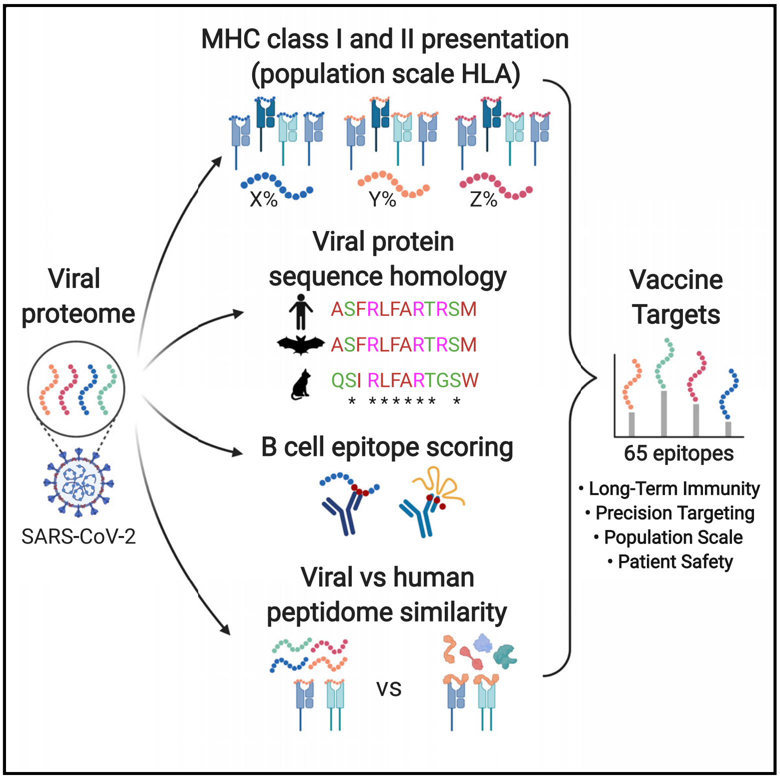
Cell Reports
Medicine

Identification of SARS-CoV-2 Vaccine Epitopes Predicted to Induce Long-Term Population-Scale Immunity

Graphical Abstract



Highlights

- Selecting optimal epitopes is essential for vaccine safety and efficacy
- We report 65 vaccine peptides predicted to drive long-term immunity in most people
- Epitopes contain domains conserved in 15 coronaviruses and newly evolved SARS2 regions
- Epitopes can be used to generate B and/or T cell vaccines (RNA and DNA)

Authors

Mark Yarmarkovich, John M. Warrington, Alvin Farrel, John M. Maris

Correspondence

maris@chop.edu

In Brief

Yarmarkovich et al. report SARS-CoV-2 peptides for use in multi-epitope vaccines. These peptides are predicted to activate CD4 and CD8 T cells, are highly dissimilar from the self-proteome, and are conserved across 15 related coronaviruses. Presented epitopes are expected to drive long-term immunity in the majority of the population.

Yarmarkovich et al., 2020, Cell Reports Medicine *1*, 100036
June 23, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.xcrm.2020.100036

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Report

Identification of SARS-CoV-2 Vaccine Epitopes Predicted to Induce Long-Term Population-Scale Immunity

Mark Yarmarkovich,[1] John M. Warrington,[1] Alvin Farrel,[1,2] and John M. Maris[1,3,4,*]
[1]Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[2]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[3]Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Lead Contact
*Correspondence: maris@chop.edu
https://doi.org/10.1016/j.xcrm.2020.100036

SUMMARY

Here we propose a SARS-CoV-2 vaccine design concept based on identification of highly conserved regions of the viral genome and newly acquired adaptations, both predicted to generate epitopes presented on major histocompatibility complex (MHC) class I and II across the vast majority of the population. We further prioritize genomic regions that generate highly dissimilar peptides from the human proteome and are also predicted to produce B cell epitopes. We propose sixty-five 33-mer peptide sequences, a subset of which can be tested using DNA or mRNA delivery strategies. These include peptides that are contained within evolutionarily divergent regions of the spike protein reported to increase infectivity through increased binding to the ACE2 receptor and within a newly evolved furin cleavage site thought to increase membrane fusion. Validation and implementation of this vaccine concept could specifically target specific vulnerabilities of SARS-CoV-2 and should engage a robust adaptive immune response in the vast majority of the population.

INTRODUCTION

The current SARS-CoV-2 pandemic has precipitated an urgent need for a safe and effective vaccine to be developed and deployed in a highly accelerated time frame as compared with standard vaccine development processes.[1] Upfront selection of epitopes most likely to induce a safe and effective immune response can accelerate these efforts. Optimally designed vaccines maximize immunogenicity toward regions of proteins that contribute most to protective immunity, while minimizing the antigenic load contributed by unnecessary protein domains that may result in autoimmunity, reactogenicity, or even enhanced infectivity. Here we present an immunogenicity map of SARS-CoV-2 generated to inform vaccine design based on analyses across five parameters: (1) stimulation of CD4 and CD8 T cells; (2) immunogenicity across the majority of human histocompatability leukocyte antigen (HLA) alleles; (3) targeting both evolutionarily conserved regions and newly divergent regions of the virus that increase infectivity; (4) targeting linear and conformational B cell epitopes; and (5) targeting viral regions with the highest degree of dissimilarity to the self-immunopeptidome, such as to maximize safety and immunogenicity. We present a list of SARS-CoV-2 minigenes and propose their use in multivalent vaccine constructs that should generate T and/or B cell epitopes that can be delivered by scalable manufacturing techniques such as DNA or nucleoside mRNA.

SARS-CoV-2 is the third coronavirus in the past two decades to acquire infectivity in humans and result in regional epidemics, and the first to cause a worldwide pandemic. The spike (S) glycoprotein of coronaviruses mediates host cell entry and dictates species tropism, with the SARS-CoV-2 S protein reported to bind its target protein angiotensin I converting enzyme 2 (ACE2) with 10- to 20-fold higher affinity than SARS-CoV in humans.[2,3] In addition, insertion of a novel protease cleavage site[4] is predicted to confer increased virulence by facilitating the cleavage necessary to expose the fusion peptide that initiates membrane fusion, enabling a crucial step of viral entry into host cells.[5,6] It is now clear that coronavirus disease 2019 (COVID-19) results when SARS-CoV-2 infects type II pneumocytes lining the pulmonary alveoli that co-express ACE2 and the transmembrane serine protease 2 (TMPRSS2)[7], likely impairing release of surfactants that maintain surface tension. This impairment hinders the ability to prevent accumulation of fluid, ultimately resulting in acute respiratory distress syndrome.[8,9] The immune response of convalescent COVID-19 patients consists of antibody-secreting cells releasing IgG and IgM antibodies, increased follicular helper T cells, and activated CD4 and CD8 T cells,[10] suggesting that a broad humoral and T cell-driven immune response mediates the clearance of infection, and that vaccination strategies directed at multiple arms of the immune response can be effective. The large size of the SARS-CoV-2 (~30 kb) suggests that selection of optimal epitopes and reduction of unnecessary antigenic load for vaccination may be essential for safety and efficacy.

Rapid deployment of antibody-based vaccination against SARS-CoV-2 raises the concern of accelerating infectivity through antibody-dependent enhancement (ADE), the facilitated viral entry into host cells mediated by subneutralizing antibodies (those capable of binding viral particles, but not neutralizing them).[11] ADE mechanisms have been described with other members of the *Coronaviridae* family.[12,13] It has already been suggested that some of the heterogeneity in COVID-19 cases may be caused by ADE from prior infection from other viruses in the coronavirus family.[14]

Although the immunogenicity map presented in this study can be used to inform multiple modalities of vaccine development, we present peptide sequences that are expected to be safe and immunogenic for use in T cell-based vaccination, and highlight B cell epitopes derived from peptides within the regions of the S protein involved in infectivity that we expect will minimize the risk for ADE. Because it has been shown that T helper (Th) cell responses are essential in humoral immune memory response,[15,16] we anticipate that the T cell epitopes generated from the peptide sequences presented here will aid the activation of CD4 T cells to drive memory B cell formation and somatic hypermutation when paired with matched B cell epitopes.

The potential of epitope-based vaccines to induce a cytolytic T cell response and drive memory B cell formation is complicated by the diversity of HLA alleles across the human population. The HLA locus is the most polymorphic region of the human genome, resulting in differential presentation of antigens to the immune system in each individual. Therefore, individual epitopes may be presented in a mutually exclusive manner across individuals, confounding the ability to immunize a population with broadly presented antigens. Whereas T cell receptors (TCRs) recognize linearized peptides anchored in the major histocompatibility complex (MHC) groove, B cell receptors (BCRs) can recognize both linear and conformational epitopes, and are therefore difficult to predict without prior knowledge of a protein structure. Here we describe an approach for prioritizing viral epitopes derived from a prioritized list of 33-mer peptides predicted to safely target the vulnerabilities of SARS-CoV-2, generate highly immunogenic epitopes on both MHC class I and II in the vast majority of the population, and maximize the likelihood that these peptides will drive an adaptive memory response.

