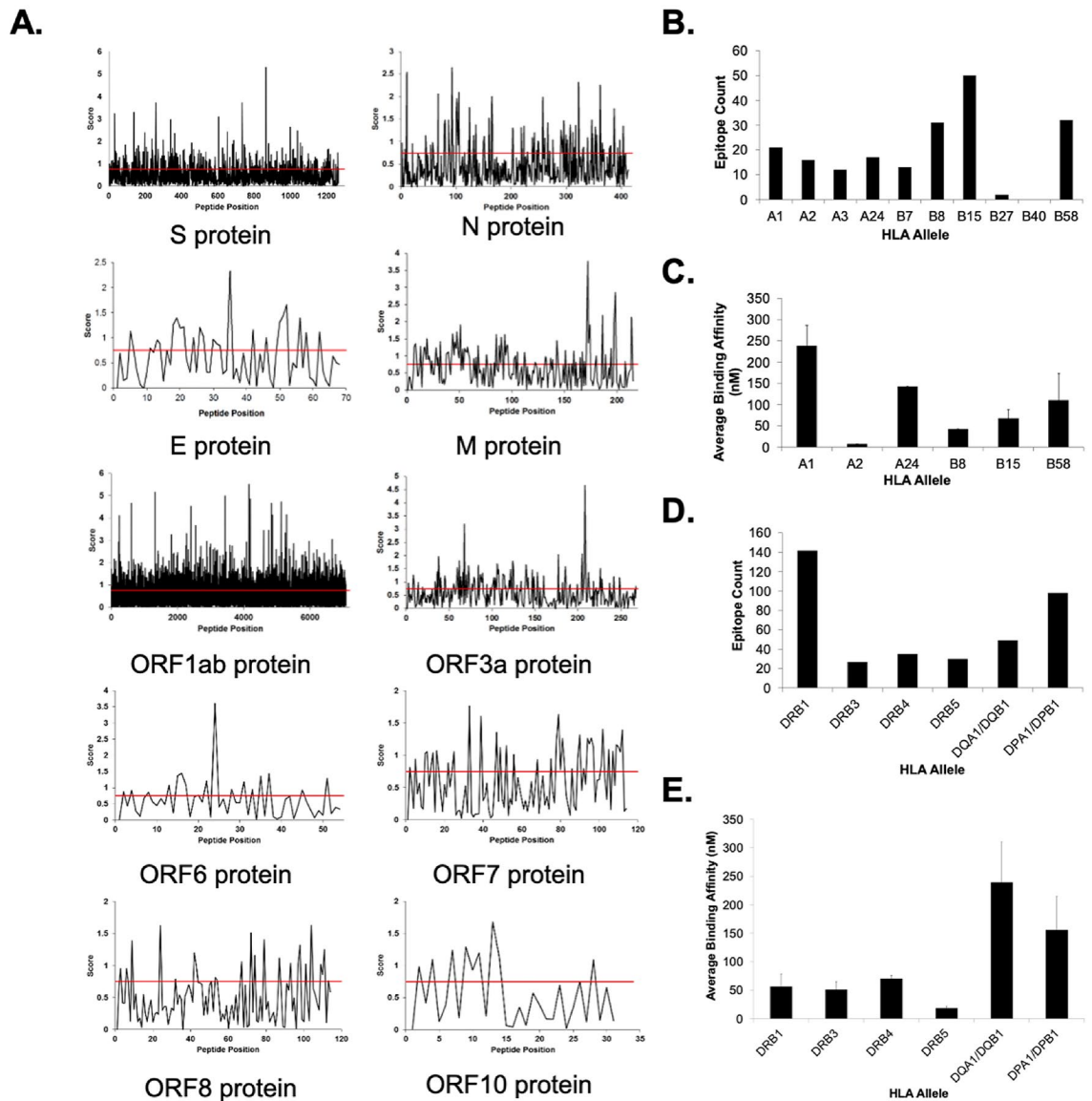


Figure 2. Immunogenicity scoring of peptides in the SARS-CoV-2 proteome with predicted HLA class I and II coverage and binding affinities. **(A)** Plots illustrating the NetCTL score for each sequential peptide across the entire amino acid sequence for each SARS-CoV-2 protein. Scores presented are the highest score identified across all HLA class I supertypes for each peptide. **(B)** Total number of predicted peptide epitopes distributed across HLA class I alleles. **(C)** Average predicted binding affinities by HLA allele for the top candidate class I peptides listed in Table 1. **(D)** Total number of predicted peptide epitopes distributed across HLA class II alleles. **(E)** Average predicted binding affinities by HLA allele for the top candidate class II peptides listed in Table 1.

rihm, we analyzed the complete amino acid sequence of each viral protein to generate sets of 9-mer peptides predicted to be recognized across at least one of the major HLA class I supertypes (Fig. 2A, Supp. Figure S2). This approach yielded a significant number of potential epitopes from each viral protein (ORF10: 9, ORF6: 17, ORF8: 23, E: 25, ORF7: 39, N: 80, M: 87, ORF3a: 87, S: 321, ORF1ab: 2814), with the number directly related to the size of the parent protein. We used the NetMHCpan 4.0 server to further refine the list of potential CD8⁺ T cell epitopes by predicting binding affinity across representative HLA class I alleles (see Methods) and assigning percentile scores to quantify binding propensity. Peptides with percentile rank scores $\leq 0.5\%$ (i.e., strong binders) were filtered using a 500 nM threshold for binding affinity to further delineate 740 candidate HLA class I epitopes from the viral proteome⁶¹. For feasibility reasons, we refined our selection to 83 candidate epitopes by excluding peptides predicted to bind only one HLA molecule (Supp. Table S1). The resultant peptides were enriched for predicted binders to HLA-B molecules (HLA-B*15:01 = 50; HLA-B*58:01 = 32; HLA-B*08:01 = 31) (Fig. 2B). A final round of selection on the basis of HLA promiscuity (i.e., predicted binding to ≥ 3 HLA molecules) and predicted antigenicity scoring using the VaxiJen 2.0 server produced a subset of five candidate peptides (four ORF1ab, one S protein) as potential targets for vaccine development (Table 1) with the hypothesis that increased HLA binding promiscuity meant broader population base coverage by those peptides. These pep-

tides were predicted to provide 74% global population coverage and had higher predicted binding affinities for HLA-B molecules (B*08:01 = 42.6 nM; B*15:01 = 67.7 nM; B*58:01 = 110.3 nM) compared to HLA-A molecules (A*01:01 = 238.6 nM; A*24:02 = 142.9 nM), with the exception of one ORF1ab-derived peptide (MMISAGFSL) that was predicted to bind HLA-A*02:01 with high affinity (IC_{50} = 6.9 nM) (Fig. 2C, Figure S3).