RESULTS

We applied our recently published methods for scoring population-scale HLA presentation of all 9-mer peptides along the length of individual oncoproteins in human cancer to analyze the population-scale HLA presentation of peptides derived from all 10 SARS-CoV-2 genes across 84 class I HLA alleles,[17] representing 99.4% of the population as calculated based on allele frequencies reported in the Bone Marrow Registry.[18] A total of 6,098 SARS-CoV-2-derived peptides were predicted to bind to no HLA class I alleles, and thus we consider them immunogenically silent. In contrast, 3,524 SARS-CoV-2 epitopes were predicted to generate strong binders with least one HLA class I allele. Indeed, peptide FVNEFYAYL was predicted to bind 30 HLA alleles, representing 90.2% of the US population (Figure 1A, top; Table S1).

We next tested various peptide sequence lengths to maximize HLA presentation on multiple alleles within a single k-mer, finding that 33 amino acids generated maximal population-scale HLA presentation. We show that 99.7% of all 9,303 possible 33-mers are predicted to generate at least one HLA class I epitope, and propose that expression and presentation of these 33-mers in dendritic cells is expected to induce an immune response across a significant proportion of the population.[19,20] We identified viral regions predicted to generate epitopes that would present across the majority of the population, highlighting a single 33-mer ISNSWLMWLIINLVQMAPISAMVRMYIFFASFY containing multiple epitopes predicted to bind 82 of the 84 HLAs alleles, suggesting that this single 33-mer can potentially induce an immune response in up to 99.4% of the population given proper antigen processing (Table S1).

Because presentation by MHC class II is necessary for robust memory B and T cell responses,[15,16] we analyzed presentation of these viral epitopes on 36 MHC class II HLA alleles, representing 92.6% of the population (Figure 1A, bottom; Table 1; Table S1). Peptides derived from the 33-mer IAMSAFAMM FVKHKHAFLCLFLLPSLATVAYFN were predicted on 24 HLA class II alleles, representing 82.1% of the US population; peptides from the same 33-mer were predicted to be presented on 74 HLA class I alleles with a population frequency rate of 98.6%, showing that a single 33-mer can contain epitopes predicted to be presented on HLA class I and II across the majority of the population. Because HLA frequencies vary significantly by population, the frequency of individual HLA alleles can be adjusted based on specific populations using the SARs-CoV-2 immunogenicity map presented here, to customize vaccine design for groups with distinct HLA allele distributions (Table S1).

Next, we sought to identify the most highly conserved regions of the SARS-CoV-2 virus, positing that conserved regions are essential to viral replication and maintaining structural integrity, while non-conserved regions can tolerate mutations and result in antigens prone to immune evasion. To do this, we compared the amino acid sequence of SARS-CoV-2 with 14 closely related mammalian alpha and beta coronaviruses (human, bat, pig, and camel) from the *Coronaviridae* family (Table S2), scoring each amino acid for conservation across the viral strains. Additionally, we scored the conservation across the 727 SARS-CoV-2 genes sequences available at the time of this analysis (Table S2), equally weighing contributions from cross-species and interhuman variation (scores normalized to 0–1, with entirely conserved regions scoring 1). As expected, evolutionary divergence was greatest in the tropism-determining S protein and lowest in ORF1ab, which contains 16 proteins involved in viral replication (Figure 1B, bottom).

We then compared predicted viral MHC-presented epitopes with self-peptides presented in normal tissue on 84 HLA alleles across the entire human proteome as listed in the UniProt database, prioritizing antigens that are most dissimilar from self-peptides based on: (1) higher predicted safety based on decreased likelihood of inducing autoimmunity due to cross-reactivity with similar self-peptides presented on MHC; and (2)

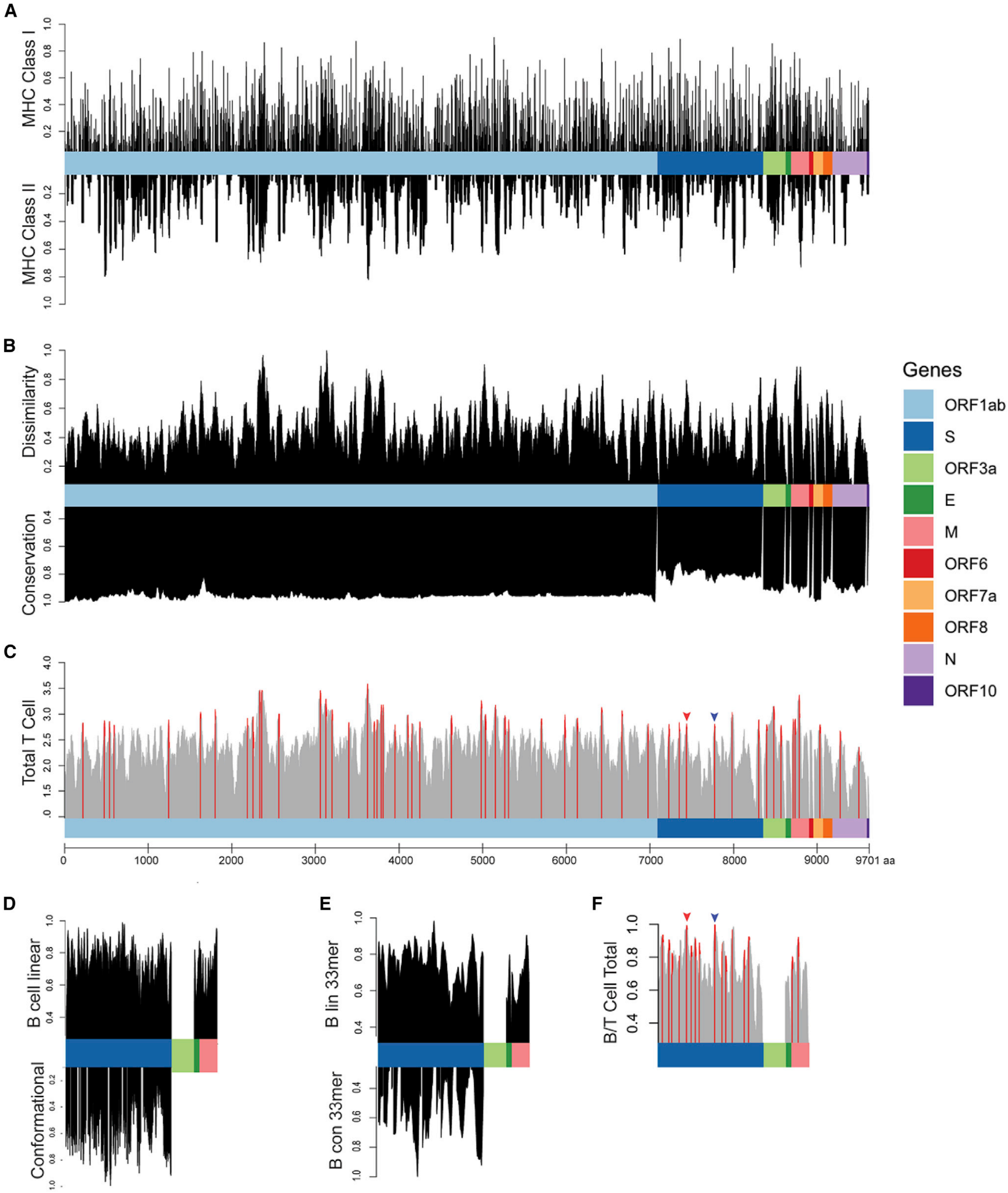


Figure 1. Epitope Scoring along SARS-CoV-2 Proteome

(A) HLA presentation of 33-mers across viral proteome. Representation of MHC class I presentation (top) and MHC class II presentation (bottom) reported as frequency of the population predicted to present peptides derived from each region of the viral proteome.

(legend continued on next page)

higher immunogenicity of dissimilar peptides based on an expected greater repertoire of antigen-specific T cells resulting from a lower degree of negative thymic selection. To analyze the similarity of the viral peptidome to human, we compared the 3,524 viral epitopes predicted to be presented on MHC against the normal human proteome on each of their MHC binding partners, testing each of 12,383 peptide/MHC pairs against the entire human proteome (85,915,364 normal peptides predicted across 84 HLA alleles). We assigned a similarity score for each peptide across all MHC peptides contained within a 33-mer, with high scoring peptides representing the highest degree of dissimilarity as compared with the space of all possible MHC epitopes derived from the normal proteome and a score of 0 representing an identical match in the human proteome (STAR Methods; Figure 1B, bottom; Table S1). We find regions of the viral proteome that are identical or highly similar to portions of the normal human proteome predicted to be presented on MHC, suggesting that an immune response mounted against these viral epitopes could result in an autoimmune response, while other high-scoring regions are highly dissimilar from self and expected to generate antigens with minimal likelihood of cross-reactivity (Table S1).

To assign an overall score for putative T cell antigens, we normalized each of our four scoring parameters (represented in Figures 1A and 1B) between 0 and 1 and summed each metric to obtain a final 33-mer peptide score, highlighting the local maxima of potentially generated epitopes scoring in the 90th percentile (55 top scoring T cell peptides) across 10 SARS-CoV-2 genes as peptide sequences for vaccination (Figure 1C; Table S3).