Prediction of CD4⁺ T cell epitopes in the SARS-CoV-2 proteome. We also sought to identify potential HLA class II peptides from SARS-CoV-2, as the stimulation of CD4⁺ T-helper cells is critical for robust vaccine-induced adaptive immune responses. Using the NetMHCIIpan 3.2 server, we identified 801 candidate HLA class II peptides from the viral proteome predicted to have high binding affinity (≤ 500 nM) and percentile rank scores $\leq 2\%$ across a reference panel of HLA molecules covering $>97\%$ of the population^{36,49}. Similar to HLA class I epitope predictions, the number of class II epitopes identified for each viral protein (ORF10: 4, E protein: 7, ORF7: 8, ORF8: 10, ORF6: 14, N: 15, M: 29, ORF3a: 31, S: 96, ORF1ab: 587) was largely proportional to protein size. After excluding peptides predicted to bind to only a single HLA molecule in our panel, we refined our selection to 211 peptides (Supp. Table S3), which were enriched for binding to HLA-DRB1 molecules ($n = 142$) (Fig. 2D). Filtering on HLA promiscuity and predicted antigenicity scores yielded a subset of 36 peptides (24 ORF1ab, 5 S protein, 2 M protein, 2 ORF7, 1 ORF3a, 1 ORF6, 1 ORF8) as CD4⁺ T cell epitopes for further study (Table 1). These peptides were predicted to collectively provide 99% population coverage and have significantly higher average binding affinities for HLA-DR alleles (DRB1 = 56.4 nM; DRB3 = 50.9 nM; DRB4 = 70.1 nM; DRB5 = 18 nM) compared to HLA-DP (155.9 nM) or HLA-DQ (238.6 nM) molecules (Fig. 2E, Figure S3). None of the peptides identified in our study (class I or class II) were predicted to be toxic or allergenic (Table S4).

Characterization of HLA class I peptide docking with HLA-B*15:01. The five candidate HLA class I peptides identified by our computational approach were predicted to provide coverage across six HLA alleles (A*01:01, A*02:01, A*24:02, B*08:01, B*15:01, B*58:01). The peptide FAMQMAYRF was the only candidate predicted to bind to A*24:02 molecules, whereas MMISAGFSL was predicted to uniquely bind A*02:01 and B*08:01 molecules. Four of the five peptides were predicted to bind A*01:01 and B*58:01 molecules, but all were predicted to bind with relatively high affinity (average IC_{50} = 67.7 nM) to HLA-B*15:01. Therefore, we performed molecular docking studies of each peptide with the molecular structure of HLA-B*15:01 (PDB: 3C9N).

All peptides were predicted to bind within the peptide binding groove, forming hydrogen bond contacts with numerous amino acid side chains (Fig. 3A). The binding motif for HLA-B*15:01 is highly selective for residues at the P2 and P9 anchor positions, with a preference for bulky hydrophobic amino acids at the C-terminus (Fig. 3B)⁶². All candidate peptides possessed terminal residues (Phe, Tyr, Leu) that fit into the hydrophobic binding pocket of the HLA groove, further supporting that these peptides should be strong binders of HLA-B*15:01 and promising candidates for vaccine development studies.

Prediction of B cell epitopes in SARS-CoV-2 proteins. An effective vaccine should stimulate both cellular and humoral immune responses against the target pathogen; therefore, we also sought to identify potential B cell epitopes from SARS-CoV-2 proteins. We limited our analysis to the primary structural proteins of the virus (S, N, M, and E), as these are the most accessible antigens for engaging B cell receptors. Using the Bepipred 1.0 algorithm, we identified 26 potential linear B cell epitopes in the S protein, 14 potential epitopes in the N protein, and 3 potential epitopes in the M protein (Table S5). No epitopes were identified in the E protein. Studies have previously shown the S protein to be the predominant target of neutralizing antibodies against coronaviruses^{63,64}, and, as our findings indicate this to likely be the case for SARS-CoV-2, we focused all subsequent analyses on the S protein. While the N protein is also a major target of the antibody response⁶⁵, it is unlikely these antibodies have any neutralizing activity based on the confinement of the N protein to the interior of intact virions. As epitope conformation can significantly influence recognition by antibodies, we also employed DiscoTope 1.1 to identify discontinuous B cell epitopes in the protein structure. Our analysis identified 16 potential structural epitopes in the S protein (9 in the S1 domain, 7 in the S2 domain), with six regions having significant overlap with our predicted linear epitopes (Table 2, Table S5). Antigenic regions identified in both analyses were modeled using the recently published structure of the SARS-CoV-2 S protein⁵⁹ to examine their accessibility for antibody binding. Epitopes in the S2 domain (P792-D796; Y1138-D1146) were clustered near the base of the spike protein, whereas regions in the S1 domain (D405-D428; N440-N450; G496-P507; D568-T573) were exposed on the protein surface (Fig. 4).

Discussion

In the face of the COVID-19 pandemic, it is imperative that safe and effective vaccines be rapidly developed in order to induce widespread herd immunity in the population and prevent the continued spread of SARS-CoV-2. Our study identified probable peptide targets of both cellular and humoral immune responses against SARS-CoV-2 using computational methodologies to investigate the entire viral proteome a priori. Studies such as these are paramount during the early stages of pandemic vaccine development given the relative scarcity of biological data available on the viral immune response, and we employed an approach that allowed us to systematically refine our predictions using increasingly stringent criteria to select a subset of the most promising epitopes for further study. The data we have curated could inform the design of a candidate peptide-based vaccine or diagnostic against SARS-CoV-2.

As selective pressures are known to introduce viral mutations that promote fitness and can lead to evasion of immune responses^{66,67}, we first sought to investigate the genetic similarity of all reported SARS-CoV-2 clinical isolates and identify a consensus sequence for use in our epitope prediction studies. The identification of amino acid mutations (and deletions) across the SARS-CoV-2 proteome was a critical step taken early in this study, as

Protein	Peptide	Residues	Antigenicity score	Predicted Alleles	Binding affinity (nM)
Class I					
S	FAMQMAYRF [#]	898–906	1.0278	A*24:02	142.9
				B*15:01	123.9
				B*58:01	23.4
ORF1ab	LSFKELLY	4758–4767	0.7234	A*01:01	371.8
				B*15:01	42.6
				B*58:01	35.7
ORF1ab	MMISAGFSL [#]	6425–6434	1.0248	A*02:01	6.9
				B*08:01	367.6
				B*15:01	16.2
ORF1ab	MSNLGMPY* [*]	2254–2262	0.9272	A*01:01	184.2
				B*15:01	74.1
				B*58:01	87.6
ORF1ab	STNVTIATY [#]	2273–2281	0.7143	A*01:01	241.1
				B*15:01	81.9
				B*58:01	294.5
Class II					
M	ASFRLFARTRSMWSE* [*]	98–112	0.7304	DRB1*01:01	19.2
				DRB1*07:01	30.9
				DRB1*08:02	53.5
				DRB1*09:01	49.9
				DRB1*11:01	12.2
				DRB5*01:01	16.3
				DPA1*02:01/DPB1*05:01	256.2
				DPA1*02:01 DPB1*14:01	387.3
M	LLQFAYANRRNFLYI* [*]	34–48	0.7387	DRB1*03:01	179.8
				DRB1*07:01	58.2
				DRB1*08:02	225.6
				DRB1*11:01	36.2
				DRB1*13:02	27.8
				DRB3*02:02	46.6
				DRB5*01:01	26.3
S	AAEIRASANLAATKM* [*]	1015–1029	0.7125	DRB1*08:02	101.3
				DRB1*13:02	23.0
				DRB3*02:02	52.7
				DQA1*01:02/DQB1*06:02	141.5
				DPA1*02:01/DPB1*14:01	327.4
S	ALQIPFAMQMAYRFN* [*]	893–907	1.0112	DRB1*09:01	52.9
				DRB1*12:01	159.5
				DRB1*15:01	50.3
S	PYRVVLSFELLHAP* [*]	507–521	0.8161	DPA1*02:01/DPB1*01:01	79.6
				DPA1*01:03/DPB1*02:01	53.3
				DPA1*01:03/DPB1*04:01	77.1
				DPA1*03:01/DPB1*04:02	92.9
S	QPVRVVLSFELLHA [#]	506–520	0.9109	DPA1*02:01/DPB1*01:01	73.2
				DPA1*01:03/DPB1*02:01	50.2
				DPA1*01:03/DPB1*04:01	71.4
				DPA1*03:01/DPB1*04:02	90.1
				DPA1*02:01/DPB1*05:01	211.1
S	YQPYRVVLSFELLH* [*]	505–519	0.9711	DPA1*02:01/DPB1*01:01	102.2
				DPA1*01:03/DPB1*04:01	93.0
				DPA1*03:01/DPB1*04:02	127.5
				DPA1*02:01/DPB1*05:01	299.3
ORF1ab	ANYIFWRNTNPIQLS [#]	7024–7038	1.0311	DRB1*04:05	89.9
				DRB1*07:01	35.2
				DRB1*13:02	13.5
Continued					