Finally, we sought to characterize B cell epitopes, assessing linear epitopes in S, matrix (M), and envelope (E) proteins that are exposed and expected to be accessible to antibodies; we also characterized conformational epitopes in the S protein for which structural data are available using BepiPred 2.0 and DiscoTope 2.0.[22,23] We found a strong concordance between linear and conformational epitope scores ($p < 2e^{-16}$). Next, we performed an agnostic scoring of individual amino acid residues in S, M, and E proteins (Figure 1D), and then used these scores to generate scores for 33-mer peptides along the length of the protein (Figure 1E). The 33-mer VGGNYNYLYRLFRKSNLKPFERDISTEIYQAGS derived from S protein at position 445 ranked the highest based on combined linear and conformational B cell epitope scoring. We combined T cell epitope scores calculated above with available B cell epitope scores derived from the S, M, and E genes, providing a list of 65 peptides predicted to stimulate both humoral and cellular adaptive immunity (Figure 1F; Table S5).

To estimate the accuracy of our predictions, we compared the 65 unique 33-mer peptides presented in Table S5 with 92 epitopes derived from the first SARS virus (SARS-CoV) in the Immune Epitope Database (IEDB; https://www.iedb.org/home_v3.php) shown to elicit T cell responses. We found a significant enrichment in immunogenic peptides contained within the 65 selected SARS-CoV-2 33-mers as compared with the 33-mers not selected (p = 0.041), and find that the 33-mer AQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAI derived from the N protein contains five immunogenic MHC class I and II antigens previously reported from SARS-CoV (GMSRIGMEV, MEVTPSGTWL, AQFAPSASAFFGMSRIGM, AFFGMSRIGMEVTPSGTW, and AQFAPSASAFFGMSR) within the single 33-mer (Table 1), demonstrating that epitopes selected using this analysis's epitopes are more likely to be processed and immunogenic based on previous studies with SARS-CoV, and supporting the hypothesis that a single 33-mer is capable of generating multiple unique epitopes presented by multiple HLA alleles. We also found that a significant proportion of the peptides present within prioritized 33-mer have been predicted to bind MHC based on structural predictions.[24]

In addition to prioritizing evolutionarily conserved regions, we sought to specifically target acquired vulnerabilities in SARS-CoV-2 by focusing on features of this coronavirus that have been shown to contribute to its increased infectivity. The receptor binding domain (RBD) of the SARS-CoV-2 S protein has been reported to have 10- to 20-fold higher binding affinity to ACE2.[2] We show that viral epitope GEVFNATRFASVYAWNRKRISNCVADYSVLYNS derived from the RBD of the S protein (position 339–372) scores in the 90.9th percentile of T epitopes and is the third of 1,546 epitopes scored in the S, E, and M genes for combined B and T cell epitopes, with presentation by MHC class I in 98.3% of the population (Figures 1C, 1F, and 2, red). Additionally, a recently evolved furin cleavage site has been reported in the SARS-CoV-2 virus, resulting in increased infectivity.[2] Indeed, we find that the SYQTQTNSP**RRAR**SVASQSIIAYTMSLGAENSV peptide containing the RRAR furin cleavage site of the S protein ranked in the 90.7th percentile of T cell epitopes and ranks first among the 1,546 combined B and T cell epitopes (Figures 1C, 1F, blue, and 2, orange), thereby targeting an additional evolutionary adaptation of SARS-CoV-2 with the highest overall scoring B and T cell epitope. Based on a recently published study that identified receptor binding hotspots deduced by comparing structures of ACE2 bound to the S protein from SARS-CoV-2 as compared with SARS-CoV,[21] we searched for 33-mers containing the five acquired residues that increase S binding to ACE2, identifying KPFERDISTEIYQ**A**GSTP**C**NGVEG**FNC**YFPLQS as the highest ranked peptide sequence

(B) Scoring of each epitope derived from the 33-mers along the length of the proteome as compared with the epitopes derived from the normal human proteome presented across 84 HLA alleles, reported as normalized scores in which the highest scoring epitopes are maximally dissimilar to self-peptides derived from normal proteins (top). Scoring for genomic conservation against 15 cross-species coronaviruses and 727 human sequences, with highest scoring regions conserved across human and other mammalian coronaviruses (bottom).

(C) Combined epitope score reported as sum of four above parameters (local maximum for epitopes with 90th percentile total score in red).

(D) Scoring of B cell epitopes for each amino acid for linear epitopes for Spike, Envelope, and Matrix proteins (top) and conformational epitopes in Spike protein (bottom).

(E) Combined scoring of 33-mer epitopes as described in (D).

(F) Combined B and T cell epitope scoring in Spike, Envelope, and Matrix proteins. Receptor binding domain epitope highlighted with red arrow and epitope containing furin cleavage site highlighted with blue arrow (Figure 2).

Table 1. Sample of Highest Scoring Viral Epitopes Suggested for Vaccination Based on MHC Class I Population-Scale Presentation, MHC Class II Population Presentation, Similarity Score, and Homology Score across 15 Mammal Species and 727 Human SARS-CoV-2 Gene Sequences

Gene Position	Epitope	HLA Class I Population Presentation	HLA Class I Alleles Bound	HLA Class I Binders	HLA Class II Population Presentation	HLA Class II Alleles Bound	HLA Class II Binders	Dissimilarity Score	Conservation Score	Combined T Cell Score	B Cell Total Score	B and T Cell Total Percentile
ORF1ab_3619	IAMSAFAMMFVKHKHAFLCLFLLPSLATVAYFN	98.6%	74	HLA-A: 0101, 0201, 0202, 0203, 0205, 0206, 0207, 0211, 0212, 0216, 0217, 0219, 0301, 1101, 2301, 2403, 2501, 2601, 2602, 2603, 2902, 3001, 3002, 3201, 3207, 6601, 6801, 6802, 6823, 6901, 8001 HLA-B: 0801, 0802, 0803, 1501, 1502, 1503, 1509, 1517, 3501, 3503, 3801, 4013, 4506, 4601, 4801, 5101, 5301, 5801, 5802, 7301, 8301 HLA-C: 0303, 0401, 0602, 0701, 0702, 0802, 1203, 1402, 1502	82.1%	24	HLA-DRB1: 0101, 0401, 0402, 0403, 0404, 0405, 0801, 0901, 1001, 1101, 1301, 1602 HLA-DPA10-DPB10: 103-201, 103-401, 103-402, 103-601, 201-101, 201-501, 301-402 HLA-DQA10-DQB10: 101-501, 102-602, 103-603, 501-201, 501-301	0.82	0.96	3.59	N/A	N/A

(*Continued on next page*)

Table 1. *Continued*

Gene Position	Epitope	HLA Class I Population Presentation	HLA Class I Alleles Bound	HLA Class I Binders	HLA Class II Population Presentation	HLA Class II Alleles Bound	HLA Class II Binders	Dissimilarity Score	Conservation Score	Combined T Cell Score	B Cell Total Score	B and T Cell Total Percentile
S_129	KVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYS	98.5%	58	HLA-A: 0101, 0201, 0202, 0206, 0211, 0212, 0216, 0217, 0301, 0302, 1101, 2301, 2402, 2403, 2602, 3001, 3101, 3207, 6601, 6823, 6901, 8001 HLA-B: 0803, 1501, 1502, 1503, 1509, 1517, 1801, 2720, 3501, 3701, 3801, 3901, 4001, 4002, 4013, 4403, 4501, 4506, 4601, 4801, 5801, 7301 HLA-C: 0303, 0401, 0501, 0602, 0701, 0702, 0802, 1203, 1402, 1502	39.0%	9	HLA-DRB1: 0403, 1302, 0405, 0404 HLA-DPA10-DPB10: 103-201, 103-401, 103-601, 301-402 HLA-DQA10-DQB10: 102-602	0.60	0.83	2.80	1.14	91%

(*Continued on next page*)

Table 1. *Continued*

Gene Position	Epitope	HLA Class I Population Presentation	HLA Class I Alleles Bound	HLA Class I Binders	HLA Class II Population Presentation	HLA Class II Alleles Bound	HLA Class II Binders	Dissimilarity Score	Conservation Score	Combined T Cell Score	B Cell Total Score	B and T Cell Total Percentile
S_252	GDSSSGWTAG AAAYYVGYLQP RTFLLKYNENGT	95.7%	53	HLA-A: 0101, 0201, 0202, 0203, 0205, 0206, 0211, 0212, 0216, 0217, 0219, 2403, 2501, 2601, 2602, 2603, 2902, 3002, 3207, 3301, 6601, 6801, 6802, 6823, 6901, 8001 HLA-B: 0801, 0802, 0803, 1402, 1502, 1503, 1517, 3501, 4013, 4501, 4506, 5703, 5801, 8301 HLA-C: 0303, 0401, 0602, 0701, 0702, 0802, 1203, 1402, 1502	68.9%	15	HLA-DRB1: 0101, 0401, 0402, 0404, 0405, 0701, 0901, 1001, 1301, 1501, 1602 HLA-DPA10-DPB10: 103-301, 301-402 HLA-DQA10-DQB10: 102-602, 501-301	0.48	0.71	2.84	0.76	81%
S_462	KPFERDISTEIYQ AGSTPCNGVEG FNCYFPLQS	74.8%	27	A: 0206, 2402, 2403, 3207, 6601, 6802, 6823 B: 0802, 1402, 1502, 1503, 2720, 3503, 4002, 4013, 4201, 4506, 4801, 8301 C: 0401, 0702, 1203, 1402	18.7%	5	DRB1: 0701, 0801, 1101, 1602 DPA10-DPB10: 201-501	0.51	0.77	2.21	1.29	75.2%

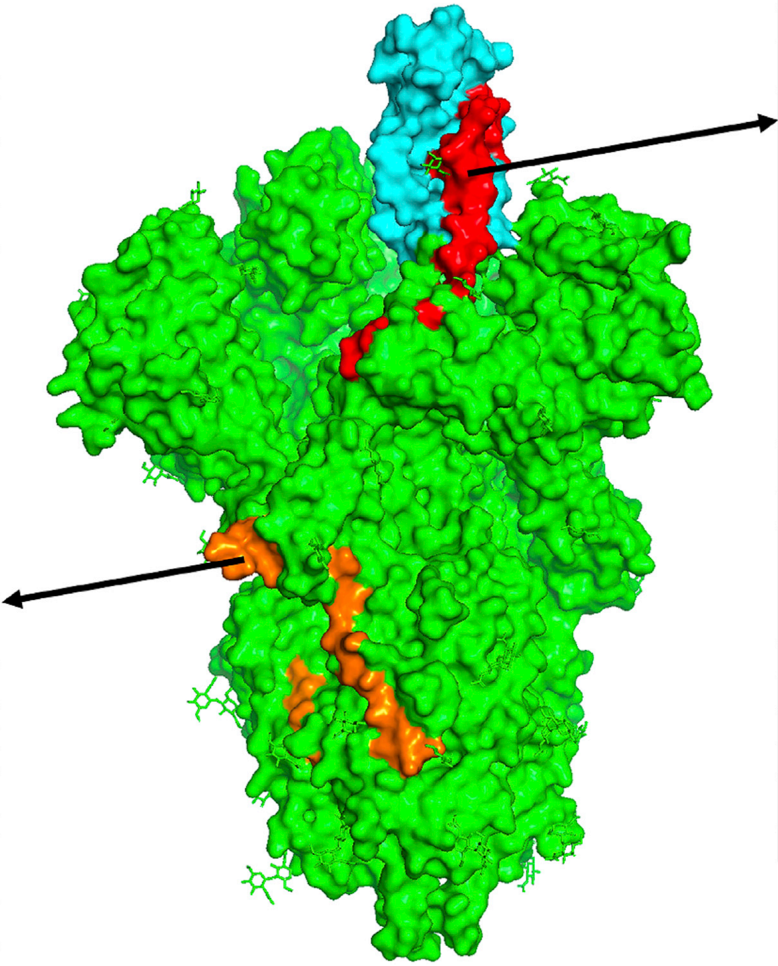
(*Continued on next page*)

Table 1. ***Continued***

Gene Position	Epitope	HLA Class I Population Presentation	HLA Class I Alleles Bound	HLA Class I Binders	HLA Class II Population Presentation	HLA Class II Alleles Bound	HLA Class II Binders	Dissimilarity Score	Conservation Score	Combined T Cell Score	B Cell Total Score	B and T Cell Total Percentile
N_305	AQFAPSASAFFG MSRIGMEVTPS GTWLTYTGAI	87.5%	40	HLA-A: 0202, 0203, 0211, 0212, 0216, 1101, 2403, 2601, 2602, 2603, 3101, 6601, 6801, 6823, 6901, 8001 HLA-B: 0803, 1502, 1503, 1517, 3501, 3503, 4402, 4403, 4506, 4801, 5301, 5703, 8301 HLA-C: 0303, 0401, 0501, 0702, 1203, 1402, 1502	6.4%	1	DRB1: 0901	0.46	0.91	2.31	N/A	N/A

Columns represent gene and position of first amino acid of 33-mer, number of HLA class I and II alleles predicted to bind at least one predicted epitope within 33-mer, list of bound alleles, the proportion of the population predicted to have at least one of these HLAs, normalized dissimilarity scores, normalized conservation scores, across the 33-mer, total T cell score, B cell score, and combined B and T cell percentile for 33-mers. Table includes S_462 in S protein containing novel receptor binding sites[21] and N_305 containing five peptides shown to be immunogenic in IEDB. N/A, not applicable.

Gene Position	S_673
Epitope	SYQTQTNSPRRARSVAS QSIIAYTMSLGAENSV
HLA Class I Population Presentation	94.5%
HLA Class I Alleles Bound	47
HLA Class I binders	A2403, A2603, A6601, B1502, A3101, A3301, B1402, B0702, B8301, A3001, B2720, C0602, C0701, B1517, B4801, B5801, C1502, B1503, B3501, B4601, C1203, B3901, B4013, A0201, A0202, A0206, A0217, A2501, A2601, A2602, A2603, A3201, A3207, A6901, B0801, B3901, B4506, C1402, A6823, C1502, B3501, B4402, B4403, B4501, A3201, A6802, C1203
HLA Class II Population Presentation	40.8%
HLA Class II Alleles Bound	1
HLA Class II binders	DQA10501-DQB10301
Dissimilarity Score	0.60
Conservation Score	0.85
Combined T Cell Score	2.80
B Cell Linear 33mer	0.99
B Cell Conformational 33mer	0.46
B and T Cell Total Percentile	100%



Gene Position	S_339
Epitope	GEVFNATRFASVYAWNRKRI SNCVADYSVLYNS
HLA Class I Population Presentation	98.3%
HLA Class I Alleles Bound	55
HLA Class I binders	B4001, B4402, B4403, A6601, A6802, A6901, A0205, B0801, B0802, B1402, C1203, C1502, B3501, A3001, A3207, B2720, B3801, B3901, B7301, C0602, C0701, C0702, A3301, A6801, A1101, A3001, A0302, A2603, A3101, A3301, B4506, C0401, A2402, C0602, B2720, A3002, A8001, A0101, A2601, A2602, A2902, B3501, C0501, A2501, A3201, A3207, B1502, B1503, B1517, C1402, A2403, A6601, B0803, A0301, A1101
HLA Class II Population Presentation	36.5%
HLA Class II Alleles Bound	8
HLA Class II binders	DRB1_0402, DRB1_0405, DRB1_1501, DPA10103-DPB10301, DPA10103-DPB10401, DPA10103-DPB10402, DPA10201-DPB10101, DPA10301-DPB10402
Dissimilarity Score	0.69
Conservation Score	0.77
Combined T Cell Score	2.81
B Cell Linear 33mer	0.83
B Cell Conformational 33mer	0.59
B and T Cell Total Percentile	99%

Figure 2. Proposed Vaccine Epitopes in SARS-CoV-2 Spike Protein

Crystal structure of SARS-CoV-2 Spike protein trimer (PDB: 6VYB) with two highlighted vaccine epitopes targeting newly evolved acquired viral vulnerabilities. First, SARS-CoV-2 receptor binding domain (cyan) has up to 10-fold higher affinity binding to the ACE2 receptor as compared with previous coronaviruses. Using our analysis, we identify a high-ranking vaccine epitope (red) within the receptor binding domain. Second, SARS-CoV-2 has acquired a novel furin cleavage site RRAR, along for increased infectivity due to improved membrane fusion (epitope containing the novel furin cleavage site highlighted in orange).

containing each of these residues (hotspots underlined; Table 1). Additionally, a D614G mutation in the S protein has been reported as a potentially dominant strain with increased pathogenicity.[25,26] We thus suggest including the highest scoring 33-mer (NTSNQVAVLYQ**G**VNCTEVPVAIHADQLTPTWRV) predicted to present this mutant epitope in a vaccine construct. Finally, it is known that mRNA transcripts proximal to the 3′ end of the *Coronaviridae* family genome show higher abundance consistent with the viral replication process, with S, E, M, and N genes shown to have significantly higher translational efficiency compared with the 5′ transcripts, with the highest expression in the N gene, and consistent with the high degree of MHC presentation as described above for the five immunogenic peptides derived from a single N protein 33-mer.[27–29] We therefore posit that viral epitopes derived from the 3′ terminus, including the S, E, M, and N genes, will have a higher representation on MHC and suggest their prioritization in a vaccine construct. Table S5 lists the highest priority viral peptides we suggest should be considered for inclusion in vaccine constructs.

DISCUSSION

Here we present a comprehensive immunogenicity map of the SARS-CoV-2 virus (Table S1) and propose sixty-five 33-mer peptide sequences predicted to generate B and T cell epitopes from a diverse sampling of viral domains across all 10 SARS-CoV-2 genes (Tables 1 and S5). Based on our computational algorithms, we expect that the highest scoring peptides will result in safe and immunogenic T cell epitopes, and that B cell epitopes should be evaluated for safety and efficacy using previously reported methods with validated subsets of these 65 epitopes.[12] DNA and mRNA vaccines have been shown to be safe and effective in preclinical studies, and can be rapidly and efficiently manufactured at scale.[30,31] Nucleoside modification of RNA has been shown to improve efficacy, which has been attributed to a reduction of RNA-induced immunogenicity.[32] We suggest that multivalent constructs composed of the SARS-CoV-2 minigenes encoding subsets of the B and/or T cell epitopes proposed here (Tables 1, S3, S4, and S5) can be used in a DNA on mRNA vaccine for expression in antigen-presenting cells.