Protein	Peptide	Residues	Antigenicity score	Predicted Alleles	Binding affinity (nM)
ORF1ab	FKWDLTAFLGLVAEWF*	2314–2328	0.8059	DQA1*05:01/DQB1*02:01	178.3
				DQA1*03:01/DQB1*03:02	425.3
				DQA1*04:01/DQB1*04:02	349.3
ORF1ab	HIQWMVMFTPLVPFW*	3125–3139	0.7238	DQA1*01:01/DQB1*05:01	293.1
				DPA1*02:01/DPB1*01:01	116.3
				DPA1*01:03/DPB1*04:01	84.6
				DPA1*03:01/DPB1*04:02	135.4
ORF1ab	IINLVQMAPISAMVR*	4048–4062	0.7682	DRB1*01:01	12.8
				DRB1*08:02	118.8
				DRB4*01:01	54.7
ORF1ab	INLVQMAPISAMVRM*	4049–4063	0.9037	DRB1*12:01	176.9
				DRB4*01:01	57.1
				DQA1*01:02/DQB1*06:02	116.5
				DPA1*02:01/DPB1*14:01	398.6
ORF1ab	IVFMCVEYCPFFIT	3758–3772	1.0267	DPA1*02:01/DPB1*01:01	116.2
				DPA1*01:03/DPB1*02:01	53.9
				DPA1*01:03/DPB1*04:01	70.9
				DPA1*03:01/DPB1*04:02	144.9
ORF1ab	IVTALRANSVAVKLQN*	4127–4141	0.7692	DRB1*08:02	115.9
				DRB1*13:02	9.4
				DRB3*02:02	19.5
				DPA1*02:01/DPB1*14:01	408.7
ORF1ab	KGRLLIRENNRVVIS*	7075–7089	0.7821	DRB1*12:01	170.9
				DRB1*13:02	9.5
				DRB1*15:01	48.2
				DRB4*01:01	58.8
ORF1ab	KSAFYILPSIISNEK*	1350–1364	0.7169	DRB1*01:01	9.3
				DRB1*04:01	49.3
				DRB1*04:05	47.5
				DRB1*08:02	96.3
ORF1ab	LIVTALRANSVAVKLQ [‡]	4126–4140	0.7473	DRB1*01:01	8.8
				DRB1*07:01	39.2
				DRB4*01:01	78.6
				DQA1*01:02/DQB1*06:02	142.5
				DPA1*02:01/DPB1*14:01	368.3
ORF1ab	NLPFKLTCATTRQVV	2737–2751	1.1632	DRB1*07:01	35.9
				DRB1*09:01	58.6
				DRB5*01:01	23.9
ORF1ab	PASRELKVTFPPDLN	1950–1964	1.0155	DPA1*02:01/DPB1*01:01	76.9
				DPA1*01:03/DPB1*02:01	48.9
				DPA1*01:03/DPB1*04:01	64.3
				DPA1*03:01/DPB1*04:02	149.5
ORF1ab	PFAMGIIAMSAFAMM*	3613–3627	0.9834	DRB1*01:01	12.3
				DRB1*09:01	57.6
				DQA1*05:01/DQB1*03:01	45.6
ORF1ab	QMNLKYAISAKNRAR [‡]	4933–4947	1.5044	DRB1*01:01	14.9
				DRB1*04:01	56.9
				DRB1*08:02	49.1
				DRB1*09:01	45.2
				DRB1*11:01	22.1
				DRB3*02:02	84.9
				DPA1*02:01/DPB1*14:01	158.3
ORF1ab	QQKLALGGSVAIKIT	6956–6970	1.2533	DRB1*01:01	12.6
				DRB1*07:01	23.4
				DRB1*09:01	32.3
				DQA1*05:01/DQB1*03:01	42.9

Continued

Protein	Peptide	Residues	Antigenicity score	Predicted Alleles	Binding affinity (nM)
ORF1ab	RFKESPFELEDIFPM [#]	6709–6723	1.2101	DPA1*02:01/DPB1*01:01	74.0
				DPA1*01:03/DPB1*02:01	65.9
				DPA1*01:03/DPB1*04:01	81.9
				DPA1*03:01/DPB1*04:02	130.6
ORF1ab	SAFAMMFVKHKHAFI	3622–3636	0.7305	DRB1*08:02	110.4
				DRB1*11:01	18.3
				DRB1*15:01	50.9
				DRB4*01:01	79.2
				DRB5*01:01	15.1
ORF1ab	SFLAHIQWMVMFTPL [#]	3121–3135	0.8215	DPA1*02:01/DPB1*01:01	103.9
				DPA1*01:03/DPB1*02:01	47.8
				DPA1*01:03/DPB1*04:01	70.7
				DPA1*03:01/DPB1*04:02	140.6
ORF1ab	SIGFDYVYNPFMIDV [*]	6155–6169	1.0823	DPA1*02:01/DPB1*01:01	108.9
				DPA1*01:03/DPB1*02:01	47.1
				DPA1*01:03/DPB1*04:01	81.9
				DPA1*03:01/DPB1*04:02	137.6
ORF1ab	TEETFKLSYGIATVR [*]	5465–5479	0.8859	DRB1*01:01	8.7
				DRB1*07:01	21.8
				DRB1*09:01	25.9
ORF1ab	VLVQSTQWSLFFFLY [*]	3593–3607	0.7309	DPA1*02:01/DPB1*01:01	77.0
				DPA1*01:03/DPB1*02:01	35.3
				DPA1*01:03/DPB1*04:01	42.3
				DPA1*03:01/DPB1*04:02	93.1
ORF1ab	VQSTQWSLFFFLYEN [*]	3595–3609	0.7509	DPA1*02:01/DPB1*01:01	107.1
				DPA1*01:03/DPB1*02:01	49.9
				DPA1*03:01/DPB1*04:02	129.8
ORF1ab	WLIINLVQMAPISAM [#]	2366–2380	0.9389	DRB1*12:01	130.6
				DRB4*01:01	65.9
				DQA1*01:02/DQB1*06:02	139.6
ORF1ab	YFNMVYMPASWVMRI [*]	3649–3663	0.7244	DRB1*01:01	8.3
				DRB1*04:05	80.2
				DRB1*07:01	38.2
				DRB1*09:01	37.4
				DRB1*12:01	184.5
				DRB1*15:01	30.1
ORF3	KKRWQLALSkgVHFV [#]	66–80	0.8172	DRB1*01:01	9.2
				DRB1*07:01	11.6
				DRB1*08:02	200.3
				DRB1*09:01	17.9
				DRB1*11:01	43.1
				DRB1*12:01	119.6
				DRB1*13:02	30.0
				DRB1*15:01	34.2
				DRB4*01:01	79.8
				DRB5*01:01	18.4
ORF6	MFHLVDFQVTIAEIL [#]	1–15	1.0366	DQA1*05:01/DQB1*02:01	192.0
				DQA1*01:01/DQB1*05:01	292.1
				DPA1*02:01/DPB1*01:01	108.3
				DPA1*01:03/DPB1*04:01	100.7
ORF7	VKHVVQLRARSVSPK [#]	71–85	1.0865	DRB1*01:01	14.3
				DRB1*08:02	150.6
				DRB1*11:01	38.3
				DRB4*01:01	86.6
Continued					