These epitopes can be used in tandem with a Toll-like receptor (TLR) agonist, such as tetanus toxoid or PADRE,[33–36] to drive activation of signals 1 and 2 in antigen-presenting cells. Constructs can be designed to contain a combination of optimal B and/or T cell epitopes, or deployed as a construct consisting of the top scoring T cell epitopes to be used in combination with the vaccines currently being developed targeting S protein in order to drive the adaptive memory response. DNA vaccine sequences can also be codon optimized to increase CpG islands, such as to increase TLR9 activation.[37]

With the third epidemic in the past two decades underway, all originating from the coronavirus family, these viruses will continue to threaten the human population, which necessitates the need for prophylactic measures against future outbreaks. The methods described here provide a rapid workflow for evaluating and prioritizing safe and immunogenic regions of a viral genome for use in vaccination. A subset of the epitopes selected here are derived from viral regions sharing a high degree of homology with other viruses in the family, and thus we expect these evolutionarily conserved regions to be essential in the infectivity and replicative life cycle across the coronavirus family. This suggests that an immune response against the aforementioned epitopes listed herein may provide more broadly protective immunity against mutated strains of SARS-CoV-2 and other coronaviruses. Additionally, we describe epitopes containing the newly acquired features of SARS-CoV-2 that confer evolutionary advantages in viral spread and infectivity. The immunogenicity map provided in Table S1 can be used to design customized multi-valent vaccines based on the HLA frequencies of specific populations. Although we suggest the use of 33-mers based on optimal MHC presentation across the population, these methods can be generalized and applied to the evaluation k-mers of various sizes depending on desired application. Because antigens may arise from the junctions between epitopes, the analyses presented here can also be used to evaluate epitope generation at the junction of specific vaccine constructs, such as to engineer linker regions that reduce the potential immunodominant epitopes elicited from irrelevant sequences.

Previous analyses of SARS-CoV-2 have predicted immunogenic epitopes based on previously reported epitopes in IEDB, sequence homology, and MHC binding predictions.[38,39] Ahmed et al.[38] present initial insight to potential SARS-CoV-2 epitopes by comparing previously detected epitopes derived from SARS-CoV. Grifoni et al.[39] extended these findings by assessing sequence homology between three host species of *betacoronaviruses* and available human strains, and performing B and T cell epitope predictions. Our analysis performed at the scale of 33-mer epitopes includes the addition of dissimilarity scoring, expands the homology search across 14 species of coronavirus and 727 SARS-CoV-2 genes sequences, and covers a wider diversity of HLA coverage across the population. We searched for peptides predicted by both groups contained within our selected epitopes, finding 27 of 100 peptides reported by Ahmed et al.[38] and 187 out of 905 peptides reported by Grifoni et al.[39] within the sixty-five 33-mers we report. We also find up to five peptides reported by Grifoni et al.[39] within a single 33-mer and up to 12 peptides reported by Ahmed et al.[38] contained in the 33-mer AQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAI described above. Taken together, these comparisons show a significant convergence on a subset of epitopes using agnostic analyses, while also reporting unique epitopes in each study. The finding that up to 12 epitopes from previous analyses are represented in a single 33-mer from our agnostic analysis further supports our prediction that cocktails of 33-mer epitopes can be used for population-scale vaccination.

By narrowing the pool of peptides selected for downstream screening, we expect that the analyses presented here will contribute to maximizing the efficiency of vaccine development. Antigenic burden from epitopes that do not contribute to viral protection can cause autoimmune reactions, reactogenicity, detraction from the efficacy of the vaccine, or result in ADE. We found that the vast majority of the SARS-CoV-2 virus is immunogenically silent on MHC class I and II and suggest these regions should be excluded from vaccine development. Although empirical testing is necessary to evaluate ADE, we suggest that antibodies directed at the RBD and furin cleavage sequences may mitigate ADE by blocking the processes needed to achieve membrane fusion. To avoid potential T cell cross-reactivities *a priori*, we selected maximally immunogenic epitopes with the highest degree of dissimilarity to the self-proteome with minimal potential of cross-reactivity that can lead to adverse reaction or weaken the efficacy of vaccination. In addition to the predicted safety of these epitopes (stemming from lack of potentially cross-reactive normal proteins), we expect that a greater repertoire of viral antigen-specific T cells will be present because of the absence of negative thymic selection. Although we prioritize epitopes with maximal dissimilarity from the human proteome, many other SARS-CoV-2 peptides show identical or nearly identical peptides presented on MHC derived from normal proteins. This implies that the inclusion of these highly similar epitopes in a vaccine could result in cross-reactive binding and potentially result in autoimmune responses.

Previously, it has been demonstrated that immunity acting through CD8 cells alone is sufficient in ameliorating infection, as demonstrated in studies showing that CD8-mediated vaccination is protective against influenza challenge in mice replete of antibodies and B cells,[40] and by human CD8 cells shown to be protective across multiple influenza strains.[41] CD8-based vaccine approaches have been shown to be particularly protective against intranasal viral transmission,[42] suggesting that nasal protection through CD8 vaccination may be relevant to SARS-CoV-2 transmission based on recent reports of ACE2 and TMPRSS2 co-expression in nasal epithelium[7] and clinical reports of SARS-CoV-2 infection in the olfactory bulb and symptoms of anosmia.[43,44] Although CD8 vaccines targeting conserved antigens in influenza did not completely block infection upon challenge with virus, they effectively reduced viral replication, morbidity, and mortality.[41,42] Taken together, these findings suggest that CD8-based immunity can be a viable strategy in quelling SARS-CoV-2. Studies demonstrating protection against multiple influenza strains imply that CD8-mediated vaccination may act more broadly than antibody responses in protecting against multiple virus family members through targeting of conserved non-structural proteins critical in the viral life cycle.

Currently, targeting CD8 epitopes has been complicated by HLA restriction of peptides and antigenic drift resulting from viral regions in which mutation is tolerable. We propose that a vaccine designed to induce CD8 responses across multiple HLA alleles covering large proportions of the population and targeting conserved regions of the virus that are highly dissimilar from the human self-peptidome can provide a safe vaccination strategy that can be rapidly tested for use alone or in combination with antibody-based vaccines in development. For example, the 33-mer ISNSWLMWLIINLVQMAPISAMVRMYIFFASFY contains epitopes predicted to be present in 99.4% of the population, scores in the 99.8th percentile in dissimilarity to the human proteome, and in the 79.3rd percentile in conservation. This 33-mer is derived from the most conserved region of the virus, ORF1ab, and encodes the NSP3 protein, which is critical to viral replication.[45] These results imply that a CD8-based vaccine including such 33-mers could induce population-scale protection targeting a critical non-structural protein and circumvent safety concerns of ADE, potentially accelerating safe vaccine development.

Although the epitopes presented here are based on computational predictions (which do not account for the multiple steps involved in antigen processing and presentation), our previous validation of peptide presentation using liquid chromatography-tandem mass spectrometry (LC-MS/MS) of peptides eluted from MHC across multiple tumors showed highly significant concordance with predicted population-scale presentation.[17] Although we expect a significant fraction of predicted antigens to be presented on MHC, binding predictions alone do not determine which antigens will elicit an immunodominant response. Although the dissimilarity scoring predicts that TCRs specific for these antigens are more likely to exist (because these TCRs are far less likely to have undergone negative thymic selection), these predictions are confounded by the TCR repertoire of a given individual and the intrinsic immunogenicity of a particular peptide, which cannot be predicted without empirical testing. Because MHC binding is a prerequisite for antigen immunogenicity, we expect that immunodominant antigens will be contained within our highest scoring epitopes. However, experimental validation will be necessary to determine the contribution of individual antigens to immunity. As a best approximation for our predictions, we show a significant enrichment of peptides previously reported in IEDB to be immunogenic in the SARS-CoV virus, contained within the 65 prioritized epitopes that we present, supporting the concept that multiple antigens derived from 33-mers can be presented across multiple HLA alleles.

We expect that the comprehensive immunogenicity map presented here can be used by the scientific community to inform the design of various vaccination modalities. We are presently designing a set of vaccine vectors and validation reagents based on these analyses that we plan to make available to the research community for testing. The 65 epitopes presented here out of the 9,303 possible 33-mers derived from SARS-CoV-2 can significantly narrow the focus of vaccine development (Table S5); these epitopes can be expressed as a single <7-kb construct, or more likely tested in various combinations delivered as a cocktail of RNA constructs encoding individual 33-mers. These vaccine constructs can be rapidly and efficiently tested for the neutralizing potential of antibodies using SARS-CoV-2 pseudovirus,[46] the formation of memory B cells, and induction of T cell activation using methods that we have recently developed for interrogating antigen specificities in a highly multiplexed manner.[47] Because SARS-CoV-2 has precipitated the need to rapidly develop and deploy vaccines in pandemic situations,[48] we suggest that this comprehensive analysis can be incorporated into a process that can be rapidly implemented when future novel viral pathogens emerge.

Limitations of Study

The *in silico* analysis of the SARS-CoV-2 genome reported here has yet to be experimentally validated. Although it is reassuring that we demonstrate enrichment of predicted epitopes from the original SARS virus previously reported in IEDB that have been shown to be immunogenic, rigorous experimental validation of our findings is required. Computational peptide MHC binding predictions do not consider critical variables in antigen presentation, such as proteasomal degradation and peptide processing. In addition, it is unclear whether the 33-mers designed to elicit a B cell will properly fold into conformations resembling the native S protein, such as to elicit a protective antibody response. We have designed multiple DNA and mRNA constructs containing combinations of 33-mers proposed here to test hypotheses that these vaccines can elicit memory and/or cytolytic T cell response and/or protective antibodies against a SARS-S-GFP pseudovirus[46] in HLA-A2 transgenic mice.[49] Construct designs are available upon request.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- METHOD DETAILS
  - ○ Population-scale HLA Class I & II Presentation
  - ○ Conservation Scoring
  - ○ Dissimilarity Scoring
  - ○ B cell Epitope Scoring

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.xcrm.2020.100036.

ACKNOWLEDGMENTS

This work was supported by a St. Baldrick's-Stand Up to Cancer Pediatric Dream Team Translational Research Grant (SU2C-AACR-DT2727 to J.M.M.) and the Beau Biden Cancer Moonshot Pediatric Immunotherapy Discovery and Development Network (NCI grant U54 CA232568 to J.M.M., M.Y., and J.M.W.). Stand Up to Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. This work was also supported by the NIH (grant R35 CA220500 to J.M.M. and A.F.), the Giulio D'Angio Endowed Chair (J.M.M.), and the Quod Erat Demonstrandum (QED) program at the Philadelphia Science Center (J.M.M. and M.Y.). Our graphical abstract was created with BioRender.com. The authors thank Drs. Paul Offit, Paul Sondel, and Crystal Mackall for their insightful discussions related to this project.

AUTHOR CONTRIBUTIONS

Conceptualization, M.Y.; Methodology, M.Y., J.M.W., and A.F.; Software, M.Y., J.M.W., and A.F.; Formal Analysis, M.Y., J.M.W., and A.F.; Investigation, M.Y., J.M.W., and A.F.; Funding Acquisition, M.Y. and J.M.M.; Writing, M.Y., J.M.W., and J.M.M.; Supervision, M.Y. and J.M.M.

DECLARATION OF INTERESTS

A provisional patent has been filed on the methods and composition of the SARS-CoV-2 peptides reported here.

Received: April 7, 2020
Revised: April 29, 2020
Accepted: May 27, 2020
Published: June 8, 2020

REFERENCES

1. Papaneri, A.B., Johnson, R.F., Wada, J., Bollinger, L., Jahrling, P.B., and Kuhn, J.H. (2015). Middle East respiratory syndrome: obstacles and prospects for vaccine development. Expert Rev. Vaccines *14*, 949–962.
2. Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260–1263.
3. Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell *181*, 281–292.e6.
4. Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N.G., and Decroly, E. (2020). The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. *176*, 104742.
5. Chen, J., Lee, K.H., Steinhauer, D.A., Stevens, D.J., Skehel, J.J., and Wiley, D.C. (1998). Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and the origin of the labile conformation. Cell *95*, 409–417.
6. Belouzard, S., Chu, V.C., and Whittaker, G.R. (2009). Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc. Natl. Acad. Sci. USA *106*, 5871–5876.
7. Ziegler, C.G.K., Allon, S.J., Nyquist, S.K., Mbano, I.M., Miao, V.N., Tzouanas, C.N., Cao, Y., Yousif, A.S., Bals, J., Hauser, B.M., et al. (2020). SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. Cell *181*, 1016–1035.
8. Zhang, H., Zhou, P., Wei, Y., Yue, H., Wang, Y., Hu, M., Zhang, S., Cao, T., Yang, C., Li, M., et al. (2020). Histopathologic Changes and SARS-CoV-2 Immunostaining in the Lung of a Patient With COVID-19. Ann. Intern. Med. *172*, 629–632.
9. Xu, H., Zhong, L., Deng, J., Peng, J., Dan, H., Zeng, X., Li, T., and Chen, Q. (2020). High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. Int. J. Oral Sci. *12*, 8.
10. Thevarajan, I., Nguyen, T.H.O., Koutsakos, M., Druce, J., Caly, L., van de Sandt, C.E., Jia, X., Nicholson, S., Catton, M., Cowie, B., et al. (2020). Breadth of concomitant immune responses prior to patient recovery: a case report of non-severe COVID-19. Nat. Med. *26*, 453–455.
11. Dejnirattisai, W., Supasa, P., Wongwiwat, W., Rouvinski, A., Barba-Spaeth, G., Duangchinda, T., Sakuntabhai, A., Cao-Lormeau, V.M., Malasit, P., Rey, F.A., et al. (2016). Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. Nat. Immunol. *17*, 1102–1108.
12. Wang, Q., Zhang, L., Kuwahara, K., Li, L., Liu, Z., Li, T., Zhu, H., Liu, J., Xu, Y., Xie, J., et al. (2016). Immunodominant SARS Coronavirus Epitopes in Humans Elicited both Enhancing and Neutralizing Effects on Infection in Non-human Primates. ACS Infect. Dis. *2*, 361–376.
13. Wan, Y., Shang, J., Sun, S., Tai, W., Chen, J., Geng, Q., He, L., Chen, Y., Wu, J., Shi, Z., et al. (2020). Molecular Mechanism for Antibody-Dependent Enhancement of Coronavirus Entry. J. Virol. *94*, e02015–e02019.
14. Tetro, J.A. (2020). Is COVID-19 receiving ADE from other coronaviruses? Microbes Infect. *22*, 72–73.
15. Alspach, E., Lussier, D.M., Miceli, A.P., Kizhvatov, I., DuPage, M., Luoma, A.M., Meng, W., Lichti, C.F., Esaulova, E., Vomund, A.N., et al. (2019). MHC-II neoantigens shape tumour immunity and response to immunotherapy. Nature *574*, 696–701.
16. McHeyzer-Williams, M., Okitsu, S., Wang, N., and McHeyzer-Williams, L. (2011). Molecular programming of B cell memory. Nat. Rev. Immunol. *12*, 24–34.
17. Yarmarkovich, M., Farrel, A., Sison, A., 3rd, di Marco, M., Raman, P., Parris, J.L., Monos, D., Lee, H., Stevanovic, S., and Maris, J.M. (2020). Immunogenicity and Immune Silence in Human Cancer. Front. Immunol. *11*, 69.
18. Gragert, L., Madbouly, A., Freeman, J., and Maiers, M. (2013). Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. Hum. Immunol. *74*, 1313–1320.
19. An, L.-L., Rodriguez, F., Harkins, S., Zhang, J., and Whitton, J.L. (2000). Quantitative and qualitative analyses of the immune responses induced by a multivalent minigene DNA vaccine. Vaccine *18*, 2132–2141.
20. Lu, Y.-C., Yao, X., Crystal, J.S., Li, Y.F., El-Gamil, M., Gross, C., Davis, L., Dudley, M.E., Yang, J.C., Samuels, Y., et al. (2014). Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions. Clin. Cancer Res. *20*, 3401–3410.
21. Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., and Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. Nature *581*, 221–224.
22. Jespersen, M.C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. *45* (*W1*), W24–W29.
23. Kringelum, J.V., Lundegaard, C., Lund, O., and Nielsen, M. (2012). Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput. Biol. *8*, e1002829.
24. Nerli, S., and Sgourakis, N.G. (2020). Structure-based modeling of SARS-CoV-2 peptide/HLA-A02 antigens. bioRxiv. https://doi.org/10.1101/2020.03.23.004176.
25. Becerra-Flores, M., and Cardozo, T. (2020). SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int. J. Clin. Pract. Published online May 6, 2020. https://doi.org/10.1111/ijcp.13525.
26. Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Foley, B., Giorgi, E.E., Bhattacharya, T., Parker, M.D., et al. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. https://doi.org/10.1101/2020.04.29.069054.
27. Irigoyen, N., Firth, A.E., Jones, J.D., Chung, B.Y., Siddell, S.G., and Brierley, I. (2016). High-resolution analysis of coronavirus gene expression by RNA sequencing and ribosome profiling. PLoS Pathog. *12*, e1005473.
28. Cheng, V.C.C., Lau, S.K.P., Woo, P.C.Y., and Yuen, K.Y. (2007). Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. Clin. Microbiol. Rev. *20*, 660–694.

29. Hiscox, J.A., Cavanagh, D., and Britton, P. (1995). Quantification of individual subgenomic mRNA species during replication of the coronavirus transmissible gastroenteritis virus. Virus Res. *36*, 119–130.

30. Dowd, K.A., Ko, S.-Y., Morabito, K.M., Yang, E.S., Pelc, R.S., DeMaso, C.R., Castilho, L.R., Abbink, P., Boyd, M., Nityanandam, R., et al. (2016). Rapid development of a DNA vaccine for Zika virus. Science *354*, 237–240.