Protein	Peptide	Residues	Antigenicity score	Predicted Alleles	Binding affinity (nM)
ORF7	NKFALTCFSTQFAFA*	52–66	1.1728	DPA1*02:01/DPB1*01:01	50.9
				DPA1*01:03/DPB1*02:01	29.1
				DPA1*01:03/DPB1*04:01	35.9
				DPA1*03:01/DPB1*04:02	80.2
				DPA1*02:01/DPB1*05:01	273.4
ORF8	SKWYIRVGARKSAPL*	43–57	0.8829	DRB1*01:01	13.7
				DRB1*08:02	87.8
				DRB1*09:01	50.7
				DRB1*11:01	15.3
				DRB5*01:01	8.8

Table 1. Top predicted HLA class I and class II T cell epitopes. *Significant sequence overlap with peptides reported in^{38,43}. *Exact peptide replicated from analyses reported in^{38,43}.

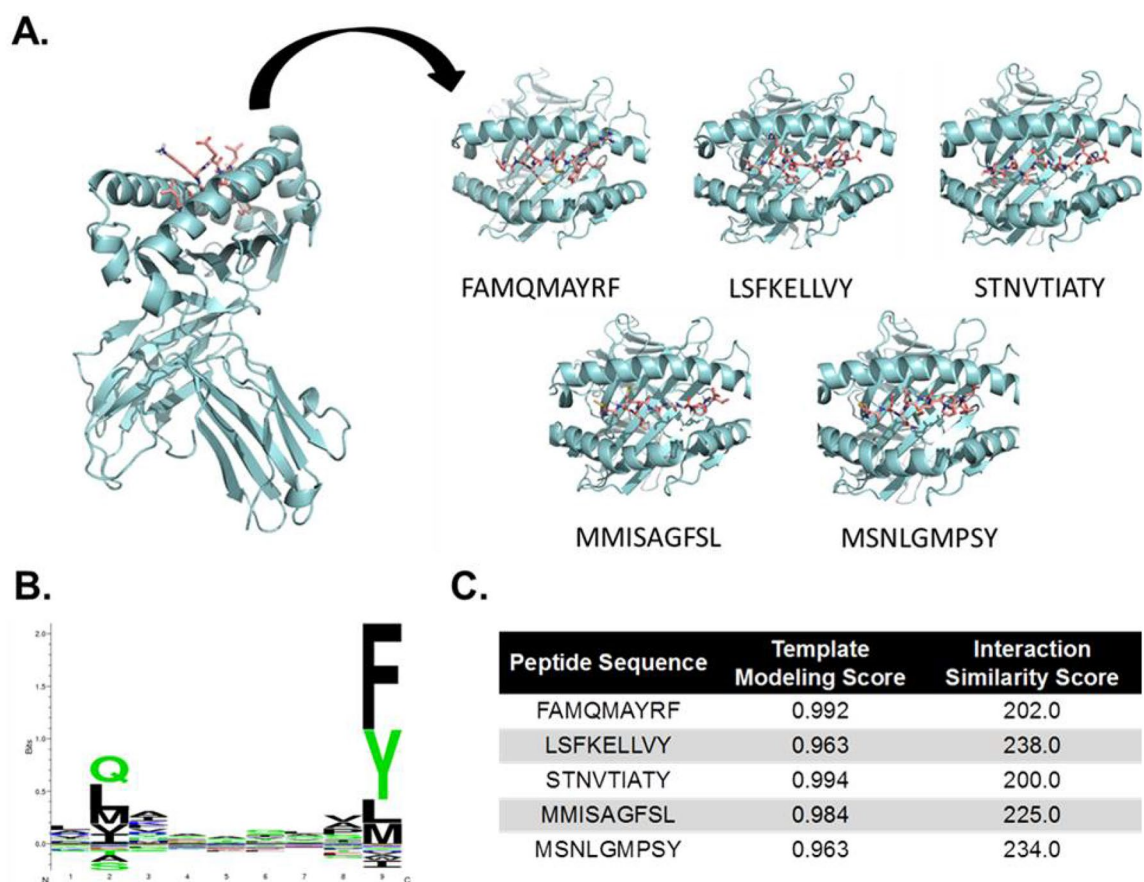


Figure 3. Docking of top predicted HLA class I peptides with a shared HLA molecule. (A) Structural docking model for each indicated peptide with the molecular structure of HLA-B*15:01 (PDB: 3C9N). Individual panels represent top-down views of the peptide binding groove. (B) Binding motif for HLA-B*15:01. (C) Template Modeling and Interaction Similarity scores for the selected peptide docking models shown in panel A^{81,82}.

we wanted to ensure the protein sequence analyzed with peptide epitope prediction algorithms was representative of the protein sequences in circulating viral variants. Mismatches between predicted peptides and viral proteins could compromise the efficacy and utility of such peptides as vaccine candidates or diagnostic agents. We identified 77 mutations/deletions across the 44 genomes of clinical isolates reported as of 27 February 2020 (Supp. Table S1). Despite these variations, the viral genomic identity was >99% conserved across all isolates. Many of these were silent mutations that did not impact the amino acid sequence, while those mutations that induced coding changes were largely limited to single isolates. As the protein coding sequences were largely conserved, the genome of the original virus isolate (Wuhan-Hu-1) was deemed a representative consensus sequence for analysis of the SARS-CoV-2 proteome.

Peptide	Residues	Bepipred score ^a	DiscoTope score ^a
DEVQRQIAPGQTGKIADYNYKLPDD	405–428	0.715	– 5.71
NLDSKVGGNYN	440–450	0.577	– 5.77
GFQPTNGVGYQP	496–507	1.01	– 5.73
DIADTT	568–573	0.853	– 5.55
PPIKD	792–796	0.936	– 3.28
VYDPLQPELDSF	1137–1148	0.747	– 4.12

Table 2. Top predicted B cell epitopes for the S protein. ^aReported scores represent the average calculated across all amino acids for the combined epitope sequence.

CD4⁺ and CD8⁺ T cell responses will likely be directed against both structural and non-structural proteins during antiviral immune responses, as all viral proteins are accessible for processing and presentation on the HLA molecules of infected cells. Therefore, we sought to identify T cell epitopes across the entire viral proteome. Our analysis identified 83 potential CD8⁺ T cell epitopes (Supp. Table S2) and 211 potential CD4⁺ T cell epitopes (Supp. Table S3), with stringent filtering for more promiscuous peptides with high predicted antigenicity yielding a subset of 5 CD8⁺ T cell epitopes and 36 CD4⁺ T cell epitopes (Table 1) as potential targets for vaccine development. A study by Grifoni and colleagues has recently reported the computational identification of 241 CD4⁺ T cell epitopes from SARS-CoV-2³⁸, and Srivastava et al. also recently reported the prediction of class II peptides from the SARS-CoV-2 proteome⁴³. Twenty-one peptides from our analysis shared sequence homology or were nested within peptides identified in these studies. Moreover, ten peptides from these initial reports were replicated in our final subset of HLA class II epitopes, supporting that these peptides may be promising vaccine targets.