31. Richner, J.M., Himansu, S., Dowd, K.A., Butler, S.L., Salazar, V., Fox, J.M., Julander, J.G., Tang, W.W., Shresta, S., Pierson, T.C., et al. (2017). Modified mRNA Vaccines Protect against Zika Virus Infection. Cell *168*, 1114–1125.e10.

32. Pardi, N., Hogan, M.J., Pelc, R.S., Muramatsu, H., Andersen, H., DeMaso, C.R., Dowd, K.A., Sutherland, L.L., Scearce, R.M., Parks, R., et al. (2017). Zika virus protection by a single low-dose nucleoside-modified mRNA vaccination. Nature *543*, 248–251.

33. Alexander, J., Fikes, J., Hoffman, S., Franke, E., Sacci, J., Appella, E., Chisari, F.V., Guidotti, L.G., Chesnut, R.W., Livingston, B., and Sette, A. (1998). The optimization of helper T lymphocyte (HTL) function in vaccine development. Immunol. Res. *18*, 79–92.

34. Hung, C.-F., Tsai, Y.-C., He, L., and Wu, T.C. (2007). DNA vaccines encoding Ii-PADRE generates potent PADRE-specific CD4+ T-cell immune responses and enhances vaccine potency. Mol. Ther. *15*, 1211–1219.

35. Zanetti, B.F., Ferreira, C.P., de Vasconcelos, J.R.C., and Han, S.W. (2019). scFv6.C4 DNA vaccine with fragment C of Tetanus toxin increases protective immunity against CEA-expressing tumor. Gene Ther. *26*, 441–454.

36. La Rosa, C., Longmate, J., Lacey, S.F., Kaltcheva, T., Sharan, R., Marsano, D., Kwon, P., Drake, J., Williams, B., Denison, S., et al. (2012). Clinical evaluation of safety and immunogenicity of PADRE-cytomegalovirus (CMV) and tetanus-CMV fusion peptide vaccines with or without PF03512676 adjuvant. J. Infect. Dis. *205*, 1294–1304.

37. Krieg, A.M. (2008). Toll-like receptor 9 (TLR9) agonists in the treatment of cancer. Oncogene *27*, 161–167.

38. Ahmed, S.F., Quadeer, A.A., and McKay, M.R. (2020). Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. Viruses *12*, 254.

39. Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., and Sette, A. (2020). A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe *27*, 671–680.e2.

40. Epstein, S.L., Lo, C.-Y., Misplon, J.A., and Bennink, J.R. (1998). Mechanism of protective immunity against influenza virus infection in mice without antibodies. J. Immunol. *160*, 322–327.

41. Koutsakos, M., Illing, P.T., Nguyen, T.H.O., Mifsud, N.A., Crawford, J.C., Rizzetto, S., Eltahla, A.A., Clemens, E.B., Sant, S., Chua, B.Y., et al. (2019). Human CD8$^+$ T cell cross-reactivity across influenza A, B and C viruses. Nat. Immunol. *20*, 613–625.

42. Price, G.E., Lo, C.-Y., Misplon, J.A., and Epstein, S.L. (2018). Reduction of influenza virus transmission from mice immunized against conserved viral antigens is influenced by route of immunization and choice of vaccine antigen. Vaccine *36* (*32 Pt B*), 4910–4918.

43. Galougahi, M.K., Ghorbani, J., Bakhshayeshkaram, M., Naeini, A.S., and Haseli, S. (2020). Olfactory bulb magnetic resonance imaging in SARS-CoV-2-induced anosmia: the first report. Acad. Radiol. *27*, 892–893.

44. Xydakis, M.S., Dehgani-Mobaraki, P., Holbrook, E.H., Geisthoff, U.W., Bauer, C., Hautefort, C., Herman, P., Manley, G.T., Lyon, D.M., and Hopkins, C. (2020). Smell and taste dysfunction in patients with COVID-19. Lancet Infect. Dis. Published online April 15, 2020. https://doi.org/10.1016/S1473-3099(20)30293-0.

45. Lei, J., Kusov, Y., and Hilgenfeld, R. (2018). Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral Res. *149*, 58–74.

46. Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., et al. (2020). Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat. Commun. *11*, 1620.

47. Overall, S.A., Toor, J.S., Hao, S., Yarmarkovich, M., O'Rourke, S.M., Morozov, G.I., Nguyen, S., Japp, A.S., Gonzalez, N., Moschidi, D., et al. (2020). High throughput pMHC-I tetramer library production using chaperone-mediated peptide exchange. Nat. Commun. *11*, 1909.

48. Lurie, N., Saville, M., Hatchett, R., and Halton, J. (2020). Developing Covid-19 Vaccines at Pandemic Speed. N. Engl. J. Med. *382*, 1969–1973.

49. Botten, J., Whitton, J.L., Barrowman, P., Sidney, J., Whitmire, J.K., Alexander, J., Kotturi, M.F., Sette, A., and Buchmeier, M.J. (2010). A multivalent vaccination strategy for the prevention of Old World arenavirus infection in humans. J. Virol. *84*, 9947–9956.

50. Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511–517.

51. Jensen, K.K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J.A., Yan, Z., Sette, A., Peters, B., and Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology *154*, 394–406.

52. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

53. Sidney, J., Steen, A., Moore, C., Ngo, S., Chung, J., Peters, B., and Sette, A. (2010). Divergent motifs but overlapping binding repertoires of six HLA-DQ molecules frequently expressed in the worldwide human population. J. Immunol. *185*, 4189–4198.

54. Solberg, O.D., Mack, S.J., Lancaster, A.K., Single, R.M., Tsai, Y., Sanchez-Mazas, A., and Thomson, G. (2008). Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum. Immunol. *69*, 443–464.

55. Gras, S., Saulquin, X., Reiser, J.-B., Debeaupuis, E., Echasserieau, K., Kissenpfennig, A., Legoux, F., Chouquet, A., Le Gorrec, M., Machillot, P., et al. (2009). Structural bases for the affinity-driven selection of a public TCR against a dominant human cytomegalovirus epitope. J. Immunol. *183*, 430–437.

56. Ishizuka, J., Stewart-Jones, G.B.E., van der Merwe, A., Bell, J.I., McMichael, A.J., and Jones, E.Y. (2008). The structural dynamics and energetics of an immunodominant T cell receptor are programmed by its Vbeta domain. Immunity *28*, 171–182.

57. Gagnon, S.J., Borbulevych, O.Y., Davis-Harrison, R.L., Baxter, T.K., Clemens, J.R., Armstrong, K.M., Turner, R.V., Damirjian, M., Biddison, W.E., and Baker, B.M. (2005). Unraveling a hotspot for TCR recognition on HLA-A2: evidence against the existence of peptide-independent TCR binding determinants. J. Mol. Biol. *353*, 556–573.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
SARS-CoV-2 Immunogenicity Map	This manuscript	Table S1
Recombinant DNA		
Vaccine constructs	This manuscript	N/A
Software and Algorithms		
netMHC 4.0	Andreatta and Nielsen[50]	http://www.cbs.dtu.dk/services/NetMHC-4.0
netMHCII 2.3	Jensen et al.[51]	https://services.healthtech.dtu.dk/service.php?NetMHCII-2.3
BepiPred 2.0	Jespersen et al.[22]	https://services.healthtech.dtu.dk/service.php?BepiPred-2.0
DiscoTope 2.0	Kringelum et al.[23]	https://services.healthtech.dtu.dk/service.php?DiscoTope-2.0
shinyNAP	Yarmarkovich et al.[17]	https://www.frontiersin.org/article/10.3389/fimmu.2020.00069/full
Clustal Omega	Sievers et al.[52]	https://www.ebi.ac.uk/Tools/msa/clustalo/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, John M. Maris (maris@email.chop.edu).

Materials Availability

Vaccine constructs and testing reagents are available from the Lead Contact, John M. Maris with a completed Materials Transfer Agreement. Please email maris@chop.edu.

Data and Code Availability

All raw data has been reported in paper and models are described in STAR Methods.

METHOD DETAILS

Population-scale HLA Class I & II Presentation

We identified potential SARS-CoV-2 epitopes by applying our recently published algorithm for scoring population-scale HLA presentation of tumor driver gene, to the SARS-CoV-2 genome (GenBank Acc#: MN908947.3).[17] All possible 33-mer amino acid sequences covering every 9-mer peptide from the 10 SARS-CoV-2 genes were generated and we employed netMHC-4.0 to predict the binding affinities of each viral 9-mer peptide across 84 HLA class I alleles.[50] We considered 9-mer peptides with binding affinities $<$ 500nM putative epitopes. MHC class II binding affinities were predicted as described above across 36 HLA class II alleles population using netMHCII 2.3.[51] All 9mers present in a 33-mer contribute to the score. 33-mer scores calculated by infering population scale hla presentation of all predicted peptides within 9-mer on class I and ii.

The frequencies of HLA class I alleles -A/B/C and HLA class II alleles -DRB1/3/4/5 were obtained from Be the Match bone marrow registry.[18] HLA class II alleles -DQA1/DQB1 and -DPA1/DPB1 were obtained from [53] and [54], respectively.