An increasing number of studies have employed predictive algorithms to identify potential HLA class I epitopes for SARS-CoV-2, although relatively few have comprehensively analyzed the entire viral proteome. A report from Feng et al. recently outlined the identification of 499 potential class I epitopes in the main structural proteins from SARS-CoV-2 but did not consider any non-structural proteins⁴¹. Grifoni and colleagues conducted a more rigorous analysis, identifying 628 unique CD8⁺ T cell epitopes across all SARS-CoV-2 proteins but focusing their analyses solely on peptides with sequence homology to known SARS-CoV epitopes³⁸. Our approach initially identified ~3,500 potential CD8⁺ T cell epitopes across all viral proteins, which we refined to a subset of 5 peptides (Table 1). Three of these peptides (i.e., FAMQMAYRE, STNVTIATY, MMISAGFSL) were replicated from previous studies^{38,43}. The MMISAGFSL peptide derived from ORF1ab was predicted to bind HLA-A*02:01 with high affinity (IC₅₀ = 6.9 nM) (Fig. 2C). Given the prevalence of this allele in the American and European populations (25–60% frequency)⁶⁸, MMISAGFSL may represent a promising epitope capable of providing broad vaccine population coverage.

We also observed a notable enrichment of epitopes predicted to bind HLA-B molecules—particularly HLA-B*15:01—as we imposed more stringent selection criteria (Fig. 2B). All five peptides identified by our approach were predicted to be relatively strong binders for this allele (IC₅₀ = 67.7 nM), with molecular docking simulations illustrating strong contacts with amino acid residues in the peptide binding groove (Fig. 3A,B). A recent computational study identified another HLA-B allele (B*15:03) as having a high capacity for presenting epitopes from SARS-CoV-2 that were conserved among other pathogenic coronaviruses⁶⁹. These data collectively suggest the HLA-B locus may be significantly associated with the immune response to SARS-CoV-2 (and potentially other coronaviruses), with further biological studies warranted to determine the true role of host genetics in SARS-CoV-2 immunology.

Lastly, we analyzed the primary structural proteins of SARS-CoV-2 (S, N, M, E proteins) for potential B cell epitopes, as an ideal vaccine would be designed to stimulate both cellular and humoral immunity. Our analysis identified potential linear B cell epitopes in all proteins except for the E protein (Table 2). The greatest number of epitopes were predicted in the surface-exposed S protein (n = 26), but a significant number of epitopes were also predicted for the N protein (n = 14). This is not surprising, as previous reports identified the N protein as a significant target of the humoral response to SARS-CoV^{70,71}. As the S protein is the predominant surface protein and has been the primary target of neutralizing antibody responses against other coronaviruses^{63,64}, we elected to focus our subsequent analyses solely on antigenic regions in the S protein. We identified 16 potential structural epitopes in the S protein structure and referenced against our linear epitope predictions to identify six regions that were independently identified by both analyses (Table 2, Fig. 4). Feng et al. recently reported the computational identification of 19 surface epitopes in the S protein using Bepipred and the Kolaskar method⁴¹, four of which had significant sequence overlap with the regions identified by our analyses.

To further evaluate the potential of these six antigenic regions as targets for antibody binding, we modeled their surface accessibility on the crystal structure of the SARS-CoV-2 spike protein⁵⁹. Four regions in the S1 domain (D405–D428; N440–N450; G496–P507; D568–T573) were solvent exposed (Fig. 4A,B), with minimal steric hindrance for antibody accessibility. The S1 domain contains the residues (N331–V524) important for virus binding to angiotensin converting enzyme 2 (ACE2) on the cell surface⁷², and studies have shown that antibodies with potent neutralizing activity against SARS-CoV target this domain^{73–75}. Indeed, three of the four S1 epitopes identified in our analyses are located in the ACE2-binding region, supporting their potential utility in vaccine development against SARS-CoV-2. Two regions were identified in the S2 “stalk” domain of the S protein (Fig. 4A,C). While V1137–F1148 is located at the base of the S protein and likely inaccessible to antibodies, P792–D796 is on the outer face of the protein and has been previously identified as part of a larger B cell epitope

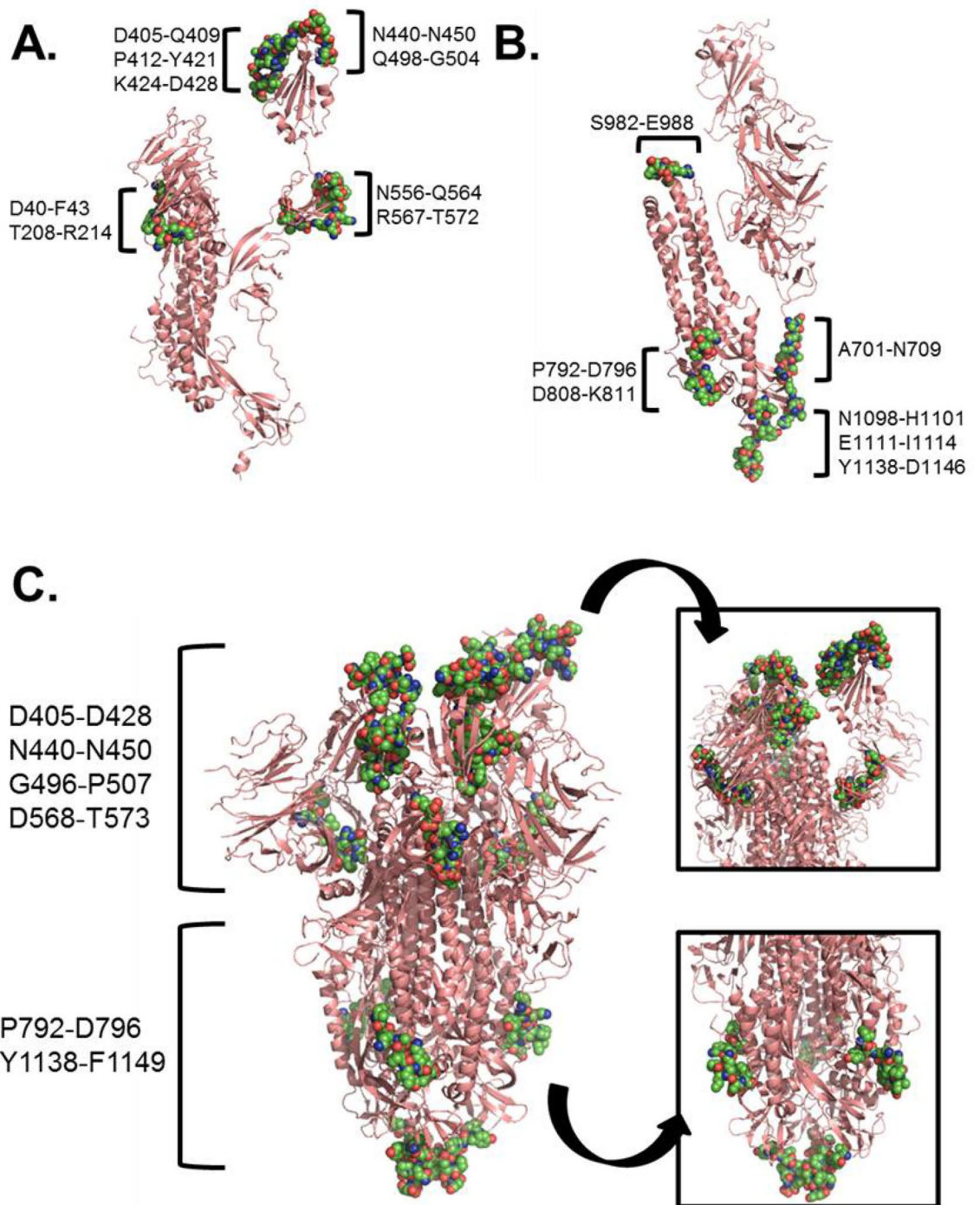


Figure 4. Modeling of predicted B cell epitopes on the crystal structure of the S glycoprotein. Predicted structural epitopes in the S1 domain (A) and S2 domain (B) highlighted on the structure of the S glycoprotein monomer (PDB: 6VSB). (C) Top predicted B cell epitopes identified by both Bepipred and DiscoTope prediction algorithms highlighted on the trimeric structure of the S glycoprotein. Inset panels show the S1 domain (upper) and S2 domain (lower). Predicted epitopes are highlighted as colored atoms (green, blue, red) on the surface of the S protein (salmon).

that is conserved with SARS-CoV³⁸. As SARS-CoV S2-specific antibodies have previously been shown to possess antiviral activity⁷³, it is interesting to speculate whether a strategy similar to targeting the influenza hemagglutinin protein stalk could be employed for developing a broadly reactive coronavirus vaccine.

Our study possessed several strengths and limitations. Rather than restricting our analyses of HLA class I and class II epitopes to specific proteins based on prior studies of SARS-CoV immunology, we investigated the complete proteome of SARS-CoV-2 using an unbiased approach. Furthermore, we employed a multi-tiered strategy for identifying putative B cell and T cell epitopes from all viral proteins studied. Our initial analyses

were performed with liberal thresholds for epitope identification, and at each additional step, we imposed more stringent selection criteria to filter these peptides to a subset of B cell and T cell epitopes for further study. Nevertheless, the results of this study are derived purely from computational methods, and it should be noted that computational algorithms can fail to capture a significant number of antigenic peptides⁷⁶. Experimental validation with biological samples will ultimately be needed.

During the early stages of a pandemic, access to sufficient biological samples may be extremely limited, so we must continue to utilize methodologies—such as computational predictive algorithms—that allow us to explore the epitope landscape for experimental vaccine development. Our approach in this study allowed us to identify and refine a manageable subset of T cell and B cell epitopes for further testing as components of a SARS-CoV-2 vaccine. Based on our results, our proposed SARS-CoV-2 vaccine formulation could contain the following: (1) one or more B cell peptide epitopes from the S protein to generate protective neutralizing antibodies; and (2) multiple HLA class I and class II-derived peptides from other viral proteins to stimulate robust CD8⁺ and CD4⁺ T cell responses. Based on global allele frequencies, these class I and class II peptides would be expected to collectively provide 74% and 99% population coverage, respectively. While such a vaccine could be readily formulated as a synthetic polypeptide or an adjuvanted peptide mixture, these strategies may not retain the epitope structural features necessary to induce a robust antibody response. Recombinant nanoparticles and assembly into VLPs represent promising alternative vaccine platforms, as they have been extensively used for the controlled display and delivery of peptide-based vaccine components^{77–80}. By omitting whole viral proteins from the vaccine formulation, a peptide-based SARS-CoV-2 vaccine containing both class I and class II peptides should have a well-tolerated safety profile and promote a balanced Th1/Th2 response that avoids the Th2-biased adverse events previously observed with experimental SARS-CoV vaccines^{19–22}. However, it should be noted that computational algorithms cannot currently predict the overall nature of an immune response or the potential for immunopathologies to develop after vaccination, as these processes are influenced by several factors (e.g., antigen dose, adjuvant system, administration route, antigen-release kinetics). Extensive biological testing of these peptides in experimental vaccine formulations will be required to ascertain information in this regard.

In summary, we have identified 41 potential T cell epitopes (5 HLA class I, 36 HLA class II) and 6 potential B cell epitopes from across the SARS-CoV-2 proteome that are predicted to have broad population coverage and could serve as the basis for designing investigational peptide-based vaccines. Further study on the biological relevance, immunogenicity, and immune response profiles of these peptides is warranted in an effort to develop a safe and effective vaccine to combat the SARS-CoV-2 pandemic.

Data availability

All data presented and analyzed were retrieved from ViPR, IEDB, and PDB as described. The tables, figures and supplementary files include all data generated and/or analyzed as a part of this study. Files of peptides and protein sequences compiled from ViPR and IEDB are available upon request.

Received: 20 May 2020; Accepted: 31 July 2020

Published online: 25 August 2020

References

1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269. <https://doi.org/10.1038/s41586-020-2008-3> (2020).
2. Chan, J. F. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514–523. [https://doi.org/10.1016/S0140-6736\(20\)30154-9](https://doi.org/10.1016/S0140-6736(20)30154-9) (2020).
3. Cucinotta, D. & Vanelli, M. WHO declares COVID-19 a pandemic. *Acta Biomed.* **91**, 157–160. <https://doi.org/10.23750/abm.v91i1.9397> (2020).
4. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395**, 507–513. [https://doi.org/10.1016/S0140-6736\(20\)30211-7](https://doi.org/10.1016/S0140-6736(20)30211-7) (2020).
5. Wang, D. *et al.* Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069. <https://doi.org/10.1001/jama.2020.1585> (2020).
6. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506. [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5) (2020).
7. World Health Organization. Coronavirus disease (COVID-19) Situation Report - 113. https://www.who.int/docs/default-source/coronavirus/situation-reports/20200512-covid-19-sitrep-113.pdf?sfvrsn=feac3b6d_2. (2020).
8. Drosten, C. *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967–1976. <https://doi.org/10.1056/NEJMoa030747> (2003).
9. Ksiazek, T. G. *et al.* A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1953–1966. <https://doi.org/10.1056/NEJMoa030781> (2003).
10. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. & Fouchier, R. A. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820. <https://doi.org/10.1056/NEJMoa1211721> (2012).
11. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679. <https://doi.org/10.1126/science.1118391> (2005).
12. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273. <https://doi.org/10.1038/s41586-020-2012-7> (2020).
13. Memish, Z. A. *et al.* Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerg. Infect. Dis.* **19**, 1819–1823. <https://doi.org/10.3201/eid1911.131172> (2013).
14. Haagmans, B. L. *et al.* Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *Lancet Infect Dis.* **14**, 140–145. [https://doi.org/10.1016/S1473-3099\(13\)70690-X](https://doi.org/10.1016/S1473-3099(13)70690-X) (2014).
15. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281–292. <https://doi.org/10.1016/j.cell.2020.02.058> (2020).
16. Weston, S. & Frieman, M. B. COVID-19: knowns, unknowns, and questions. *mSphere*. <https://doi.org/10.1128/mSphere.00203-20> (2020).