Conservation Scoring

We obtained all 727 unique protein sequences categorized by each of the 10 SARS-CoV-2 genes available from the NCBI as of 25 March 2020. All sequences were aligned using Clustal Omega[52] and each position summed for homology. In addition to human sequences, we scored each amino acid position for homology across 15 species of related coronavirus found in bats, pigs, camels, mice, and humans (SARS-CoV, SARS-CoV-2, and MERS). Each amino acid was scored up to 100% conservation. 33-mer peptides were then scored in Equation 1:

$$C = \frac{\sum_{1}^{33} A_i - \mathrm{Y}}{Z - Y} \qquad [1]$$

Where C is the 33-mer conservation score, A is the conservation percentage of an amino acid position, Y is the minimum 33-mer conservation percentage sum, and Z is the maximum 33-mer conservation percentage sum. In the same way, we ranked the conservation across 274 SARS-CoV-2 amino acid sequences available at the time of this study. A final conservation score was generated by averaging the conservation scores from cross-species and interhuman variation and 33-mer peptides with the highest score were considered the most conserved.

Dissimilarity Scoring

3,524 viral epitopes were compared against the normal human proteome on each of their MHC binding partners, testing a total of 12, 383 peptide/MHC pairs against the entire human proteome (85,915,364 normal peptides across HLAs), assigning a similarity score for each peptide. Residues in the same position of the viral and human peptides with a perfect match, similar amino acid classification, or different polarity, were assigned scores of five, two, or negative two respectively. Similarity scores were calculated based on amino acid classification and hydrophobicity were determined using non-anchor residues on MHC (Figure S1A). The canonical TCR-interaction hotspots (residues four through six) were double weighted.[55–57] The similarity scores generated for each viral peptide were converted to Z-scores and peptides with a $p < 0.0001$ were selected for comparison to viral epitopes (Figure S1B). The overall dissimilarity score for the viral peptide was then calculated using Equation 2:

$$S_{Sim} = Z_{Max} - \left(Z_{Top} + \frac{N_{Sig}}{1000} \frac{\overline{Z_{Sig}}}{Z_{max}} \right) \qquad [2]$$

where S_{Sim} is the overall dissimilarity score for the viral peptide, Z_{Max} is the highest possible Z-score given a perfect sequence match to the viral peptide, Z_{Top} is the highest Z-score from the human proteome, N_{Sig} is the number of statistically significant peptides from the human proteome, and $\overline{Z_{Sig}}$ is the mean Z-score from the statistically significant peptides given a $p < 0.001$.

B cell Epitope Scoring

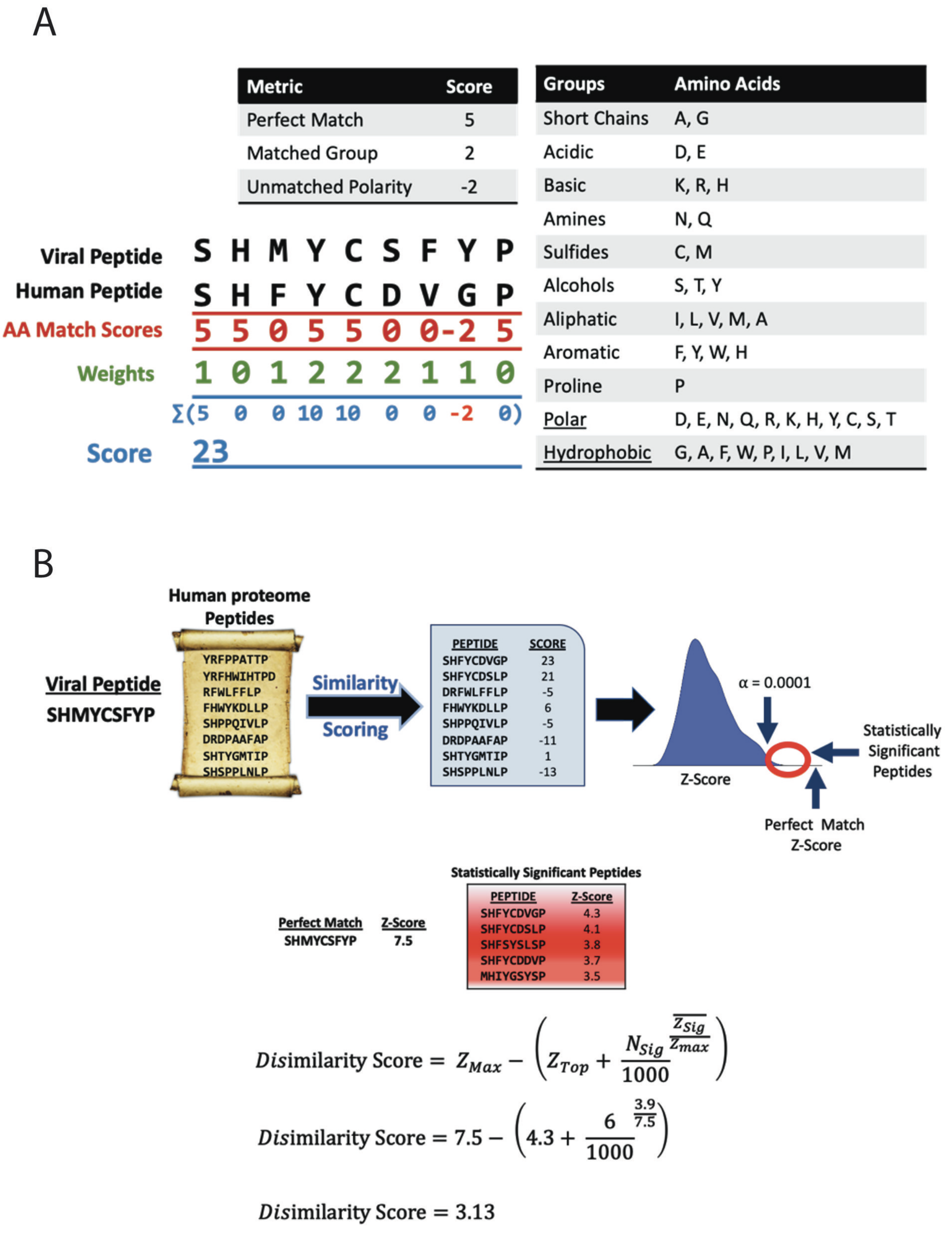
We used BepiPred 2.0 and DiscoTope 2.0[22,23] to score individual amino acid residues, assessing linear epitopes in Matrix, Envelope, and Spike proteins, and conformational epitopes for Spike protein, based on published structure (PDB 6VYB). To we summed and normalized linear and conformational, using separate normalizations for proteins in which only linear predictions were available.

Cell Reports Medicine, Volume *1*

Supplemental Information

Identification of SARS-CoV-2 Vaccine Epitopes Predicted to Induce Long-Term Population-Scale Immunity

Mark Yarmarkovich, John M. Warrington, Alvin Farrel, and John M. Maris



Metric	Score
Perfect Match	5
Matched Group	2
Unmatched Polarity	-2

Groups	Amino Acids
Short Chains	A, G
Acidic	D, E
Basic	K, R, H
Amines	N, Q
Sulfides	C, M
Alcohols	S, T, Y
Aliphatic	I, L, V, M, A
Aromatic	F, Y, W, H
Proline	P
Polar	D, E, N, Q, R, K, H, Y, C, S, T
Hydrophobic	G, A, F, W, P, I, L, V, M

PEPTIDE	SCORE
SHFYCDVGP	23
SHFYCDSLP	21
DRFWLFFLP	-5
FHWYKDLLP	6
SHPPQIVLP	-5
DRDPAAFAP	-11
SHTYGMTIP	1
SHSPPLNLP	-13

Perfect Match	Z-Score
SHMYCSFYP	7.5

Statistically Significant Peptides

PEPTIDE	Z-Score
SHFYCDVGP	4.3
SHFYCDSLP	4.1
SHFSYSLSP	3.8
SHFYCDDVP	3.7
MHIYGSYSP	3.5

$$Disimilarity\ Score = Z_{Max} - \left(Z_{Top} + \frac{N_{Sig}^{\frac{\overline{Z_{Sig}}}{Z_{max}}}}{1000} \right)$$

$$Disimilarity\ Score = 7.5 - \left(4.3 + \frac{6^{\frac{3.9}{7.5}}}{1000} \right)$$

$$Disimilarity\ Score = 3.13$$

Figure S1. Dissimilarity Scoring, related to STAR Methods Dissimilarity Scoring. A) 3,524 viral epitopes (12,383 total peptide/MHC pairs) were compared against the normal human proteome. Non-anchor residues were used to calculate similarity scores based on amino acid classifications as described in methods. Residues in the same position of the viral and human peptides with a perfect match, similar amino acid classification, or different polarity, were assigned scores of five, two, or negative two, respectively. B) Each viral peptide/HLA pair was compared against the set of normal peptides presented on the same MHC. Dissimilarity score for each viral peptide was calculated by comparing against the most similar group of peptides with p < 0.0001 and reported as the difference in Z-scores between the viral peptide and closest-scoring peptides.