17. World Health Organization. Draft landscape of COVID-19 candidate vaccines. <https://www.who.int/who-documents-detail/draft-landscape-of-covid-19-candidate-vaccines>. (2020).
18. Poland, G. A. Tortoises, hares, and vaccines: a cautionary note for SARS-CoV-2 vaccine development. *Vaccine* **38**, 4219–4220 (2020).
19. Tseng, C. T. *et al.* Immunization with SARS coronavirus vaccines leads to pulmonary immunopathology on challenge with the SARS virus. *PLoS ONE* **7**, e35421. <https://doi.org/10.1371/journal.pone.0035421> (2012).
20. Deming, D. *et al.* Vaccine efficacy in senescent mice challenged with recombinant SARS-CoV bearing epidemic and zoonotic spike variants. *PLoS Med.* **3**, e525. <https://doi.org/10.1371/journal.pmed.0030525> (2006).
21. Yasui, F. *et al.* Prior immunization with severe acute respiratory syndrome (SARS)-associated coronavirus (SARS-CoV) nucleocapsid protein causes severe pneumonia in mice infected with SARS-CoV. *J Immunol.* **181**, 6337–6348. <https://doi.org/10.4049/jimmunol.181.9.6337> (2008).
22. Bolles, M. *et al.* A double-inactivated severe acute respiratory syndrome coronavirus vaccine provides incomplete protection in mice and induces increased eosinophilic proinflammatory pulmonary response upon challenge. *J Virol.* **85**, 12201–12215. <https://doi.org/10.1128/JVI.06048-11> (2011).
23. Weingartl, H. *et al.* Immunization with modified vaccinia virus Ankara-based recombinant vaccine against severe acute respiratory syndrome is associated with enhanced hepatitis in ferrets. *J Virol.* **78**, 12672–12676. <https://doi.org/10.1128/JVI.78.22.12672-12676.2004> (2004).
24. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8) (2020).
25. Grifoni, A. *et al.* Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* <https://doi.org/10.1016/j.cell.2020.05.015> (2020).
26. Braun, J. *et al.* Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. Preprint at <https://doi.org/10.1101/2020.04.17.20061440> (2020).
27. Peng, Y. *et al.* Broad and strong memory CD4⁺ and CD8⁺ T cells induced by SARS-CoV-2 in UK convalescent COVID-19 patients. Preprint at <https://doi.org/10.1101/2020.06.05.134551> (2020).
28. Channappanavar, R., Fett, C., Zhao, J., Meyerholz, D. K. & Perlman, S. Virus-specific memory CD8 T cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *J Virol.* **88**, 11034–11044. <https://doi.org/10.1128/JVI.01505-14> (2014).
29. Ng, O. W. *et al.* Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. *Vaccine* **34**, 2008–2014. <https://doi.org/10.1016/j.vaccine.2016.02.063> (2016).
30. Zhao, J. *et al.* Airway memory CD4(+) T cells mediate protective immunity against emerging respiratory coronaviruses. *Immunity* **44**, 1379–1391. <https://doi.org/10.1016/j.immuni.2016.05.006> (2016).
31. Lorente, E. *et al.* Structural and nonstructural viral proteins are targets of T-helper immune response against human respiratory syncytial virus. *Mol. Cell Proteomics* **15**, 2141–2151. <https://doi.org/10.1074/mcp.M115.057356> (2016).
32. Ip, P. P. *et al.* Alphavirus-based vaccines encoding nonstructural proteins of hepatitis C virus induce robust and protective T-cell responses. *Mol. Ther.* **22**, 881–890. <https://doi.org/10.1038/mt.2013.287> (2014).
33. Henriques, H. R. *et al.* Targeting the non-structural protein 1 from dengue virus to a dendritic cell population confers protective immunity to lethal virus challenge. *PLoS Negl. Trop. Dis.* **7**, e2330. <https://doi.org/10.1371/journal.pntd.0002330> (2013).
34. Tomar, N. & De, R. K. Immunoinformatics: a brief review. *Methods Mol. Biol.* **1184**, 23–55. https://doi.org/10.1007/978-1-4939-1115-8_3 (2014).
35. Backert, L. & Kohlbacher, O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med.* **7**, 119. <https://doi.org/10.1186/s13073-015-0245-0> (2015).
36. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**, 394–406. <https://doi.org/10.1111/imm.12889> (2018).
37. Tahir Ul Qamar, M. *et al.* Epitope-based peptide vaccine design and target site depiction against Middle East Respiratory Syndrome Coronavirus: an immune-informatics study. *J Transl Med.* **17**, 362. <https://doi.org/10.1186/s12967-019-2116-8> (2019).
38. Grifoni, A. *et al.* A Sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe* **27**, 671–680 e672. <https://doi.org/10.1016/j.chom.2020.03.002> (2020).
39. Fast, E., Altman, R. B. & Chen, B. Potential T-cell and B-cell Epitopes of 2019-nCoV. Preprint at <https://doi.org/10.1101/2020.02.19.955484> (2020).
40. Seema, M. T cell epitope-based vaccine design for pandemic novel coronavirus 2019-nCoV. Preprint at https://chemrxiv.org/articles/T_Cell_Epitope-Based_Vaccine_Design_for_Pandemic_Novel_Coronavirus_2019-nCoV/12029523 (2020).
41. Feng, Y.-E. *et al.* Multi-epitope vaccine design using an immunoinformatics approach for 2019 novel coronavirus in China (SARS-CoV-2). Preprint at 2020, <https://doi.org/10.1101/2020.03.03.962332> (2019).
42. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucl. Acids Res.* **47**, W636–W641. <https://doi.org/10.1093/nar/gkz268> (2019).
43. Srivastava, S. *et al.* Structural basis to design multi-epitope vaccines against Novel Coronavirus 19 (COVID19) infection, the ongoing pandemic emergency: an in silico approach. Preprint at <https://doi.org/10.1101/2020.04.01.019299> (2020).
44. Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinform.* **8**, 424 (2007).
45. Larsen, M. V. *et al.* An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* **35**, 2295–2303. <https://doi.org/10.1002/eji.200425811> (2005).
46. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1–13. <https://doi.org/10.1007/s00251-008-0341-z> (2009).
47. Jurtz, V. *et al.* NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol.* **199**, 3360–3368. <https://doi.org/10.4049/jimmunol.1700893> (2017).
48. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 33. <https://doi.org/10.1186/s13073-016-0288-x> (2016).
49. Greenbaum, J. *et al.* Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different super-types and a surprising degree of repertoire sharing across supertypes. *Immunogenetics* **63**, 325–335 (2011).
50. Doytchinova, I. A. & Flower, D. R. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinform.* **8**, 4. <https://doi.org/10.1186/1471-2105-8-4> (2007).
51. Doytchinova, I. A. & Flower, D. R. Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine* **25**, 856–866. <https://doi.org/10.1016/j.vaccine.2006.09.032> (2007).
52. Bui, H. H. *et al.* Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinform.* **7**, 153 (2006).
53. Gupta, S. *et al.* In silico approach for predicting toxicity of peptides and proteins. *PLoS ONE* **8**, e73957. <https://doi.org/10.1371/journal.pone.0073957> (2013).
54. Maurer-Stroh, S. *et al.* AllerCatPro-prediction of protein allergenicity potential from the protein sequence. *Bioinformatics* **35**, 3020–3027. <https://doi.org/10.1093/bioinformatics/btz029> (2019).
55. Ko, J., Park, H., Heo, L. & Seok, C. GalaxyWEB server for protein structure prediction and refinement. *Nucl. Acids Res.* **40**, W294–297. <https://doi.org/10.1093/nar/gks493> (2012).

56. Roder, G., Kristensen, O., Kastrop, J. S., Buus, S. & Gajhede, M. Structure of a SARS coronavirus-derived peptide bound to the human major histocompatibility complex class I molecule HLA-B*1501. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **64**, 459–462. <https://doi.org/10.1107/S1744309108012396> (2008).
57. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612. <https://doi.org/10.1002/jcc.20084> (2004).
58. Larsen, J. E., Lund, O. & Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2**, 2. <https://doi.org/10.1186/1745-7580-2-2> (2006).
59. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263. <https://doi.org/10.1126/science.abb2507> (2020).
60. Haste Andersen, P., Nielsen, M. & Lund, O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* **15**, 2558–2567. <https://doi.org/10.1110/ps.062405906> (2006).
61. Sette, A. *et al.* The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* **153**, 5586–5592 (1994).
62. Roder, G. *et al.* Crystal structures of two peptide-HLA-B*1501 complexes; structural characterization of the HLA-B62 supertype. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 1300–1310. <https://doi.org/10.1107/S0907444906027636> (2006).
63. Okba, N. M. A. *et al.* SARS-CoV-2 specific antibody responses in COVID-19 patients. Preprint at <https://doi.org/10.1101/2020.03.18.20038059> (2020).
64. Wang, Q. *et al.* Immunodominant SARS coronavirus epitopes in humans elicited both enhancing and neutralizing effects on infection in non-human primates. *ACS Infect. Dis.* **2**, 361–376. <https://doi.org/10.1021/acsinfectdis.6b00006> (2016).
65. Zhang, L. *et al.* Anti-SARS-CoV-2 virus antibody levels in convalescent plasma of six donors who have recovered from COVID-19. *Aging* **12**, 6536–6542. <https://doi.org/10.18632/aging.103102> (2020).
66. Doud, M. B., Hensley, S. E. & Bloom, J. D. Complete mapping of viral escape from neutralizing antibodies. *PLoS Pathog.* **13**, e1006271. <https://doi.org/10.1371/journal.ppat.1006271> (2017).
67. Keck, M. L., Wrensch, F., Pierce, B. G., Baumert, T. F. & Fong, S. K. H. Mapping determinants of virus neutralization and viral escape for rational design of a hepatitis C virus vaccine. *Front. Immunol.* **9**, 1194. <https://doi.org/10.3389/fimmu.2018.01194> (2018).
68. Ellis, J. M. *et al.* Frequencies of HLA-A2 alleles in five U.S. population groups. Predominance of A*02011 and identification of HLA-A*0231. *Human Immunol.* **61**, 334–340 (2000).
69. Nguyen, A. *et al.* Human leukocyte antigen susceptibility map for SARS-CoV-2. Preprint at <https://doi.org/10.1101/2020.03.22.20040600> (2020).
70. Huang, L. R. *et al.* Evaluation of antibody responses against SARS coronavirus nucleocapsid or spike proteins by immunoblotting or ELISA. *J. Med. Virol.* **73**, 338–346. <https://doi.org/10.1002/jmv.20096> (2004).
71. Qiu, M. *et al.* Antibody responses to individual proteins of SARS coronavirus and their neutralization activities. *Microbes Infect.* **7**, 882–889. <https://doi.org/10.1016/j.micinf.2005.02.006> (2005).
72. Tai, W. *et al.* Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell Mol. Immunol.* **17**, 613–620 (2020).
73. Zeng, F. *et al.* Quantitative comparison of the efficiency of antibodies against S1 and S2 subunit of SARS coronavirus spike protein in virus neutralization and blocking of receptor binding: implications for the functional roles of S2 subunit. *FEBS Lett.* **580**, 5612–5620. <https://doi.org/10.1016/j.febslet.2006.08.085> (2006).
74. Berry, J. D. *et al.* Neutralizing epitopes of the SARS-CoV S-protein cluster independent of repertoire, antigen structure or mAb technology. *MAbs* **2**, 53–66. <https://doi.org/10.4161/mabs.2.1.10788> (2010).
75. He, Y. *et al.* Identification and characterization of novel neutralizing epitopes in the receptor-binding domain of SARS-CoV spike protein: revealing the critical antigenic determinants in inactivated SARS-CoV vaccine. *Vaccine* **24**, 5498–5508. <https://doi.org/10.1016/j.vaccine.2006.04.054> (2006).
76. Johnson, K. L., Ovsyannikova, I. G., Mason, C. J., Bergen, H. R. III. & Poland, G. A. Discovery of naturally processed and HLA-presented class I peptides from vaccinia virus infection using mass spectrometry for vaccine development. *Vaccine* **28**, 38–47 (2009).
77. Zhang, L. *et al.* Development of autologous C5 vaccine nanoparticles to reduce intravascular hemolysis in vivo. *ACS Chem Biol.* **12**, 539–547. <https://doi.org/10.1021/acscchembio.6b00994> (2017).
78. Brune, K. D. *et al.* Plug-and-display: decoration of Virus-Like Particles via isopeptide bonds for modular immunization. *Sci. Rep.* **6**, 19234. <https://doi.org/10.1038/srep19234> (2016).
79. Zhai, L. *et al.* A novel candidate HPV vaccine: MS2 phage VLP displaying a tandem HPV L2 peptide offers similar protection in mice to Gardasil-9. *Antiviral Res.* **147**, 116–123. <https://doi.org/10.1016/j.antiviral.2017.09.012> (2017).
80. McCarthy, D. P., Hunter, Z. N., Chackerian, B., Shea, L. D. & Miller, S. D. Targeted immunomodulation using antigen-conjugated nanoparticles. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.* **6**, 298–315. <https://doi.org/10.1002/wnan.1263> (2014).
81. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710. <https://doi.org/10.1002/prot.20264> (2004).
82. Lee, H., Heo, L., Lee, M. S. & Seok, C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucl. Acids Res.* **43**, W431–435. <https://doi.org/10.1093/nar/gkv495> (2015).

Acknowledgements

The authors would like to thank Caroline L. Vitse for editorial assistance with this manuscript. The research presented here was not supported by any specific funding source.

Author contributions

S.N.C., I.G.O., R.B.K., and G.A.P. developed the computational workflow used in this analysis; S.N.C. retrieved genomic sequences from databases and carried out the analysis; R.B.K. and I.G.O. supervised the project; S.N.C., I.G.O., R.B.K., and G.A.P. interpreted the results; S.N.C. drafted the manuscript with significant input from I.G.O., R.B.K., and G.A.P.; all authors reviewed and approved the final version of the paper.

Competing interests

Dr. Poland is the chair of a Safety Evaluation Committee for novel investigational vaccine trials being conducted by Merck Research Laboratories. Dr. Poland offers consultative advice on vaccine development to Merck & Co., Medicago, GlaxoSmithKline, Sanofi Pasteur, Emergent Biosolutions, Dynavax, Genentech, Eli Lilly and Company, Janssen Global Services LLC, Kentucky Bioprocessing, and Genavant Sciences, Inc. Drs. Poland, Kennedy, and Ovsyannikova hold patents related to vaccinia, influenza, and measles peptide vaccines. Drs. Poland, Kennedy, and Ovsyannikova have received grant funding from ICW Ventures for preclinical studies on a peptide-based COVID-19 vaccine. Dr. Kennedy has received funding from Merck Research Laboratories to study waning

immunity to mumps vaccine. These activities have been reviewed by the Mayo Clinic Conflict of Interest Review Board and are conducted in compliance with Mayo Clinic Conflict of Interest policies. This research has been reviewed by the Mayo Clinic Conflict of Interest Review Board and was conducted in compliance with Mayo Clinic Conflict of Interest policies. All other authors declare no competing financial interests. This research has been reviewed by the Mayo Clinic Conflict of Interest Review Board and was conducted in compliance with Mayo Clinic Conflict of Interest policies.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-70864-8>.

Correspondence and requests for materials should be addressed to G.A.P.

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